Journal of
Marine Science and Engineering

MDPI

Article

DA-YOLOv7: A Deep Learning-Driven High-Performance Underwater Sonar Image Target Recognition Model

Zhe Chen [1,2], Guohao Xie [1,2,3], Xiaofang Deng [1,2,*], Jie Peng [1,2] and Hongbing Qiu [1,2,*]

[1] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; chenzhe@mail.nwpu.edu.cn (Z.C.); 22172301007@mails.guet.edu.cn (G.X.); pengjie@guet.edu.cn (J.P.)
[2] Cognitive Radio and Information Processing Key Laboratory Authorized by China's Ministry of Education Foundation, Guilin University of Electronic Technology, Guilin 541004, China
[3] School of Ocean Engineering, Guilin University of Electronic Technology, Beihai 536000, China
* Correspondence: xfdeng@guet.edu.cn (X.D.); qiuhb@guet.edu.cn (H.Q.)

Abstract: Affected by the complex underwater environment and the limitations of low-resolution sonar image data and small sample sizes, traditional image recognition algorithms have difficulties achieving accurate sonar image recognition. The research builds on YOLOv7 and devises an innovative fast recognition model designed explicitly for sonar images, namely the Dual Attention Mechanism YOLOv7 model (DA-YOLOv7), to tackle such challenges. New modules such as the Omni-Directional Convolution Channel Prior Convolutional Attention Efficient Layer Aggregation Network (OA-ELAN), Spatial Pyramid Pooling Channel Shuffling and Pixel-level Convolution Bilat-eral-branch Transformer (SPPCSPCBiFormer), and Ghost-Shuffle Convolution Enhanced Layer Aggregation Network-High performance (G-ELAN-H) are central to its design, which reduce the computational burden and enhance the accuracy in detecting small targets and capturing local features and crucial information. The study adopts transfer learning to deal with the lack of sonar image samples. By pre-training the large-scale Underwater Acoustic Target Detection Dataset (UATD dataset), DA-YOLOV7 obtains initial weights, fine-tuned on the smaller Smaller Common Sonar Target Detection Dataset (SCTD dataset), thereby reducing the risk of overfitting which is commonly encountered in small datasets. The experimental results on the UATD, the Underwater Optical Target Detection Intelligent Algorithm Competition 2021 Dataset (URPC), and SCTD datasets show that DA-YOLOV7 exhibits outstanding performance, with mAP@0.5 scores reaching 89.4%, 89.9%, and 99.15%, respectively. In addition, the model maintains real-time speed while having superior accuracy and recall rates compared to existing mainstream target recognition models. These findings establish the superiority of DA-YOLOV7 in sonar image analysis tasks.

Keywords: deep learning; underwater images; underwater target recognition; sonar images; improved YOLO

check for updates

Citation: Chen, Z.; Xie, G.; Deng, X.; Peng, J.; Qiu, H. DA-YOLOv7: A Deep Learning-Driven High-Performance Underwater Sonar Image Target Recognition Model. *J. Mar. Sci. Eng.* **2024**, *12*, 1606. https://doi.org/10.3390/jmse12091606

Academic Editors: Yan-Tsung Peng, Wenqi Ren, Jingchun Zhou and Qiuping Jiang

Received: 6 August 2024
Revised: 7 September 2024
Accepted: 7 September 2024
Published: 10 September 2024

1. Introduction

In underwater target detection, there is an increasing demand for high-precision recognition in marine resource development, underwater construction, and marine ecological monitoring. Sonar image target identification plays a crucial role [1–4]. However, traditional image recognition algorithms face challenges due to the complex underwater environment, low-resolution sonar image data, and small sample sizes, which result in difficulties in achieving accurate recognition. Therefore, there is a need to develop more effective methods to improve the performance of sonar image target recognition.

This research aims to tackle these challenges by leveraging deep learning techniques. Specifically, we strive to build a high-performance underwater sonar image target recognition model. This model should be able to handle the complexities of the underwater environment, enhance the accuracy of target recognition, and solve the problem of limited sample sizes. The novelty of this research lies in leveraging deep learning techniques to build a high-performance underwater sonar image target recognition model, namely the Dual Attention Mechanism YOLOv7 model (DA-YOLOv7). This model incorporates new modules such as the Omni-Directional Convolution Channel Prior Convolutional Attention Efficient Layer Aggregation Network (OA-ELAN), Spatial Pyramid Pooling Channel Shuffling and Pixel-level Convolution Bilateral-branch Transformer (SPPCSPCBiFormer), and Ghost-Shuffle Convolution Enhanced Layer Aggregation Network-High performance (G-ELAN-H) to enhance the model's ability to handle the complexities of the underwater environment, improve the accuracy of target recognition, and address the issue of limited sample sizes. By doing so, we hope to significantly contribute to underwater target detection and provide more reliable and efficient solutions for various applications.

Inherent in sonar imagery lies its origin in imaging sonars. When operating as an active sonar system, the process unfolds through the following steps: (I) The sonar system initiates the emission of sound waves; (II) these waves traverse through water, bouncing off underwater targets and returning; (III) the reflected echoes retrace their path back to the sonar system; and (IV) image formation ensues from the sophisticated processing of these echoes. The imaging procedure is susceptible to medium effects. Due to the complexity and inherent uncertainty of the underwater environment, echo signals often encounter challenges such as attenuation and distortion, resulting in lower contrast, reduced resolution, blurred target boundaries, and scarcity of discernible features in the generated sonar images [5–7].

Traditional sonar image target identification methodologies predominantly employ pixel-based features, grayscale values, or prior assumptions about the targets [8–12], often resulting in limited accuracy. Recently, inspired by the exceptional performance of deep convolutional neural networks (DCNNs) in optical image object detection [13–16], researchers have extensively investigated the application of deep learning in enhancing sonar image recognition. For instance, Fan and colleagues [17] designed a 32-layer deep residual network, substituting the Residual Network (Resnet50/101) in the Mask Regions with Convolutional Neural Network features (R-CNN) to enhance detection efficiency while minimizing trainable parameters, with implications for real-time operations and embedded systems. Zhu and co-authors [18] advanced this by integrating the Swin Transformer and Deformable Convolutional Networks (DCN) to create a Swin Transformer Based Anchor-Free Network (STAFNet). This no-anchor detection model surpasses conventional CNN-based solutions like Faster R-CNN and Fully Convolutional One-Stage Object Detection (FCOS) in detecting victims, vessels, and aircraft within their forward-facing sonar dataset. Zhou et al. [19] enhanced the detection and recognition of underwater biological targets in the Underwater Optical Target Detection Intelligent Algorithm Competition 2021 Dataset (URPC) through the fusion of image enhancement techniques, expanded Visual Geometry Group 16-layer network (VGG16) feature extraction, and a Faster R-CNN incorporating feature maps. Chen and collaborators [20] introduced IMA (Invert Multi-Class AdaBoost), a weighting scheme based on sample distribution, to mitigate noise-induced degradation in detection accuracy. Qiao and his team [21] developed a real-time and precise classifier for underwater targets, utilizing a combination of Local Wavelet Acoustic Pattern (LWAP) and Multilayer Perceptron (MLP) networks to tackle the complexities of underwater target classification. While these methods leverage CNNs to extract features, cascade convolutional layers, pool dimensions, and classify them through fully connected layers, they still exhibit room for improvement in accuracy when confronted with intricate sonar imagery.

Additionally, some scholars have adopted the YOLO (You Only Look Once) framework, reframing the sonar image target recognition problem as a regression task that directly predicts target locations and categories, optimizing recognition speed and precision. Notable examples include the work of [22,23], which integrated YOLOv3's principles by incorporating feature pyramids for multi-scale feature aggregation. Fan and his team [24] built upon the Adaptive Spatial Feature Fusion (ASFF) concept [25] to refine the YOLOv4 backbone, enhancing the feature fusion capabilities. Zhang et al. [26] further enhanced the YOLOv5 backbone, enhancing detection speed. Cheng et al. [27] studied underwater target recognition in the context of small datasets, proposing a new approach that combines diffusion models with YOLOv7. Zheng [28] improved YOLOv8 by introducing the Scale and Channel Efficient Module Attention (ScEMA) module, enhancing the model's sensitivity to small targets and objects across various scales. Despite the algorithm's advantages in speed, YOLO has been found to occasionally miss detections or produce false positives, particularly in complex environments and with limited datasets, raising concerns about its stability.

Traditional image recognition algorithms have a relatively complex process to achieve accurate recognition when dealing with low-resolution sonar image data and small samples. In addition, the existing deep learning-based models still have room for improvement in accuracy when handling complex sonar images. For example, the YOLO framework has problems with omissions and false detections in complex environments and limited datasets and lacks stability. Meanwhile, with the continuous advancement in technology, although YOLOv10 shows an excellent performance in the field of optical images, in the field of sonar images, because underwater sonar targets are usually small in scale, YOLOv7 has more advantages due to its excellent recognition ability for small targets. However, the insufficiency of available public sonar image datasets severely limits the application of deep learning models in this field. To address these challenges, the researchers of this study have decided to make improvements based on YOLOv7 and propose to use DA-YOLOv7 to solve these problems.

Our approach addresses the drawbacks of low-resolution sonar imagery, unclear target features, and limited sample availability by introducing the Dual Attention Mechanism YOLOv7 model, abbreviated as DA-YOLOv7. To address the drop in accuracy in complex environments, we develop the OA-ELAN module, integrating Omni-Directional Convolution (ODConv) and Channel Prior Convolutional Attention (CPCA) attention mechanisms, thereby enhancing the model's capacity for capturing crucial local features and improving the recognition of object shapes and edges. Focusing on the intricacy of sonar image targets, we devise the SPPCSPCBiFormer and G-ELAN-H modules, leveraging the ability of the Transformer to handle long-range dependencies. These innovations enhance the model's adaptability to multi-scale targets and reduce computational demands while improving the detection accuracy for small and dense targets. Transfer learning is employed to counteract the scarcity of sonar image samples. We put the data from the diverse Underwater Acoustic Target Detection Dataset (UATD) [29] into the DA-YOLOv7 model for pre-training. During this process, the model learns to extract features and make predictions based on these data. After the training, the pre-trained weights optimized for the UATD dataset are obtained. The pre-trained model can be used for transfer learning. In this case, fine-tuning the pre-trained weights on the smaller Common Sonar Target Detection Dataset (SCTD) [30] can alleviate the overfitting issues typically associated with small datasets. To verify the effectiveness of our proposed method, we conducted rigorous comparative evaluations on various real-world sonar image datasets, including UATD, SCTD, and URPC [31].

To conclude, the core innovations of this paper are embodied in the following key contributions:

- The development of the DA-YOLOv7 model represents a significant advancement, striking a balance between computational efficiency and enhanced feature extraction capabilities. By systematically refining YOLOv7's modules and incorporating innovative elements such as the OA-ELAN, SPPCSPCBIFormer, and G-ELAN-H units, the model not only fortifies its ability to discern crucial information but also streamlines multi-scale target processing and integrates long-range dependencies seamlessly. These enhancements collectively foster enhanced generalization, diversification, and robustness within the overall architecture of the model.
- In response to the lack of datasets encountered in sonar image target recognition tasks, this investigation employs a transfer learning training approach. DA-YOLOv7

initiates by acquiring pre-trained weights through an extensive pretraining phase on the UATD dataset. These pre-trained weights are then strategically employed to refine the model's training on the smaller SCTD dataset, thereby effectively addressing the common issue of diminished recognition accuracy due to the scarcity of data in the small datasets.

- Our analysis meticulously investigates its performance across various scenarios to strengthen the credibility of YOLOv7's enhanced capabilities for underwater sonar image target recognition. Our proposed method includes rigorous testing on low-resolution images sourced from the UATD dataset and color images from the URPC dataset, showcasing its dependable and versatile nature in different imaging circumstances.

The structure of the remainder of this paper is as follows: Section 2 mainly introduces YOLOv7 and the various modules for its improvement, including OA-ELAN, SPPCSPCBiFormer, G-ELAN-H, etc. The improvements in these modules aim to enhance the model's performance when handling the underwater sonar image target recognition task. Section 3 elaborates on the structure and improvements in the DA-YOLOv7 network. Through the upgrades of multiple modules, DA-YOLOv7 is committed to addressing the existing problems in YOLOv7 in underwater sonar image target recognition and enhancing the performance and adaptability of the model. Section 4 is dedicated to the comparative analysis of DA-YOLOv7 with leading sonar image recognition technologies, presenting the results on multiple real datasets. The summary and conclusion of our investigation results are presented in Section 5.

2. Related Works

We selected YOLOv7 for our research because it has shown promising performances in various object detection tasks. However, we modified this model to better suit the underwater sonar image target recognition requirements. The newer versions of YOLO may have their advantages, but YOLOv7 provides a solid foundation that allows us to make targeted improvements to address the unique challenges in this domain. Our modifications include adding the OA-ELAN, SPPCSPCBiFormer, and G-ELAN-H modules to enhance the model's feature extraction, target recognition, and generalization capabilities in the underwater context.

2.1. YOLOv7

Figure 1 depicts the network structure diagram of YOLOv7, encompassing the input layer, backbone, head, and prediction layer. The adaptive bounding box calculation approach is employed in the input layer to ensure that the input color image is uniformly scaled to the 640 × 640 size specification. The backbone comprises the Convolutional Block with Convolution-Batch Normalization-SiLU activation function module (CBS), the Efficient Layer Aggregation Network module (ELAN), and the Multi-scale Layer (MP). The input image undergoes 3 × 3 convolutions with strides of 1 and then with strides of 2 successively to compress the spatial dimension and increase the number of channels. The ELAN module consists of multiple CBSs, which enhances feature extraction by enlarging the receptive field. The MP layer is accountable for adjusting the channel ratio, among which MP1 utilizes max pooling to reduce the number of channels, and MP2 compresses the space and increases the number of channels via convolution and downsampling. This design enables the ELAN and MP layers to be alternately reused multiple times, progressively expanding the receptive field and enriching the feature expression. Eventually, the backbone constructs a multi-scale feature pyramid, and these feature maps are passed to different prediction heads, each focusing on detecting a specific target size and generating the corresponding bounding box, category probability, and confidence through the convolutional layer.

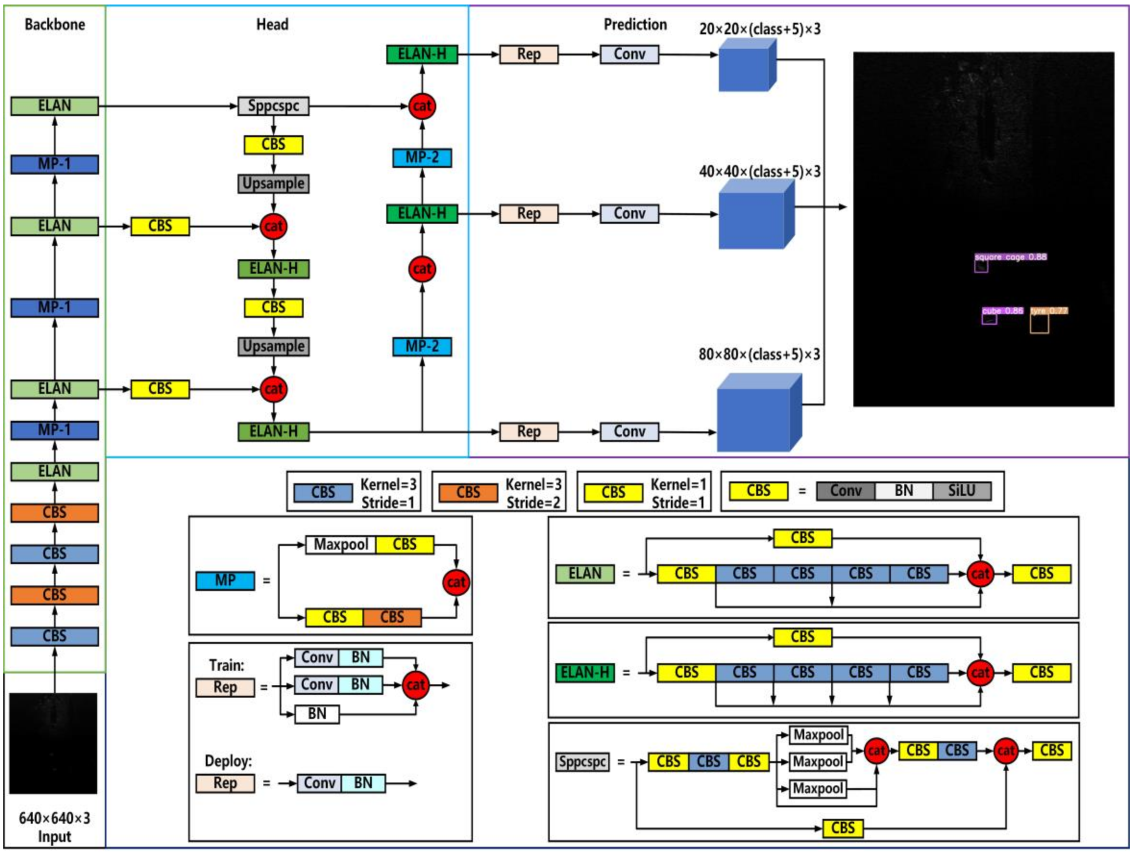


Figure 1. Structure of the YOLOv7.

2.2. OA-ELAN

The effect of YOLOv7 in processing optical images with high contrast and apparent target features is exceptionally excellent. However, problems of false recognition and missed recognition are prone to occur when processing sonar images with low contrast and blurred target features. For this reason, we designed a new OA-ELAN module. Specifically, we replaced the ordinary convolution in the CBS module after the cat operation in the original ELAN module with the ODConv module and incorporated the innovative CPCA attention mechanism.

This change is mainly reflected in the following two aspects: First, in the CBS module after the cat operation in the ELAN module, we utilized the advanced ODConv module to replace the traditional convolution operation. The ODConv module, through its parallel strategy and the combination of a multi-attention mechanism, can learn a more abundant and flexible four-dimensional convolution kernel space, significantly enhancing the model's feature extraction ability, especially in dealing with small targets. Secondly, we introduced a brand-new CPCA attention mechanism. Through its unique design, the CPCA mechanism can flexibly project attention among different channels and regions, enabling the model to precisely focus on crucial single-channel information and deeply understand and ingeniously utilize the inherent interaction relationships between the different channels, further optimizing the feature representation. Our approach has greatly enhanced the accuracy of target recognition and the stability of the model. Especially when facing the complex and changeable background of sonar images, it plays a significant role in effectively reducing false recognitions and missed recognitions.

2.2.1. ODConv

To solve the problems of false recognition, missed recognition, and low recognition accuracy, the limitations of dealing with multi-scale targets, being interfered with by background noise, inaccurate target positioning, and poor performance in the complex and changeable underwater environment existing in the object recognition of underwater sonar images by YOLOv7, we newly added the convolutional operation method of ODConv (Figure 2). ODConv can divide the input signal into high-frequency and low-frequency parts and, through processing with different convolutional kernels, it can better capture the local details and global information of the image, enhance the model's understanding and processing ability for complex images, and improve the recognition accuracy of the targets; it can enable the model to learn different features, especially for small target recognition, and the accurate detection of underwater sonar targets of different scales; its anti-noise ability can filter out background noise by separating high- and low-frequency information, allowing the model to focus on the target itself and reducing the occurrence of false recognition and missed recognition; the depth understanding ability of ODConv helps the model understand the depth differences of the objects and improves the accuracy of target positioning, further improving the recognition performance; in addition, ODConv has solid environmental adaptability and flexibility, can maintain a good performance in different underwater environments, enhance the robustness of object recognition, and ensure the stable operation of the model.

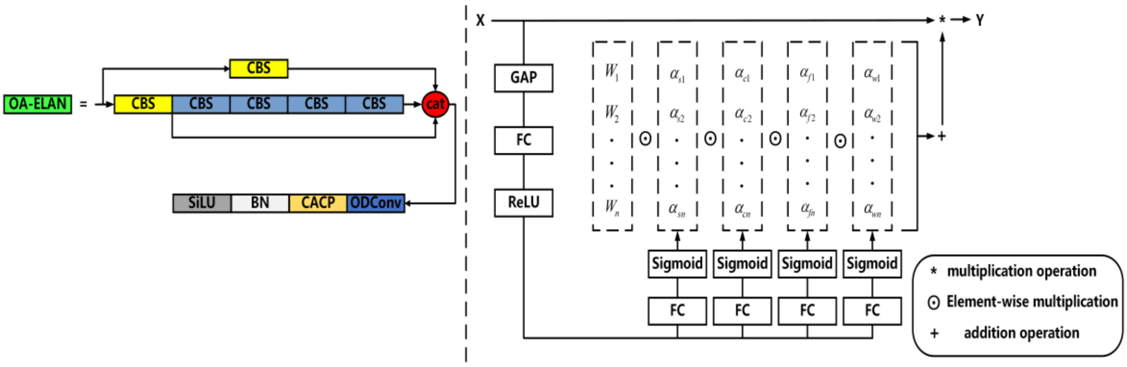


Figure 2. (**left**): OA-ELAN structure diagram, (**right**): ODConv Structure diagram.

2.2.2. CPCA Attention Mechanism

The Channel Prior Convolutional Attention Module is depicted in Figure 3, which sequentially undertakes channel and spatial attention. With an intermediate feature map, $F \in R^{C\times H\times W}$, provided as the input, the channel attention module (CA) initially infers a 1D channel attention map, $M_c \in R^{C\times 1\times 1}$. Following the procedure, the operation of element-wise multiplication between M_c and input feature, F yields channel attention values broadcast along the spatial dimensions to generate a refined feature, $F_c \in R^{C\times H\times W}$, enriched with channel-specific focus. Subsequently, the spatial attention module (SA) applies its processing on F_c to yield a three-dimensional spatial attention map, $M_s \in R^{C\times H\times W}$. The final output feature, $\hat{F} \in R^{C\times H\times W}$, is computed by the element-wise multiplication of M_s with F_c, where C signifies the number of channels, H represents the height, and W denotes the width. In summary, the entire attention mechanism is encapsulated within the following formulaic description:

$$F_c = CA(F) \otimes F \tag{1}$$

$$\hat{F} = SA(F_c) \otimes F_c \tag{2}$$

where $\otimes$ means the element-by-element multiplication.

The channel attention module is central in generating a channel-wise attention map to decipher the complex interdependencies within feature channels. Employing average and max pooling operations effectively consolidates the feature maps' spatial context, thereby generating two unique spatial context descriptors. These descriptors are then processed through a shared MLP, characterized by a single hidden layer of fixed dimension,

$R^{c/r\times1\times1}$, with r denoting the reduction ratio to ensure efficient parameter utilization. The computation of the channel attention map is completed through the integration of MLP outputs using element-wise summation, its systematic process being mathematically formalized as follows:

$$CA(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$

where σ represents the sigmoid function.

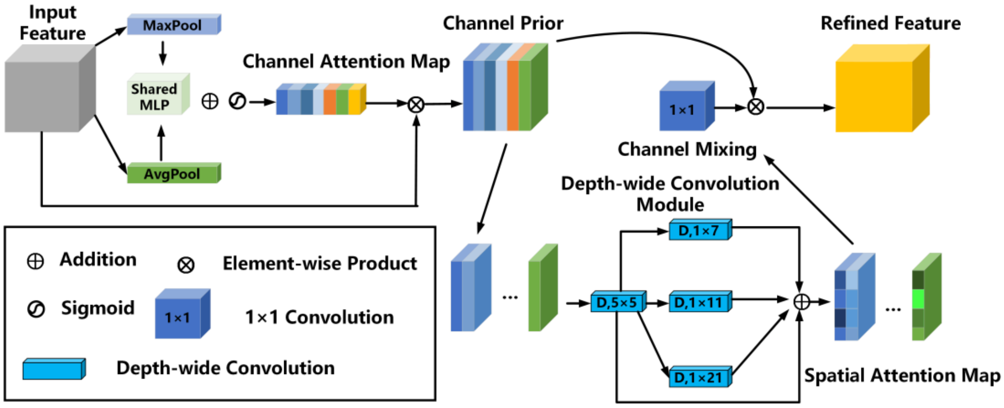


Figure 3. The CPCA attention mechanism.

Spatial attention maps are generated by a process that involves extracting the spatial relationships facilitated by the deep convolution layers that simultaneously preserve inter-channel correlations and capture intricate spatial dependencies. A multi-scale architecture is adopted to optimize this, allowing for convolution operations to effectively discern spatial relationships without compromising computational efficiency. The spatial attention module concludes with a 1 × 1 convolutional layer, which mixes channels and produces a more refined attention map. The mathematical description of the spatial attention calculation proceeds as follows:

$$SA(F) = Conv_{1\times1}\left(\sum_{i=0}^{3} Branch_i(DwConv(F))\right) \tag{4}$$

The model architecture incorporates Depthwise Convolution ($DwConv$), a deep convolutional component, alongside branches designated as $Branch_i$ and $i \in \{0,1,2,3\}$, collectively forming branch *i-th*. Notably, $Branch_0$ constitutes a direct connection path, functioning as a shortcut that omits any intermediate transformations or processing, allowing the input to flow uninterrupted to the output. $Conv_{1\times1}$ is a 1×1 convolution. This design choice contributes to the overall architectural integrity while maintaining computational efficiency.

In response to the challenges faced by YOLOv7 during underwater sonar image object recognition, such as limited feature extraction, vulnerability to noise and interference, and inadequate adaptation to diverse target dimensions, we employed a series of strategic improvements. Primarily, we replaced the ELAN module in the backbone with the advanced OA-ELAN, enhancing its structural robustness. The introduction of the ODConv module within the Contextual CBS bolstered the model's capacity to detect objects across multiple scales and locations, reducing false positives and missing recognitions in complex scenarios with multi-scale objects. It facilitated the dynamic capture of spatial and channel-specific details thus optimizing feature representation, particularly for precise small target identification.

The OA-ELAN module is a novel design that replaces the ordinary convolution in the original ELAN module with the ODConv module and incorporates the CPCA attention mechanism. This innovation enhances the model's feature extraction ability, especially for small targets. It reduces false and missed recognitions, a significant improvement over the traditional methods that struggle with the low contrast and blurred target features of sonar images.

The CPCA attention mechanism further improves the feature expression ability of the model, facilitating the extraction of critical features, emphasizing target significance, and enhancing the model's understanding of underwater sonar imagery. Our method improved recognition and recognition accuracy, reduced noise and interference resilience, and enhanced robustness in complex underwater environments. Additionally, the mechanism adaptively adjusted attention weights based on target size and shape, enabling more effective recognition across a spectrum of target dimensions. The model's transfer learning and generalization capabilities were significantly enhanced by fostering general feature representations.

Notably, these enhancements were achieved without compromising computational efficiency, as both the ODConv and CPCA mechanisms maintained high computational effectiveness, ensuring the model's runtime speed. Consequently, these modifications substantially improved YOLOv7's underwater sonar image object recognition performance, particularly in tackling complex backgrounds, noise, and small target recognition tasks.

2.3. SPPCSPCBiFormer Module

To better recognize and understand the target in complex scenes, we have incorporated the SPPCSPCBiFormer module (Figure 4). This module is integrated with the Spatial Pyramid Pooling Channel Shuffling and Pixel-level Convolution (SPPCSPC) module and the Bilateral-branch Transformer (BiFormer) module.

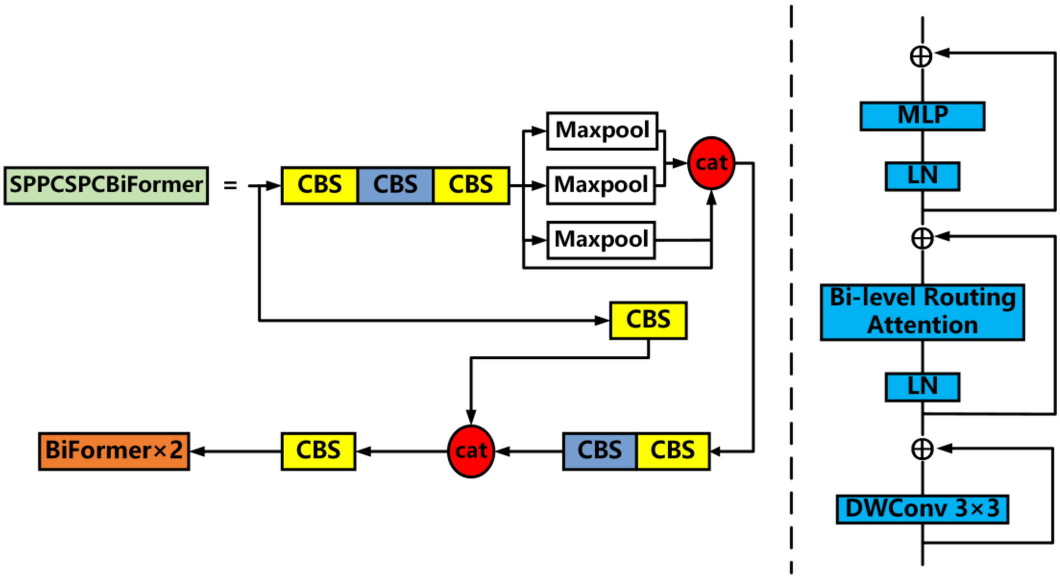


Figure 4. Left: SPPCSPC structure, **Right**: BiFormer structure.

Among them, the Spatial Pyramid Pooling (SPP) in the SPPCSPC module generates multi-scale feature maps by conducting pooling operations of different scales on the input image, thereby effectively handling the variations in the input size and enhancing the invariance in the model to image scaling, rotation, and other transformations; Channel Shuffling and Pixel-level Convolution (CSPC) incorporates channel shuffling to promote the interaction between channels and pixel-level convolution to extract local features, which contributes to improving the model's ability to represent features of high-dimensional data.

SPPCSPCBiFormer not only inherits the advantage of the SPPCSPC module in its ability to better understand the diversity and complexity of the target in complex scenes but also integrates the strength of the self-attention mechanism of the BiFormer module in understanding the complex relationships between targets, further enriching the feature representation and enabling the model to understand the input data from multiple perspectives and levels.

This combination of multiple modules boosts the model's generalization ability, enabling it to adapt to targets of various scales and handle complex dependency relationships. It integrates multi-scale feature extraction, long-distance dependency capture, and efficient feature fusion, significantly enhancing the model's ability to recognize and understand the target.

As illustrated in Figure 4, the BiFormer architecture commences with a Depthwise Convolution, which inherently encodes spatial relationships. Following this, the model employs the Bi-directional Relationship Aggregation (BRA) module, followed by a two-layer MLP with an expansion rate of e, the latter of which is dedicated to modeling cross-position relationships and generating per-position embeddings, thereby enriching the feature representation.

A feedforward neural network layer is deployed after the multi-head attention layer to enhance the expressive power of the extracted features. Residual connections are strategically employed to mitigate the issue of vanishing gradients and optimize the model's training efficiency.

The BRA module is the core part of the BiFormer module. It processes the input features through multiple parallel attention heads and each attention head can learn different feature representations, thereby improving the flexibility and expression ability of the model. The steps are as follows:

Region partitioning and input projection: Given an input feature map $X \in R^{H \times W \times C}$, we first partition it into $S \times S$ non-overlapping regions, such that each region contains $\frac{HW}{S^2}$ feature vectors. This step is accomplished by reshaping X into $X^R \in R^{S^2 \times \frac{HW}{S^2} \times C}$. Then, we use a linear projection to derive the query, key, and value tensors Q, K, $V \in R^{S^2 \times \frac{HW}{S^2} \times C}$:

$$Q = W^r W^q, K = X^r W^k, V = X^r W^v \tag{5}$$

where W^q, W^q, $W^v \in R^{C \times C}$ are the projection weights of query, keys, and values.

- First, the regional-level queries and keys Q^r, $K^r \in R^{S^2 \times C}$ are derived by applying the average value of each region to Q and K, respectively. Then, we derived the adjacency matrix $A^r \in R^{S^2 \times S^2}$ of the region-to-region affinity graph through the matrix multiplication between Q^r and the transpose of K^r as follows:

$$A^r = Q^r (k^r)^T \tag{6}$$

Among them, the semantic relevance of two regions is measured in the adjacency matrix A^r.

The core step that we undertake next is to prune the affinity graph by merely retaining the top-k connections of each region. Specifically, we employ the row-wise *topk* operator to obtain the routing index matrix $I_r \in N^{S^2 \times K}$ as follows:

$$I^r = topkIndex(A^r) \tag{7}$$

- Token-level Attention: Our approach harnesses the region-to-region routing index matrix I^r to implement a meticulous token-to-token attention mechanism. Each query token within region i extends its attention to all the key–value pairs that are found within the aggregated routing regions indexed by $I^r_{(i,1)}, I^r_{(i,2)}, \ldots, I^r_{(i,k)}$, representing the union of k regions. The process of attention computation begins with the collection of the key and value tensors, which are fundamental components in this procedure, as follows:

$$K^g = gather(K, I^r), V^g = gather(V, I^r) \tag{8}$$

where $K^g, V^g \in R^{S^2 \times \frac{kHW}{S^2} \times C}$ is the clustered key–value tensor.

We can then focus our attention on the collected key–value pairs:

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{9}$$

Here, we introduce the local context enhancement terms $LCE(V)$, as shown in [32]. The function $LCE(\cdot)$ is parameterized using a deep convolution, and we set the kernel size to 5.

The SPPCSPCBIFormer module enhances the ability to handle targets of different scales, enabling a more comprehensive understanding of the features of targets in complex scenes, thereby improving recognition accuracy. Regarding long-range dependency capture, the Transformer structure helps handle dense or overlapping targets. In terms of enhancement of the generalization ability, introducing multiple attention mechanisms and Transformer structures enables the model to adapt to the variability of the underwater environment and consistently maintain good performance.

2.4. G-ELAN-H

To overcome the issues of inadequate target recognition accuracy, weak generalization ability, and high computational cost existing in the YOLOv7 model, we substituted some of the Convs in all the ELAN-H modules of YOLOv7 with Ghost-Shuffle Convolution (GSConv), thereby constituting the G-ELAN-H module. As shown in Figure 5, the structure of the G-ELAN-H module reveals its design. In this manner, GSConv can assist the model in learning features more effectively, significantly enhancing the target recognition accuracy, and the recognition effect is particularly notable for small targets or densely arranged targets; it can strengthen the model's processing capability for targets of diverse shapes and sizes and enhance its generalization ability; and it can decrease the computational cost, accelerate the running speed of the model, and achieve faster image processing.

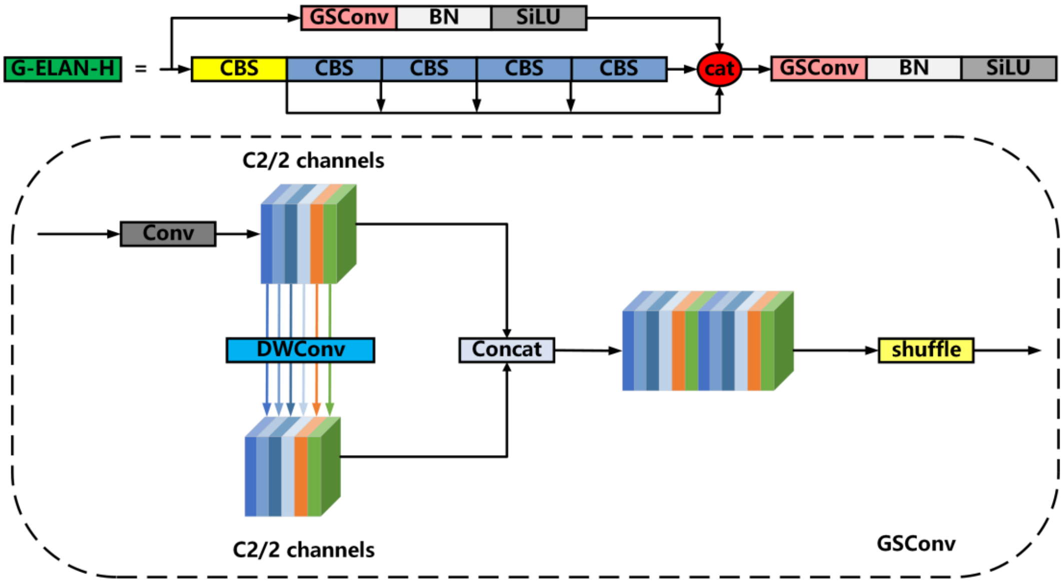


Figure 5. Structure diagram of G-ELAN-H.

3. Proposed Method

3.1. DA-YOLOv7 Network

The DA-YOLOv7 model, as shown in Figure 6, aims to address the problems in underwater sonar image target recognition of YOLOv7 and enhance the performance and adaptability of the model. Through improvements in the multiple modules, it strives to overcome the limitations of YOLOv7 in handling sonar images with low contrast and blurred target features, as well as the issues of insufficient target recognition accuracy, poor generalization ability, and high computational cost.

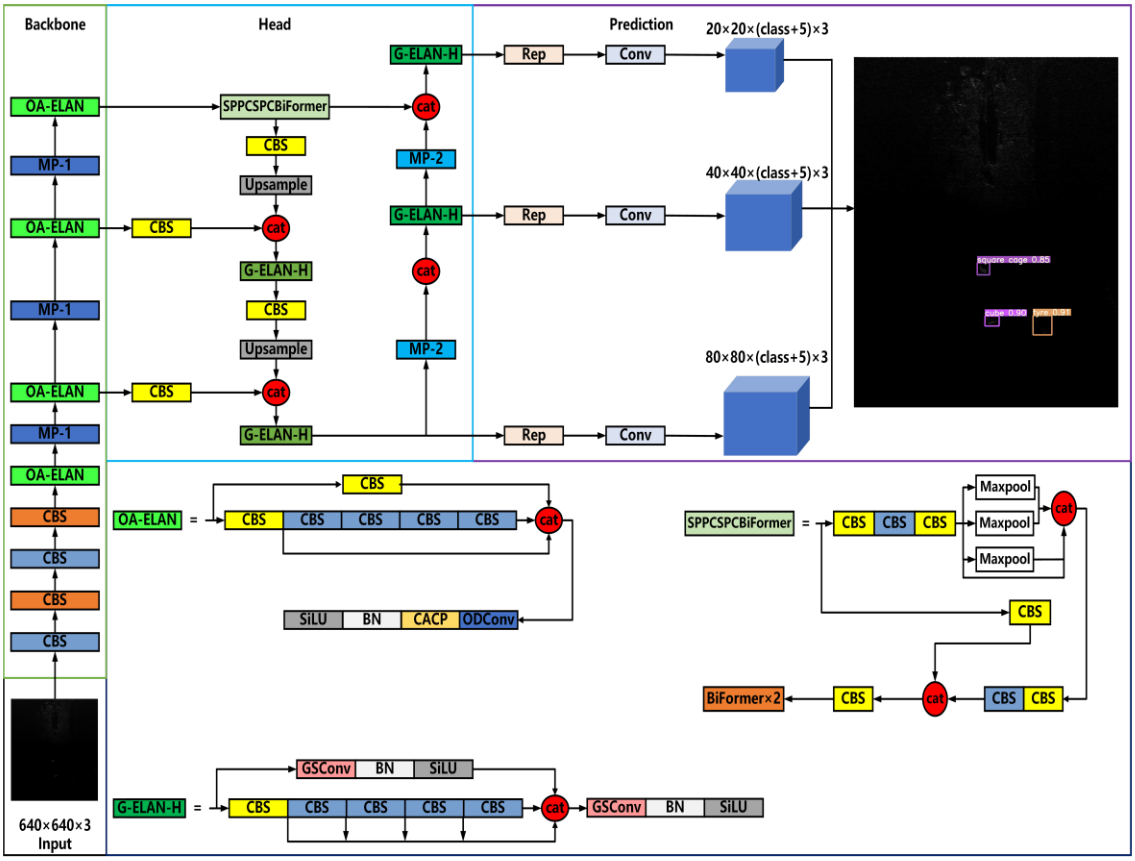


Figure 6. The DA-YOLOv7 network.

In the OA-ELAN module, we made significant modifications. We replaced the ordinary convolution in the CBS module after the cat operation in the original ELAN module with the ODConv module and incorporated the CPCA attention mechanism. This change aims to enhance the model's feature extraction ability, especially for small targets. The ODConv module can better capture the image's local details and global information, reducing the problems of false recognition and missed recognition in sonar images with low contrast and blurred target features. The CPCA attention mechanism helps the model to precisely focus on crucial information, improving the accuracy of target recognition and the stability of the model.

The SPPCSPCBiFormer module integrates the SPPCSPC and the BiFormer modules to improve the model's ability to identify and understand targets in complex scenes. The SPPCSPC module generates multi-scale feature maps and enhances the model's invariance to image transformations. In contrast, the BiFormer module leverages the self-attention mechanism to handle the complex relationships between targets. This integration enables the model to understand the input data from multiple perspectives and levels, enhancing its processing and generalization ability for targets of different scales.

In the G-ELAN-H module, we substituted some of the Convs in all the ELAN-H modules of YOLOv7 with Ghost-Shuffle Convolution (GSConv). This modification helps the model learn features more effectively, significantly enhancing the target recognition accuracy, especially for small or densely arranged targets. It also strengthens the model's processing capability for targets of diverse shapes and sizes, enhances its generalization ability, reduces the computational cost, and speeds up the running speed of the model.

These comprehensive improvements enable DA-YOLOv7 to exhibit extremely high accuracy, robustness, and computing speed when detecting small targets, analyzing dense scenes, and coping with changeable environments, significantly enhancing the performance of underwater sonar image target recognition.

3.2. Comparison of Related Work

To present the characteristics and differences of the related works more clearly, Table 1 summarizes and compares the associated works of YOLOv7 and its improved models.

Table 1. Comparison Table of YOLOv7 and its Improved Models.

Related Work	Description
YOLOv7	The network features an input layer for uniform image scaling to 640 × 640, a backbone with CBS, ELAN, and an MP, constructing a multi-scale feature pyramid. The output from the previous stage feeds into various prediction heads for precise object detection.
OA-ELAN	YOLOv7's modifications target low-contrast and blurred sonar images to reduce false and missed detections. The CBS module integrates ODConv, enhancing local and global feature capture, while the CPCA attention mechanism refines feature representation, boosts recognition accuracy, and stabilizes the model.
SPPCSPCBiFormer Module	The system integrates the SPPCSPC and BiFormer modules for enhanced target recognition in complex scenes. The SPPCSPC module boosts feature representation and model invariance to image transformations through multi-scale pooling and channel shuffling. The BiFormer module leverages bidirectional feature aggregation and multi-layer perceptrons to improve multi-scale target handling and scene nderstanding.
G-ELAN-H	To address YOLOv7's accuracy, generalization, and computational efficiency issues, GSConv replaces some convolutions in ELAN-H modules. Our approach enhances feature learning, particularly for small or clustered targets, and improves the model's handling of diverse target shapes and sizes. It also reduces computational demands and accelerates processing speed.
DA-YOLOv7 Network	Enhancements across several modules aim to tackle YOLOv7's underwater sonar image recognition challenges, enhancing model performance and adaptability. The OA-ELAN module strengthens feature extraction for small targets, minimizing false and missed detections. The SPPCSPCBiFormer module boosts multi-scale target recognition in complex scenes. The G-ELAN-H module refines accuracy, improves generalization, reduces computational costs, and speeds up image processing.

It can be seen from the table that YOLOv7 is the basic model, and the subsequent improvement works mainly focus on improving the performance of the model when processing sonar images. The OA-ELAN module enhances the model's ability to capture local features and critical information by introducing the ODConv and CPCA attention mechanisms, reducing misidentification and missed identification. The SPPCSPCBiFormer module integrates the advantages of the SPPCSPC module and the BiFormer module, improving the model's ability to handle multi-scale targets and understand complex scenes. The G-ELAN-H module improves the target recognition accuracy and generalization ability and reduces the computational cost by replacing the convolution operation. The DA-YOLOv7 network integrates these improvements, enhancing the underwater sonar image target recognition task.

4. Results and Discussion

To evaluate the effectiveness of DA-YOLOv7 in natural underwater environments, we used multiple underwater sonar image datasets, such as the lake bottom and shallow water areas of UATD, the diverse biomes of URPC, and the marine test scenarios of SCTD, to achieve comprehensive qualitative and quantitative analysis. We compared DA-YOLOv7 with the current underwater sonar image recognition models through these datasets.

4.1. Experimental Environment

We conducted rigorous experimental evaluations on the UATD, URPC, and SCTD datasets to substantiate the model's effectiveness to verify its recognition capabilities. The experimental setup included a Windows 11 operating environment, with the deep learning computations handled by PyTorch version 1.4.0. The hardware specifications comprised an Intel Core i7 12,700 H processor, 32 gigabytes of Random Access Memory (RAM), and a powerful NVIDIA GeForce GTX 3070Ti graphics processing unit (GPU).

4.2. Experimental Indicators

The assessment criteria were P*recision*, Re*call*, mAP, and FPS using the following formulas:

$$\mathrm{P}recision = \frac{TP}{TP + FP} \tag{10}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{11}$$

Here, TP represents predicting the correct answer, FP stands for mistakenly predicting from other classes to this class, and FN indicates that this category of labels is predicted to be of the other category of labels.

mAP is an abbreviation of average accuracy and an indicator of recognition accuracy in target recognition as follows:

$$AP = \int_0^1 P(r)dr \tag{12}$$

$$mAP = \frac{\sum_{n=1}^{N} AP_n}{N} \tag{13}$$

The following metrics are employed for evaluating the neural network's performance: precision (p), recall rate (r), a function characterized by parameter r (p), and the total number of object categories (n). The average accuracy in target identification, denoted as AP_n, is another critical aspect. The Mean Average Precision (MAP) is assessed using the following two distinct measures: mAP@0.5, which imposes a moderate recognition requirement, and mAP@0.5 : 0.95, which imposes a more stringent standard. These four evaluative parameters collectively determine the network's efficacy, and recognition speed is quantified through the measurement of frames per second (FPS).

4.3. UATD Dataset

The study employs the Open-Source Underwater Acoustic Target Recognition (UATD) dataset, featuring over nine thousand images captured utilizing a Tritech Gemini 1200 ik sonar, a recognized standard for MFLS data acquisition. The dataset is characterized by its extensive nature, encompassing raw sonar imagery sourced from various lake and shallow water settings. It serves as a valuable resource, featuring a diverse range of ten target objects, including cubes, spheres, cylinders, mannequins, tires, circular and square cages, metal buckets, aircraft models, and remotely operated vehicles (ROVs), thereby providing a robust empirical basis for research purposes.

4.3.1. Ablation Experiments on the UATD Dataset

We carried out comprehensive tests and ablation studies on each component of DA-YOLOv7 using the UATD dataset to delve into the impact of the modules detailed in Section 2 on the model's performance. As shown in Table 2, the notation "√" indicates the experimental outcomes when the respective module is engaged.

Table 2. Ablation comparison of model performance improvements in the UAATD dataset.

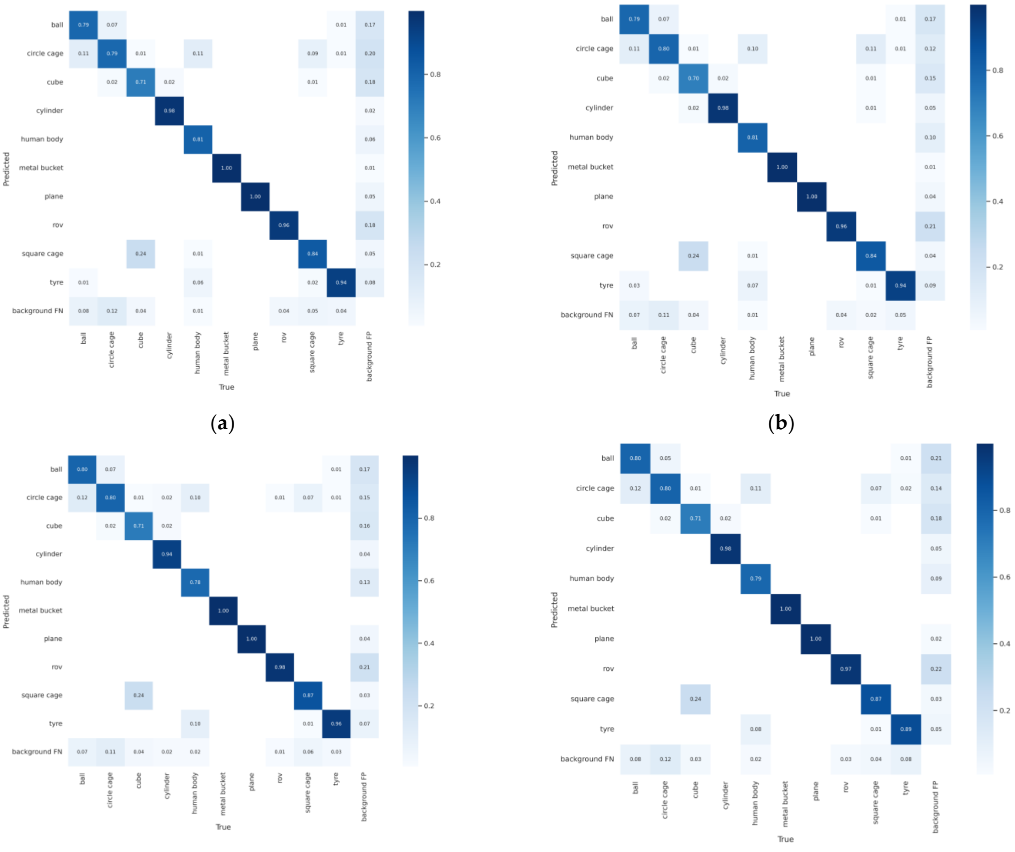
Model	OA-ELAN	SPPCSPCBiFormer	G-ELAN-H	Precision (%)	Recall(%)	mAP@0.5 (%)	mAP@0.5:0.95(%)
YOLOv7				85.7	83.0	81.2	34.4
	√			87.1	84.6	83.6	35.5
	√	√		91.2	88.9	87.2	37.2
	√	√	√	92.8	89.7	89.4	37.8

The original YOLOv7 model was taken as the baseline, with a precision of 85.7%, a recall rate of 83.0%, a mAP@0.5 of 81.2%, and a mAP@0.5:0.95 of 34.4%. When the OA-ELAN module was introduced alone, the precision increased to 87.1%, the recall rate increased to 84.6%, the mAP@0.5 increased to 83.6%, and the mAP@0.5:0.95 increased to 35.5%. This finding indicates that the OA-ELAN module is highly effective in optimizing feature extraction and classification capabilities and improving target detection accuracy under different confidence thresholds.

When both the OA-ELAN and SPPCSPCBiFormer modules were introduced simultaneously, the precision significantly increased to 91.2%, the recall rate increased to 88.9%, the mAP@0.5 increased dramatically to 87.2%, and the mAP@0.5:0.95 increased to 37.2%. This result shows the synergy of the two modules, enhancing the processing of complex features, reducing missed and false detections, and improving the detection performance in a broader range of confidence levels.

When integrating the three modules of OA-ELAN, SPPCSPCBiFormer, and G-ELAN-H, the precision reached as high as 92.8%, the recall rate reached 89.7%, the accuracy of mAP@0.5 significantly increased to 89.4%, and the mAP@0.5:0.95 increased to 37.8%. This finding proves that adding the G-ELAN-H module further improves the model architecture, enabling the model to achieve comprehensive and significant performance improvements across the entire confidence range, with more vital generalization ability and robustness. In conclusion, each module's individual and combined effects significantly improve the model performance.

To comprehensively compare YOLOv7 and our approach, we inspected the confusion matrices and precision-recall (PR) curves of the four models in question. As depicted in Figures 7 and 8, our methodology exhibits superior performance, demonstrating enhanced accuracy and a more balanced recognition capability for target and background classes.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 7. Confusion matrix of the ablation model: (**a**) YOLOv7; (**b**) YOLOv7 + OA-ELAN; (**c**) YOLOv7 + OA-ELAN + SPPCSPCBiFormer; (**d**) DA-YOLOv7.

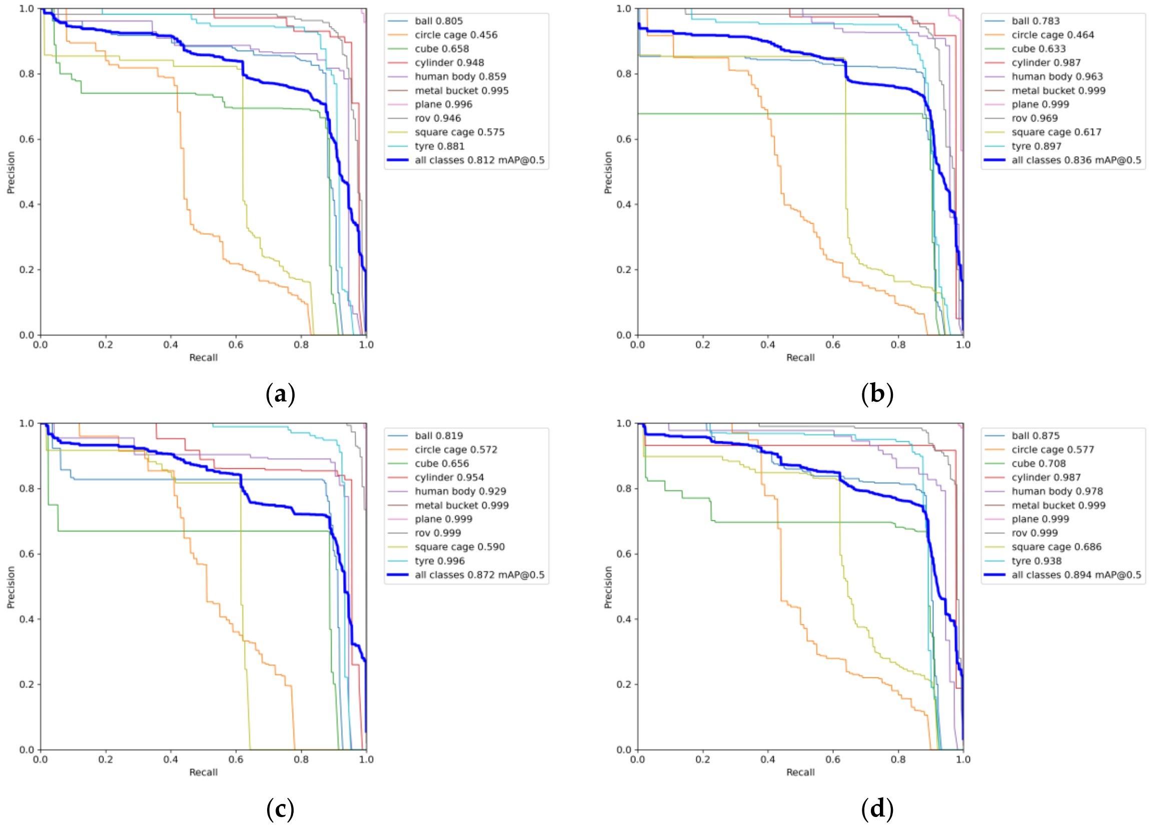


Figure 8. The PR curve: (**a**) YOLOv7; (**b**) YOLOv7 + OA-ELAN; (**c**) YOLOv7 + OA-ELAN + SPPCSPCBiFormer; (**d**) DA-YOLOv7.

It can be observed from Figure 8 that in Figure 8a, namely YOLOv7, the precision and recall rates of various targets are at the baseline level, indicating that the original YOLOv7 model has a specific performance in handling these targets, but there is still room for improvement. In Figure 8b, namely YOLOv7 + OA-ELAN, there are some significant improvements compared to Figure 8a. For example, the precision of the circle cage has increased from 0.456 to 0.464, and the precision of the cylinder has significantly increased from 0.948 to 0.987, which shows that the OA-ELAN module plays a vital role in enhancing the recognition accuracy of specific targets.

The improvements are more prominent in Figure 8c, namely YOLOv7 + OA-ELAN + SPPCSPCBiFormer. For example, the cube has increased from 0.633 to 0.656, and the precision for the human body has also increased from 0.963 to 0.978. These results indicate that the combined effect of the SPPCSPCBiFormer and OA-ELAN modules has brought significant performance improvements and effectively optimized the recognition ability of complex targets.

Figure 8d, DA-YOLOv7, integrates all the improvement modules and performs excellently in almost all target categories. For instance, the precision of a square cage has increased from 0.617 to 0.686, and the precision of the tire has increased from 0.897 to 0.938. This finding fully demonstrates the synergy of the three modules, OA-ELAN, SPPCSPCBiFormer, and G-ELAN-H, enabling the model to recognize various targets more accurately and comprehensively.

4.3.2. Analysis of UATD Dataset Results

This experiment evaluates the DA-YOLOv7 model's accuracy and recall performance for detecting ten distinct target categories. Figure 8d illustrates the trend of changes in these metrics when employing a forward-looking sonar for recognition, with a mean average precision (mAP) at 0.5, reaching 89.4% for all categories. Moreover, Figure 7d presents a confusion matrix showcasing the DA-YOLOv7's recognition outcomes for various target classes, reinforcing the model's recognition capabilities.

While the model demonstrates strong performance in detecting various targets, it occasionally encounters background noise in forward-looking sonar images, causing the misidentification of certain background regions as targets. Therefore, further investigations into sonar image denoising is deemed essential.

Finally, we trained and tested multiple popular target recognition models along with the original YOLOv7, and the comparative performance of these models is presented in Table 3.

Table 3. Target recognition results of each model in the previsual sonar images.

Model	Backbone	mAP@0.5 (%)	FPS
FasterR-CNN	Resnet	83.9	44.1
FasterR-CNN	Resnet-50	82.9	32.9
FasterR-CNN	Resnet-101	81.8	26.6
YOLOv3	Darknnet-53	80.1	49.8
YOLOv3	MobilenetV2	78.7	93.4
YOLOv5	CSPDarknet-53	81.5	83
YOLOv7	ELAN-Net	85.3	62.11
YOLOv8	CSPDarknet-c2f	85.6	49
CCW-YOLOv5	CSPDarknet-53	85.3	54
DA-YOLOv7	OA-ELAN-Net	89.4	97.09

It can be learned from Table 3 that, in the target recognition results of forward-looking sonar images, different models have their characteristics. In the FasterR-CNN series, the mAP@0.5 value gradually decreases as the backbone network changes, and the frame rate also gradually drops. In YOLOv3, the model with the Darknet-53 backbone network has a higher accuracy than the MobilenetV2 backbone network, but the latter has a better frame rate. For YOLOv5, YOLOv7, and YOLOv8, which adopt different backbone networks, respectively, the mAP@0.5 values vary.

However, our DA-YOLOv7 model has significant advantages. When using OA-ELAN-Net as the backbone network, the model achieves a mAP@0.5 as high as 89.4%, and its accuracy rate is significantly ahead of the other models. At the same time, the frame rate reaches 97.09, and the processing speed is breakneck. Combining the roles of different modules as described before, the OA-ELAN module can capture the critical features of the target more accurately by optimizing feature extraction and classification to improve the accuracy rate. The SPPCSPCBiFormer module enhances the processing of complex features and the utilization of global information, helping the model better understand the target context and overall features and improving accuracy. The G-ELAN-H module improves the model architecture, optimizes the depth and width of the network, and promotes feature fusion. While improving the accuracy rate, it ensures the efficient operation of the model and significantly improves the processing speed. These modules work together to make DA-YOLOv7 have both high accuracy and processing speed, showing excellent comprehensive performance.

As depicted in Figure 9, the value steadily decreases as the number of iterations rises, ultimately attaining convergence after 200 iterations. This stability reflects the optimization process of the model.

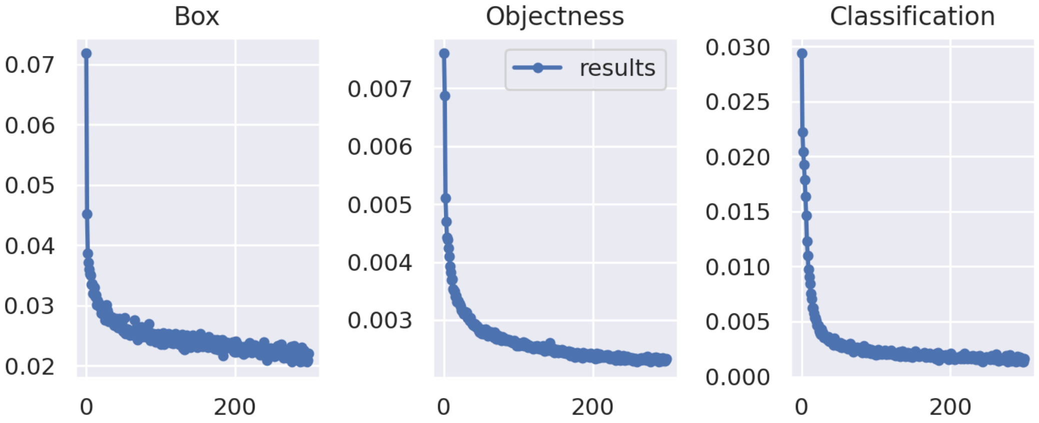


Figure 9. Curve of the change in loss value on the UATD dataset.

To visually illustrate the target recognition performance of the DA-YOLOv7 model developed in this chapter under real-world circumstances, Figure 10 presents the prediction results of various targets in multi-beam forward-looking sonar images extracted from the UATD validation set.

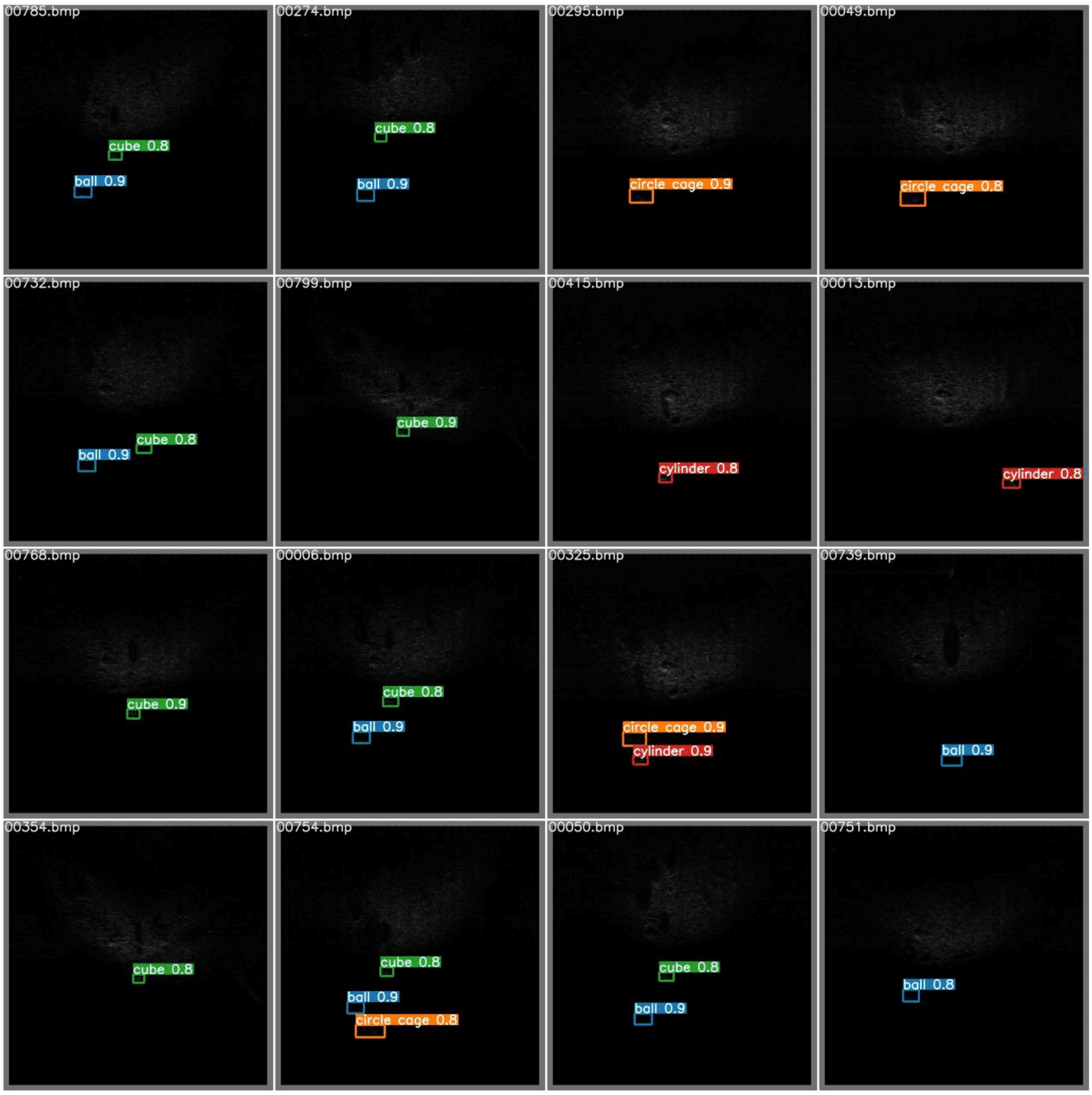
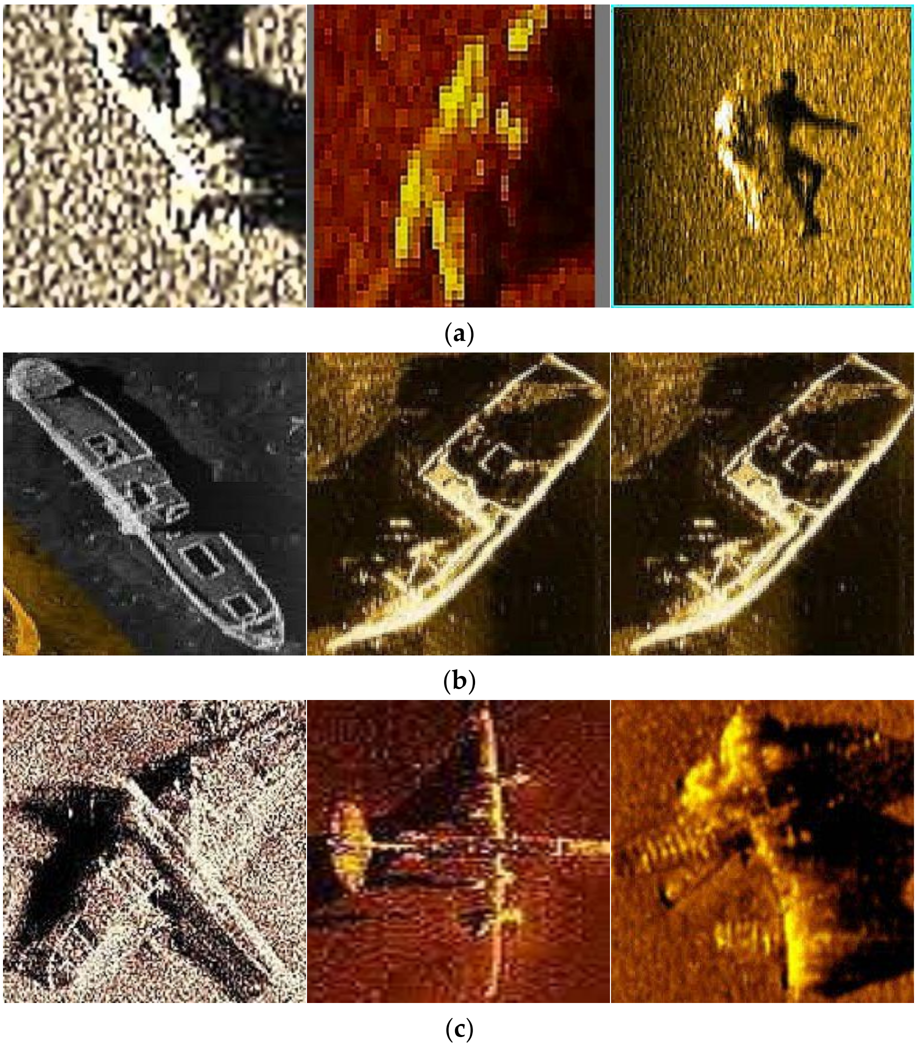


Figure 10. Prediction results of various targets in UATD multi-beam forward-looking sonar images.

Figure 10 shows many multi-beam forward-looking sonar images and their corresponding target annotations. From the figure, the advantages of our model can be observed. The model can not only accurately identify various targets, such as “ball”, “cube”, “cylinder”, etc., but also performs well in complex target combination situations and can still make accurate judgments when multiple targets of the same or different types exist simultaneously. Moreover, it performs stably and consistently in various images. Whether a single target or a complex multi-target scene, it can maintain a high accuracy rate and effectively distinguish easily confused targets. Its strong generalization ability enables it to complete the target recognition task excellently in different multi-beam forward-looking sonar image scenarios.

4.4. SCTD Dataset

We collected an SCTD dataset showing the sonar imaging results of different targets in different underwater maritime trials to conduct image processing and recognition research. These sonar images were filtered and classified, resulting in sonar images of the following three fundamental target categories: human, ship, and airplane. We selected three images of each type for display, as shown in Figure 11.


(**a**)

(**b**)

(**c**)

Figure 11. SCTD sonar image dataset: (**a**) human; (**b**) ship; (**c**) aircraft.

The SCTD dataset is categorized into three classes—ships, airplanes, and humans. Numerous uncertainties, such as complex backgrounds, object fragmentation, breakage, and deformation, augment the difficulty of recognition. Each represents one of the three underwater targets. These images shall be utilized as the input of the deep learning model. Through manual annotation, the target images are classified into three categories, respectively. The training set and the validation set are partitioned in a 9:1 ratio.

4.4.1. Training Strategy for the SCTD Dataset

The SCTD dataset presently encompasses a total of 357 images. Owing to the limited number of samples, we adopted a pre-trained model to tackle the problem of small sample datasets, and the flowchart is presented in Figure 12. This experiment aims to demonstrate that, by utilizing a pre-trained model, we can effectively tackle the issue of the limited number of samples in the SCTD dataset and enhance the performance and generalization capability of the model on small sample datasets.

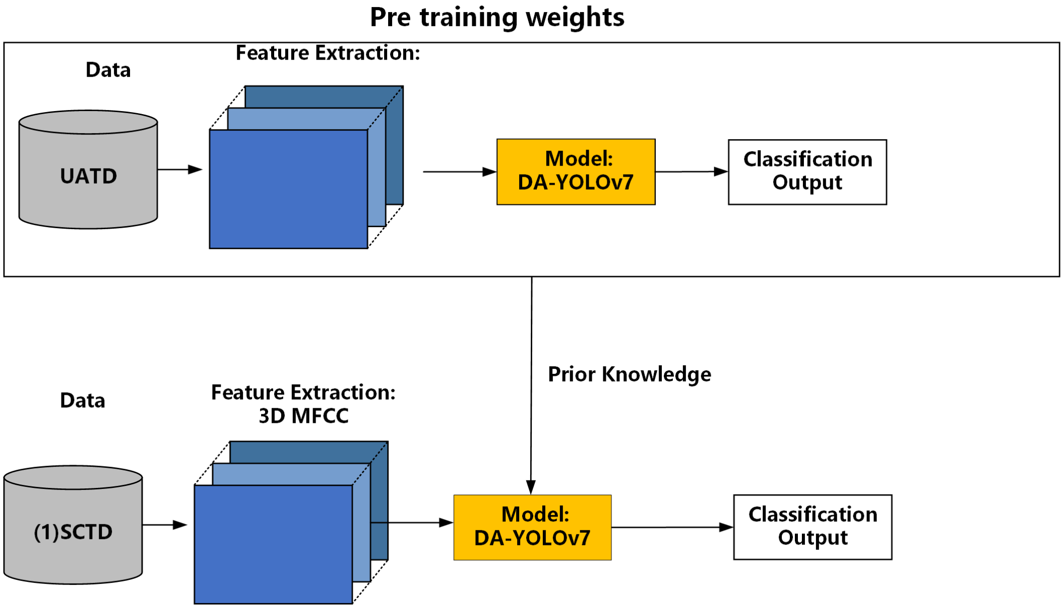


Figure 12. Flowchart of the training strategy for the SCTD dataset.

When dealing with a small sample dataset like SCTD, we adopted an innovative approach of pre-training the DA-YOLOv7 model on the UATD large dataset to obtain pre-trained weights. This approach has many advantages. Firstly, the UATD large dataset is rich and diverse, enabling DA-YOLOv7 to learn extensive and general features and patterns. When these pre-trained weights are used for training on the SCTD small dataset, the model can quickly adapt to the new data, significantly reducing the training time and computational cost. Secondly, the knowledge acquired by DA-YOLOv7 from the UATD dataset can provide valuable prior information for its learning on the SCTD small dataset, helping the model to extract more representative features from the limited data, enhancing its understanding and learning ability of the small dataset, and thereby improving the generalization performance.

Regarding the issue of low accuracy on small datasets, the pre-trained weights play a crucial role. By relying on the general feature representation learned on the UATD dataset, DA-YOLOv7 can more accurately capture the critical information when confronted with the SCTD small dataset, effectively reducing the possibility of overfitting and thereby significantly improving the accuracy. In conclusion, by leveraging the UATD large dataset to pre-train DA-YOLOv7 and then applying it to the training of the SCTD small dataset, an efficient and feasible approach is provided to address the challenges of small sample datasets.

4.4.2. Model Performance Comparison

Table 4 presents the performance comparison results between the proposed model and other outstanding recognition models on the dataset SCTD. The evaluation metrics of this model encompass AP for each category, mAP@0.5, and FPS.

Table 4. Model performance comparison on SCTD.

Model	Aircraft-AP (%)	Human-AP (%)	Ship-AP (%)	mAP@0.5 (%)	FPS
YOLOv4-tiny	88.16	79.35	59.44	75.65	199
YOLOv5I	88.37	85.95	70.64	81.66	51
EfficientDetD0	93.83	84.37	89.62	89.27	28
YOLOX-s	98.86	92.96	86.41	92.72	90
YOLOX-m	99.89	93.53	91.50	94.98	63
SDNET	99.39	93.53	90.74	95.20	138
DA-YOLOv7 (pre training)	100	100	97.00	99.15	125

As indicated in Table 4, compared to other recognition models, this model exhibits higher performance on the sonar dataset SCTD.

The DA-YOLOv7 model has significant advantages when transferred to the SCTD small dataset after pre-training with the UATD large dataset. In terms of performance improvements, after pre-training with the UATD large-scale dataset, the average precision (AP) of the DA-YOLOv7 model for the three categories of aircraft, human, and ship dramatically improves. For example, compared to the model trained only on the SCTD small dataset, aircraft-AP increased from 88.16% to 100%, human-AP increased from 79.35% to 100%, and ship-AP increased from 59.44% to 97.00%, indicating a significant enhancement in the model's ability to recognize various types of targets. At the same time, using a large-capacity UATD dataset for pretraining helps the model learn more extensive visual patterns and features, improving its generalization ability, maintaining extremely high accuracy on the SCTD small dataset, and demonstrating good adaptability to new scenes and targets.

In addition, pretraining also brings a faster convergence speed. Because the model has already learned the basic features of a large amount of data, it can achieve the desired performance faster on the SCTD small dataset, saving training time and computing resources. Moreover, due to the rich diversity of the UATD dataset, the model undergoes a large amount of noise and complex situations in the pretraining stage, making it more resistant to overfitting and able to maintain stable performance on the small-scale SCTD dataset. Although the running speed (FPS) of DA-YOLOv7 is not the fastest (125 FPS), its significant advantage in accuracy achieves a good balance between accuracy and speed. The experimental results demonstrate the novelty and superiority of DA-YOLOv7. Compared to other models, DA-YOLOv7 achieves higher accuracy and faster speed, especially on the SCTD dataset with small sample sizes. The pretraining strategy on the UATD dataset and the transfer learning to the SCTD dataset effectively address the challenge of limited samples, a novel approach in this field. Additionally, the model's performance improvements in mAP@0.5 and recall rate on the SCTD dataset showcase its excellent ability to recognize targets in complex underwater environments.

For applications with extremely high accuracy requirements, the high AP value of DA-YOLOv7 far outweighs its minor disadvantage in speed. In conclusion, through pretraining with the UATD large dataset, DA-YOLOv7 performs exceptionally well on the SCTD small dataset and is an excellent choice for object recognition tasks. Figure 13 shows the improved model's enhanced recognition effect on the SCTD dataset.

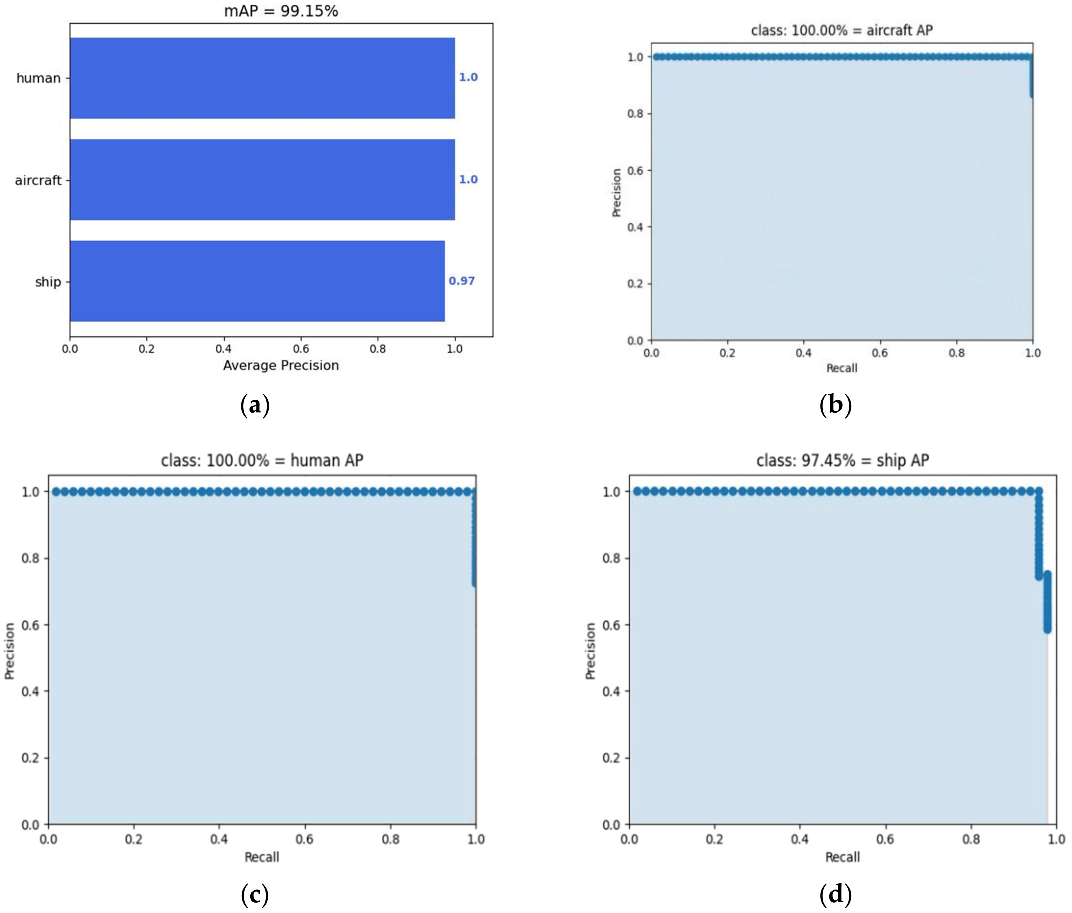
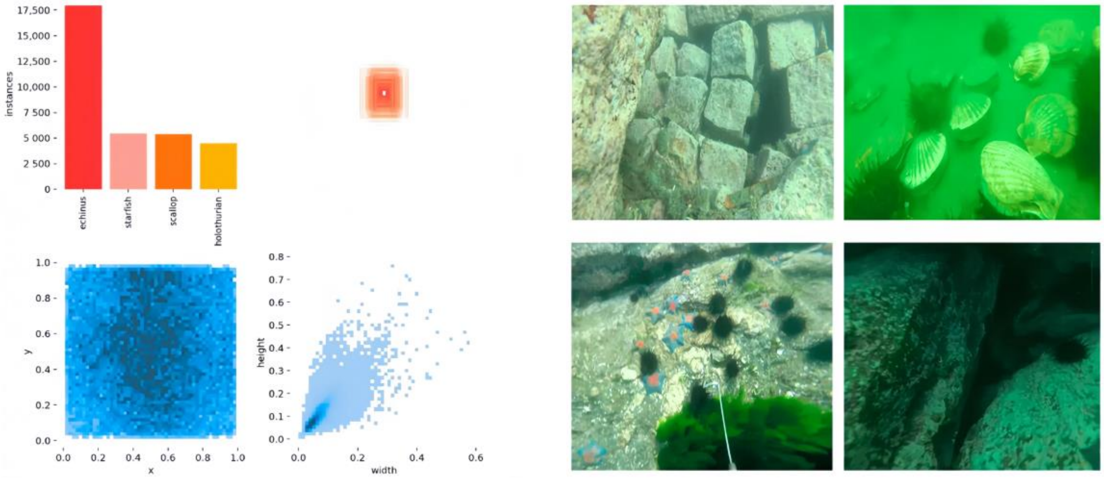


Figure 13. The effect of recognition on the SCTD dataset: (**a**) SCTD mAP Results; (**b**) SCTD aircraft-AP results; (**c**) SCTD human-AP results; (**d**) SCTD ship-AP results.

4.5. Experimental Results and Analysis of the Underwater Optical Target Detection Intelligent Algorithm Competition 2021 Dataset (URPC)

To reinforce the model's superiority and generalizability across diverse scenarios, we utilize the URPC dataset for training and evaluation. The dataset, comprising 5543 images distributed across four classes, urchins, sea cucumbers, scallops, and starfish, providing a comprehensive benchmark. The dataset is divided into a training set (80%) and a testing set (20%), totaling 4434 images for training and 1109 for testing. It captures intricate challenges, including visual occlusions due to underwater organism aggregations, variations in lighting, and motion-induced blurring, accurately reflecting the underwater environment and enhancing the model's adaptability. However, the dataset presents imbalanced category distributions and variable resolutions, posing considerable challenges during training. Figure 14a, a category statistics chart, reveals that sea urchins predominate, followed by scallops and starfish, with sea cucumbers being the least represented. The box plot indicates that the target sizes are relatively consistent, and the regularized target position map highlights the horizontal densification, contrasted with vertical dispersion. The normalized target size map further illustrates a concentration of the target sizes, mainly consisting of smaller objects. A selection of sample images from the URPC dataset is showcased in Figure 14b.



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

Figure 14. The sample information of URPC is as follows: (**a**) Labels: The upper left corner shows the distribution of categories; the upper right corner presents the visualization of all box sizes; the lower left corner indicates the distribution of the box centroid position; the lower right corner depicts the distribution of the box aspect ratio. (**b**) Example images.

Comparative Experiments on the URPC Dataset

To further substantiate the superiority of our proposed DA-YOLOv7 model, we conducted comparative analyses with renowned object recognition models, including SSD, RetinaNet, Faster-RCNN, YOLOX, YOLOv5s, YOLOv6, YOLOv7, YOLOv8n, and YOLOv8s. Both training and testing were conducted using the URPC dataset, with performance metrics such as accuracy, recall rate, and mAP@0.5 being evaluated. The comparison outcomes are presented in the table below.

It can be seen from Table 5 that on the URPC dataset, the performance of each target recognition model varies. The indicators of traditional models such as SSD, RetinaNet, and Faster-RCNN are relatively low. Among the YOLO series, YOLOX performs poorly and, although YOLOv5s, YOLOv6, YOLOv7, YOLOv8n, and YOLOv8s gradually improve in performance, there is still a significant gap compared to our DA-YOLOv7 model.

Table 5. Performance comparison of object recognition models on the URPC dataset.

Model	Precision (%)	Recall (%)	mAp@0.5 (%)
SSD	74.2	68.7	75.4
RetinaNet	75.2	66.8	73.4
Faster-RCNN	78.8	73.1	76.1
YOLOX	73.3	64.1	69.5
YOLOv5s	78.9	75.3	80.8
YOLOv6	81.7	79.1	80.4
YOLOv7	82.9	78.3	83.2
YOLOv8n	83.8	79.2	85.7
YOLOv8s	86.4	82.1	87.1
DA-YOLOv7	90.0	84.4	89.9

Our DA-YOLOv7 model has extremely remarkable advantages. Its precision is as high as 90.0%, far exceeding other models, so it can identify targets more accurately and reduce misjudgments. The recall rate is 84.4%, indicating that it can effectively capture more real targets and reduce the situation of missed detections. The mAp@0.5 reaches 89.9%, demonstrating its stable and excellent detection performance under different thresholds. The DA-YOLOv7 model can perform outstandingly because of its unique architectural design and elaborate optimization strategy. It adopts advanced feature extraction and fusion techniques, extracting key and discriminative features from complex data, thereby achieving precise positioning and recognition of targets. At the same time, its efficient training algorithm and parameter adjustment enable the model to maintain excellent performance and generalization ability when facing various complex scenes and target changes. Overall, the DA-YOLOv7 model has shown a decisive advantage and excellent comprehensive performance, far exceeding other models on the URPC dataset.

Figure 15 vividly showcases the recognition performance of the updated model in four distinct scenarios. It demonstrates the model's remarkable capacity for precise recognition even under arduous underwater conditions. Whether dealing with sparse or dense targets, the model exhibits its superiority. In sparse target situations, it accurately identifies and locates individual targets. In dense target scenarios, it can distinguish multiple targets despite the complexity.

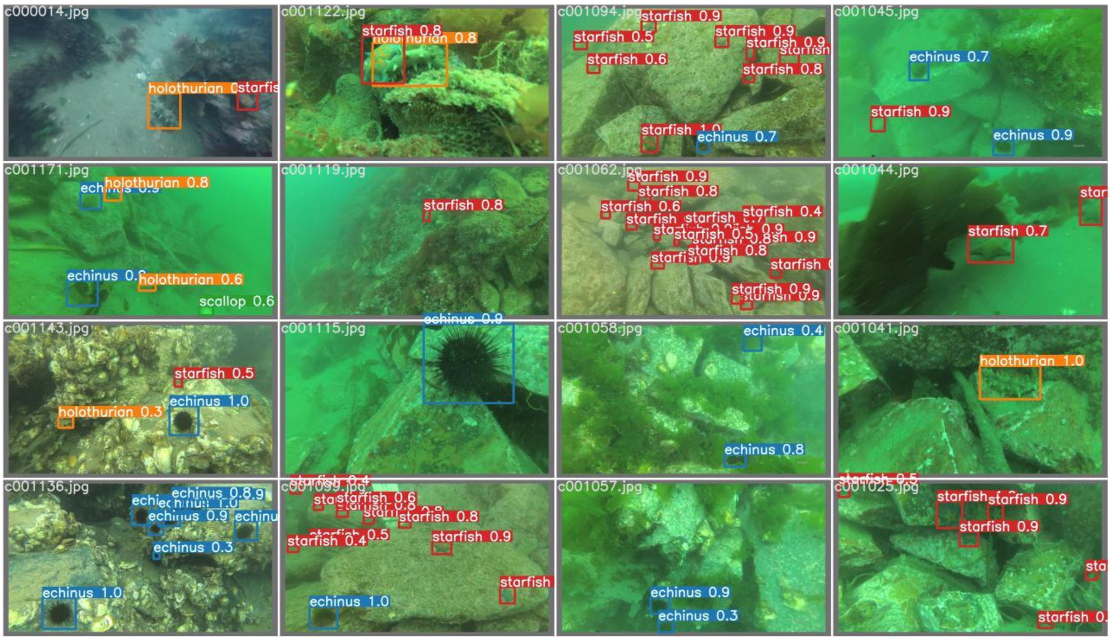


Figure 15. Recognition results in multiple underwater scenes.

The performance of the updated model in these four scenarios highlights its advanced design and training. This advanced design and training hold great promise for various applications in underwater exploration, marine biology research, and underwater robotics. The model's ability to handle different underwater conditions with precision and reliability makes it a valuable tool for these fields, enabling more accurate exploration, better understanding of marine life, and improved performance of underwater robots.

The DA-YOLOv7 model successfully identified the targets in diverse underwater scenes, manifesting its robust recognition performance. As depicted in Figure 16, with the augmentation in the number of iterations, the values exhibit a stable descending trend and eventual stabilization, attaining convergence after 200 iterations.

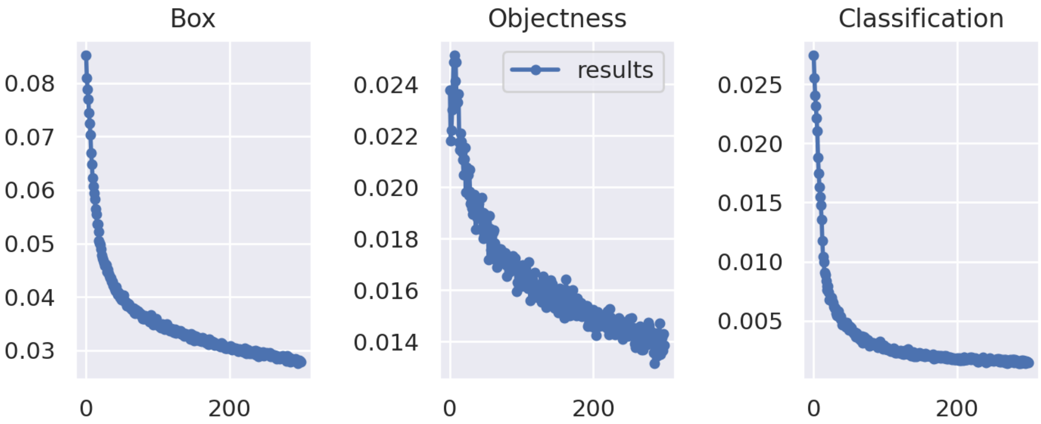


Figure 16. RCurve of loss value changes on the UPRC dataset.

5. Conclusions

This study proposed an improved DA-YOLOv7 model based on YOLOv7 to tackle the challenges in target recognition within complex underwater environments and the issue of low recognition accuracy resulting from small samples of underwater sonar images. The following significant results were obtained:

1. With innovative improvements in YOLOv7, the OA-ELAN module, the SPPCSPCBiFormer module, and the G-ELAN-H module were introduced, significantly enhancing the model's performance. The OA-ELAN module augments the ability to extract local features and capture critical information, thereby improving the recognition accuracy of features such as object shapes and edges; the SPPCSPCBiFormer module enhances the ability to capture long-range correlations and contextual information in the input feature map, enhancing the processing and generalization ability of multi-scale targets; the G-ELAN-H module boosts the model's adaptability to various target sizes and shapes, further enhancing the generalization performance.
2. For the problem of small samples of underwater sonar images, pre-training and transfer learning methods were adopted. The large-scale UATD dataset was used to pre-train DA-YOLOv7 to obtain pre-trained weights, which were then applied to the training of the small sample SCTD dataset, effectively avoiding overfitting and improving the generalization ability of small sample data.
3. Experiments on the UATD, URPC, and SCTD datasets showed that DA-YOLOv7 performed excellently in accuracy, recall rate, and running speed. Especially in the mAP@0.5 metric, it improved by 8.2% compared to YOLOv7, reaching 89.4%, and the running speed was 97.09 FPS, achieving a balance between accuracy and real-time performance, establishing a new benchmark for underwater sonar image target recognition.
4. DA-YOLOv7 demonstrated excellent robustness and accuracy in different underwater environments, such as the lake bottom and shallow water areas of UATD, the diverse biological communities of URPC, and the ocean test scenarios of SCTD, showing broad application potential.

Future work on DA-YOLOv7 includes reducing the dependence on training data, enhancing the model's adaptability to new environments and targets, improving the model's anti-noise ability and computational efficiency to adapt to resource-constrained environments, simplifying the model structure to increase maintainability, exploring more effective data augmentation and balancing strategies to enhance generalization ability, testing the model's generalization performance on a more diverse underwater sonar image dataset, and developing more effective noise suppression techniques, optimizing the model structure to reduce computational costs, and exploring data augmentation strategies to improve generalization ability and model interpretability. These efforts aim to optimize the network structure further, enhance the anti-noise ability, adapt to more diverse underwater environments, and support the ability of automated monitoring and resource management in underwater environments.

Author Contributions: Conceptualization, Z.C. and H.Q.; methodology, X.D. and G.X.; software, J.P.; validation, X.D., J.P. and H.Q.; investigation, Z.C. and G.X.; writing—original draft preparation, Z.C.; writing—review and editing, H.Q. and X.D.; funding acquisition, X.D. and H.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Guangxi Science and Technology Base and Talent Project (No. GuikeAD21220098) and the 2021 Open Fund project of the Key Laboratory of Cognitive Radio and Information Processing of the Ministry of Education (No. CRKL210102). We also thank to the support of Guangxi Key Laboratory of Wireless Broadband Communication and Signal Processing Project (No. GXKL06240114), Beihai City Science and Technology Bureau Project (No. Bei ke He 2023158004) and the Innovation Project of Guangxi Graduate Education (No. YCSW2024344).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

Abbreviations

The following abbreviations are used in this manuscript:

DA-YOLOv7	The Dual Attention Mechanism YOLOv7 model
OA-ELAN	Omni-Directional Convolution Channel Prior Convolutional Attention Efficient Layer Aggregation Network
SPPCSPCBiFormer	Spatial Pyramid Pooling Channel Shuffling and Pixel-level Convolution Bilateral-branch Transformer
G-ELAN-H	Ghost-Shuffle Convolution Enhanced Layer Aggregation Network-High performance
DCNNs	Deep convolutional neural networks
R-CNN	Regions with Convolutional Neural Network features
DCN	Deformable Convolutional Networks
STAFNet	Swin Transformer Based Anchor-Free Network
URPC	The Underwater Optical Target Detection Intelligent Algorithm Competition 2021 Dataset
FCOS	Fully Convolutional One-Stage Object Detection
VGG16	Visual Geometry Group 16-layer network
IMA	Invert Multi-Class AdaBoost
LWAP	Local Wavelet Acoustic Pattern
MLP	Multilayer Perceptron
YOLO	You Only Look Once
ASFF	Adaptive Spatial Feature Fusion
ScEMA	Scale and Channel Efficient Module Attention
ODConv	Omni-Directional Convolution
CPCA	Channel Prior Convolutional Attention
UATD	Underwater Acoustic Target Detection Dataset
SCTD	Smaller Common Sonar Target Detection Dataset
CBS	Convolution-Batch Normalization-SiLU activation function
ELAN	Efficient Layer Aggregation Network
MP	The Multi-scale Layer
SPPCSPC	Spatial Pyramid Pooling Channel Shuffling and Pixel-level Convolution
BiFormer	The Bilateral-branch Transformer
SPP	Spatial Pyramid Pooling
CSPC	Channel Shuffling and Pixel-level Convolution

References

1. Yin, Z.; Zhang, S.; Sun, R.; Ding, Y.; Guo, Y. Sonar image target detection based on deep learning. In Proceedings of the 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballar, India, 29–30 April 2023; pp. 1–8.
2. Huang, C.; Zhao, J.; Zhang, H.; Yu, Y. Seg2Sonar: A Full-Class Sample Synthesis Method Applied to Underwater Sonar Image Target Detection, Recognition, and Segmentation Tasks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–19. [CrossRef]
3. Steiniger, Y.; Groen, J.; Stoppe, J.; Kraus, D.; Meisen, T. A study on modern deep learning detection algorithms for automatic target recognition in sidescan sonar images. In Proceedings of the 6th Underwater Acoustics Conference and Exhibition, Virtual, 21–24 June 2021; pp. 1–6.
4. Li, J.; Chen, L.; Shen, J.; Xiao, X.; Liu, X.; Sun, X.; Wang, X.; Li, D. Improved neural network with spatial pyramid pooling and online datasets preprocessing for underwater target detection based on side scan sonar imagery. *Remote Sens.* **2023**, *15*, 440. [CrossRef]
5. Vijaya Kumar, D.T.T.; Mahammad Shafi, R. A fast feature selection technique for real-time face detection using hybrid optimized region based convolutional neural network. *Multimed. Tools Appl.* **2022**, *82*, 13719–13732. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
7. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Chen, Z.; Wang, H.; Shen, J.; Dong, X. Underwater Object Detection by Combining the Spectral Residual and Three-Frame Algorithm. In *Advances in Computer Science and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1109–1114.
9. Villar, S.A.; Acosta, G.G.; Solari, F.J. OS-CFAR Process in 2-D for Object Segmentation from Side-scan Sonar Data. In Proceedings of the 2015 XVI Workshop on Information Processing and Control (RPIC), Cordoba, Argentina, 6–9 October 2015; pp. 1–6.
10. Mukherjee, K.; Gupta, S.; Ray, A.; Phoha, S. Symbolic Analysis of Sonar Data for Underwater Target Detection. *IEEE J. Ocean. Eng.* **2011**, *36*, 219–230. [CrossRef]
11. Midtgaard, O.; Hansen, R.E.; Sæbø, T.O.; Myers, V.; Dubberley, J.R.; Quidu, I. Change Detection Using Synthetic Aperture Sonar: Preliminary Results from the Larvik Trial. In Proceedings of the OCEANS'11 MTS/IEEE KONA, Waikoloa, HI, USA, 19–22 September 2011; pp. 1–6.
12. Raghuvanshi, D.S.; Dutt, I.; Vaidya, R.J. Design and analysis of a novel sonar-based obstacle-avoidance system for the visually impaired and unmanned systems. In Proceedings of the 2014 International Conference on Embedded Systems (ICES), Coimbatore, India, 3–5 July 2014; pp. 238–243.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
14. 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.
15. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Wei, X. YOLOv6: A single-stage object detection framework for industrial applications. *J. Mar. Sci. Eng.* **2022**, *11*, 67721.
16. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
17. Fan, Z.; Xia, W.; Liu, X.; Li, H. Detection and segmentation of underwater objects from forward-looking sonar based on a modified Mask R-CNN. *Signal Image Video Process.* **2021**, *15*, 1135–1143. [CrossRef]
18. Zhu, X.; Liang, Y.; Zhang, J.; Chen, Z. STAFNet: Swin Transformer Based Anchor-Free Network for Detection of Forward-looking Sonar Imagery. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022. [CrossRef]
19. Zhou, H.; Huang, H.; Yang, X.; Zhang, L.; Qi, L. Faster R-CNN for marine organism detection and recognition using data augmentation. In Proceedings of the International Conference on Video and Image Processing, Singapore, 27–29 December 2017; pp. 56–62.
20. 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; pp. 1–8.
21. 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]
22. Yan, W. Sonar Image Target Detection and Recognition Based on Convolution Neural Network. *Mob. Inf. Syst.* **2021**, *2021*, 5589154. [CrossRef]
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
24. Fan, X.; Lu, L.; Shi, P.; Zhang, X. A novel sonar target detection and classification algorithm. *Multimed. Tools Appl.* **2022**, *81*, 10091–10106. [CrossRef]

25. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
26. Zhang, H.; Tian, M.; Shao, G.; Cheng, J.; Liu, J. Target Detection of Forward-Looking Sonar Image Based on Improved YOLOv5. *IEEE Access* **2022**, *10*, 18023–18034. [CrossRef]
27. Cheng, C.; Hou, X.; Wen, X.; Liu, W.; Zhang, F. Small-Sample Underwater Target Detection: A Joint Approach Utilizing Diffusion and YOLOv7 Model. *Remote Sens.* **2023**, *15*, 4772. [CrossRef]
28. Zheng, L.; Hu, T.; Zhu, J. Underwater Sonar Target Detection Based on Improved ScEMA-YOLOv8. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–4. [CrossRef]
29. Xie, K.; Yang, J.; Qiu, K. A Dataset with Multibeam Forward-Looking Sonar for Underwater Object Detection. *Sci. Data* **2022**, *9*, 739. [CrossRef] [PubMed]
30. Zhou, Y.; Chen, S.; Wu, K.; Ning, M.; Chen, H.; Zhang, P. SCTD1.0: Common Sonar Target Detection Dataset. *Ship Sci. Technol.* **2021**, *48*, 334–339.
31. Dong, J.; Yang, M.; Xie, Z.; Cai, L. Overview of Underwater Image Object Detection Dataset and Detection Algorithms. *J. Ocean. Technol.* **2022**, *41*, 60–72.
32. Ren, S.; Zhou, D.; He, S.; Feng, J.; Wang, X. Shunted self-attention via multi-scale token aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 7–8 May 2022; pp. 10853–10862.

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.