Journal of
Marine Science and Engineering

MDPI

Article

Sonar Image Target Detection Based on Simulated Stain-like Noise and Shadow Enhancement in Optical Images under Zero-Shot Learning

Jier Xi and Xiufen Ye *

College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; xijier@hrbeu.edu.cn
* Correspondence: yexiufen@hrbeu.edu.cn

Abstract: There are many challenges in using side-scan sonar (SSS) images to detect objects. The challenge of object detection and recognition in sonar data is greater than in optical images due to the sparsity of detectable targets. The complexity of real-world underwater scanning presents additional difficulties, as different angles produce sonar images of varying characteristics. This heterogeneity makes it difficult for algorithms to accurately identify and detect sonar objects. To solve these problems, this paper presents a novel method for sonar image target detection based on a transformer and YOLOv7. Thus, two data augmentation techniques are introduced to improve the performance of the detection system. The first technique applies stain-like noise to the training optical image data to simulate the real sonar image environment. The second technique adds multiple shadows to the optical image and 3D data targets to represent the direction of the target in the sonar image. The proposed method is evaluated on a public sonar image dataset, and the experimental results demonstrate that the proposed method outperforms the state-of-the-art methods in terms of accuracy and speed. The experimental results show that our method achieves better precision.

Keywords: sonar image; transformer; stain-like noise; multiple shadows

check for updates

Citation: Xi, J.; Ye, X. Sonar Image Target Detection Based on Simulated Stain-like Noise and Shadow Enhancement in Optical Images under Zero-Shot Learning. *J. Mar. Sci. Eng.* **2024**, *12*, 352. https://doi.org/10.3390/jmse12020352

Academic Editor: Rafael Morales

Received: 1 January 2024
Revised: 6 February 2024
Accepted: 14 February 2024
Published: 18 February 2024

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

1. Introduction

The advancement of sonar technology has yielded remarkable achievements in underwater exploration [1] and target detection [2,3]. Compared with the limitations of optical sensors in detecting targets, such as short detection distances and poor underwater visibility, SSS-based target detection methods have become increasingly popular and effective. These methods [4–6] have proven to be more effective in terms of distance and visibility, overcoming the limitations of traditional optical sensors. The lack of sonar sample data and image quality remains a common problem in sonar target recognition. Researchers have developed various approaches to address these challenges, such as utilizing deep learning algorithms [7–10] to enhance the image quality of sonar data and applying transfer learning [11,12], allowing for more robust detection and recognition results. These methods have shown promising results in simulated underwater environments, but more research is needed to validate their effectiveness in real-world scenarios or simulated environments. However, the high cost associated with underwater experiments [13], including the deployment and recycling of underwater targets, the use of diverse sonar devices, and the search for suitable experimental areas, has resulted in a lack of available samples. As a result, it is challenging to obtain sufficient data to evaluate the performance of target recognition algorithms in real-world scenarios.

Some researchers have employed deep convolutional neural networks (DCNNs) using style transfer to simulate the environment [4,11,12,14]. This has highly improved the performance of sonar image detection. Due to the complex underwater environment, some key features will be lost in the simulated environment. Several experts in the field have utilized semantic segmentation to classify targets in sonar images [15–17]. However, most of them have focused on image classification. Hence, this paper provides a comprehensive approach that considers the complexities of the underwater environment and employs feature enhancement techniques for accurate target detection. This approach is combined with the semantic segmentation method in an optical image dataset to address the lack of available samples and enable the evaluation of target recognition algorithms in real-world scenarios. This makes up for the loss of some features in the simulation environment despite the high cost associated with underwater experiments.

First, we used the semantic segmentation method [18,19] on optical images to extract the target. However, the limitation of optical single targets in images causes a low performance in object detection. Therefore, we propose using shadow enhancement on targets to solve the problem of sonar target features in the real environment to improve the performance of sonar target detection. Furthermore, we added stain-like noise on targets to simulate mud and sand obstruction and mutilated targets in the real environment. Finally, considering the style difference between sonar images and optical images, we used style transfer to enhance object features via frequency analysis in real sonar images.

The remainder of this paper is organized as follows: In Section 2, we provide an overview of the existing methods and highlight their shortcomings. In Section 3, we present our proposed methods, which combine data augmentation and simulation techniques. These techniques are based on shadow enhancement and the addition of stain-like noise to the data. Section 4 presents a comparison of our proposed methods with existing approaches, along with the training results. Additionally, we provide an analysis of the experiments we designed and conducted, along with a comparison of their results.

2. Related Works

Scholars have researched sonar synthetic image datasets [2,3] and zero-shot learning methods [4,11,14,20] to augment samples to overcome the shortage of samples and simulate sonar images. Pre-trained DCNNs and fine-tuning techniques are powerful methods for sonar image detection [21,22]. William et al. [16] present an approach for merging SSS data and bathymetry information to improve automatic shipwreck identification. The method combines raw SSS images with a 2D relief map into a composite RGB image and uses a supervised image segmentation approach to identify shipwrecks. Zhao et al. [4] utilized a combination of 3D modeling, amplified data, equipment noise, and image mechanisms to extract target features and simulate target damages and postures using a DCNN and a fine-tuning style-transfer method. Their approach achieved a precision of 85.3% and a recall of 94.5%. Li et al. [6] identified texture features as domain-specific features and proposed to narrow the domain gap by removing these features. This method successfully transferred knowledge from optical images to sonar image classification tasks. The approach shows promise for improving the performance of sonar image classification tasks. Lee et al. [12] employed StyleBankNet [23] to perform style transfer simulations on optical images of the human body, which improved sonar object detection and achieved a precision of 86%. The authors generated samples using CAD but noted that significant simulation work was required to generate sufficient samples. This approach shows promise for enhancing sonar object detection using simulated optical images. Song et al. [24] introduced an effective sonar segmentation approach that leverages speckle noise analysis for pixel-wise classification. This method involves a single-stream deep neural network (DNN) with multiple side outputs to optimize edge segmentation. Huo et al. [25] utilized a transfer learning method to leverage knowledge from the ImageNet dataset to classify underwater targets in an SSS image dataset they built. They proposed using a semisynthetic data generation method during the transfer process to produce sonar images that effectively compensate for insufficient data. Ochal et al. [20] conducted a comparison of multiple supervised and semi-supervised few-shot learning (FSL) methods using underwater optical and SSS imagery. The results indicate that FSL methods have significant advantages over simple transfer learning methods, such as fine-tuning a pre-trained model for underwater target classification. Yu et al. [9] proposed a novel method for underwater target recognition, integrating a transformer module and YOLOv5. The method also incorporates an attention mechanism to improve both accuracy and efficiency. Xu et al. [22] proposed an active instance segmentation method combining a region-based convolution neural network (R-CNN) and balanced sampling. The method has benefits when a limited number of labeled samples are available, leading to better results for underwater shipwreck detection.

These enhancements make the methods well-suited for underwater environments where target recognition is challenging. However, target features cannot be properly expressed without considering the image environment (the state of the target, such as target damage and corruption, target postures, etc.), and a simulated image cannot properly present sonar features. Many studies have focused on sample amplification and image-processing mechanisms for underwater target recognition but have not sufficiently considered the challenges posed by real-world underwater environments, such as mud and sand obstruction, missing target parts, multiple target states, and shadows and reflections in sonar data.

3. Our Methods

Our method is based on yolov7 and a transformer backbone model to address the issue by enhancing multiple shadows on the target. The detection process in a DCNN involves the relationship between the target feature A and the model feature T.

For the contributions of this study, we define three feature sets including optical target image features, shadow enhancement features, and random stain-like noise features to describe the feature mapping process. By adding random stain-like noise to the target image, the optical image is simulated for a sediment-covered, mutilated target to improve the uncertainty of target states. Moreover, an image-processing method based on the existing style transfer method is proposed for data training to more closely represent the real data and enhance the object features. From the perspective of feature matching, the more features of the target A contained in model T, the higher the similarity.

3.1. Problem Definitions and Our Framework

A lack of samples is a common problem in target detection in sonar images, which leads to low model performance. Many methods are applied to transfer optical data to sonar data to improve target detection performance, but these methods do not fully consider the underwater environment. Given this issue, the key to successful deep learning work is preparing datasets with appropriate target features. In this section, we focus on our main contributions to this field, which include extracting complex features from datasets and utilizing zero-shot learning for target detection.

We define three types of features, D, S, and T, to describe the feature mapping process. $D_{(x,y)} = \{(x_i, y_j) | y_j = \{x_1, x_2, \ldots, x_n\}, j = 1, 2, \ldots, m\}$ denotes the domain of the optical target features. y_m indicates the mth image feature. x_n expresses the nth feature. $S_{(x,y)}$ denotes the domain of the shadow enhancement features that extends from D. $T_{(x,y)}$ denotes the domain of random stain-like noise to extend features S. The optical image features of the target set $D_{(x,y)}$ are expressed as $D_{(x,y)} = \begin{pmatrix} x_{11}, x_{12}, \ldots, 0, \ldots x_{1n} \\ x_{21}, x_{22}, \ldots, 0, \ldots x_{2n} \\ \ldots\ldots \\ x_{m1}, x_{m2}, \ldots, 0, \ldots x_{mn} \end{pmatrix}$. To standardize the dimensions of all features, zero-padding is used to ensure that all images have the same dimensions. The target is extracted from the background, so the background is empty and presents as zero. The values $(x_{11}, x_{12}, \ldots, x_{1n})$ denote the features of the original target. By augmenting the features $D_{(x,y)}$ with shadow features, the features $S_{(x,y)}$ are obtained. Simultaneously, a simulation of the complex underwater environment is performed with the addition of random stain-like noise using the method to obtain the features $T'_{(x,y)}$. We define an equation to describe the generated training data on one target, which combines the features (by summing the features). The features can be expressed as follows:

$$T_{(x,y)} = D_{(x,y)} \bigoplus S_{(x,y)} \bigoplus T'_{(x,y)} \tag{1}$$

The detection process aims to map the relationships between the real sonar target features A and features $T_{(x,y)}$ for sonar target detection. We consider three major aspects in the dataset design: (1) defining the dataset and augmentation from the optical image to extend multiple shadows on the same target; (2) transferring the optical image to a sonar-style image; (3) and designing stain-like noise on the target to simulate mud and sand obstruction.

The data processing to generate the training data in our experiments is shown in Figure 1.

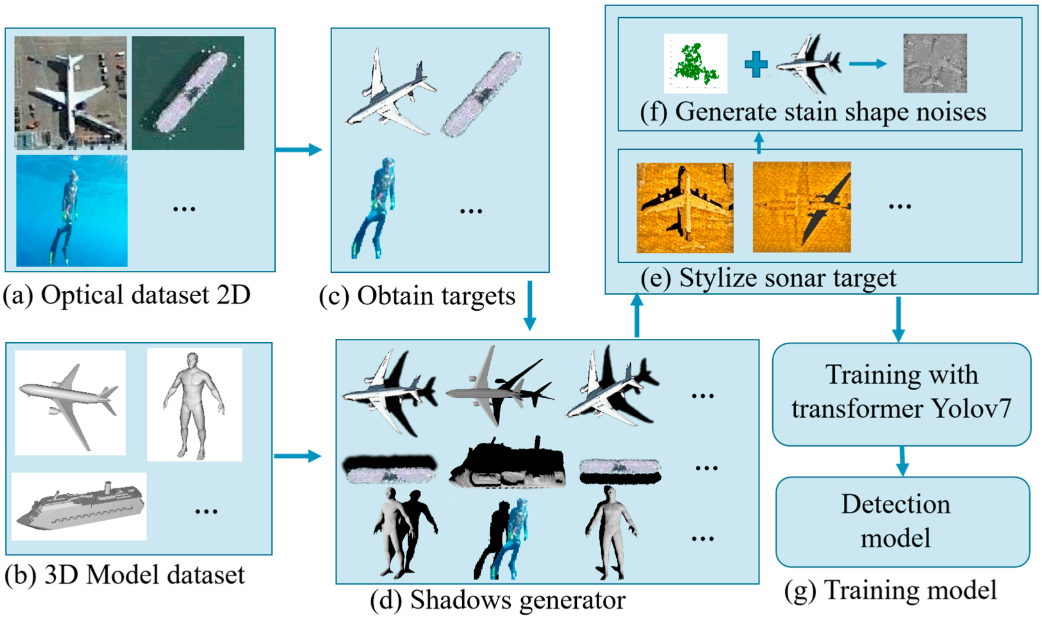


Figure 1. Data processing to generate training data on one target.

The process in Figure 1 includes four parts: First, different optical image datasets are integrated into target categories, such as airplanes, ships, cars, etc. Second, multiple shadows are generated and the direction for simulating the SSS image targets on the dataset is adjusted. Third, noise is generated on the targets to simulate covered and incomplete targets, whereby sediment occlusion on the seabed is simulated by adding stain-like noise. Fourth, yolov7 is used as the framework, and the transformer is used as the backbone method for data training. The detection model is then used to detect target objects in the sonar data.

3.2. Feature Enhancement and Augmentation Methods

The most existing methods focus on amplifying samples from an optical image dataset with less consideration of the optical background, which can impact the detection performance. The target shadow plays an important role in real sonar target image detection. Examples of real sonar images are shown in Figure 2.

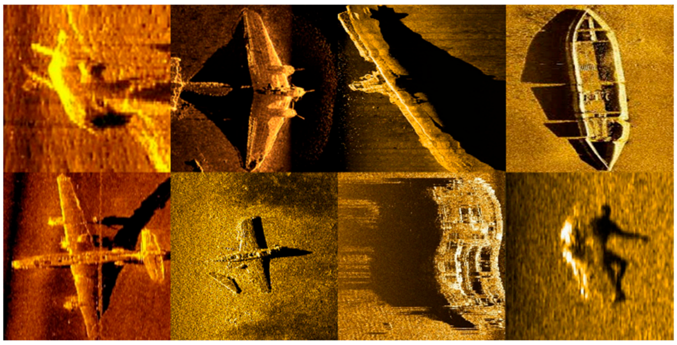
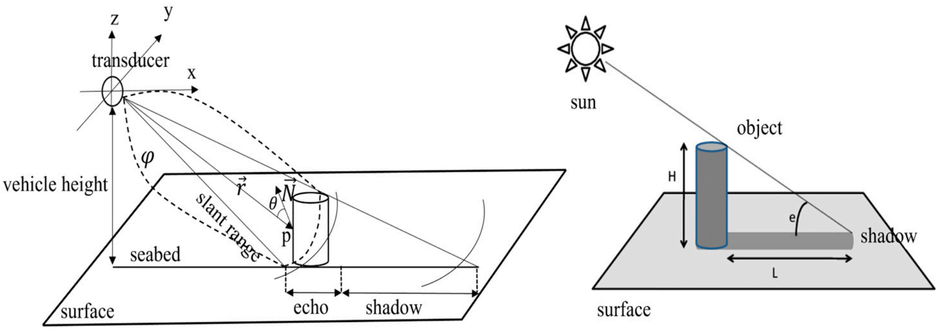


Figure 2. Real sonar images with shadows.

The shadow between an underwater acoustics image and an optical image are illustrated in [26]. Observation geometry given by the range and elevation angle is important for interpreting the highlight and shadow in an image. Examples of a shadow feature in sonar and optical images are shown in Figure 3.



(a) The shadow of sonar image (b) The shadow of optical image

Figure 3. Shadow features in sonar and optical images.

As shown in Figure 3, the target shadow in a real SSS image is always on the left or right side. The processes for a sonar shadow and an optical shadow are very similar [27]. The side-scan image formation process is briefly sketched in Figure 3a. The intensity of the corresponding pixel of the side-scan image depends on the amount of energy scattered back from the surface point. The traditional Lambertian model [28] permits us to derive the returned intensity from the parameters defining the observed scene [29]. The point p and intensity I can be expressed as follows:

$$\begin{cases} I(p) = K\varphi(p)R(p)|\cos(\theta(p))|, \\ \qquad \vec{r} = (x, 0, Z(x,y)), \\ \vec{N} = (-\frac{\partial Z}{\partial x}(x,y), -\frac{\partial Z}{\partial x}(x,y), 1) \end{cases} \tag{2}$$

where φ represents the intensity of the illuminating sound wave at point p, R is the reflectivity of the object, θ is the incidence angle of the wave front, and $K \in [0,1]$ is a normalization constant. To obtain the maximum intensity, return Imax at any surface point, K is set as 1, and the reflectivity and incident intensity values are both 1 for the optimal surface orientation, with respect to the incident illumination. $\vec{N}$ and $\vec{r}$ are a coordinate system relative to the sensor (Figure 3a). To simplify the process in our experiment, we defined the seafloor as a flat surface, denoted as $C_{surface}$, with a constant value to express the surface intensity. Under this assumption and the combination of expressions in (2), $\frac{\partial Z}{\partial x}$ and $\frac{\partial Z}{\partial y}$ yield an expression that depends on Z when applying finite difference methods on the gradients. The intensity returned from an object point can be represented with the following expression:

$$\begin{cases} I(x,y) = \dfrac{-Z+\frac{x^2}{-Z}\cdot(1+>(\partial_y Z>)^2)}{\sqrt{x^2+Z^2}\cdot\sqrt{1+\left(\partial_y Z\right)^2+\frac{x^2}{Z^2}\cdot\left(1+\left(\partial_y Z\right)^2\right)^2}}, \\ I(x,y) = C_{surface}, (surface\ intensity) \end{cases} \quad (3)$$

A shadow is a critical feature in deep learning detection work. The shadow of an object has discrepancies due to the object's posture and sonar position. We propose a method that uses a large amount of optical data and 3D model data [30] to improve an object's features via shadow enhancement in the training data. First, we split the optical target and background to reduce interference when the deep learning network extracts target features. Second, we generate a target image with multiple shadow features via a 3D model by adjusting the object and light position.

From the overall process of target extraction and shadow as shown generator in Figure 1. The goal of a deep learning network is to extract object features. Since our experiments were based on zero-shot learning, the backgrounds of the optical images lack features from the sonar images. We adopted finetuning DeepLabV3 [31,32] as a semantic segmentation method to extract target features from the optical image. The optical image was segmented and processed to obtain an image containing only the target, thus improving the model's recognition rate. In addition, we adopted 3D data and employed the fine-tuning exponential shadow maps (ESM) method [33–35], combining lighting and object position techniques for shadow simulation. Figure 1a–d show the processes of segmentation and shadow enhancement in our experiment.

The research on real sonar data showed that simulating the features of sonar images using two-dimensional image shadow simulation cannot fully simulate shadows. A comparison between real sonar images and stylized 3D-generated images is shown in Figure 4.

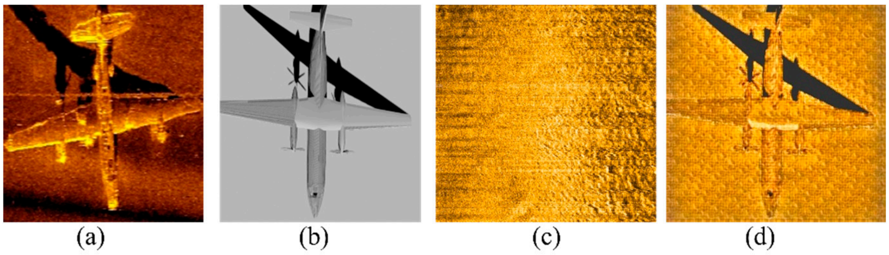


Figure 4. Three-dimensional-model-generated sonar-style images. (**a**) Sonar image; (**b**) 3D model simulating shadow; (**c**) sonar-style image; (**d**) and sonar-style image.

In a two-dimensional image, we defined function f as the shadow enhancement function. $g(x)$ is the semantic segmentation function. x is the original optical image. A is the enhanced features, which can be expressed as $Z = f_i^j(x, g(x))$, $i \in \{0, 1, \ldots, 360\}$, $j \in N$. In function f_i^j, i is shadow angle, and j is shadow width. Z is the entire enhanced features from one original image.

The rotation matrix $R(\alpha, \beta, \gamma) = R_z(\alpha)R_y(\beta)R_x(\gamma)$ is used to calculate the target with its shadow in the image.

$$R_z(\alpha) = \begin{bmatrix} cos\alpha & -sin\alpha & 0 \\ sin\alpha & cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}, R_y(\beta) = \begin{bmatrix} cos\beta & 0 & sin\beta \\ 0 & 1 & 0 \\ -sin\beta & 0 & cos\beta \end{bmatrix}, R_x(\gamma) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\gamma & -sin\gamma \\ 0 & sin\gamma & cos\gamma 1 \end{bmatrix}.$$

In the matrix, $R_z(\alpha)$ represents the rotation of an object around the z-axis by α degrees. $R_y(\beta)$ and $R_x(\gamma)$ represent the rotations of the y-axis and x-axis.

In the three-dimensional model, $R'' = R(\alpha, \beta, \gamma)$. The transformation move matrix is defined as $T'' = \begin{bmatrix} 1 & 0 & 0 & \Delta x \\ 0 & 1 & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

In the two-dimensional image, $R^{'} = R_z(\alpha)$. The transformation matrix is defined as $T' = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix}$.

The shadow calculation process can be expressed as follows:

$$\begin{cases} \ddot{A} = \dot{A}_{(i,j)} \cdot R' \cdot T' + A_{(i,j)}, if \min\left(\dot{A}_{(i,j)}, A_{(i,j)}\right),\ (i,j) > 0 \\ \dddot{A} = f_{CSM}(O \cdot R'' \cdot T'') \end{cases} \tag{4}$$

where $\ddot{A}$ is the final image matrix with the shadow in the two-dimensional image. Where $A_{(i,j)}$ is the target without the background and $\dot{A}_{(i,j)}$ denotes the target shadow generated from $A_{(i,j)}$. (i,j) represents the coordinate positions. $\dddot{A}$ is the final image matrix with the shadow in the three-dimensional model. O is the target matrix in the 3D model. Figure 5 displays the generated 3D data shadow compared with the real sonar data.

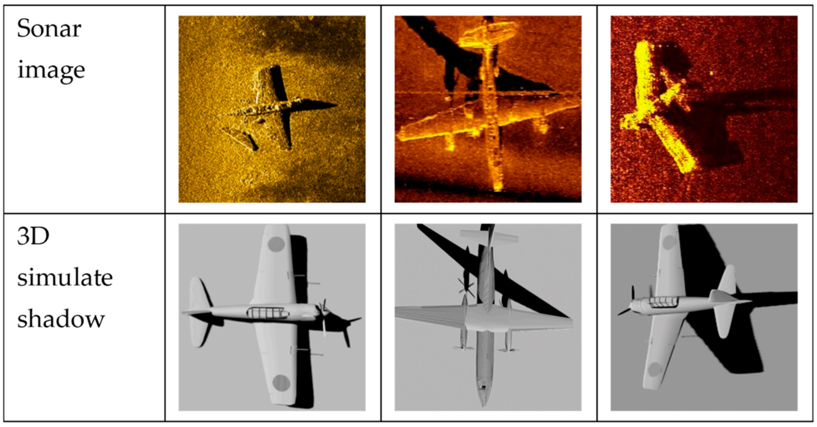


Figure 5. Shadow image sample generated via 3D model and real sonar data.

Figure 6 displays the airplane sample data for the feature expansion of the target image in Figure 1c, using shadow feature enhancement methods.

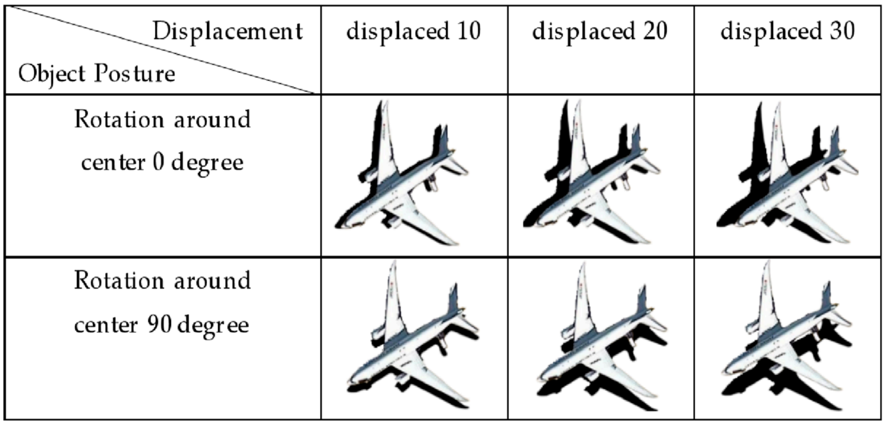
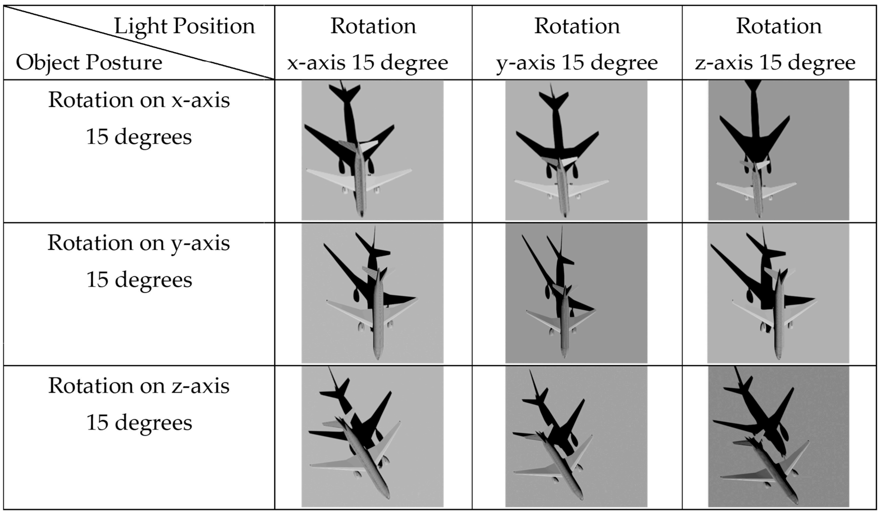

Displacement
Object Posture
displaced 10
displaced 20
displaced 30
Rotation around center 0 degree
Rotation around center 90 degree

Figure 6. Sample data for shadow feature expansion in two-dimensional image.

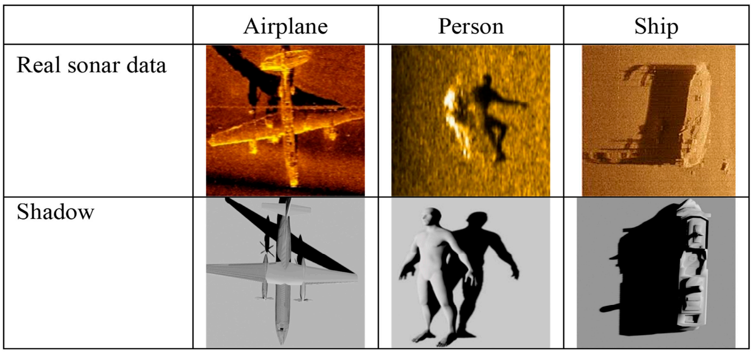
The displacement generally depends on the center of the original image. We define a 10 pixel displacement in a (512, 512) image in the examples. The rotation angle is around the image center as the axis, and the angle of the object's shadow is determined by the object's orientation and simulated lighting. Figure 7 displays the sample data for the shadow feature expansion of the 3D target using shadow feature enhancement methods.



Light Position / Object Posture	Rotation x-axis 15 degree	Rotation y-axis 15 degree	Rotation z-axis 15 degree
Rotation on x-axis 15 degrees			
Rotation on y-axis 15 degrees			
Rotation on z-axis 15 degrees			

Figure 7. Sample data for shadow feature expansion in three-dimensional model.

The simulated shadow enhancement with different targets is shown in Figure 8.



	Airplane	Person	Ship
Real sonar data			
Shadow			

Figure 8. Shadow enhancement with different targets.

3.3. *Stain-like Noise Method*

Many scholars extract targets from optical images and simulate defects, which can effectively replicate the defects in the targets. However, these defect simulation methods are limited to the targets and do not adequately represent the surrounding environment and shadows.

Optical and acoustic images of an aircraft target were derived from [36] to better understand the real environment, as shown in Figure 9.

Many of the targets to be detected in actual sonar image applications are incomplete or defective targets. The diagram in Figure 10 illustrates varying degrees of burial of the targets by sediments, resulting in minimal obstruction, moderate obstruction, and significant obstruction of the targets.

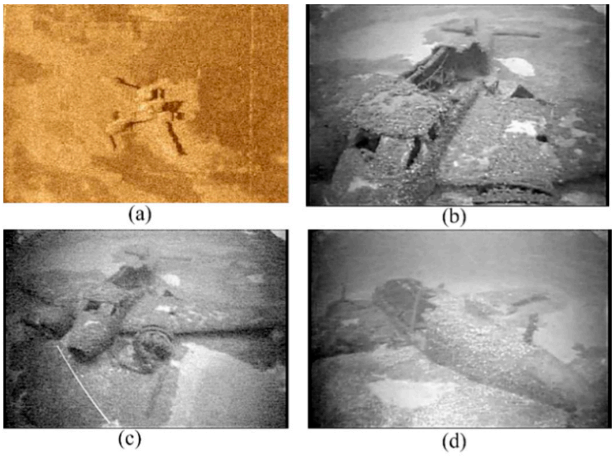


Figure 9. Examples of incomplete targets. (**a**) Original sonar image; (**b**) original optical image; (**c**) front-view optical image of the target; (**d**) rear-view optical image of the target.

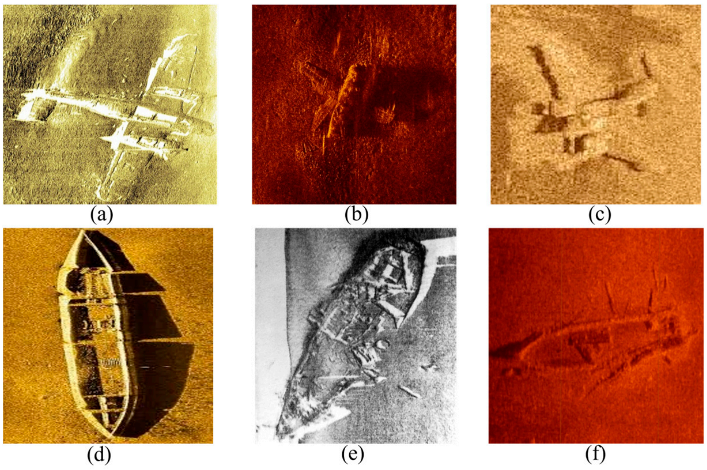


Figure 10. Examples of varying degrees of burial of the targets by sediments. (**a**) Minimal obstruction of airplane; (**b**) moderate obstruction of airplane; (**c**) significant obstruction of airplane; (**d**) minimal obstruction of ship; (**e**) moderate obstruction of ship; (**f**) and significant obstruction of ship.

With the rapid development of DCNNs, object features can be easily extracted from data. A DCNN network is capable of extracting object features even from zero-shot learning, with minimal real-world conditions. This is because the training samples used for a DCNN are almost perfect and may not accurately represent the complexities and variabilities presented in real-world conditions. We propose a random stain-like noise method to simulate the damage, occlusion, and other factors in real sonar imaging targets underwater, which can effectively improve recognition efficiency. The proposed method was proven to be effective in the experiments. The single process of generating data with stain-like noise is shown in Figure 11.

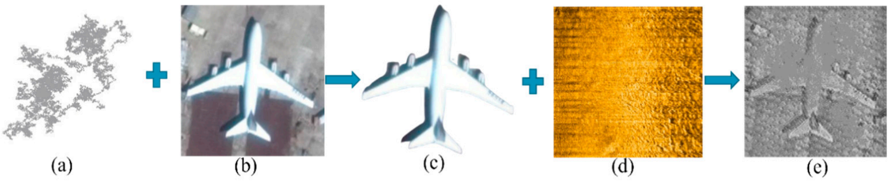
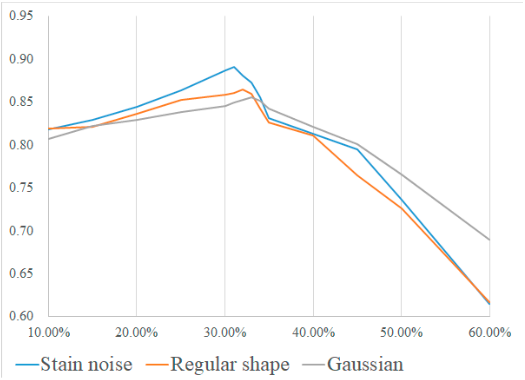


Figure 11. The single process of generating data with stain-like noise. (**a**) Random stain-like noise; (**b**) original image; (**c**) optical target; (**d**) sonar image background; (**e**) and sonar-style target with stain-like noise.

Our study compared the recognition performance for different types of noise, and we found that random stain-like noise resulted in the highest performance, as depicted in Figure 11. However, stain-like noise can lead to overfitting, which affects the recognition accuracy. To address this, we finetuned the data and achieved a peak performance of 0.89 mean average precision (mAP) [37] when the noise occupied approximately 31% of the target image in our experimental data.

Figure 12 compares the performance trends for the different noise types in the noise-occupied area on the target. The noise-occupied area on the target ranged from 10% to 60%. We found that the difference between a noise-occupied area of less than 10% and no noise was minimal. The performance greatly decreased with a noise-occupied area of over 50% due to overfitting.



X: The proporation of target image area occupied by noise.
Y: mAP value (IOU = 50%)

Figure 12. Performance trends for different noise types.

The process of stain-like noise is exhibited in Algorithm 1.

Algorithm 1: Process of stain noise

```
Input: Scale $n$ is the number of stain points. $\theta$ is the area ratio.
C is a constant value which expresses the pixels of one stain point.
Steps: Directions of walk (up, down, right, left). Number of walks.
Output: The image with random stain points.
Initialize: Size of image (width, height). The maximum area proportion of stain points
In image. Calculate $n$ by $\theta$.
for $i$ in scale($n$) do
    for step in walks do
        if is over the maximum area proportion then
            Return image
        end
        if is the direction being walked then
            Update the direction
        end
        Update stain noise in image
    end
end
```

An example with different parameters is shown in Figure 13.

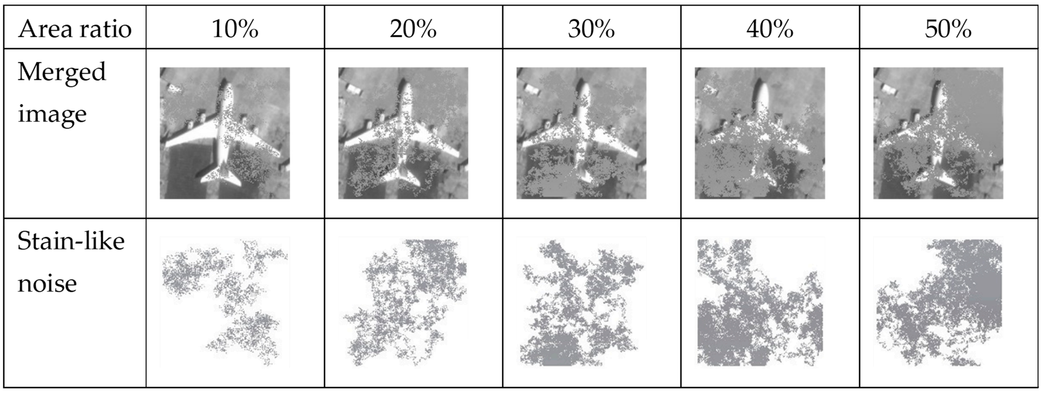


Figure 13. Example of target with different area ratios with stain-like noise.

The simulation process for generating stain-like noise data can be expressed as the following equation:

$$\begin{cases} g(x,y) = x \odot \left(C \cdot \sum_{i,j}^{i<h,j<w} y_{(i,j)} \cdot A_{(i,j)} \right) \\ A_{(i,j)} = \begin{bmatrix} 0,\ldots,0 \\ 0,\ldots,0 \\ \ldots, a_{i,j}, \ldots \\ 0,\ldots,0 \\ 0,\ldots,0 \end{bmatrix} \end{cases} \tag{5}$$

We use $g(x,y)$ to express the generated stain-like points on the target. x is the original image. C is a constant that expresses the pixels of one stain-like point. y is one stain-like point, and (i,j) is its position. $A_{(i,j)}$ is the image matrix of generated stain-like points. Zeros are used in $A_{(i,j)}$ to create an empty background for the stain-like image. h is the image height. w is the image width.

4. Experiment and Analysis

In this section, we present a series of experiments to compare our proposed method with the existing methods. The experiments were conducted on different datasets. We report our method's performance using several evaluation metrics commonly used in the field.

In this study, we adopted precision, recall, and mAP to evaluate the model's performance. True positive (TP) means that the network detection is a target and is correct. False positive (FP) means that if a sample does not belong to a class but is predicted to, it is considered a false positive. False negative (FN) means that if a sample belongs to a class but is predicted not to, it is considered a false negative.

Precision signifies the proportion of accurately predicted positive samples to the total number of predicted positive samples:

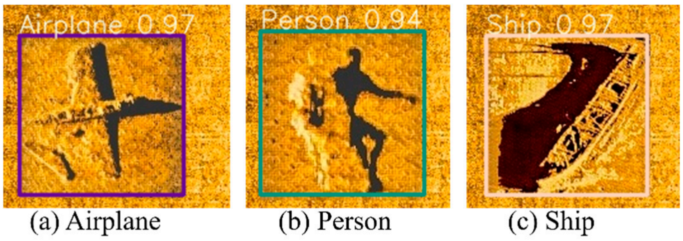
$$Precision = \frac{TP}{TP + FP}$$

Recall signifies the proportion of correctly predicted positive samples to the overall number of positive samples:

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

With the results of our experiments, we believe that our method has the potential to be used in real-world applications and can contribute to the advancement of the underwater detection field.

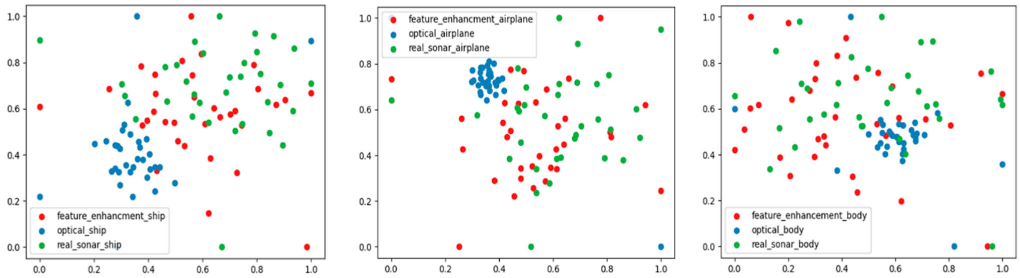
Our model can be fitted to customized target sizes, which can be defined in the training data. We adjusted the target size to (128,128) in the training data. Figure 14 shows examples of the detection of different, real-sonar targets selected from our test results.



(a) Airplane (b) Person (c) Ship

Figure 14. Detection of different real sonar targets.

In our experiments, we extracted target features from our trained model with different types of images with the t-distributed stochastic neighbor embedding (t-SNE) method [38] to better understand the impact of enhancement features on the target in an optical image, as shown in Figure 15.



(a) Feature distribution of ship (b) Feature distribution of airplane (c) Feature distribution of body

Figure 15. Features extracted on different targets with t-SNE distribution.

Based on the results, by enhancing the optical image targets, the enhanced target images have similar distributions to the real sonar targets.

4.1. Experimental Data

To increase the diversity of the target forms, we trained our model on a portion of the VOC2012 dataset, the human pose and NWPU VHR-10 [39,40] image dataset, and the pascal and human pose 3D dataset [41]. We conducted a comparative experiment on the different datasets under the same batch, verifying 29 real aircraft wreck sonar images, five real body sonar images, and 43 real shipwreck sonar images that are publicly available on the internet. The results show that our model is effective in detecting both aircraft wrecks and shipwrecks in sonar images and can be used for practical applications in underwater target recognition. We used three types of targets with shadow enhancement and stain-like noise in our experiments, including an airplane, person, ship, and others. The enhanced training dataset and test data is shown in Table 1.

Table 1. Training and test data.

Class	Training Data (Optical Image and 3D-Model)	Test Data (Real Sonar Image)
Airplane	3648	29
Person	3180	5
Ship	3608	43
Others	2800	9

Table 2 compares the performance of our method with that of state-of-the-art methods.

Table 2. Comparison with existing methods' performances.

Model	Precision	Recall	mAP (IOU = 0.5)
StyleBank + fastrcnn [12]	0.860	0.705	0.786
Whitening and coloring transform [14]	0.875	0.836	0.75
Improved style transfer + yolov5 [4]	0.853	0.945	0.876
Yolov5 + style transfer + regular-shaped noise [11]	0.899	0.861	0.865
Our method: Yolov7 (transformer backbone) + stain-shaped noise	0.903	0.857	0.891

In the comparison table, the precision with our method is increased by 0.004 compared with the existing methods' highest precision. Our method's recall is decreased by 0.088 compared with the existing methods' top recall, and its mAP is increased by 0.015 compared with the existing top mAP.

4.2. Experiment with Different Noise Types

We employed the Yolov7 framework–transformer backbone model on different types of noise datasets to investigate the performance of each noise type, including Gaussian, salt and pepper, regular shapes, stain-like shapes, and no noise. Examples of the noise types are shown in Figure 16.

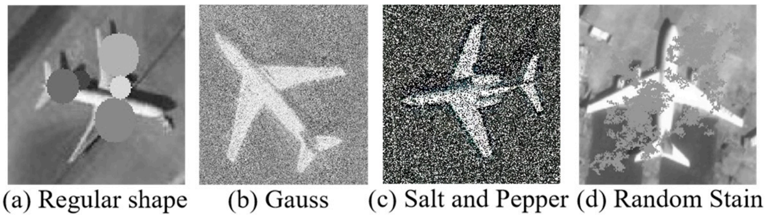


Figure 16. Examples of noise types.

Table 3 presents the performances of the different types of noise in the Yolov7 framework.

Table 3. Comparison between performances of different noise types.

Noise Type	No Noise	Gaussian	Salt and Pepper	Regular Shape	Stain-like Shape
mAP	0.739	0.803	0.806	0.816	0.824

Our analysis of the experimental results reveals that the highest mAP achieved for the recognition of stain-like noise was 0.824.

4.3. Experiment with Different Models

We conducted experiments using different models on the same dataset with shadow enhancement and random stain-like noise to further verify the detection performance. The comparison is shown in Table 4. Our experiments reveal that the combination of the two models exhibits better detection performance. The results show that the Yolov7 framework–transformer backbone model has significant potential to enhance object detection accuracy in various real-world applications.

Table 4. Comparison of different models' performances.

Model	Yolov5	Yolov7	Yolov5 (Transformer Backbone)	Yolov7 (Transformer Backbone)
mAP	0.742	0.815	0.843	0.891

The results show that the Yolov7 framework–transformer backbone model achieved the highest recognition mAP of 0.891. We obtained the best performance in all model comparison experiments using stain-like shapes and shadow enhancement as the training dataset.

4.4. Experiment on Shadow Enhancement

To verify whether the target shadow features increased the detection performance, we conducted an experiment using the Yolov7 framework to compare two datasets: one with shadow enhancement features and another without. The results of the comparison are presented in Table 5.

Table 5. Comparison of performance with and without shadow enhancement.

Data	No Shadow Enhancement	Shadow Enhancement
mAP	0.763	0.806

Our experimental results demonstrate that shadow enhancement is an effective data augmentation technique for improving the performance of sonar target detection models. Using simulation methods based on shadow enhancement can improve the model's ability to generalize real-world scenarios, resulting in a higher recognition mAP of up to 0.806. It should be noted that the detection result uses the model without real data in the training phase (only enhanced optical image and 3D-models are in the training dataset).

5. Conclusions

In this paper, we applied a transformer as the backbone model of Yolov7 to improve the underwater detection performance, despite a lack of training data. We addressed the design considerations for complex underwater scenarios, the limitations of lost features with style transfer, and targets covered by mud and sand. Hence, we proposed a method that merges stain-like noise on a simulated target to overcome the constraints of the real environment. Furthermore, we removed the background from optical target images to focus the training model on target features and reduce useless information. Additionally, we used shadow enhancements on the targets in two-dimensional images and a CSM shadow generator on a 3D model. The method addressed the key features of the target shadows, which would otherwise be missing when directly using optical object style transfer. Using comparison experiments, we demonstrated that our proposed method could achieve a better target detection performance than other methods that do not include shape noise fusion and key feature enhancement in the training data.

Future research could, for instance, investigate the relationship between the percentage of noise occupying the target and the dataset size, target number, and target categories.

Author Contributions: Conceptualization, J.X. and X.Y.; methodology, J.X.; software, J.X.; validation, J.X. and X.Y.; formal analysis, J.X.; investigation, J.X.; resources, J.X. and Y.X; data curation, J.X. and X.Y.; writing—original draft preparation, J.X.; writing—review and editing, J.X. and X.Y.; visualization, J.X.; supervision, X.Y.; project administration, J.X.; funding acquisition, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant no. 42276187) and the Fundamental Research Funds for the Central Universities (grant no. 3072022FSC0401).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the experimental data and code can be found at https://github.com/xijier/SonarDetection. (accessed on 5 February 2024).

Acknowledgments: We would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions that greatly improved the quality of this paper. We thank the researchers who have published sonar and optical image datasets. https://www.edgetech.com/underwater-technology-gallery/. https://sandysea.org/projects/visualisierung. (accessed on 5 February 2024).

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

References

1. Sahoo, A.; Dwivedy, S.K.; Robi, P.S. Advancements in the field of autonomous underwater vehicle. *Ocean Eng.* **2019**, *181*, 145–160. [CrossRef]
2. Wang, J.; Li, H.; Huo, G.; Li, C.; Wei, Y. Multi-Modal Multi-Stage Underwater Side-Scan Sonar Target Recognition Based on Synthetic Images. *Remote Sens.* **2023**, *15*, 1303. [CrossRef]
3. Er, M.J.; Chen, J.; Zhang, Y.; Gao, W. Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review. *Sensors* **2023**, *23*, 1990. [CrossRef]
4. Huang, C.; Zhao, J.; Yu, Y.; Zhang, H. Comprehensive sample augmentation by fully considering SSS imaging mechanism and environment for shipwreck detection under zero real samples. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5906814. [CrossRef]
5. Long, H.; Shen, L.; Wang, Z.; Chen, J. Underwater Forward-Looking Sonar Images Target Detection via Speckle Reduction and Scene Prior. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
6. Li, C.; Ye, X.; Xi, J.; Jia, Y. A Texture Feature Removal Network for Sonar Image Classification and Detection. *Remote Sens.* **2023**, *15*, 616. [CrossRef]
7. Neupane, D.; Seok, J. A review on deep learning-based approaches for automatic sonar target recognition. *Electronics* **2020**, *9*, 1972. [CrossRef]
8. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A Systematic Review and Analysis of Deep Learning-based Underwater Object Detection. *Neurocomputing* **2023**, *527*, 204–232. [CrossRef]
9. 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]
10. Ma, Q.; Jiang, L.; Yu, W.; Jin, R.; Wu, Z.; Xu, F. Training with noise adversarial network: A generalization method for object detection on sonar image. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; WACV: Bentley, Australia, 2020; pp. 729–738.
11. Xi, J.; Ye, X.; Li, C. Sonar Image Target Detection Based on Style Transfer Learning and Random Shape of Noise under Zero Shot Target. *Remote Sens.* **2022**, *14*, 6260. [CrossRef]
12. Lee, S.; Park, B.; Kim, A. Deep learning based object detection via style-transferred underwater sonar images. *IFAC-Pap.* **2019**, *52*, 152–155. [CrossRef]
13. Greene, A.; Rahman, A.F.; Kline, R.; Rahman, M.S. Side scan sonar: A cost-efficient alternative method for measuring seagrass cover in shallow environments. *Estuar. Coast. Shelf Sci.* **2018**, *207*, 250–258. [CrossRef]
14. Li, C.; Ye, X.; Cao, D.; Hou, J.; Yang, H. Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples. *Appl. Acoust.* **2021**, *173*, 107691. [CrossRef]
15. Gerg, I.D.; Monga, V. Deep Multi-Look Sequence Processing for Synthetic Aperture Sonar Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
16. Ard, W.; Barbalata, C. Sonar Image Composition for Semantic Segmentation Using Machine Learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; WACV: Bentley, Australia, 2023; pp. 248–254.
17. Zhao, Y.; Guo, P.; Sun, Z.; Chen, X.; Gao, H. Residualgan: Resize-residual dualgan for cross-domain remote sensing images semantic segmentation. *Remote Sens.* **2023**, *15*, 1428. [CrossRef]
18. Yan, M.; Kezierbieke, G. The research review of image semantic segmentation method in high-resolution remote sensing image interpretation. *J. Front. Comput. Sci. Technol.* **2023**. [CrossRef]
19. Wang, J.; Chen, X.; Jiang, W.; Hua, L.; Liu, J.; Sui, H. PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *119*, 103309. [CrossRef]
20. Ochal, M.; Vazquez, J.; Petillot, Y.; Wang, S. *A Comparison of Few-Shot Learning Methods for Underwater Optical and Sonar Image Classification*; Global Oceans 2020, Singapore–US Gulf Coast; IEEE: Piscataway, NJ, USA, 2020; pp. 1–10.
21. Song, Y.; He, B.; Liu, P. Real-time object detection for AUVs using self-cascaded convolutional neural networks. *IEEE J. Ocean. Eng.* **2019**, *46*, 56–67. [CrossRef]

22. Xu, F.; Huang, J.; Wu, J.; Jiang, L. Active Mask-Box Scoring R-CNN for Sonar Image Instance Segmentation. *Electronics* **2022**, *11*, 2048. [CrossRef]
23. Chen, D.; Yuan, L.; Liao, J.; Yu, N.; Hua, G. Stylebank: An explicit representation for neural image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–27 July 2017; pp. 1897–1906.
24. Song, Y.; Liu, P. Segmentation of sonar images with intensity inhomogeneity based on improved MRF. *Appl. Acoust.* **2020**, *158*, 107051. [CrossRef]
25. Huo, G.; Wu, Z.; Li, J. Underwater Object Classification in Sidescan Sonar Images Using Deep Transfer Learning and Semisynthetic Training Data. *IEEE Access* **2020**, *8*, 47407–47418. [CrossRef]
26. Kolev, N. (Ed.) *Sonar Systems*; BoD–Books on Demand: Norderstedt, Germany, 2011.
27. Aykin, M.D.; Negahdaripour, S. Forward-look 2-D sonar image formation and 3-D reconstruction. In *2013 OCEANS-San Diego*; IEEE: Piscataway, NJ, USA, 2013; pp. 1–10.
28. Ragheb, H.; Hancock, E.R. Surface radiance correction for shape from shading. *Pattern Recognit.* **2005**, *38*, 1574–1595. [CrossRef]
29. Coiras, E.; Petillot, Y.; Lane, D.M. Multiresolution 3-D reconstruction from side-scan sonar images. *IEEE Trans. Image Process.* **2007**, *16*, 382–390. [CrossRef]
30. Xiang, Y.; Mottaghi, R.; Savarese, S. Beyond pascal: A benchmark for 3d object detection in the wild. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, 24–26 March 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 75–82.
31. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
32. Liu, Y.; Yu, J.; Han, Y. Understanding the effective receptive field in semantic image segmentation. *Multimed. Tools Appl.* **2018**, *77*, 22159–22171. [CrossRef]
33. Annen, T.; Mertens, T.; Seidel, H.P.; Flerackers, E.; Kautz, J. Exponential shadow maps. In *Graphics Interface*; ACM Press: New York, NY, USA, 2008; pp. 155–161.
34. Wann Jensen, H.; Marschner, S.R.; Levoy, M.; Hanrahan, P. A practical model for subsurface light transport. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*; ACM: New York, NY, USA, 2023; pp. 319–326.
35. Lokovic, T.; Veach, E. Deep shadow maps. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*; ACM: New York, NY, USA, 2023; pp. 311–318.
36. Available online: https://www.shipwreckworld.com/articles/gallery (accessed on 5 February 2024).
37. Beitzel, S.M.; Jensen, E.C.; Frieder, O. Map. In *Encycl. Database Systems*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1691–1692.
38. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
39. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
40. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
41. Güler, R.A.; Neverova, N.; Kokkinos, I. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 7297–7306.

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.