


Article

# Underwater Image Translation via Multi-Scale Generative Adversarial Network

Dongmei Yang<sup>1</sup>, Tianzi Zhang<sup>1</sup>, Boquan Li<sup>2,\*</sup>, Menghao Li<sup>1</sup>, Weijing Chen<sup>1</sup> , Xiaoqing Li<sup>1</sup>  
and Xingmei Wang<sup>1,3,\*</sup>

<sup>1</sup> College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China; yangdongmei@hrbeu.edu.cn (D.Y.); zhangtianzi@hrbeu.edu.cn (T.Z.); limenghao@hrbeu.edu.cn (M.L.); chenweijing@hrbeu.edu.cn (W.C.); lixiaoqing@hrbeu.edu.cn (X.L.)

<sup>2</sup> School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore

<sup>3</sup> National Key Laboratory of Underwater Acoustic Technology, Harbin Engineering University, Harbin 150001, China

\* Correspondence: boquanli@smu.edu.sg (B.L.); wangxingmei@hrbeu.edu.cn (X.W.)

**Abstract:** The role that underwater image translation plays assists in generating rare images for marine applications. However, such translation tasks are still challenging due to data lacking, insufficient feature extraction ability, and the loss of content details. To address these issues, we propose a novel multi-scale image translation model based on style-independent discriminators and attention modules (SID-AM-MSITM), which learns the mapping relationship between two unpaired images for translation. We introduce Convolution Block Attention Modules (CBAM) to the generators and discriminators of SID-AM-MSITM to improve its feature extraction ability. Moreover, we construct style-independent discriminators that enable the discriminant results of SID-AM-MSITM to be not affected by the style of images and retain content details. Through ablation experiments and comparative experiments, we demonstrate that attention modules and style-independent discriminators are introduced reasonably and SID-AM-MSITM performs better than multiple baseline methods.

**Keywords:** underwater image translation; generative adversarial network; convolution block attention module; style-independent discriminator



**Citation:** Yang, D.; Zhang, T.; Li, B.; Li, M.; Chen, W.; Li, X.; Wang, X. Underwater Image Translation via Multi-Scale Generative Adversarial Network. *J. Mar. Sci. Eng.* **2023**, *11*, 1929. <https://doi.org/10.3390/jmse11101929>

Academic Editors: Fausto Pedro García Márquez, Jingchun Zhou, Wenqi Ren, Qiuping Jiang and Yan-Tsung Peng

Received: 31 August 2023

Revised: 27 September 2023

Accepted: 2 October 2023

Published: 6 October 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Underwater images play critical roles in diverse marine-related military and scientific applications, such as seabed sediment classification [1], submarine cable detection [2], and mine recognition [3]. However, the complex underwater environment limits the use of camera devices, including Kinect units [4] and binocular stereo cameras [5], which makes it difficult to obtain real underwater images. Intuitively, image translation [6] provides a viable direction to obtain such scarce data. Specifically, image translation methods translate source-domain non-underwater images into target-domain underwater ones, which re-assigns particular attributes of underwater images. The translated underwater images are valuable for advanced visual tasks such as target detection, 3D reconstruction, and target segmentation [7,8].

Existing image translation methods are generally categorized into two groups, i.e., conventional and Generative Adversarial Network (GAN)-based [9] methods. First, conventional methods [10,11] extract low-level features to transfer input images' texture or devise various Convolutional Neural Network-based (CNN) [12], such as image style transfer [13], to make use of semantic content for image translation. Second, GAN-based methods, such as supervised models including Pix2Pix [14] and Pix2PixHD [15], and unsupervised models including StyleGAN [16] and StarGAN [17], use generators to translate images from one

image domain to another image domain. By comparison, GAN-based methods do not require researchers to design complex loss functions, which saves manpower.

In view of the immense potential of GAN, researchers attempt to apply it to underwater image translation tasks. Li et al. [18] propose an unsupervised WaterGAN that uses in-air RGB images and depth maps to generate corresponding realistic underwater images. Wang et al. [19] use an unsupervised image translation method that also takes in-air RGB-D images to generate realistic underwater images. Li et al. [20] work to generate images with underwater style using in-air RGB images.

Although existing methods have achieved a certain success, they still face multiple challenges. (1) Data lacking. The necessary paired in-air and depth images are too scarce to train translation models. In the absence of data, using the general image translation methods, it is difficult to achieve good results. (2) Insufficient feature extraction ability. Underwater optics images and underwater sonar images present obvious colors, which reduce the visibility of objects in translated images and reduce the quality of the translated images. These translated images limit the performance of subsequent advanced computer vision tasks [21]. Therefore, the colors of the underwater images pose challenges to the feature extraction ability of image translation. (3) Loss of content details. GAN models are sensitive to the style of images (such as color and texture) [22], which makes image translation models ignore the content information of images. Similar to the lack of feature extraction ability, the loss of content details also limits the performance of subsequent advanced computer vision tasks.

With the aim of addressing the above challenges, in this paper, we propose a multi-scale underwater image translation model based on style-independent discriminators and attention modules (SID-AM-MSITM). Specifically, our contributions mainly include the following:

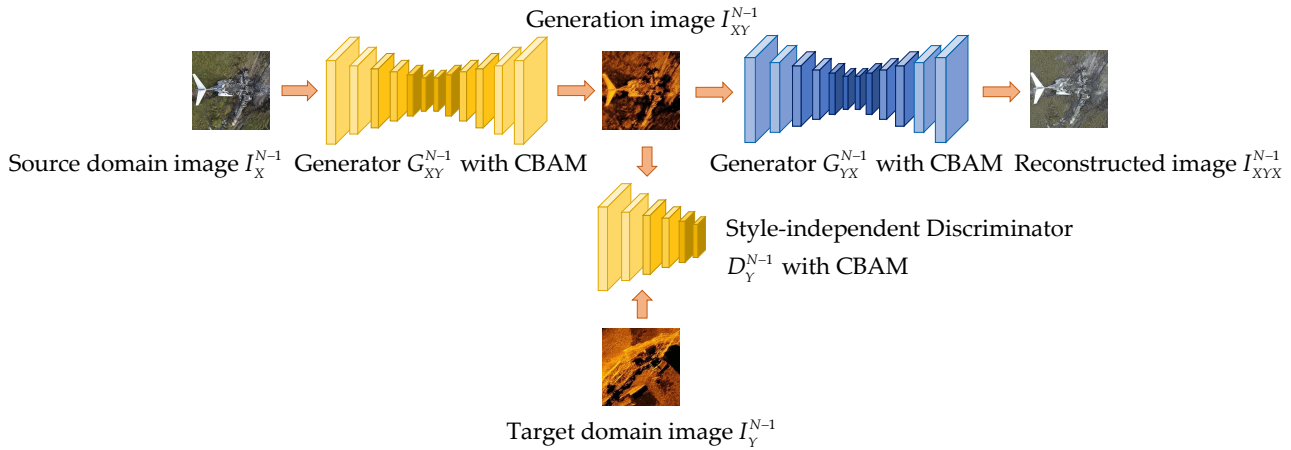
1. In response to the data lacking challenge, we construct the backbone model of SID-AM-MSITM based on a fundamental image translator, TuiGAN [23]. TuiGAN conducts image translation tasks based on only two unpaired images, and we thus make further improvements on its encoders and decoders.
2. In response to the challenge of insufficient feature extraction ability, we propose to apply Convolution Block Attention Modules (CBAM) [24] to the generators and discriminators of SID-AM-MSITM. CBAM assigns the weight distribution of feature maps in the two dimensions of channel and space and increases the weight of important features, so as to make SID-AM-MSITM pay attention to meaningful information.
3. In response to the loss of content details, we further improve SID-AM-MSITM by constructing style-independent discriminators. The discriminators give similar results when discriminating images with the same content and different styles, so as to make SID-AM-MSITM focus on the content information instead of the style information.
4. We conduct systematical experiments based on multiple datasets, including submarine, underwater optics, sunken ship, crashed plane, and underwater sonar images. Compared with multiple baseline models, SID-AM-MSITM improves the ability to access effective information and retain content details.

The rest of this paper is organized as follows: Section 2 details the methodology of SID-AM-MSITM. Section 3 presents our ablation and comparative experiments, and their corresponding analysis. Section 4 concludes this work.

## 2. Methodology

In this section, we will present our proposed SID-AM-MSITM in detail. Based on TuiGAN as a backbone architecture, SID-AM-MSITM improves its generators and discriminators by introducing CBAM and makes further improvements by devising style-independent discriminators. Figure 1 shows the architecture of SID-AM-MSITM and the process of translating a non-underwater image into an underwater image.

In the underwater image translation task, the source domain image means the non-underwater image  $I_X$ , and the target domain image means the underwater image  $I_Y$ . We use SID-AM-MSITM to translate non-underwater images into underwater images, also known as generating underwater images. Using SID-AM-MSITM, we also reconstruct translated underwater images into non-underwater images. The original image represents the initial image that has not been processed by SID-AM-MSITM.



**Figure 1.** The architecture of SID-AM-MSITM (the process of translating a non-underwater image into an underwater image).

As in TuiGAN, generally, SID-AM-MSITM downsamples two images into different scales  $\{I_X^0, I_X^1, \dots, I_X^N, I_Y^0, I_Y^1, \dots, I_Y^N\}$ , and each scale corresponds to two generators  $\{G_{XY}^n, G_{YX}^n\}$  and two discriminators  $\{D_Y^n, D_X^n\}$ . The generators  $\{G_{XY}^0, G_{XY}^1, \dots, G_{XY}^{N-1}\}$  utilize the downsampled source domain images  $\{I_X^0, I_X^1, \dots, I_X^{N-1}\}$  as well as their previous-scale upsampled generated images  $\{I_{XY}^{1\uparrow}, I_{XY}^{2\uparrow}, \dots, I_{XY}^{N\uparrow}\}$  to generate new images. At the lowest scale  $N$ , the previous-scale upsampled generated image is replaced with an image with pixel values of 0.

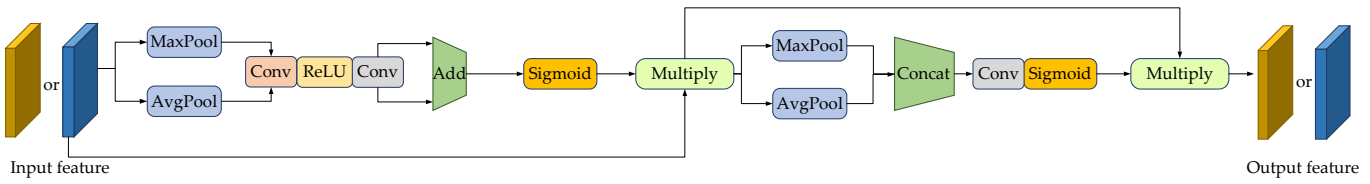
The discriminators  $\{D_Y^0, D_Y^1, \dots, D_Y^N\}$  learn the distribution of the target domain using a variety of loss functions for model training, including adversarial loss WGAN-GP [25], cycle-consistency loss, identity loss, and total variation loss [26]. Among them, the cycle-consistency loss helps SID-AM-MSITM to avoid the mode collapse, the identity loss helps it to align colors and textures, and TV loss smooths the generated images. Finally, the translated underwater image  $I_{XY}^0$  is obtained at the highest scale. In Figure 1, the discriminator processes images translated by a generator of the same color as it.

In order to obtain the generators  $\{G_{YX}^0, G_{YX}^1, \dots, G_{YX}^N\}$ , we also train the generators  $\{G_{YX}^0, G_{YX}^1, \dots, G_{YX}^N\}$  and discriminators  $\{D_X^0, D_X^1, \dots, D_X^N\}$  in a similar way.

### 2.1. Generators and Discriminators with CBAM Modules

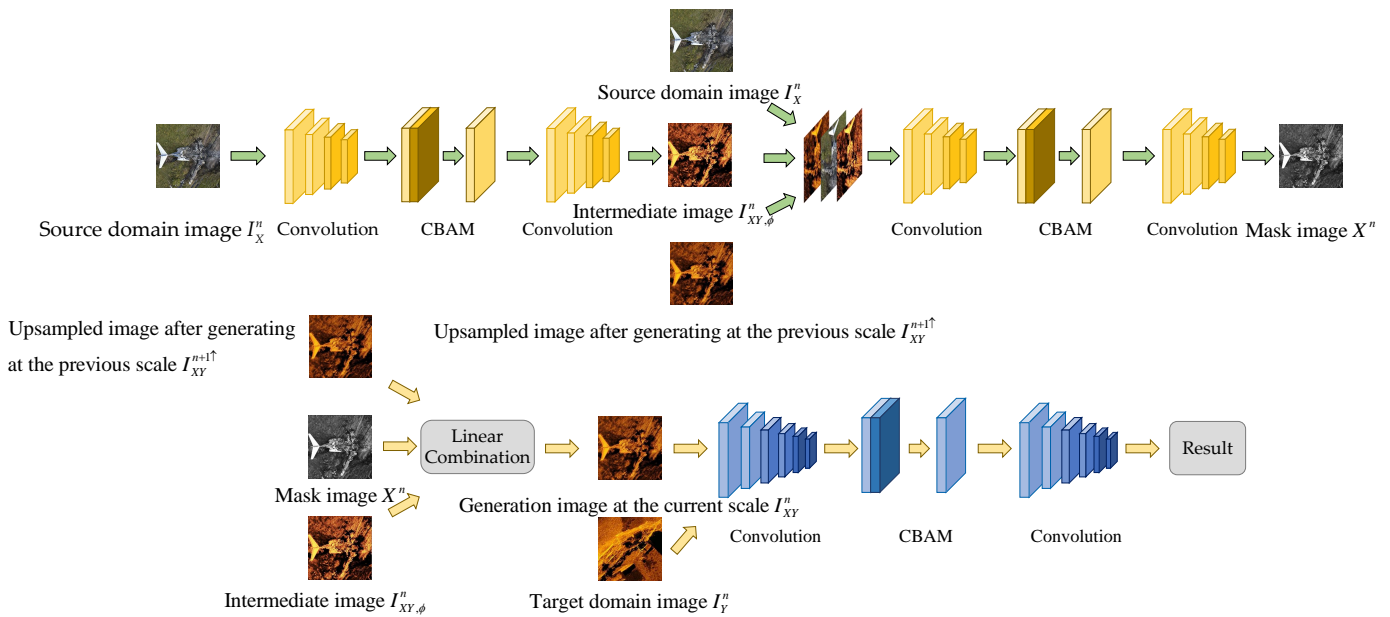
We start with presenting generators and discriminators with CBAM modules. CBAM enables SID-AM-MSITM to focus on critical features in a given image, so as to improve the quality of generated images [27].

As illustrated in Figure 2, each CBAM contains two modules, a channel attention module [28] and a spatial attention module. Specifically, the channel attention module enables SID-AM-MSITM to focus on critical features in a given image, so as to obtain accurate weights of channel features. The spatial attention module performs Max Pooling and Average Pooling on the spatial dimension of compressed features. The yellow blocks represent the input features of the source domain, and the blue blocks represent the input features of the target domain.



**Figure 2.** The structure of CBAM modules. CBAM represents the Convolution Block Attention Module. Conv represents convolution operations.

Figure 3 shows the generator and discriminator of SID-AM-MSITM, where the generator  $G_{XY}^n$  implements the translation of a source domain image  $I_X^n$  to a generated image  $I_{XY}^n$ . Firstly, SID-AM-MSITM simply processes  $I_X^n$  from source domain  $X$  to obtain an intermediate image  $I_{XY,\phi}^n$  through the CBAM module and convolution operations. Then, SID-AM-MSITM utilizes  $I_{XY,\phi}^n$ ,  $I_X^n$ , and an upsampled image after generating at the previous scale  $I_{XY}^{n+1\uparrow}$  to concatenate in the direction of the channel, and obtain a mask image  $X^n$  through the CBAM module and convolution operations. Finally, a generated image  $I_{XY}^n$  is obtained using the linear combination of  $X^n$ ,  $I_{XY,\phi}^n$ , and  $I_{XY}^{n+1\uparrow}$ .  $I_{XY}^n$  and the target domain image  $I_Y^n$  are input into the discriminator  $D_Y^n$  to obtain discriminant results for model training, where  $0 \leq n < N$ .



**Figure 3.** Generator and discriminator of SID-AM-MSITM (the process of translating a source domain image  $I_X^n$  to a generated image  $I_{XY}^n$ ).

$G_{XY}^n$  and  $G_{YX}^n$  share the same architecture but with different weights. The working principle of  $G_{XY}^n$  is as follows:

$$I_{XY,\phi}^n = \phi(I_X^n), \quad (1)$$

$$X^n = A^n(I_{XY,\phi}^n, I_X^n, I_{XY}^{n+1\uparrow}), \quad (2)$$

$$I_{XY}^n = X^n \otimes I_{XY}^{n+1\uparrow} + (1 - X^n) \otimes I_{XY,\phi}^n, \quad (3)$$

where  $0 \leq n < N$ ,  $\otimes$  represents pixel-level multiplication. At the lowest scale  $N$ ,  $I_{XY}^{n+1}$  is replaced with an image with pixel values of 0. First, SID-AM-MSITM uses the encoder  $\phi$  to preprocess  $I_X^n$  to  $I_{XY,\phi}^n$ . Then, SID-AM-MSITM uses the encoder  $A^n$  to generate mask  $X^n$ . Finally, SID-AM-MSITM uses the linear combination to obtain output  $I_{XY}^n$ .

Similarly, the implementation of the translation of  $I_Y \rightarrow I_{YX}$  at scale  $n$  is as follows:

$$I_{YX,\phi}^n = \phi'(I_Y^n), \quad (4)$$

$$Y^n = A'^n(I_{YX,\phi}^n, I_Y^n, I_{YX}^{n+1\uparrow}), \quad (5)$$

$$I_{YX}^n = Y^n \otimes I_{YX}^{n+1\uparrow} + (1 - Y^n) \otimes I_{YX,\phi}^n, \quad (6)$$

where  $0 \leq n < N$ . At the lowest scale  $N$ ,  $I_{YX}^{n+1\uparrow}$  is also replaced with an image with pixel values of 0.

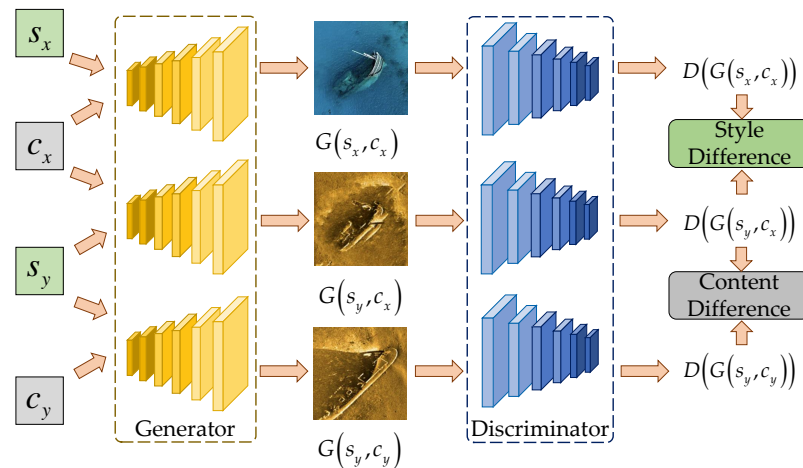
In this way, the generators focus on the regions that synthesize current scale details in the images. Meanwhile, it maintains the previously learned global structures as unaffected.

## 2.2. Style-Independent Discriminators

Next, we present our proposed style-independent discriminators, which focus on the images' content information rather than their style information, so as to enable SIM-AM-MSITM to avoid losing the content details in non-underwater images.

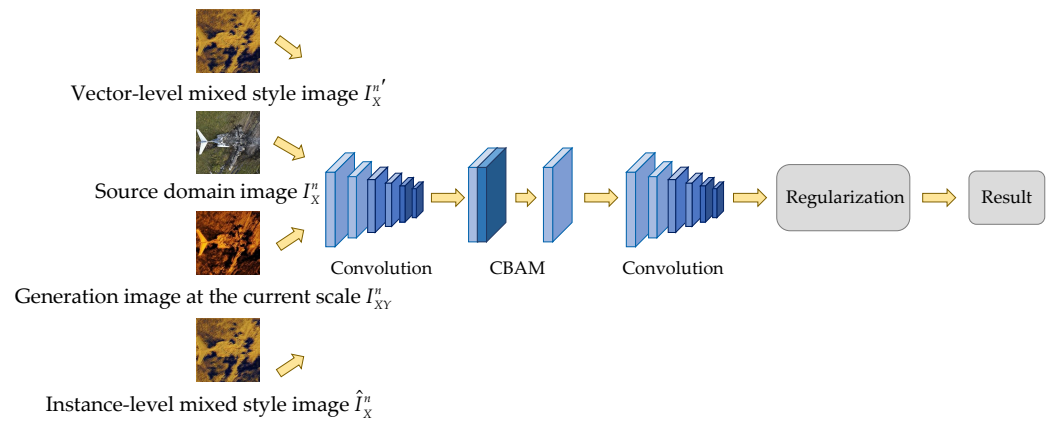
When two sets of images share the same content but different styles, it is ideal to make discriminators give similar discriminant results to the images. Thus, SID-AM-MSITM uses instance-level as well as vector-level style difference losses to train style-independent discriminators.

Figure 4 illustrates style and content differences. As illustrated in the figure, the first two generated images share the same content information, while the latter two share the same style information.  $s_x$  and  $s_y$  represent the style information of images  $x$  and  $y$  respectively.  $c_x$  and  $c_y$  represent the content information of images  $x$  and  $y$ , respectively.  $G(s, c)$  indicates the image generated using style information  $s$  and content information  $c$ .  $D(G(s, c))$  represents the discriminant result of  $G(s, c)$  given by a discriminator.



**Figure 4.** The illustration of style and content differences.

Figure 5 shows a style-independent discriminator  $D_Y^n$  of SIM-AM-MSITM, which requires a source domain image  $I_X^n$ , a generation image at current scale  $I_{XY}^n$ , an instance-level mixed style image  $\hat{I}_X^n$ , and a vector-level mixed style image  $I_X^{n\prime}$  to make the discriminator  $D_Y^n$  style-independent.



**Figure 5.** The training process of SID-AM-MSITM's style-independent discriminator  $D_Y^n$ .

Discriminator  $D_X^n$  has a similar structure and requires a target domain image  $I_Y^n$ , a generation image at current scale  $I_{YX}^n$ , an instance-level mixed style images  $\hat{I}_Y^n$ , and a vector-level mixed style image  $I_Y^{n'}$ , where  $0 \leq n \leq N$ .

Then, we will describe how style-independent discriminators eliminate style differences at the instance level as well as the vector level, respectively. Instance-level style difference refers to the style difference between the image obtained by stylizing its pixels and the original image. Vector-level style difference refers to the style difference between the image obtained by stylizing its encoded vectors and the original image.

### 2.2.1. Instance-Level Style-Independent Discriminators

Instance-level style-independent discriminators use a special regularization term, so as to reduce style differences between the images obtained by stylizing its pixels and the original images.

We first adjust weight  $\alpha$  to gradually increase the proportion of generated images among the mixed ones at multi-scales. Then, we make discriminators to reduce the differences between the original non-underwater images or the underwater images and the final mixed images, which are constrained by a consistency loss. Such progress is formulated as follows:

$$\hat{I}_X^n = \alpha I_X^n + (1 - \alpha) I_{XY}^n, 0 \leq n \leq N, \quad (7)$$

$$\mathcal{L}_{con} = \|D_Y^n(I_X^n) - D_Y^n(\hat{I}_X^n)\|_1, \quad (8)$$

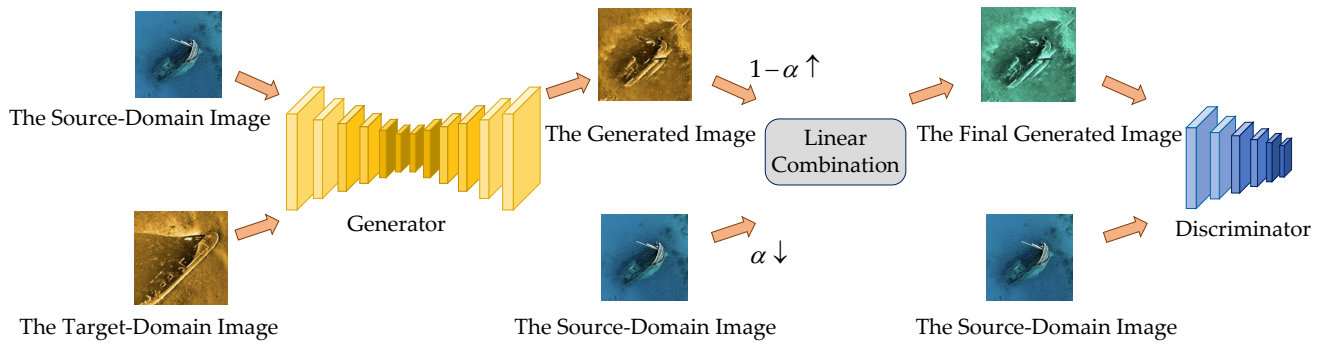
$$\hat{I}_Y^n = \alpha I_Y^n + (1 - \alpha) I_{YX}^n, 0 \leq n \leq N, \quad (9)$$

$$\mathcal{L}_{con} = \|D_X^n(I_Y^n) - D_X^n(\hat{I}_Y^n)\|_1, \quad (10)$$

where  $0 \leq n \leq N$ ,  $0 \leq \alpha < 1$ .  $I_X^n$  and  $I_Y^n$  represent the source-domain and target-domain images at the current scale, respectively.  $I_{XY}^n$  and  $I_{YX}^n$  denote the generation images at the current scale.  $\alpha$  indicates the weight of the linear combination and gradually becomes smaller as the scale rises.  $\mathcal{L}_{con}$  uses an  $L_1$  paradigm to process the instance-level style-independent loss, and  $D(\cdot)$  represents the discriminant results of discriminators.

As the scale rises, the style of instance-level mixed-style images  $\hat{I}_X^n$  is closer to the target-domain ones and away from the source-domain ones. The style of instance-level mixed-style images  $\hat{I}_Y^n$  is closer to the source-domain ones and away from the target-domain ones. Discriminators penalize the distances between the source-domain image outputs or the target-domain image outputs and the mixed-style image outputs.

Figure 6 shows the training process of instance-level style-independent discriminators  $D_Y$ .  $\uparrow$  and  $\downarrow$  indicate the value rising and descending as the scale rises, respectively.



**Figure 6.** The training process of instance-level style-independent discriminators.

### 2.2.2. Vector-Level Style-Independent Discriminators

Based on the above instance-level discriminators, it is not enough to generate images since this is limited by style-independent pixels. Therefore, we devise vector-level style-independent discriminators that further mix the encoded vectors of the source-domain and target-domain images at each scale. We put the mixed encoded vectors into a decoder, and utilize its generated images as well as the source-domain ones or the target-domain ones for model training.

First, we encode the source domain-images and the target-domain images, respectively using VGG 19 [29] and then process them using AdaIN [30]. The results obtained are linearly combined with the encoded vectors of the source- or target-domain images. Then, we put the results of the linear combinations into a decoder to get the vector-level mixed-style images  $I_X^{n'}$  and  $I_Y^{n'}$ . The decoder is a convolutional network that is symmetric to VGG 19 and upsamples the mixed encoded vectors into images. Finally, we utilize discriminators to penalize the distances between the source-domain image outputs or the target-domain image outputs and the mixed-style image outputs. Such progress is formulated as follows:

$$I_X^{n'} = \text{Decoder}(\alpha \text{Encoder}(I_X^n) + (1 - \alpha) \text{AdaIN}(\text{Encoder}(I_X^n), \text{Encoder}(I_Y^n))), \quad (11)$$

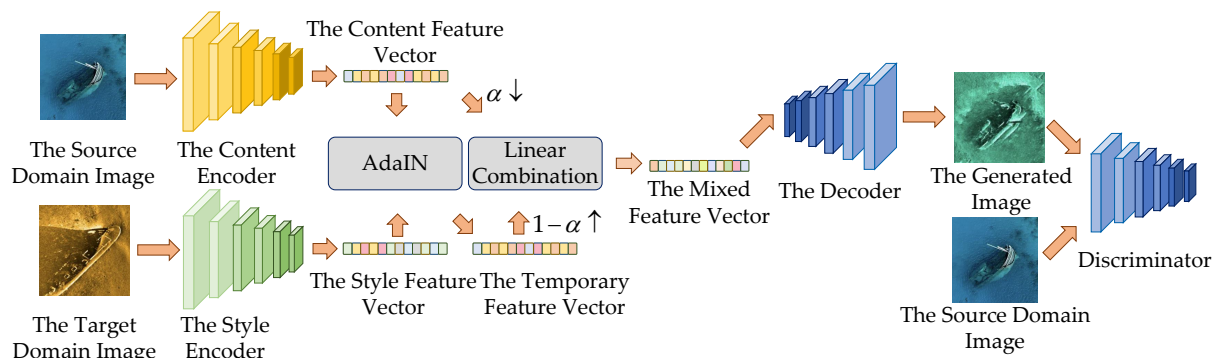
$$\mathcal{L}_{con} = \|D_Y(I_X^n) - D_Y(I_X^{n'})\|_1, \quad (12)$$

$$I_Y^{n'} = \text{Decoder}(\alpha \text{Encoder}(I_Y^n) + (1 - \alpha) \text{AdaIN}(\text{Encoder}(I_Y^n), \text{Encoder}(I_X^n))), \quad (13)$$

$$\mathcal{L}_{con} = \|D_X(I_Y^n) - D_X(I_Y^{n'})\|_1, \quad (14)$$

where  $0 \leq n \leq N$ ,  $0 \leq \alpha < 1$ .  $I_X^n$  and  $I_Y^n$  represent the source-domain and target-domain images at the current scale, respectively.  $\text{Encoder}(\cdot)$  is VGG 19 and  $\text{Decoder}(\cdot)$  is symmetric to VGG 19.  $\alpha$  is the weight coefficient that becomes smaller as the scale rises.  $\mathcal{L}_{con}$  uses an  $L_1$  paradigm to process the vector-level style-independent loss.  $D(\cdot)$  represents the discriminant results of discriminators.

Figure 7 shows the training process of vector-level style-independent discriminators.  $\uparrow$  and  $\downarrow$  indicate the value rising and descending as the scale rises, respectively.



**Figure 7.** The training process of vector-level style-independent discriminators.

### 2.3. Implementation

To implement SID-AM-MSITM, we utilize Adam [31] as its optimizer and LeakyReLU [32] as its activation function. At the lowest scale, images of the model are  $100 \times 100$ -pixel ones. And at the highest scale, the size of images is  $250 \times 250$  pixels. The model uses 6 scales.

## 3. Experiment and Result Analysis

In this section, we will present our experiments and the corresponding results analysis. We first introduce the evaluation metrics and then present the ablation and comparative experiments, respectively.

### 3.1. Evaluation Metric

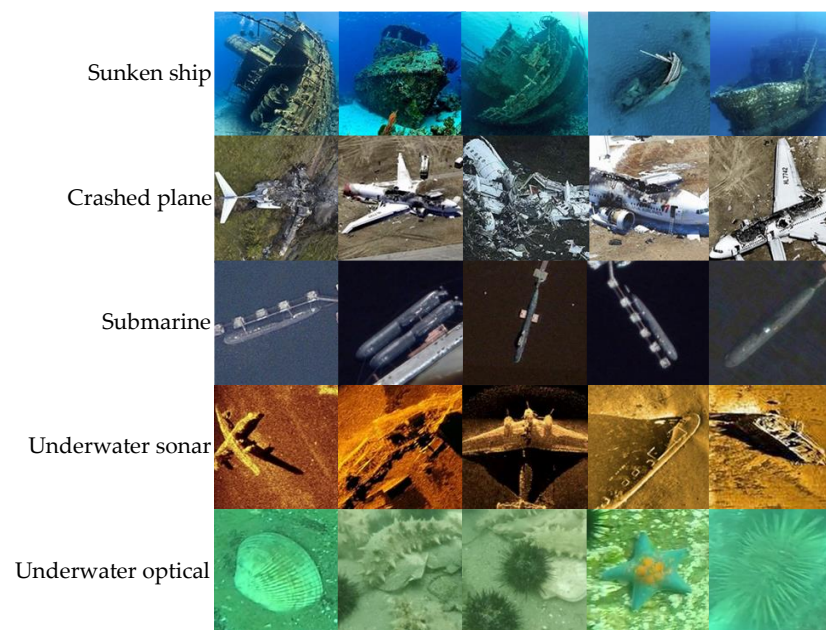
We utilize four metrics for quantitative evaluation, including peak signal-to-noise ratio (PSNR) [33], structure similarity index measure (SSIM) [34], information entropy (Entropy) [35], and single image Fréchet Inception distance (SIFID) [36].

- (1) PSNR: PSNR measures the distance between the distributions of two images. We use PSNR to calculate the distance between the source-domain images and reconstructed images. A larger PSNR value indicates a smaller difference between the two images.
- (2) SSIM: SSIM measures the similarity of two images. The value of SSIM is between 0 and 1, and a larger SSIM indicates a better reconstruction effect, which suggests the translation effect of an image translation model.
- (3) Entropy: Information entropy measures the complexity of an image. Larger information entropy indicates complex images that contain more information.
- (4) SIFID: Single Image Fréchet Inception Distance (SIFID) is a special type of Fréchet Inception distance (FID) [37]. It measures the deviation between the feature distribution of two single images, and smaller SIFID indicates the better effect of generated images.

### 3.2. Ablation Experiment

We use five different datasets, including submarine, underwater optics, sunken ship, crashed plane, and underwater sonar datasets. The underwater optics images are from the URPC2020 dataset [38]. These images are difficult to collect and are not large in number. Specifically, the submarine, sunken ship, and crashed plane images are content categories, and the underwater sonar and underwater optics images are style ones. Meanwhile, the submarine, sunken ship, and crashed plane images are non-underwater images, and the underwater sonar and underwater optics images are underwater ones.

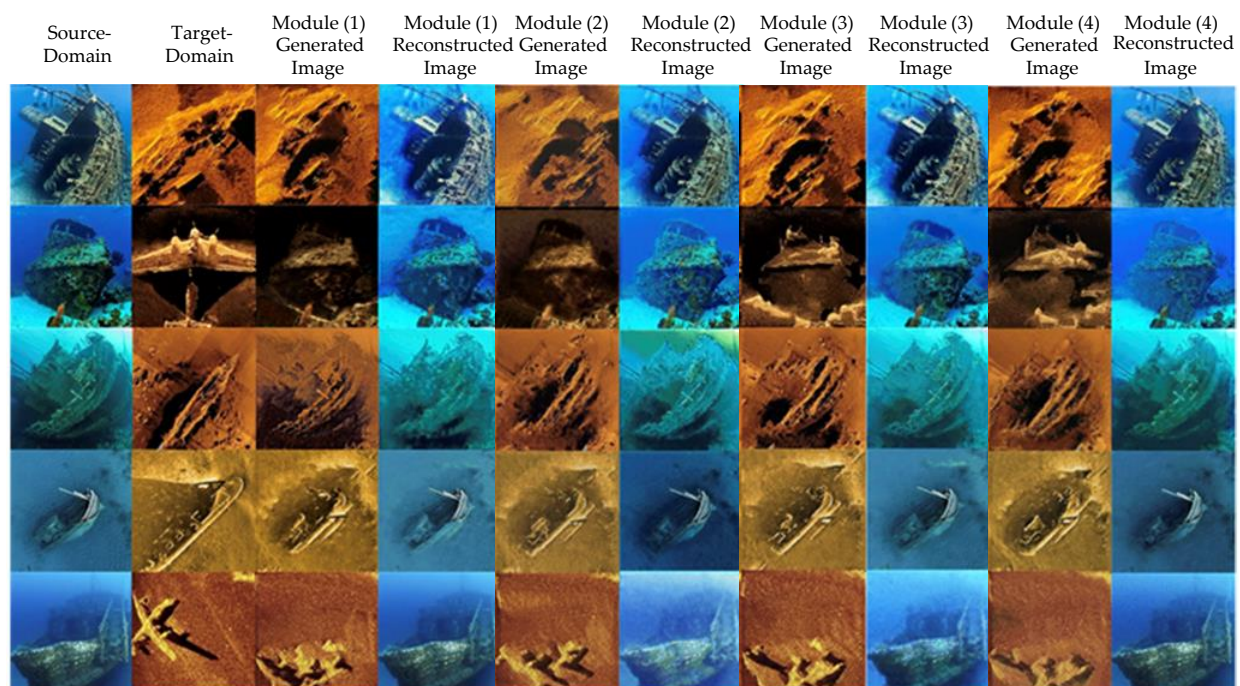
Figure 8 presents the utilized datasets, which are grouped into six different combinations, i.e., (1) sunken ships with underwater sonars, (2) sunken ships with underwater optics, (3) crashed planes with underwater sonars, (4) crashed planes with underwater optics, (5) submarines with underwater sonars, and (6) submarines with underwater optics.



**Figure 8.** Datasets.

Based on the above six combinations of images, we perform ablation experiments on four modules, including (1) TuiGAN only, (2) TuiGAN with CBAM, (3) TuiGAN with style-independent discriminators, and (4) SID-AM-MSITM, so as to comprehensively evaluate the improvements we have made.

Figures 9–14, respectively, present the generated and reconstructed images based on different combinations of datasets. In the figures, the first-column and second-column images are source-domain and target-domain ones, respectively. The third to the tenth columns, respectively, present the images generated or reconstructed using TuiGAN, TuiGAN with CBAM modules, TuiGAN with style-independent discriminators, and SID-AM-MSITM.



**Figure 9.** Generated and reconstructed images using the sunken ships and the underwater sonars.

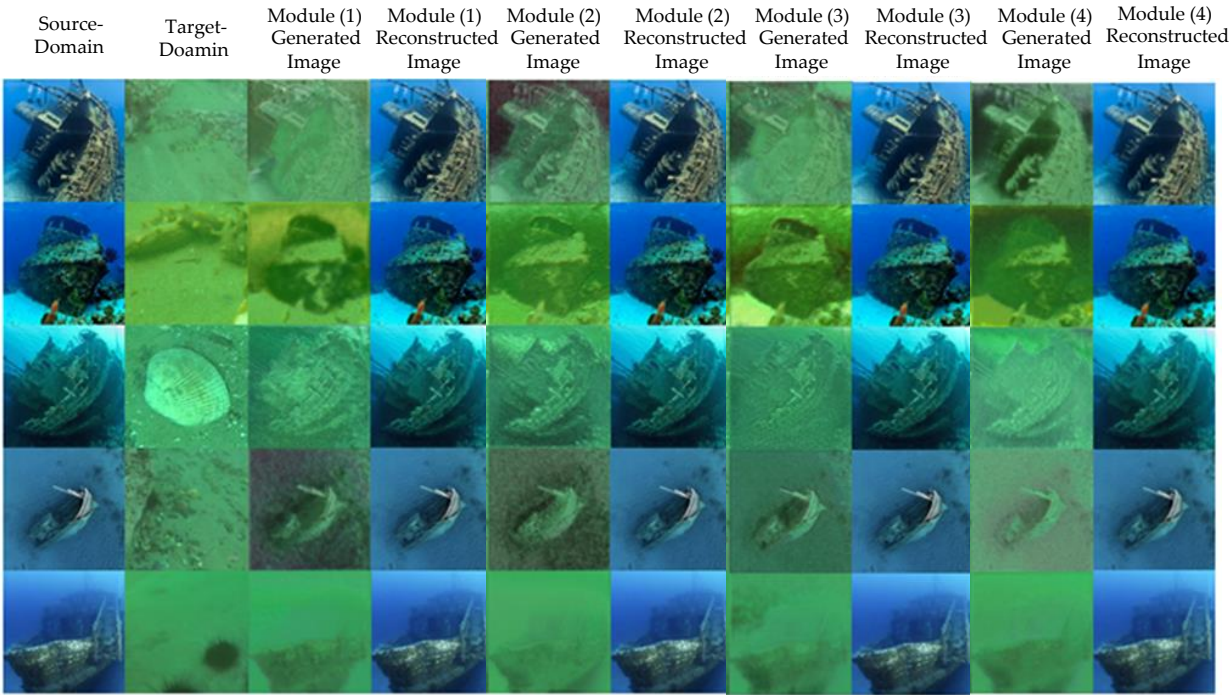


Figure 10. Generated and reconstructed images using the sunken ships and the underwater optics.

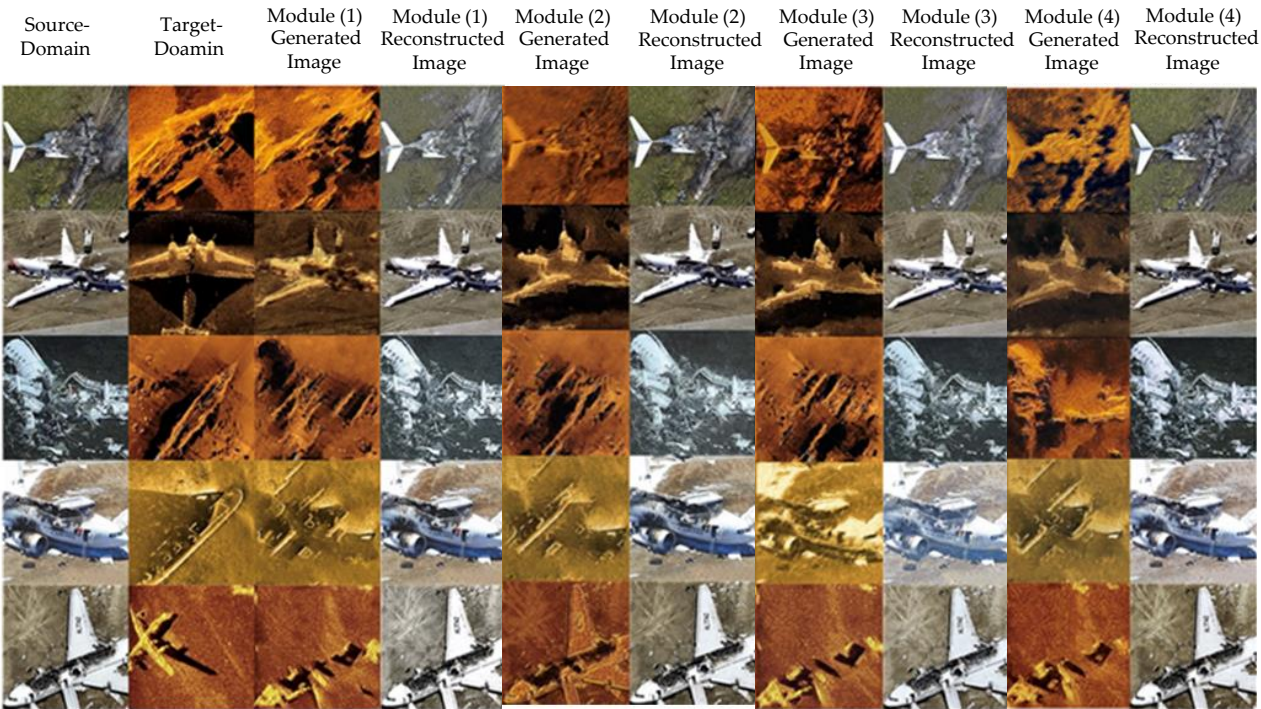


Figure 11. Generated and reconstructed images using the crashed planes and the underwater sonars.

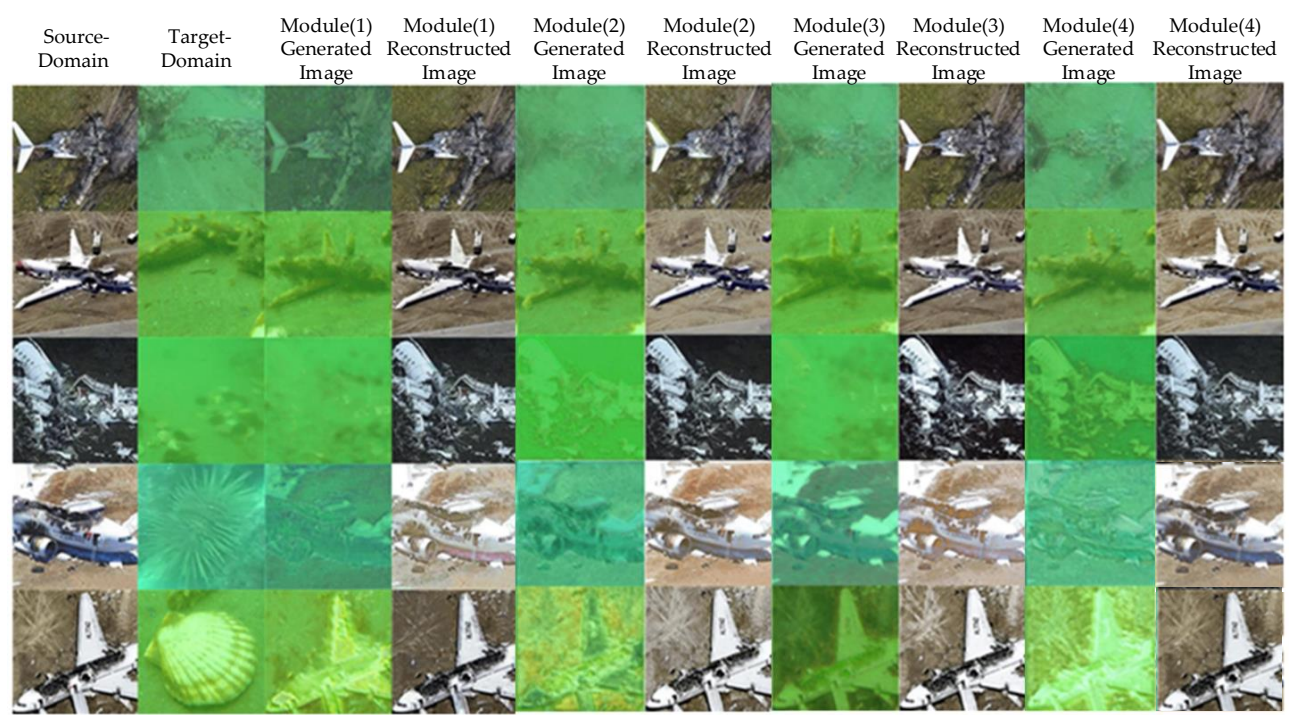


Figure 12. Generated and reconstructed images using the crashed planes and the underwater optics.

It is observed that SID-AM-MSITM has the ability to translate non-underwater images into underwater sonar images and underwater optics images, and objects such as submarines, crashed planes, and sunken ships are evident in underwater images. Meanwhile, SID-AM-MSITM is also capable of reconstructing translated underwater images into non-underwater images with little difference from the original non-underwater images.

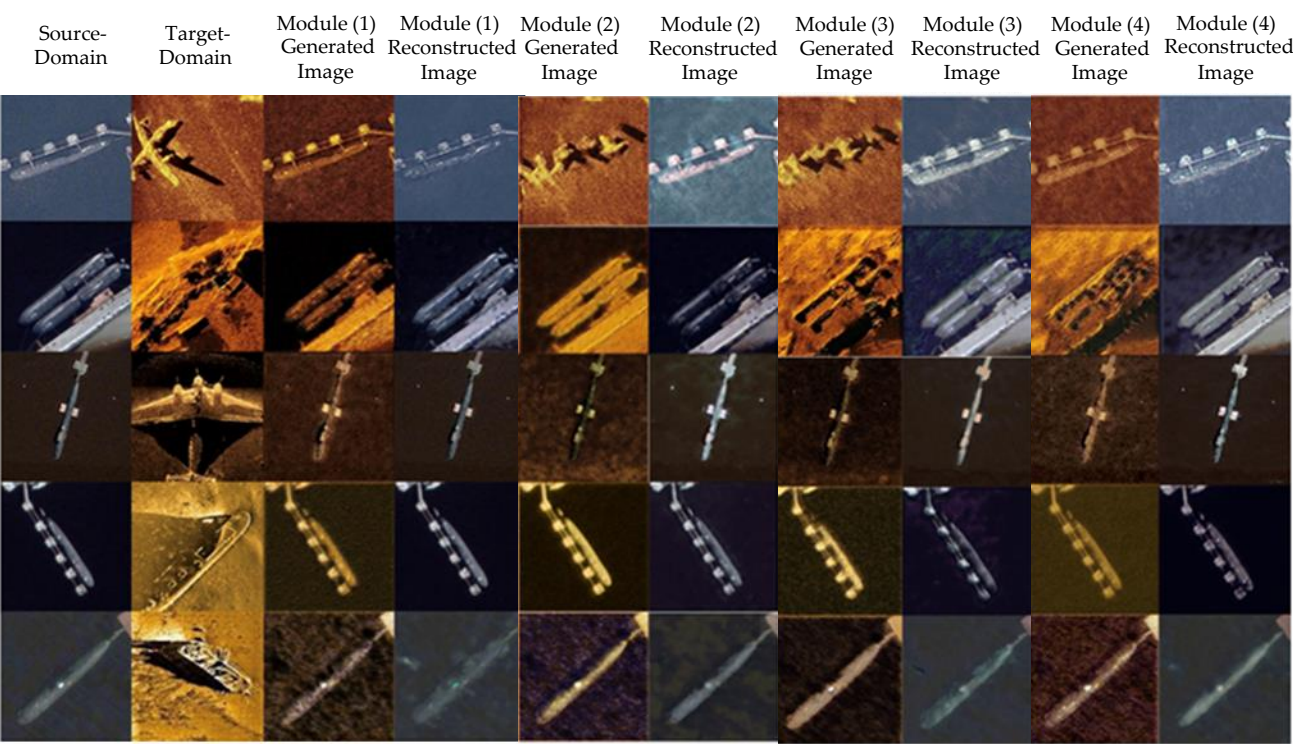
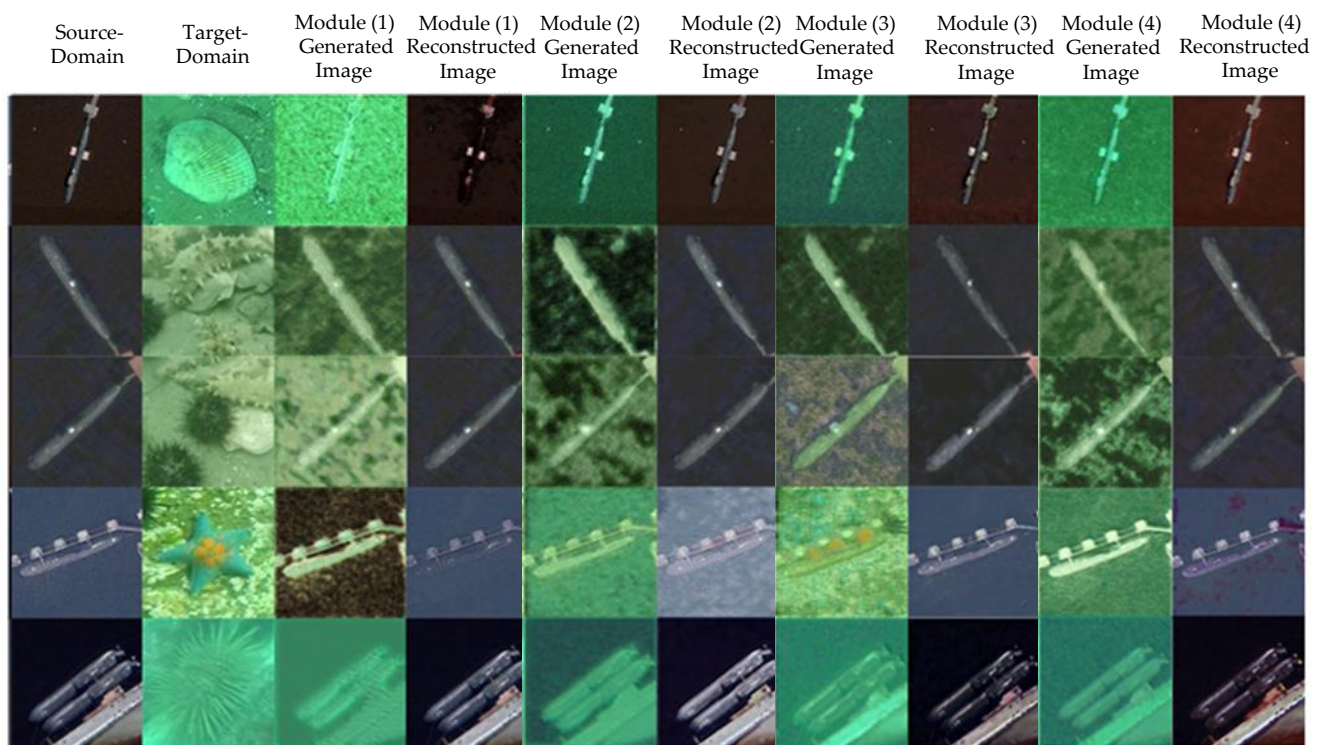


Figure 13. Generated and reconstructed images using the submarines and the underwater sonars.



**Figure 14.** Generated and reconstructed images using the submarines and the underwater optics.

Table 1 presents the PSNR results of each model. It is observed that the PSNR values of the images reconstructed using the TuiGAN with CBAM modules on the three combinations of the datasets are higher than that of TuiGAN, and the maximum difference reaches 8.63. The PSNR values of the images reconstructed using TuiGAN with style-independent discriminators on the five combinations of datasets are higher than that of TuiGAN, and the maximum difference reaches 4.7. The PSNR values of the images reconstructed using SID-AM-MSITM on all combinations of the datasets are not less than that of TuiGAN, and the maximum difference reaches 4.58. The promising results indicate that the combination of CBAM modules and style-independent discriminators significantly improves the effective information acquisition ability of backbone TuiGAN and is suitable for all combinations of datasets.

**Table 1.** Ablation experiments on different datasets (PSNR).

Category	TuiGAN	TuiGAN with CBAM Modules	TuiGAN with Style-Independent Discriminators	SID-AM-MSITM
Sunken Ship + Sonar	17.20	20.28	19.53	22.77
Sunken Ship + Optics	21.24	18.91	19.98	22.87
Crashed plane + Sonar	20.02	23.00	20.50	24.26
Crashed plane + Optics	26.46	25.04	29.17	26.46
Submarine + Sonar	27.27	25.88	31.11	31.85
Submarine + Optics	23.20	31.83	27.90	25.96

Table 2 presents the SSIM results of each model. SSIM is also a metric to measure the effect of model reconstruction. It is observed that the SSIM results of the images reconstructed using TuiGAN with CBAM modules on four combinations of datasets are higher than that of TuiGAN, and the maximum difference reaches 0.17. The SSIM results of the images reconstructed using TuiGAN with style-independent discriminators on all combinations of datasets are not less than that of TuiGAN, and the maximum difference reaches 0.09. The SSIM values of the images reconstructed using SID-AM-MSITM on all combinations of

datasets are higher than that of TuiGAN, and the maximum difference reaches 0.15. These promising results also indicate that CBAM modules and style-independent discriminators improve the ability to access effective information.

**Table 2.** Ablation experiments on different datasets (SSIM).

Category	TuiGAN	TuiGAN with CBAM Modules	TuiGAN with Style-Independent Discriminators	SID-AM-MSITM
Sunken Ship + Sonar	0.68	0.82	0.75	0.83
Sunken Ship + Optics	0.85	0.88	0.87	0.90
Crashed plane + Sonar	0.83	0.91	0.83	0.90
Crashed plane + Optics	0.91	0.89	0.92	0.92
Submarine + Sonar	0.85	0.84	0.88	0.89
Submarine + Optics	0.70	0.87	0.79	0.84

Table 3 presents the Entropy results of each model. It is observed that the Entropy result of the images generated using TuiGAN with CBAM modules on only one combination of datasets is higher than that of TuiGAN, and the difference reaches 0.21. The Entropy results of the images generated using TuiGAN with style-independent discriminators on four combinations of datasets are higher than that of TuiGAN, and the maximum difference reaches 0.86. The Entropy results of the images reconstructed using SID-AM-MSITM on all combinations of datasets are higher than that of TuiGAN, and the maximum difference reaches 0.78. These promising results indicate that style-independent discriminators improve the diversity of generated images. The style-independent discriminators improve TuiGAN's ability to retain content details and are suitable for the combinations of all datasets.

In summary, based on the above ablation results, our proposed SID-AM-MSITM achieves promising underwater image translation performance in terms of PSNR, SSIM, and Entropy. The ablation experiments demonstrate that CBAM modules enhance the feature extraction ability of the network, so as to enhance the ability to access effective information. Moreover, we prove that style-independent discriminators improve the diversity of the generated images without weakening the reconstruction performance, which indicates SID-AM-MSITM retains the content details of non-underwater images.

**Table 3.** Ablation experiments on different datasets (Entropy).

Category	TuiGAN	TuiGAN with CBAM Modules	TuiGAN with Style-Independent Discriminators	SID-AM-MSITM
Sunken Ship + Sonar	6.75	6.70	7.12	7.00
Sunken Ship + Optics	6.00	6.21	6.86	6.78
Crashed plane + Sonar	7.31	6.53	7.23	7.38
Crashed plane + Optics	5.57	5.53	5.70	5.63
Submarine + Sonar	6.47	5.54	6.96	6.59
Submarine + Optics	6.32	6.29	5.44	6.33

### 3.3. Comparative Experiment

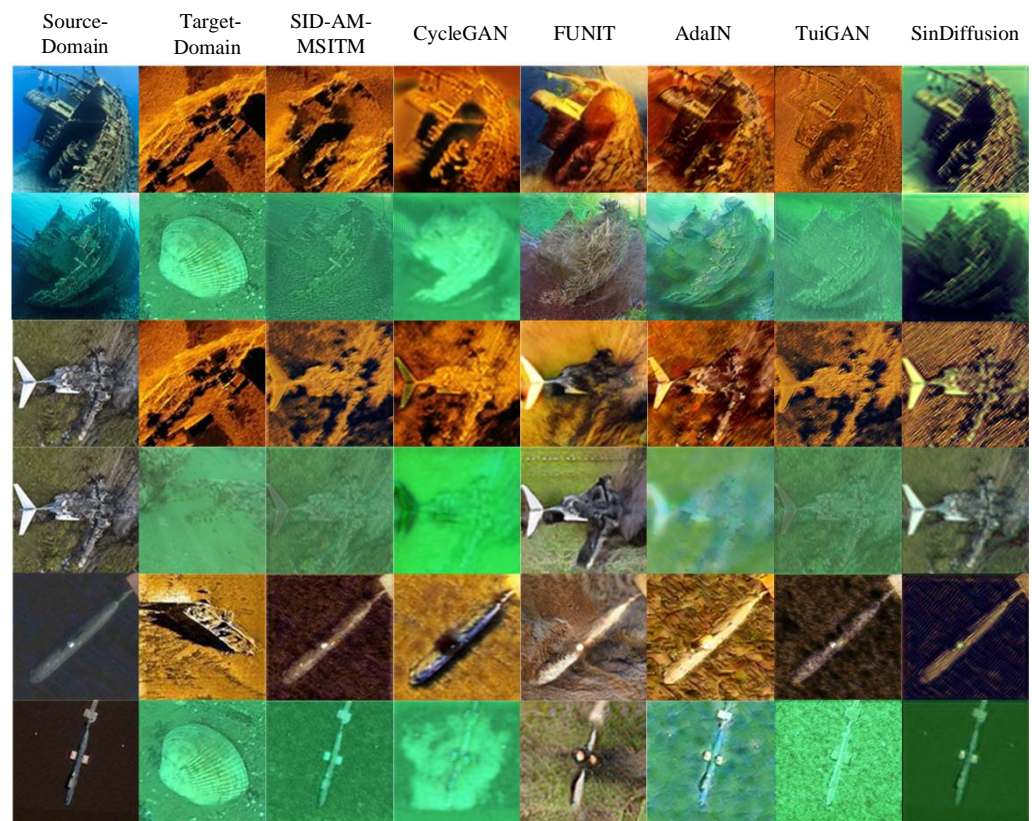
The above ablation experiments demonstrate the overall effect of SID-AM-MSITM. In the following, we further compare it with multiple advanced image translation models, including CycleGAN [39], FUNIT [40], AdaIN [30], and SinDiffusion [41]. These models are selected as baselines since they present promising performance and cover general image translation models, CycleGAN, FUNIT, and AdaIN, as well as the emerging SinDiffusion.

- (1) CycleGAN: CycleGAN is one of the most typical translation models using cycle consistency. The model assumes the potential correspondence between source-domain and target-domain images.

- (2) FUNIT: FUNIT is an unsupervised few-shot image translation model that achieves satisfactory performance based on limited data.
- (3) AdaIN: AdaIN is an image translation model that achieves real-time and arbitrary style transfer.
- (4) SinDiffusion: SinDiffusion is a diffusion model that works on a single natural image.

Figure 15 presents the comparison results between the images translated using SID-AM-MSITM and other baseline models. Through visual effect comparison, it is observed that SID-AM-MSITM has learned the style of target-domain images and retains the content of source-domain images. Moreover, the images translated using SID-AM-MSITM show little difference between adjacent pixels as well as excellent smoothness, which is superior to CycleGAN which shows obvious adjacent pixels difference after amplification. Compared with FUNIT and AdaIN, SID-AM-MSITM retains the content of source-domain images and learns better texture and color information from the target domain.

Next, we use SIFID to quantitatively compare SID-AM-MSITM with these baselines. Table 4 shows the SIFID results. It is observed that SID-AM-MSITM achieves the best (smallest) SIFID values. For example, the SIFID values of the images translated using SID-AM-MSITM are roughly  $0.02 \times 10^{-2}$  to  $0.058 \times 10^{-2}$  smaller than that of CycleGAN,  $17.408 \times 10^{-2}$  to  $17.55 \times 10^{-2}$  smaller than that of FUNIT,  $9.31 \times 10^{-2}$  to  $9.418 \times 10^{-2}$  smaller than that of AdaIN,  $0.002 \times 10^{-2}$  to  $0.018 \times 10^{-2}$  smaller than that of TuiGAN, and  $1.118 \times 10^{-2}$  to  $1.25 \times 10^{-2}$  smaller than that of SinDiffusion. This demonstrates that the images translated using SID-AM-MSITM are closer to the source-domain images and retain more content information than other models.



**Figure 15.** The comparison results of the images generated using SID-AM-MSITM and other models.

**Table 4.** SIFID comparison of underwater images generated using SID-AM-MSITM and other baseline models ( $\times 10^{-2}$ ).

Our Model	CycleGAN	FUNIT	AdaIN	TuiGAN	SinDiffusion
0.092	0.130	17.5	9.51	0.101	1.21
0.054	0.080	17.6	9.41	0.072	1.23
0.050	0.108	17.6	9.36	0.052	1.30
0.015	0.035	17.5	9.35	0.017	1.20

In summary, SID-AM-MSITM is superior to multiple baseline models in improving the ability to access effective information and avoiding the loss of content details.

#### 4. Conclusions

In this work, we propose a novel multi-scale image translation model with attention modules and style-independent discriminators (SID-AM-MSITM), to complete the underwater image translation task. We use a multi-scale generative adversarial network, TuiGAN, to construct a backbone architecture, which translates images from low scales to high scales. We introduce CBAM modules into the generators and discriminators at multi-scales and devise style-independent discriminators to improve the generative and discriminant effects. Based on systematical ablation and comparative experiments, we demonstrate that SID-AM-MSITM has the ability to acquire effective information and retain the content details of non-underwater images during the underwater image translation process, and it requires only two unpaired images to complete the image translation.

However, there are still some problems in the current research. In the use of style-independent discriminators, SID-AM-MSITM uses the number between 0 and 1 in the linear combination to achieve the translation from the source domain to the target domain. We will continue to study whether there is a more appropriate interval to train style-independent discriminators. We only select several source-domain and target-domain images as the dataset, which has certain limitations. In order to measure the performance of the model comprehensively, we will use other underwater target images to verify the versatility of SID-AM-MSITM, such as the UIEB database [42].

**Author Contributions:** Conceptualization, D.Y., T.Z. and M.L.; Data curation, M.L. and X.W.; Formal analysis, T.Z. and M.L.; Funding acquisition, D.Y. and X.W.; Investigation, M.L., W.C. and X.L.; Methodology, D.Y., T.Z., M.L. and X.W.; Project administration, B.L. and X.W.; Resources, M.L. and X.W.; Software, M.L. and W.C.; Supervision, B.L. and X.W.; Validation, T.Z. and M.L.; Visualization, M.L.; Writing—original draft, D.Y. and M.L.; Writing—review and editing, T.Z., B.L., W.C., X.L. and X.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by a grant from the Key Laboratory of Avionics System Integrated Technology, the Fundamental Research Funds for the Central Universities in China, Grant No. 3072022JC0601, and the Ministry of Industry and Information Technology High-tech Ship Project [2019] No. 331.

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available from the corresponding author upon reasonable request.

**Acknowledgments:** The authors are grateful to the editors and anonymous reviewers for their constructive comments.

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

## Abbreviations

The following abbreviations are used in this manuscript:

SID-AM-MSITM	multi-scale image translation model based on style-independent discriminators and attention modules
CBAM	convolution block attention module
GAN	generative adversarial network
CNN	convolutional neural network
LeakyReLU	leaky rectified linear unit
CycleGAN	cycle-consistent adversarial network
TV loss	total variation loss
PSNR	peak signal-to-noise ratio
SSIM	structure similarity index measure
Entropy	information entropy
AdaIN	adaptive instance normalization
CycleGAN	cycle-consistent adversarial networks
FID	the Fréchet Inception distance
SIFID	single image Fréchet Inception distance

## References

1. Zhao, Y.; Zhu, K.; Zhao, T.; Zheng, L.; Deng, X. Small-Sample Seabed Sediment Classification Based on Deep Learning. *Remote Sens.* **2023**, *15*, 2178. [\[CrossRef\]](#)
2. Chen, B.; Li, R.; Bai, W.; Zhang, X.; Li, J.; Guo, R. Research on recognition method of optical detection image of underwater robot for submarine cable. In Proceedings of the 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 11–13 October 2019; pp. 1973–1976.
3. Teng, B.; Zhao, H. Underwater target recognition methods based on the framework of deep learning: A survey. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420976307. [\[CrossRef\]](#)
4. Cruz, L.; Lucio, D.; Velho, L. Kinect and rgbd images: Challenges and applications. In Proceedings of the IEEE 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials, Ouro Preto, Brazil, 22–25 August 2012; pp. 36–49.
5. Yang, L.; Wang, B.; Zhang, R.; Zhou, H.; Wang, R. Analysis on location accuracy for the binocular stereo vision system. *IEEE Photonics J.* **2017**, *10*, 1–16. [\[CrossRef\]](#)
6. Lin, E. Comparative Analysis of Pix2Pix and CycleGAN for Image-to-Image Translation. *Highlights Sci. Eng. Technol.* **2023**, *39*, 915–925. [\[CrossRef\]](#)
7. Multi-view underwater image enhancement method via embedded fusion mechanism. *Eng. Appl. Artif. Intell.* **2023**, *121*, 105946. [\[CrossRef\]](#)
8. Zhou, J.; Liu, Q.; Jiang, Q.; Ren, W.; Lam, K.M.; Zhang, W. Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction. *Int. J. Comput. Vis.* **2023**. [\[CrossRef\]](#)
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 8–13 December 2014; Volume 27.
10. Hertzmann, A.; Jacobs, C.E.; Oliver, N.; Curless, B.; Salesin, D.H. Image analogies. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 12–17 August 2001; pp. 327–340.
11. Resales; Achan; Frey. Unsupervised image translation. In Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, 14–17 October 2003; pp. 472–478.
12. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
13. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
14. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
15. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
16. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
17. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.

18. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394. [[CrossRef](#)]
19. Wang, N.; Zhou, Y.; Han, F.; Zhu, H.; Yao, J. UWGAN: Underwater GAN for real-world underwater color restoration and dehazing. *arXiv* **2019**, arXiv:1912.10269.
20. Li, N.; Zheng, Z.; Zhang, S.; Yu, Z.; Zheng, H.; Zheng, B. The synthesis of unpaired underwater images using a multistyle generative adversarial network. *IEEE Access* **2018**, *6*, 54241–54257. [[CrossRef](#)]
21. Zhou, J.; Li, B.; Zhang, D.; Yuan, J.; Zhang, W.; Cai, Z.; Shi, J. UGIF-Net: An Efficient Fully Guided Information Flow Network for Underwater Image Enhancement. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 1–17. [[CrossRef](#)]
22. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* **2018**, arXiv:1811.12231.
23. Lin, J.; Pang, Y.; Xia, Y.; Chen, Z.; Luo, J. Tuigan: Learning versatile image-to-image translation with two unpaired images. In Proceedings of the 16th European Conference of Computer Vision (ECCV 2020), Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16; Springer: Cham, Switzerland, 2020; pp. 18–35.
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
25. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
26. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.
27. You, Q.; Wan, C.; Sun, J.; Shen, J.; Ye, H.; Yu, Q. Fundus image enhancement method based on CycleGAN. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 4500–4503.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Xu, J.; Li, Z.; Du, B.; Zhang, M.; Liu, J. Reluplex made more practical: Leaky ReLU. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020; pp. 1–7.
33. Najafipour, A.; Babaee, A.; Shahrtash, S.M. Comparing the trustworthiness of signal-to-noise ratio and peak signal-to-noise ratio in processing noisy partial discharge signals. *IET Sci. Meas. Technol.* **2013**, *7*, 112–118. [[CrossRef](#)]
34. Khadtare, M.S. GPU based image quality assessment using structural similarity (SSIM) index. In *Emerging Research Surrounding Power Consumption and Performance Issues in Utility Computing*; IGI Global: Hershey, PA, USA, 2016; pp. 276–282.
35. Xu, N.; Zhuang, J.; Xiao, J.; Peng, C. Regional Differential Information Entropy for Super-Resolution Image Quality Assessment. *arXiv* **2021**, arXiv:2107.03642.
36. Shaham, T.R.; Dekel, T.; Michaeli, T. Singan: Learning a generative model from a single natural image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4570–4580.
37. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
38. Zhang, J.; Zhang, J.; Zhou, K.; Zhang, Y.; Chen, H.; Yan, X. An Improved YOLOv5-Based Underwater Object-Detection Framework. *Sensors* **2023**, *23*, 3693. [[CrossRef](#)] [[PubMed](#)]
39. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
40. Liu, M.Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10551–10560.
41. Wang, W.; Bao, J.; Zhou, W.; Chen, D.; Chen, D.; Yuan, L.; Li, H. Sindiffusion: Learning a diffusion model from a single natural image. *arXiv* **2022**, arXiv:2211.12445.
42. Zhou, J.; Pang, L.; Zhang, D.; Zhang, W. Underwater Image Enhancement Method via Multi-Interval Subhistogram Perspective Equalization. *IEEE J. Ocean. Eng.* **2023**, *48*, 474–488. [[CrossRef](#)]

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