applied sciences

MDPI

Article

Horizon Targeted Loss-Based Diverse Realistic Marine Image Generation Method Using a Multimodal Style Transfer Network for Training Autonomous Vessels

Jisun Park, Tae Hyeok Choi and Kyungeun Cho *

Department of Multimedia Engineering, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea; jisun@dongguk.edu (J.P.); xogur6889@dgu.ac.kr (T.H.C.)
* Correspondence: cke@dongguk.edu; Tel.: +82-2-2260-3834

Abstract: Studies on virtual-to-realistic image style transfer have been conducted to minimize the difference between virtual simulators and real-world environments and improve the training of artificial intelligence (AI)-based autonomous driving models using virtual simulators. However, when applying an image style transfer network architecture that achieves good performance using land-based data for autonomous vehicles to marine data for autonomous vessels, structures such as horizon lines and autonomous vessel shapes often lose their structural consistency. Marine data exhibit substantial environmental complexity, which depends on the size, position, and direction of the vessels because there are no lanes such as those for cars, and the colors of the sky and ocean are similar. To overcome these limitations, we propose a virtual-to-realistic marine image style transfer method using horizon-targeted loss for marine data. Horizon-targeted loss helps distinguish the structure of the horizon within the input and output images by comparing the segmented shape. Additionally, the design of the proposed network architecture involves a one-to-many style mapping technique, which is based on the multimodal style transfer method to generate marine images of diverse styles using a single network. Experiments demonstrate that the proposed method preserves the structural shapes on the horizon more accurately than existing algorithms. Moreover, the object detection accuracy using various augmented training data was higher than that observed in the case of training using only virtual data. The proposed method allows us to generate realistic data to train AI models of vision-based autonomous vessels by actualizing and augmenting virtual images acquired from virtual autonomous vessel simulators.

Keywords: style transfer; autonomous vessels; horizon targeted loss

Citation: Park, J.; Choi, T.H.; Cho, K. Horizon Targeted Loss-Based Diverse Realistic Marine Image Generation Method Using a Multimodal Style Transfer Network for Training Autonomous Vessels. *Appl. Sci.* **2022**, *12*, 1253. https://doi.org/10.3390/app12031253

Academic Editor: Francesco Bianconi

Received: 8 October 2021
Accepted: 21 January 2022
Published: 25 January 2022

1. Introduction

Recent technological advances in artificial intelligence (AI) have led to improvements in the field of autonomous driving. Autonomous vehicles learn diverse scenarios using open-source simulators, such as AirSim [1], CARLAR [2], and SVL Simulator [3], by training terrain and weather condition variables to respond in a manner that reflects real driving environments. The trained AI model is subsequently mounted on a real vehicle. Research similar to that of autonomous vehicles is now underway with respect to marine vessels, considering the effect of waves, buoyancy, water currents, and wind currents. The automation of navigation necessitates a robot to control the rudder and sails, steering the sailing yacht, making tactical decisions regarding the sailing routes, and performing docking maneuvers in ports. Virtual simulators of autonomous vessels, such as Freefloating Gazebos [4], VREP [5], RobotX Simulator [6], and USVSim [7], when used to simulate rudder adjustments according to the wind and tide, face the problem of a very low level of representativeness of marine graphics. Consequently, the performance of AI models trained to perform vision-based object tracking or pathfinding in a virtual ocean environment deteriorates when mounted in a real environment, owing to the differences

between the simulated and real environments. Changing the environment that is already built into a new style is limited because it requires manual effort.

Therefore, we propose a diverse realistic marine image generation method that uses virtual images. Figure 1 depicts the conversion of image data obtained from a virtual simulator to provide realistic images pertaining to a variety of marine environments. Furthermore, this study generates realistic data that can train an artificially intelligent and vision-based autonomous vessel model by enhancing the virtual images of various styles obtained through the virtual autonomous vessel simulator, which is suitable for a photo-real world.

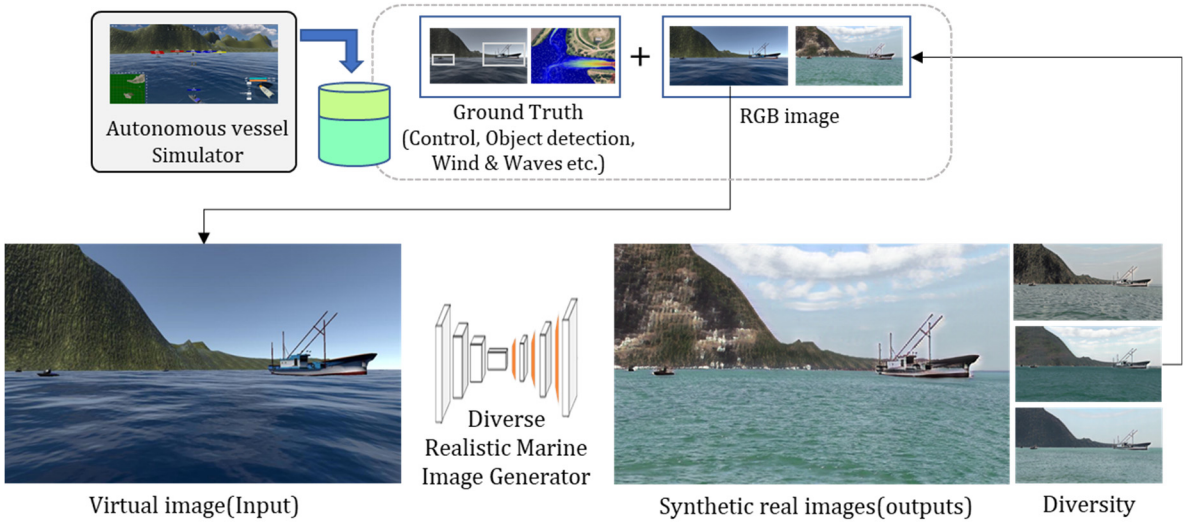


Figure 1. Concept of realistic generation of diverse marine images for training autonomous vessels.

Existing virtual simulators for autonomous vehicles have achieved good performance in transforming data from virtual images to those in realistic images through the generative adversarial network (GAN) [8], such as the use of virtual environment images obtained in the GTA game simulator and real-world driving data from Cityscapes [9]. However, when AI model architectures developed using images from land-based data are applied to marine images, the structural shapes within the images often collapse. On land, the locations and orientations of roads and vehicles are uniform; therefore, land-based images are nearly identical in structure; however, in the case of the ocean, the locations and orientations of the vessels differ widely owing to the absence of lanes. In addition, it is easy for neural networks to extract and learn features concerning land-based images because the color differences between elements, such as the sky, road, and vehicle, are evident. However, it is difficult to differentiate between the colors of the sky and sea in ocean-based images because they may comprise a range of similar blue colors, such as in the images depicted in Figure 2.

(a) High structural diversity (b) Low color difference

Figure 2. Examples of marine data wherein image-feature extraction is difficult; (**a**) example of high structural diversity, and (**b**) example of low color difference.

To overcome these limitations, we propose a method for the realistic generation of diverse marine images based on horizon-targeted loss that can preserve the shapes of the horizon and the vessel. Horizon-targeted loss calculates the difference in structural forms in the input and output images in the detected vessel areas based on the horizon and reflects them in the loss function, thereby enabling the network to prevent loss of the structural form of the relatively complex horizon and vessel. Moreover, this loss enhances the AI-based learning performance of autonomous vessels by enhancing the marine image extracted by this network from a virtual simulator such that it resembles a variety of realistic images.

The main contributions of this study are summarized as follows.

- A specialized style transfer method for marine images is proposed.
- A novel horizon-targeted loss is designed to enhance and preserve the shapes of the ocean horizon and the vessel.
- The accuracy of representing structural forms is improved through the style transfer of marine data.
- A method for the generation of diverse and realistic marine images is proposed that uses one-to-many style mapping and is based on multimodal style transfer.

The remainder of this paper is organized as follows. Related works concerning autonomous vessel simulators and image-style transfers are outlined in Section 2. The proposed diverse realistic marine image generation framework is introduced in Section 3. The experimental results and analyses are presented in Section 4. Finally, the proposed framework is presented in Section 5.

2. Related Works

This section summarizes existing studies concerning simulators of autonomous vessels and style-transfer approaches. Subsequently, the necessity of the proposed diverse and realistic marine image generation method is explained.

2.1. Simulators of Autonomous Vessels

Autonomous vessels are currently used for applications such as search-and-rescue operations, inspecting bridge structures, maintaining security, and monitoring the environment. Most autonomous AI testing is conducted through simulators focused on modeling physical conditions, such as waves, buoyancy, water currents, and wind currents, and they do not account for realistic visual images obtained from RGB sensors. However, visual images play an important role in autonomous vessel control. For example, object detection models can be used to identify other boats or objects in images; furthermore, there is potential for the utilization of segmentation models in detecting shorelines or separating water from other structures. As listed in Table 1, autonomous-vessel simulators [4–7] can simulate vessel control using waves, buoyancy, water currents, and wind-current conditions, but the visual representativeness of these simulators in terms of reality is low, and they do not incorporate diversity. Therefore, in this study, we propose a virtual-to-realistic marine image style transfer method to provide diverse and realistic marine RGB images by converting virtual images to real images, which overcomes the limitations of the lack of visual quality and diversity exhibited by existing autonomous vessel simulators.

Table 1. Existing simulators for autonomous vessels.

Simulator	Waves	Buoyancy	Water Currents	Wind Currents	Camera (RGB)
Freefloating Gazebo [4]	O	O	O	X	O
VREP [5]	O	O	X	X	O
RobotX Simulator [6]	O	O	X	O	O
USVSim [7]	O	O	O	O	O

2.2. Image-to-Image Translation for Style Transfer

Image-to-image translation has been studied extensively and applied in various fields such as photorealistic transfer [8], semantic synthesis [9], and data augmentation [10]. This section introduces the existing image-to-image translation approaches for style transfer and is organized in the following order: Section 2.2.1. Supervised-learning Approach, Section 2.2.2. Unsupervised-learning Approach for a Single Style, and Section 2.2.3. Unsupervised-learning Approach for Multiple Styles.

2.2.1. Supervised-Learning Approach

Since the creation of the generative adaptive network (GAN) [11] algorithm, studies focusing on image-to-image translation using urban scene dataset [12] such as Pix2pix [13] and SPADE [14] have been conducted actively. Supervised image-to-image translation aims to translate source images into a target domain using many aligned image pairs as the source and target domains during training. However, the paired dataset used in supervised learning cannot be built to reflect new styles. For example, in a realistic style transfer from a virtual image, it is impossible to obtain the same structure as that of the corresponding real-world image. To address this, unsupervised learning techniques, such as CycleGAN [15], have been proposed.

2.2.2. Unsupervised-Learning Approach for a Single Style

Unlike Pix2pix [13] and SPADE [14], CycleGAN [15] does not require paired data from the X and Y domains for training. Using adversarial loss, the data from the X domain were mapped to the Y domain without a paired dataset. $G_{xy} : X \rightarrow Y$ is trained such that the distribution of data from $G_{xy}(x)$ is identical to that from Y. Furthermore, $G_{yx}(y)$ works in conjunction with reverse mapping, given by $G_{yx} : Y \rightarrow X$, and introduces cycle-consistency loss that forces $G_{yx}\left(G_{xy}(x)\right)$ to resemble X.

Similar to the concept of dual learning in image translation, DualGAN [16] and DiscoGAN [17] have been proposed to train two cross-domain transfer GANs with two cyclic losses simultaneously. However, these networks can map only one style per network. To solve this problem, research concerning multimodal image-to-image translation has been conducted using disentangled representations.

2.2.3. Unsupervised-Learning Approach for Multiple Styles

The main concept involved in multimodal image-to-image translation learning for diverse style transfer is the conversion of one picture into a variety of images, which contrasts with a one-to-one mapping process. Disentangled representations [18–20] present a solution to the one-to-one domain mapping problem. They have facilitated advances in multimodal image-to-image translation, notably through the DRIT [21] and MUNIT [22] methods, which assume that image data from different domains can be mapped to a single identical latent space.

DRIT [21] and MUNIT [22] assume that the latent feature space for the image data is composited into contents and styles, wherein the content is shared regardless of the domain, and the style is domain-specific. Through adversarial loss, the multimodal approach applies weight-sharing and a discriminator to force feature representations to be mapped onto the same shared space, as well as to guarantee that the corresponding feature representations encode the same information for both domains. With the cross-cycle consistency loss, the multimodal approach implements a forward–backward function by swapping domain representations. On training completion, the networks can input a different-attribute feature vector randomly sampled from the specific attribute space to generate diverse outputs.

However, these methods are unable to solving problems involved in marine images. Because of the high structural diversity and low color difference of the area near the horizon, errors in identifying structural forms occur frequently. Therefore, we propose a horizon-targeted loss based on the MUNIT [22] architecture to reduce the error in identifying the

horizon line and vessel shapes. The proposed method can generate diverse realistic images from virtual RGB images without losing shape coherence in oceanic data, such that various realistic images and ocean conditions can be provided to autonomous vessels.

3. Proposed Diverse and Realistic Marine Image Generation Framework

We propose a framework for diverse and realistic marine image generation using horizon targeted loss to enhance and preserve the shapes of the ocean horizon and vessels for generating realistic training data for anonymous vessels. This framework is divided into two parts, as illustrated in Figure 3. First, the content and style features are extracted from the virtual input image to the generator using content and style encoders, and realistic synthetic images are generated by the decoder based on the content and style features. Secondly, we reduce the loss by detecting the shape of vessels and extracting horizon regions to minimize the structural errors of the ocean horizon and vessels caused by the complexity of marine images.

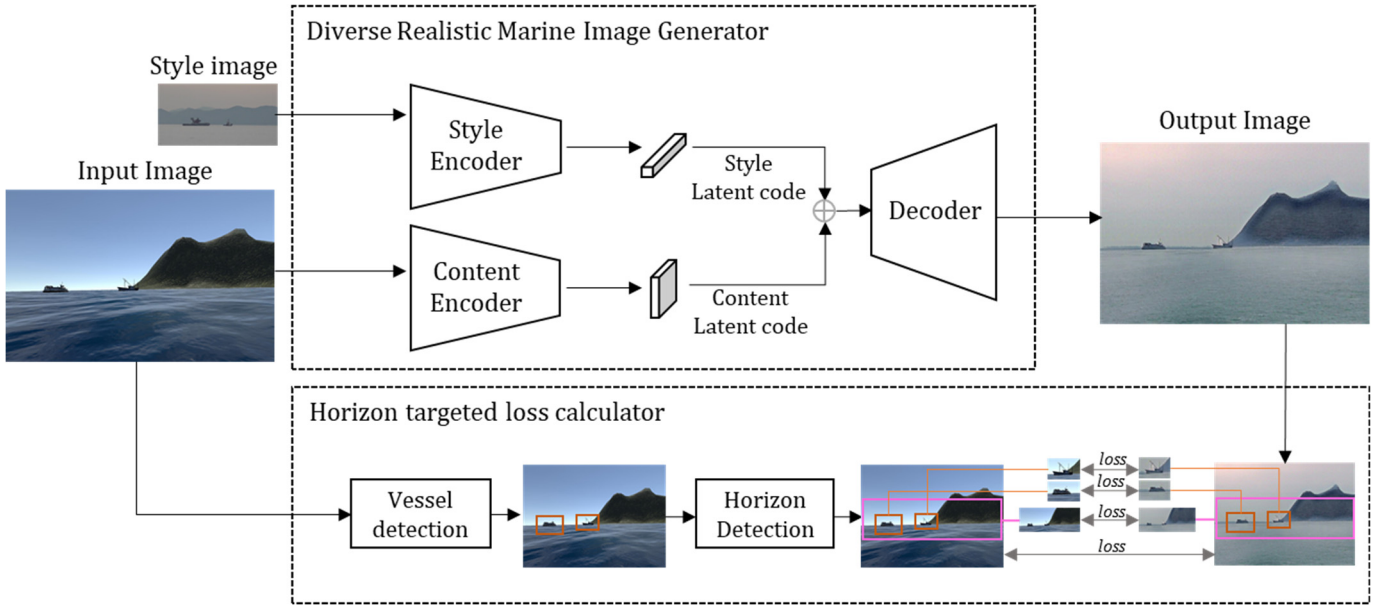


Figure 3. Architecture of the proposed diverse realistic marine image generation framework.

3.1. Diverse Realistic Marine Image Generator

We customized the MUNIT [22] by adding horizon-targeted loss to enable the generator to transform a virtual image into a corresponding realistic image. When training the generator, each feature is extracted using the style encoders $E_{style}^{virtual}$ and E_{style}^{real}, and the content encoders, $E_{con}^{virtual}$ and E_{con}^{real}, in the virtual and real-world domains, respectively. The extracted features intersect with $G_{virtual}\left(E_{style}^{real}(x_{real}),\ E_{content}^{virtual}(x_{virtual})\right)$ and $G_{real}\left(E_{style}^{virtual}(x_{virtual}),\ E_{content}^{real}(x_{real})\right)$ through the decoder to synthesize each content and style aspect, wherein the content is shared regardless of the domain, and the style is domain-specific. To generate the stylized image output, the intermediate result is once again provided as input to the encoder and decoder and subsequently synthesized to return to the style of the input image. The network is trained by calculating the loss based on the differences between the output and input images, as illustrated in Figure 4.

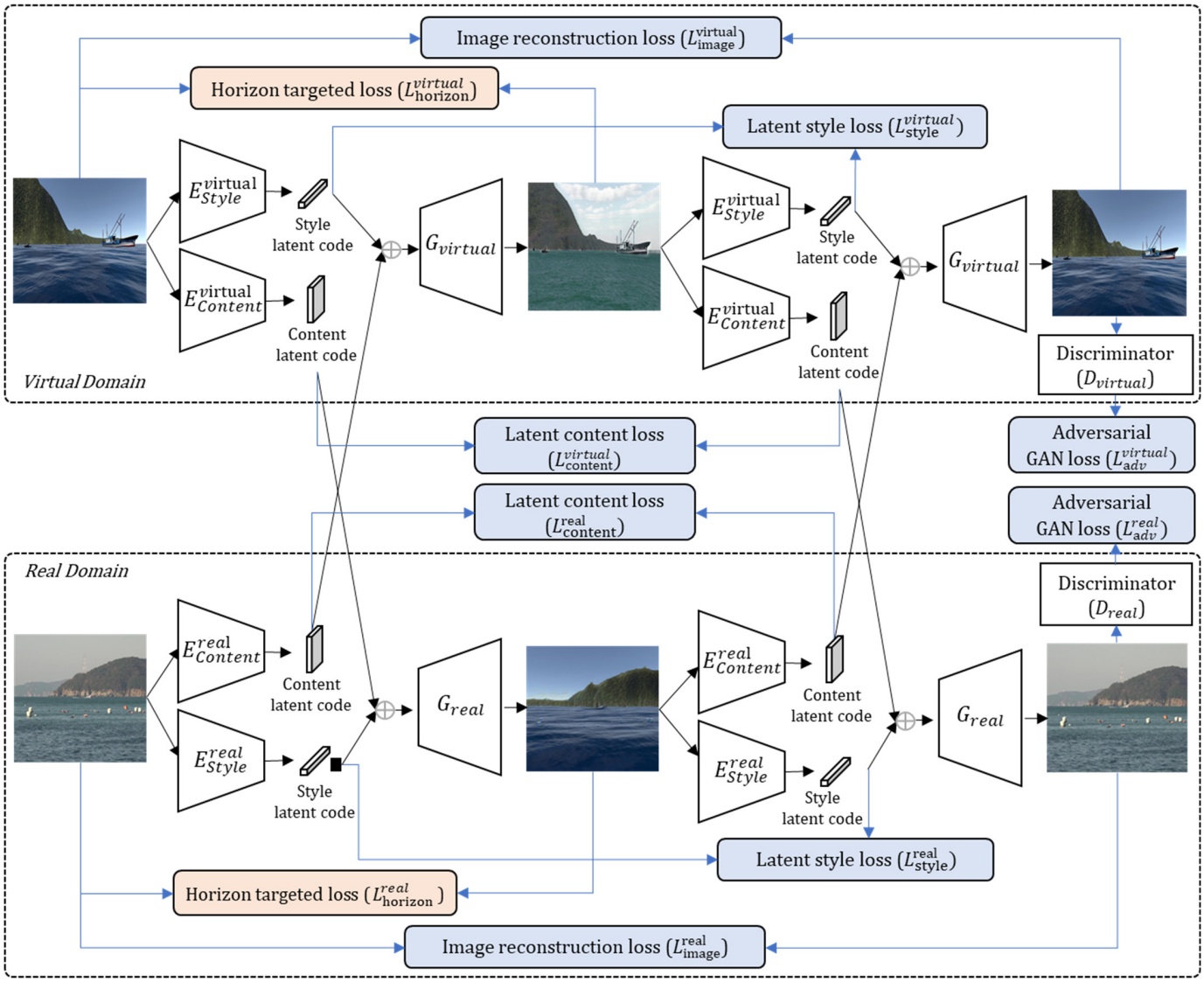


Figure 4. Architecture of training process of the generator.

There are five main types of loss functions: image reconstruction loss, latent content loss, latent style loss, adversarial GAN loss, and horizon-targeted loss. In image reconstruction loss, $L_{image}^{virtual}$ and L_{image}^{real} are loss functions that calculate and reflect the difference between images converted to other domains by the encoder and decoder. In latent content and style loss, $L_{content}^{virtual}$, $L_{content}^{real}$, $L_{style}^{virtual}$, and L_{style}^{real} are loss functions for the latent space. In horizon-targeted loss, $L_{horizon}^{virtual}$ and $L_{horizon}^{real}$ are loss functions that calculate and reflect the difference between images converted to other domains based on the RoI (region of interest) containing the horizon. In adversarial GAN loss, $L_{adv}^{virtual}$ and L_{adv}^{real} are loss functions that adjust the feature distribution of the input image to match that of the target domain.

(1) Image reconstruction loss: Given an image sampled from the data distribution, we reconstruct the entire image after encoding and decoding. L_{image}^{real} is defined in a similar manner.

$$L_{image}^{virtual} = \mathbb{E}_{x_1 \sim p(x_1)} \left[\| G_{virtual}(E_{content}^{virtual}(x_1), E_{style}^{virtual}(x_1)) - x_1 \| \right] \quad (1)$$

(2) Latent content loss: Given a latent content code sampled from the latent distribution, we calculate the latent content code required to reconstruct it after decoding and encoding. $L_{content}^{real}$ is defined in a similar manner.

$$L_{content}^{virtual} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim p(s_2)} \left[\| E_{content}^{real}(G_{real}(c_1, s_2) - c_1 \| \right]_1 \quad (2)$$

(3) Latent style loss: Given a latent style code sampled from the latent distribution, we calculate the latent style code required to reconstruct it after decoding and encoding. $L_{style}^{virtual}$ is defined in a similar manner.

$$L_{style}^{real} = \mathbb{E}_{c_1 \sim p(c_1), s_1 \sim p(s_1)} \left[\| E_{style}^{real}(G_{real}(c_1, s_2) - s_1 \| \right] \quad (3)$$

(4) Adversarial GAN loss: We employ GANs to match the distribution of virtual-domain images to that of the real-domain data. L_{adv}^{real} is defined in a similar manner.

$$L_{adv}^{real} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim p(s_2)}[log(1 - D_{real}(G_{real}(c_1, s_2)))] + \mathbb{E}_{x_2 \sim p(x_2)}[log\ D_{real}(x_2)] \quad (4)$$

(5) Horizon-targeted loss: We calculate the loss by comparing the RoI in the input data to that in the output data. $L_{horizon}^{real}$ is defined in a similar manner.

$$L_{horizon}^{virtual} = \mathbb{E}_{x_1 \sim p(x_1)} \left[\| G_{virtual}(E_{content}^{virtual}(\text{RoI\_}x_1),\ E_{style}^{virtual}(\text{RoI\_}x_1)) - \text{RoI\_}x_1 \| \right] \quad (5)$$

(6) Total loss: The objective function of our conditional GAN model is based on that of MUNIT [22]. We propose the introduction of horizon-targeted loss, which focuses on the horizon area to prevent losing shape coherence, via the following minimax game.

$$\min_{E_{style},\ E_{content},\ \ G_{real},\ G_{virtual}} \max_{D_{real},\ D_{virtual}} = \mathcal{L}\left(E_{style},\ E_{content},\ \ G_{real},\ D_{virtual},\ \ G_{real},\ D_{virtual}\right)$$

$$= \mathcal{L}_{GAN}^{virtual} + \mathcal{L}_{GAN}^{real} + \lambda_x\left(\mathcal{L}_{image}^{virtual} + \mathcal{L}_{image}^{real}\right) + \lambda_c\left(\mathcal{L}_{content}^{virtual} + \mathcal{L}_{content}^{real}\right) + \lambda_s\left(\mathcal{L}_{style}^{virtual} + \mathcal{L}_{style}^{real}\right) \quad (6)$$
$$+ \lambda_h\left(\mathcal{L}_{horizion}^{virtual} + \mathcal{L}_{horizion}^{real}\right)$$

In the above equation, $\lambda_x, \lambda_c,\ \lambda_s$, and λ_h are weights that control the importance of the reconstruction terms.

3.2. Horizon Targeted Loss Calculator

We introduce a horizon-targeted loss function because many marine images do not exhibit a clear difference in colors between the sky and sea; unlike vehicles on land, there are no roads for vessels, thereby resulting in diverse vessel positions, directions, and sizes. The proposed horizon-targeted loss can consider the detected horizon and vessel area to reduce shape inconsistency by comparing segmented RoI images of the horizon and vessel in real- and virtual-domain data. Thus, the network can learn to identify the horizon more proficiently. A flowchart depicting the calculation of horizon-targeted loss is outlined in Figure 5.

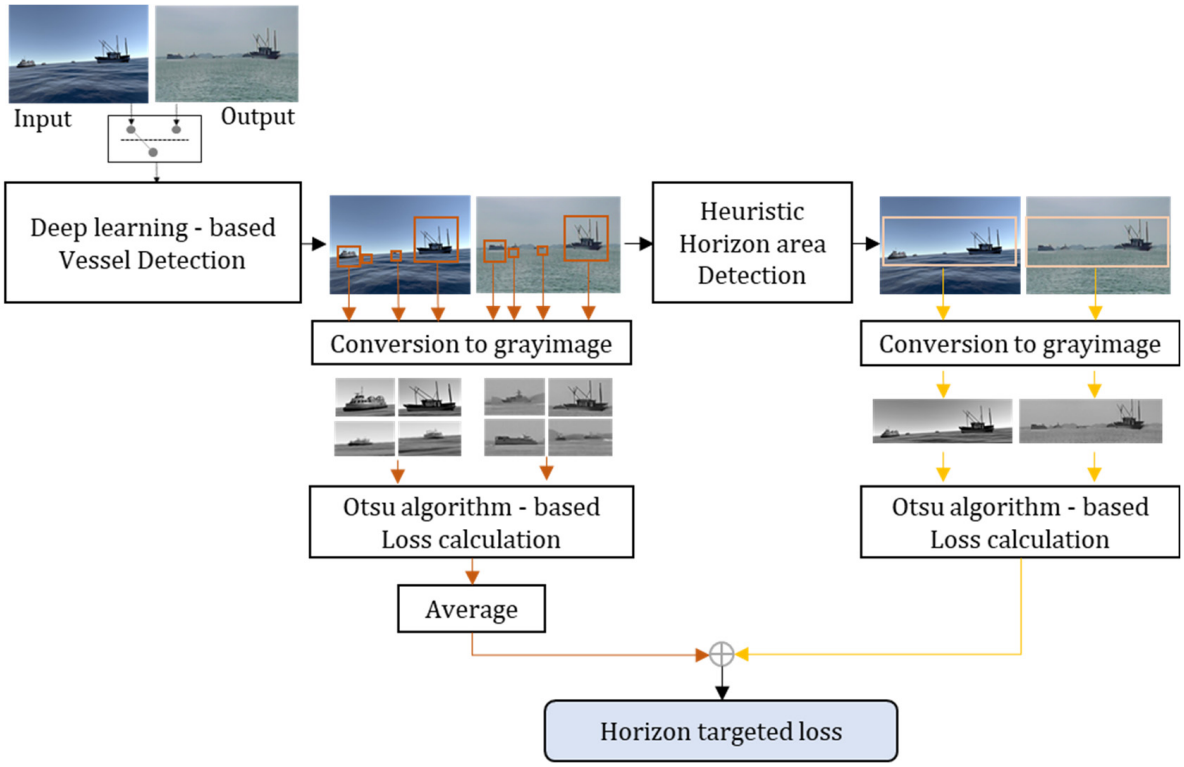


Figure 5. Flowchart depicting horizon-targeted loss calculation process.

First, the positions and bounding boxes of the vessels in a single input image (virtual or realistic) are detected using a deep learning based object detection algorithm such as YOLO [23]. Second, a horizon candidate area is extracted based on the position of the vessel located at the center of the detected positions, because we can assume that the horizon line extends across all vessels. After extracting RoI images according to the extracted vessel and horizon regions, they are converted into black and white images to enable a comparison of the structural characteristics of the images rather than that of their colors. The regions are roughly divided according to the objects in the image using the Otsu algorithm. Finally, the loss is calculated by comparing each region extracted from the real image with that extracted from the virtual image.

Algorithm 1 describes the procedure for horizon-targeted loss calculation in detail. After obtaining the detected vessel RoI list by YOLO [23], a horizon candidate region is extracted considering the highest vessel position in the image. The height of the horizon candidate region is measured from the highest vessel position to approximately one-third of the height of the input image on either side, which can be expected in the horizontal area. For example, if the input height is 1080 pixels, the height of the candidate region is 300 pixels. Subsequently, the loss is calculated by comparing the segmented result of the input and output images of the vessels and the horizon area. If there are no vessels, we do not calculate the horizontal targeted loss.

3.3. Discriminator

We employed multi-scale discriminators introduced in pix2pixHD [24] to update the generators such that they produced both realistic local features and correct global structures. A discriminator with a large receptive field is required to produce high-quality images. Generally, the network depth or the kernel size increases. However, in such cases, substantial memory is required for training, and overfitting may occur. Therefore, a multi-scale discriminator with the same structure, but managing different image scales, was used.

Algorithm 1 Horizon-targeted loss calculation

Input: $Image_{virtual/real}$, $Image_{realistic}$
Output: *Horizon targeted loss*
$minY \leftarrow ImageHegiht$
$detectedObjectList \leftarrow$ Yolo_Based_Object_Detection_Network$(Image_{virtual})$
If $detectedObjectList <> null$ **then**
  For $i := 0 \rightarrow len(detectedObjectList)$ **do**
    If $object ==' vessel'$ **then**
      $vesselList.append(detectedObjectList[i])$
      $RoI_{vessel}list \leftarrow$ RoI$_{\text{vessel}}$ListExtract $(vesselList)$
      If $detectedObjectList[i].y < minY$ **then**
        $minY \leftarrow DetectedObjectList[i].y$
      End
    End
  End
$RoI_{horizon} \leftarrow$ RoI$_{\text{horizon}}$Extract $(minY)$
$grayImage_{virtual/real} \leftarrow$ Convert_to_gray$(Image_{virtual/real})$
$grayImage_{realistic} \leftarrow$ Convert_to_gray$(Image_{realistic})$
$segementedImage_{virtual/real} \leftarrow$ Otsu$(Image_{virtual/real})$
$segementedImage_{realistic} \leftarrow$ Otsu $(Image_{realistic})$
  For $i := 0 \rightarrow len(RoI_{vessel}list)$ **do**
    $vesselLoss \leftarrow$
$vesselLoss +$ CampareImageWithRoI$(RoI_{vessel}list[i], Image_{virtual/real,} Image_{realistic})$
  End
  $vesselLoss \leftarrow vesselLoss/len(RoI_{vessel}list)$
  $horizonLoss \leftarrow$ CampareImageWithRoI$(RoI_{horizon}, Image_{virtual/real,} Image_{realistic})$
  $Horizon\ targeted\ loss \leftarrow vesselLoss + horizonLoss$
Else $Horizon\ targeted\ loss \leftarrow 0$

4. Experiments and Analysis

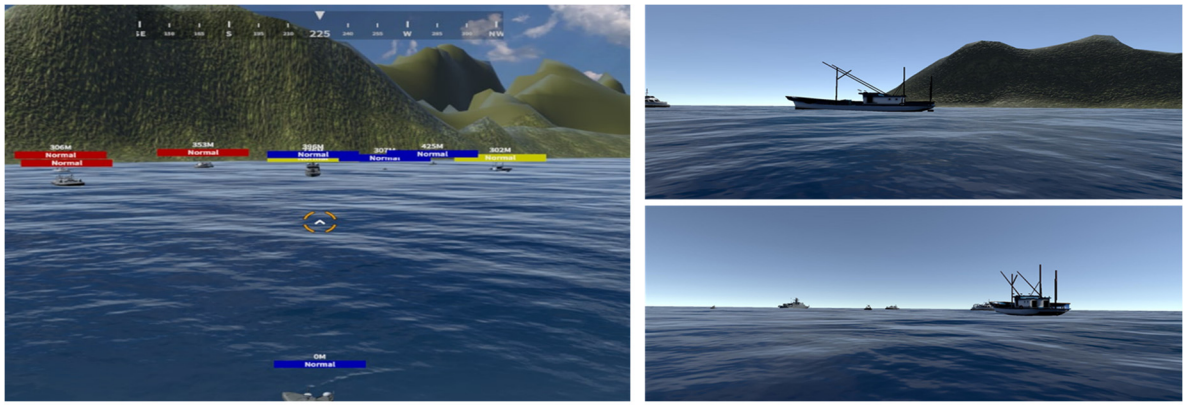
This section describes the experiments conducted to verify the performance and analysis of the results of the proposed diverse realistic marine image generation framework.

4.1. Experiment Settings

Figure 6 illustrates the experimental environment for testing the proposed method which was implemented in Unity to obtain 2000 virtual oceanic images. Furthermore, we collected 2000 real-world images of the South Korean Ocean. The real and virtual ocean image dimensions were 1920 × 1080, and the images did not exhibit paired matching. The distributions of the real and virtual datasets for training and testing are summarized in Table 2.

Table 2. Training data overview, including the distribution of images for training and testing.

Ocean Dataset	Real Ocean Data	Virtual Ocean Data
Training	1800	1800
Testing	200	200
Total Number of Images	2000	2000



(a) Our simulator (b) Camera sensor data

Figure 6. Experimental environment: (**a**) virtual vessel-simulator, and (**b**) virtual camera sensor data.

4.2. Evaluation Metrics

We utilized a Frechet Inception Distance (*FID*) [25] in our evaluation, which is a popular metric for evaluating image generation tasks. This metric calculates the distance between feature vectors calculated for real and generated images. The *FID* metrics were defined as follows:

$$FID = \| \mu_X - \mu_Y \|^2 + \mathrm{Tr}\left(\sum X + \sum Y - 2\sqrt{\sum X \sum Y}\right) \quad (7)$$

4.3. Quantitative Evaluation of Diverse Photorealistic Marine Image Generator

Figure 7 illustrates the results of converting domains A -> B and B -> A. The figure reveals that the generator can maintain the structures within each image, such as the horizon, ship, and mountain shapes, on conversion. The trained network is one, and we can generate new diverse style images.

We applied FID to quantitatively evaluate the performance of the proposed method. To obtain a lower score, a model must be able to generate images that are similar with real world image data. While virtual images, which obtained by our simulator yielded an FID of 109.422, the realistic images which are generated by the proposed method yielded an FID of 102.210. Table 3 shows our results are closer to the real data than those of virtual images.

Table 3. Quantitative evaluation result of the proposed method.

Method	FID (↓)
Virtual images vs. Real images	109.422
Realistic images (ours) vs. Real images	102.210

Figure 8 illustrates the results of converting the virtual ocean simulator images of four frames into four different styles. The results indicated that the styles of the sky and sea were well applied.

(a) Domain A (Virtual image) -> Domain B (Real image) Results

(b) Domain B (Real image) -> Domain A (Virtual image) Results

Figure 7. Results of the proposed generator: (**a**) results of image translation from the virtual to real domain, and (**b**) results of image translation from the real to virtual domain.

4.4. Horizion Targeted Loss Effects

The result of applying horizon-targeted loss is revealed in Figure 9. Notably, in the case of MUNIT [22], the line of the horizon is curved into a wave shape; this is not the case with the proposed method.

4.5. Comparison of Object Detection Performance

After converting the virtual data to real-world data, three types of data were available: (1) real-world data, (2) virtual data, and (3) realistic data. We trained the YOLO [23] network, which is a deep learning algorithm for object detection in images, on real-data images to detect the positions of ships and their bounding boxes. We subsequently tested the network on the three types of data and compared the object-detection accuracies observed, as illustrated in Figure 10. The dimensions of the input image were 512 × 910. We trained the YOLO [23] network using 500 real training data images. Subsequently, we tested 200 images each of real, virtual, and realistic data.

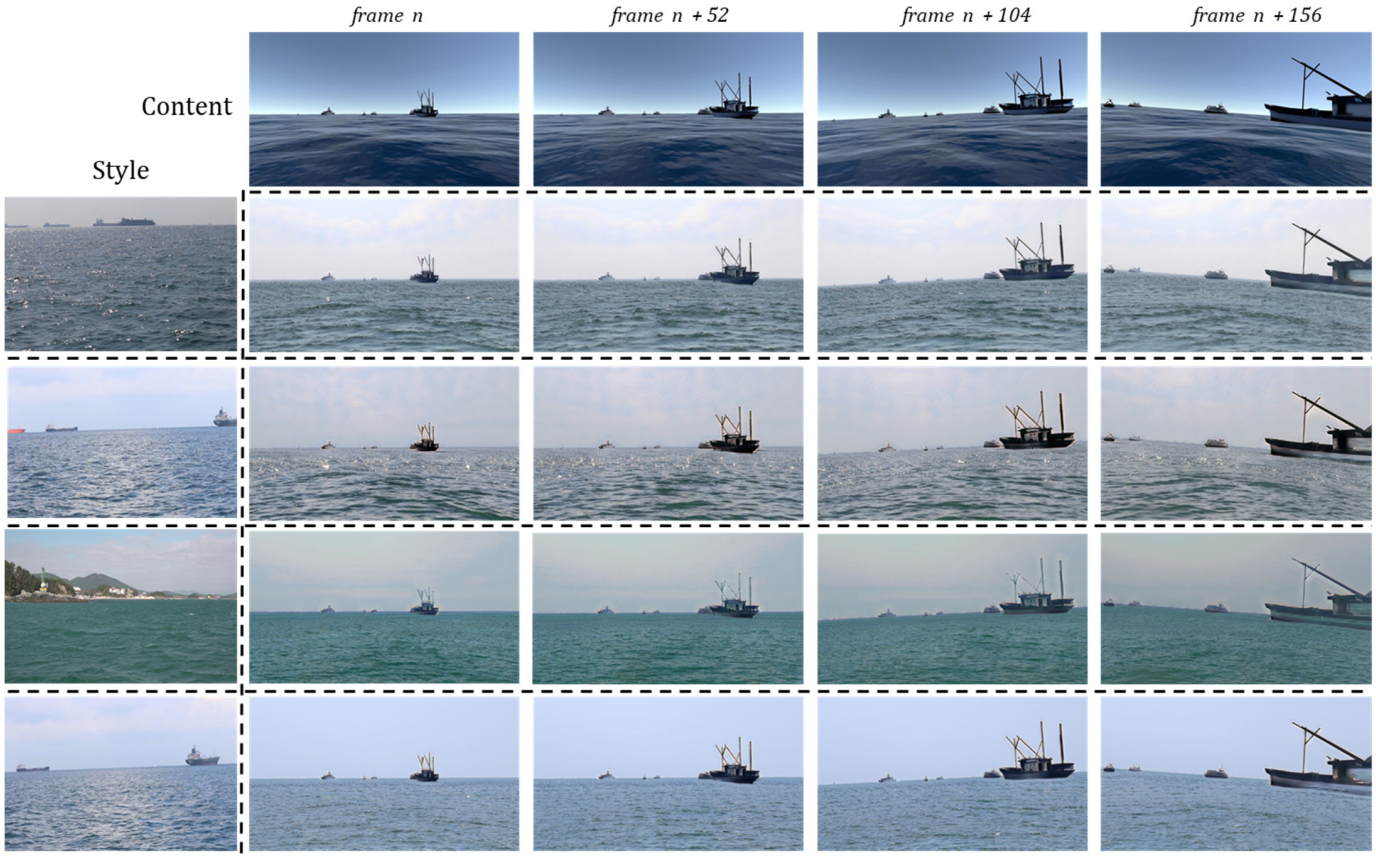


Figure 8. Compared results of four different style image mapping.

Table 4, which lists the object detection accuracies obtained for the three types of data, reveals that the object detection accuracy is higher on testing within the case of the realistic data than that obtained within the case of the virtual data. This indicates that we generated realistic images that resembled the real data more closely.

Table 4. Comparison results of object detection performance.

Data Type	Object Detection Accuracy
Real data	85%
Virtual data	79%
Realistic data	81%

Figure 9. Horizon-targeted loss effect results; the first column shows that MUNIT [22] and our proposed method can generate realistic images which do not include the horizon line. However, the second and the third column show our proposed method maintain the shape of horizon line in a yellow box and vessel shape in a green box more clearly than MUNIT [22] in the case of horizon scene.

(a) Detection result using real data (b) Detection result using virtual data (c) Detection result using realistic data

Figure 10. Ship detection results using three different data types: real, virtual and realistic.

4.6. Inference Time

The proposed method, which applies horizon-targeted loss, does not require semantic information from the input. Because the algorithm requires considerable time to compare losses in the training stage, the training time increases, and no significant difference between MUNIT [22] and the proposed algorithm is observed after training. Therefore, unlike in MUNIT [22], no additional computation is required at the time of inference because the input image and network layers remain the same. Furthermore, the accuracy of image conversion was improved by reducing the error in identifying the horizon area. Table 5 shows that we achieved an inference time of 0.032 s (31.25 fps) with 640 × 480-pixel images on a GeForce GTX 3090 Ti graphics card.

Table 5. Inference time comparison.

Method	Average Inference Time
MUNIT [22]	0.034 s (29.41 fps)
Ours	0.032 s (31.25 fps)

5. Conclusions

In this study, we propose a virtual-to-realistic marine image style transfer method using horizon-targeted loss for marine data that can preserve the shapes of the horizon and the vessel. Horizon-targeted loss focuses on distinguishing the horizon from the other structural forms in the input and output images, based on a comparison of the segmented shape. Experiments reveal that the proposed method preserves the structural shapes on the horizon more accurately than the existing algorithms. In addition, a higher object detection accuracy is observed on learning using the augmented learning data of various walkdown styles compared with learning using virtual data alone. The proposed method allows us to generate realistic data to train AI models of vision-based autonomous vessels by actualizing and augmenting virtual images acquired from virtual autonomous-vessel simulators. A comparison between the proposed and MUNIT [22] methods via visual assessment and quantitative analysis reveals that our method achieves better performance in maintaining the shapes of the horizon and vessels. In future work, we will modify the proposed model to obtain high-resolution images in real time.

Author Contributions: Conceptualization, methodology, and writing—original draft preparation, J.P.; software, T.H.C.; project administration and funding acquisition, K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Future Challenge Program through the Agency for Defense Development funded by the Defense Acquisition Program Administration.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and Service Robotics*; Springer: Cham, Switzerland, 2017; pp. 621–635. [CrossRef]
2. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In Proceedings of the Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017; pp. 1–16.
3. Rong, G.; Shin, B.H.; Tabatabaee, H.; Lu, Q.; Lemke, S.; Možeiko, M.; Boise, E.; Uhm, G.; Kim, T.H.; Kim, S.; et al. Lgsvl simulator: A high fidelity simulator for autonomous driving. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–6.
4. Kermorgant, O. A dynamic simulator for underwater vehicle-manipulators. In *International Conference on Simulation, Modeling, and Programming for Autonomous Robots*; Springer: Cham, Switzerland, 2014; pp. 25–36.

5. Rohmer, E.; Singh, S.P.; Freese, M. V-REP: A versatile and scalable robot simulation framework. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 1321–1326.
6. RobotX Simulator. Available online: https://bitbucket.org/osrf/vmrc/overview (accessed on 30 December 2018).
7. Unmanned Surface Vehicle Simulator. Available online: https://github.com/disaster-robotics-proalertas/usv_sim_lsa (accessed on 30 December 2018).
8. Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; Ha, J.W. Photorealistic style transfer via wavelet transforms. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9036–9045.
9. Li, Y.; Tang, S.; Zhang, R.; Zhang, Y.; Li, J.; Yan, S. Asymmetric GAN for Unpaired Image-to-Image Translation. *IEEE Trans. Image Process.* **2019**, *28*, 5881–5896. [CrossRef]
10. Manzo, M.; Pellino, S. Voting in Transfer Learning System for Ground-Based Cloud Classification. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 542–553. [CrossRef]
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Processing Syst.* **2014**, *27*, 1–9.
12. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
13. 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.
14. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
15. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 16th IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2242–2251.
16. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2849–2857.
17. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
18. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2180–2188.
19. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. Beta-Vae: Learning Basic Visual Concepts with a Constrained Variational Framework. 2016. Available online: https://openreview.net/forum?id=Sy2fzU9gl (accessed on 1 December 2021).
20. Kim, H.; Mnih, A. Disentangling by factorising. In Proceedings of the International Conference on Machine Learning, Hanoi, Vietnam, 28–30 September 2018; pp. 2649–2658.
21. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–51.
22. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
23. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
24. 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.
25. 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. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 1–12.