electronics

MDPI

Article

A Ship Detection Method in Infrared Remote Sensing Images Based on Image Generation and Causal Inference

Yongmei Zhang [1,*], Ruiqi Li [1], Zhirong Du [2] and Qing Ye [1]

[1] School of Computer Science and Technology, North China University of Technology, Beijing 100144, China; lrq@mail.ncut.edu.cn (R.L.); yeqing@ncut.edu.cn (Q.Y.)
[2] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; dzr@mail.ncut.edu.cn
* Correspondence: zhangym@ncut.edu.cn

Abstract: Aiming at the problems of scarce public infrared ship data and the difficulty of obtaining them, a ship image generation method based on improved StyleGAN2 is proposed. The mapping network in StyleGAN2 is replaced with a Variational Auto-Encoder, enabling the generated latent variables to retain original image information while reducing computational complexity. This benefits the construction of the image. Additionally, a self-attention mechanism is introduced to capture dependency information between distant features, generating more detailed object representation. By reducing the number of input noises in the generator, the quality of the generated images is effectively enhanced. Experimental results show that the images generated by the proposed method closely resemble the structure, content and data distribution of the original real images, achieving a higher level of detail. Regarding ship detection methods based on deep learning, they often suffer from complex detection networks, numerous parameters, poor interpretability, and limited real-time performance. To address these issues, a lightweight multi-class ship detection method for infrared remote sensing images is designed. This method aims to improve real-time performance while maintaining accurate ship detection. Based on ship detection, an interpretable ship detection approach based on causal reasoning is presented. By integrating singular value decomposition with the Transformer architecture, the model focuses on causal ship features associated with labels in the images. This enhances the model's robustness against non-causal information, such as background details, and improves its interpretability.

Keywords: StyleGAN2; image generation; object detection; infrared remote sensing images; causal reasoning

check for updates

Citation: Zhang, Y.; Li, R.; Du, Z.; Ye, Q. A Ship Detection Method in Infrared Remote Sensing Images Based on Image Generation and Causal Inference. *Electronics* **2024**, *13*, 1293. https://doi.org/10.3390/electronics13071293

Academic Editor: Manohar Das

Received: 21 February 2024
Revised: 24 March 2024
Accepted: 29 March 2024
Published: 30 March 2024

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

1. Introduction

Ship detection is a pivotal concern in both military and civilian arenas. In recent years, infrared detection technology has developed rapidly. The characteristics of non-contact and passive detection have made infrared detection technology a research hot-spot in the field of military reconnaissance and civil detection. The infrared detection technology has received extensive attention from all over the world. The accurate and real-time detection of ships is one of the most important means to ensure maritime safety. Traditional ship detection methods mainly rely on manual inspection and radar monitoring, but these methods entail an excessive consumption of human resources and have a limited monitoring range. Efficient ship detection is of great significance not only in military fields such as maritime combat scheduling, maritime rights, and interest maintenance, but also in civil fields such as maritime search and rescue and port management [1].

Efficient object detection and image quality are inseparable [2,3]. The study of the generation of high-quality images is of great practical significance. With the continuous development of deep learning, deep generation has become an important research direction

in the field of computer vision [4,5]. In 2014, Goodfellow et al. proposed the Generative Adversarial Network (GAN) [6], which is one of the main technologies for deep generation in computer vision. Subsequently, many variants have been developed and derived, such as Conditional Generative Adversarial Network (CGAN) with constraints of both generators and discriminators [7], Deep Convolution Generative Adversarial Network (DCGAN) [8], and Self-Attention Generative Adversarial Network (SAGAN), which combines the self-attention (SA) module with GAN [9], etc. These GAN variants improve the performance and stability of GAN by changing the network structure, loss function, etc.

In recent years, GAN has achieved great success in the field of computer vision [10,11], bringing about breakthroughs in image generation, image restoration, and image conversion. However, the distribution of infrared data generated by a typical GAN differs significantly from that of original data. Reducing this distribution difference in the generated data poses a challenging problem for data generation methods.

Infrared remote sensing images obtain object-related information by detecting the infrared radiation energy of the object, and have obvious advantages over visible light in terms of night detection and object state monitoring. In an infrared remote sensing image, because the object and the background show different radiation characteristics, it is possible to distinguish between different objects and their surrounding backgrounds. Infrared remote sensing is an important source of ship detection data because of its unique advantages, such as all-weather operation, good concealment, anti-reconnaissance, penetration through rain and fog, as well as strong environmental adaptability and anti-interference ability [12].

Most existing ship detection methods rely on visible light images or SAR images. However, ship detection methods based on infrared images are extremely rare. One of the important reasons for this is the scarcity of public infrared ship data due to the sensitivity and challenges associated with obtaining infrared images. Additionally, when compared to the detection tasks of natural images, ship detection using infrared images faces challenges such as poor-quality images, complex scenes and weather conditions, small objects, and weak semantic features. Given the limited availability of ship samples in infrared remote sensing images, it becomes difficult to ensure the robustness and accuracy of ship object detection in complex scenes. Currently, the precision and real-time performance of infrared ship detection methods cannot fully satisfy the requirements of practical applications. Consequently, infrared ship detection technology remains a focal point and challenge in contemporary research.

StyleGAN2 [13] is an important model for image generation, producing high-quality and diverse images. However, there are still some challenges in the generation of infrared remote sensing ship images [14], including unclear details in the generated images and a lack of realism. Therefore, this paper proposes a method for ship image generation based on an improved StyleGAN2 network, generating infrared remote sensing ship images that are closer to the original images. Additionally, an infrared remote sensing ship detection method is designed for multi-class lightweight use, achieving better detection results for multi-class ship objects in complex scenes. Based on ship detection, an interpretable method of ship detection based on causal reasoning is proposed to identify ship objects, and the effectiveness of this method is verified through experimental results.

The main contributions can be summarized as follows:

(1) Replacing the mapping network of the original StyleGAN2 network. To reduce the computational complexity and the number of parameters of the original StyleGAN2 network while preserving the original image information in the generated latent space, this paper replaces the mapping network of the original StyleGAN2 with a Variational Auto-Encoder (VAE). The latent space generated through VAE retains the original image information, which is beneficial for the generation of infrared remote sensing ship images;

(2) Introducing a self-attention mechanism in addition to replacing the mapping network of the original StyleGAN2 network addresses the issue of lacking low-level detailed features and the reduction in useful information extraction during the convolution

process. This results in the generation of infrared remote sensing ship images that lack edges, textures, and other features. Furthermore, in the generator, the input of noise has been reduced from two noises per feature block to one noise, effectively preventing excessive noise from affecting the quality of generated images;

(3) Conducting comparative experiments on an infrared remote sensing ship image dataset. In this paper, the effectiveness of the proposed method is verified through ablation experiments and comparative experiments with mainstream image generation methods. Subjective and objective evaluation metrics show that the proposed method has a good effect, which is beneficial for addressing the problem of a lack of ship samples in infrared remote sensing images. Frechet Inception Distance (FID) is added to the common evaluation metrics of image generation methods to further assess the diversity of generated images and the similarity of the distribution between generated images and original real images;

(4) The ship detection method based on deep learning has the problems of a complex network, a large number of parameters, and poor interpretability. This paper designs a lightweight multi-class ship detection method for infrared remote sensing images [15], which can detect multi-class ship objects in complex scenes. An interpretability method for ship detection based on causal reasoning is proposed. Singular value decomposition and Transformer are combined to reduce the dimensionality and classify the data, and the model detection results are further explained to improve the interpretability of the model.

The remaining structure of this paper is as follows. Section 2 introduces related works. The proposed improved StyleGAN2 network image generation method is described in Section 3. The multi-class ship detection method of infrared remote sensing images is shown in Section 4. The results and analysis of the comparative experiments are presented in Section 5, and Section 6 summarizes the work of this paper.

2. Related Works

This section briefly introduces the related work of traditional image generation methods, image generation methods based on deep learning, ship detection methods for infrared remote sensing images based on deep learning, and causal reasoning methods.

2.1. Traditional Image Generation Methods

Image generation has always been an important research direction in the field of computer vision. Typical applications include data augmentation, animation generation, and face generation, among others. Traditional image generation methods rely on the explicit modeling of data distribution, but the quality of the generated images is not high. Traditional image generation methods can be roughly divided into probability-based methods and energy-based methods.

Probability-based image generation methods involve constructing a probability estimation model for the entire image space and then sampling from the probability distribution of that space. The key to this method lies in obtaining a parameterized representation of the object probability distribution. The object distribution is generally characterized by a limited number of data samples. Each sample is a pulse signal, and the data distribution is expressed by a combination of these pulse signals. This pulse combination constitutes an observation of the object distribution. The probability-based generation method aims to model the object distribution through observations and then sample from this parametric model. This includes Principal Component Analysis [16], Independent Component Analysis [17], and Gaussian Mixture models [18]. These methods assume that the image distribution follows a specific type of simple distribution, and use parameter estimation methods to fit the parameters.

Energy-based image generation methods are more versatile. They capture dependencies between variables through a well-defined scalar energy function. Compared to probability-based methods, energy-based methods do not require normalization, thus

avoiding the issue of integrating over the entire space. Energy-based methods offer great flexibility in terms of architecture and training criteria. Commonly used energy-based methods for image generation include Hidden Markov Models [19], Markov Random Fields [20], and Restricted Boltzmann Machines [21].

2.2. Deep Learning-Based Image Generation Methods

With the increase in image data volume and the enhancement of computing power, the quality of image generation is constantly improving with the powerful representation ability of deep learning [22,23]. An image generation model based on deep learning can generate images of different scales and resolutions [24,25]. According to the different representations of probability distribution, the existing image generation methods based on deep learning can be divided into likelihood-based image generation methods and implicit image generation methods.

Likelihood-based image generation methods [26,27] learn the probability distribution of the data using the maximum likelihood criterion, including auto-regressive models, flow models, energy-based models, VAE, etc. Implicit image generation methods [28], such as GAN, do not explicitly express the probability distribution, and instead achieve direct transformation from noise to samples through adversarial learning. GAN possesses powerful image generation capabilities, and variant methods of GAN further enhance the performance of image generation.

In addition, there is a recently emerging image generation model called diffusion model [29], which has a more flexible model architecture and accurate log-likelihood calculation capacity compared to GAN. The diffusion model mainly includes forward and reverse diffusion processes, where random noises are gradually added to the samples in the forward diffusion process, and then samples are generated from the noise through the reverse diffusion process. However, the diffusion model relies on the diffusion steps of long Markov chains to generate samples. The computational cost is high, and the sampling is slower than GAN. Therefore, GAN-based methods are still the mainstream in the current field of image generation.

2.3. Infrared Remote Sensing Image Ship Detection Methods Based on Deep Learning

Traditional infrared remote sensing image ship detection methods are mainly divided into two steps. First, sea and land segmentation is performed, and then the ship object is detected. Traditional infrared remote sensing image ship detection methods have good real-time performance, but they find it difficult to extract robust artificial features, and the false alarm rate is high in complex and changeable marine scenes. Therefore, ship detection methods based on deep learning have become a research hot-spot. In ship object detection based on deep learning, common models are mainly divided into the feature extraction of candidate regions and regression-based models [30]. The main difference between these two types of models is whether they generate candidate regions. The model based on candidate region extraction first generates candidate regions and then performs feature extraction and object detection. The regression-based model directly processes the image features and then outputs the detection results.

Research on the detection of infrared remote sensing ships is extremely rare both domestically and internationally. In the literature [31], a ship detection method of thermal infrared remote sensing image based on YOLO in complex backgrounds has been studied. Gray stretch is used to expand the dataset, and the image with a resolution of 30 m is sampled to 10 m by the BiCubic interpolation method. Ships with similar features but located in different positions are marked, and the improved YOLOv5s model is used to quickly detect the ship candidate regions. This method labels the dataset by manually screening the ships that meet the features of the dataset and then labeling them. The workload is large, and the method has a weak detection ability for ships close to the coast.

Ref. [32] improved CenterNet to detect infrared ship images. The backbone network is replaced with ResNet50 for feature extraction. After up-sampling, an encoder is added

to further process the feature map extracted from the backbone network. To improve the adaptability of the network model to ship objects, dilated convolution is introduced into the encoder. This method performs well on backgrounds close to the coast, but no further detection experiments are carried out on multi-class ships and complex scenes (ports, multi-ship activity on the sea surface).

Ship detection methods based on deep learning can automatically extract features, and their detection accuracy and robustness are excellent, but their real-time performance is poor. Due to the difficulty of obtaining infrared remote sensing image sources, it is difficult to construct an infrared remote sensing ship dataset. The deep learning model has high computational complexity and high storage resource requirements, which makes it difficult to apply in practice. Therefore, how to improve the real-time performance of the detection method while ensuring performance is the challenge and focus of the current ship detection method based on deep learning.

2.4. Causal Inference Methods

According to the false correlation between the background and the object, Refs. [33,34] used the neural network to display the position of the object, obtained the weight value of each pixel point or used the category activation map to highlight the object, then masked the object and classified it. According to the classification results, the focus of the model is the object rather than the background information.

In Ref. [35], the current status of the combination of causal reasoning and machine learning is reviewed. Current causal reasoning is still in the initial stages of development. Due to the complexity of its neural network, the data involved are no longer just one-dimensional and two-dimensional data. Object detection based on deep learning involves high-dimensional data and massive parameters, making it extremely difficult to process these data.

Cui Peng et al. [36] proposed the concept of stable learning (reaching a consensus between causal reasoning and machine learning) and published corresponding papers. This paper mainly aims at extracting the relevant features of the object category. Unrelated features and false associations are removed, and predictions are made only based on features that have a causal relationship with the label. The main idea is to weight the image and irrelevant features independently to achieve the statistical independence of the essential features and irrelevant features of the image, such that, when detecting the object in the image, more attention will be paid to the relevant features of the image object rather than the irrelevant features. The stable learning network they designed can give more accurate predictions based on the essential features, and improve the generalization of the model. The experimental results show that the model performs well with OOD data.

In summary, traditional image generation methods mostly rely on feature representations to generate images, which are limited in their ability to generate complex images. The emergence of deep learning has improved the expressive ability of complex features, allowing image generation technology to undergo a breakthrough. Currently, image generation methods are mainly based on GAN, such as CGAN, DCGAN, SAGAN, StyleGAN, etc. Compared to traditional methods, these networks can generate more complex and higher-quality images. However, generating images, such as infrared remote sensing ship images with poor quality, limited content, and complex backgrounds remains challenging. Moreover, these deep learning methods often suffer from a high number of parameters and computational complexity.

In view of the scarcity of infrared remote sensing ship image samples and the difficulty obtaining them, this paper presents an improved StyleGAN2 data generation method for generating infrared remote sensing image ship objects. Aiming at the problems of large parameters, poor real-time performance, and poor interpretability in infrared remote sensing image ship detection methods based on deep learning, this paper designs a lightweight multi-class ship detection method for infrared remote sensing images, and proposes a ship

detection interpretability method based on causal reasoning to improve the interpretability of the model.

3. An Image Generation Method Based on Improved StyleGAN2

StyleGAN2 was proposed by the NVIDIA team in 2020 and is a variant of GAN [37]. Its purpose is to address the issue of noticeable droplet artifacts in the images generated by StyleGAN [38]. The overall network architecture of StyleGAN2 is similar to that of StyleGAN.

In terms of network architecture, StyleGAN2 removes the normalization operation after the input and eliminates the averaging operation in the style module. Additionally, it moves the noise outside the module. There are also differences in the noise input between StyleGAN2 and StyleGAN. While both utilize style noise, StyleGAN achieves this by adding the noise vector to the style vector, whereas StyleGAN2 directly adds the noise vector to the feature maps of each convolution layer.

In addition, StyleGAN2 also improves regularization. Batch optimization is used in the loss function to reduce the frequency of regularization optimization. When intermediate variables or spatial features undergo small-scale directional changes, the image features also undergo the same magnitude of variation. Therefore, path length regularization is introduced to penalize high-frequency oscillations in the style space. Path length regularization and style noise are incorporated into the progressive growing network, along with efficient convolution algorithms such as non-local operations and separable convolution, to accelerate the training of the generator and discriminator, while simultaneously improving the quality and diversity of the generated images.

Among the existing image generation methods based on deep learning, StyleGAN2 achieves better performance in generating high-resolution images and style transformations. This method has achieved good results in the generation of a face dataset [39]. The dataset used in this paper consists of infrared remote sensing images of ships, which have lower image quality and less content richness compared to visible light images (such as a human face dataset). Moreover, the complex background of the images makes it difficult to generate infrared remote sensing ship images using the StyleGAN2 network model. Therefore, this paper improves the StyleGAN2 network model. By replacing the mapping network of the original StyleGAN2 network model and introducing the SA mechanism, the generated infrared ship images contain more original image information, which is beneficial for generating ship details.

3.1. Improved StyleGAN2 Network Structure

Considering the construction of the object structure within the generated image, this paper employs an encoder to replace the mapping network, generating hidden variables that encapsulate genuine image information. As the resolution of the generated image escalates, convolution operations persist, influencing the generation of intricate image details. The SA mechanism significantly diminishes the distance between remotely dependent features, facilitating their effective utilization. This paper introduces SA. Considering the size and parameters of the original model while adding the SA mechanism, SA is only added to the 64*64 resolution layer and the 1024*1024 resolution layer to increase the connection between long-distance feature information, such that the feature details of the object in the generated image are more complete. In addition, considering that the texture features of infrared remote sensing ship images are fewer and the contrast between them and the background is lower, the noise input is reduced in the generator, and two noises are input from each feature block to one noise, so as to prevent too much noise from affecting the generation of high-quality images. The improved StyleGAN2 network structure is shown in Figure 1.

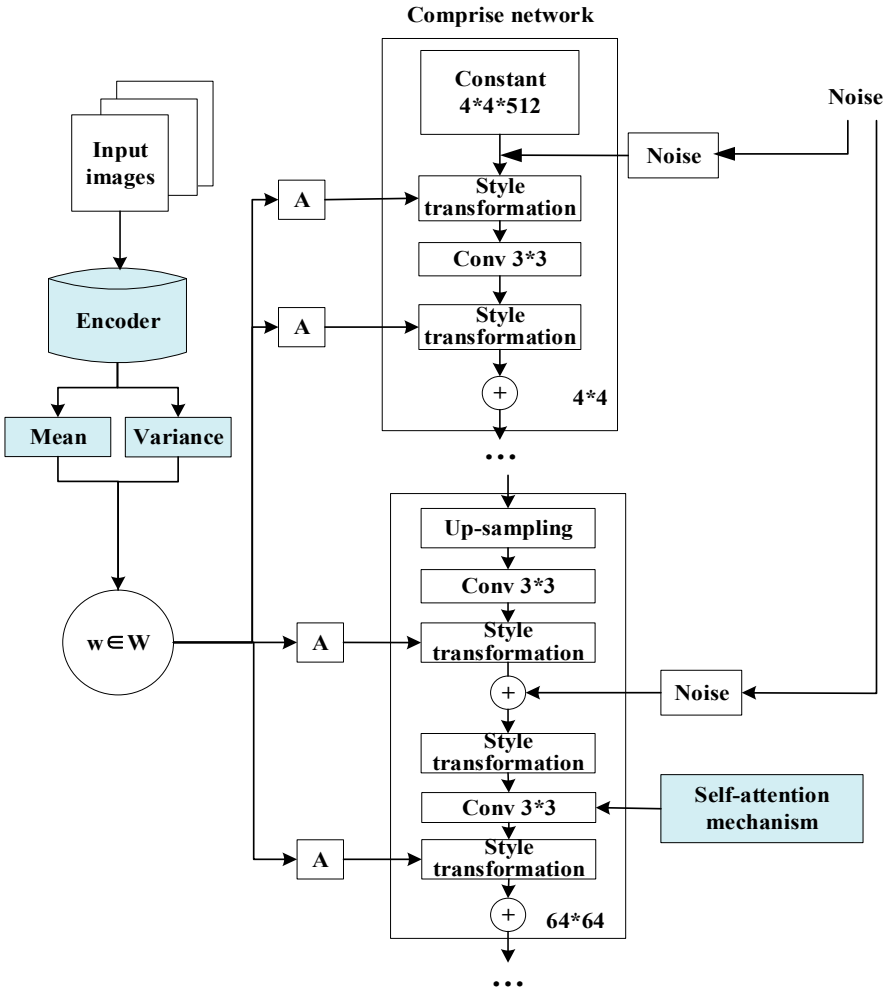


Figure 1. The network structure of the improved StyleGAN2.

3.1.1. Mapping Network Based on the Original Image Information Encoding

The mapping network in StyleGAN2 utilizes eight fully connected layers, which increases the computational complexity and the number of parameters for the StyleGAN2 network. To reduce the computational complexity and the number of parameters, this paper suggests substituting the StyleGAN2 mapping network with a VAE. The original image is directly inputted into the encoder of the VAE, which consists of a three-layer neural network. The encoder calculates the mean encoding and the variance encoding that controls the degree of noise interference. Subsequently, the original encodings are superimposed with appropriately weighted noise encodings to derive the intermediate hidden variable, which serves as the output of the mapping network. This intermediate hidden variable serves as the input for the StyleGAN2 generator network. This substitution omits the need for normalization, and compared with the original mapping network of StyleGAN2 using eight fully connected layers, it reduces the number of network layers and neurons per layer, thereby achieving the goal of reducing computational and parametric requirements.

VAE is a powerful generative model in deep learning, with a wide range of applications in data generation and potential space exploration. VAE [40] takes random samples from a specific distribution as input and generates corresponding images. Instead of a discriminator, VAE uses an encoder to estimate a specific distribution. The overall structure of VAE is similar to that of an auto-encoder, but the intermediate latent vectors are random vectors from a specific distribution. Moreover, the latent space generated by the encoder in VAE carries the original image information, which is beneficial for the generation of infrared remote sensing ship images. The architecture of VAE is shown in Figure 2.

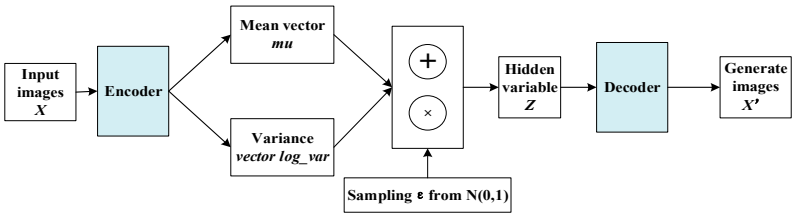


Figure 2. The architecture of VAE.

As can be seen from Figure 2, VAE has an encoder–decoder structure. In the figure, "+" and "×" represent addition and multiplication operations. After encoding the samples using the encoder, the mean vector *mu* and variance vector *log_var* are generated. A hidden variable Z is generated by multiplying a randomly sampled vector from the standard normal distribution by the variance vector *log_var* and adding the mean vector *mu*, as shown in Equation (1). The generated hidden variable is then input into the decoder for sample data generation.

$$Z = mu + \exp(log_var) \times \varepsilon \quad (1)$$

In Equation (1), *mu* and *log_var* represent the mean vector and variance vector generated by the encoder, respectively.ε is a vector randomly sampled from the normal distribution. Multiplying by ε in Equation (1) ensures the continuity of the latent space and the goodness of the structure.

3.1.2. A Generator Network Introducing SA Mechanisms

In the StyleGAN2 network, the image generation resolution ranges from 4*4 to 1024*1024. As the resolution increases, so does the network depth required for image generation. However, the training of deep networks is challenging, and network performance can deteriorate. To address this problem, StyleGAN2 incorporates skip connections to connect various layers. This approach mitigates the difficulty in training deep networks and partially alleviates the decline in network performance.

Nevertheless, during the convolution process, there is still a lack of low-level detailed feature information, and useful information tends to decrease as convolutions progress. As a result, the generated infrared remote sensing images of ships lack features such as edges and textures. To rectify this, this paper introduces an SA mechanism based on replacing the mapping network of StyleGAN2. The SA mechanism performs exceptionally well in capturing long-range dependencies between features [41]. In the domain of image generation, combining the feature extraction network with the SA mechanism has been shown to improve the quality of generated images.

Considering the complexity of the StyleGAN2 network and the training time, the SA mechanism is only incorporated after the convolutions at resolutions of 64*64 and 1024*1024 in this paper. This approach effectively preserves important feature information in the feature maps and enhances the long-range dependencies between feature information. As a result, it improves the generation of image details without significantly increasing the network burden. The SA module's structure is shown in Figure 3.

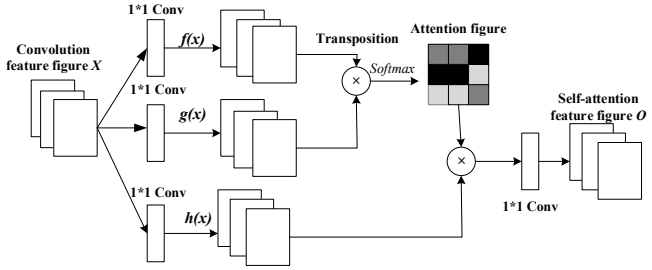


Figure 3. The SA module structure.

In Figure 3, $\otimes$ represents the dot product operation of the matrix. *f(x)*, *g(x)* and *h(x)* represent three linear transformations. The feature map X of the previous hidden layer is transformed into three feature spaces f, g and h, by 1*1 convolution operations. The calculation of the three linear transformations is shown in Equations (2)–(4).

$$f(x) = W_f \times x \tag{2}$$

$$g(x) = W_g \times x \tag{3}$$

$$h(x) = W_h \times x \tag{4}$$

In Equations (2)–(4), W_f, W_g, and W_h represent the weights of the convolution layers corresponding to the three linear transformations.

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, s_{ij} = f(x_i)^T g(x_j) \tag{5}$$

The f feature space and g feature space are used to generate attention maps by using Equation (5). In Equation (5), T denotes the transposition operation, and $\beta_{j,i}$ represents the relationship weight of the i-th position with the j-th position, which is calculated by *Softmax*. Finally, the point product of $h(x)$ and the normalized attention weight matrix is multiplied, and then the SA feature map O is obtained through a 1*1 convolution layer.

3.2. An Infrared Ship Image Generation Method Based on Improved StyleGAN2 Network

The specific steps of the infrared ship image generation method based on the improved StyleGAN2 network proposed in this paper are as follows:

(1) The input infrared remote sensing image is preprocessed, the image data are converted into high-dimensional vector data, and the real infrared remote sensing ship image is encoded. Through the mapping network based on the original image information coding, namely, VAE, the real infrared remote sensing ship image is mapped to the potential space, the feature information of the real infrared remote sensing ship image is extracted, and the real sample hidden variables close to the normal distribution are output. The encoding process is detailed below. ① The sample image X outputs two m-dimensional vectors through the encoder; ② Assuming that the latent normal distribution can generate the input image, ε is sampled from the standard normal distribution N (0, 1), and then the spatial hidden variable is obtained by Equation (2); ③ KL divergence is used to measure the similarity between the distribution of the hidden space and the normal distribution, so that the distribution of the generated potential space is as close as possible to the normal distribution;

(2) The generated hidden variable Z is broadcasted as the input of the AdaIN module in the Synthesis Network for the network training of the Synthesis Network. The Synthesis Network consists of multiple Synthesis Blocks. The input of each Synthesis Block is composed of noise, the output of the previous style block (the input of the first style block is a 512-dimensional constant), and style transformation. Each Synthesis Block includes up-sampling and convolution kernel AdaIN operations. The original image X and the hidden variable Z are used as the input of the Synthesis Network. After the convolution operation with resolutions of 64*64 and 1024*1024, SA is added to generate the image, $Y' = Synthesis\ Network_SA(X, Z)$;

(3) The infrared remote sensing ship image generated by the generator and the real infrared ship image are simultaneously input into the discriminator for judgment, and the results are fed back to the generator;

(4) The discriminator cost function is updated and optimized. Similarly, the generator cost function is also optimized;

(5) Repeat the above steps to complete all iterations of model training. The number of iterations is set to be 1600 times, and the generated image is output.

To verify the effectiveness and practicability of the proposed image generation method, a multi-class ship detection method in infrared remote sensing images is further studied and designed. LabelImg is used to label the ship in the generated image to generate a file with.xml as the suffix.

4. A Multi-Class Ship Detection Method for Infrared Remote Sensing Images

Because infrared ship objects are small and there are various complex scene disturbances, detection needs to be highly accurate and robust. Addressing the issues of low detection efficiency and high false alarm rates in traditional methods, as well as the large number of parameters in ship detection models based on deep learning, the author of this paper has proposed a lightweight multi-class ship detection method for infrared remote sensing images. On this basis, this paper proposes an interpretability method for ship detection based on causal reasoning, which combines low-rank decomposition with the Transformer to improve the interpretability of the model.

4.1. A Lightweight Multi-Class Ship Detection Method in Infrared Remote Sensing Images

At present, some commonly used deep learning methods have high computational complexity, complex models, and need to rely on a large amount of computing resources. It is difficult to achieve a balance between detection accuracy, speed, and storage space. Although a compressed depth model has emerged, the chip cost required for hardware is still high. With the development of deep models, more and more lightweight networks have achieved remarkable results in object detection.

YOLOv5 uses a lightweight model design. YOLOv5 can achieve high-precision object detection while maintaining real-time performance, especially in small object detection. YOLOv5 adopts a series of optimization strategies, which have been greatly optimized in terms of running speed and memory footprint, and can run quickly on different hardware devices. Because of the need to detect small object ships, this paper uses the YOLOv5l model for ship detection.

In this paper, the YOLOv5l model is improved upon. The overall network model of multi-class ship detection in lightweight infrared remote sensing images is shown in Figure 4. This paper mainly simplifies the model structure and reduces the model parameters. In the current lightweight network, considering the balance of speed and accuracy, the basic structure used in the ShuffleNet v2 network is selected [42], and this basic structure is named ShuffleUnit. The basic structure employs deep separable convolution, channel splitting and channel cleaning techniques to reduce the parameters numbers and calculations while ensuring effective feature extraction. In this paper, the coordinated attention mechanism (CA) is introduced, and the attention mechanism is added to the backbone network and the neck network at the same time to improve the sensitivity of the model to the object features. The EIOU Loss bounding box loss function is used to select high-quality anchor boxes for detecting ship objects more quickly and accurately.

Compared with the two-stage detection method based on Faster R-CNN, and the one-stage object detection algorithm based on YOLO v3, YOLO v4, YOLO v5s, YOLO v5l, YOLO x-s, and YOLO v7, the experimental results show that the lightweight infrared remote sensing image multi-class ship detection method can accurately detect infrared remote sensing image ships in complex scenes, and can maintain a good balance in detection accuracy, speed, and storage space.

4.2. A Ship Detection Interpretability Method Based on Causal Reasoning

In deep learning, the learning of objects mainly lies in the expression learning of object features, but the relationship between the specific features learned and the objects is in a black box state. How to better explain the relationship between them is a research hot-spot.

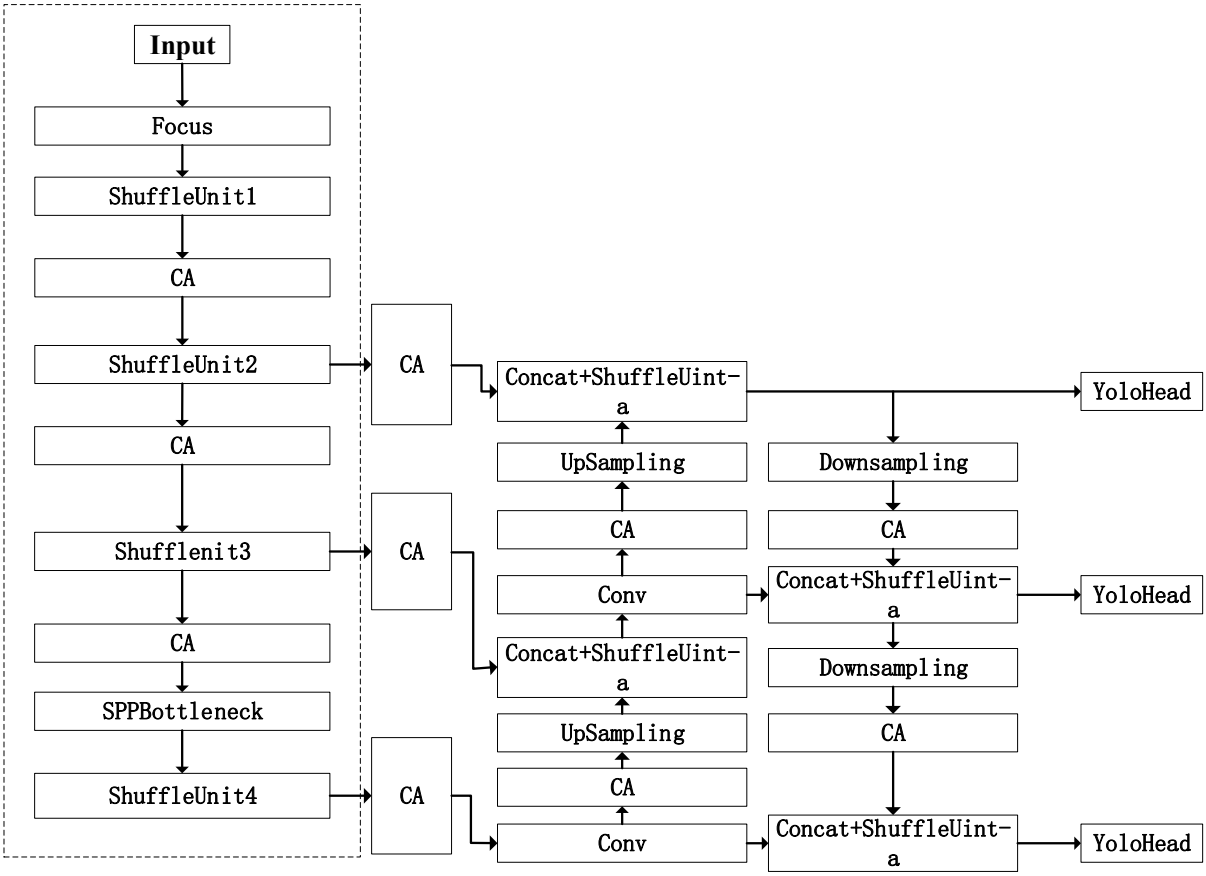


Figure 4. The overall network model of lightweight multi-class ship detection in infrared remote sensing images.

Compared with traditional machine learning, deep learning models have higher computational efficiency and higher accuracy of object detection, and show good performance in various fields. However, many deep learning models are black boxes with poor interpretability, because they are more interested in the correlation between input and output, rather than causality, which is not conducive to model structure optimization. The purpose of interpretability research is to enable humans to understand the operating mechanism of the model and the reasons for making certain inferences, so as to provide important auxiliary information for model improvement, data analysis, and decision-making.

Generally speaking, causality refers to the relationship between results and causes. Causal reasoning is a process of drawing conclusions about causality based on the environment in which the influence occurs. Causal reasoning is a powerful modeling tool for interpretive analysis, which can help restore causal associations in data, guide machine learning, and achieve interpretable and stable predictions. The core idea of the causal model is to pay more attention to the causal information corresponding to the result when the deep learning model extracts the high-dimensional features of the image, and try to ignore the false correlation between the background and the object. In recent years, deep learning models have been widely used to identify the causal relationships in data, rather than correlations.

At present, the interpretability methods are mainly based on the interpretation classification model, but there are few studies on the interpretability of advanced tasks such as object detection and semantic segmentation. At present, the causal reasoning of object detection is mainly intended to find the object in the image through the detection algorithm and describe its contour. That is, the object position and shape are roughly given by the pixel point, and then the object part is masked or removed, and then recognized. We then

observe the change of the recognition result to judge the pattern closely related to the given class in the input, in order to explain the mechanism of model classification.

In 2017, Vaswani et al. [43] proposed the Transformer model, which integrates the attention mechanism into the model. Transformer has shown strong recognition ability in natural language processing, sentiment analysis and other fields. However, Transformer's excellent recognition ability and training speed require a large number of parameters to support it. As an effective model compression method, low-rank decomposition has achieved results by adding it to the deep learning model. Although low-rank matrix factorization compresses the model, it has an impact on the recognition ability of the model. In this paper, low-rank decomposition is combined with Transformer to reduce the influence of the low-rank decomposition method on model recognition ability while maintaining the compression model to reduce the number of Transformer parameters. Ref. [44] proposed to train the model from the perspective of causality. When the classifier is classified, it should pay attention to the causal information of the image and ignore the irrelevant background information of the image. In this paper, the interpretability of ship detection in infrared remote sensing images is analyzed by referring to Ref. [44].

The singular value decomposition (SVD) method in low-rank decomposition is an important linear algebraic technique that can effectively reduce the amount of calculation in matrix operations and has a fast convergence speed. Especially for larger matrices, the convergence performance is better. This paper uses the SVD method. SVD is performed on the high-dimensional matrix representing the image, and only the first k larger singular values are retained to obtain a k-dimensional low dimensional representation, thus retaining as much of the main information of the image as possible and greatly reducing the storage requirements of the image. The SVD method is mainly used to sparsely represent the image information, which can retain the important information in the image, play a data compression role in the image, and then make the model more interpretable for feature learning. The specific steps of the presented interpretable method of ship detection based on causal reasoning proposed in this paper are as follows:

(1) The singular value curve of the image is drawn to determine the distribution range of the information in the image;
(2) The singular value of the image is selected, and the appropriate singular value k is selected by image reconstruction;
(3) The image after singular value decomposition is combined with Transformer.

The singular value decomposition method is defined as follows. A custom layer named SVDLinear is defined to perform SVD decomposition on the input matrix. This layer contains three parameters: the number of input features, the number of output features, and the number of singular values. In the forward propagation process, we multiply the input matrix by the left singular vector matrix, the singular value matrix, and then the right singular vector matrix to obtain the output matrix. In the initialization process, we randomly initialize the left singular vector matrix and the right singular vector matrix, as well as the singular value vector obtained by normal distribution sampling.

In this paper, the defined SVDLinear is embedded into the Transformer model, and the specific combination position is the penultimate layer of the Transformer decoder; that is, before the fully connected layer. After SVDLinear, a fully connected layer is used to map the hidden layer feature sequence back to the output feature space to obtain the output sequence, and then the fully connected layer is used to classify it.

Based on Ref. [44], and using the idea of data dimensionality reduction and data compression, the input image is encoded into a vector sequence, and each vector corresponds to the input data to construct a deep network. The deep network mainly relies on the Transformer structure, and it uses the multi-head SA mechanism and sparse coding operation to yield low-dimensional and sparse data. The model parameters are trained by back-propagation. The experimental results verify that the method is interpretable for its process.

(4) Train the model and save weight parameters.
(5) Load the weight parameters and model, input an image, and classify and recognize the input image.

5. Experiments and Analysis

5.1. Environment Configuration and Datasets

The deep learning framework used in this paper is PyTorch 1.11.0. The CPU is 20 Intel(R) Xeon(R) Silver 4210 CPU @ 2.20 GHz. The operating system version is Centos 7. The GPU version used is NVIDIA GeForce RTX 2080Ti with a memory size of 12 GB. The GPU memory is 32 GB.

The infrared ship images used in this paper are part of the infrared sea ship dataset produced by Arrow Optoelectronics Technology Co., Ltd. in Wuhan, Hubei, China. Using infrared devices with different resolutions and focal lengths, the dataset collected 8000 ship images with different resolutions from scenes at sea, in ports, and on the seaside, captured at various times. Seven types of ship objects in the image, namely, liner, bulk carrier, warship, sailboat, canoe, container ship, and fishing boat, are marked. All tag information is saved in the form of an XML file.

The clarity and size of the infrared ship images in this dataset vary, meaning they cannot be used directly to verify the image generation method proposed in this paper. Therefore, this paper selects a clear image with a resolution of 1024*1024 from the data set to form a new data set, and a total of 421 infrared ship images are screened out.

The experimental data set used for object detection is mainly composed of two parts. One part is the infrared ship image generated by the image generation method in this paper, and the other part is the infrared sea ship data set made by Arrow Optoelectronics Technology Co., Ltd. The data set contains seven types of ship objects in various simple and complex scenes, such as sea, port, and seaside. The number distribution of the seven types of ships in the experimental data set is shown in Table 1.

Table 1. The number distribution of the seven kinds of ships in the experimental data set.

Class	Liner	Bulk Carrier	Warship	Sailboat	Canoe	Container Ship	Fishing Boat
Amount	1099	4103	1761	3675	1381	489	6514

5.2. Parameters Setting and Evaluation Metrics

(1) Parameters

According to a large number of experiments, in the image generation experiment of this paper, the learning rate is 0.002, the optimizer uses Aadm the batch size is set to 4, and the single experiment is trained 1600 times. In the object detection experiment, the learning rate is an adaptive learning value, which is mainly adjusted by judging the size of batch_size. The upper and lower limits of the two learning rates are set. The upper limit is 0.01, and the lower limit is 0.01 times the maximum learning rate. During the training process, the "cos" method was used to adjust the way the learning rate decreased, and the SGD optimizer was used to train 300 times, 16 images each time.

(2) Evaluation metrics

To verify the quality of the generated infrared ship image, and determine whether the generated image conforms to the characteristics of infrared data in terms of style, content, and features, this paper evaluates the generated image from subjective and objective perspectives. The subjective evaluation method is an evaluation method based on human subjective consciousness, which can visually observe the gap between the generated image and the original image from the image content. Objectively, structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and Frechet Inception Distance (FID) were used for evaluation.

SSIM is a metric used to measure the similarity of two images. Compared with PSNR, SSIM is more in line with human visual characteristics in evaluating image quality. From the perspective of image composition, SSIM defines structural information as independent of brightness and contrast, reflecting the properties of object structure in the scene, and model distortion as a combination of three different factors: brightness, contrast, and structure. The mean is used as the estimation of brightness, the standard deviation is used as the estimation of contrast, and the covariance is used as the measure of structural similarity [45]. The image is generated from the three aspects of brightness, contrast, and structure. The calculation of brightness, contrast and structure in SSIM is shown in Equations (6)–(8).

$$l(X,Y) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_2} \tag{6}$$

$$c(X,Y) = \frac{2\sigma_X\sigma_Y + C_3}{\sigma_X^2 + \sigma_Y^2 + C_2} \tag{7}$$

$$s(X,Y) = \frac{2\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3} \tag{8}$$

In Equation (6), μ_X and μ_Y represent the mean of the original image X and the generated image Y, respectively. In Equation (7), σ_X and σ_Y represent the variance of the original image X and the generated image Y, respectively. In Equation (8), σ_{XY} represents the covariance of the original image X and the generated image Y, and C_1, C_2 and C_3 are constants. The SSIM value is calculated from the above calculation results of brightness, contrast and structure. The specific calculation equations are shown in Equations (9)–(12).

$$\mu_X = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X(i,j) \tag{9}$$

$$\sigma_X^2 = \frac{1}{H \times W - 1}\sum_{i=1}^{H}\sum_{j=1}^{W} \left(X(i,\mathrm{j}) - \mu_X\right)^2 \tag{10}$$

$$\sigma_{XY} = \frac{1}{H \times W - 1}\sum_{i=1}^{H}\sum_{j=1}^{W} \left((X(i,j) - \mu_x)(Y(i,j) - \mu_y)\right) \tag{11}$$

$$SSIM(X,Y) = l(X,Y) \cdot c(X,Y) \cdot s(X,Y) \tag{12}$$

In Equations (9)–(12), H and W represent the length and width of the image. The value range of SSIM is [0, 1]. A larger value of SSIM indicates that the images are more similar and the quality of the generated image is higher.

PSNR is a measure of image quality. Since the PSNR value has limitations, it is generally used to measure the image quality reference value between the maximum signal and the background noise [46]. The degree of distortion of the generated image is evaluated in this paper in comparison to the original image. A larger value of PSNR indicates less image distortion and more realistic generated image content. The calculation equation for PSNR is shown in Equations (13) and (14).

$$PSNR = 10\log_{10}\left(\frac{2^n - 1}{MSE}\right) \tag{13}$$

$$MSE = \frac{1}{H * W}\sum_{i=1}^{H}\sum_{j=1}^{W} \left(X(i,j) - Y(i,j)\right)^2 \tag{14}$$

In Equations (13) and (14), $2^n - 1$ represents the maximum possible pixel value in the image, $X(i,j)$ and $Y(i,j)$ represent the generated image and the original real image, respectively, and H and W represent the length and width of the image respectively.

The infrared image generated in this paper is required for later object detection, and the quality of the image has an impact on the object detection result. As a result, this paper uses FID to evaluate the quality and diversity of generated infrared images based on SSIM and PSNR evaluation metrics [47]. FID is a statistical aspect of the computer vision features of the original image used to measure the similarity between the two sets of images. It is a measure of the distance between the feature vectors of the original real image and the generated image. The distance between the two is calculated by using the mean and covariance matrix, and the calculation equation is shown in Equation (15).

$$FID(x,g) = ||\mu_x - \mu_g||_2^2 + Tr(\sum x + \sum g - (2\sum x \sum g)^{\frac{1}{2}}) \quad (15)$$

μ_x is the mean value of the original real image features, and μ_g is the mean value of the generated image features.$\sum x$ and $\sum g$ are the variances of the original real image features and the generated image features, respectively, and T_r is the trace of the square matrix.

FID is often used to evaluate the quality of images generated by Generative Adversarial Network (GAN). The smaller the FID value, the closer the generated image is to the original real image, or the closer the distribution of the generated image to the distribution of the original real image. If the original real image used for testing has high definition and variety, it means that the generated image has high quality and good diversity. FID is more robust to noise.

This paper selects Precision (P), Recall (R), Average Precision (AP), Mean Average Precision (mAP), and parameter numbers as the evaluation metrics of the object detection method.

5.3. Experiment Results

5.3.1. Comparison Results of Image Generation Ablation Experiments

The effects of different training periods on the generation of the same image are shown in Figure 5. The images generated when epoch was set at 400, 600, 1200, and 1600 were selected, respectively.

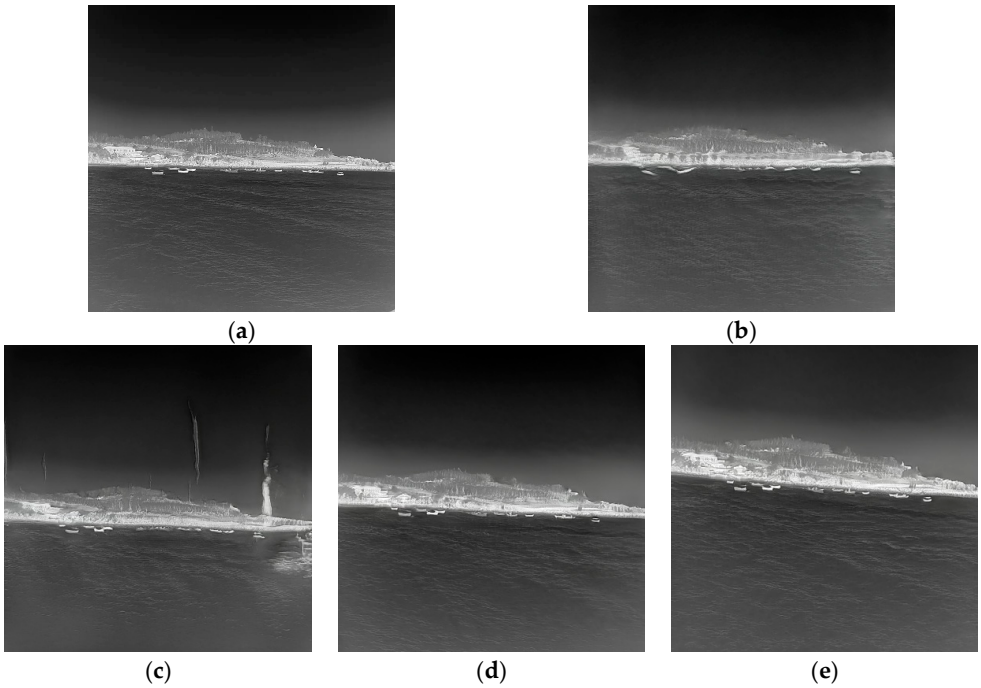


Figure 5. Images generated by training with different epochs: (**a**) original real image; (**b**) epoch =400; (**c**) epoch = 600; (**d**) epoch = 1200; (**e**) epoch = 1600.

From Figure 5, it can be observed that when the model was trained for 400 iterations, the generated images began to exhibit background and object features. By the time the training reached 600 iterations, the generated images started to approximate the shape of the object ship in the original real images, but there were significant artifacts present. The appearance of artifacts in the generated images may be influenced by neighboring ship images within the grid, leading to a phenomenon of concatenation between images.

At 1200 iterations of training, it can be observed that the generated images roughly resembled the ship shapes in the original real images, and the background of the infrared ship image was more detailed. When the training reached 1600 iterations, both in terms of content and style, the generated images were extremely close to the original real images, making it almost impossible to distinguish between the original real images and generated images with the human eye. The ships were noticeably sharper, exhibiting higher visual quality.

The images generated in this paper are diverse, including single ship objects and multiple ship object images in various complex scenes such as tailing scenes, port and seaside scenes, and wave scenes. A comparison between the generated image and the original real image is shown in Figure 6. The left side is the original real image, and the right side is the image generated by the proposed method.

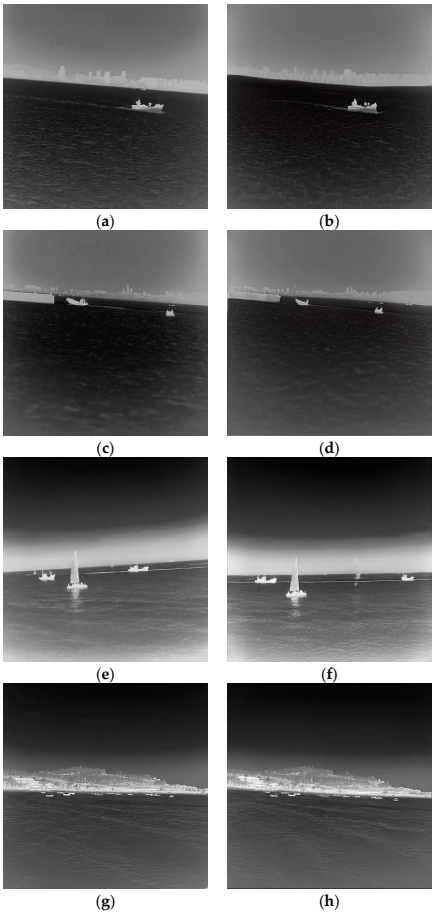
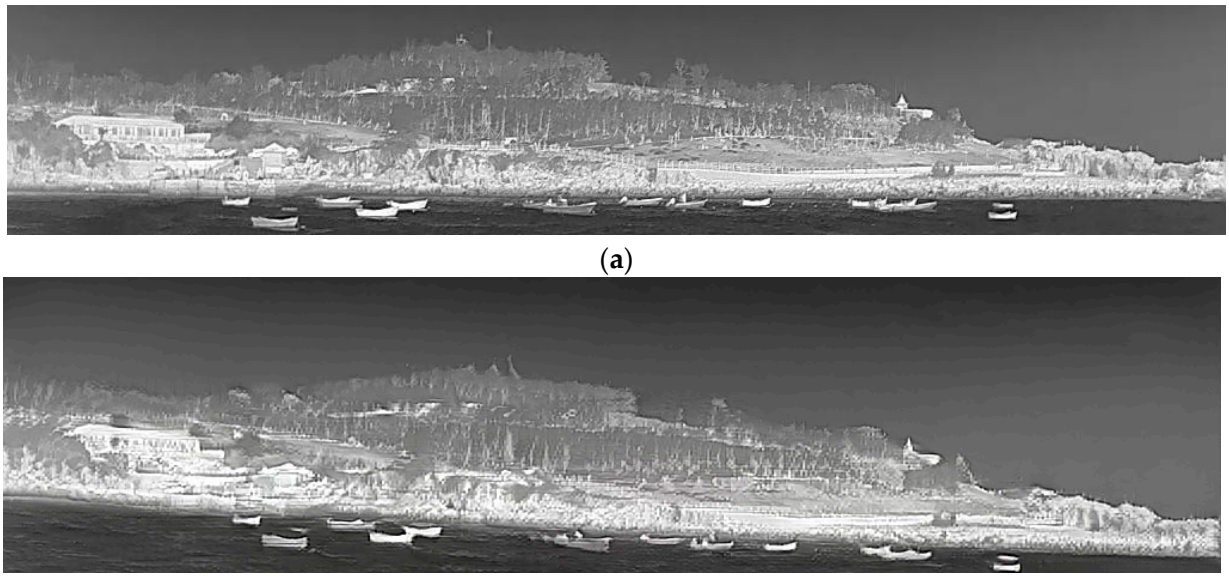


Figure 6. Comparison of the original real image and the generated image: (**a**) The original real image of a single ship object in the trailing scene. (**b**) The image generated by the proposed method. (**c**) The original real images of multiple ship objects in the port. (**d**) The image generated by the proposed method. (**e**) The original real image of multiple ship targets in the wave scene. (**f**) The image generated by the proposed method. (**g**) The original real image of multiple ship targets in the seaside scene. (**h**) The image generated by the proposed method.

From the four sets of comparison images in Figure 6, it can be observed that in the first set, the generated image has less clear background texture compared to the original real image, and there is a phenomenon of the texture sticking together. In the second set, the edges of the ship in the generated image are not well generated, and the rear part of the ship shows a poor generation effect. In the third set, the generated image contains a small amount of artifacts, but when looking at the sailboat class in the generated images, it is almost the same as the original real image. In the fourth set, the generated image contains multiple small ship objects, and overall, the shape and position of the ship object are basically consistent with the original real image.

In order to further observe and compare, this paper took partial images of only the ship and background from the comparison images in Figure 6g,h. The local detailed image after interception is shown in Figure 7.



(**a**)

(**b**)

Figure 7. The local details of the fourth set of comparison figures: (**a**) Partial detail-enlarged image of the original real image; (**b**) partial detail-enlarged image of the presented method.

Figure 7 shows that the image generated by the proposed method is very close to the original real image in terms of the shape, position and background of the object. The generated image contains all the contents of the original real image. Almost all the ship classes of small objects are generated. The corresponding position, the shape of the ship, and the background part are relatively real, and the generated image is very similar in terms of the overall brightness to the original real image. However, detailed parts of the image background are not completely generated. For instance, the windows of the houses in the original real images are clearly visible, while the windows in the generated images are somewhat blurry. The experiment environment is the same for all the ablation experiments conducted in this paper. The results of the ablation experiments are given in Table 2.

Table 2. Ablation experiment results.

Methods	SSIM	PSNR	FID
StyleGAN2	0.82	22.86	55.60
VAE + StyleGAN2	0.84	23.52	54.39
StyleGAN2 + SA	0.83	23.34	47.66
Proposed method	0.86	24.84	40.49

From Table 2, it can be observed that the SSIM and PSNR values improved after replacing the mapping network in the original StyleGAN2 with VAE. This suggests that in the process of generating infrared remote sensing images, the hidden variables containing the original image information have a greater influence on the image generation in the style network compared to the decoupled noise. The addition of the SA attention mechanism to the original StyleGAN2 network also leads to improvements in the SSIM and PSNR values, but the increase is smaller. When VAE or the SA is added separately to the StyleGAN2 network, the FID values are lower than those of the original StyleGAN2 network. This indicates that the images generated by the StyleGAN2 network with the addition of VAE or the SA are closer to the original images, with the SA yielding better results.

The proposed method incorporates both VAE and SA mechanisms. On the same experiment data set, the SSIM and PSNR values of the proposed method are higher than those of the original StyleGAN2 network and the StyleGAN2 network with either VAE or SA mechanisms. Moreover, the FID value of the proposed method is significantly lower than those of the original StyleGAN2 network and the StyleGAN2 network with either VAE or SA. This indicates that the proposed method performs better in terms of image style and diversity generation.

The infrared ship data set contains diverse image scenes with different shapes and varying numbers of objects. Different generative algorithms yield different results for image generation in different scenes. Objective evaluations were conducted on the generated images in various scenes, and the evaluation results are shown in Table 3.

Table 3. Evaluation results of different image generation methods.

Methods	SSIM			PSNR		
	Single-Objective	Multi-Objective	Complex Scenarios	Single-Objective	Multi-Objective	Complex Scenarios
StyleGAN2	0.73	0.63	0.23	25.37	18.58	14.80
VAE + StyleGAN2	0.80	0.77	0.28	29.71	26.97	15.06
StyleGAN2 + SA	0.75	0.67	0.25	27.62	19.89	15.33
Proposed method	0.85	0.81	0.31	28.11	27.87	20.85

It can be seen from Table 3 that the SSIM value and PNSR value of the proposed method are higher than the SSIM value and PNSR value of the original StyleGAN2 network in single-objective, multi-objective and complex scenarios. It shows that the proposed method can effectively improve generalization performance and generate diverse infrared remote sensing ship images.

5.3.2. The Experiment Comparison Results and Analysis of the Proposed Method and Mainstream Methods

In terms of visual perception evaluation, the proposed method is compared with StyleGAN and StyleGAN2. Figure 8 shows comparative results of the original real image, StyleGAN, StyleGAN2, and the image generated by the proposed method when training 1600 times under the same experimental environment and experiment parameter settings.

From Figure 8, it can be observed that the StyleGAN method produces inferior background effects compared to the original real image, and backgrounds are generated that do not exist, which may be caused by the generation of artifacts during the training process. The generated images from the StyleGAN2 method clearly show that the details of the ship itself are not well generated, and the background of the generated image is relatively blurry compared to the original real image background. In comparison to the original real image, the generated images from the proposed method have a closer background and similar ship details, showing the highest similarity to the original real image. Compared with StyleGAN and StyleGAN2, the proposed method generates better infrared remote sensing ship images.

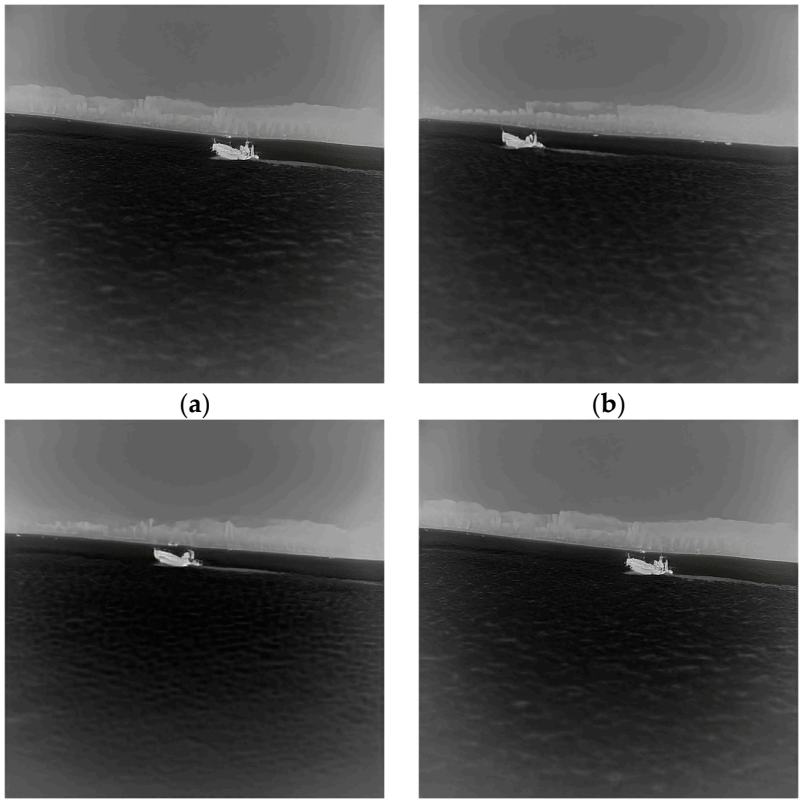


Figure 8. The images generated by different methods: (**a**) the original real image; (**b**) the image generated by StyleGAN; (**c**) the image generated by StyleGAN2; (**d**) the image generated by the proposed method.

Considering the evaluation metrics of SSIM, PSNR and FID, the proposed method is compared with the mainstream methods ProGAN, StyleGAN and StyleGAN2 for generating high-resolution images. The comparison results are shown in Table 4.

Table 4. Comparison of evaluation metrics.

Methods	SSIM	PSNR	FID
ProGAN	0.76	19.51	64.40
StyleGAN	0.78	22.55	70.19
StyleGAN2	0.82	22.86	55.60
Proposed method	0.86	24.84	40.49

From Table 4, it can be observed that the proposed method performs better than the other methods in terms of SSIM, PSNR, and FID evaluation metrics. Specifically, the SSIM value of the proposed method is 0.04 higher than that of the StyleGAN2 method, 0.08 higher than that of the StyleGAN method, and 0.1 higher than that of the ProGAN method. In terms of the PSNR metric, the proposed method has a value of 24.84, which is higher than the values of the other three methods. This indicates that the proposed method performs better in generating infrared remote sensing ship images, and the generated images are closer to original real images. Compared with the proposed method, ProGAN and StyleGAN2, the FID value of StyleGAN is higher, which indicates that the diversity of images generated by StyleGAN is poor, which is related to the single color and limited clarity of infrared remote sensing images.

Comprehensive subjective evaluation, as well as SSIM, PSNR, and FID objective evaluation indicators, all verify that the generated image of the proposed method has a

better effect, which is more similar to the original real image in terms of structure, content, distribution, and style, and is closer to the original real image in terms of detail.

5.3.3. Experiment Comparison Results and Analysis of Lightweight Multi-Class Ship Detection Method in Infrared Remote Sensing Images

The detection results of three typical single-type multi-targets, multi-type multi- targets, and multi-type multi-targets moored at dock on a vast sea surface are shown in Figure 9. It can be seen from (a) and (b) that both methods can detect ships, but in contrast, the detection rate of the proposed method is higher than that of YOLOv5l. It can be seen from (c) and (d) that the proposed method offers a great improvement in the detection of small objects compared with YOLOv5l, indicating that the proposed method has a strong capacity for the infrared remote sensing of small object. Observing (e) and (f), it can be seen that the proposed method detects large objects that are not detected in the original model and some mutually occluded objects, indicating that the method also has better detection ability for occluded ship objects. For the small ship objects missed by the original YOLOv5l, the missed detection of large objects in complex scenes, and the occluded ship objects, the proposed method can achieve accurate detection, indicating that the method is robust in small objects and complex scenes.

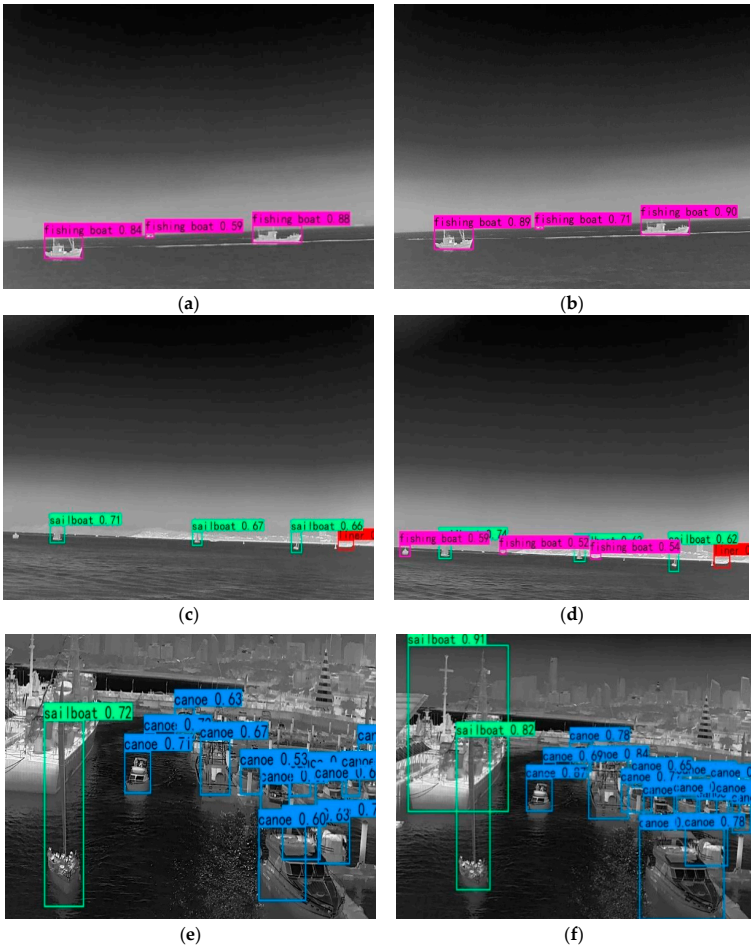


Figure 9. The ship detection results of YOLOv5l and the proposed method: (**a**) The detection results of YOLOv5l. (**b**) The detection results of the proposed method. (**c**) The detection results of YOLOv5l. (**d**) The detection results of the proposed method. (**e**) The detection results of YOLOv5l. (**f**) The detection results of the proposed method. Different colors represent different types of ships.

5.3.4. Experiment Results and Analysis of the Ship Detection Interpretability Method Based on Causal Reasoning

We visualized the sorting of image singular values, and the results are shown in Figure 10. The abscissa represents the sorting, and the ordinate is the singular value. It can be seen from Figure 10 that the main information of the image is basically distributed between 0 and 10, and the information is basically stable after 10. It can be seen from this that most of the information of the image is contained in a small number of singular values. For an image, finding the boundary of the singular value accurately can reduce the dimension and compress the image without affecting the structure of the image.

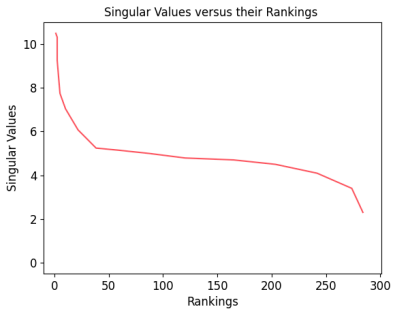
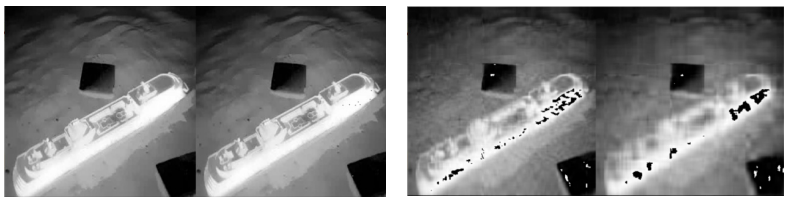


Figure 10. Visualization result of image singular value.

Due to the limited space, this paper only gives different singular values of a ship image, as shown in Figure 11. It can be seen from Figure 11 that when the singular value $k = 200$, the ship in the image is clear and close to the original image. When $k = 150$, black dots begin to appear on the contour of the ship. When $k = 20$, although the shape of the ship can still be seen in the image, the clarity is obviously reduced, and the contour of the ship begins to blur. When $k = 10$, the image is very unclear, the contour of the ship is blurred and a continuous black shape appears. In this paper, after many experiments, different images are selected for image reconstruction, and the k value is finally selected as 200.



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

Figure 11. The reconstructed images for different values of k: (**a**) The reconstructed images for singular values of $k = 200$ and $k = 150$. (**b**) The reconstructed images for singular values of $k = 20$ and $k = 10$.

The training set loss and validation set loss curves of SVD combined with Transformer are shown in Figure 12. It can be seen from Figure 12 that after the low-rank decomposition of the image, combined with the deep learning model, the accuracy of the recognition results is higher than that of the original Transformer model. The training set loss and verification set loss of the combination method of SVD and Transformer show a downward trend as a whole. Among them, although the loss of the training set has occasionally increased during the training process, it has shown a downward trend as a whole. The loss of the validation set first decreased and then increased slightly, but after 50 times, the overall trend was slowly decreasing. The training set loss curve of the Transformer model itself shows a problem of first decreasing and then increasing. After training about 170 times, the phenomenon of sudden drop and sudden rise occurs. The verification set loss is also rising and falling as a whole. In summary, it can be seen that the method of combining SVD with Transformer in this paper has better loss than the original Transformer

model on both the training set and the verification set, indicating that the method achieves model compression while maintaining model recognition performance.

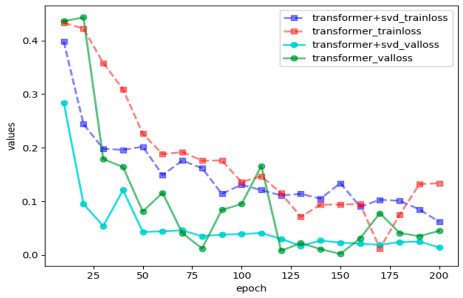


Figure 12. The loss curves of validation set and test set.

In Figure 13, three sets of Transformers can be seen, along with the ship recognition results of the combination of SVD and Transformer. After the compression of image information, only most of the information of the image is retained, which improves the accuracy of recognition. Figure 13 shows that the important information about the ship contained in the image plays a key role in the recognition process, and the ship recognition results are obtained based on this crucial information. By making the model pay more attention to the causal features related to the label in the image, the robustness of the model to non-causal information such as background information is improved, and the interpretability of the recognition results is improved.

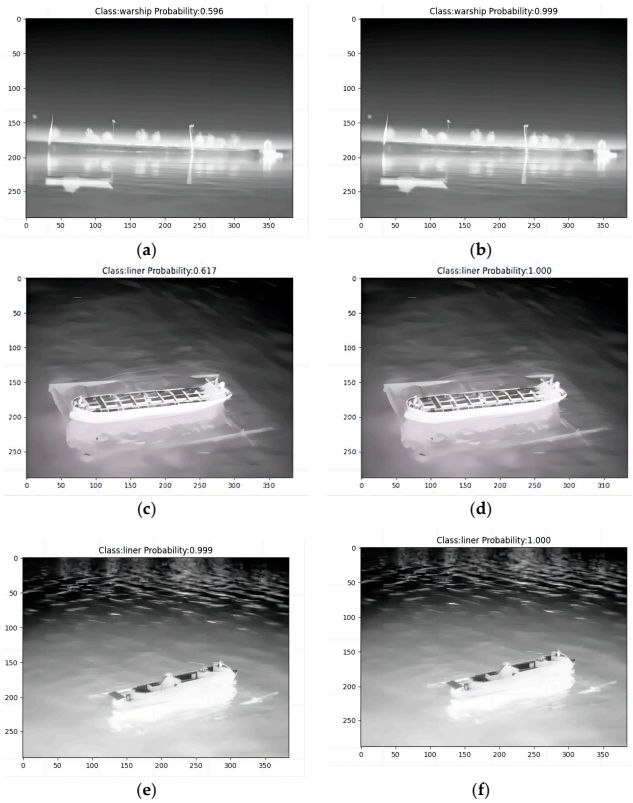


Figure 13. The ship detection results of YOLOv5l and the proposed method: (**a**) The recognition results of Transformer. (**b**) The recognition results of the proposed method. (**c**) The recognition results of Transformer. (**d**) The recognition results of the proposed method. (**e**) The recognition results of Transformer. (**f**) The recognition results of the proposed method.

6. Conclusions

Addressing the scarcity of public data on infrared ships resulting from the sensitivity and inaccessibility of infrared images, this paper proposes an enhanced StyleGAN2 network image generation method. The encoder is used to replace the mapping network so that the generated hidden variables carry the original image information. The SA mechanism is introduced into the generative network to obtain the dependence information between distant features. The effect of the image generated by the proposed method is better, which is more similar to the original image in terms of structure, content, distribution and style, and is closer to the original real image in detail.

In view of the problems posed by the large number of parameters and poor real-time performance of deep learning models, this paper designs a multi-class lightweight infrared remote sensing ship detection method, which achieves better detection results for multi-class ship objects in complex scenes. An interpretability method for ship detection in infrared remote sensing images based on causal reasoning is proposed, which combines low-rank decomposition with the Transformer to improve the interpretability of the model. The experimental results show that the method achieves model compression while maintaining model recognition performance. By making the model pay more attention to the causal features related to the label in the image, the robustness of the model to non-causal information such as background information is improved, and the interpretability of the recognition results is improved. In the future, the data set will be further enriched, more complex scenes will be added, and the ship detection performance using infrared remote sensing images under different illumination, weather, and terrain will be improved.

Author Contributions: Conceptualization, methodology, data curation, validation, writing original draft preparation R.L., Y.Z., Z.D. and Q.Y.; writing review and editing, Z.D.; supervision, project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Fund of China (61371143) and the Graduate Education Reform Project in North China University of Technology (217051360023XN269-20).

Data Availability Statement: The data can be obtained by the corresponding author on reasonable request.

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

References

1. Chang, L.; Chen, Y.T.; Wang, J.H.; Chang, Y.L. Modified Yolov3 for ship detection with visible and infrared images. *Electronics* **2022**, *11*, 739. [CrossRef]
2. Huang, Y.; Liu, R.W.; Liu, J. A two-step image stabilization method for promoting visual quality in vision-enabled maritime surveillance systems. *IET Intell. Transp. Syst.* **2023**, *17*, 435–449. [CrossRef]
3. Zhang, Z.; Gao, Q.; Liu, L.; He, Y. A high-quality rice leaf disease image data augmentation method based on a dual GAN. *IEEE Access* **2023**, *11*, 21176–21191. [CrossRef]
4. Yang, W.J.; Chen, B.X.; Yang, J.F. CTDP: Depacking with guided depth upsampling networks for realization of multiview 3D video. In Proceedings of the Future of Information and Communication Conference, San Francisco, CA, USA, 2–3 March 2023; pp. 136–152.
5. Tan, M.K.; Xu, S.K.; Zhang, S.H.; Chen, Q. A review on deep adversarial visual generation. *J. Image Graph.* **2021**, *26*, 2751–2766.
6. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
7. Abu-Srhan, A.; Abushariah, M.A.M.; Al-Kadi, O.S. The effect of loss function on conditional generative adversarial networks. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 6977–6988. [CrossRef]
8. Gao, H.; Zhang, Y.; Lv, W.; Yin, J.; Qasim, T.; Wang, D. A deep convolutional generative adversarial networks-based method for defect detection in small sample industrial parts images. *Appl. Sci.* **2022**, *12*, 6569. [CrossRef]
9. Phan, H.; Nguyen, H.L.; Chen, O.Y.; Koch, P.; Duong, N.Q.K.; McLoughlin, I.; Mertins, A. Self-attention generative adversarial network for speech enhancement. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7103–7107.
10. Chan, E.R.; Lin, C.Z.; Chan, M.A.; Nagano, K.; Pan, B.; Mello, S.; Gallo, O.; Guibas, L.; Tremblay, J.; Khamis, S.; et al. Efficient geometry-aware 3D generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16102–16112.

11. Brophy, E.; Wang, Z.; She, Q.; Ward, T. Generative adversarial networks in time series: A systematic literature review. *ACM Comput. Surv.* **2023**, *55*, 1–31. [CrossRef]
12. Han, Y.; Liao, J.; Lu, T.; Pu, T.; Peng, Z. KCPNet: Knowledge-driven context perception networks for ship detection in infrared imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000219. [CrossRef]
13. Kawai, N.; Koike, H. Facial mask completion using StyleGAN2 preserving features of the person. *IEICE Trans. Inf. Syst.* **2023**, *106*, 1627–1637. [CrossRef]
14. Li, L.; Yu, J.; Chen, F. TISD: A three bands thermal infrared dataset for all day ship detection in spaceborne imagery. *Remote Sens.* **2022**, *14*, 5297. [CrossRef]
15. Zhang, Y.M.; Li, R.Q. A lightweight multi-target detection method for infrared remote sensing image ships. *J. Netw. Intell.* **2023**, *8*, 535–545.
16. Turk, M.A.; Pentland, A.P. Face recognition using eigenfaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, 3–6 June 1991; pp. 586–591.
17. Comon, P. Independent component analysis, a new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
18. Permuter, H.; Francos, J.; Jermyn, I.H. Gaussian mixture models of texture and colour for image database retrieval. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 21 May 2003; pp. 569–573.
19. Rabiner, L.; Juang, B. An introduction to hidden markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16. [CrossRef]
20. Cross, G.R.; Jain, A.K. Markov random field texture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **1983**, *5*, 25–39. [CrossRef] [PubMed]
21. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
22. Fang, Z.; Fu, Y.; Liu, L.X. A dual of transformer features-related map-intelligent generation method. *J. Image Graph.* **2023**, *28*, 3281–3294.
23. Huang, S.Y.; Wu, W.; Yang, Y.; Li, H.X.; Wang, B. A low-exposure image enhancement based on progressive dual network model. *Chin. J. Comput.* **2021**, *44*, 384–394.
24. Wang, Y.H.; He, Y.; Wang, Z. Overview of text-to-image generation methods based on deep learning. *Comput. Eng. Appl.* **2022**, *58*, 50–67.
25. Nishio, M. Machine learning/deep learning in medical image processing. *Appl. Sci.* **2021**, *11*, 11483. [CrossRef]
26. Huang, M.; Mao, Z.; Chen, Z.; Zhang, Y. Towards accurate image coding: Improved autoregressive image generation with dynamic vector quantization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22596–22605.
27. Mak, H.W.L.; Han, R.; Yin, H.H.F. Application of Variational AutoEncoder (VAE) model and image processing approaches in game design. *Sensors* **2023**, *23*, 3457. [CrossRef]
28. Zhou, T.; Li, Q.; Lu, H.; Cheng, Q.; Zhang, X. GAN review: Models and medical image fusion applications. *Inf. Fusion* **2023**, *91*, 134–148. [CrossRef]
29. Ho, J.; Saharia, C.; Chan, W.; Fleet, D.J.; Norouzi, M.; Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **2022**, *23*, 1–33.
30. Wang, S.M. Research on Intelligent Detection Technology of Optical Fiber End Face Based on Feature Fusion. Guangdong University of Technology, Guangzhou, China, 2020.
31. Li, L.; Jiang, L.; Zhang, J.; Wang, S.; Chen, F. A complete YOLO-based ship detection method for thermal infrared remote sensing images under complex backgrounds. *Remote Sens.* **2022**, *14*, 1534. [CrossRef]
32. Miao, C.K.; Lou, S.L.; Gong, W.F. Infrared ship target detection algorithm based on improved centernet. *Laser Infrared* **2022**, *52*, 1717–1722.
33. Karras, T.; Aittala, M.; Laine, S.; Harkonen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 852–863.
34. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
35. Li, J.N.; Xiong, R.B.; Lan, Y.Y.; Pang, L.; Guo, J.F.; Cheng, X.Q. Overview of the frontier progress of causal machine learning. *J. Comput. Res. Dev.* **2023**, *60*, 59–84.
36. Cui, P.; Athey, S. Stable learning establishes some common ground between causal inference and machine learning. *Nat. Mach. Intell.* **2022**, *4*, 110–115. [CrossRef]
37. Shao, F.; Luo, Y.; Zhang, L.; Ye, L.; Tang, S.; Yang, Y.; Xiao, J. Improving weakly supervised object localization via causal intervention. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 20–24 October 2021; pp. 3321–3329.
38. Gao, G.; Li, X.; Du, Z. Custom attribute image generation based on improved StyleGAN2. In Proceedings of the 2023 15th International Conference on Machine Learning and Computing, New York, NY, USA, 17–20 February 2023; pp. 335–340.
39. Sundar, S.; Sumathy, S. An effective deep learning model for grading abnormalities in retinal fundus images using Variational Auto-Encoders. *Int. J. Imaging Syst. Technol.* **2023**, *33*, 92–107. [CrossRef]

40. Li, Y.Z.; Wang, Y.; Huang, Y.H.; Xiang, P.; Liu, W.-X.; Lai, Q.Q.; Gao, Y.Y.; Xu, M.S.; Guo, Y.F. RSU-Net: U-net based on residual and self-attention mechanism in the segmentation of cardiac magnetic resonance images. *Comput. Methods Programs Biomed.* **2023**, *231*, 107437. [CrossRef]
41. Mi, Z.; Jiang, X.; Sun, T.; Xu, K. GAN-generated image detection with self-attention mechanism against gan generator defect. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 969–981. [CrossRef]
42. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 122–138.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
44. Yu, Y.; Buchanan, S.; Pai, D.; Chu, T.; Wu, Z.; Tong, S.; Haeffele, B.D.; Ma, Y. White-Box Transformers via Sparse Rate Reduction. *arXiv* **2023**, arXiv:2306.01129.
45. Wang, H.; Li, Y.; Ding, S.; Pan, X.; Gao, Z.; Wan, S.; Feng, J. Adaptive denoising for magnetic resonance image based on nonlocal structural similarity and lowrank sparse representation. *Clust. Comput.* **2023**, *26*, 2933–2946. [CrossRef]
46. Dziembowski, A.; Mieloch, D.; Stankowski, J.; Grzelka, A. IV-PSNR:the objective quality metric for immersive video applications. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7575–7591. [CrossRef]
47. Lee, J.; Lee, M. FIDGAN: A generative adversarial network with an inception distance. In Proceedings of the 2023 International Conference on Artificial Intelligence in Information and Communication, Bali, Indonesia, 20–23 February 2023; pp. 397–400.

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.