Journal of *Marine Science and Engineering*

MDPI

*Article*

# Underwater Target Tracking Method Based on Forward-Looking Sonar Data

**Wenjing Zeng [1], Renzhe Li [2], Heng Zhou [2,3] and Tiedong Zhang [4,*]**

[1] Zhuhai College of Science and Technology, Zhuhai 519082, China; zenwenjing@163.com
[2] Science and Technology on Underwater Vehicles Laboratory, Harbin Engineering University, Harbin 150001, China; lirenzhe@hrbeu.edu.cn (R.L.); zhouheng@cssrc.com.cn (H.Z.)
[3] China Ship Scientific Research Center, Wuxi 214000, China
[4] School of Ocean Engineering and Technology, Sun Yat-sen University, Zhuhai 519082, China
* Correspondence: zhangtd5@mail.sysu.edu.cn

**Abstract:** Underwater dynamic targets often display significant blurriness in their forward-looking sonar imagery, accompanied by sparse feature representation. This phenomenon presents several challenges, including disturbances in the trajectories of underwater targets and alterations in target identification throughout the tracking process, thereby complicating the continuous monitoring of moving targets. This research proposes a new framework for underwater acoustic data interpolation and underwater object tracking. Considering the character of underwater acoustic images, a Swin Transformer is integrated in the architecture of the YOLOv5 network; then, an improved Deep Simple Online and Real-time Tracking method is developed. By enlarging the bounding box output generated by the detector and subsequently integrating it into the tracker, the sensing horizon of the tracker is broadened. This strategy enables the extraction of noise features surrounding the target, thereby augmenting the target's characteristics and improving the stability of the tracking process. The experimental results demonstrate that the proposed method effectively reduces the frequency of changes in target identification numbers, minimizes the occurrence of trajectory interruptions, and decreases the overall percentage of trajectory interruptions. Additionally, it significantly enhances tracking stability, particularly in scenarios involving intersecting target paths and encounters.

**Keywords:** underwater perception; forward-looking sonar (FLS); sonar imaging; target detection; target identification number (TIN); deep simple online and real-time tracking method (DeepSORT)

check for updates

Received: 3 January 2025
Revised: 15 February 2025
Accepted: 20 February 2025
Published: 25 February 2025

**Citation:** Zeng, W.; Li, R.; Zhou, H.; Zhang, T. Underwater Target Tracking Method Based on Forward-Looking Sonar Data. *J. Mar. Sci. Eng.* **2025**, *13*, 430. https://doi.org/10.3390/jmse13030430

**Copyright:** © 2025 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

Underwater motion target-sensing technology holds immense potential across diverse applications, including marine scientific exploration and geophysical surveys. However, the inherent complexities of the underwater environment present significant challenges for its implementation. Current underwater target sensing methods primarily fall into two categories: optical vision or acoustic vision detection. Optical vision detection, often employing underwater cameras, captures detailed images of targets, revealing intricate features of the underwater terrain. However, the rapid attenuation of light in water severely limits the field of view and compromises image quality, particularly in turbid conditions. In contrast, acoustic waves experience slower decay in water, enabling a wider field of view that is less susceptible to water quality degradation. This advantage makes acoustic

detection a robust and effective solution for tasks in turbid waters and long-distance target identification.

Sonar imagery is notably different from optical imagery due to its higher noise levels, and targets within sonar imagery are prone to considerable deformation during movement, resulting in significant changes to their appearance. Moreover, the target region often exhibits sparse features, which complicates the extraction of meaningful information for tracking purposes using feature extraction networks. Traditional tracking methods often rely on denoised sonar images; however, the scattered noise surrounding the target, which frequently arises from the target's shape or material, may contain valuable feature information that is essential for accurate tracking. In contrast to conventional feature extraction techniques, neural network methodologies are capable of extracting high-level semantic features relevant to the target through various configurations of network layers, thus enabling end-to-end feature extraction. Acknowledging these challenges, this paper introduces a new framework for tracking underwater objects based on acoustic vision images and presents an enhanced tracking method based on DeepSORT. The characteristics of underwater acoustic images are systematically analyzed, with YOLOv5 serving as the foundational network. To further enhance the network's capability for global image information fusion, the block architecture derived from the Swin Transformer network is integrated. The results indicate that the proposed method effectively prevents interruptions in trajectory and alterations in identity during the process of underwater target tracking, thereby optimizing the tracking performance of underwater acoustic targets. The main contributions of this paper can be summarized as follows:

(1) In view of the problems of low contrast, high noise levels in underwater acoustic images, and the insufficient global image information fusion capability in the CNN structure of the YOLOv5 network, the C3 structure in the network was replaced with an STR block to improve the detection accuracy of objects in the sonar images.
(2) A DeepSORT tracking improvement framework based on extended bounding boxes was designed. For issues such as trajectory interruption and ID switch in the field of underwater target tracking, an extension strategy for detecting bounding boxes was proposed, and the features of the scattered noise around the target were extracted to compensate for the sparsity of the target's own features. Experimental comparison results show that the proposed method can effectively suppress trajectory interruption and ID switch issues during the tracking process, and improves the stability of the tracking network.

## 2. Related Work

The inherent physical properties of sound waves impose several limitations on acoustic imaging. These limitations include significant noise interference, abrupt luminance variations, ambiguous target regions, and target profile deformation. Such challenges often hinder the effectiveness of certain target tracking methodologies. Consequently, these issues have become the focus of extensive research.

### *2.1. Traditional Target Tracking Methods of Acoustic Images*

Williams [1] proposed a sophisticated tracking method that employs a Kalman filter (KF) to discern multiple targets by analyzing clustered sonar returns, effectively tracking both stationary and moving obstacles. This approach laid the groundwork for subsequent advancements in the field. Furthering this research, Dai [2] utilized temporal feature measures to provide a detailed quantitative analysis of moving target behavior across several scans, a technique that was successfully validated through tracking experiments involving a diver. Continuing the evolution of tracking methodologies, Lane [3] discussed

an innovative tracking approach for underwater targets, drawing on the principles of optical flow theory. By constructing a tracking tree to manage and store tracking information, Lane's method significantly bolstered the robustness and reliability of the tracking process. Chantler [4] and Ruiz [5] introduced distinct methodologies for classification and obstacle tracking, focusing on the robustness of inter-frame feature measure classifiers for underwater sector scan sonar images. Building on this foundation, Petillot [6] developed a tracker that integrates segmentation and object-based feature extraction, enhancing tracking accuracy and robustness through the application of an extended Kalman filter. Dandan [7] proposed a target tracking method based on variable image templates, where target features were extracted using surfacelet transform, and a particle filter was employed to estimate the moving state of targets. Perry [8,9] harnessed machine learning techniques for underwater target detection, leveraging the self-learning capabilities of neural networks to analyze feature variations in acoustic images and effectively distinguish target regions. Blanding [10] minimized false target alarms by integrating intermittent sensor data from unmanned underwater vehicles with launch platform data within a maximum likelihood–probabilistic data association tracking framework. Clark [11,12] improved underwater target tracking with a method based on the probability hypothesis density filter, fusing predicted target positions with trajectory data. Their experiments demonstrated superior tracking stability compared to the Kalman filter. Darko [13] investigated the performance of three multitarget trackers in active sonar environments, presenting simulation studies of these algorithms. Handegard [14] presented an automatic tracking method for fish populations using a forward-looking sonar (FLS), evaluating the tracker across three test datasets with varying target sizes, observation ranges, and densities. Yue [15] proposed a single-target tracking method for non-complex backgrounds, combining particle filters with correlation matching techniques. Quidu [16] utilized statistical deviations in small patches of acoustic vision sequence information to detect targets in front of AUVs, implementing object tracking algorithms with multiple Kalman filters to simultaneous monitor several objects. Their experimental results aligned with theoretical analysis. DeMarco [17] discussed diver detection and tracking using a high-frequency FLS, achieving cluster classification by matching observed trajectories with trained hidden Markov models, effectively distinguishing divers from stationary targets amidst noisy sonar imagery. Karoui [18] proposed an efficient tool for detecting and tracking sea surface obstacles through FLS image processing, achieving target tracking in Cartesian coordinates using the Kalman filter and object association via the joint probabilistic data association filter, with results derived from real data. Hurtós [19] presented two FLS-based detectors combined with strategic planning and control to detect, follow, and map underwater chains, with experimental trials showing satisfactory accuracy. AI Muallim [20] proposed a robust wake detection algorithm to enhance diver tracking in acoustic vision, fine-tuning the Kalman filter to achieve stable diver tracks in tests. Danxiang [21] introduced an innovative dual-frequency identification sonar imaging technique, employing the Nearest Neighbor search algorithm with extended Kalman filtering, achieving a recognition error rate of less than 5%. Xiufen [22] introduced a tracking methodology for small moving targets based on a FLS, significantly enhancing tracking accuracy and real-time performance. Xingmei [23] presented an adaptive particle swarm optimization algorithm for tracking multiple underwater objects, demonstrating superior accuracy and speed. Fuchs [24] established a classifier using a Convolutional Neural Network, showing that transfer learning is effective for underwater object classification. Mingwei [25] proposed a cloud-like model for data association in multiple targets, integrating the inherent ambiguity and randomness of qualitative concepts, with their experiments indicating enhanced clustering accuracy over traditional algorithms. Danxiang [26] proposed a multitarget tracking methodology integrating Sequential Monte

Carlo Probability Hypothesis Density filtering with the auction track recognition algorithm, though it showed limitations under certain conditions. Tiedong [27] described an online processing framework based on FLS images and presented a novel tracking approach using a Gaussian particle filter to address persistent multiple-target tracking in cluttered environments. Jue [28] proposed a method for underwater target identification using local features and a feature tracking algorithm for acoustic image sequences, with experiments confirming the algorithm's ability to accurately track potential targets. Christensen [29] introduced an automated boulder detection system employing deep learning techniques, showing not only significantly faster processing times than human operators but also high accuracy.

### 2.2. Deep Learning-Based Target Tracking Methods of Acoustic Images

In recent years, deep learning methods, leveraging their outstanding feature modeling capabilities, have achieved significant success in the field of target tracking. Horimoto [30] utilized YOLOv2 to detect turtles in multibeam sonar imagery, and the method successfully detected turtles in sonar images with 88% precision and 64% recall. Yue [31] proposed a lightweight convolutional neural network tracking method based on a two-layer convolutional neural network, as well as a target tracking method based on an improved fully convolutional Siamese network using AlexNet. Compared to deep neural networks, the lightweight convolutional neural network and the improved fully convolutional Siamese network he proposed can address the deformation caused by the flipping motion of frogmen during their movement. Igor [32] focused on finding a robust and reliable sonar image processing method for the detection and tracking of human divers using convolutional neural networks. The performance of these algorithms was compared on a set of sonar recordings to determine their reliability and applicability in a real-time operation. Xinglong [33] proposed a method for detecting small underwater targets and their shadows. By adopting improved tracking based on rotation estimation, the problem of losing track of distant targets due to the movement of the sonar carrier was solved. Zhikang [34] proposed a method for forward-looking sonar target tracking and positioning based on an improved Siamese network. It solves the problems of target detection and tracking under conditions of low image quality and low information density in forward-looking sonars, and achieves accurate, coherent, and stable detection and tracking. Xiufen [35] proposed a Mobilenetv3-YOLOv4-Sonar algorithm. It was improved by a Convolutional Block Attention Module and modified Squeeze-and-Excitation Network. Experiments illustrate that the accuracy of the model increased. In view of the problem that the efficient detection model Single Shot Detector MobileNet v2 has low detection accuracy in underwater multi-scale targets in acoustic images, Baoqi [36] proposed a feature extraction module Extended Selective Kern, and the experimental results show that the method is suitable for acoustic underwater multi-scale target detection tasks. Yongcan [37] proposed a real-time automatic target recognition method for acoustic images, and a transformer module and YOLOv5s network were put forward to meet the requirements of accuracy and efficiency for underwater target recognition. Ruoyu [38] optimized the one-stage detection algorithm YOLO by combining Swin Transformer blocks and layers. The results verified that the combination of the lightweight network YOLO and the well-performing network Swin Transformer can achieve more accurate detection precision and meanwhile meet the requirements of real-time detection. Xiang [39] used the YOLOv3 network to determine obstacle candidate areas in a sonar image. Their experimental results showed that the proposed algorithms improved object detection accuracy and the processing speed of sonar images. Xin [40] proposed a single transformer-based generative adversarial network to improve the accuracy of analysis tasks on sonar images. The results showed that it achieved

significantly better despeckling performance. Sehwa [41] proposed a novel method for the 3D detection and tracking of moving marine life, utilizing YOLOv3 to identify the target's position in sonar imagery. The method was validated through tank experiments. Ken Sinkou [42] proposed an object detection method based on YOLOv7C. The backbone network and neck network were improved to handle complex backgrounds in sonar images. Their experimental results showed that the proposed algorithms maintained high detection accuracy.

According to the above existing research, it can be found that due to the limitations of underwater target sample data, there are still relatively few deep learning-based studies in the field of acoustic image target detection and tracking, especially for the dynamic tracking of small targets in acoustic images. Although many scholars have carried out research on underwater target detection and tracking based on underwater optical sensors, sonar images themselves have characteristics such as a low pixel count, low resolution, poor imaging quality (significantly affected by noise), low contrast, and blurred edges. These factors directly impact the research on sonar image target detection and tracking, particularly for small moving targets. So, these deep learning networks should be applied in the acoustic images further, and many algorithms still have ample room for improvement. Therefore, the target tracking and detection technology of sonar images still needs further research.

## 3. FLS Overview

The acoustic images are obtained using BlueView Sonar, a type of FLS developed by Teledyne [43]. The sonar is a compact fully featured 2D multibeam imaging sonar characterized by its low power consumption (Figure 1). It is designed to provide optimal acoustic performance to enhance image quality and range capabilities. Consequently, it can operate effectively while in motion or from a stationary position, thereby delivering real-time imagery and data. The detailed specifications of the sonar are presented in Table 1.

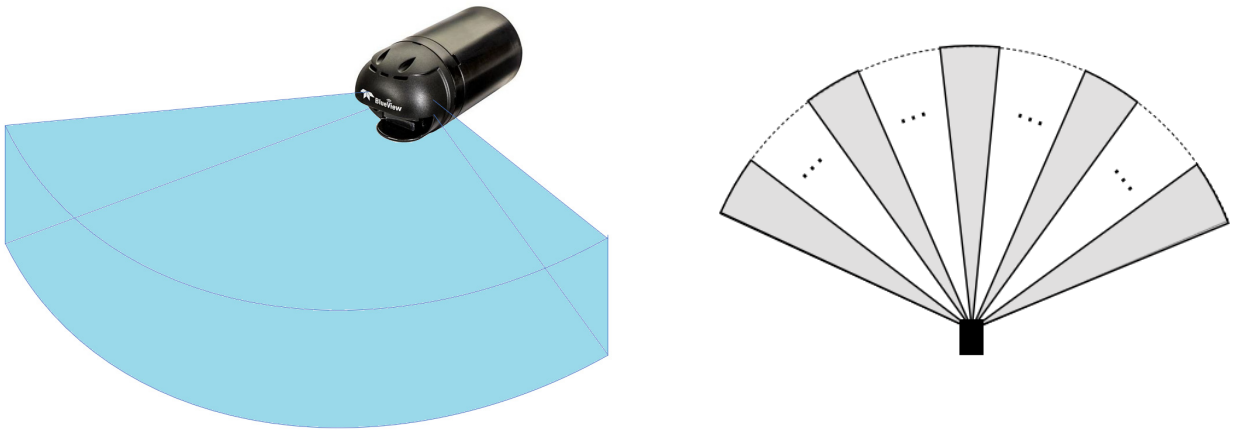


**Figure 1.** Diagram showing scanning procedure [44].

**Table 1.** Sonar specifications.

| Operating Frequency | Horizontal Beam Width | Vertical Beam Width | Maximum Range | Range Resolution | Field of View | Weight |
|---|---|---|---|---|---|---|
| 900 KHz | 1° | 12° | 100 m | 1.3 cm | 130° | 2.5 kg in air, 1.2 kg in water |

Acoustic images are generated based on the intensity of the echoes received from the three-dimensional spatial environment. Although this imaging sonar offers a significant

advantage in terms of range compared to conventional visual methods, it suffers from several drawbacks:

(1) The physical limitations associated with the transducer size restrict the quantity of transducers that can be incorporated into an array. As a result, the resolution of the images produced by the FLS is compromised, leading to a diminished grayscale representation of the target area. This reduction in resolution complicates the process of discerning finer details within the target.

(2) The scattering properties of different regions on the target surface demonstrate variability, which is affected by factors such as shape, material composition, and the spatial relationship between the target and the sonar system. Furthermore, the angle at which acoustic waves strike the target may change as a result of the target's movement, leading to the emergence of distinct regions within the acoustic image of the same target. These regions frequently appear as disjointed segments in acoustic imagery.

(3) Multipath propagation is a significant phenomenon in acoustic imaging, characterized by the occurrence of reflected acoustic waves that may exhibit higher energy levels than those reflected from obstacles. This phenomenon can result in the inaccurate or incomplete detection of targets, thereby complicating the processing of acoustic images.

To illustrate these concepts, Figure 2 presents the acoustic image captured under underwater conditions, emphasizing the distinct characteristics of acoustic images relative to optical images. As a result, it is imperative to modify specific image-processing methodologies that are conventionally utilized for optical images in order to attain optimal results in acoustic imaging.

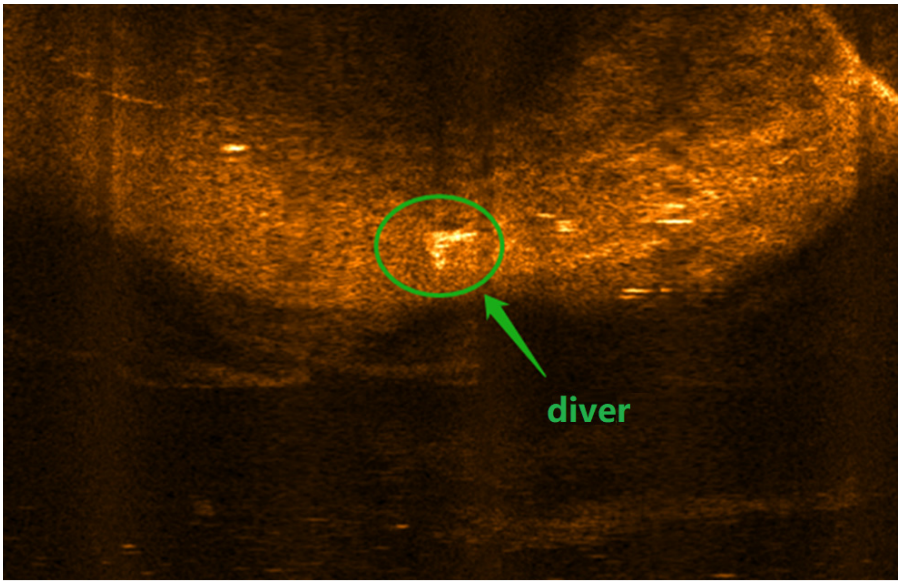



**Figure 2.** Acoustic image of target.

## 4. Design of YOLOv5 Network Model for Acoustic Object Tracking

### *4.1. Swin Transformer Basic Principle*

The Swin Transformer (STr) architecture was developed in 2021 [45], and successfully achieves a favorable balance between accuracy and inference speed, thus enhancing its applicability for tasks such as target detection and instance segmentation. The overall STr structure is illustrated in Figure 3a.

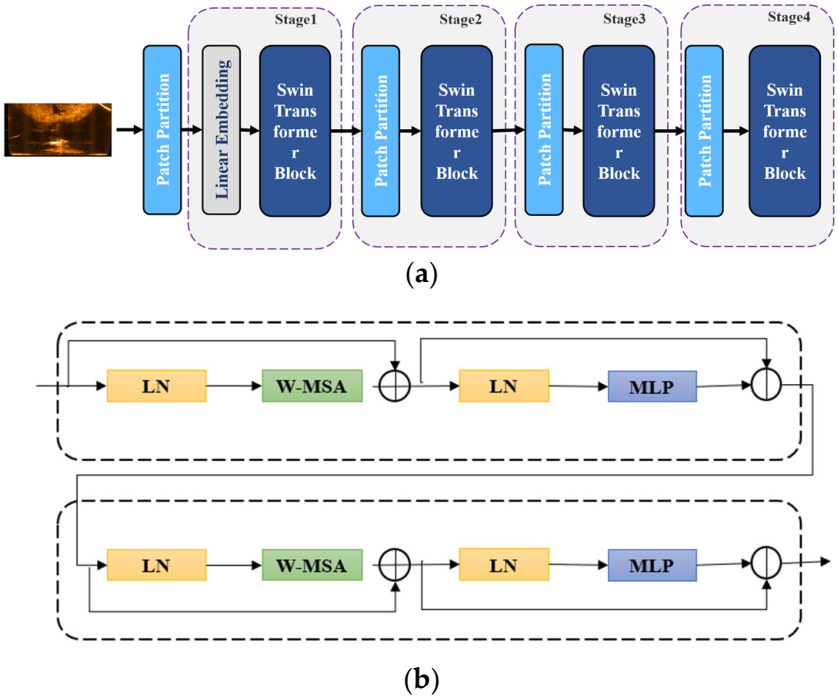



**Figure 3.** (**a**) STr architecture [45]. (**b**) STr block [45].

The STr architecture consists of four distinct stages, each integrating a Linear Embedding layer in conjunction with an STr block configuration. The details of the STr block structure are presented in Figure 3b.

### *4.2. Improvement in YOLOv5 Based on Swin Transformer*

The YOLO network was first introduced in 2016 [46]. Subsequently, in 2020, YOLOv5 was released [47], which improved upon the YOLO detection architecture by integrating various algorithmic optimization techniques derived from CNNs. An adaptive mechanism is incorporated to calculate anchor frames in YOLOv5, so the initial anchor frame size can be adjusted in response to variations in the dataset. The architecture of the YOLOv5 network is primarily structured around three main components: the backbone, neck, and head. The fundamental framework is illustrated in Figure 4.

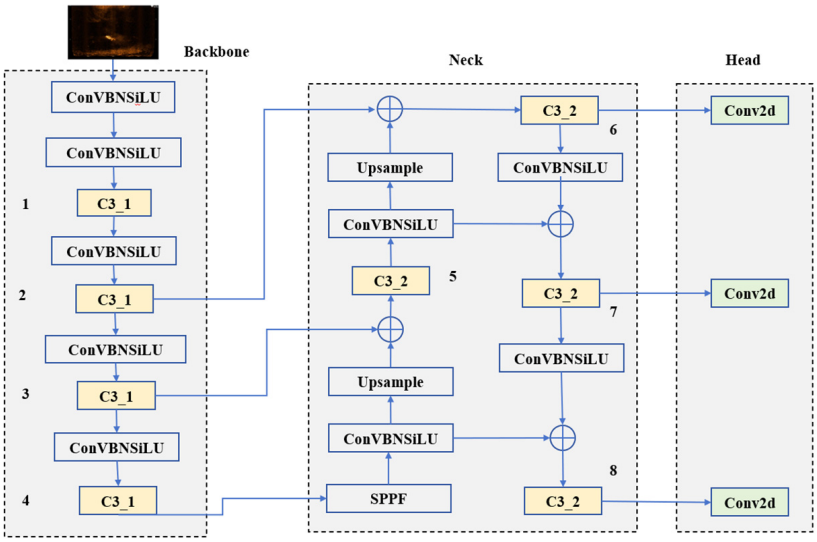



**Figure 4.** YOLOv5 architecture.

In Figure 4, there are eight C3 structures in the YOLOv5 architecture. Those located in the backbone are designated for feature extraction, and those situated in the neck are employed for feature fusion. For the purpose of clarity, these C3 structures are assigned numerical identifiers. Those in the backbone are labeled as C3_1, and those in the neck are labeled as C3_2. Due to the adaptable nature of the STr block, it can be customized to meet diverse network requirements. As a result, the C3 structures in YOLOv5 are intended to be replaced with STr blocks, thereby improving the model's ability to integrate global information. Subsequently, three distinct replacement strategies are developed, as shown in Figure 5.

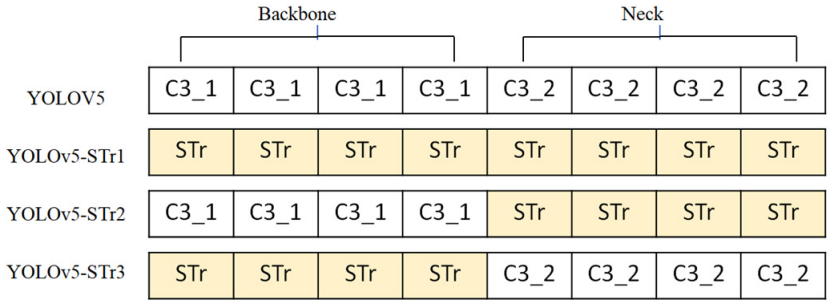



**Figure 5.** YOLOv5 incorporating STr block.

### *4.3. Experiment Results and Analysis*

To evaluate the influence of the STr block on feature representation at different stages of the network, a comprehensive comparative analysis was performed. This analysis encompassed three refined schemes and five distinct configurations of the YOLOv5 network architecture. Mean average precision (mAP), recall (Re), and frames per second (Fps) were applied to estimate the accuracy of the current metamodel. The results are presented in Table 2, Figures 6–8.

**Table 2.** Comparison of test results of different networks.

| Network Architecture | mAP (IOU = 0.5) | Re (IOU = 0.65) | Fps |
|---|---|---|---|
| YOLOv5n (Y5n) | 92.6% | 91.1% | 30.6 |
| YOLOv5s (Y5s) | 93.5% | 93.6% | 29.8 |
| YOLOv5m (Y5m) | 93.3% | 93.1% | 29.8 |
| YOLOv5l (Y51) | 92.3% | 93.9% | 29.4 |
| YOLOv5x (Y5x) | 91.9% | 93.9% | 20.1 |
| YOLOv5-STr1 (Y5S1) | 91.4% | 89.4% | 27.7 |
| YOLOv5-STr2 (Y5S2) | 93.3% | 93.7% | 28.2 |
| YOLOv5-STr3 (Y5S3) | 94.4% | 93.6% | 28.0 |

As presented in Table 2, the Y5S3 model attains a notable mAP of 94.4%, surpassing the performance of the conventional YOLOv5 architecture. It is shown that the substitution of the C3 component in the backbone with the STr block architecture significantly enhances mAP performance. In contrast, the mAP obtained by the Y5S2 model is 93.3%, which is similar to the performance of the standard YOLOv5 network. However, the Y5S1 model does not exhibit significant improvements and generally performs below the average relative to the other models.

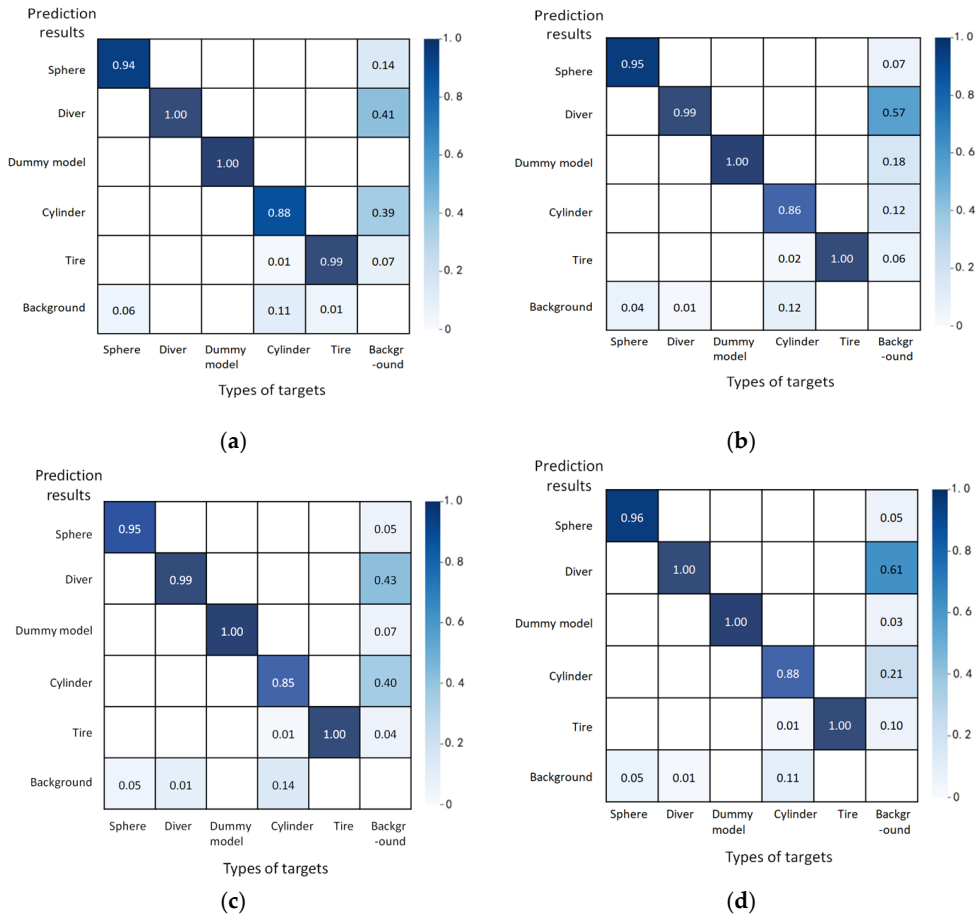



**Figure 6.** Confusion matrix results. (**a**) Results obtained by Y5s; (**b**) results obtained by Y5S1; (**c**) results obtained by Y5S2; (**d**) results obtained by Y5S3.

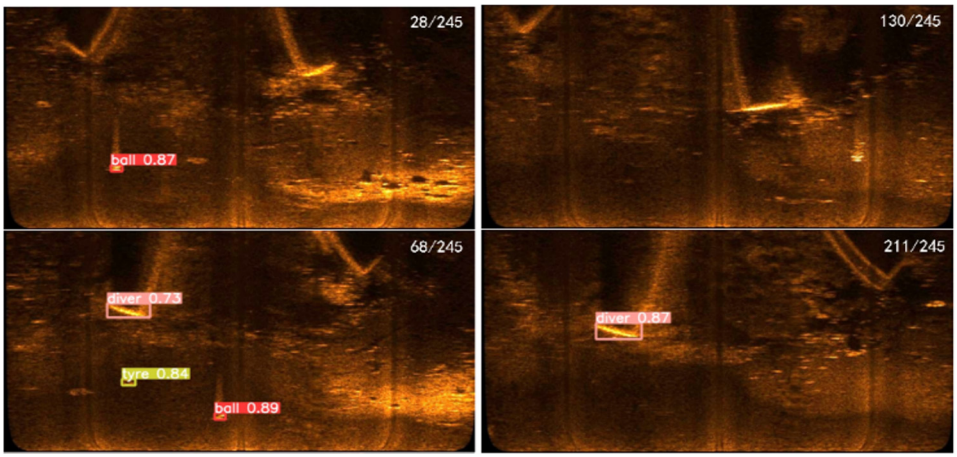



**Figure 7.** Results obtained by Y5s.

The Re obtained by the Y5S2 and Y5S3 models are relatively comparable and closely aligned with those of the Y51and Y5x models. In contrast, the Re value of the Y5S1 architecture significantly decreases to 89.4%, which is lower than that of the original Y5s model. As illustrated in Figure 6, for some targets, such as the ball, the diver, and the single column, both Y5S1 and Y5S2 show a lower detection accuracy in comparison to Y5s. Conversely, Y5S3 shows a marked enhancement in detection precision. It achieves an impressive 100% detection accuracy for the targets (divers, dummies, and tires) with no instances of misdetection recorded. This performance consistently exceeds the results

obtained by the standard Y5s model. Figures 7 and 8 further demonstrate that the conventional Y5s model experiences both missed and incorrect detections in certain image frames, while the Y5S3 network maintains a high level of detection accuracy.

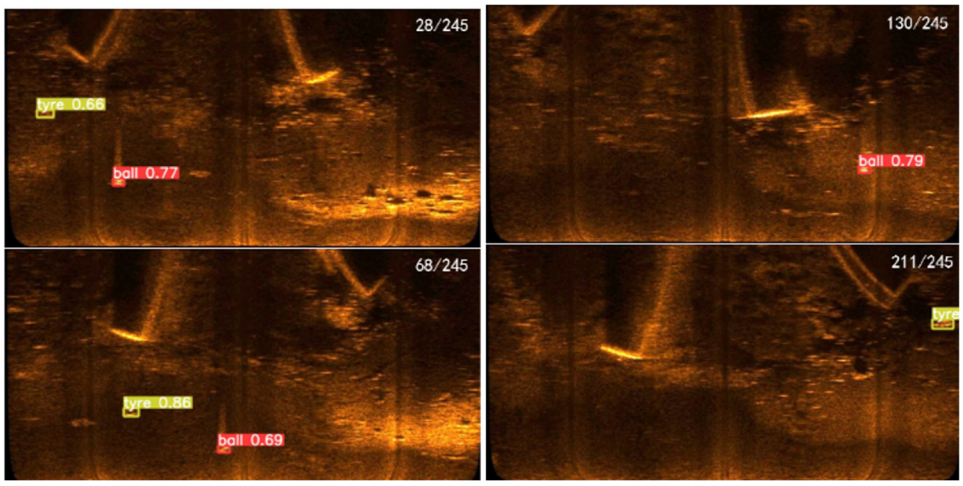
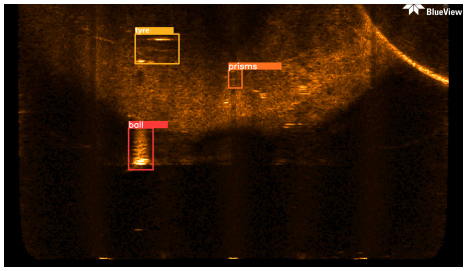
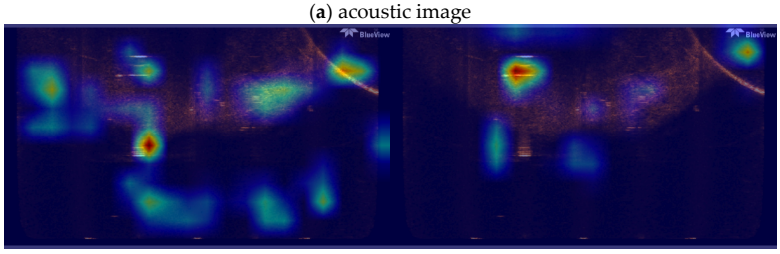



**Figure 8.** Results obtained by Y5S3.

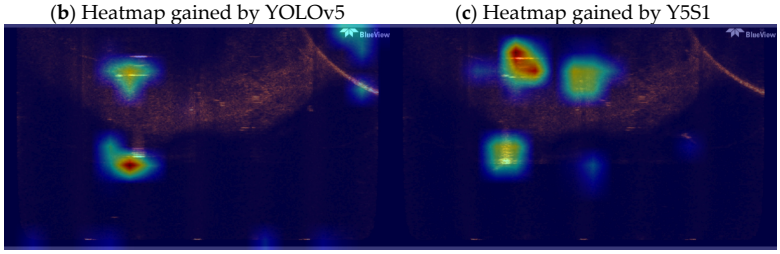
In order to further analyze the differences in the acoustic image feature information extraction capabilities among Y5S1, Y5S2, and Y5S3, the network heatmaps were calculated, as shown in Figure 9. It can be seen from Figure 9 that all three replacement plans can, to some extent, increase the network's attention to the target. However, both Y5S1 and Y5S2 have the problem of target omission, while Y5S3 shows the most attention to the target in the acoustic image and the smallest focus offset, showing that this structure effectively improves the network's ability to extract target feature information.




(**a**) acoustic image



(**b**) Heatmap gained by YOLOv5 (**c**) Heatmap gained by Y5S1



(**d**) Heatmap gained by Y5S2 (**e**) Heatmap gained by Y5S3

**Figure 9.** Heatmaps.

In summary, the results indicate that the global feature information from the STr block structure plays a crucial role in enhancing the performance of the feature extraction component within the detection network. Consequently, the Y5S3 network exhibits notable advancements. These improvements effectively reduce both misclassifications and missed detections in detection networks, while simultaneously maintaining the essential consistency of the network's Re and inference speed.

## 5. Tracker Design Based on DEEPSORT

### *5.1. DeepSORT Basic Principle*

The Simple Online and Real-time Tracking (SORT) algorithm employs a basic KF to address the correlation of the data between consecutive frames [48] and employs the Hungarian algorithm to measure correlation. This method has demonstrated impressive performance at high frame rates. In 2017, DeepSORT was developed [49], which enhanced the SORT framework by incorporating a deep neural network-based object detector to improve detection accuracy.

The Mahalanobis distance is defined as follows:

$$d^{(1)}(i,j) = \left(d_j - y_{\mathrm{i}}\right)^{\mathrm{T}} {S_i}^{-1} \left(d_j - y_i\right), \tag{1}$$

where $(y_i, S_i)$ represents the projection of the $i-$th track distribution into measurement space. $d_j$ represents the $j-$th bounding box detection.

In accordance with Formula (1), a binary variable ${b_{i,j}}^{(1)}$ is established to signify whether an association is permissible, and it is defined as

$${b_{i,j}}^{(1)} = \mathbb{1}\left[d^{(1)}(i,j) \le t^{(1)}\right], \tag{2}$$

where $t^{(1)}$ represents the corresponding Mahalanobis threshold value.

It is proposed that the appearance feature vector associated with the $j$th detection is denoted as $r_j$, while the $k$th appearance feature vector of the $i$th track is represented as ${r_k}^{(i)}$. As a result, the minimum cosine distance $d^{(2)}(i,j)$ between the $i$th track and the $j$th detection is defined as follows:

$$d^{(2)}(i,j) = min\left\{1 - {r_j}^{\mathrm{T}} {r_k}^{(i)} \middle| {r_k}^{(i)} \in R_i\right\}, \tag{3}$$

where $R_i = \left\{{r_k}^{(i)}\right\}_{k=1}^{100}$.

In accordance with Formula (3), the binary variable ${b_{i,j}}^{(2)}$ is defined as

$${b_{i,j}}^{(2)} = \mathbb{1}\left[d^{(2)}(i,j) \le t^{(2)}\right], \tag{4}$$

where $t^{(2)}$ represents the corresponding cosine distance threshold.

Formulas (1) and (3) provide complementary approaches to tackle different aspects of the assignment problem. As a result, the cost matrix $c_{i,j}$ is formulated as

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j), \tag{5}$$

This approach incorporates the appearance information of the target, leveraging the robust feature extraction capabilities of deep learning, which leads to improved performance in person re-identification (ReID). By incorporating appearance data, the tracking efficacy of the algorithm is greatly enhanced, particularly in scenarios involving occlusion and interactions, thereby mitigating issues related to identity confusion and trajectory interruption.

### 5.2. Improvement in DeepSORT

Given the distinctive attributes of acoustic detection sensors and the complexities inherent in their operational mechanisms, sonar images are prone to considerable interference. Traditional methodologies typically employ denoised sonar imagery for tracking processes; however, it inadvertently removes target scatter noise, which is essential for conveying critical information about the shape and material properties of the target, ultimately leading to a reduction in the discernible features of the target region. To address this issue, an improved DeepSORT algorithm, referred to as ExDeepSORT, is proposed, as illustrated in Figure 10. This algorithm increases the dimensions of the output target bounding boxes produced by the detector by a predetermined ratio, thereby increasing the network's ability to identify target feature interference. This improvement not only enriches the primary features of the target but also enhances the overall stability of the tracking network.

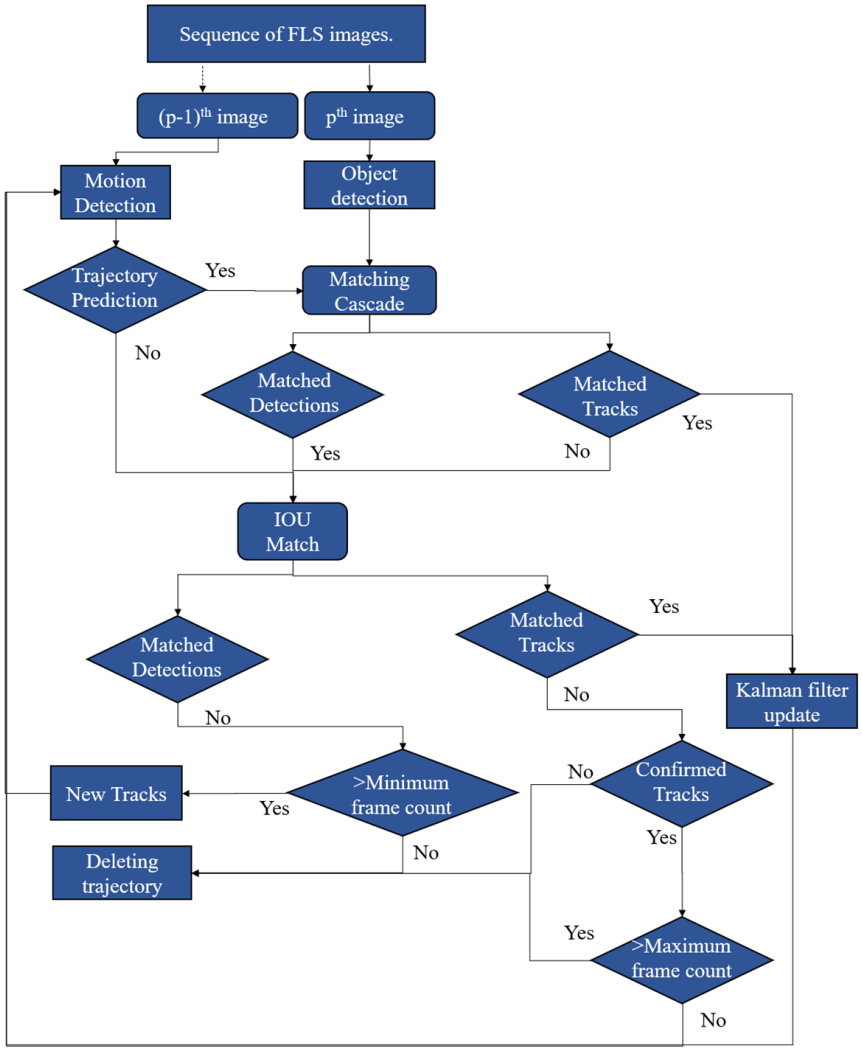



**Figure 10.** DeepSORT tracing framework.

The algorithmic framework is structured into three primary phases: the collection of evaluation metrics for the target box within the current frame, the gathering of prediction metrics for the target box in the same frame, and the execution of target matching. Initially, the sonar image is processed by the detector to facilitate target identification, which subsequently determines the coordinates of the center point ($x_c$, $y_c$) and the dimensions (width $w$, height $h$) of the target box. In the subsequent step, the identified target box is

proportionally enlarged based on the specified expansion rate $R$, resulting in the extended target box $(x_{exp}, y_{exp}, w_{exp}, h_{exp})$, as described by the following equation:

$$\begin{cases} x_{exp} = x_c \\ y_{exp} = y_c \\ w_{exp} = w \times R \\ h_{exp} = h \times R \end{cases}, \tag{6}$$

Subsequently, a predictive bounding box is established, and KF is employed to anticipate motion based on the information derived from the preceding frame. The parameters of the KF state are denoted as

$$X = \left[x_c, y_c, w, h, \dot{x_c}, \dot{y_c}, \dot{w}, \dot{h}\right]^T, \tag{7}$$

After that, the expanded predictive bounding box is further expanded, as denoted in Equation (6), and the target is aligned in accordance with the methodologies outlined in Equations (1)–(5), employing both the expanded target box and the anticipated target box, as illustrated in Figure 11.

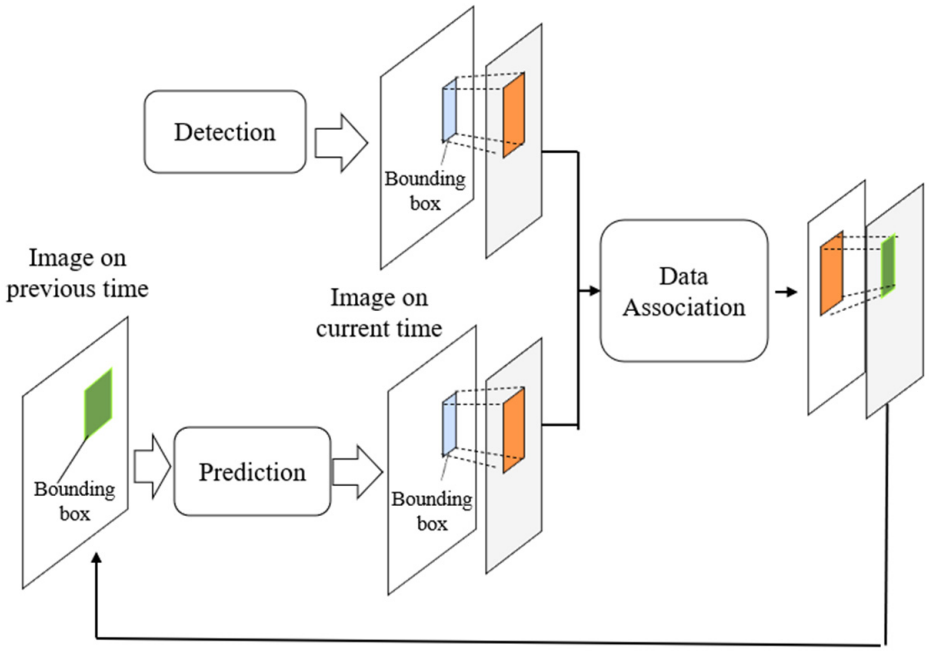



**Figure 11.** The extended target frame of the DeepSORT structure.

As shown in Figure 12, the expansion of detection boxes enhances the network's ability to recognize feature interference around the target, consequently augmenting the feature information of the target relative to the DeepSORT tracking framework.

To determine the best expansion ratio (R) and evaluate the impact of the target box expansion technique on the efficacy of the feature extraction network, the MobileNet and ResNet50 architectures were used as classification networks for the experiments. Figure 13 illustrates the trend in classification accuracy during the training process of the network.

In Figure 13, "acc" refers to classification accuracy. It indicates that an increase in the size of the target box correlates with an enhancement in the network's classification accuracy, leading to greater convergence stability. When the R value is set to 3, the network achieves its peak classification accuracy; thus, this research designates the expansion ratio as 3.

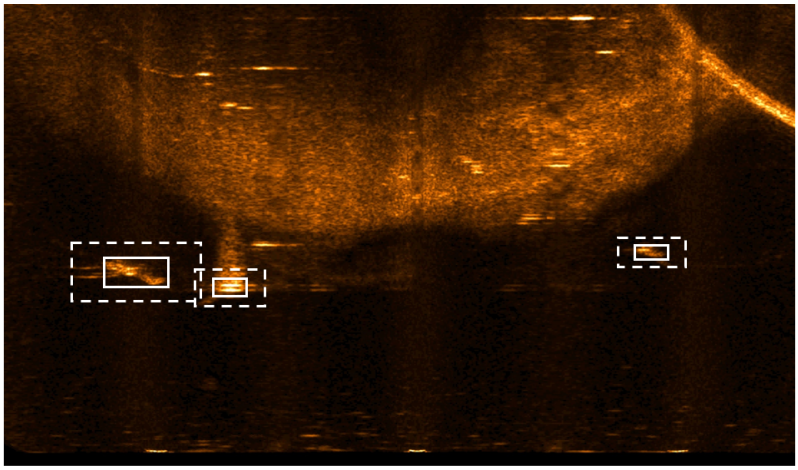
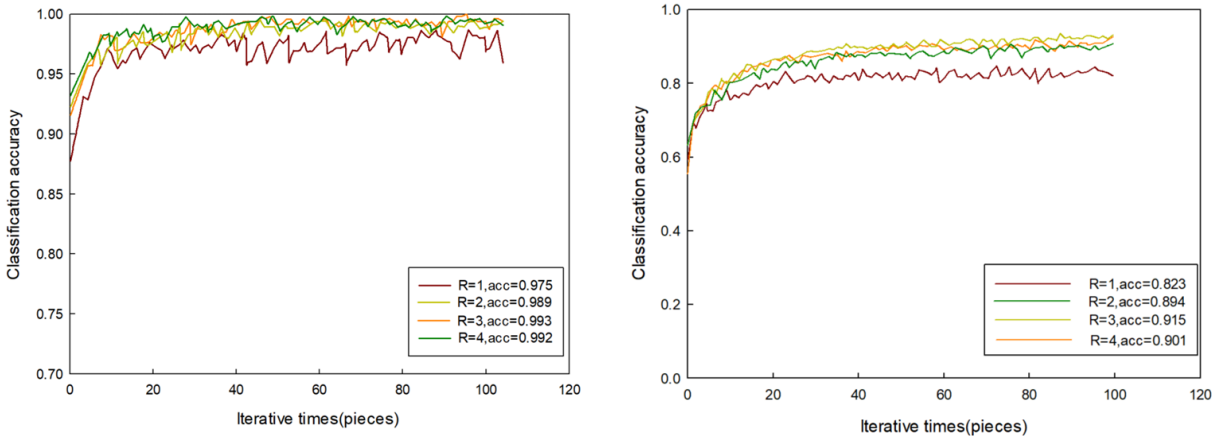


**Figure 12.** Original target box (solid line) and expanded target box (dashed line) in sonar imagery.




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

**Figure 13.** Comparison of classification accuracy. (**a**) Results of the ResNet50 network; (**b**) results of the MobileNet network.

## 6. Experimental Test

### *6.1. Constructing the Experimental Datasets*

Table 3 provides an overview of the objective tracking dataset mentioned in this article. The dataset encompasses images derived from tank experiments, open-lake field trials, and generative augmentation. The tank experimental data comprise 16 sonar image sequences of single target motion, amounting to a total of 1710 images. Additionally, the augmentation image data consist of 20 image sequences of single target motion, totaling 4000 images, as well as 10 image sequences of two-target motion, which collectively yield 2000 images.

**Table 3.** Target tracking dataset.

| Data Origination | Trajectory Classification | Number of Sequences | Total Frames |
|---|---|---|---|
| Data Expansion | One Object | 20 | 4000 |
| | Two Objects | 10 | 2000 |
| Tank Experiment | One Object | 16 | 1710 |
| Lake Experiment | Two Objects | 1 | 245 |

6.1.1. Acquisition of Acoustic Images in Tank

Acoustic imaging data of five typical targets were acquired in the tank of the underwater robotics key laboratory at Harbin Engineering University. Five typical types of targets, including a sphere, diver, dummy model, single cylinder, and tire, were utilized in the experimental process. Figure 14 presents the environmental configuration of the data acquisition test, and Figure 15 displays the acoustic imaging examples for each of the five categories of targets, sequenced beginning with the sphere, then the diver, dummy model, single cylinder, and tire.

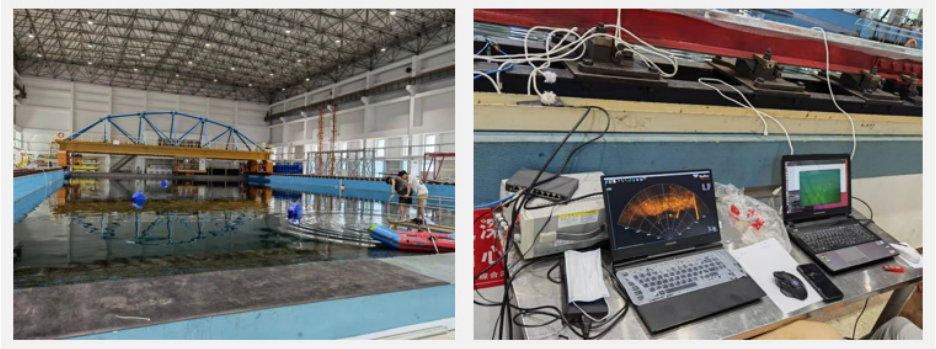


**Figure 14.** Data collection environment.



**Figure 15.** Five types of targets and their sonar images.

In the test, a multibeam sonar was strategically positioned around the side of the tank, operating at a 1 m water depth and 10-degree angle. The target was maneuvered underwater by traction provided by a surface vessel. Four different trajectory types were created for the different targets, as shown in Figure 16. For the diver, several movement patterns were utilized, such as stationary rotation, lateral oscillatory swimming, longitudinal oscillatory swimming, large circular swimming, and swimming close to the tank's bottom. These diverse movements were intentionally designed to meet the variability requirements of the dataset and to improve the generalizability of the experimental results.

The sonar image video sequence collected from this experiment lasted about six hours. The dataset comprised 415 images of spheres, 772 images of divers, 443 images of dummies, 304 images of single columns, and 457 images of tires, culminating in a total of 2391 sonar images.

6.1.2. Acquisition of Acoustic Images in Lake

To further obtain underwater acoustic image data, an outdoor experiment was carried out in a lake. In the experiment, a sphere and a tire were tethered to buoys using ropes, with the two spheres positioned approximately 10 m apart on the water's surface, suspended about 1 m below the surface. A multibeam FLS was mounted on the side of the ship, and its tilt angle was adjusted to 10°. The ship then moved through the area where the targets

were situated at a moderate speed, capturing the acoustic imaging data of the targets. The experimental situation is shown in Figure 17.

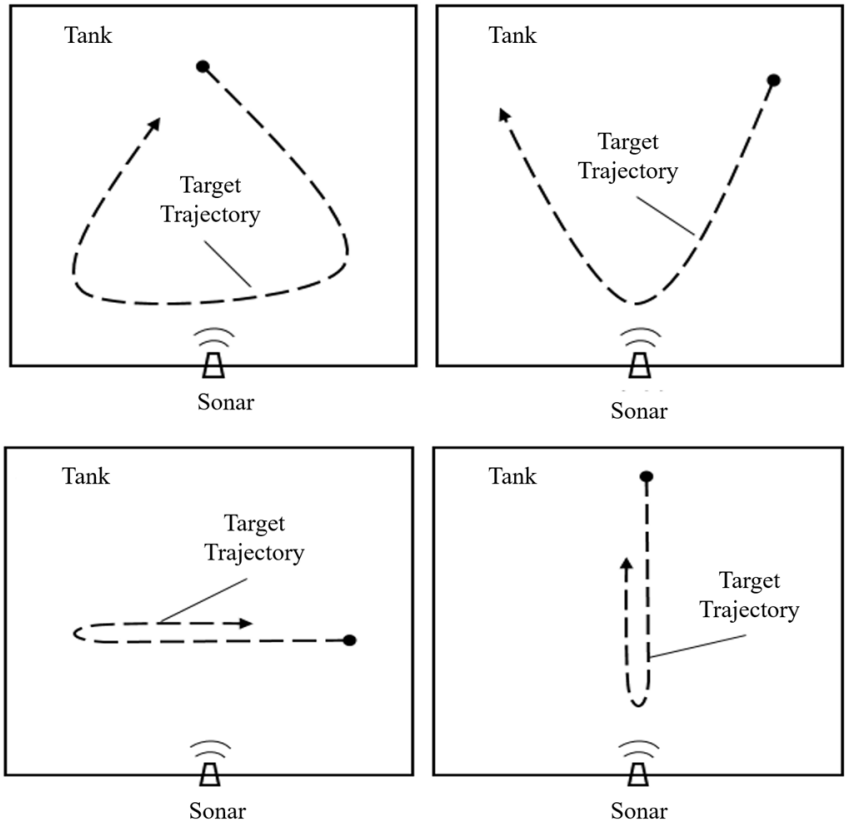
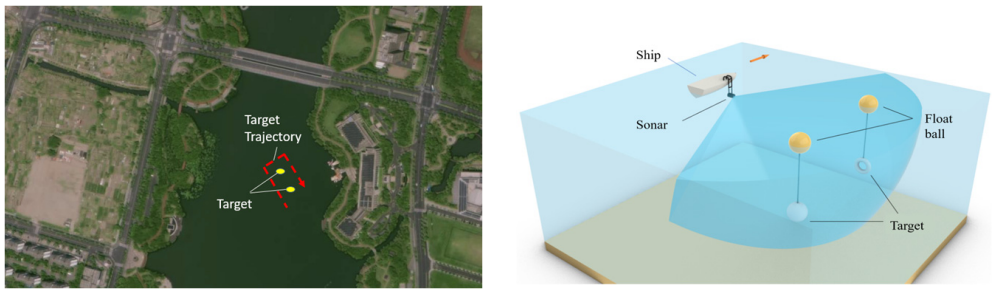



**Figure 16.** Underwater object trajectory.




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

**Figure 17.** Layout of test waters and object placement. (**a**) Movement trajectory; (**b**) object placement.

Figure 18 presents a series of sonar images captured from the lake. In contrast to those obtained in the tank, there is much interference in them. Certain targets (such as tires) demonstrate echo intensities that are comparable to the surrounding noise, rendering them susceptible to being obscured by environmental disturbances. As shown in Figure 18d, the tire is entirely masked by background noise. Furthermore, numerous unidentified targets are present in the water, including fish and bottom protrusions, as illustrated in Figure 18a–c, respectively. These unidentified targets are characterized by high-intensity bright spots, and their shapes and sizes resemble those of identifiable targets, which may lead to misidentification by the detection network.

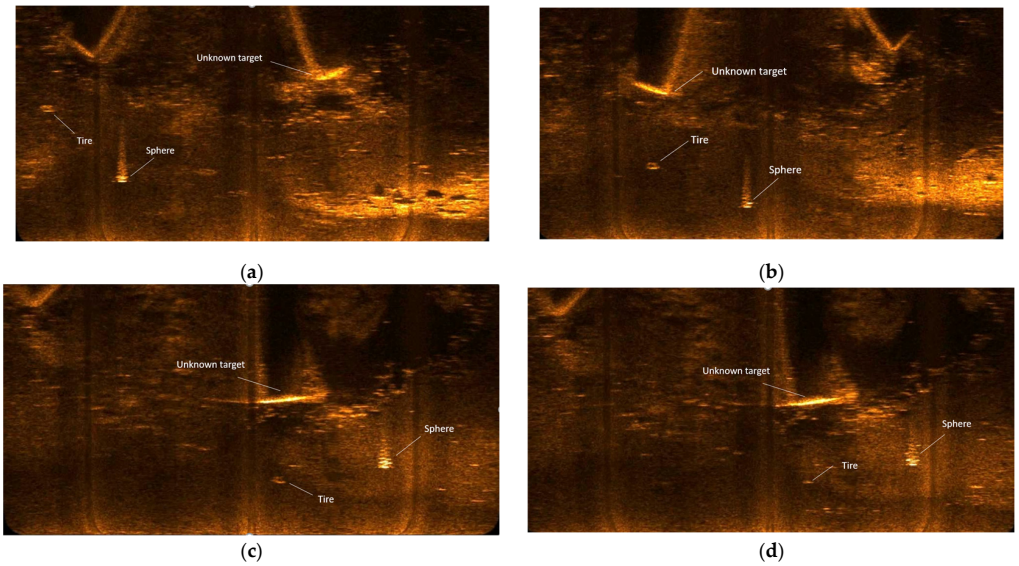



**Figure 18.** Sonar image sequence of targets in lake. (**a**) The 30th image; (**b**) the 90th image; (**c**) the 130th image; (**d**) the 170th image.

#### 6.1.3. Acoustic Image Dataset Extension Based on Pix2PixHD

The acquisition of moving target acoustic imaging data frequently presents significant challenges, particularly in the presence of interferences such as occlusion and target disappearance during motion. To address this issue, the Pix2PixHD model was employed for the transfer of image styles, facilitating the generation of a sequence of sonar image data for evaluation across various interference backgrounds. In accordance with the training requirements of the Pix2PixHD network, it is essential that images from dataset A (real sonar imagery) and dataset B (simulated sonar imagery) are appropriately paired. In addition to the necessity of varying image styles, it is essential to closely preserve the target's position, size, and orientation within the images. To achieve this, acoustic simulation software was employed to produce simulated sonar images of the underwater targets. To maintain consistency with the parameters of the experimental equipment and environment, the target's location was predetermined, enabling the generation of simulated sonar images that correspond to those in dataset A. The parameter settings for the simulation software are outlined in Table 4, while the simulation results for the spherical target are illustrated in Figure 19.

**Table 4.** Parameter settings in simulation software.

| Parameter | Unit | Value |
|---|---|---|
| Distance between sonar and the seabed | m | 9 |
| Inclination angle of sonar | ° | 10 |
| Field of view | ° | 130 |
| Range resolution | m | 0.02 |
| Transmission frequency | kHz | 900 |

Figure 20 provides a comparative analysis of the differences among actual, simulated, and generated acoustic images related to different targets. The results indicate that according to subjective assessment, the generated acoustic images exhibit a notable resemblance to the real acoustic image.

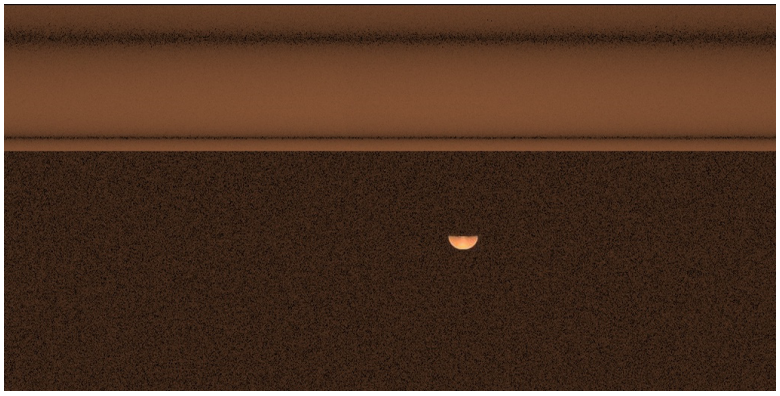


**Figure 19.** Sonar-simulated image (sphere).

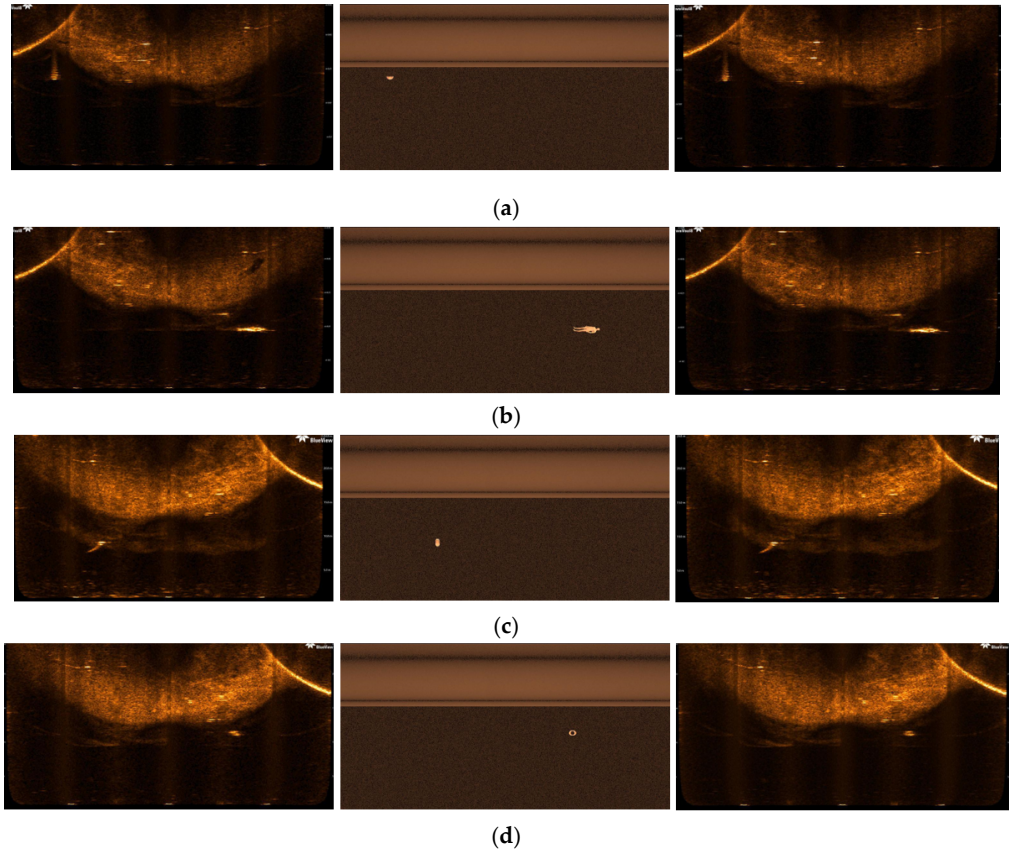
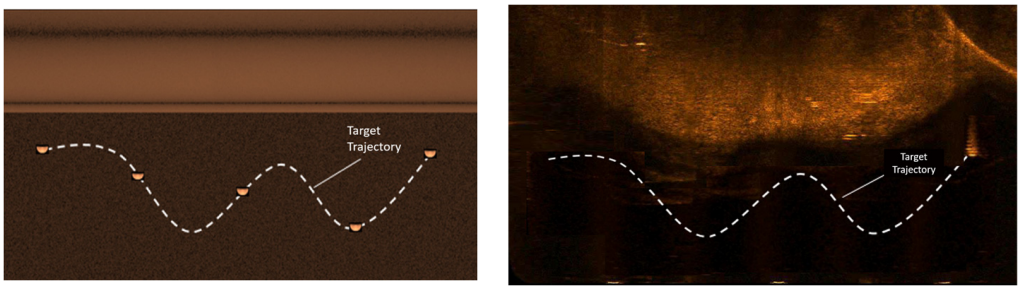


**Figure 20.** Comparison results of different targets in genuine acoustic images (**left**), simulated images (**middle**), and generated images (**right**). (**a**) Sphere target; (**b**) dummy model; (**c**) cylinder target; (**d**) tire target.

This paper focuses on improving raw forward sonar image data, collected through experimental methods, into datasets designed for target analysis, specifically in the realms of detection and tracking. The detection dataset was employed to train detection algorithms, whereas the tracking dataset served to assess the accuracy of the tracking network. Consequently, it was feasible to produce dynamic sequences of sonar images depicting moving targets that followed various trajectories by modifying the position and dimensions of the target within the simulated images, as illustrated in Figure 21.

In accordance with the method previously discussed, this paper utilizes generative augmentation to create an improved target tracking dataset comprising five distinct targets, each adhering to one of two specified motion trajectories, as illustrated in Figure 22. These motion trajectories consist of linear and curvilinear movements, with the curvilinear path

characterized by a cosine function. Each scenario incorporates instances of target occlusion, specifically one or two occurrences, which serve to assess the network's tracking efficacy in the presence of occlusion. Partial images are presented in Figures 23 and 24.




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

**Figure 21.** Sonar image sequences generated by Pix2PixHD net. (**a**) Movement trajectory of object in the simulated image; (**b**) movement trajectory of object in the generated image.

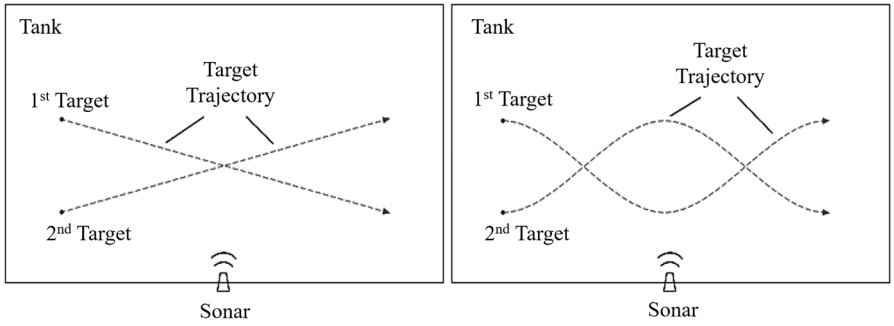



**Figure 22.** Object trajectory based on expanded data.

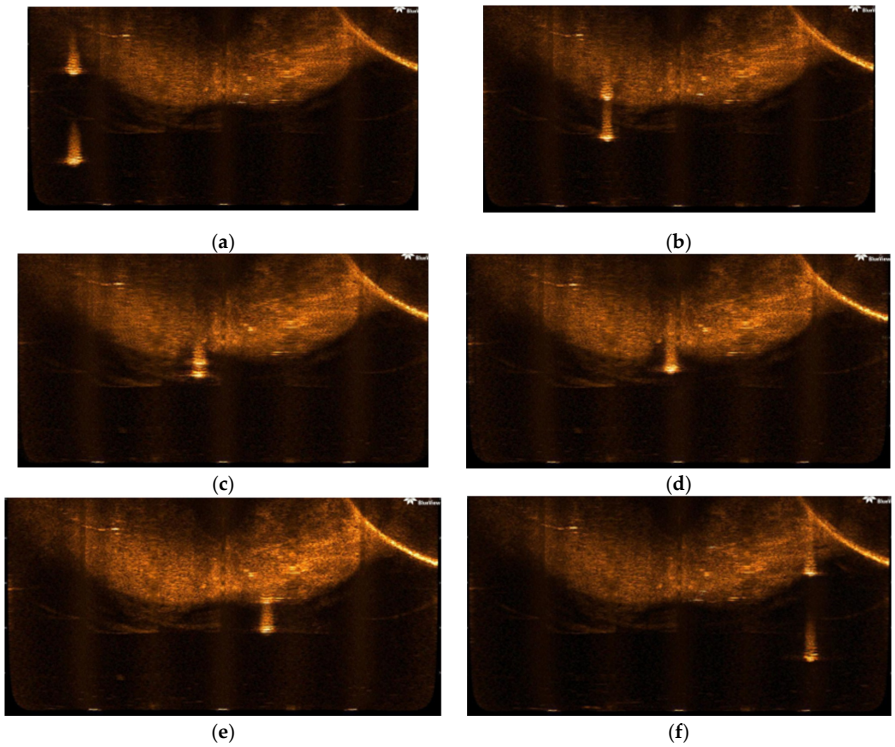



**Figure 23.** Image sequences of intersection trajectory (spherical). (**a**) The 10th image; (**b**) the 60th image; (**c**) the 90th image; (**d**) the 100th image; (**e**) the 130th image; (**f**) the 180th image.

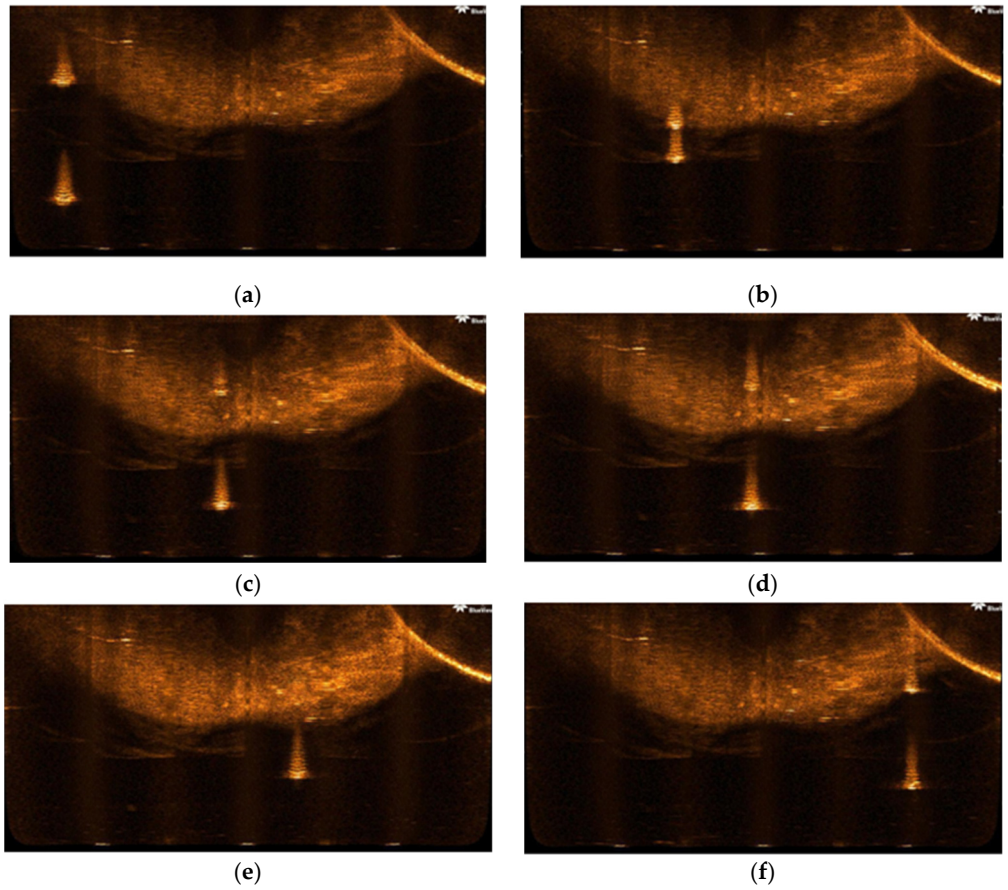


**Figure 24.** Image sequences of the straight line-crossing trajectory (spherical). (**a**) The 10th image; (**b**) the 60th image; (**c**) the 90th image; (**d**) the 100th image; (**e**) the 130th image; (**f**) the 180th image.

#### 6.1.4. Ablation Experiment

To verify that the constructed dataset is practical and effective, YOLOv5 was used as the test network, and two sets of ablation experiments were conducted, respectively. The experimental results are shown in Table 5.

**Table 5.** Ablation experiment results.

| No. | Real Data | Extension Data | mAP (IOU = 0.5) | Re (IOU = 0.5) |
|---|---|---|---|---|
| 1 | ✓ | - | 69.88% | 46.59% |
| 2 | ✓ | ✓ | 87.46% | 52.73% |

As can be seen from Table 5, for the dataset composed of real data, the average precision for underwater target detection was 69.88%, and the recall rate was 46.59%. After successively adding augmented data, the mAP value increased to 87.46%, and the recall rate also reached 52.73%. The results showed that the datasets composed of raw data, traditionally augmented data, and generatively augmented data had a positive impact on the performance of the mAP and Recall metrics.

### *6.2. Underwater Target Tracking Evaluation Criteria*

The effectiveness of underwater acoustic target tracking is evaluated by two main performance indexes: Frag Ratio and ID Switch. As shown in Figure 25, it is described in detail as follows:

(1) ID Switch: An ID switch is defined as an alteration in the identification number of a target along a singular trajectory line.

(2) Frag Ratio: A trajectory interruption is identified as the absence of an assigned ID for the target within a single trajectory line. The frequency of frames exhibiting trajectory interruptions is quantified as a ratio of the total frames within the trajectory, known as the trajectory interruption proportion, as detailed in the following sections.

$$Frag\ ratio = \frac{F_{break}}{F_{total}} \tag{8}$$

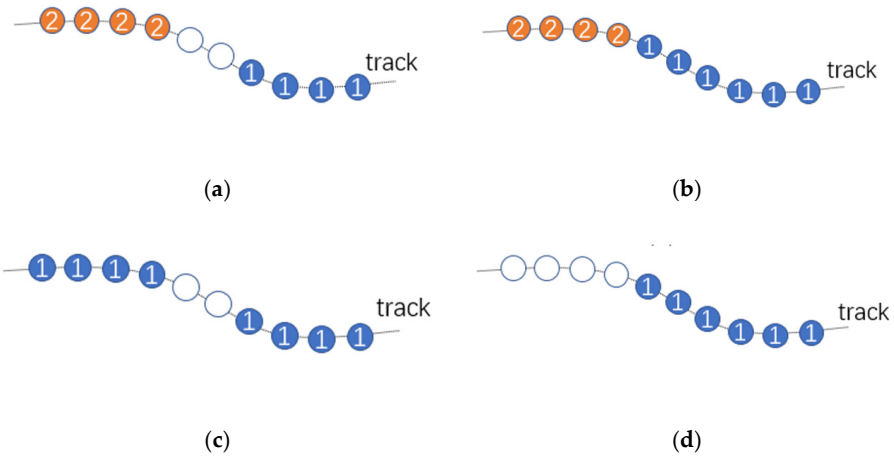



**Figure 25.** An overview of indexes used to assess target tracking performance. (**a**) ID Switch in the case of trajectory interruption (ID Switch = 1, Frag Ratio = 0.2); (**b**) ID Switch in the case of trajectory interruption (ID Switch = 1, Frag Ratio = 0.0); (**c**) trajectory interruption without ID Switch (Frag Ratio = 0.2); (**d**) trajectory interruption without ID Switch (Frag Ratio = 0.4).

### *6.3. Underwater Target Tracking Experiment Based on Extension Datasets*

To verify the effectiveness of the improved method, a testing environment was set up based on PyTorch (1.10.0). The constructed image dataset was used for training and testing. The dataset was divided into three categories, a training set, validation set, and testing set, in the ratio of 7:2:1. In the experiment, the number of training epochs was set to 200, and mosaic data augmentation was disabled during the last 10 epochs. The batch size was set to 16, and the input image size was set to 1024 × 512. The SGD optimizer was used for training. The computer configuration and environmental parameters used are shown in Table 6.

**Table 6.** The configuration of the training environment.

| | | |
|---|---|---|
| Hardware | GPU | NVIDIA GeForce RTX2080 Ti (*Harbin Ship Electronics Market., Harbin, China*) |
| | CPU | Intel(R) Core(TM) i9-9900K (*Harbin Ship Electronics Market., Harbin, China*) |
| Parameters | Operation System | Windows 11 |
| | CUDA | 12.2.79 |
| | CUDNN | 7.6.5 |
| | Python | 3.8.5 |
| | Pytorch | 1.10.0 |

#### 6.3.1. Single-Target Tracking Experiment

In the single-object tracking experiment, five types of targets were utilized, each accompanied by four sets of image sequences. According to the results presented in

Table 7, the proposed method achieved a trajectory interruption rate of 9.09%, which is a 13.60% improvement over the SORT method and a 0.41% improvement over the traditional DeepSORT method, suggesting that the performance of the proposed method is comparable to that of traditional DeepSORT in the context of single-object tracking. Regarding ID change metrics, the proposed method recorded 12 ID changes, which signifies a reduction of 71 compared to the SORT method and a decrease of 19 in relation to the traditional DeepSORT method, thereby underscoring a notable optimization effect. The results also show that the proposed method is slower than other methods. Considering the navigation tasks in AUV platforms, it is still a suitable approach. Consequently, it is evident that the proposed method demonstrates robust performance in single-object tracking, as evidenced by testing with the expanded dataset.

**Table 7.** Comparison results of single-object tracking.

| Tracking Framework | Frag Ratio | ID Switch | Fps |
|---|---|---|---|
| SORT | 22.69% | 83 | 13.68 |
| DeepSORT | 9.50% | 31 | 7.35 |
| ExDeepSORT | 9.09% | 12 | 6.98 |

From Figure 26a, it can be seen that the motion path of the dummy model is modeled using a cosine function. In Figure 26b–d, it is shown that underwater reverberation noise causes disruptions in the tracking paths generated by the three methods. Among them, the SORT method exhibits the most severe interruptions in the motion trajectory points, and there are numerous instances of ID changes occurring during the tracking process. In contrast, the DeepSORT method successfully maintains only three valid IDs throughout the tracking process, and the issues of trajectory interruption and ID changes are significantly optimized. The approach presented in this paper results in a more stable and continuous tracking trajectory, further minimizing ID changes in comparison to the traditional DeepSORT method.

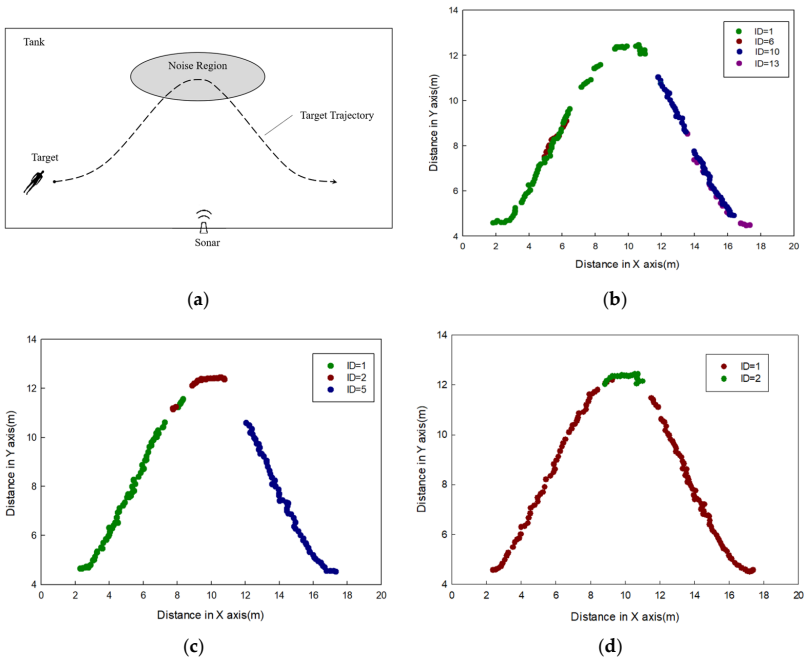



**Figure 26.** Tracking results of the dummy model. (**a**) The target motion trajectory; (**b**) results obtained by the SORT method; (**c**) results obtained by the DeepSORT method; (**d**) results obtained by the ExDeepSORT method.

From Table 8, it can be seen that compared to other targets, the single cylinder and tire exhibit a higher susceptibility to trajectory disruptions and ID changes. This phenomenon can be attributed to the diminished imaging intensity associated with single cylinders and tires, which yields lower grayscale values in the images and less pronounced features. As a result, these targets are more likely to be misidentified due to background interference, leading to missed detections and interruptions in the tracking process. Furthermore, as indicated in Table 9, the proposed methodology demonstrates a reduced frequency of ID changes while tracking various targets, thereby achieving enhanced tracking stability.

**Table 8.** Comparison single-object tracking results of Frag Ratio.

| Target | SORT | DeepSORT | ExDeepSORT |
|---|---|---|---|
| Sphere | 15.8% | 8.9% | 8.1% |
| Diver | 9.5% | 5.3% | 4.9% |
| Prop person | 9.5% | 3.3% | 3.4% |
| Single Cylinder | 43.3% | 12.4% | 11.8% |
| Tire | 22.3% | 13.5% | 13.1% |

**Table 9.** Comparison single-object tracking results of ID Switch.

| Target | SORT | DeepSORT | ExDeepSORT |
|---|---|---|---|
| Sphere | 5 | 1 | 1 |
| Diver | 3 | 1 | 0 |
| Prop person | 6 | 4 | 2 |
| Single Cylinder | 36 | 21 | 8 |
| Tire | 36 | 9 | 1 |

In Figure 27, it can be seen that during the tracking process of the dummy model, only two valid target IDs appeared by the ExDeepSORT method, which indicates a relatively stable tracking process. Before the target entered the reverberation noise zone, the stability of the target tracking was commendable, with the ID remaining consistent. However, once the target entered the reverberation area, there were significant changes in the target's image features, leading to some interruptions in the tracking process, and the ID also changed. Once the target left the reverberation zone, due to the setting of the maximum survival frame count parameter in the tracking framework, the ID returned to its previous value and remained consistent in the subsequent tracking process.

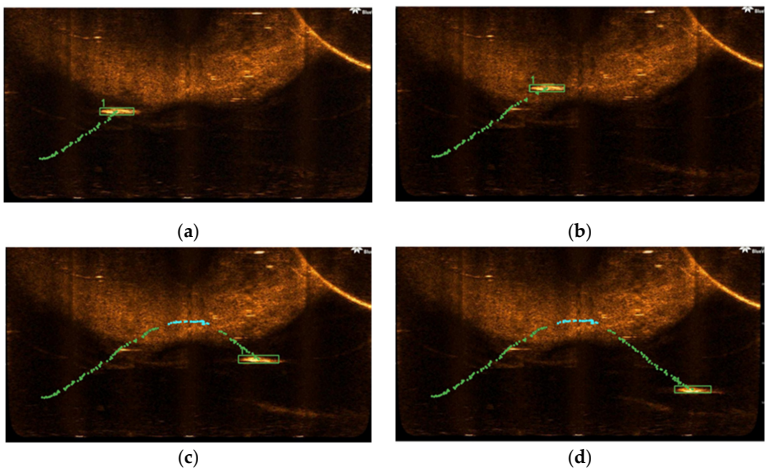



**Figure 27.** Tracking results of a dummy model by ExDeepSORT. (**a**) The 60th image; (**b**) the 90th image; (**c**) the 150th image; (**d**) the 180th image.

6.3.2. Double-Target Tracking Experiment

In the double-object tracking experiment, five types of targets were utilized, each associated with two sets of image sequences. The motion paths of the targets were represented as straight-line and curved trajectories described by different colored lines, with at least one instance of target occlusion occurring during the motion. As shown in Table 10, the proposed method achieved a trajectory interruption rate of 14.97%, reflecting a 15.47% improvement over the SORT method and a 1.44% improvement over the traditional DeepSORT method. Regarding the ID change index, the proposed method recorded 28 ID changes, which is 81 fewer than the SORT method and 31 fewer than the traditional DeepSORT method. Since target occlusion is the primary factor contributing to ID changes in dual-target motion, the test results suggest that the target box expansion strategy allows the tracking network to effectively recover the ID after an occlusion occurs.

**Table 10.** Comparison results of objects tracking.

| Tracking Framework | Frag Ratio | ID Switch |
|---|---|---|
| SORT | 30.44% | 109 |
| DeepSORT | 16.41% | 59 |
| ExDeepSORT | 14.97% | 28 |

In Figure 28a, the trajectory of the dummy model is modeled through a double cosine function. Figure 28b indicates that the target's ID is subject to change due to occlusion in the tracking outcomes produced by the SORT method. In Figure 28c, it is shown that in the traditional DeepSORT method, target 1 successfully regained its ID after the initial occlusion, but ID changes were observed in other occlusion scenarios. Throughout the tracking process without occlusion, it demonstrates superior ID retention compared to the SORT method. The results in Figure 28d indicate that in the tracking results of the proposed method, both target 1 and target 2 effectively maintain their IDs after experiencing two occlusions, with no changes in ID recorded. The entire tracking process is characterized by consistency and stability, demonstrating excellent ID retention. While tracking target 2, all three methods encountered interruptions due to the area being classified as a submerged reverberation zone, where reverberation noise compromised the clarity of target imaging information, thereby complicating identification and causing interruptions in the trajectory.

From Tables 11 and 12, it can be seen that all indexes increased compared to the results obtained from single-target tracking. This indicates that when target occlusion occurs during motion, trajectory interruptions and ID changes become more frequent. However, the proposed method has demonstrated an improvement in the stability of the tracking process for different types of targets. Specifically, in the tracking process of the dummy model, the DeepSORT method shows a high number of ID switches compared to the SORT method (as shown in Figure 29c). This is because the SORT method experiences a higher frequency of trajectory interruptions than the DeepSORT method, which results in a lower number of ID changes due to the larger number of affected frames. Overall, for double targets, the proposed method not only enables continuous and stable tracking but also effectively recovers IDs following instances of target occlusion, thereby improving the stability of the tracking process.

From Figure 29a, it can be seen that when the trajectories of target 1 and target 2 cross each other, the tracking framework based on ExDeepSORT demonstrates a robust capability for identity recovery. From Figure 29b,c, it can be observed that when target 1 moves into the reverberation zone, changes in the image features of the target region result in a shift in the target ID. Furthermore, the pronounced intensity of reverberation noise diminishes the prominence of target features, leading to multiple disruptions in tracking.

From Figure 29d, it can be seen that when target 1 leaves the reverberation zone, its ID is restored to its previous value, demonstrating effective ID recovery during the subsequent intersection of trajectories.

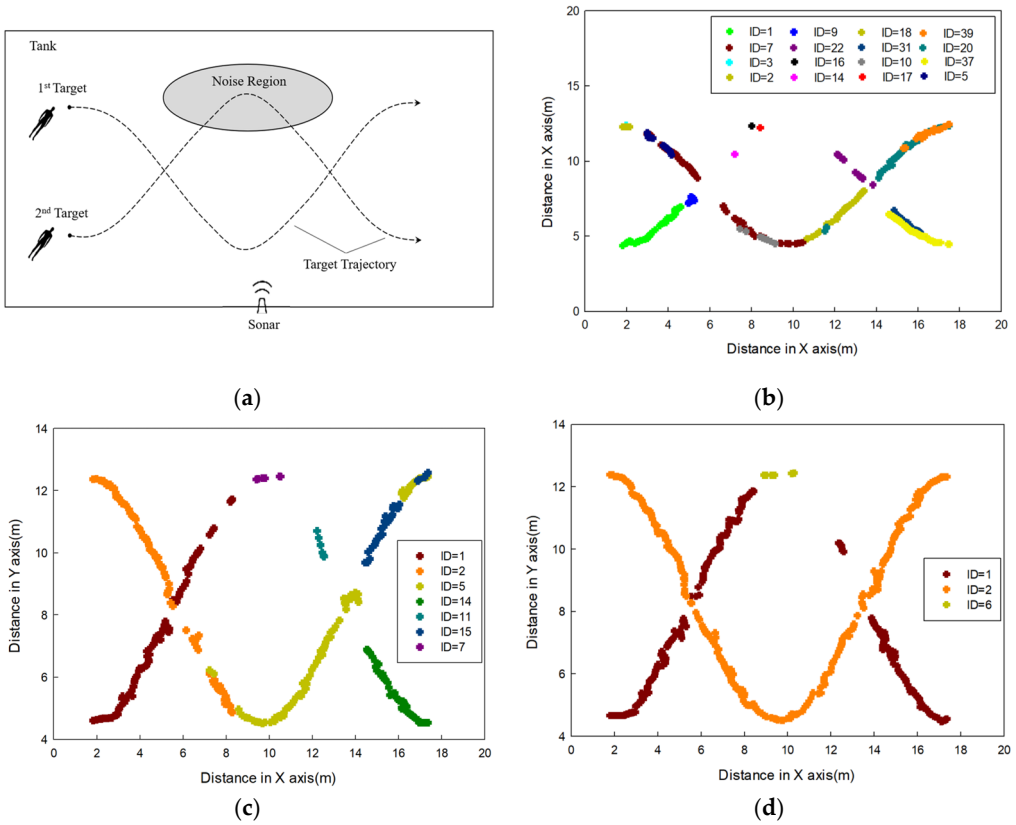



**Figure 28.** Tracking results of the dummy model. (**a**) The target motion trajectory; (**b**) results obtained by the SORT method; (**c**) results obtained by the DeepSORT method; (**d**) results obtained by the ExDeepSORT method.

**Table 11.** Comparison double-object tracking results of Frag Ratio.

| Target | SORT | DeepSORT | ExDeepSORT |
|---|---|---|---|
| Sphere | 36.4% | 22.6% | 19.1% |
| Diver | 21.8% | 13.5% | 11.5% |
| Dummy model | 27.4% | 18.3% | 15.7% |
| Single cylinder | 59.6% | 29.5% | 28.5% |
| Tire | 34.6% | 23.8% | 23.1% |

**Table 12.** Comparison double-object tracking results of ID Switch.

| Target | SORT | DeepSORT | ExDeepSORT |
|---|---|---|---|
| Sphere | 30 | 21 | 7 |
| Diver | 18 | 13 | 4 |
| Dummy model | 37 | 48 | 27 |
| Single cylinder | 68 | 47 | 33 |
| Tire | 77 | 27 | 26 |

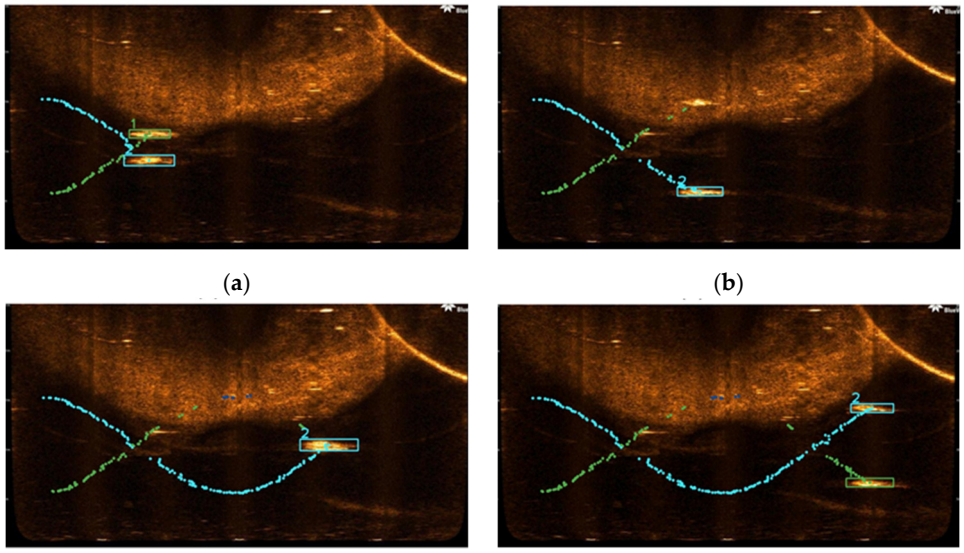


**Figure 29.** Tracking results of the dummy model. (**a**) The 60th image; (**b**) the 90th image; (**c**) the 150th image; (**d**) the 180th image.

#### 6.3.3. Underwater Target Tracking Trial Based on Pond Test Data

The sonar image sequence obtained from the pool experiments consists of 16 distinct segments, which include various objects such as spheres, divers, a dummy model, individual cylinders, and tires, amounting to a total of 1710 images. Given the challenges associated with capturing multitarget motion sonar sequences in controlled laboratory environments, this specific segment of the sonar image sequence exclusively contains sequences characterized by single-target motion.

In Table 13, the proposed methodology demonstrates a trajectory interruption ratio of 10.30%, which represents a substantial reduction of 26.91% in comparison to the SORT method and a decrease of 6.86% relative to the traditional DeepSORT method. This suggests that the proposed approach exhibits superior optimization capabilities, thereby enhancing the continuity of target tracking. Furthermore, the method recorded only 19 ID changes across 1710 sonar sequences, reflecting a remarkable reduction of 50 times in comparison to the SORT method and 14 times relative to the traditional DeepSORT method. This indicates the method's effectiveness in minimizing ID alterations throughout the tracking process. Consequently, the proposed methodology substantially enhances the stability and continuity of underwater acoustic target tracking.

**Table 13.** Tracking comparison results based on tank experiment data.

| Tracking Framework | Frag Ratio | ID Switch |
|---|---|---|
| SORT | 37.21% | 69 |
| DeepSORT | 17.16% | 33 |
| ExDeepSORT | 10.30% | 19 |

Figure 30a illustrates the trajectory of the diver, which is characterized by a reciprocating motion that occurs perpendicular to the sonar beam. Figure 30b–d provide a comparative analysis of tracking efficiency across different frameworks for the diver. In this Figure, the vertical axis denotes the number of frames, reflecting the temporal progression of the trajectory, while the horizontal axis represents the coordinate value of the trajectory point in relation to the image width from the left edge. The trajectory points are colored to correspond with distinct target IDs.

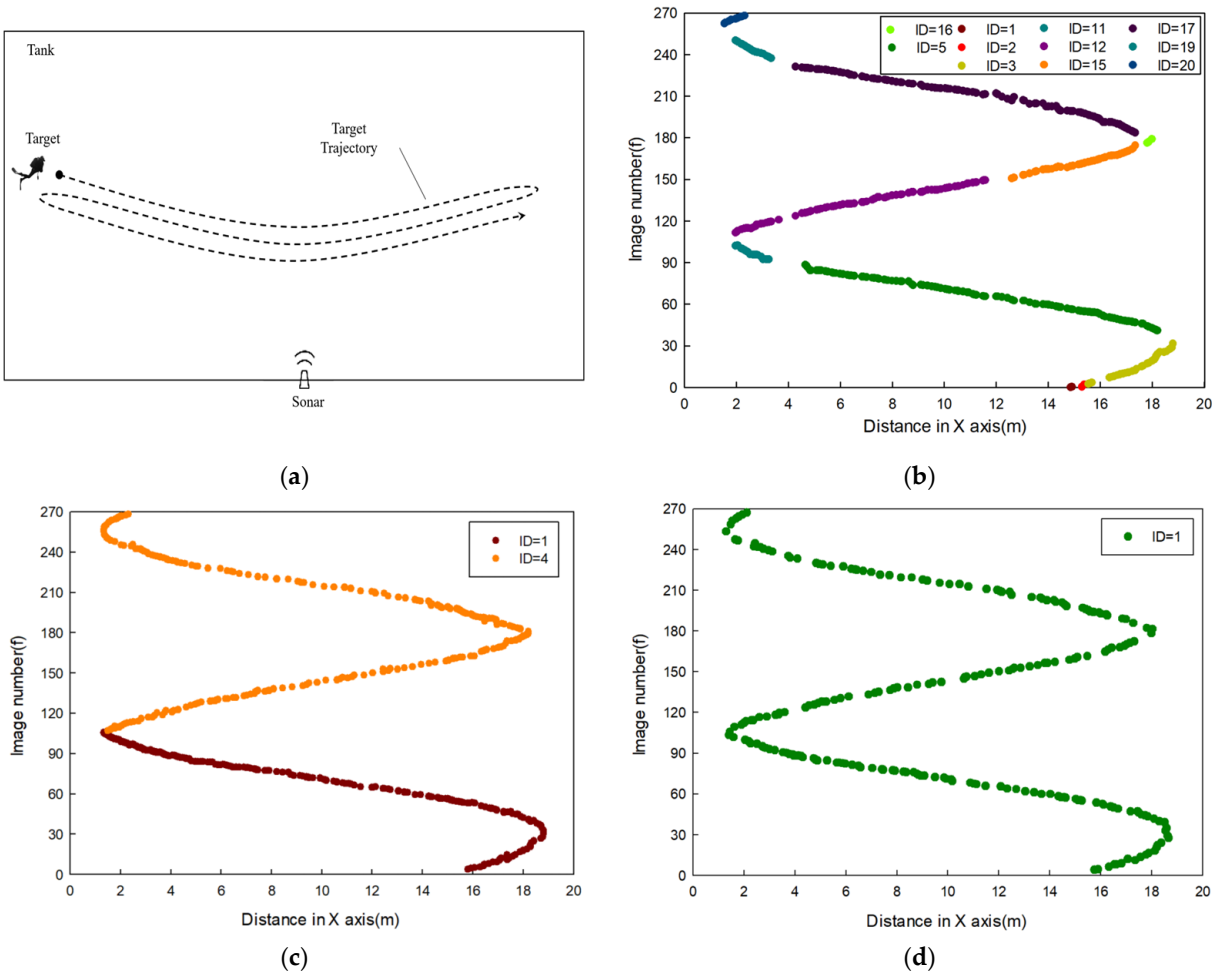



**Figure 30.** Tracking results of the diver. (**a**) The target motion trajectory; (**b**) results obtained by the SORT method; (**c**) results obtained by the DeepSORT method; (**d**) results obtained by the ExDeepSORT method.

In Figure 30b, the SORT method exhibits frequent fluctuations in the identification of tracking targets during its application, which cause interruptions in the tracking of trajectories at various intervals and ultimately result in suboptimal tracking performance. In contrast, as shown in Figure 30c, the traditional DeepSORT technique experiences only a single modification in tracking target identification throughout its operation, indicating a significant enhancement in trajectory continuity compared to the SORT method. Furthermore, as shown in Figure 30d, the proposed methodology maintains consistent tracking target identification throughout its entire process, resulting in a stable trajectory and achieving a reliable and steady tracking of the diver's target.

In Table 14, the proposed methodology demonstrates the ability to maintain relatively low trajectory interruption rates across a range of objectives. Nevertheless, the single cylinder and tire targets exhibit higher rates of trajectory interruption in comparison to the other targets. This phenomenon can be attributed to the diminished imaging intensity associated with single cylinders and tires in aquatic environments, which results in lower gray levels within the images and less pronounced features. Consequently, these targets are more susceptible to confusion with background interference, leading to instances of missed detection. This is the principal factor contributing to the elevated trajectory interruption rate observed for the single cylinder target relative to the other targets.

In Table 15, a significant reduction in the frequency of ID changes across all object categories is shown, with the exception of the spheres and dummy model, during the target

tracking process utilizing the methodology presented in this paper. For the sphere, due to its distinct regional characteristics, the traditional DeepSORT method also maintains a low level of ID changes, consistent with the results of the proposed method. For the dummy model, the influence of its contour shape and material composition means that there are no significant feature noises around the imaging region of the dummy model. Therefore, after the target bounding box is expanded, the main features of the target do not provide new feature information. Instead, background noise is incorporated into the receptive field, which affects target identification during the tracking process. Although the proposed method exhibits a slightly higher ID change rate in the dummy model tracking process compared to the traditional DeepSORT results, it still remains at a relatively low frequency of changes. Thus, the tracking process of the proposed method still demonstrates excellent stability.

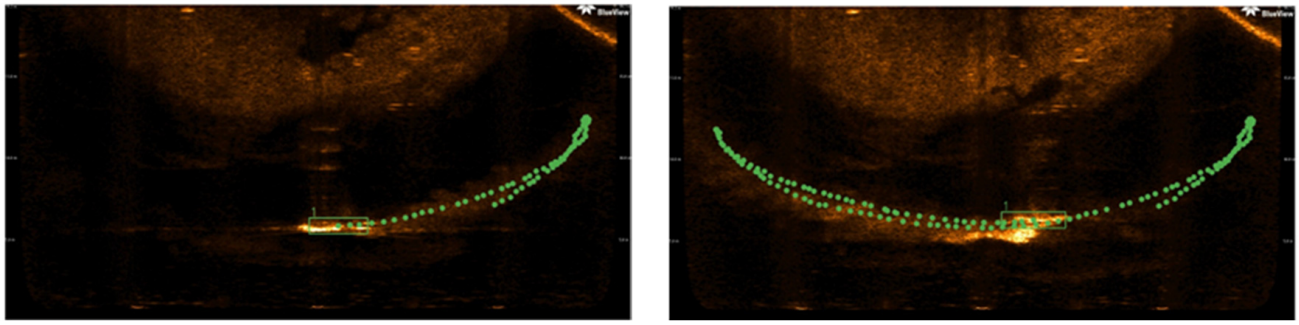
**Table 14.** Comparison results of Frag Ratio based on tank experiment data.

| Target | SORT | DeepSORT | ExDeepSORT |
|---|---|---|---|
| Sphere | 9.2% | 4.6% | 3.9% |
| Diver | 20.8% | 4.0% | 0.4% |
| Dummy model | 23.3% | 7.4% | 6.1% |
| Single cylinder | 61.9% | 36.4% | 20.1% |
| Tire | 32.5% | 16.8% | 12.1% |

**Table 15.** Comparison results of ID Switch based on tank experiment data.

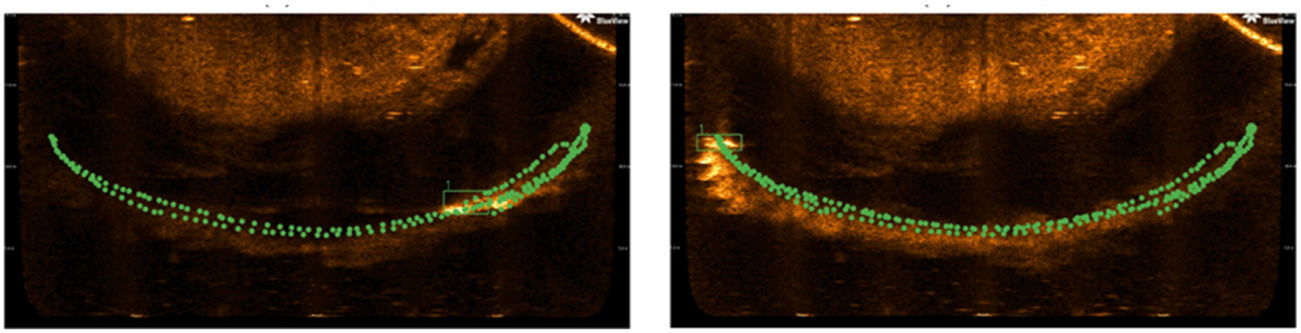
| Target | SORT | DeepSORT | ExDeepSORT |
|---|---|---|---|
| Sphere | 7 | 2 | 2 |
| Diver | 16 | 10 | 0 |
| Dummy model | 19 | 1 | 3 |
| Single cylinder | 16 | 13 | 10 |
| Tire | 13 | 8 | 4 |

In Figure 31, the target ID remains consistent throughout the tracking of a single diver using the method outlined in this paper. Notably, the method successfully maintains ID stability even when the diver swims in the opposite direction. In summary, the proposed approach displays robust tracking stability throughout the entire duration of the tracking process.



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

**Figure 31.** *Cont.*



(**c**) (**d**)

**Figure 31.** Tracking results of single diver in a pool using ExDeepSORT. (**a**) The 70th image; (**b**) the 150th image; (**c**) the 200th image; (**d**) the 250th image.

### *6.4. Underwater Target Tracking Experiment Based on Sea Trial Data*

The sonar image sequence captured in the sea trial consists of 245 frames. Table 14 presents a comparison of the tracking results achieved by SORT, traditional DeepSORT, and ExDeepSORT.

In Table 16, the proposed methodology demonstrates merely two occurrences of slight modifications in target identification within sonar image sequences obtained from a real maritime environment, which is characterized by dynamic backgrounds. This indicates a significant improvement over both the SORT and traditional DeepSORT methods in mitigating changes in target identification and maintaining tracking stability. The proposed approach exhibits a trajectory interruption rate of 20.00%, which surpasses the performance of the other methods. It is important to highlight that trajectory interruptions primarily arise when the target moves through the reverberation zone, where reverberation noise interferes with or obscures the target, thereby complicating the system's detection capabilities.

**Table 16.** Tracking results based on sea experiment data.

| Tracking Framework | Frag Ratio | ID Switch |
|---|---|---|
| SORT | 33.67% | 12 |
| DeepSORT | 21.22% | 8 |
| ExDeepSORT | 20.00% | 2 |

In Figure 32a, the relative trajectories of the two targets illustrated are approximately parallel, with a significant reverberation region evident in the latter portion of the trajectory. Figure 32b–d indicate that as the targets pass through the reverberation zone, the tracking stability of the three methodologies experiences varying degrees of decline. This decline is characterized by interruptions in the trajectory, which can be attributed to the reverberation noise's intensity approaching or exceeding the imaging intensity of the targets. Such conditions complicate the differentiation of target image features and may even result in complete obscuration by the noise. The SORT method is particularly prone to generating numerous identity alterations throughout the tracking process, leading to considerable instability. In contrast, the traditional DeepSORT method encounters similar difficulties, as demonstrated by the tire target's identity changing while traversing the reverberation zone, followed by multiple identity changes in subsequent tracking phases. Conversely, the ExDeepSORT method has the ability to recover identity after passing through the reverberation zone, thereby maintaining a continuous and stable tracking process thereafter. The tracking results of the ExDeepSORT method applied to actual underwater sonar image sequences are presented in Figure 33.

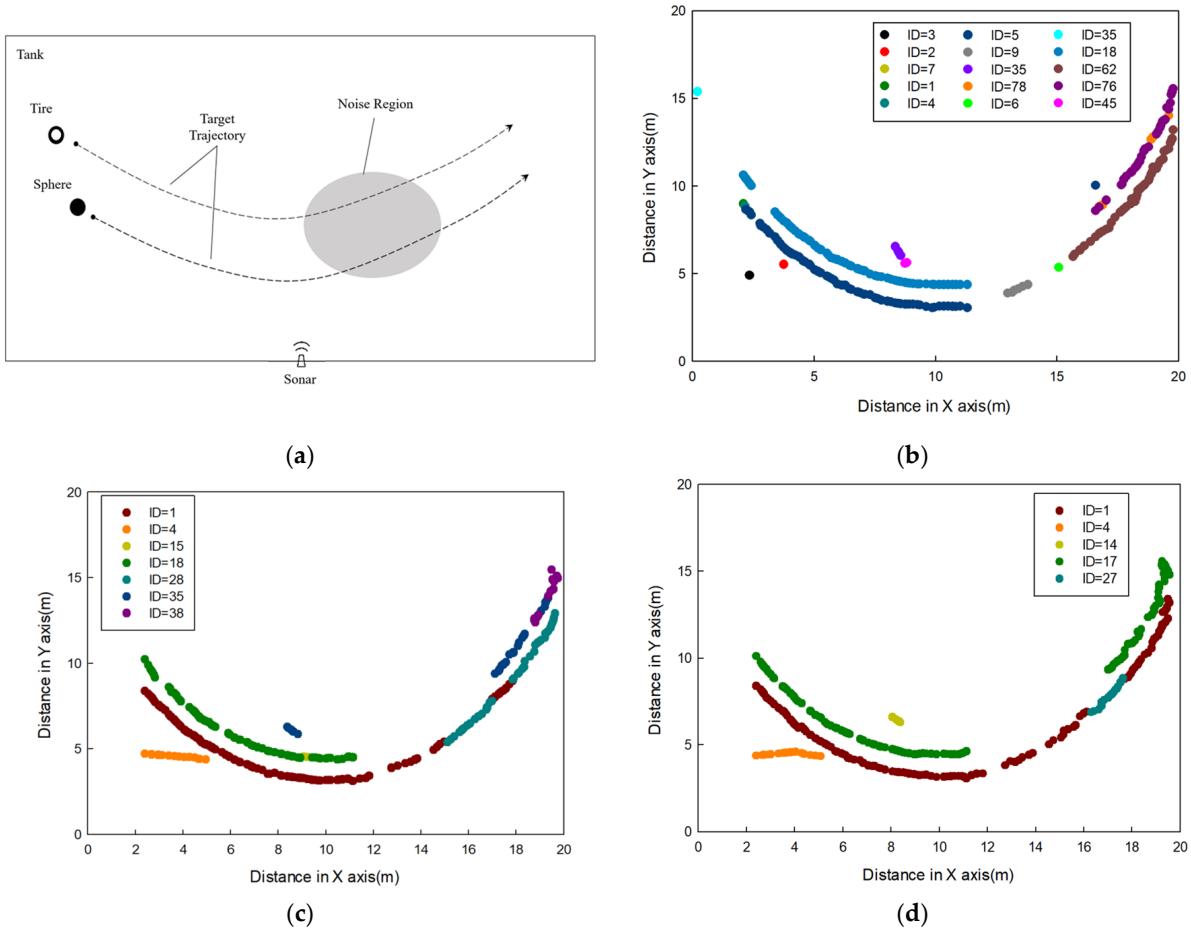



**Figure 32.** Tracking results based on sea trial data. (**a**) Target motion trajectory; (**b**) results obtained by SORT method; (**c**) results obtained by DeepSORT method; (**d**) results obtained by ExDeepSORT method.

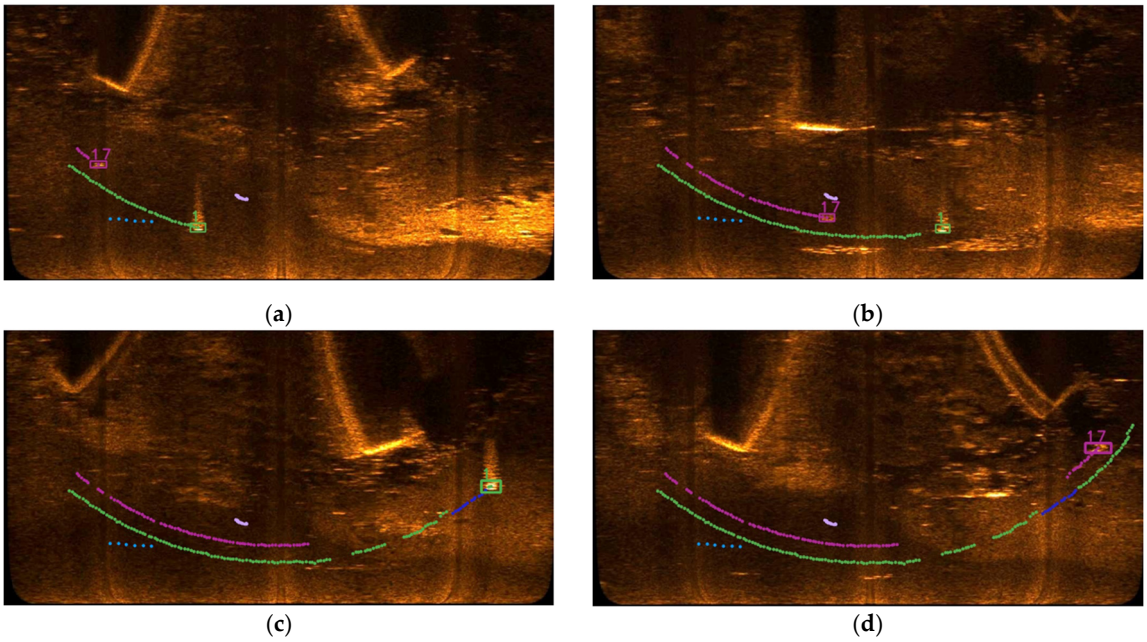



**Figure 33.** Tracking results obtained by ExDeepSORT method. (**a**) The 50th image; (**b**) the 100th image; (**c**) the 150th image; (**d**) the 200th image.

## 7. Conclusions

To meet the requirements of acoustic images, this paper introduced an underwater target tracking method. First, taking YOLOv5 as the base network, the STR block structure from the Swin Transformer network was introduced, which enhanced the global image information fusion capability of the YOLOv5 network and improved its detection performance for underwater targets. By increasing the dimensions of the target bounding box, the receptive field of the tracking network is expanded, which facilitates the capture of characteristic noise surrounding the target and enhances the information related to the target's features. A comparative analysis was conducted utilizing collected sonar sequences. The results indicate that the proposed methodology significantly enhances the stability of target tracking within sonar sequences, exhibiting consistent tracking performance even in complex scenarios characterized by path intersections and encounters.

For the target tracking task in the acoustic images, small samples are still an important factor limiting its accuracy. For this problem, acoustic image generation, transfer learning, and other methods should still be studied. At the same time, other advanced block-based or modular strategies require more in-depth consideration, such as denoising using 3D filtering and block matching block-based CNN models. In addition, pre-Gaussian filtering is also expected to improve the quality of images. In the future, we will try to improve the accuracy of acoustic image target tracking in multiple ways such as with data augmentation, the visualization of feature maps, attention mechanisms, and modern deep-learning-based trackers.

**Author Contributions:** Conceptualization, R.L. and W.Z.; methodology, R.L.; software, R.L.; validation, R.L. and W.Z.; investigation, W.Z.; data curation, R.L. and T.Z.; writing—original draft preparation, W.Z. and R.L.; writing—review and editing, W.Z. and H.Z.; supervision, W.Z. and T.Z.; project administration, W.Z. and T.Z.; funding acquisition, W.Z. and T.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 51879061; Guangdong Provincial General Colleges and Universities Characteristic Innovation Project, grant number 2023KTSCX215; Fundamental Research Funds for the Central Universities, grant number 23ptpy101 and 23xkjc012; Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), grant number SML2023SP228; Zhuhai Basic and Applied Basic Research Foundation, grant number 2320004002789; Guangdong Provincial General Colleges and Universities Key Scientific Research Platform and Project, grant number 2022KTSX188.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

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

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

## References

1. Williams Glen, N.; Lagace Glenn, E. A collision avoidance controller for autonomous underwater vehicles. In Proceedings of the Symposium on Autonomous Underwater Vehicle Technology, Washington, DC, USA, 5–6 June 1990.
2. Dai, D.; Chantler, M.J.; Lane, D.M.; Williams, N. A Spatial-Temporal Approach for Segmentation of Moving and Static Objects in Sector Scan Sonar Image Sequences. In Proceedings of the 5th International Conference on Image Processing and its Applications, Stevenage, UK, 4–6 July 1995.
3. Lane, D.M.; Chantler, M.J.; Dai, D. Robust Tracking of Multiple Objects in Sector-Scan Sonar Image Sequences Using Optical Flow Motion Estimation. *IEEE J. Ocean. Eng.* **1998**, *23*, 31–46. [CrossRef]
4. Chantler, M.J.; Stoner, J.P. Automatic Interpretation of Sonar Image Sequences Using Temporal Feature Measures. *IEEE J. Ocean. Eng.* **1997**, *22*, 29–34. [CrossRef]

5. Ruiz, I.T.; Lane, D.; Chantler, M. A Comparison of Inter-Frame Feature Measures for Robust Object Classification in Sector Scan Sonar Image Sequences. *IEEE J. Ocean. Eng.* **1999**, *24*, 458–469. [CrossRef]
6. Petillot, Y.; Tena-Ruiz, I.; Lane, D.M. Underwater vehicle obstacle avoidance and path planning using a multi-beam forward looking sonar. *IEEE J. Ocean. Eng.* **2001**, *26*, 240–251. [CrossRef]
7. Liu, D. Sonar Image Target Detection and Tracking Based on Multi-Resolution Analysis. Ph.D. Thesis, Harbin Engineering University, Harbin, China, 2011.
8. Perry, S.W.; Ling, G. Pulse-Length-Tolerant Features and Detectors for Sector-Scan Sonar Imagery. *IEEE J. Ocean. Eng.* **2004**, *29*, 35–45. [CrossRef]
9. Perry, S.W.; Ling, G. A Recurrent Neural Network for Detecting Objects in Sequences of Sector-Scan Sonar Images. *IEEE J. Ocean. Eng.* **2004**, *29*, 47–52. [CrossRef]
10. Blanding, W.R.; Willett, P.K.; Bar-Shalom, Y.; Lynch, R.S. UUV-platform cooperation in covert tracking. In *Signal and Data Processing of Small Targets 2005*; SPIE: Bellingham, WA, USA, 2005; Volume 5913, pp. 607–615.
11. Clark, D.E.; Bell, J. Bayesian Multiple Target Tracking in Forward Scan Sonar Images Using the PHD filter. *IEE Proc. Radar Sonar Navig.* **2005**, *152*, 327–334. [CrossRef]
12. Clark, D.E.; Tena-Ruiz, I.; Petillot, Y.; Bell, J. Multiple Target Tracking and Data Association in Sonar Images. In Proceedings of the 2006 IEEE Seminar on Target Tracking: Algorithms and Applications, Birmingham, UK, 19 June 2006; pp. 149–154.
13. Musicki, D.; Wang, X.; Ellem, R.; Fletcher, F. Efficient active sonar multitarget tracking. In Proceedings of the OCEANS 2006-Asia Pacific, Singapore, 16–19 May 2006; IEEE: New York, NY, USA, 2006; pp. 1–8.
14. Handegard, N.O.; Williams, K. Automated tracking of fish in trawls using the DIDSON (dual frequency identification sonar). *ICES J. Mar. Sci.* **2008**, *65*, 636–644. [CrossRef]
15. Yue, M. Research of Underwater Object Detection and Tracking Based on Sonar. Master's Thesis, Harbin Engineering University, Harbin, China, 2008.
16. Quidu, I.; Jaulin, L.; Bertholom, A.; Dupas, Y. Robust multitarget tracking in forward-looking sonar image sequences using navigational data. *IEEE J. Ocean. Eng.* **2012**, *37*, 417–430. [CrossRef]
17. DeMarco, K.J.; West, M.E. Sonar-Based Detection and Tracking of a Diver for Underwater Human-Robot Interaction Scenarios. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 13–16 October 2013; pp. 2378–2383.
18. Karoui, I.; Quidu, I.; Legris, M. Automatic sea-surface obstacle detection and tracking in forward-looking sonar image sequences. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4661–4669. [CrossRef]
19. Hurtós, N.; Palomeras, N.; Carrera, A.; Carreras, M. Autonomous detection, following and mapping of an underwater chain using sonar. *Ocean Eng.* **2017**, *130*, 336–350. [CrossRef]
20. AI Muallim, M.T.; Duzenli, O. Improve Divers Tracking and Classification in Sonar Images Using Robust Diver Wake Detection Algorithm. In Proceedings of the 19th International Conference on Machine Learning and Computing, Vancouver, BC, Canada, 7–8 August 2017; pp. 85–90.
21. Jing, D.; Han, J.; Wang, X.; Wang, G.; Tong, J.; Shen, W.; Zhang, J. A method to estimate the abundance of fish based on dual-frequency identification sonar (DIDSON) imaging. *Fish. Sci.* **2017**, *83*, 685–697. [CrossRef]
22. Ye, X.; Sun, Y.; Li, C. FCN and Siamese network for small target tracking in forward-looking sonar images. In Proceedings of the OCEANS 2018, Charleston, SC, USA, 10 January 2019.
23. Wang, X.; Wang, G.; Wu, Y. An adaptive particle swarm optimization for underwater target tracking in forward looking sonar image sequences. *IEEE Access* **2018**, *6*, 46833–46843. [CrossRef]
24. Fuchs, L.R.; Gällström, A.; Folkesson, J. Object recognition in forward looking sonar images using transfer learning. In Proceedings of the 2018 IEEE/OES Autonomous Underwater Vehicle Workshop, Porto, Portugal, 6 June 2019.
25. Sheng, M.; Tang, S.; Qin, H.; Wan, L. Clustering Cloud-Like Model-Based Targets Underwater Tracking for AUVs. *Sensors* **2019**, *19*, 370–375. [CrossRef]
26. Jing, D.; Han, J.; Xu, Z.; Chen, Y. Underwater multi-target tracking using imaging sonar. *J. Zhejiang Univ. (Eng. Sci.)* **2019**, *53*, 753–760.
27. Zhang, T.; Liu, S.; He, X.; Huang, H.; Hao, K. Underwater target tracking using forward-looking sonar for autonomous underwater vehicles. *Sensors* **2019**, *20*, 102–103. [CrossRef] [PubMed]
28. Gao, J.; Gu, Y.; Zhu, P. Feature Tracking for Target Identification in Acoustic Image Sequences. *Complexity* **2021**, *2021*, 8885821. [CrossRef]
29. Christensen, J.H.; Mogensen, L.V.; Ravn, O. Side-scan sonar imaging: Automatic boulder identification. In Proceedings of the Oceans 2021, San Diego, CA, USA, 15 February 2022.
30. Horimoto, H.; Maki, T.; Kofuji, K.; Ishihara, T. Autonomous sea turtle detection using multi-beam imaging sonar: Toward autonomous tracking. In Proceedings of the 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV), Porto, Portugal, 6 June 2019.

31. Sun, Y. Forward-Looking Sonar Underwater Target Tracking Technology Based on Deep Learning. Master's Thesis, Harbin Engineering University, Harbin, China, 2019.
32. Kvasić, I.; Mišković, N.; Vukić, Z. Convolutional Neural Network Architectures for Sonar-Based Diver Detection and Tracking. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019.
33. Ma, X. Research on Object Detection and Tracking Based on Forward-Looking Sonar. Master's Thesis, Harbin Engineering University, Harbin, China, 2020.
34. Pian, Z. Forward-Looking Sonar Underwater Target Detection and Tracking Technology Based on Deep Learning. Master's Thesis, Harbin Engineering University, Harbin, China, 2020.
35. Ye, X.; Zhang, W.; Li, Y.; Luo, W. Mobilenetv3-YOLOv4-Sonar: Object Detection Model Based on Lightweight Network for Forward-Looking Sonar Image. In Proceedings of the OCEANS 2021, San Diego, CA, USA, 15 February 2022.
36. Li, B.; Huang, H.; Liu, J.; Liu, Z.; Wei, L. Synthetic Aperture Sonar Underwater Multi-scale Target Efficient Detection Model Based on Improved Single Shot Detector. *J. Electron. Inf. Technol.* **2021**, *43*, 2854–2862.
37. Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; Ma, J. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. *Remote Sens.* **2021**, *13*, 3555. [CrossRef]
38. Chen, R.; Zhan, S.; Chen, Y. Underwater target detection algorithm based on YOLO and Swin transformer for sonar images. In Proceedings of the OCEANS 2022, Hampton Roads, VA, USA, 19 December 2022.
39. Cao, X.; Ren, L.; Sun, C. Research on Obstacle Detection and Avoidance of Autonomous Underwater Vehicle Based on Forward-Looking Sonar. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 9198–9208. [CrossRef]
40. Zhou, X.; Tian, K.; Zhou, Z.; Ning, B.; Wang, Y. SID-TGAN: A Transformer-Based Generative Adversarial Network for Sonar Image Despeckling. *Remote Sens.* **2023**, *15*, 5072. [CrossRef]
41. Chun, S.; Kawamura, C.; Ohkuma, K.; Maki, T. 3D Detection and Tracking of a Moving Object by an Autonomous Underwater Vehicle With a Multibeam Imaging Sonar: Toward Continuous Observation of Marine Life. *IEEE Robot. Autom. Lett.* **2024**, *9*, 3037–3044. [CrossRef]
42. Qin, K.S.; Liu, D.; Wang, F.; Zhou, J.; Yang, J.; Zhang, W. Improved YOLOv7 model for underwater sonar image object detection. *J. Vis. Commun. Image Represent.* **2024**, *100*, 104124. [CrossRef]
43. M-Series Mk2 Sonars Operator's Manual. TELEDYNE BlueView Press, Version 2. Available online: https://www.manualslib.com/brand/teledyne/sonar.html (accessed on 4 May 2020).
44. M-Series Sonars Quick Start Guide. TELEDYNE BlueView Press. Available online: http://ocean-innovations.net/OceanInnovationsNEW/TeledyneBlueview/M-Series-QuickStart-Guide-rev-.pdf (accessed on 2 July 2020).
45. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
46. 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.
47. Jocher, G. Available online: https://github.com/ultralytics/yolov5 (accessed on 19 June 2020).
48. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
49. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017.


**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.