applied sciences

MDPI

Article

SAITI-DCGAN: Self-Attention Based Deep Convolutional Generative Adversarial Networks for Data Augmentation of Infrared Thermal Images

Zhichao Wu [1], Changyun Wei [1,*], Yu Xia [1] and Ze Ji [2]

[1] College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213200, China; wuzhichao656@gmail.com (Z.W.); mr_xia_yu@163.com (Y.X.)
[2] School of Engineering, Cardiff University, Cardiff CF24 3AA, UK; jiz1@cardiff.ac.uk
* Correspondence: c.wei@hhu.edu.cn

Abstract: Defect detection plays a crucial role in industrial production, and the implementation of this technology has significant implications for improving both product quality and processing efficiency. However, the limited availability of defect samples for training deep-learning-based object detection models within industrial processes poses challenges for model training. In this paper, we propose a novel deep convolutional generative adversarial network with self-attention mechanism for the data augmentation of infrared thermal images for the application of aluminum foil sealing. To further expand its applicability, the proposed method is designed not only to address the specific needs of aluminum foil sealing but also to serve as a robust framework that can be adapted to a wide range of industrial defect detection tasks. To be specific, the proposed approach integrates a self-attention module into the generator, adopts spectral normalization in both the generator and discriminator, and introduces a two time-scale update rule to coordinate the training process of these components. The experimental results validated the superiority of the proposed approach in terms of the synthesized image quality and diversity. The results show that our approach can capture intricate details and distinctive features of defect images of aluminum foil sealing. Furthermore, ablation experiments demonstrated that the combination of self-attention, spectral normalization, and two time-scale update rules significantly enhanced the quality of image generation, while achieving a balance between stability and training efficiency. This innovative framework marks a notable technical breakthrough in the field of industrial defect detection and image synthesis, offering broad application prospects.

Keywords: infrared thermography; aluminum foil sealing; data augmentation; self-attention; GAN

Citation: Wu, Z.; Wei, C.; Xia, Y.; Ji, Z. SAITI-DCGAN: Self-Attention Based Deep Convolutional Generative Adversarial Networks for Data Augmentation of Infrared Thermal Images. *Appl. Sci.* **2024**, *14*, 11391. https://doi.org/10.3390/app142311391

Academic Editor: Gianluigi Ferrari

Received: 30 October 2024
Revised: 5 December 2024
Accepted: 5 December 2024
Published: 6 December 2024

1. Introduction

With advancements in industrial manufacturing capabilities, the utilization of aluminum foil packaging has become widespread and convenient across various sectors, including food packages, medical bottles, and chemical containers. This packaging technique involves the application of heat to the aluminum foil, resulting in localized melting and a viscous state at the bottle mouth, as depicted in Figure 1. Following the application of a specific pressure until the sealing process is complete, the foil cools down and forms a tight seal. Nonetheless, issues such as loose seals and leakage commonly arise during the aluminum foil sealing process. Detecting these minuscule defects and liquid leaks with the naked eye within a short time frame poses difficulties. Improper sealing can lead to product degradation, compromising consumer safety and resulting in substantial economic losses. For instance, a single bottle leaking can cause entire batches to be rejected, and in the case of pesticide leakage, it can pose serious health risks to personnel handling the goods, potentially endangering their lives. Therefore, even the rarest of defects must be strictly avoided to prevent such occurrences. Consequently, conducting comprehensive sealing tests on aluminum foil packaging is of utmost importance. By implementing suitable testing techniques, such as infrared imaging technology, the sealing performance of aluminum foil can be accurately evaluated, thus ensuring product integrity and safety. (This not only protects consumers from potential health risks but also minimizes the economic losses associated with product recalls and contamination incidents.)

Food package
PET medicine bottle
Cosmetic cream cap
Milk bottle
Motor oil cap
Chemical containers

Figure 1. Diverse applications of aluminum foil sealing in industrial packaging.

Traditional methods for detecting the closure area of aluminum foil predominantly rely on manual inspections or single-point temperature measurements, as well as water detection methods, among others. Within the confines of a bottle, it is impossible to visually ascertain the presence of aluminum foil without unscrewing the cap. Nevertheless, this approach not only elevates production costs but also diminishes processing efficiency. On the other hand, the single-point temperature measurement method merely identifies the absence of aluminum foil and fails to assess the sealing performance of all contact surfaces, potentially resulting in missed detections. To effectively evaluate the sealing performance of the complete closed area, encompassing all contact surfaces of the aluminum foil sealing, an alternative solution utilizing infrared imaging technology can be employed. Infrared imaging enables comprehensive visualization of the shape and temperature distribution of the measured thermal images, addressing the shortcomings associated with the traditional methods of tightness detection in aluminum foil seals, including limited accuracy, high labor costs, and inherent defects.

In recent years, deep-learning-based methods have proven to be effective for defect detection, due to their ease of use, cost-effectiveness, real-time capabilities, and accurate target detection [1–3]. However, similarly to other supervised deep learning approaches, object detection requires a significant amount of training data. To achieve high accuracy and generalization, an algorithm needs to be trained on diverse image datasets that contain various types of defects, allowing it to learn relevant features effectively. Unfortunately, in many industrial environments [4,5], the majority of products meet production standards, and equipment typically remains in normal condition throughout its life-cycle. As a result, acquiring a substantial amount of labeled defect data becomes a challenging and time-consuming task, often accompanied by high costs or even feasibility issues. Effective neural network training for defect detection usually requires hundreds to thousands of images to achieve robust performance. However, production facilities for pesticides and pharmaceuticals are often located in remote areas, making it logistically challenging to collect and label large datasets. Moreover, the occurrence of defects is relatively rare, meaning that even when access to production sites is possible, the number of defect samples available may still be limited. Therefore, overcoming the challenge of limited defect sample size is critical for the successful implementation of object detection in defect detection applications.

Researchers have devoted considerable efforts to addressing the issue of inadequate samples in industrial settings by developing data augmentation methods aimed at enhancing accuracy. Wan et al. [6] utilized a contrastive learning model to tackle the issue of insufficient samples. Cao et al. [7] applied a domain-shared convolutional neural network (CNN) to overcome the challenge of limited samples in machine fault diagnosis, particularly in cases with time-varying speed. Zhang et al. [8] introduced a prototype matching network model to handle cross-domain diagnosis with limited data. In a similar vein, Hu et al. [9] proposed a meta-learning model for task ranking, providing a solution for fault diagnosis with a small sample size. However, many of the aforementioned studies on insufficient samples heavily depended on the data distribution or learning strategies from other fields [10]. To address data scarcity, generative adversarial networks (GANs) have emerged as a commonly used data augmentation method [11]. Gao et al. [12] presented a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) to augment low-dimensional fault data. Chen et al. [13] developed an end-to-end generative adversarial network (PadGAN) for generating low b-value diffusion magnetic resonance imaging (dMRI) data, improving the data quality and detail of macaque brain dMRI images. Shang et al. [14] introduced a GAN-based method with fused attention and perceptual quality enhancement to improve the quality and detail of solar coronal images, enhancing the observation and prediction of solar activities. Wang et al. [15] proposed a super-resolution image reconstruction method using cascaded generative adversarial networks (GANs) for Sentinel-2 and Gaofen-2 images, enhancing image quality and detail. Liu et al. [16] proposed a CVAEGAN method for fault diagnosis, incorporating self-modulation into the generator based on the input and feedback from the discriminator. A conditional variational autoencoder (AE) was integrated into the conditional Wasserstein GAN to generate higher-quality synthetic images. Nevertheless, GANs suffer from two major drawbacks: inconspicuous defect features and unstable training. The generator is not incentivized to learn the difference between local and global features, resulting in inconspicuous features in the synthesized images. Additionally, during training, the density ratio estimation of the discriminator is inaccurate and unstable, impeding the generator's ability to effectively learn the multimodal structure of the target distribution. Consequently, the GAN framework is inherently unstable and susceptible to loops during training. Further research and innovation are necessary to address these challenges and improve the effectiveness of data augmentation techniques in industrial fault diagnosis.

To address the challenges outlined above, this study proposes a novel deep convolutional generative adversarial network with self-attention mechanism, named SAITI-DCGAN, to facilitate the efficient and high-quality generation of defect images for aluminum foil sealing. This approach addresses the issue of insufficient training samples of equipment or product defects in industrial processes, accomplishing this through the synthesis of image samples. Our approach can leverage the knowledge acquired from existing defect samples by integrating an enhanced deep convolutional generative adversarial network, self-attention mechanism learning, spectrum normalization, and data augmentation via two scale-time update rules. This synthesis results in the autonomous creation of novel defect sample data, without relying on external information from other domains. The primary contributions of this work are summarized as follows:

- The proposed approach integrate a self-attention mechanism into the generator of GANs to assess characteristics through weight assignment and selective information extraction. The objective is to improve the fidelity of infrared thermal images and produce synthetic counterparts that possess a heightened sense of realism.
- The adoption of spectral normalization entails a weight normalization that is seamlessly integrated with network adjustment. By imposing Lipschitz constraints on the parameter matrices of both the generator and discriminator, this approach effectively enhances the stability of the training process.
- The model incorporates a two time-scale update rule to expedite the generator's adaptation process in response to the discriminator's feedback. This integration aims to

improve the equilibrium between the generator and the discriminator. Consequently, it can accelerate the overall convergence, reduce the training duration, mitigate the likelihood of mode collapse, and foster increased diversity in the outputs generated by the generator.

The remainder of this paper is structured as follows: Section 2 introduces the preliminaries of data augmentation, the GAN architecture, and the attention mechanism. Section 3 provides a detailed description of the proposed approach. In Section 4, we conduct experimental and comparative verification analyses of the proposed model. Section 5 discusses the implications and insights gained from the results. Section 6 addresses potential threats to the validity of our study. Finally, we conclude the work in Section 7.

2. Preliminaries

2.1. Data Augmentation

Data augmentation is the most direct and effective way to increase the number of data samples. Methods can be divided into two categories: image mixing and image generation. In the first category, new images are generated by randomly erasing existing pixels [17,18] or mixing with other image information, such as pixels [19,20], patches [21,22], and manifold structures [23,24]. These methods usually do not solely rely on deep models or lightweight models. With regard to image generation, GAN-based generative models [25] have been widely used in the field of data augmentation. GANs can decouple images into the structure and style of latent layers. The decoupled information can be reconstructed to generate a new image. Our proposed approach also belongs to the method of image generation. The difference is that our approach incorporates a self-attention network, resulting in remarkably diverse generated images.

2.2. Deep Convolutional Generative Adversarial Networks

GAN models are widely employed across computer vision, natural language processing, and other fields. The primary structure of GANs comprises a generator, denoted as G, and a discriminator, denoted as D, as illustrated in Figure 2.

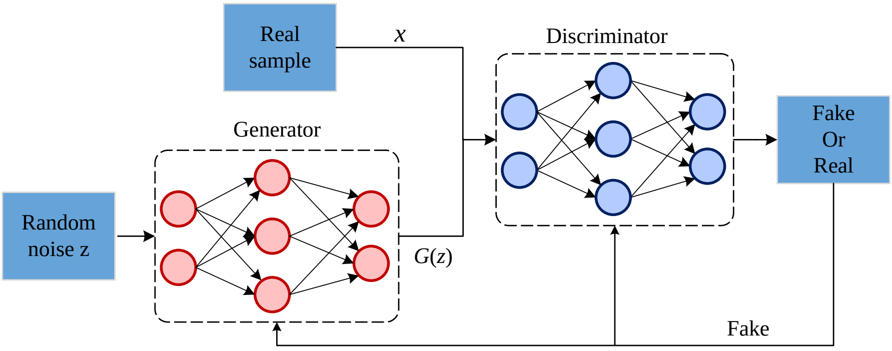


Figure 2. The basic structure of generative adversarial networks (GANs).

The typical process of sample generation unfolds as follows. Initially, noise z is input into the generator G to produce synthetic data $G(z)$, which resemble real data x, with the intention of deceiving the discriminator. Subsequently, both the generated synthetic data and real samples are fed into the initial discriminator D until the discriminator becomes unable to distinguish between the genuine and synthetic data. It is commonly assumed that random noise z follows a Gaussian distribution. Ultimately, if the input data from the generator closely approximate the real data, the goal of learning the approximate distribution of genuine images is achieved.

GANs are non-cooperative game-based models optimized through the maximization and minimization of probabilistic outputs. The generator and discriminator iteratively update their model parameters until reaching a Nash equilibrium state. This state refers to a condition where neither the generator nor the discriminator can improve their performance by unilaterally changing their strategies. Specifically, this means that the generator produces samples that the discriminator cannot reliably distinguish from real data, while the discriminator maximizes its ability to differentiate between real and generated data given the current capabilities of the generator. This concept is fundamental to understanding the training dynamics of GANs. In this state, the generator is capable of producing samples akin to real data, while the discriminator struggles to accurately discern the authenticity of the generator's input data. This process can be formalized as a non-cooperative game that involves maximizing and minimizing the value function, denoted as $V(D, G)$. The generator's objective is to maximize this value function, making it challenging for the discriminator to differentiate between samples generated by the generator and real data. Conversely, the discriminator aims to minimize the value function, enhancing its ability to distinguish between generator-produced samples and genuine data. The objective function is expressed as follows:

$$\begin{aligned}\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] \\ + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))],\end{aligned} \tag{1}$$

where $p_{data}(x)$ denotes the distribution of real data x, $P_z(z)$ indicates the prior distribution of the noise vector z, and $D(x)$ and $D(G(z))$ represent the probabilities of real data and generated data, respectively.

However, GANs can be unstable during training, leading to instances where the generator produces invalid outputs. A DCGAN [26] shares the same goal as traditional GANs. The primary distinction lies in the fact that a DCGAN utilizes multiple convolutional blocks to construct its generator and discriminator. Specifically, the generator employs a deconvolutional architecture, while the discriminator employs a convolutional architecture. The structures of the generator and discriminator of the DCGAN are illustrated in Figure 3.

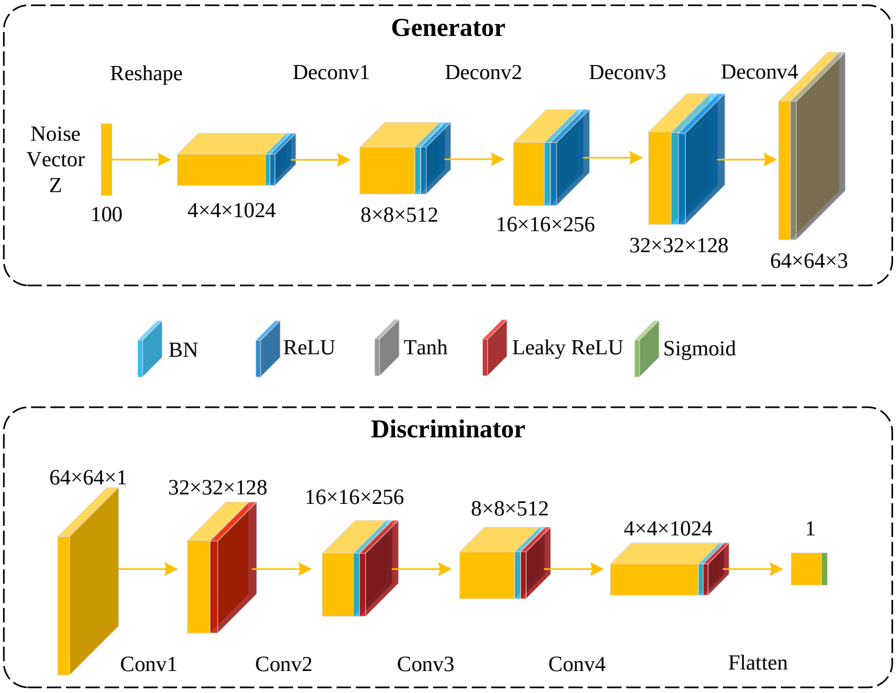


Figure 3. The structures of the generator and discriminator of a DCGAN.

We can see that BN indicates a batch normalization operation, ensuring that its mean and unit variance are zero, thereby stabilizing learning. ReLU, leaky ReLU, and tanh represent three distinct activation functions that aim to expedite model learning and saturate the color space for comprehensive coverage. The aforementioned structure possesses a remarkable capability to generate high-quality images.

2.3. Attention Models

Recently, attention mechanisms have become an integral part of models designed to capture global dependencies [27–30]. In particular, self-attention [31,32], also known as intra-attention, computes the response at a specific position in a sequence by attending to all positions within the same sequence. Vaswani et al. [33] demonstrated that state-of-the-art results can be achieved in machine translation models using self-attention alone. Parmar et al. [34] introduced an image Transformer model that incorporates self-attention into an autoregressive framework for image generation. Wang et al. [35] formalized self-attention as a non-local operation to model spatio-temporal dependencies in video sequences. Despite these advancements, self-attention has yet to be explored in the context of DCGANs. Although an AttnGAN [36] employs attention for word embeddings in the input sequence, it does not use self-attention for internal model states. Thus, our proposed approach aims at efficiently discovering global, long-range dependencies within the internal representations of images.

3. Materials and Methods

The main structure of the proposed SAITI-DCGAN approach can be summarized as follows. A self-attention module is integrated into the generator, and spectral normalization is adopted for both the generator and discriminator. In addition, the two time-scale update rule is applied to both the generator and discriminator, utilizing distinct learning rates. The general framework of the proposed approach is depicted in Figure 4.

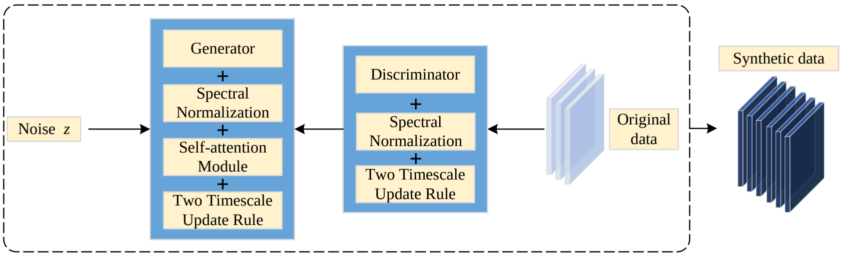


Figure 4. The general framework of our proposed SAITI-DCGAN approach.

3.1. Self-Attention DCGAN

The traditional DCGAN generates high-quality and intricate features by utilizing fixed spatial local information within an image, exhibiting remarkable performance for texture features. Nevertheless, capturing specific geometric features remains challenging [37]. To address this issue, a self-attention mechanism is introduced into the DCGAN framework in this work. During the image generation process, the generator coordinates fine details from each position with distant parts of the image, allowing it to disregard irrelevant information and enhance the significance of key feature details. This combination of the attention mechanism is applied to the middle layer of the generator, enabling it to effectively extract local image features and enhance the generation diversity. The network structure of the self-attention module is illustrated in Figure 5.

The image features from the previous hidden layer $x \in \mathbb{R}^{C \times N}$ are first transformed into two feature spaces f and g to calculate the attention, where $f(x) = W_f x$, and $g(x) = W_g x$. Here, C indicates the number of channels, and N represents the number of feature locations. The transpose of $f(x_i)$ is multiplied by $g(x_j)$ to obtain the correlation S_{ij}. Then, S_{ij} are normalized using Softmax to generate the attention feature, which is calculated as follows:

$$\beta_{j,i} = \frac{\exp\left(S_{ij}\right)}{\sum_{i=1}^{n} \exp\left(S_{ij}\right)}, \text{where } S_{ij} = f(x_i)^T g(x_j). \tag{2}$$

Here, $\beta_{j,i}$ represent the extent to which the model focuses its attention on the i-th position when synthesizing the j-th region. It is important to note that the layer's output is denoted as $o = (o_1, o_2, \ldots, o_j, \ldots, o_N)$,

$$o_j = v\left(\sum_{i=1}^{N} \beta_{j,i} h(x_i)\right), \text{where } h(x_i) = W_h x_i, v(x_i) = W_v x_i. \tag{3}$$

Here, the weight matrices $W_g \in \mathbb{R}^{\overline{C}\times C}$, $W_f \in \mathbb{R}^{\overline{C}\times C}$, $W_h \in \mathbb{R}^{\overline{C}\times C}$, and $W_v \in \mathbb{R}^{\overline{C}\times C}$ implement 1×1 convolutions. After multiple training iterations on ImageNet, no significant performance degradation was observed when reducing the number of channels in $\overline{C}$ to C/k, where $(k = 1, 2, 4, 8)$. To enhance the memory efficiency, we chose $k = 8$ (i.e., $\overline{C} = C/8$) for all experiments.

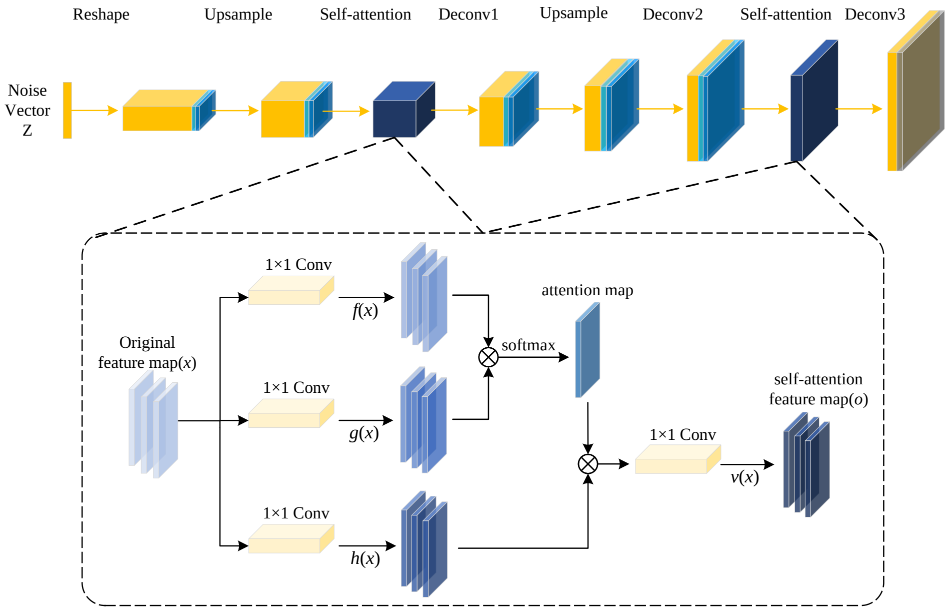


Figure 5. The network structure of the self-attention module with DCGAN.

Additionally, the output of the attention layer is scaled by a factor and then added to the input feature map. As a result, the final output is obtained as follows:

$$y_i = \gamma o_i + x_i. \tag{4}$$

Here, γ indicates a scalar value that can be learned and starts with an initial value of 0. The purpose of introducing the learnable γ is to prompt the network to initially depend on information from local neighborhoods, which is simpler, and then gradually acquire knowledge from non-local sources to assign greater importance. This strategy is adopted to facilitate the sequential learning of simpler tasks before gradually transitioning to complex tasks. In the proposed approach, an attention module is incorporated into both the generators and discriminators, which are trained alternately using a minimized form of the adversarial loss function [38–40].

$$\begin{aligned} L_D = &- \mathbb{E}_{(x,y)\sim p_{data}}[\min(0, -1 + D(x,y))] \\ &- \mathbb{E}_{z\sim p_z, y\sim p_{data}}[\min(0, -1 - D(G(z), y))], \\ L_G = &-\mathbb{E}_{z\sim p_z, y\sim p_{data}} D(G(z), y), \end{aligned} \tag{5}$$

where p_{data} represents the sample distribution, and $\mathbb{E}$ denotes the expected distance. $D(x,y)$ is the discriminator that takes (x,y) as input and outputs a scalar, and $G(z)$ is the generator in which a sample z can be drawn from a distribution p_z to the input space.

Compared to existing GAN models that have introduced self-attention mechanisms [16,37], our approach has several distinct features: it is specifically designed for industrial production, focusing on enhancing infrared thermal images for industrial defect detection. We integrate the self-attention module into the middle layer of the generator to capture long-range dependencies and enhance the feature extraction. Our model uses a learnable scaling factor γ to gradually transition from local to global information, and we reduce the number of channels in the self-attention module to C/8 to improve the memory efficiency, without significant performance degradation.

The generator is composed of a 7-layer convolutional network, as shown in Table 1. It incorporates two self-attention modules, three deconvolution layers, and two upsampling layers. The 262,144-dimensional feature vector is transformed into 256 × 256 × 3 images as the input. Furthermore, the deconvolution layer adopts the LeakyReLU activation function and batch normalization, while the output layer employs the hyperbolic tangent activation function (tanh).

Table 1. Detailed information of the proposed generator.

Input Image Size	Convolution Operation	Output Image Size
1 × 262,144	Reshape (BN)	64 × 64 × 64
64 × 64 × 64	Upsample:2	64 × 128 × 128
64 × 128 × 128	Self-attention	64 × 128 × 128
64 × 128 × 128	Deconv2d:64 3 × 3 stride = 1 (BN ReLU)	64 × 128 × 128
64 × 128 × 128	Upsample:2	64 × 256 × 256
64 × 256 × 256	Deconv2d:32 3 × 3 stride = 1 (BN ReLU)	32 × 256 × 256
32 × 256 × 256	Self-attention	32 × 256 × 256
32 × 256 × 256	Deconv2d:3 3×3 stride = 1 (Tanh)	3 × 256 × 256

3.2. Spectral Normalization for Generator and Discriminator

In stable GAN training methods, spectral normalization has recently been implemented in the discriminator, as introduced in the study in [41]. This technique constrains the Lipschitz constant of the discriminator by bounding the spectral norm of each layer. The spectral norm serves as an approximation of the Lipschitz constant for a composite function, which can be thought of as a sequence of convolutional layers and activation functions. According to the Lipschitz continuity theory, a composite function retains Lipschitz continuity if each individual function within it maintains this property. In the discriminator, the activation function is LeakyReLU, which inherently satisfies Lipschitz continuity and contributes to the stable training of the GAN network. Notably, unlike other normalization methods, spectral normalization does not necessitate additional hyperparameter tuning (in our experiments, setting the spectral norm of all weight layers to 1 yielded satisfactory results). Consequently, every convolutional layer must be designed to satisfy Lipschitz continuity. When considering an input h and a parameter matrix A for a convolutional layer, the spectral norm $\sigma(A)$ is calculated as follows:

$$\sigma(A) := \max_{h:h\neq 0} \frac{||Ah||_2}{||h||_2} = \max_{||h||_2\leq 1} ||Ah||_2. \tag{6}$$

This value is equivalent to the largest singular value of matrix A, and it corresponds to the Lipschitz continuity constant, which is also equal to the spectral norm of the parameter matrix. By determining the spectral norm $\sigma(W)$ of the parameter matrix W for the convolutional layer, we ensure that the maximum singular value of the normalized convolutional layer parameter matrix W_{SN} remains at 1. Consequently, for a linear layer, the norm can be expressed as $g(h) = Wh$, and its Lipschitz constant is calculated as $||g||_{\text{Lip}} = sup_h \sigma(\nabla g(h)) = sup_h \sigma(W) = \sigma(W)$. The computation of W_{SN} is outlined as follows:

$$W_{\text{SN}}(W) = \frac{W}{\sigma(W)}. \tag{7}$$

Recent research has emphasized the significance of modulating generators for improving the performance of GANs. In particular, studies have shown that applying spectrum normalization to generators can yield several benefits [42]. By employing spectrum normalization, the risk of parameter escalation is minimized and potential gradient irregularities are avoided. Empirical evidence suggests that this technique encourages the generator to rely on a smaller number of discriminators per update, resulting in a significant reduction in computational training costs. Additionally, spectrum normalization has been observed to enhance training stability.

To provide additional information, the DCGAN incorporates spectrum normalization in its discriminator's convolutional network, as shown in Table 2. The input consists of images with dimensions of 256 × 256 × 3. The convolutional kernel size is set to 3 × 3, and the channel lengths progress as 16, 32, 64, 128, and 128. To improve training stability, all layers are augmented with batch normalization (BN) and spectral normalization (SN). The LeakyReLU activation function from ref. [43] with a slope of 0.2 is utilized. Additionally, 25% of the input data channels are randomly dropped, aiding the model in achieving improved generalization over the training data. Ultimately, linear and sigmoid functions are employed for classification purposes.

Table 2. Detailed information of the proposed discriminator.

Input Image Size	Convolution Operation	Output Image Size
3 × 256 × 256	Conv2d:16 3 × 3 stride = 2 (BN SN 0.2 leakyReLU Dropout)	16 × 256 × 256
16×256 × 256	Conv2d:32 3 × 3 stride = 2 (BN SN 0.2 leakyReLU Dropout)	32 × 256 × 256
32 × 256 × 256	Conv2d:64 3 × 3 stride = 2 (BN SN 0.2 leakyReLU Dropout)	64 × 256 × 256
64 × 256 × 256	Conv2d:128 3 × 3 stride = 2 (BN SN 0.2 leakyReLU Dropout)	128 × 256 × 256
128 × 256 × 256	Flatten (Liner Sigmoid)	1 × 262,144

3.3. Two Time-Scale Update Rule

In previous research, it has been noted that the orthogonalization of the discriminator [44] can hinder the learning progress of GANs. Specifically, the process of training the orthogonalized discriminator often requires multiple updates of the discriminator for every update of the generator. To tackle this challenge, Heusel et al. [45] proposed a solution known as the two time-scale update rule (TTUR), which involves assigning distinct learning rates to the generator and discriminator. The TTUR approach aims to mitigate the issue of slow learning when training the orthogonalized discriminator.

The TTUR method suggests employing fewer discriminator update steps for each generator update step whenever possible. By adopting this technique, it becomes feasible to achieve improved outcomes within the same time frame. In the context of the two time-scale update rule, the learning rates $a(n)$ and $b(n)$ are utilized for generator and discriminator updates, as follows:

$$\begin{aligned} w_{n+1} &= w_n + b(n)(g(\theta_n, w_n) + M_n^{(w)}) \\ \theta_{n+1} &= \theta_n + a(n)(h(\theta_n, w_n) + M_n^{(\theta)}), \end{aligned} \tag{8}$$

where w and θ represent the parameter vectors of the discriminator $D(., w)$ and the generator $G(., \theta)$, respectively. In addition, $g(\theta,w)$ and $h(\theta, w)$ indicate the true gradient of the discriminator and generator, and $M^{(w)}$ and $M^{(\theta)}$ are random variables.

The flowchart presented in Figure 6 illustrates the proposed approach, which progresses through four distinct steps. Firstly, the integration of the self-attention module occurs in both the low-dimensional and high-dimensional convolutional layers of the generator. This integration plays a crucial role in generating the synthetic data. Secondly, spectral normalization is employed to train both the generator and the discriminator. The discriminator evaluates the authenticity of both the synthetic and real data. Thirdly, different learning rates are assigned to the generator and discriminator. Lastly, a cyclic training process is executed for the discriminator and generator, continuously improving the quality of the synthesized images until a state of Nash equilibrium has been achieved. The losses associated with the generator and discriminator guide the parameter updates in their respective networks.

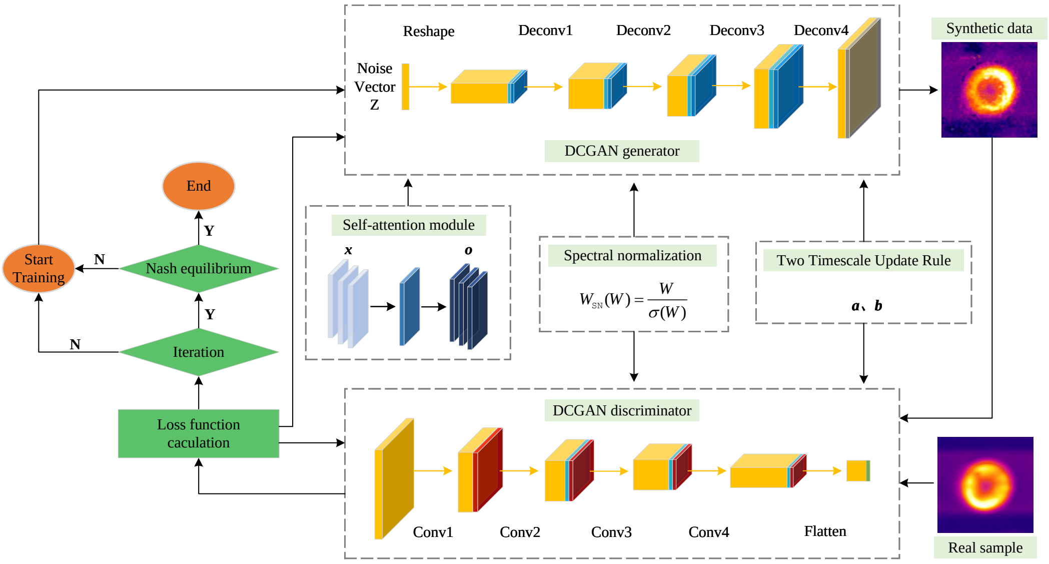


Figure 6. Flowchart of the proposed SAITI-DIGAN approach.

4. Results

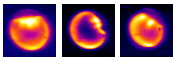
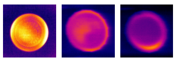
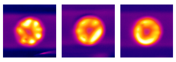
To verify the effectiveness of our SAITI-DCGAN approach, a set of qualitative and quantitative comparison experiments were conducted. The total number of epochs and the learning rates of the generator and discriminator were set to 5000, 0.0002, and 0.0001, respectively. The hardware configuration for the experiments included an AMD (R) Ryzen (TM) 5 5600H @ 3.30 GHz CPU, manufactured by Advanced Micro Devices, Inc., Austin, TX, USA; an NVIDIA GeForce RTX 3050 GPU, and 16 GB of memory, both manufactured by NVIDIA Corporation, Santa Clara, CA, USA. The experiments were conducted within the Python 3.6 operating environment, utilizing the PyTorch platform (version 1.13.1+cu117) for model implementation. The operating system employed was Windows 11.

4.1. Data Collection

The dataset of aluminum foil sealing was collected on-site through infrared thermal imaging. In the absence of designated domain experts, we engaged several senior engineers from within the company, who possess substantial practical experience in pertinent fields, to contribute to the collection process. Guided by these experts, we performed an exhaustive classification of on-site acquired samples and meticulously selected representative instances of both normal and anomalous cases to compile our training and testing datasets. This methodological approach enriched the diversity and complexity of the dataset, thereby ensuring its fidelity to real-world scenarios.

A total of three categories of aluminum foil sealing defect images were collected, with the resolution spanning from 100 × 100 pixels to 384 × 288 pixels. Infrared thermal images were captured within a specific spectral range of 7.5 to 14 μm. To enhance the visual contrast and improve feature extraction, we generated 24-bit false color images by mapping different infrared bands to colors in the visible spectrum. Specifically, the red channel captured high-temperature features, the green channel captured mid-high temperature features, and the blue channel captured low-temperature features. This enhanced contrast helped the network more effectively distinguish different features. The RGB values were mapped within the range of 0 to 255 to ensure compatibility with standard image processing techniques. To prepare the data for model training, we resized the height and width of the defect images to 256 × 256 pixels, preserving the aspect ratio to maintain the fundamental morphological characteristics of the defects. During the preprocessing phase of the dataset, random rotation, flipping, and Gaussian blurring techniques were employed to enhance the quality of the generated images. Representative samples with various defects are presented in Table 3. To be specific, the types are divided into the following categories: broken aluminum foil (BAF), missing aluminum foil (MAF), and loose or crooked caps (LCC).

Table 3. Collected dataset images of aluminum foil sealing.

Type	Description	Example
BAF	broken aluminum foil	
MAF	missing aluminum foil	
LCC	loose or crooked caps	

4.2. *Qualitative Evaluation*

Images generated by the GAN, DCGAN, and our SAITI-DCGAN approach are illustrated in Figure 7. In the left-most column, various types of infrared thermal images of aluminum foil sealing are displayed. Specifically, it includes images of broken aluminum foil within the bottle cap (BAF), images that depict the absence of aluminum foil in the bottle cap (MAF), and images showcasing instances where the bottle cap was either inadequately tightened or skewed during the tightening process (LCC). The final row presents the generation outcomes obtained from the GAN, DCGAN, and our proposed SAITI-DCGAN approach. Given the large number of test samples, for each type of defect generation, the input used for comparison in Figure 7 was randomly selected from the normal samples in the test set. All samples were rigorously screened and classified under the guidance of domain experts, ensuring that they not only accurately reflected both normal and abnormal conditions in real-world application scenarios but also maintained internal consistency within each category of images. Consequently, even randomly selected samples, particularly the normal samples, possessed characteristics that were representative of their respective categories. Furthermore, to ensure the fairness and directness of the model output comparisons, all tests were conducted at the same iteration stage on identical inputs. This meant that the different network architectures faced exactly the same input conditions at the same training stage, thereby making the comparison of output results more equitable and comparable. The results shown are the outputs of the networks at different iteration stages for the same input. Further analysis of these results will be provided based on the subsequent experimental findings.

We can observe that in the GAN model, during the initial 100 training iterations for the three defect cases (i.e., BAF, MAF, and LCC), the generated images exhibited significant blurriness and noise, lacking distinct forms for the defects. After 1000 iterations, the generated images still retained some blurriness, but they showcased enhanced content and sharpness compared to the 100 iteration stage. By the time the training reached 1500 iterations, the generated images began to exhibit the rudimentary form of the defects. With further training iterations, at around 2000 iterations, the generated images became clearer, containing richer detail information than that at 1500 iterations. The images produced after 2000 iterations were close to the actual defect images. With the original DCGAN model, the internal defect structure of the generated images from 100 to 500 training iterations remained chaotic. Stability in shape was gradually achieved around 750 training iterations, occasionally accompanied by aberrant images. This pattern was particularly evident in the BAF and LCC cases, while well-defined MAF defect images emerged as the training progressed to 1750 iterations. In comparison, the generated images of our SAITI-DCGAN yielded a stable and clear structure after approximately 750 training iterations with respect to all three categories.

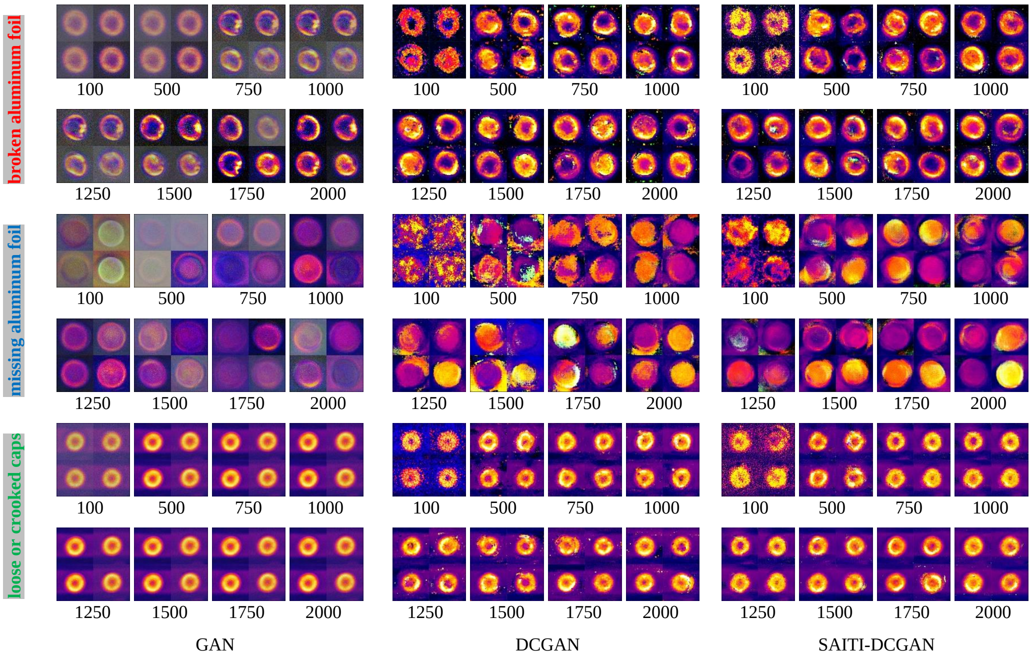


Figure 7. Comparison of the existing image generation models and the model proposed in this paper (the numbers below the images correspond to the model's iterations).

In evaluating the generated anomaly samples, we collaborated with industry experts to conduct a detailed comparative analysis. Leveraging their extensive experience in defect detection and image processing, the experts provided critical guidance and feedback throughout the process. Their input was essential in ensuring the scientific rigor and reliability of our findings.

Firstly, the GAN model was easily influenced by noise during the training process. Moreover, the images generated by the GAN for the three types of defects exhibited a monotonous style, lacking diversity and randomness. Notably, the generated LCC defect images differed greatly from the original training set, indicating that the GAN model failed to effectively capture the image features. With respect to the DCGAN model, the generated images displayed considerable instability and were prone to collapsing. Even after 2000 training iterations, the generated defect images remained less effective compared to the original images. In contrast, our proposed SAITI-DCGAN model could surpass the above two models by learning local features and also by presenting superior texture details.

4.3. Quantitative Evaluation

4.3.1. Evaluation Metrics

Considering the potential impact of subjective factors such as individual variances and preferences, the qualitative evaluation of experimental outcomes may inherently exhibit a degree of bias. To attain a more objective and accurate assessment of the generated image quality, this study incorporated two evaluation metrics: inception score (IS) [46] and Frechet inception distance (FID) [46].

The IS represents a metric employed to assess the diversity and quality of generated images. This metric involves categorizing the images generated by a classification model, typically employing the Inception V3 model. The IS value is derived by computing the entropy of the softmax output distribution for the generated images, combined with the exponent of the average class probability. The higher the value of IS, the better the diversity and quality of the generated images. The IS can be calculated by

$$\mathrm{IS}(G) = exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x)||p(y))), \tag{9}$$

where $x \sim p_g$ denotes the generated image, and $p(y|x)$ indicates that the generated image x is input into the initial model to obtain a 1000-dimensional vector y, i.e., the probability distribution of the image represents various types. Moreover, $p(y)$ means the average corresponding probability distribution vector obtained by N generated images input into the initial model, i.e., the marginal distribution of the image generated by the generator in all categories.

The FID is utilized to quantify the dissimilarity between generated and real images. This assessment involves passing both generic and authentic data through a pre-trained Inception V3 model [47] originally trained on the ImageNet dataset [48] to extract visually relevant features. The mean and covariance of the authentic and generated features are represented by (M_t, C_t) and (M_g, C_g), respectively, and T_r denotes the trace of the covariance matrix of the feature vectors of the real data. This metric can be calculated by

$$\mathrm{FID} = ||M_t - M_g||_2^2 + T_r(C_t + C_g - 2(C_t C_g)^{\frac{1}{2}}). \tag{10}$$

A lower FID value indicates that the generated images are closer to real images in terms of visual quality, diversity, and distribution of features.

4.3.2. Evaluation Results

Here, we employed the IS, FID, and loss function as quantitative evaluation metrics. The generator's loss function typically hinges on the discriminator's predictions concerning the generated samples. Specifically, the generator aims to minimize the probability of the discriminator classifying the generated sample as fake. Conversely, the discriminator's loss function typically relies on its classification outcomes for both real and generated samples. Figure 8 illustrates the fluctuation patterns of the generator's and discriminator's loss functions across different model variants. The graph indicates that the discriminator's loss function experienced notable fluctuations during the early stages. On the other hand, the generator's loss function exhibited significant fluctuations in the beginning due to its limited capacity to create authentic defect images. This made it susceptible to being identified as a fake sample by the discriminator. As the training progressed, the generator's loss function gradually increased, signifying an improved ability to generate samples resembling real defect images. In comparison with our proposed approach, we can see that the error between the maximum value of the loss function of the generator and the discriminator was one to three times in the GAN and DCGAN models. The loss of the generator and discriminator of our approach was significantly lower than for the other two models, and it also demonstrated robust convergence and performance.

Figure 9 depicts the performance for the FID and IS values across different models and defect cases. In all three defect cases, the FID value consistently decreased as the optimization and training of the model progressed. This trend signifies that the quality of the generated images steadily approached that of real images, reflecting an overall enhancement in the performance of the generated model. Conversely, the IS value exhibited a more intricate pattern of change. The IS value tended to increase over time, indicating an enhancement in the quality and diversity of the generated images. However, changes in the IS can be complex due to its potential sensitivity to specific categories of generated images during diversity evaluation. Consequently, the improvement in the quality of certain categories of generated images may be very slow. We can see that the GAN model showed a sharp downward trend in terms of the IS value for the LCC defect, similar to the trend observed for the BAF defect with the DCGAN model. This indicates that the training process of these models was not stable. In contrast, when employing imbalanced learning rates for training both the generator and discriminator within our proposed model, the generated images consistently exhibited a monotonic change with respect to the FID and the IS values throughout the training process.

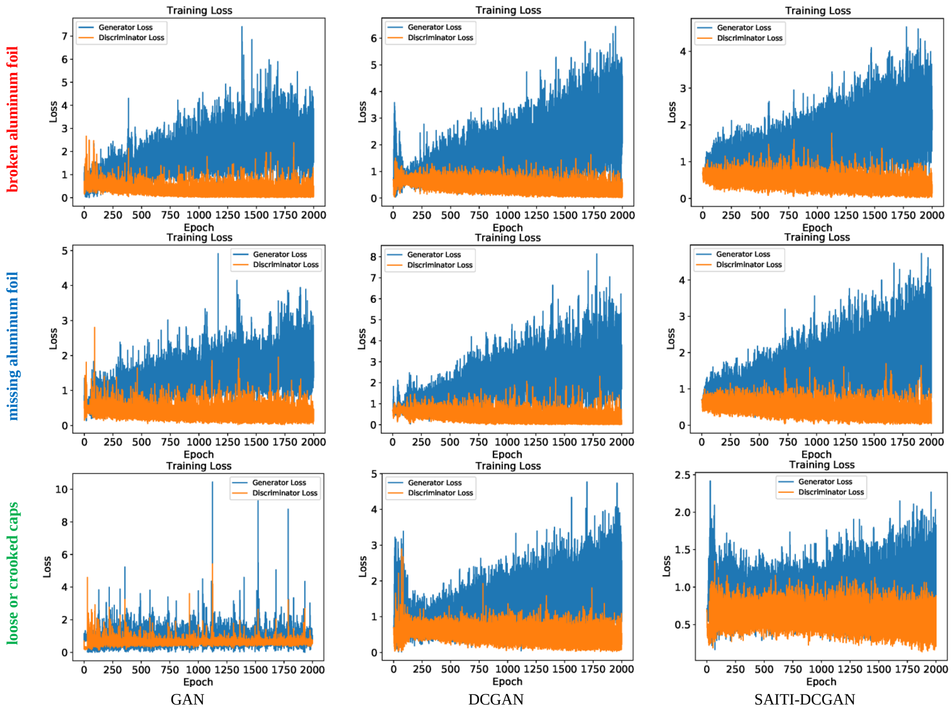


Figure 8. Comparison of the loss function of the generator and discriminator under different models.

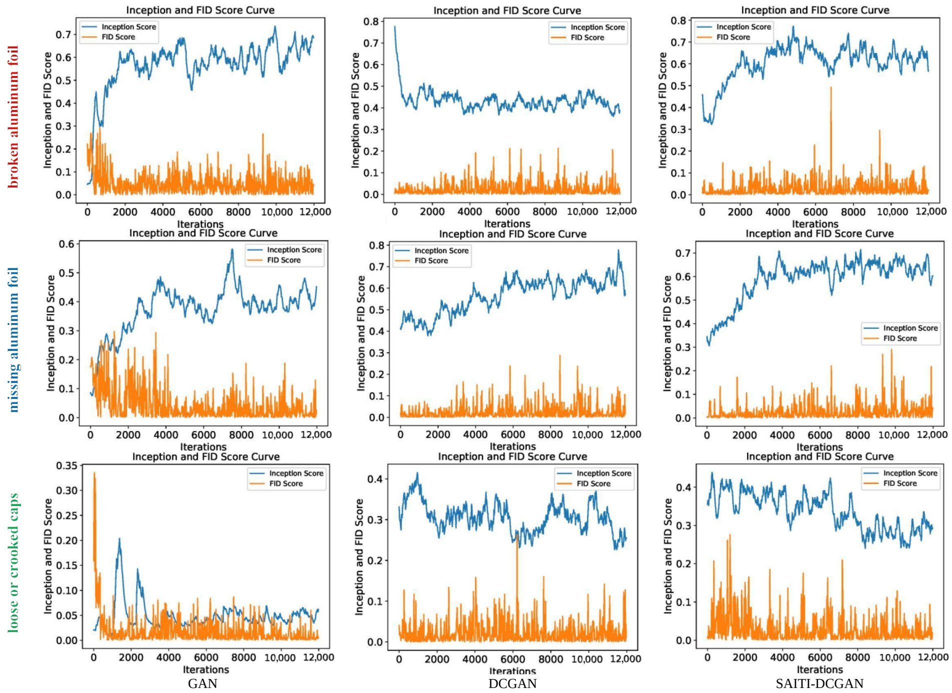


Figure 9. Comparison of the IS and FID metrics under different models.

We also detail the FID and IS values in Table 4. We can see that the image generation model introduced in this paper achieved better results.

For the IS value, our model outperformed both the DCGAN and GAN models for all defect types, showing substantial improvements. For the BAF defect, our model significantly outperformed both comparison models. For MAF and LCC defects, our model also demonstrated notable improvements over the DCGAN and GAN models.

Regarding the FID values, our model achieved impressively low scores for all defect types, significantly lower than those of the DCGAN and GAN models. These results indicate that the proposed model consistently produced superior images compared to the existing models.

Table 4. Results of the different models for the three defects.

Type	Description	FID (Avg)			Inception Score (Avg)		
		GAN	DCGAN	SAITI-DCGAN	GAN	DCGAN	SAITI-DCGAN
BAF	broken aluminum foil	0.050169	0.028795	**0.025064**	0.567541	0.433728	**0.606708**
MAF	missing aluminum foil	0.044233	0.02598	**0.025588**	0.373805	0.56516	**0.590326**
LCC	loose and crooked caps	0.22409	0.028959	**0.019037**	0.048396	0.306051	**0.336719**

4.4. Ablation Experiments

In this paper, we also conducted ablation experiments to validate the effectiveness of the individual modules. The use of each module is indicated by ✓, while its absence is denoted by ×. All three modules generated images of size 256 × 256 × 3. Table 3 presents a subset of the original defect images, while Table 5 displays the synthesized images of different defects (i.e., BAF, MAF, and LCC) under different modules.

The first three rows show images generated with one module removed, while the last three rows display images generated with two modules removed. Overall, the image quality remained acceptable, but images generated using the SA module exhibited a higher coherence in layout and better alignment with the original images, enhancing their similarity.

According to Table 5, the image quality was slightly lower when only the SN and TTUR modules were used compared to when only the SA module was included. This is expected, as the SN module stabilizes the training process but does not improve the image details. Table 6 presents a comparative analysis of the training losses of the generator and discriminator under the different module configurations, and discriminator loss comparisons under the different modules. In Table 6, the loss curves of the generator and discriminator are represented by the blue and orange lines, respectively. Models with the SN module exhibited less loss fluctuation, indicating enhanced training stability. The TTUR module, by using a smaller learning rate for the discriminator and a larger one for the generator, improves the data distribution representation and adversarial learning stability, leading to better generator quality and reduced loss. The generator's loss value varied the most, suggesting that the model effectively leveraged the capabilities of the SA, SN, and TTUR modules to enhance the image quality.

When one module was removed, as shown in the first three rows, removing the SN module had a minor impact on the image quality and training stability. However, when two modules were removed, as demonstrated in the last three rows, the decline in image quality and training stability was more pronounced. Specifically, removing both the SA and SN modules led to a significant deterioration in visual quality and training instability, characterized by large fluctuations in training loss (Table 6). Removing the SA and TTUR modules resulted in a substantial drop in image quality due to the lack of key image feature capture and optimized learning rate adjustment (Table 5).

Table 5. Synthesized images of different defects under different modules.

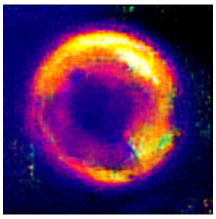
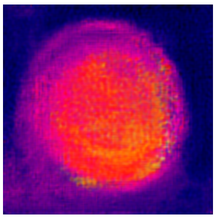
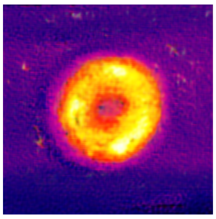
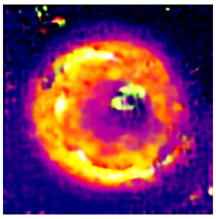
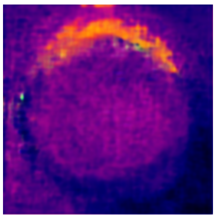
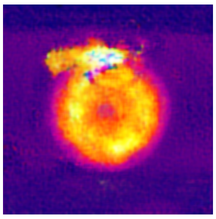
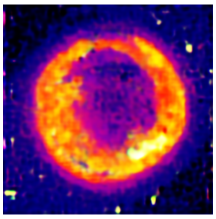
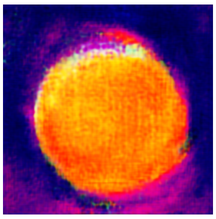
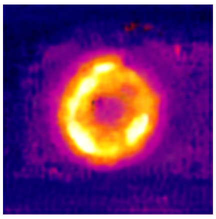
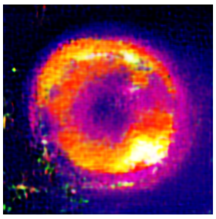
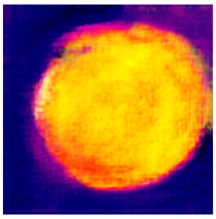
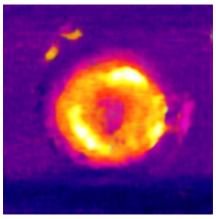
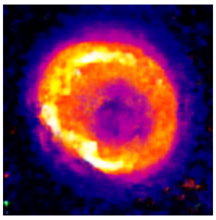
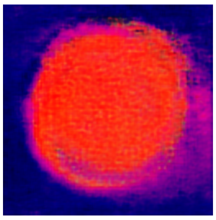
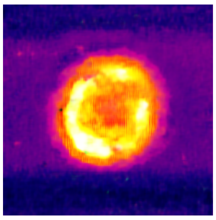
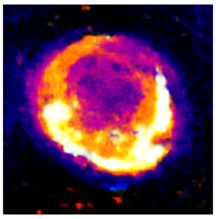
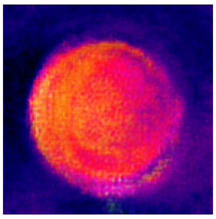
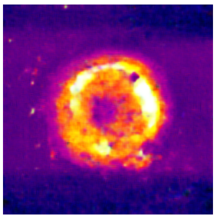
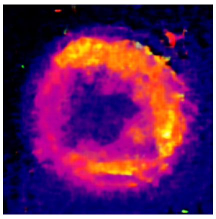
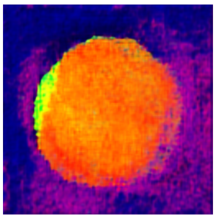
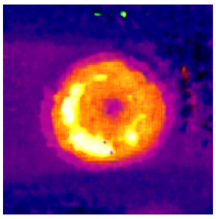
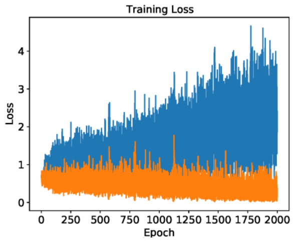
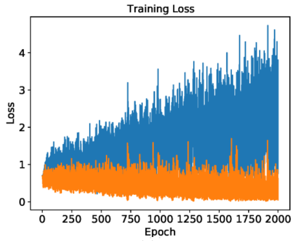
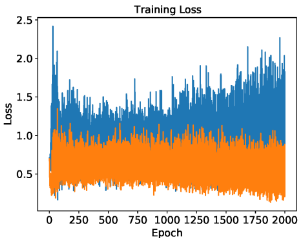
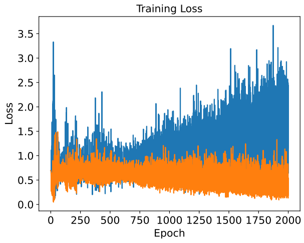
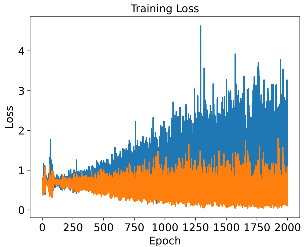
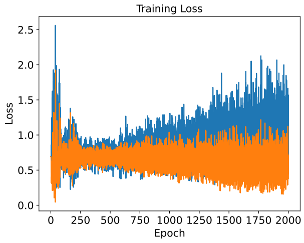
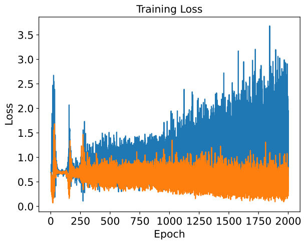
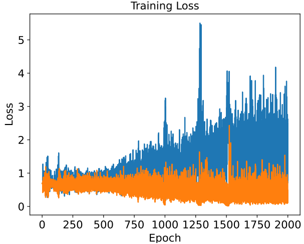
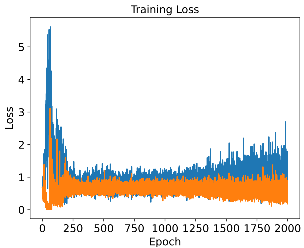
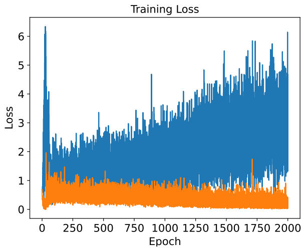
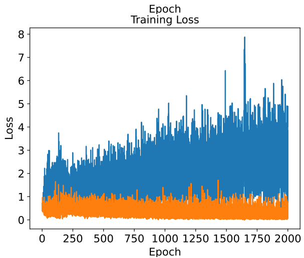
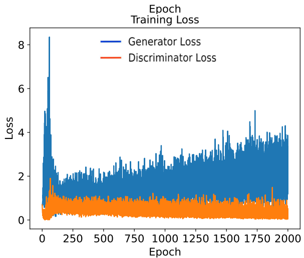
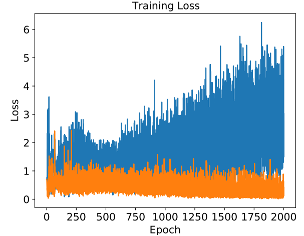
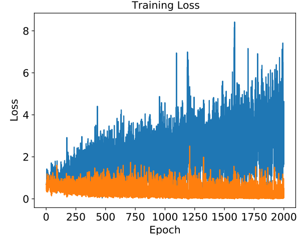
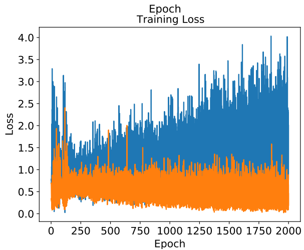
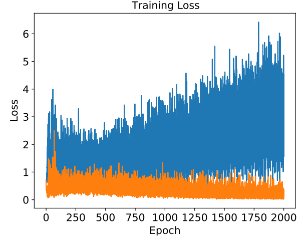
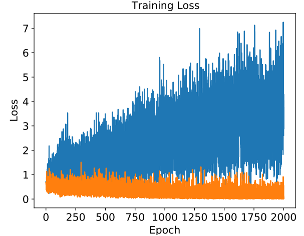
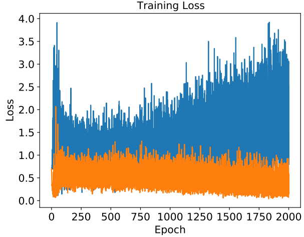
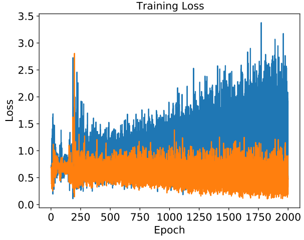
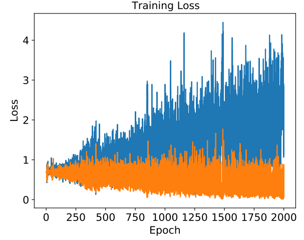
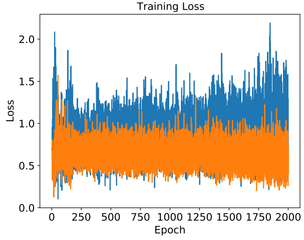
Module			BAF	MAF	LCC
SA	SN	TTUR			
✓	✓	✓			
×	✓	✓			
✓	×	✓			
✓	✓	×			
✓	×	×			
×	✓	×			
×	×	✓			

Table 6. Training loss of the generator and discriminator under different modules.

Module			BAF	MAF	LCC
SA	SN	TTUR			
✓	✓	✓			
×	✓	✓			
✓	×	✓			
✓	✓	×			
✓	×	×			
×	✓	×			
×	×	✓			

5. Discussion

In this paper, we evaluated the SAITI-DCGAN (spectral and spatial attention improved thermal image deep convolutional generative adversarial network) using self-collected infrared thermal images. Our assessment included both qualitative analysis and quantitative metrics such as the inception score (IS) and Frechet inception distance (FID). The results indicate that our model generated more realistic and higher-quality defect images compared to the traditional GAN and DCGAN models, because of the spectral–spatial attention mechanism that enhanced the feature extraction and image fidelity.

While the trained discriminator can extract useful features for defect detection via transfer learning, its primary role in distinguishing real from generated images limits its fine-grained recognition capabilities. For optimal defect detection, we recommend using the generated images for data augmentation and employing specialized detection models for precise identification.

Future work will focus on refining the attention mechanisms to improve the image realism, integrating multi-source data for richer training sets, and developing customized algorithms for specific defect types. This approach aims to enhance the application of infrared thermal imaging in non-destructive testing.

6. Threats to Validity

The dataset used in this study consists of self-collected infrared thermal images. The images provide valuable controlled data but may not fully represent the diverse variations found in industrial settings. This limitation could affect the model's adaptability to real-world conditions, impacting the generalizability of our findings.

For quantitative evaluation, we used the Frechet inception distance (FID) metric based on a pre-trained Inception V3 model. While FID is widely accepted, its reliance on a network trained on natural RGB images may introduce biases when applied to infrared thermography, potentially leading to inaccurate assessments of image quality due to mismatches with unique characteristics like temperature gradients and material emissivity.

To ensure practical applicability and robustness, additional validation methods are necessary. Human expert evaluations can provide critical insights into the realism and diagnostic value of generated images. Real-world testing in industrial environments helps identify limitations and areas for improvement. Integrating user feedback and existing inspection workflows further refines a model's utility.

In summary, while the current dataset and FID metric offer valuable insights, they highlight the need for a more comprehensive validation approach. Future work should expand the dataset and develop metrics specific to infrared thermal imaging to better meet the demands of industrial defect detection.

7. Conclusions and Future Outlook

In this paper, we proposed a novel deep convolutional generative adversarial network that incorporates self-attention, spectral normalization, and two time-scale update rules, named SAITI-DCGAN, to enhance the synthesis quality of infrared thermal images for aluminum foil sealing. Specifically, a self-attention module was embedded into the generator, and spectral normalization was adopted for both the generator and discriminator. In addition, a two time-scale update rule was applied to coordinate the training of the generator and discriminator. With respect to the functionalities of these modules, we integrated the self-attention into the generator to capture the global dependencies of the images, thereby enhancing the quality and variety of the generated images. Subsequently, we elaborated on the roles and advantages of spectral normalization and how to apply it to generators and discriminators for stabilizing the training process, with the aim of preventing gradient anomalies and pattern crashes. The two time-scale update rule was adopted to expedite the generator's adaptation to the discriminator feedback. This mechanism can improve the balance between the generator and discriminator and enhance convergence.

In the experiments, we utilized self-collected infrared thermal images of aluminum foil sealing as a dataset for comprehensive verification. We considered both qualitative and quantitative metrics, including the inception score and the Frechet inception distance, to demonstrate the superiority of our proposed SAITI-DCGAN model in terms of image quality and diversity. Compared to traditional GAN and DCGAN models, the images generated by our SAITI-DCGAN were more realistic and of higher quality, as it could capture intricate details and distinctive features of the defect images of aluminum foil sealing. In addition, we also conducted ablation experiments to verify the roles of the self-attention, spectral normalization, and two time-scale update rules in improving the model performance. The results indicate that the combination of these modules significantly enhanced the image generation quality and achieved a balance between stability and training efficiency.

In conclusion, the proposed SAITI-DCGAN model achieved remarkable results in the quality and diversity of synthesized infrared thermal images for aluminum foil sealing. Our work also provides a strong impetus for the application and innovation of infrared thermal imaging technology. In the future, we will continue to explore and optimize the model, so as to meet the requirements of various practical application scenarios in the industrial field.

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

Funding: This research was funded by the National Natural Science Foundation of China grant number 52371275.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Fang, H.; Xia, M.; Liu, H.; Chang, Y.; Wang, L.; Liu, X. Automatic zipper tape defect detection using two-stage multi-scale convolutional networks. *Neurocomputing* **2021**, *422*, 34–50. [CrossRef]
2. Zeng, N.; Li, H.; Peng, Y. A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease. *Neural Comput. Appl.* **2023**, *35*, 11599–11610. [CrossRef]
3. Zeng, N.; Li, H.; Wang, Z.; Liu, W.; Liu, S.; Alsaadi, F.E.; Liu, X. Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip. *Neurocomputing* **2021**, *425*, 173–180. [CrossRef]
4. Bai, D.; Li, G.; Jiang, D.; Yun, J.; Tao, B.; Jiang, G.; Sun, Y.; Ju, Z. Surface defect detection methods for industrial products with imbalanced samples: A review of progress in the 2020s. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107697. [CrossRef]
5. Cho, E.; Jeon, B.; Park, I.K. Synthesizing Industrial Defect Images Under Data Imbalance. *IEEE Access* **2023**, *11*, 111335–111346. [CrossRef]
6. Wan, W.; Chen, J.; Zhou, Z.; Shi, Z. Self-Supervised Simple Siamese Framework for Fault Diagnosis of Rotating Machinery with Unlabeled Samples. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 6380–6392. [CrossRef]
7. Cao, H.; Shao, H.; Zhong, X.; Deng, Q.; Yang, X.; Xuan, J. Unsupervised domain-share CNN for machine fault transfer diagnosis from steady speeds to time-varying speeds. *J. Manuf. Syst.* **2022**, *62*, 186–198. [CrossRef]
8. Zhang, T.; Jiao, J.; Lin, J.; Li, H.; Hua, J.; He, D. Uncertainty-based contrastive prototype-matching network towards cross-domain fault diagnosis with small data. *Knowl.-Based Syst.* **2022**, *254*, 109651. [CrossRef]
9. Hu, Y.; Liu, R.; Li, X.; Chen, D.; Hu, Q. Task-sequencing meta learning for intelligent few-shot fault diagnosis with limited data. *IEEE Trans. Ind. Inform.* **2021**, *18*, 3894–3904. [CrossRef]
10. Pan, T.; Chen, J.; Zhang, T.; Liu, S.; He, S.; Lv, H. Generative adversarial network in mechanical fault diagnosis under small sample: A systematic review on applications and future perspectives. *ISA Trans.* **2021**, *128*, 1–10. [CrossRef]

11. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
12. Gao, X.; Deng, F.; Yue, X. Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. *Neurocomputing* **2020**, *396*, 487–494. [CrossRef]
13. Chen, Y.; Zhang, L.; Xue, X.; Lu, X.; Li, H.; Wang, Q. PadGAN: An End-to-End dMRI Data Augmentation Method for Macaque Brain. *Appl. Sci.* **2024**, *14*, 3229. [CrossRef]
14. Shang, Z.; Li, R. Enhanced Solar Coronal Imaging: A GAN Approach with Fused Attention and Perceptual Quality Enhancement. *Appl. Sci.* **2024**, *14*, 4054. [CrossRef]
15. Wang, X.; Ao, Z.; Li, R.; Fu, Y.; Xue, Y.; Ge, Y. Super-Resolution Image Reconstruction Method between Sentinel-2 and Gaofen-2 Based on Cascaded Generative Adversarial Networks. *Appl. Sci.* **2024**, *14*, 5013. [CrossRef]
16. Liu, Y.; Jiang, H.; Wang, Y.; Wu, Z.; Liu, S. A conditional variational autoencoding generative adversarial networks with self-modulation for rolling bearing fault diagnosis. *Measurement* **2022**, *192*, 110888. [CrossRef]
17. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008. [CrossRef]
18. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
19. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412. [CrossRef]
20. Berthelot, D.; Carlini, N.; Goodfellow, I.J.; Papernot, N.; Oliver, A.; Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. *arXiv* **2019**, arXiv:1905.02249. [CrossRef]
21. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
22. Kim, J.H.; Choo, W.; Song, H.O. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5275–5285.
23. Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6438–6447.
24. Mangla, P.; Kumari, N.; Sinha, A.; Singh, M.; Krishnamurthy, B.; Balasubramanian, V.N. Charting the right manifold: Manifold mixup for few-shot learning. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2218–2227.
25. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
26. Fang, W.; Zhang, F.; Sheng, V.S.; Ding, Y. A Method for Improving CNN-Based Image Recognition Using DCGAN. *Comput. Mater. Contin.* **2018**, *57*, 167.
27. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473. [CrossRef]
28. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2048–2057.
29. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
30. Chen, X.; Mishra, N.; Rohaninejad, M.; Abbeel, P. Pixelsnail: An improved autoregressive generative model. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 864–872.
31. Cheng, J.; Dong, L.; Lapata, M. Long short-term memory-networks for machine reading. *arXiv* **2016**, arXiv:1601.06733. [CrossRef]
32. Parikh, A.P.; Täckström, O.; Das, D.; Uszkoreit, J. A decomposable attention model for natural language inference. *arXiv* **2016**, arXiv:1606.01933. [CrossRef]
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
34. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
35. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7794–7803.
36. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1316–1324.
37. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363. [CrossRef]
38. Lim, J.H.; Ye, J.C. Geometric gan. *arXiv* **2017**, arXiv:1705.02894. [CrossRef]
39. Tran, D.; Ranganath, R.; Blei, D.M. Deep and hierarchical implicit models. *arXiv* **2017**, arXiv:1702.08896. [CrossRef]

40. Miyato, T.; Koyama, M. cGANs with projection discriminator. *arXiv* **2018**, arXiv:1802.05637. [CrossRef]
41. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv* **2018**, arXiv:1802.05957. [CrossRef]
42. Odena, A.; Buckman, J.; Olsson, C.; Brown, T.; Olah, C.; Raffel, C.; Goodfellow, I. Is generator conditioning causally related to GAN performance? In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 3849–3858.
43. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv* **2015**, arXiv.1505.00853. [CrossRef]
44. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
45. 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. Process. Syst.* **2017**, *30*.
46. Chong, M.J.; Forsyth, D. Effectively unbiased fid and inception score and where to find them. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6070–6079.
47. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
48. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE conference on computer vision and pattern recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

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.