SSS Underwater Target Image Samples Augmentation Based on the Cross-Domain Mapping Relationship of Images of the Same Physical Object

Yulin Tang, *Member, IEEE*, Liming Wang, Shaofeng Bian, Shaohua Jin, *Member, IEEE*, Yuting Dong, *Member, IEEE*, Houpu Li, and Bing Ji

***Abstract*—Side-scan sonar (SSS) image sample augmentation plays an important role in improving the effect of deep-learning-based underwater target detection. However, the existing sample augmentation methods for cross-domain conversion always result in weak representativeness of the augmented samples since the targets in the nondomain images are similar but not exactly the same as the actual underwater target to be detected. In this article, an augmentation method for SSS image samples of underwater targets based on the cross-domain mapping relationship of images of the same object is proposed. A physical model of the actual underwater target was first constructed using three-dimensional printing. A series of optical images and SSS images of underwater targets can be obtained by using the actual measurement of underwater targets under different conditions. To achieve the augmentation of SSS target samples, a single-cycle-consistency network structure with a channel and spatial attention and generative adversarial networks with least squares loss was designed for efficient and robust conversion of information between optical and SSS acoustic samples. To verify the effectiveness of the proposed method in generating high-quality samples, underwater targets were detected using the detection model trained by the generated samples. The experimental results revealed that the proposed method achieved impressive performance with a more than 5.8% improvement in average precision value for zero-sample underwater mine target detection and 4.3% for few-sample shipwreck target detection, compared with using only real SSS data.**

***Index Terms*—Generative adversarial network (GaN), mapping relationship, physical target model, sample augmentation, side-scan sonar (SSS), underwater target detection.**

Manuscript received 7 May 2023; revised 13 June 2023 and 27 June 2023; accepted 29 June 2023. Date of publication 5 July 2023; date of current version 19 July 2023. This work was supported in part by the National Science Foundation for Outstanding Youth under Grant 42122025 and in part by the National Natural Science Foundation of China under Grant 41876103, Grant 42176186, Grant 41974005, Grant 41971416, and Grant 42074074. *(Corresponding author: Shaofeng Bian.)*

Yulin Tang, Liming Wang, Houpu Li, and Bing Ji are with the College of Electrical Engineering, Naval University of Engineering, Wuhan 430033, China (e-mail: 21000601@nue.edu.cn; icesoar@163.com; lihoupu1985@126.com; jibing1978@126.com).

Shaofeng Bian and Yuting Dong are with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: sfbian@sina.com; dongyt@cug.edu.cn).

Shaohua Jin is with the Department of Oceanography and Hydrography, Dalian Naval Academy, Dalian 116018, China (e-mail: jsh_1978@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3292327

I. Introduction

UNDERWATER target detection plays an essential role in fields, such as navigation safety, marine investigation, maritime search and rescue, and military tasks [1]. At present, most underwater target detection methods are based on acoustic detection, magnetic detection, optical detection, and electric detection. Among them, acoustic detection is the most common underwater target detection method because of the advantages of sound waves with respect to underwater imaging conditions, propagation distance, and range [2], [3]. Side-scan sonar (SSS) is widely used in underwater target detection because it has a wider sweep amplitude and higher imaging resolution than other acoustic devices, in addition to the small size and low price of the SSS equipment [4], [5].

Current underwater target detection based on SSS images mostly depends on manual visual interpretation, which has the disadvantages of being inefficient, slow, and highly subjective [6]. Therefore, it is necessary to study automatic detection methods for underwater targets using SSS images [7], [8].

Some researchers have performed the automatic detection of underwater targets using machine learning methods that combine handcrafted features with a classifier and achieved good detection results under certain conditions [9], [10]. However, because of the influence of the complex seabed environment and measurement conditions, SSS images usually have the features of low resolution, poor characteristics, complex noise, and severe distortion, which limit the detection accuracy of the traditional machine learning methods [11], [12], [13]. In recent years, the performance of target detection methods based on deep learning has far exceeded that of the traditional machine learning methods, which has attracted widespread attention in the field of underwater exploration [14], [15], [16]. However, a target detection model based on the deep convolutional neural networks (DCNNs) requires a large number of training samples, and highly representative samples are the key to achieving high-performance detection [17]. Moreover, because of the high cost and low speed of data collection as well as a lack of targets, the number of available SSS images is highly insufficient, and the representation in samples is not sufficient [18]. Hence, it is important to augment the number of underwater target SSS images using a small sample.

Inspired by the sample augmentation techniques for optical images, most methods for augmenting SSS underwater target samples have been developed using transfer learning [19], [20], [21]. For example, Lee et al. [22] carried out style transfer using StyleBankNet to generate simulation data with an SSS image style for the detection of drowning people. Li et al. [18] obtained simulated Sonar images from optical targets without the use of any target sonar image using a transfer-learning method for style transfer. Huo et al. [23] increased the size of a sample by fusing the target samples generated from optical images with the existing background images. Huang et al. [24] proposed an SSS image augmentation method for shipwreck targets by taking into account the complete SSS imaging mechanism and environmental impacts of target diversity, target texture, imaging resolution, device environmental noise, and background, realizing a detection accuracy of 95% for SSS shipwreck targets.

The above methods demonstrate the potential of underwater target SSS image generation based on style transfer, but the cross-domain optical-sonar image pairs input into the network during the training of the migration conversion were not derived from the same target. In addition to learning the conversion relationship between optical images and sonar images, a model should also learn to eliminate the interference of system errors caused by different targets, which increases the computation of the network without improving performance, and cannot fully consider the influence of seven factors, i.e., the acoustic transmitting unit, acoustic propagation medium, acoustic reflection target, acoustic reflection background field, acoustic receiving unit, noise, and data postprocessing [25], [26]. Moreover, the generated samples are weak in terms of representativeness, leading to limited improvement in the generalization ability and accuracy of a DCNN-based target detection network.

As the most commonly used technique for style transfer of nondomain images, generative adversarial networks (GANs) [27], [28], [29], [30] has received extensive attention. Li et al. [31] transformed underwater, small-target optical images into synthetic aperture sonar (SAS) images using an improved CycleGAN (CG) based on cyclic consistency. Chen and Summers [32] used GAN-based networks in unsupervised feature learning for SAS image seabed classification to generate realistic SAS images of different seafloor bottom types. Karjalainen et al. [33] proposed a GAN-based approach to add simulated contacts into real SSS images, generating images that even experts could not distinguish as generated or real. Reed et al. [34] proposed a method that couples an optical renderer with GAN to synthesize realistic SAS images of the subsea targets, achieving high levels of SAS image realism while retaining control over image geometry and parameters. Jiang et al. [35] proposed a GAN-based semantic image synthesis model that can generate cost-effective high-quality SSS images efficiently. In general, the above methods demonstrate the advantages of style transfer based on GANs in SSS image sample augmentation, and GAN is the preferred approach for the method proposed in this article. However, the data used in the above methods were all cross-domain data, and the targets were not generated from the same physical object. Therefore, there are noises generated by systematic errors, and the generated SSS images still need improvement in terms of their realism.

In summary, due to the high cost of data acquisition, slow speed, and lack of targets, SSS underwater target images are scarce, and there is an urgent need to explore sample augmentation techniques for SSS images. The main limitation of the traditional SSS underwater target image sample augmentation methods is that the conversion models are built based on nonidentical entities, which will introduce systematic biases in the models, resulting in poor sample augmentation performance and ultimately affecting the performance of the target detection models.

Aiming at these limitations, this article proposes an augmentation method for underwater target SSS image samples based on the cross-domain mapping relationship of the same object. A physical model of the underwater target was first created using three-dimensional (3-D) printing technology [36] and the target itself. According to the mechanisms of underwater target imaging, the optical images of the physical model in different imaging modes were obtained using an optical camera at multiple viewing angles, heights, and distances, and the mapping relationship of the object-to-optical images was established. The SSS images of the underwater physical model were obtained through sea-based experiments, and the mapping relationship of the object-to-SSS image was established. Therefore, the mapping relationship from optical images to SSS images of the same object was established using the designed GAN, and the style conversion model using a cross-domain mapping of the same object was obtained. Finally, the conversion model was used to convert other optical images of the target into SSS images to achieve high-quality sample augmentation and solve the limitation of image generation quality caused by the cross-domain mapping relationship of different entities. This process provides a high-quality dataset for improving the performance of target detection models based on deep learning. The main contributions of this article are as follows.

1) An augmentation method for underwater targets SSS image samples based on the cross-domain mapping relationship of the same object is proposed. Using 3-D printing technology, a physical model of underwater targets with a small-sample size or even no samples was made, and the real mapping relationship between optical and SSS images of the target is established, which solves the problem that the cross-domain mapping relationships of nonidentical target objects restricts the performance of sample augmentation.
2) The conversion model of GAN based on cyclic consistency is given. A single-cycle consistent network structure was designed to ensure the training efficiency of the mode, and a channel and spatial attention (CSA) module was integrated into the generator to reduce information diffusion while enhancing cross-dimensional interaction and improving the quality of generated images. A loss function based on least squares generative adversarial network (LSGAN) was designed to improve the training stability and avoid mode collapse, thus enabling the high-quality-domain conversion between optical and SSS images and

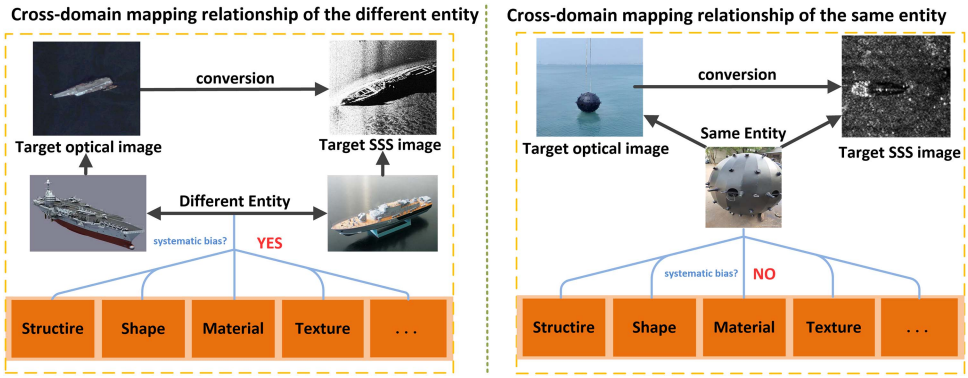

Fig. 1. Comparison of a mapping relationship based on images of the same object and a mapping relationship based on images of different objects.

achieving the augmentation of underwater target SSS images with a small-sample size.

3) The method is verified by two experiments, and the cross-domain mapping relationship of the optical and SSS images of the sample-free underwater targets is established, and the high-quality conversion of images between two domains is realized. Compared with the images of the traditional sample augmentation method, the generated SSS underwater target images have high definition, complete details, and strong realism. This improves the detection performance of underwater target detection models based on deep learning to demonstrate the effectiveness of the method and provides a new approach to the highly representative augmentation of target samples and the construction of high-performance underwater target detection model with a small sample or no sample.

II. Methods

To augment a sample of underwater target images, it is necessary to consider the unique features of the target and the characteristics of an SSS image to enhance the authenticity of the generated samples. To obtain the unique features of the target, previous methods referred to the optical images of targets with similar features, including texture, shape, intensity, background, and noise, and the targets, were of the same category but were not the same target because it is almost impossible for the targets in the nondomain dataset to be the same target objects as the actual target object in the real sonar image. As shown in the left part of Fig. 1, the ships shown in the optical and SSS images are not the same ship. They are different in shape, material, texture, internal structure, and other aspects, and a forced conversion is bound to produce systematic deviations. In addition to learning the conversion relationship between optics and acoustics, the conversion network must also eliminate the interference of systematic errors caused by different entities, which results in a low level of authenticity of the generated image. In this study, an optical image of the target was obtained according to the imaging methods and mechanisms of an underwater target by creating a single target object, and the actual mapping relationship between the object and its optical image was established. According to the actual measurement mode of SSS, the target SSS image was obtained through sea trials to establish the real mapping relationship between the object and its SSS image. This avoids the systematic error between different entities to establish the real cross-domain mapping relationship of optical images to SSS images and obtain the real correspondence and interaction between target features and image features involved in the process of converting images of the target from optical images to SSS images to generate higher quality images.

A. Sample Augmentation Based on the Cross-Domain Mapping Relationship of the Same Object

A large number of strongly representative data samples are needed to establish high-performance deep-learning models. However, for SSS underwater target data, lack of samples, low resolution, poor and sparse features, and complex noise are the key factors restricting the establishment of high-performance SSS underwater target deep-learning models. Although the method of transfer learning through multidomain images and sample augmentation taking into account the SSS imaging mechanism exists, the target characteristics and marine environment can be used to augment the sample of SSS images, the essence of which is based on semiexperience and semimodeling simulation. Moreover, there is a strong correlation between the noise, texture, and resolution of the background and the target, which is difficult to quantify using mathematical expressions. In addition, these experiences are not necessarily applicable to SSS images of underwater targets that have a small-sample size or no sample. Therefore, an augmentation method for small samples of underwater target images based on the cross-domain mapping relationship of the same physical object is proposed, and the process flow is shown in Fig. 2.

1) A physical model of the underwater target is made using 3-D printing technology, which provides a physical basis for the subsequent optical and SSS image acquisition of the same target.
2) A series of optical images of the physical model at different angles, heights, and positions are obtained using unmanned aerial vehicles (UAVs) and high-altitude

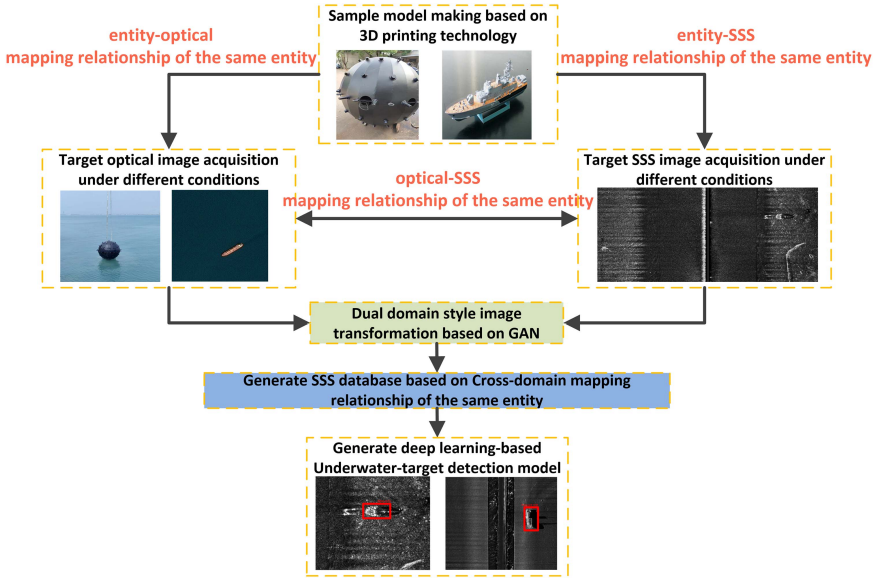

Fig. 2. Process of sample augmentation technology of SSS underwater targets based on the cross-domain mapping relationship of the same physical object.

imaging with an optical camera that follows the methods and mechanisms of underwater target imaging, and the object-to-optical image mapping relationship is established.

3) The SSS images of the physical model are obtained using SSS according to the actual SSS measurement method, and the mapping relationship of the object-to-SSS image is established, which provides a dataset for training the GAN.
4) The GAN is trained using the optical and SSS images obtained from the same physical object, and the cross-domain real mapping relationship is extracted to form a cross-domain conversion model to complete the dual-domain conversion of optical–SSS images, thereby realizing the sample augmentation of underwater target SSS images and establishing a high-quality and representative underwater target SSS image database.
5) A deep-learning-based underwater target detection model is trained by using the augmented underwater target SSS image sample as the training set, and the real underwater target SSS images are used to qualitatively evaluate the performance of the detection model.

When acquiring underwater target SSS images, theoretically, the richer the SSS images under different conditions, the better the performance. In practice, because of the difficulty and cost of offshore operations, the difficulty and cost of image acquisition were also high, and only a few real samples were obtained in this experiment. However, this is in line with the theme of this article, namely solving the problem of augmenting SSS image samples that are small or do not exist.

B. Three-Dimensional Printing Technology

The core idea of this study is to establish the cross-domain true mapping relationship between the optical and acoustic images of the same underwater target for targets with a small-sample size or no samples, and 3-D printing technology is an important part of this idea. For targets with few or even no underwater samples, such as mines, 3-D printing was used to physically manufacture the objects, which enables the conditions needed to acquire optical and acoustic images of the same target. The principles, production processes, and precautions of the 3-D printing used in this article are briefly reviewed in this section.

The 3-D printing is based on light-curing molding technology, also known as stereolithography (SLA), which is based on the principle of photopolymerization and uses laser beams to transform liquid photosensitive resin into a solid state. The scanning and curing are repeated layer-by-layer until the entire part is manufactured to obtain a 3-D physical model [36]. The specific process of 3-D printing is shown as follows.

First, the style, material, force, structure, and other elements of the existing underwater target are analyzed, and the attributes of the physical model are made to be as consistent as possible with those of the real target to provide a basis for obtaining the SSS images of the underwater target with the same characteristics.

Second, the 3-D modeling software (3-Ds MAX) is used to design a digital model of the target and create engineering drawings based on the density of printing materials and the production characteristics of 3-D printing.

Then, 3-D slicing software (SOLIDWORKS) is used to refine the model data, which includes mold modification, shelling, slicing, and support, and generate the STL files.

Finally, the generated file is imported into an industrial-grade 3-D printer for SLA-based 3-D printing using additive manufacturing to obtain a physical underwater target model.

Because the 3-D-printed physical model of the underwater target is subsequently used to acquire the target optical and SSS images, in addition to the basic elements, such as the shape, size, and texture of the underwater target itself, it is also necessary to consider factors in the manufacturing process that increase the realism of the underwater target during optical imaging and actual sea deployment. These include the following.

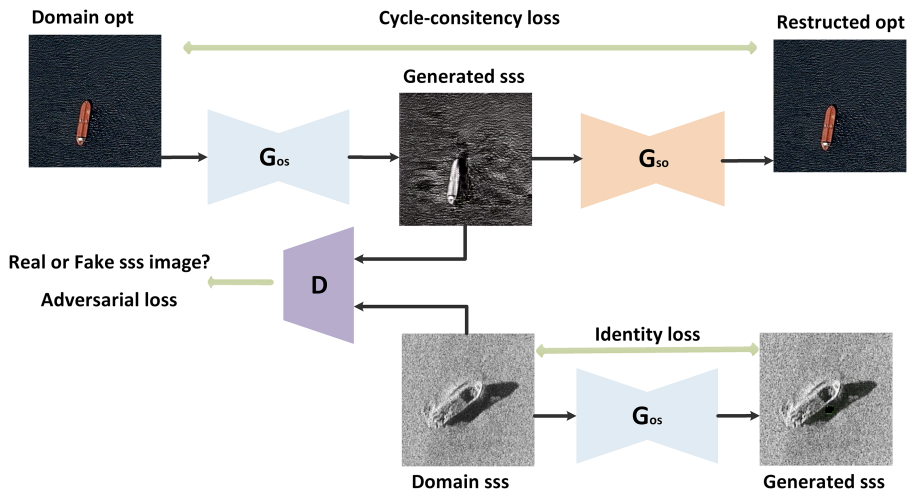

Fig. 3. Structure of the proposed GAN.

1) Background: When taking optical images, the underlying structure of SSS images should be imitated as much as possible, that is, overhead images should be taken, and a single uncluttered area should be selected for the background to avoid the impact of clutter, such as buildings, trees, or crowds, on the quality of the subsequent image-domain conversion.
2) Density: It is best to choose a rigid metal material that meets the material properties of the actual target and is not prone to corrosion by seawater. The most important factor is to ensure that the target is not swept away by the current under complex sea conditions.
3) Structure: The internal structure needs to be supported well. Holes are reserved in the structure so that the model can sink smoothly after it has been fully filled with water while saving production costs. To ensure that the shape matches that of the real target, it is necessary to reserve a ring. One purpose of this ring is to facilitate the attachment of a float to ensure that it is easy to find and salvage after the target has sunk to the seabed. The other purpose is for safety: attaching an anchor chain under the target can effectively keep it from being caught by nearby fishermen.

C. Proposed GAN Network

The use of GAN to achieve domain conversion of underwater targets is an important part of the proposed method, and the use of networks is the key to achieving high-quality sample augmentation. Because of the large difference in the styles of optical and SSS images, the traditional GAN outputs poor quality results so much so that the images from the two domains do not have any relationship. To create a strong correlation between the generated and output images, GAN based on single-cycle consistency is used that focuses on the conversion task from the optical domain to the acoustic domain while ensuring training efficiency. In addition, the CSA module is integrated in the generator, enhancing cross-dimensional interaction, which reduces information diffusion. Finally, a combined loss function based on LSGAN was designed to improve the quality of generated images and the stability of model training. The architecture of the proposed network is shown in Fig. 3.

1) Network Structure: The main body of the model proposed in this article consists of two generators (G_{os} and G_{so}) and a discriminator (D). The generator that converts an image from domain O to an image in domain S is called G_{os}, the generator that converts an image from domain S to domain O is called G_{so}, and the discriminator that discriminates images belonging to domain S from "fake" images is called D.

After the SSS image (S) has been obtained from the original input image, optical image (O), by the generator, the generated image S is used as input to the generator to obtain the image reconstructed optical image (RO), which should be the same as image O. Images O and RO will eventually become consistent, allowing the image to loop through a cycle and back to the starting point without change.

The generator structure is shown as follows. First, the features of the input image are extracted using three convolutional layers, and after each convolutional layer, the instance normalization (IN) operation is used along with the activation function. Second, the attention mechanism, the CSA module, is used to conduct global learning of the image channels and spatial features to establish the interaction among local detail features and global features, and realize multiscale feature fusion through skip connections. Six residual networks are then used to retain the input data features while further extracting image information. Next, two transpose convolutions are used for sample loading. Finally, the data pass through a convolutional layer, and the obtained image matrix is activated by the activation function to obtain the final output image.

The discriminator uses five convolutional layers to extract features from the input image, and after each convolutional layer, the IN operation is also used as the activation function, where the last convolutional layer directly returns the linear operation result.

2) CSA Module: The key to generating high-quality images is to fully learn the target's detail features and the background

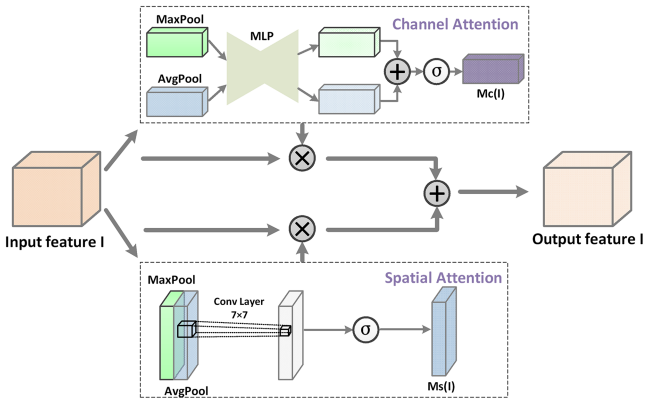

Fig. 4. Structure of the CSA module.

features of SSS images. To better learn the global information and local features of the input image as well as enhance the interaction between channels and space, a CSA module was designed in this study to improve the network performance by reducing information diffusion while augmenting global cross-layer interactions. The CSA module consists of channel attention and spatial attention, and the architecture is shown in Fig. 4.

In the proposed method, parallel calculation was used for CSA modules. First, the weight coefficients of the respective outputs are multiplied elementwise with the initial input feature map I ($H \times W \times C$), and then the result is summed, element-by-element, to increase the cross-dimensional receptor domain while improving efficiency. H, W, and C are the length, width, and channel number of the feature I

$$I_{\text{output}} = (M_c(I) \otimes I) + (M_s(I) \otimes I). \tag{1}$$

Here, $\otimes$ represents the elementwise multiplication, and *Mc(I)* and *Ms(I)* are the feature outputs for the channel attention and spatial attention, respectively. A detailed description of each module is given in the following sections.

a) Channel attention: Channel attention emphasizes to which features the model should pay attention. Each of its channels contains a specific characteristic response. First, the input feature I is passed through a global max pooling layer and global average pooling layer based on the width and height, respectively, to obtain two $1\times1\times C$ images. Second, they are fed into a two-layer multilayer perceptron (MLP), which is a shared two-layer neural network, where the number of neurons in the first layer is C/r (here, r is the reduction rate) and the activation function is ReLU. In addition, the number of neurons in the second layer is C. Then, the MLPs output features are summed elementwise and activated by a sigmoid function to generate the final channel attention feature, i.e., M_c. Finally, M_c and feature I are multiplied elementwise to obtain the new feature after scaling

$$\begin{aligned} M_c(I) &= \sigma(\text{MLP}(\text{AvgPool}(I)) + \text{MLP}(\text{MaxPool}(I))) \\ &= \sigma(W_1(W_0(I^C_{\text{Avg}})) + W_1(W_0(I^C_{\text{Max}}))). \end{aligned} \tag{2}$$

Here, σ is the sigmoid activation function, and $W_0 \in R^{C/r\times C}$ and $W_1 \in R^{C\times C/r}$ are the weights of the MLP shared network.

b) Spatial attention: Spatial attention emphasizes the features the model focuses on based on the location, namely, enhancing or suppressing features at different spatial locations.

First, the input feature I is passed through a channel-based global max pooling layer and global average pooling layer to obtain two $H\times W\times1$ feature maps. Then, the channel-based concatenation operation is performed. Second, the image is reduced to one channel by convolution with a 7×7 convolution kernel. Then, a spatial attention feature, namely *Ms*, is generated via the sigmoid activation function. Finally, *Ms* and feature I are multiplied elementwise to obtain the final feature

$$\begin{aligned} M_s(I) &= \sigma(f^{7\times7}([\text{AvgPool}(I), \text{MaxPool}(I)])) \\ &= \sigma(f^{7\times7}([I^S_{\text{Avg}}, I^S_{\text{Max}}])). \end{aligned} \tag{3}$$

Here, σ is the sigmoid activation function, and $f^{7\times7}$ represents the 7×7 convolution operation of the convolution kernel.

3) Loss Function: The appropriate loss functions play a crucial role in improving the quality of images generated by the GAN. The proposed network architecture in Fig. 3 reveals that the loss function is composed of LSGAN loss, cyclic-consistency loss, and identity loss. LSGAN loss directs the generator to generate a more realistic image of the target domain; cyclic-consistency loss directs the generator to generate images as similar as possible to the input images; and identity loss limits the generator so that it ignores the input data.

a) LSGAN loss: The traditional GAN uses cross entropy as the loss function, which does not optimize the images that are judged to be real by the discriminator, even if these images are still far from the decision boundary of the discriminator. This will result in low-quality images generated by the generator and unstable model training. To address this, the objective function in LSGAN is used as the loss function of the model, namely, the least squares loss is used as the loss function

$$\begin{aligned} \min_D V_{\text{LSGAN}}(D) = &\frac{1}{2}E_{s\sim P_{\text{data}}(s)}[(D(s)-b)^2] \\ &+ \frac{1}{2}E_{z\sim P_z(z)}[(D(G(z))-a)^2] \end{aligned} \tag{4}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}E_{z\sim P_z(z)}[(D(G(z))-c)^2]. \tag{5}$$

In the objective function of discriminator D, the real data and generated data are assigned labels b and a, where $b = 1$ indicates the real data, and $a = 0$ indicates the generated data. The discriminator is optimized by minimizing the error between the data generated by the discriminator and 0 as well as the error between the real data s and 1. In the objective function of generator *Gos*, the generated data are assigned label c, and by minimizing the error between the data z generated by the generator and 1, the generator is trained to successfully trick the discriminator to obtain a high score for $c = 1$. Therefore, (4) and (5) above can be converted into the following two equations:

$$\begin{aligned} \min_D V_{\text{LSGAN}}(D) = &\frac{1}{2}E_{s\sim P_{\text{data}}(s)}[(D(s)-1)^2] \\ &+ \frac{1}{2}E_{z\sim P_z(z)}[(D(G(z)))^2] \end{aligned} \tag{6}$$

$$\min_G V_{\text{LSGAN}}(G_{os}) = \frac{1}{2}E_{z\sim P_z(z)}[(D_{os}(G(z))-1)^2]. \quad (7)$$

b) Cyclic-consistency loss: To achieve cyclic consistency, the following criterion should be met during conversion from domain O to domain S:

$$x \Rightarrow G_{os}(x) \Rightarrow G_{so}(G_{os}(x)) \approx x. \quad (8)$$

The mathematical formula is expressed as follows:

$$L_{\text{cyc}}(G_{os}, G_{so}) = E_{x\sim P_{\text{data}}(x)}[\|G_{so}(G_{os}(x)) - x\|_1]. \quad (9)$$

Here, the one norm is the matrix one norm, which is the value that is the largest sum of the absolute values of the elements in the column vectors of all matrices, as follows:

$$X_1 = \max_j \sum_{i=1}^{m} |a_{i,j}|. \quad (10)$$

c) Identity loss: Identity loss is used to limit the situation where the generator modifies the image pixels independently regardless of the input data, which means that if a domain *S* image is fed into generator G_{os}, then it should return to its original form as much as possible. The loss function is expressed as follows:

$$L_{\text{Identity}}(G_{os}) = E_{s\sim P_{\text{data}}(s)}[\|(G_{os}(s) - s)\|_1]. \quad (11)$$

Therefore, the total loss function of the GAN is as follows:

$$\text{Loss}_{\text{cyc}} = \min_D V_{\text{LSGAN}}(D) + \min_G V_{\text{LSGAN}}(G_{os}) + \lambda_1 L_{\text{cyc}}(G_{os}, G_{so}) + \lambda_2 L_{\text{Identity}}(G_{os}). \quad (12)$$

Here, λ_1 and λ_2 are nonnegative hyperparameters to adjust the influences of the losses on the overall results. We weight each loss to balance the importance of each component.

III. Experiments

To evaluate the feasibility and effectiveness of the proposed method, the experiments described in this section consisted of two main parts. The first set of experiments took no-sample mine targets as the experimental objects to analyze and evaluate the feasibility of the SSS image sample augmentation method based on the real mapping relationship proposed in this article. The second set of experiments were based on the GAN model proposed in this article to analyze and verify its effectiveness for the sample augmentation of few-sample SSS images.

A. Experiments on the Sample Augmentation Strategies

To evaluate the effectiveness of underwater target SSS image augmentation based on the cross-domain mapping relationship of the same physical object, a mine target with no-sample images was selected as the research object. After the optical and SSS images were acquired using the 3-D-printed physical model of the mine, the quality of the generated images of different small-sample datasets was qualitatively and quantitatively analyzed to evaluate the effect of the proposed method on the sample augmentation strategy for small-sample datasets. The effects of the augmented SSS image datasets on the detection performance of the YOLOv5 target detection model were analyzed and evaluated.

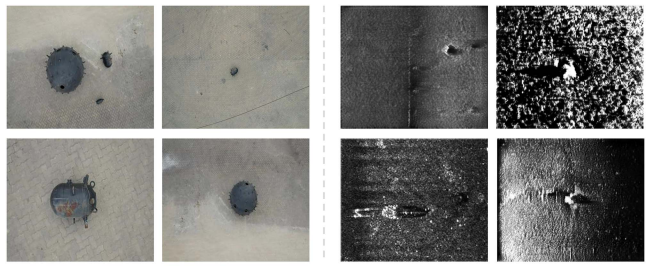

Fig. 5. Example images of several mine targets. (a) Optical images. (b) SSS images.

1) Datasets: The experimental dataset consisted of optical images and SSS images obtained from the physical model of the mine. The specific preparation process was as follows: a physical model of the mine target was made using 3-D printing. Next, optical images of the mine target were obtained by aerial photography using the DJI Air 2S UAV at different altitudes and angles. A total of 150 images were taken, and the 100 images with higher quality were selected. In the waters of Haitang Bay, Sanya, China, an autonomous underwater vehicle (AUV) equipped with a Shark-S455D SSS was used to scan the mine target and obtain a total of 60 SSS images. The 50 images with the best screening quality were then selected. Examples of optical and SSS images of some mine targets are shown in Fig. 5.

2) Evaluation Indicators: According to Xu et al. [37], the Fréchet inception distance (FID), kernel maximum mean discrepancy (MMD), and the 1-nearest neighbor (1-NN) are better than other indicators for evaluating the clarity, diversity of features, and authenticity of the synthetic samples.

The FID is a measure that calculates the distance between the real image and the feature vector of the generated image and is used to measure the similarity of two sets of images. The FID is calculated as follows:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\sum r + \sum g - 2\sqrt{\sum r \sum g}\right). \quad (13)$$

Here, μ_r and μ_g are the mean vectors of the two distributions, respectively, Σr and Σg are their covariance matrices, $\|$ denotes the norm of a vector, and Tr is the trace of a matrix. The lower the FID values, the better the image augmentation results.

The MMD measures the similarity between two feature distributions based on a statistical test of the maximum mean squared difference, mapping the real and generated sets to a kernel space with a fixed kernel function, and then computing the mean difference between the two distributions. The MMD is calculated as follows:

$$\text{MMD}^2(X, Y) = E[K(X_i, X_j) - 2K(X_i, Y_j) + K(Y_i, Y_j)]. \quad (14)$$

Here, *X* represents the set of real images, X_i and X_j are the samples drawn from *X*, *Y* represents the set of generated images, and Y_i and Y_j are the samples drawn from *Y*. *E* stands for

expectation, and K is the Gaussian kernel. Lower MMD values indicate more effective image augmentation.

The 1-NN uses a binary classifier to calculate the similarity between two image sets by mixing n real sets (labeled 1) with n generated sets (labeled 0) and then randomly dividing them into a training set *T1* (number is $2n-1$) and test set *T2* (number is 1), using *T1* to train the classifier and *T2* to obtain the accuracy of classification. The above steps are cycled $2n$ times, each time selecting a different *T2*, and finally calculating the average classification accuracy. The closer this accuracy is to 0.5, the better.

Meanwhile, considering that the purpose of this study was to augment small SSS underwater target image datasets to improve the performance of deep-learning-based target detection models, a comparative experiment was carried out using a deep-learning-based target detection model. At present, many suitable target detection models exist. The YOLOv5 model, which has the features of high speed, a lightweight architecture, and easy deployment, was chosen for the evaluation experiments. The images generated by our GAN model were used as the training set and fed to the YOLOv5 network. Then, the real images were used as the validation set to evaluate the effectiveness of the generated images in training the detection network using Precision (P), Recall (R), and Average Precision (AP). P measures the proportion of correctly detected objects out of all the objects predicted by the model; R measures the proportion of correctly detected objects out of all the actual objects present in the dataset

$$P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$$
$$R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}). \tag{15}$$

True Positives (TP) represent the number of correctly detected objects, False Positives (FP) represent the number of objects incorrectly identified as positives, and False Negatives (FN) represent the number of objects that was not detected by the model.

AP evaluates the tradeoff between P and R at different thresholds. It is the area under the precision–recall curve, and it provides an overall performance measure for the model

$$\mathrm{AP} = \int_0^1 P(R)dR. \tag{16}$$

3) Design of the Experiments: All model training was implemented based on the PyTorch framework, and two NVIDIA GeForce RTX 3090s with a parallel memory of 48 GB was used.

To verify the performance of the proposed augmentation method for small-sample datasets, 50 real SSS mine images were divided into a training set and evaluation set using a ratio of 3:2. Of these, the 30 images of the training set were divided into three groups of 10, 20, and 30 images each. The 100 optical images were divided into a training set and conversion set using a ratio of 3:2. Of these, the 60 images of the training set were divided into three groups of 20, 40, and 60 images each. The evaluation set of SSS images and the conversion set of optical images were used for a quantitative analysis of the quality of the generated images.

To reduce performance fluctuation during training and stabilize the model training, historical data were cached in this experiment during training. A list was used to store the last ten images, one of which was randomly chosen by the discriminator for discrimination in each training iteration, so that the discriminator was able to generate images at any point in time. Table I lists the parameters of model training. In Table I, λ_1 and λ_2 are the parameters of the loss function (12) and β is the parameter for Adam optimization.

4) Experiments and Analysis: Although the original mine SSS images were created from scratch from actual field SSS images and optical images, the number of images was an order of magnitude smaller than datasets published on the Internet. Hence, the conventional augmentation functions, such as rotation, clipping, scaling, left and right shifts, flipping, and noise addition, were adopted before model training to increase the dataset size by ten times for each group of training samples. In the noise addition process, due to the significant influence of speckle noise on the quality of SSS images, this experiment incorporated Rayleigh noise and salt-and-pepper noise with zero mean and standard deviations of 30 and 60, respectively. To evaluate the augmentation performance of the sample with a small number of real samples, the real SSS images were divided into three groups of 10, 20, and 30 images, and the corresponding real optical images were divided into three groups of 20, 40, and 60 images, as listed in Table II.

a) Quantitative analysis: After expanding the various numbers of real samples to ten times, they were input to the GAN for training (see Table II). The 40 images of the mine-optical image conversion set were input to the trained GAN for SSS image generation after $10\times$ data augmentation. For each of the generated 400 SSS images and the 20 real SSS mine images in the evaluation set, the FID, MMD, and 1-NN values were calculated. These three indicators were used to assess clarity, variety, and difference between the generated image and real image. Smaller values of these indicators indicate better performance, where 1-NN values closer to 0.5 indicate better performance. The final quantitative test results are shown in Table III.

As Table III reveals, even when there were only 10, 20, and 30 real SSS images, the SSS images generated using the proposed method in this article achieved good results in terms of FID, MMD, and 1-NN values, which indicates that the generated images have high clarity and realism, and have only a small difference with respect to the real SSS images, proving the effectiveness of the proposed method. However, a comparison of the values reveals that, as the number of real SSS samples increases, the values of all indicators improve, which also shows the importance of sample size, and demonstrates that it is necessary to augment the sample.

b) Qualitative analysis: Fig. 6 shows the conversion of three typical representative mine-optical SSS images of large, high-quality, and small models after training on the three groups of real SSS images with different numbers of images.

By comparing the images in Fig. 6(b)–(d), it can be seen that the model trained on group 1 was able to generate the basic contours, but the image still needs to be enhanced in terms of detailed features, such as texture and shadow. The model

TABLE I
TRAINING PARAMETERS FOR THE MODEL

Item	batch size	crop_size	λ_1 / λ_2	Lr	Lr police	Lr decay iters	epoch	optimizer	β
Parameters	16	256	10/0.5	0.0002	Linear	50	1000	Adam	0.5

TABLE II
DISTRIBUTION OF THE MINE TARGET IMAGE SAMPLES IN DIFFERENT GROUPS

Group	Real SSS	SSS After Augmentation	Real Optical	Optical After Augmentation	Model-Generated Images	Validation Sets
1	10	100	20	200	400	20
2	20	200	40	400	400	20
3	30	300	60	600	400	20

TABLE III
PERFORMANCE OF SSS IMAGES GENERATED BY THE MODEL AFTER TRAINING WITH DIFFERENT GROUPS OF REAL SAMPLES

Group	FID↓	MMD↓	1-NN↓0.5
1	172.12	0.311	0.74
2	165.33	0.284	0.71
3	158.64	0.247	0.70

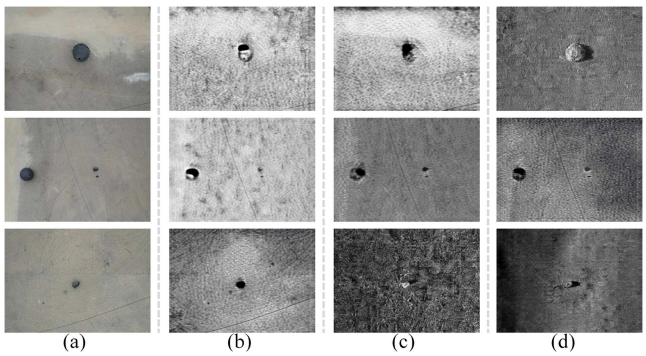

Fig. 6. Results of optical-to-SSS image conversion using three sets of real samples for models trained on different numbers of real samples. (a) Input optical images and converted results of the (b) model trained on group l, (c) model trained on group 2, and (d) model trained on group 3.

trained on group 2 obtained better results than that trained on group 1. Compared with the results of the models trained on the first two groups, the model trained on group 3 generated target images that is significantly more realistic and clear in terms of texture, shadow, and background, indicating that an increase in the number of real samples does indeed improve the quality of the generated sample images. In general, the models trained on real SSS mine images with different small samples basically realized the cross-domain conversion from optical to acoustic images of mine targets, which proves that the proposed method can realize the generation of underwater target samples for targets with no-sample images.

c) Performance of the target detection model: The purpose of this study was to augment small SSS underwater target image datasets to improve the performance of deep-learning-based target detection models. Hence, a deep-learning-based target detection model was used to carry out the comparative experiment. At present, there are many target detection models. Given the purpose of this study mentioned above, the YOLOv5 model, which has the features of high speed, a lightweight architecture, and easy deployment, was used for the experiment.

TABLE IV
COMPOSITION OF TRAINING AND VERIFICATION DATASETS

Group	Real/data-enhanced mine image	Generated mine image
A	30/300	–
B	–	300
C	30/300	300
verify	20	–

Using mine targets as objects, three datasets were designed to train the YOLOv5 model, as listed in Table IV. To ensure the model training performance and improve training efficiency, the ratio of the training set to the test set was set to 4:1, of which 5% of the training set was used as the verification set, and fivefold cross validation was used for model training. The initial learning rate of training was set to 0.0001, and warm-up training with a step size of 5 was performed before training began, while the learning rate was adjusted by a 1-D linear interpolation. In the training process, the cosine annealing algorithm was used to adjust the learning rate in real time. The number of training epochs was set to 1200, and the batch size was set to 32 according to the default values.

Datasets A, B, and C were, respectively, one containing only real SSS data, one containing only data generated by the method proposed in this article, and one containing the real data and generated data, in which the real mine SSS images were augmented by a factor of 10. The generated mine data were filtered, and a total of 300 images were eliminated because the augmentation failed. The performance of the trained model was evaluated by the 20 real mine SSS images in the evaluation set.

The loss values and AP0.5:0.95 values of the model during training on the three datasets are shown in Fig. 7.

Fig. 7 shows that the loss value of the models decreased as the number of training steps increases in the three sets of experiments and finally stabilized, which proves that there was no overfitting of the models to each dataset. Among them, the loss and AP values of the models that were trained on groups B and C, which contained the augmented samples obtained using the proposed method, were significantly higher than those trained on group A, which only included the real data, demonstrating the effectiveness of increasing the data generated by the proposed method to improve the training performance of the model.

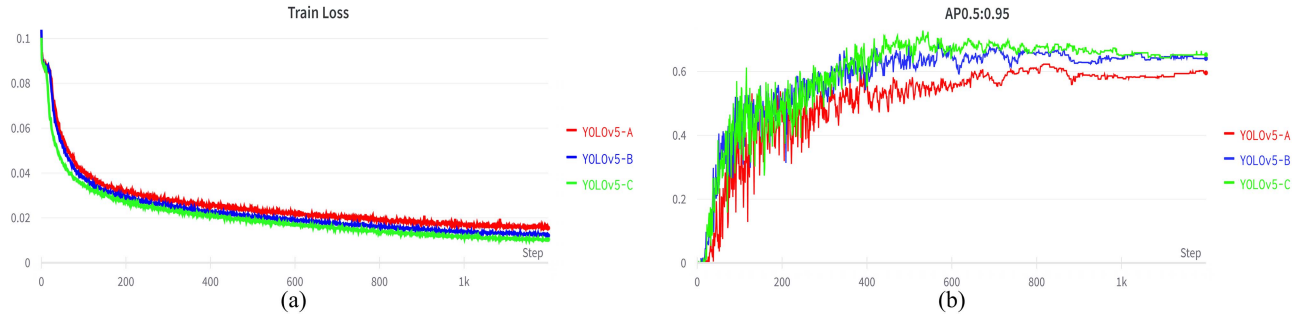

Fig. 7. Loss and AP values during the training process of the three groups of models. (a) Loss value of the training process. (b) AP0.5:0.95 for the training process.

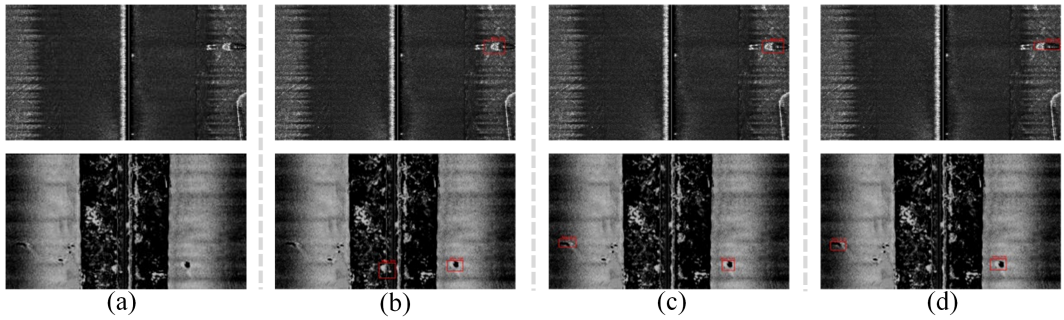

Fig. 8. Detection results of the three models on the mine SSS images. (a) Original SSS waterfall image and detection results of (b) YOLOv5-A, (c) YOLOv5-B, and (d) YOLOv5-C.

TABLE V
DETECTION PERFORMANCE ON REAL SSS IMAGES OF YOLOV5 MODELS TRAINED ON DIFFERENT TRAINING SETS

	Recall	Precision	AP0.5	AP0.5: 0.95
YOLOv5-A	80.14	84.02	74.08	47.62
YOLOv5-B	86.51	88.02	78.14	57.44
YOLOv5-C	87.63	88.21	79.96	58.35

Comparing models trained on groups B and C, it can be found that the final AP values of the two groups of experiments were not much different, which proves that training with augmented samples using the proposed method is the key to improving the performance of the YOLOv5 model. The images produced by the proposed method are almost completely consistent with the real images in terms of the authenticity and diversity of features, which demonstrates the effectiveness of the proposed method.

The trained model was tested using 20 real SSS mine images, and the recall, precision, and AP, which are widely used metrics in the field of target detection and evaluation, were used to evaluate the model.

Table V once again presents that the recall, precision, and AP values of the model trained using the images generated by the proposed method were significantly higher than those of the model trained using only real SSS data. It was proved that the generated data play a key role in improving the performance of the model. The difference between the evaluation metrics of YOLOv5-C, which was trained using real data and generated data, and YOLOv5-B, which was trained only on generated data, is not substantial, which proves that the improvement in model performance is mainly due to the use of the proposed method to generate data. Alternatively, it proves that the images generated by the proposed method meet the requirements of realism and diversity.

The real mine SSS images obtained in the Sanya sea trial (group 1) and Guangzhou lake test (group 2) were detected by the three trained models, and some of the results are compared in Fig. 8.

By comparing the images of Fig. 8(b)–(d) for group 1, it can be found that the confidence of the YOLOv5-A model trained to detect mine targets using only dataset A was only 65%, and there was no good identification of the shadow of the mine target in terms of positioning accuracy. The YOLOv5-B and YOLOv5-C models trained using the augmented datasets B and C, respectively, obtained better performance in the shadow identification of shipwreck targets. Both the positioning accuracy and confidence were significantly higher than those of the models trained only using real data, obtaining confidence levels of 83% and 86%, respectively.

By comparing the images of Fig. 8(b)–(d) for group 2, it can be found that, because of the large amount of rubble in the scanning area, the YOLOv5-A model failed to successfully identify the mine target and misidentified a mine-like reef. Moreover, there were several missed alarms and false alarms. By contrast, YOLOv5-B and YOLOv5-C successfully identified the mine

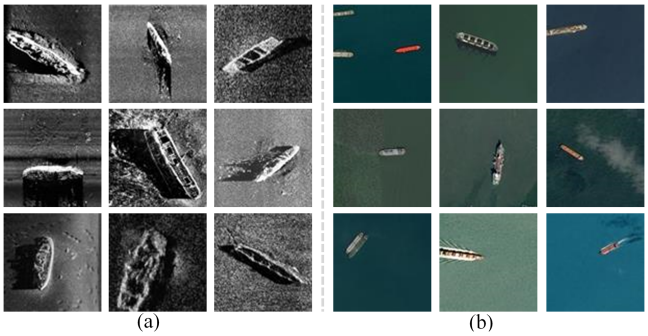

Fig. 9. Examples from the datasets. (a) Example images from the SSS shipwreck dataset. (b) Examples from the satellite optical ship image dataset.

targets, achieving a higher positioning accuracy and confidence scores of 81% and 82%, respectively. Nevertheless, both models had false alarms: they misidentified the small mine-like reef on the port side as a mine target. However, in actual operation, a missed target alarm may lead to more serious consequences than a false alarm. Moreover, how to reduce the rate of false alarms in subsequent development is a direction of future research.

In summary, the proposed method first created mine target images from scratch and established a real mapping relationship, and second, the target detection model trained using the augmented dataset yielded higher recognition accuracy and positioning accuracy, which proves that the proposed method achieves the high-quality sample augmentation of targets with no-sample images and improves the performance of underwater target detection models, which basically meets the needs of practical applications and can be applied to other types of underwater targets.

B. Experiments on the Proposed GAN

Dual-domain image conversion based on GAN is an important component of the method proposed in this article; therefore, this experiment was carried out to evaluate the performance of the proposed GAN. The proposed model was mainly evaluated through the performance of the dual-domain conversion of nondomain optical ship images and SSS shipwreck images. This evaluation included a comparison with conventional GANs, qualitative and quantitative analyses of the quality of the generated images, the effect of the generated images on the detection performance of the YOLOv5 target detection model, and qualitative and quantitative analyses of the effectiveness of the strategies used in the GAN through ablation experiments.

1) Datasets: The datasets consisted mainly of SSS shipwreck images and satellite optical ship data. The SSS shipwreck dataset consists of 600 SSS shipwreck images obtained in the Yellow Sea, Bohai Sea, East China Sea, and South China Sea by various hydrographic departments and manufacturers using the conventional Chinese and foreign-manufactured SSS instruments and equipment, such as the Klein3000, EdgeTech4200, Yellowfin, and Hydra series. Some of the samples are shown in Fig. 9(a). The satellite optical ship image dataset was taken from HRSC2016, which is a high-resolution ship dataset from Google Earth. A total of 5000 representative images were selected for this experiment, and a part of the samples is shown in Fig. 9(b).

TABLE VI
PERFORMANCE OF IMAGES GENERATED USING DIFFERENT CONVENTIONAL MODELS

Group	Model	FID↓	MMD↓	1-NN↓0.5
1	pix2pix	214.77	0.315	0.97
2	CG-ResNet-06	153.65	0.192	0.82
3	CG-ResNet-09	148.71	0.164	0.83
4	CG-UNet-128	155.12	0.137	0.77
5	CG-UNet-256	133.69	0.151	0.79
6	DualGAN	130.17	0.160	0.74
7	DiscoGAN	129.98	0.149	0.75
8	Our Method	123.12	0.105	0.72

2) Design of the Experiments: The evaluation indicators and experimental configuration used in this evaluation were consistent with those described in Section III-A2. Because the resolution of the original images was large and most of the pixels were background, it was difficult to train the model on the HRSC2016 data. Hence, all optical image data were uniformly set to a resolution of 250×250 and included the target. The SSS shipwreck images were divided into training and evaluation sets using a ratio of 5:1, and the ship optical images were divided into training and conversion sets using a ratio of 9:1. When introducing cached historical data for training, the number of images stored in the list was adjusted to the previous 100 images because the data samples were larger than the mine images in experiment A.

3) Experiments and Analysis:

a) Quantitative analysis: In this section, the training process and performance of the model are first analyzed and evaluated. The network proposed in this article was inspired by CG [38]. Hence, the model proposed in this article was compared with the CG model with different structures (i.e., the generator adopted ResNet-06, ResNet-09, UNet-128, and UNet-256 backbone networks).

As Fig. 10 shows, the loss values of the five models decreased as the number of training steps increased and eventually stabilized to reach the fitted state. Among them, the proposed network had the lowest cyclic-consistency loss, LSGAN loss, and identity loss, and was the most stable over the whole training process. There were no other networks that exhibited large amplitude changes during training.

Because the proposed model belongs to models that transfer the image style of two domains using unsupervised learning, the above model was compared in terms of image generation with the conventional methods pix2pix [39], DualGAN [40], and DiscoGAN [41], and the comparison target consisted of the 500 SSS shipwreck images generated by optical conversion from the HRSC2016 conversion set and the 100 images from the real SSS evaluation set. They were quantitatively analyzed and the FID, MMD, and 1-NN were calculated, respectively. The final quantitative test results are presented in Table VI.

By comparing groups 2, 3, 4, and 5, it can be seen that when the model structure was not complex, more parameters improved performance. On the contrary, because of the low resolution, less distinct features, and other characteristics of the

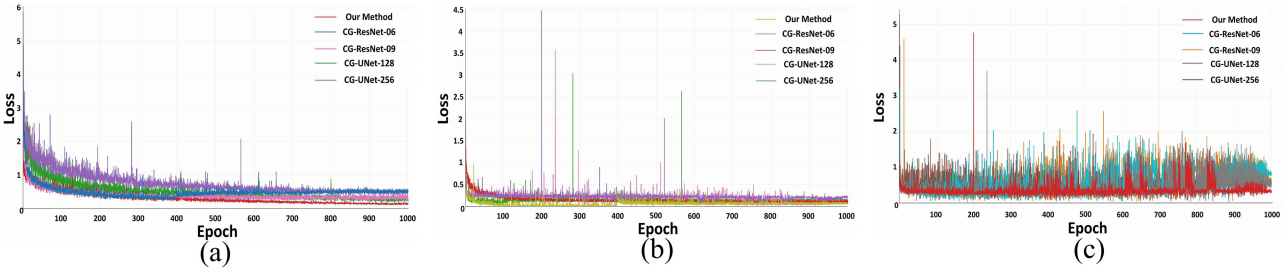

Fig. 10. Loss values of the training process of the five models. (a) Cyclic-consistency loss. (b) Identity loss. (c) LSGAN loss.

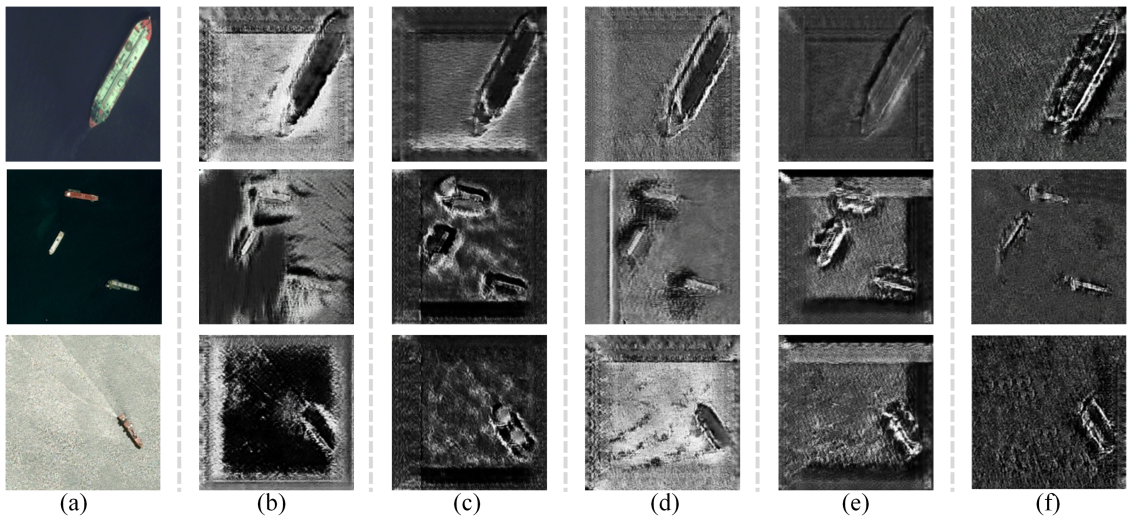

Fig. 11. Results of converting SSS images using five models for optical image conversion. (a) Input optical images and converted results obtained by (b) CG-ResNet-06, (c) CG-ResNet-09, (d) CG-UNet-128, (e) CG-UNet-256, and (f) proposed method.

SSS images, when a more complex model structure is used to generate the images, the performance of generation does not necessarily improve. The results for group 1 demonstrate that the quality of the images generated by the pix2pix model was the least satisfactory, possibly because the model needed paired datasets as input for training. This experiment used different domain images of the same target, but in addition to the difference in the background, there were also multiple differences in the size, orientation, texture, and resolution, and the pairs cannot be considered to be ideal pairwise images. For groups 2 to 8, the models could achieve better results than the model of group 1 because of the use of unsupervised learning, which does not require paired dual-domain images to generate high-quality images. As can be seen from the results of groups 6 and 7 (DualGAN and DiscoGAN, which are related to CG), the difference in performance between the network in the optical and SSS dual-domain image conversion tasks of shipwreck targets was not large and both can perform the task well.

Compared with the other groups, it can be seen that for group 8, the FID and MMD values obtained using the model proposed in this article were the lowest, and the 1-NN value was the closest to 0.5, which proves that the images generated by this model had a higher fitting degree with the real SSS shipwreck images than the above models, obtaining better clarity, detail, and realism while maintaining a lower probability of mode collapse.

b) Qualitative analysis: Fig. 11 shows the conversion of five models on three typical representative optical images with large targets, multiple targets, and small targets.

From Fig. 11, it can be seen that the five models have basically realized the cross-domain conversion from the optical domain to the acoustic domain and augmented the sample. Among them, CG-ResNet-06 [see Fig. 11(b)] and CG-ResNet-09 [see Fig. 11(c)] could not generate the texture characteristics of the shipwreck well and black holes appeared on the background. In Fig. 11(d), although the background was better generated, there are still white bars and boxes in the background. In Fig. 11(e), in the generation of the texture features of multiple targets and small targets, CG-UNet-256 achieved a certain improvement over the previous models, but there are black boxes on the border, which may be because the shadow features of the shipwreck target were incorrectly learned as background information. Our method, when compared with the previous four models, realized good results both in the generation of texture features and background features of the image.

c) Performance of the target detection model: For the shipwreck targets evaluation, the strategy of the experiment was consistent with that in Section III-A4.c. Three sets of datasets

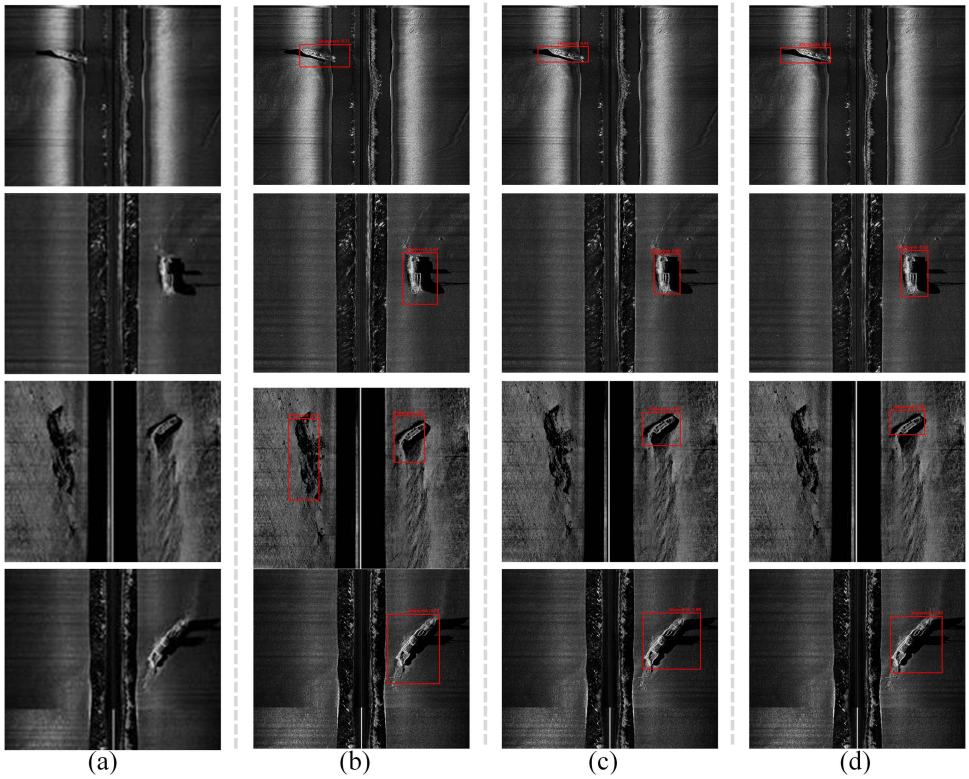

Fig. 12. Comparison of the detection results of the three models on several real SSS targets. (a) Original input SSS images and the detection results of (b) YOLOv5-A, (c) YOLOv5-B, and (d) YOLOv5-C.

TABLE VII
COMPOSITION OF THE TRAINING AND VERIFICATION DATASETS

Group	Real shipwreck image	Generated shipwreck image
A	500	–
B	–	2000
C	500	2000
verify	100	–

TABLE VIII
REAL IMAGE DETECTION PERFORMANCE OF THE YOLOV5 NETWORK TRAINED ON DIFFERENT TRAINING SETS

	Recall	Precision	AP0.5	AP0.5: 0.95
YOLOv5-A	79.12	84.14	80.43	46.21
YOLOv5-B	82.47	88.29	83.68	50.66
YOLOv5-C	83.22	88.65	84.71	51.14

were designed to train the YOLOv5 model, as presented in Table VII. One contains only real SSS data, one contains only data generated by the proposed model, and one contains real data and generated data. Moreover, 100 real SSS images were selected to evaluate the performance of the trained model. Among them, the generated shipwreck data were filtered to exclude images that failed to be augmented.

The trained model was evaluated using 100 real SSS images, and the recall, precision, and average accuracy, which are used in the field of target detection, along with the AP, were used to evaluate the model. The results are reported in Table VIII.

Table VIII presents that the recall, precision, and AP values of the model trained using the images generated by the proposed method were higher than those trained using only real SSS data. This proves that the generated data played a key role in the performance improvement of the model, and the conclusion is consistent with that drawn in Section III-A4.c, demonstrating the effectiveness of the proposed method.

As can be seen from Fig. 12, the YOLOv5 models trained with datasets A, B, and C could identify real submarine shipwreck targets, but comparing the images in Fig. 12(b)–(d), it was found that the confidence of the model trained only using dataset A was on average 65% in shipwreck target recognition. Moreover, the positioning accuracy needs to be enhanced and the shadow of the shipwreck was not well identified. In terms of the identification accuracy, the reef target in the images of group 3 was mistakenly identified as a shipwreck target. The model trained using the augmented datasets B and C yielded better results for the shadow identification of the shipwreck targets, and the positioning accuracy and confidence were substantially higher than the models trained using only the real dataset. The average confidence level was 90%.

The above experiments show that the image samples augmented by the proposed method had a realism, detail, and integrity that are more similar to the real SSS images. Moreover, the purpose of improving the detection performance of the target detection model based on deep learning was achieved.

d) Ablation experiment and evaluation: To evaluate the role of each module in the performance of the proposed model,

TABLE IX
PERFORMANCE OF THE VARIOUS IMPROVEMENTS OF THE PROPOSED GAN

Group	CSA Module		LSGAN	FID↓	MMD↓	1-NN↓0.5
	Channel	Spatial				
1	–	–	–	136.52	0.211	0.91
2	√	–	–	132.95	0.169	0.81
3	–	√	–	133.01	0.152	0.80
4	√	√	–	126.84	0.131	0.77
5	–	–	√	130.98	0.134	0.78
6	√	√	√	123.12	0.102	0.72

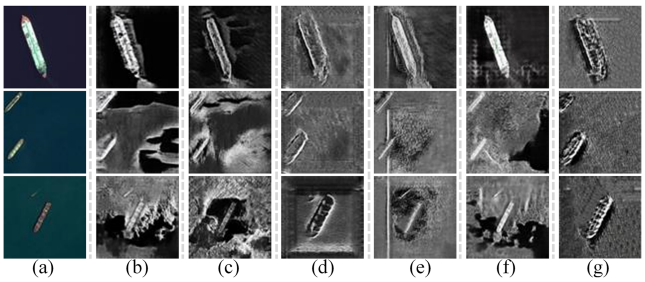

Fig. 13. Conversion performance on optical images obtained by six models. (a) Input optical images and SSS images converted using the model in (b) group 1, (c) group 2, (d) group 3, (e) group 4, (f) group 5, and (g) group 6.

the controlled variable method was used to perform ablation experiments on the CSA module and LSGAN loss function. Six groups of control experiments were designed, and the experimental configuration, training dataset, and evaluation dataset were consistent with those in Section III-B2. The experimental results are reported in Table IX.

A comparison of groups 1–4 reveals that the quality of the images generated by the model was higher after the attention mechanism was incorporated. Group 4, which integrated the CSA mechanisms, outperformed groups 2 and 3, which only used a CSA module, respectively, proving the effectiveness of the CSA module proposed in this article. By comparing group 5 with group 1, the superiority of the proposed LSGAN loss function can be seen. A comparison of group 6 with groups 4 and 5 reveals that the performance of the model after adding the CSA module and LSGAN loss function is better than that of a single strategy, which plays a crucial role in the overall performance improvement of the model and reflecting the effectiveness of the method proposed in this article.

The conversion results of six models trained using different strategies on several optical images are shown in Fig. 13. From Fig. 13, it can be seen that the data generated by the model in group 1 have the lowest realism. A comparison of group 2 with group 1 reveals that the model using the channel attention module can extract more detailed features when generating targets, but it needs to be strengthened in terms of the generation of background features. By comparing group 3 with group 1, it can be seen that the model using the spatial attention module performs better in generating the background of the image, but there are still black holes in the background, and the generation of target detail features is mediocre. A comparison of group 4 with group 1 reveals that the model using the CSA module significantly improves its ability to generate target detail features and background features. By comparing group 5 with group 1, it can be seen that the model using the LSGAN loss function generated the target image better, but there are still obvious boxes at the edge of the background, which is not sufficiently natural. Comparing group 6 with group 1 reveals that the model that integrates the CSA module and LSGAN loss function generated target images with high definition, complete detail features, and high realism both in terms of detailed features, such as the texture and edges of the shipwreck targets or the background features, which proves the effectiveness of the proposed method.

IV. DISCUSSION

A. Advantages of the Method

The proposed method realized the augmentation of samples with few or even no underwater target sonar images based on the true cross-domain mapping relationship of the same physical object, which largely solves the problems in the existing inter-domain style transfer methods, such as the problem that target in the nondomain images is not the target to be detected or the constructed conversion model is not accurate, resulting in weak representativeness of augmented samples and poor performance of the target detection model. The proposed method provides a new approach to the augmentation of strongly representative target samples and construction of high-performance underwater target detection models. The following is a further discussion of the innovations proposed in this article.

1) Sample Augmentation Based on Cross-Domain Mapping of the Same Physical Model: The essence of a target detection model training based on deep learning is to imitate the human optic nerve in order to extract the shallow and deep features of a target from a large number of data samples. The number and representative strength of the samples are the key factors that determine the performance of the model. The cross-domain/style transfer method was implemented for sample augmentation based on the nonreal mapping relationships of nonidentical target entities, which leads to systematic bias, resulting in low authenticity of the generated images. Moreover, SSS images are a nonlinear mapping of a sound wave emission unit, propagation medium, target reflection, background reflection, receiving unit, noise, and data postprocessing unit in image space. It is difficult to qualitatively simulate all these elements, and hence, sample augmentation based on the real mapping relationship is particularly meaningful, especially for underwater targets with few samples and or no samples and an incomplete understanding of attributes.

2) Proposed GAN: The sample augmentation model based on GAN actually finds the intrinsic implicit relationship between a series of influencing factors, such as SSS image target, background, texture, noise, echo intensity, and so on, through the confrontation and game between the generator and the discriminator so as to extract the target and background features and generate a new image. To a certain extent, the influence of feature selection in image enhancement techniques that needs to rely on artificial experience and imagination is avoided. The key here is to input optical and SSS images of the same object and

TABLE X
COMPARISON OF SONAR TARGET DETECTION ACCURACY OBTAINED USING THE PROPOSED METHOD AND EXISTING SAMPLE AUGMENTATION METHODS

Num	Method	Object	Detection-Net	Accuracy	References
1	Our Method	Mine Shipwreck	YOLOv5	89.22% G 87.18% G (AP)	/
2	Optical Image MetaStyle	Shipwreck	YOLOv5	96.2% G (MAP)	[24]
3	Optical Image Photowct	Airplane Shipwreck Others	DCNN	75.1% G (MAP)	[18]
4	Optical Image	Airplane Shipwreck Others	DCNN	93.83% R 97.76% RG (OA)	[23]
5	Optical Image SM-CycleGAN	sphere	ResNet-18	94.47%RG (Precision)	[31]
6	3D Modeling GAN	Tire	YOLO	86.6% G (TP)	[42]
7	CAD Model Stylebanknet	Drowning Victim	Faster R-CNN	77% G (AP)	[22]
8	SPADE	Shipwreck	PSPNet	88.51% G (mIoU)	[35]
9	Noise Adversarial Network	Airplane Shipwreck Others	Faster R-CNN	81.7% R 88.3% RG (MAP)	[43]

(G means using only generated data for training, R means using only real sonar data for training, and RG means using superimposed data of both.)

establish a real mapping relationship to facilitate the authenticity of the image.

B. Comparison With Existing Methods

The small-sample size of underwater target data is a common problem in the field of underwater target detection. Most current research is carried out in the form of transfer learning and data augmentation. Table X lists the representative work in the field of augmentation of underwater target sonar image samples in recent years, which may be mainly classified into four types from the perspective of implementation: optical image + style transfer (2, 3, and 4), 3-D modeling + style transfer (6 and 7), image enhancement (9 and 10), and GAN models (5, 6, and 8).

Among them, optical image + style transfer models (2, 3, and 4) need to collect a large number of real optical images and use transfer learning to achieve SSS image conversion. This method must find other datasets of targets that are similar to underwater targets, and this is only applicable to a few common underwater targets, such as ships. In addition, a real mapping relationship cannot be established for targets that are of the same type but not exactly the same, which often restricts the quality of the generated images. By contrast, 3-D modeling + style transfer models (5 and 6) obtain the shape of the target through simulation and then obtain texture, noise, and other features through style transfer. The factors affecting the quality of SSS images are strongly correlated with each other, and it is difficult to qualitatively describe the mathematical relationship between them. Image enhancement (9) enhances sonar image data by introducing a noise model, but a single dimension often yields a limited improvement in accuracy. GAN models (5, 6, and 8) also need to train the network through multidomain images, and the real mapping relationship of the target multidomain images cannot be established. In addition, methods 6–8 require the images to be segmented and labeled.

Because publicly available datasets in the field of SSS image detection are extremely rare, most existing studies use self-generated data, and the detection performance is related to the model, the number and representativeness of the training samples, and the complexity of the target in the test images. In addition, the evaluation indicators used by different models are different, and hence, the accuracy in Table X is only of reference significance.

The proposed method belongs to the style transfer method based on GAN models in which the GAN is trained by a small number of real optical and SSS images of the same target to obtain high-quality SSS images of underwater targets. In contrast to the above methods, the training data used in this study have a real mapping relationship between the target in the images from both domains, which can achieve high-quality sample augmentation with few or even no samples and provides high-quality data support for the final high-performance target detection model based on deep learning.

C. Limitations of the Method

1) High Cost and Complex Process: In practice, the method proposed in this article has problems with high cost and complex processes, as it involves the building of the physical model and real-world sea trials to obtain the SSS images. However, compared with the traditional sample augmentation, which is achieved only through optical images, hand-drawn images, and so on, this method at least has the characteristics of real SSS images, and the real mapping relationship between the two domains for the target is established. Although 3-D printing can create targets with known materials, textures, and structures,

it cannot produce targets with unknown properties. However, in general, the method proposed is this article can be used for the augmentation of deep-learning training data when there are few or even no samples and obtain high-performance detection models.

Moreover, although the 3-D printing technology used in this article was able to build a physical model of the underwater targets, it is surely not the simplest and most efficient method, and the process flow could be further optimized to achieve the aim of reducing costs and increasing efficiency.

2) Optimizing the GAN Model: Although the GAN designed in this article achieved good performance in the augmentation of underwater target samples, there is a room for improvement in the FID, MMD, and 1-NN values of the generated samples. Moreover, in some cases where the difference between the images in the two domains was too large, the target feature was generated in the background or the background was generated in the target feature, so it will be necessary to further optimize the GAN model.

3) Obtaining Representative Data: Although the method proposed in this article achieved good detection accuracy on the YOLOv5 model, the samples used for generator training were high-quality data obtained in simple scenes, single scenes, and a good sea state. In contrast, when the SSS is actually working at sea under complex sea conditions, the attitude of the SSS itself will be seriously affected by waves and surges and the sonar data will be heavily disturbed by noise. Hence, the quality of the data obtained could be very poor. Therefore, it will be necessary to further improve the representativeness of the samples by acquiring a small amount of representative data for different attribute targets, different marine environments, multiple backgrounds, multiple angles, and different equipment situations.

4) Generating a Specified Sample: Although the method proposed in this article augmented the sample based on the real mapping relationship, the acquisition of augmented sample images was random, and it was not possible to obtain targeted images with a specified background, noise texture, or other factors, as in [27]. It is possible that a combination of the method proposed in this article with the approach in [27] would achieve good results.

D. Applicability of the Proposed Method

The proposed method demonstrates extensive versatility and applicability. It goes beyond underwater SSS target sample augmentation and can be effectively applied to other domains as well. For instance, in the field of UAV optical imagery and satellite remote sensing, the method can establish real mapping relationships and conduct sample augmentation, thereby improving the accuracy of remote sensing data processing and UAV image analysis. In the medical domain, by establishing cross-domain mappings between X-ray images and CT scans, the method can provide comprehensive and accurate information for medical image analysis and diagnosis. Additionally, in the field of facial recognition, by mapping and augmenting optical facial images with infrared facial images, the method enhances the robustness and accuracy of facial recognition systems. Hence, the proposed method holds significant potential across various domains, providing valuable support for related research and applications.

V. Conclusion

An augmentation method for SSS image samples of the underwater target based on the cross-domain mapping relationship of the same object was proposed in this article, aiming to address the problem in the existing sample augmentation methods for cross-domain conversion/style transfer, which is that the targets in the nondomain images are not exactly the same as the actual underwater target that is to be detected. This work makes the following three main contributions.

1) An augmentation method for SSS image samples of an underwater target based on the cross-domain mapping relationship of the same physical object is proposed. A physical model of the underwater target was built using 3-D printing technology; the mapping relationship of the physical model-to-optical image was established using the same imaging methods and mechanisms use for underwater target imaging, and the physical model-to-SSS image was established using the actual measurement mode of SSS. Finally, the cross-domain mapping relationship of the optical—SSS images of the target was established, addressing the restrictions in the sample augmentation performance caused by the fact that the target is an underwater target but is not exactly the same as the one to be detected in the nondomain image.
2) GAN based on circular consistency was designed. A single-cycle consistent network structure was designed to ensure the training efficiency of the model. The CSA module was integrated into the generator to reduce information diffusion and enhance cross-dimension interaction, thereby improving the quality of the generated images. The loss function, which is based on an LSGAN, was designed to stabilize training and avoid mode collapse. The cross-domain conversion of target optical images and SSS images was realized, and the high-quality augmentation of samples was achieved, which improved the accuracy of the target detection model based on deep learning.
3) The above method was verified by real sea experiments. The real cross-domain mapping relationship between optical and SSS images of underwater targets was established, and the conversion of dual-domain images was realized. Compared with the images generated by the traditional sample augmentation methods, the SSS underwater target images generated by this method had higher definition, more complete detail features, and higher realism, which achieves the purpose of the high-quality augmentation of samples with few or even no images, proving the effectiveness and necessity of the method. This method could be adapted for other types of underwater targets because of its value as a reference.

In summary, the proposed underwater target SSS image sample augmentation method in this article realized the high-quality

augmentation of underwater target SSS image samples with few or even no images. The proposed approach improves the detection accuracy of the model and provides a new way to augment underwater target samples with strong representativeness as well as construct high-performance underwater target detection models, which will have practical guiding significance and value.

REFERENCES

[1] L. Henriksen, "Real-time underwater object detection based on an electrically scanned high-resolution sonar," in *Proc. IEEE Symp. Auton. Underwater Veh. Technol.*, 1994, pp. 99–104.

[2] M. Rizal Arshad, "Recent advancement in sensor technology for underwater applications," *Indian J. Geo-Mar. Sci.*, vol. 38, no. 3, pp. 267–273, 2009.

[3] A. Greene, A. F. Rahman, R. Kline, and M. S. Rahman, "Side scan sonar: A cost-efficient alternative method for measuring seagrass cover in shallow environments," *Estuarine, Coastal Shelf Sci.*, vol. 207, pp. 250–258, Jul. 2018.

[4] D. Buscombe, "Shallow water benthic imaging and substrate characterization using recreational-grade sidescan-sonar," *Environ. Model. Softw.*, vol. 89, pp. 1–18, Mar. 2017.

[5] H. J. Flowers and J. E. Hightower, "A novel approach to surveying sturgeon using side-scan sonar and occupancy modeling," *Mar. Coastal Fisheries*, vol. 5, no. 1, pp. 211–223, Jan. 2013.

[6] B. Zhu, X. Wang, Z. Chu, Y. Yang, and J. Shi, "Active learning for recognition of shipwreck target in side-scan sonar image," *Remote Sens.*, vol. 11, no. 3, Jan. 2019, Art. no. 243.

[7] F. Yang, Z. Du, and Z. Wu, "Object recognizing on sonar image based on histogram and geometric feature," *Bull. Mar. Sci.*, vol. 25, pp. 64–69, 2006.

[8] X. Shang, J. Zhao, and H. Zhang, "Automatic overlapping area determination and segmentation for multiple side scan sonar images mosaic," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2886–2900, Feb. 2021.

[9] F. Langner, C. Knauer, W. Jans, and A. Ebert, "Side scan sonar image resolution and automatic object detection, classification and identification," in *Proc. Oceans Europe*, 2009, pp. 1–8.

[10] J. C. Isaacs, "Sonar automatic target recognition for underwater UXO remediation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 134–140.

[11] Z. Wang, S. Zhang, C. Zhang, and B. Wang, "RPFNet: Recurrent pyramid frequency feature fusion network for instance segmentation in side-scan sonar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published, doi: 10.1109/JSTARS.2023.3266383.

[12] F. Yuan, F. Xiao, K. Zhang, Y. Huang, and E. Chen, "Noise reduction for sonar images by statistical analysis and fields of experts," *J. Vis. Commun. Image Representation*, vol. 74, Jan. 2021, Art. no. 102995.

[13] N. Kumar, U. Mitra, and S. S. Narayanan, "Robust object classification in underwater sidescan sonar images by using reliability-aware fusion of shadow features," *IEEE J. Ocean. Eng.*, vol. 40, no. 3, pp. 592–606, Jul. 2015.

[14] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, "Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5786–5795, May 2021.

[15] J. M. Topple and J. A. Fawcett, "MiNet: Efficient deep learning automatic target recognition for small autonomous vehicles," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 1014–1018, Jun. 2021.

[16] H.-T. Nguyen, E.-H. Lee, and S. Lee, "Study on the classification performance of underwater sonar image classification based on convolutional neural networks for detecting a submerged human body," *Sensors*, vol. 20, no. 1, 2020, Art. no. 94.

[17] P. Feldens, A. Darr, A. Feldens, and F. Tauber, "Detection of boulders in side scan sonar mosaics by a neural network," *Geosciences*, vol. 9, no. 4, Apr. 2019, Art. no. 159.

[18] C. Li, X. Ye, D. Cao, J. Hou, and H. Yang, "Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples," *Appl. Acoust.*, vol. 173, Feb. 2021, Art. no. 107691.

[19] E. Coiras, P.-Y. Mignotte, Y. R. Petillot, J. Bell, and K. Lebart, "Supervised target detection and classification by training on augmented reality data," *IET Radar, Sonar Navig.*, vol. 1, no. 1, pp. 83–90, Mar. 2007.

[20] N. Kapetanovic, N. Miskovic, and A. Tahirovic, "Saliency and anomaly: Transition of concepts from natural images to side-scan sonar images," in *IFAC-PapersOnLine*, vol. 53, pp. 14558–14563, 2020.

[21] F. Li, R. Feng, W. Han, and L. Wang, "An augmentation attention mechanism for high-spatial-resolution remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3862–3878, Jul. 2020.

[22] S. Lee, B. Park, and A. Kim, "Deep learning from shallow dives: Sonar image generation and training for underwater object detection," 2018, *arXiv.1810.07990*.

[23] G. Huo, Z. Wu, and J. Li, "Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data," *IEEE Access*, vol. 8, pp. 47407–47418, 2020.

[24] C. Huang, J. Zhao, Y. Yu, and H. Zhang, "Comprehensive sample augmentation by fully considering SSS imaging mechanism and environment for shipwreck detection under zero real samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Oct. 2021.

[25] P. S. Chavez Jr. et al., "Processing, mosaicking and management of the Monterey Bay digital sidescan-sonar images," *Mar. Geol.*, vol. 181, no. 1/3, pp. 305–315, Mar. 2002.

[26] P. Blondel, *The Handbook of Sidescan Sonar*. Berlin, Germany: Springer, 2010, pp. 7–83.

[27] Y.-C. Chang, S.-K. Hsu, and C.-H. Tsai, "Sidescan sonar image processing: Correcting brightness variation and patching gaps," *J. Mar. Sci. Technol.*, vol. 18, no. 6, pp. 785–789, Jan. 2010.

[28] N. Bore and J. Folkesson, "Modeling and simulation of sidescan using conditional generative adversarial network," *IEEE J. Ocean. Eng.*, vol. 46, no. 1, pp. 195–205, Jan. 2021.

[29] Y. Xu, X. Wang, K. Wang, J. Shi, and W. Sun, "Underwater sonar image classification using generative adversarial network and convolutional neural network," *IET Image Process.*, vol. 14, no. 12, pp. 2819–2825, 2020.

[30] Y. Steiniger, D. Kraus, and T. Meisen, "Survey on deep learning based computer vision for sonar imagery," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105157.

[31] B.-Q. Li, H.-N. Huang, J.-Y. Liu, and Y. Li, "Optical image-to-underwater small target synthetic aperture sonar image translation algorithm based on improved CycleGAN," *Acta Electronica Sinica*, vol. 49, no. 9, pp. 1746–1753, 2021.

[32] J. L. Chen and J. E. Summers, "Deep neural networks for learning classification features and generative models from synthetic aperture sonar big data," *Proc. Meetings Acoust.*, vol. 29, 2016, Art. no. 032001.

[33] A. I. Karjalainen, R. Mitchell, and J. Vazquez, "Training and validation of automatic target recognition systems using generative adversarial networks," in *Proc. Sensor Signal Process. Defence Conf.*, 2019, pp. 1–5.

[34] A. Reed, I. Gerg, J. McKay, D. Brown, D. Williams, and S. Jayasuriya, "Coupling rendering and generative adversarial networks for artificial SAS image generation," 2019, *arXiv:1909.06436*.

[35] Y. Jiang, B. Ku, W. Kim, and H. Ko, "Side-scan sonar image synthesis based on generative adversarial network for images in multiple frequencies," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1505–1509, Sep. 2021.

[36] B. Berman, "3D printing: The new industrial revolution," *IEEE Eng. Manage. Rev.*, vol. 41, no. 4, pp. 72–80, Dec. 2013.

[37] Q. Xu et al., "An empirical study on evaluation metrics of generative adversarial networks," 2018, *arXiv:1806.07755*.

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.

[40] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2868–2876.

[41] T. Kim, M. Cha, H. Kim, J. W. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 1857–1865.

[42] M. Sung et al., "Realistic sonar image simulation using deep learning for underwater object detection," *Int. J. Control, Autom. Syst.*, vol. 18, no. 3, pp. 523–534, 2020.

[43] Q. Ma, L. Jiang, W. Yu, R. Jin, Z. Wu, and F. Xu, "Training with noise adversarial network: A generalization method for object detection on sonar image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 718–727.

Yulin Tang (Member, IEEE) received the B.S. degree from Dalian Naval Academy, Dalian, China. He is currently working toward the Ph.D. degree with the Naval University of Engineering, Wuhan, China.

His main research direction is side-scan sonar image processing, underwater target detection, AUV system, and computer vision.

Liming Wang is a Professor with the Naval University of Engineering, Wuhan, China, the Doctoral Supervisor and Chief Scientist of National Key R&D Program, and expert in the fields of intelligent equipment and unmanned platforms. He is mainly engaged in the fields of ship intelligent monitoring technology, intelligent detection robots, etc.

Shaofeng Bian is a Professor with the Naval University of Engineering, Wuhan, China. He got founded by National Science Foundation for Distinguished Young Scholars in 2001.

Dr. Bian was a recipient of an Alexander von Humboldt Research Fellowship in 1996.

Shaohua Jin (Member, IEEE) received the postdoctoral degree from the Second National Institute of Oceanography in 2019.

He is an Associate Professor with the Department of Hydrography and Cartography, Dalian Naval Academy, Dalian, China. His main research direction is marine and maritime geodetic survey data processing.

Yuting Dong (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2018.

From 2019 to 2020, she was a Postdoctor with German Aerospace Center (DLR). She is currently a Professor with the China University of Geosciences, Wuhan, China. Her research interests include SAR image processing and InSAR topographic mapping.

Dr. Dong was a recipient of an Alexander von Humboldt Research Fellowship in 2022.

Houpu Li is a Professor with the Naval University of Engineering, Wuhan, China. He got founded by National Science Foundation for Outstanding Young Scholars in 2021. He specializes in the mathematical analysis of geodesy.

Bing Ji received the Ph.D. degree from the Naval University of Engineering, Wuhan, China, in 2011.

He is currently an Associate Professor with the Naval University of Engineering, Wuhan, China. His research interests include GNSS technology and underwater navigation.