

## Article

# IoT-Enabled Few-Shot Image Generation for Power Scene Defect Detection Based on Self-Attention and Global–Local Fusion

Yi Chen <sup>1</sup>, Yunfeng Yan <sup>2,3</sup>, Xianbo Wang <sup>3</sup> and Yi Zheng <sup>1,\*</sup>

<sup>1</sup> College of Electrical Engineering, Zhejiang University, Hangzhou 310058, China; morningone@126.com  
<sup>2</sup> School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China; yvonnech@zju.edu.cn  
<sup>3</sup> Hainan Institute of Zhejiang University, Sanya 572025, China; xianbowang@zju.edu.cn  
\* Correspondence: zhengyipds@zju.edu.cn

**Abstract:** Defect detection in power scenarios is a critical task that plays a significant role in ensuring the safety, reliability, and efficiency of power systems. The existing technology requires enhancement in its learning ability from large volumes of data to achieve ideal detection effect results. Power scene data involve privacy and security issues, and there is an imbalance in the number of samples across different defect categories, all of which will affect the performance of defect detection models. With the emergence of the Internet of Things (IoT), the integration of IoT with machine learning offers a new direction for defect detection in power equipment. Meanwhile, a generative adversarial network based on multi-view fusion and self-attention is proposed for few-shot image generation, named MVSA-GAN. The IoT devices capture real-time data from the power scene, which are then used to train the MVSA-GAN model, enabling it to generate realistic and diverse defect data. The designed self-attention encoder focuses on the relevant features of different parts of the image to capture the contextual information of the input image and improve the authenticity and coherence of the image. A multi-view feature fusion module is proposed to capture the complex structure and texture of the power scene through the selective fusion of global and local features, and improve the authenticity and diversity of generated images. Experiments show that the few-shot image generation method proposed in this paper can generate real and diverse defect data for power scene defects. The proposed method achieved FID and LPIPS scores of 67.87 and 0.179, surpassing SOTA methods, such as FIGR and DAWSON.

**Keywords:** few-shot image generation; power scenarios; self-attention encoder; multi-view feature fusion



**Citation:** Chen, Y.; Yan, Y.; Wang, X.; Zheng, Y. IoT-Enabled Few-Shot Image Generation for Power Scene Defect Detection Based on Self-Attention and Global–Local Fusion. *Sensors* **2023**, *23*, 6531. <https://doi.org/10.3390/s23146531>

Academic Editors: Wenbo Wang, Dinh Thai Hoang, Shaohan Feng, Shimin Gong and Yingbo Huang

Received: 22 May 2023  
Revised: 13 June 2023  
Accepted: 14 July 2023  
Published: 19 July 2023



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

## 1. Introduction

The significance of defect detection in power equipment is crucial for the safety and stable development of both a country and society at large. With the emergence of the Internet of Things (IoT) [1], the integration of IoT with technologies such as artificial intelligence (AI) and deep learning (DL) is driving the advancement of machine detection methods.

The power industry embodies a highly intricate and hazardous environment, necessitating the imposition of stringent accuracy standards for machine inspection algorithms [2]. The defect detection method [3–5], which is based on deep learning, can independently extract multi-level and multi-angle features from the original data without artificial feature extraction. It can have higher detection accuracy and stronger generalization ability through feature learning. However, in order to improve the generalization ability and accuracy of the detection model, it relies on diverse and high-quality power scene defect data for training.

IoT devices [6,7], such as sensors and smart meters, can be deployed in power systems to gather real-time information on various parameters, such as voltage, current, temperature, and other relevant factors [8–10]. These IoT devices continuously collect data from the power scene, providing valuable insights into the operation and performance of the power equipment.

The present defect dataset for electrical equipment exhibits two primary limitations. First, a comprehensive dataset encompassing diverse defect categories on a large scale is currently inadequate. The existing defect dataset [11,12] suffers from an imbalanced distribution of images across different defect categories, and the total number of samples is limited. These shortcomings significantly compromise the effectiveness of the detection model. Second, defect image acquisition is difficult. Typically, image acquisition in the electric power domain involves issues related to data protection and security, thereby constraining the diversity of the scene variability. Such limitations impede the enhancement of the generalization capacity of the model, thereby hindering its efficacy. Based on the above challenges, this paper uses a few-shot image generation method to enrich the defect data of power equipment.

Few-shot image generation methods [13–16] are mainly divided into three categories: optimization-based methods, metric-based methods, and fusion-based methods. The optimization-based method [17,18] is similar to the migration model, where images of known categories are used to train the model, and images of unknown categories are used to fine-tune the model to achieve image generation. Inspired by matching networks, metric-based methods combine the matching process with VAEs [19,20]. Fusion-based methods fuse high-level and low-level features and then decode the features back to realistic images of the same class. Electrical scenes can be complex and dynamic, with multiple objects and lighting conditions that can change rapidly. This can make it difficult to capture a wide variety of defects and produce high-quality images consistent with the entire scene. Simultaneously generating defect images in power scenarios requires accurate detection and localization of defects. Compared with the former two, the fusion-based method has improved visual quality, better diversity, faster training speed, and better generalization ability. Therefore, this paper realizes the efficient generation of power equipment defect images based on the fusion-based small-sample image generation method.

In order to better realize the feature extraction of defects in complex power scenarios, a generative adversarial network based on **multi-view** fusion and **self-attention** is proposed (MVSA-GAN). Self-attention [21] can selectively fuse features from multiple generators according to the correlation with the input image, improving the diversity and authenticity of generated images. In addition, the complexity of the electrical scene and the richness of defect types should be reflected in the generated images at the same time. The fusion of global and local features, on the one hand, allows the model to generate images that are consistent with the overall scene. On the other hand, by fusing features of different spatial scales, the model is able to capture fine and coarse details of the input and improve the visual quality of the generated images.

Our contributions can be summarized as follows:

- A self-attention encoder (SAEncoder) is designed to help improve the quality of generated images via a more robust encoding of input images. By leveraging the self-attention mechanism, the encoder can effectively capture high-level features and contextual information of the input image, which can then be utilized to guide the image generation process.
- A global–local fusion module (GLFM) is designed to enhance the quality, diversity, and realism of generated images. By capturing both global and local features of input images, the GLFM can effectively guide the image generation process, enabling the creation of high-quality images with improved visual fidelity.
- Our study addresses the challenges of defect detection in electric power scenes by proposing a novel few-shot image generation network that leverages the self-attention mechanism and multi-view fusion, named MVSA-GAN. Specifically, our proposed

network is designed to generate high-quality images with clear defect positions and diverse backgrounds, even when the available power defect scene data are limited.

The rest of the paper is arranged as follows: The current methods of generative adversarial networks (GANs) and few-shot image generation are described in Section 2. The proposed method for multi-exposure image fusion is introduced in Section 3. Additionally, the effectiveness of this method is verified through a computer simulation in Section 4. Finally, the conclusion is described in Section 5.

## 2. Related Works

### 2.1. Generative Adversarial Networks

GAN [22] is a branch of the generative model, which trains the model through adversarial learning, and uses mutual game learning of the generative model and the discriminative model to produce a fairly good output. Its training process is as follows:

$$\min_G \max_D V(D, G) = E_{[x \sim \text{real}]} \log^{D(x)} + E_{[x \sim \text{fake}]} \log^{(1-D(G(z)))} \quad (1)$$

where  $G$  represents the unknown parameter set in the generator,  $D$  represents the unknown parameter set in the discriminator,  $E$  represents the expectation, and  $V(D, G)$  represents the cross-entropy function of the entire training process. Overall, for the function  $V(D, G)$ , the parameter set of  $G$  is regarded as a symbolic constant first, and the parameter set  $D$  of the discriminator is regarded as a variable to find the maximum value of the function  $V$ . The maximum value is a symbolic constant algebraic expression containing  $G$ . Then, we find the minimum value of the function to obtain the optimal parameter value of the generator [23,24].

Due to the advantages of GANs in fitting data distributions, it is effective in image generation, image editing, and image–image translation [25,26]. The reason is mainly due to the unlimited supply of training images, which requires a large amount of training data as support. However, in the case of limited data, the discriminator is prone to overfitting, which makes it difficult for the model to converge [27,28]. Recently, some researchers [29,30] have proposed some advanced data augmentation strategies for training GANs with limited data, but these methods are mainly designed for unconditional generation. To solve the disadvantages of generative adversarial network training on limited datasets, this paper attempts to solve this problem with the few-shot learning paradigm. The GAN network generates different images for a new category by limiting the data category; we hope that GAN can generate large numbers of real and diverse images with only a few images in this category.

### 2.2. Few-Shot Image Generation

Few-shot image generation is a subfield of computer vision that aims to generate high-quality images from a small number of training examples. The goal is to develop algorithms that can generate realistic images of objects, scenes, and other visual content with a limited amount of data. Optimization-based methods [31,32] involve optimizing a model to generate high-quality images by minimizing a loss function that measures the difference between the generated and real images. These methods typically involve training a generative model, such as a generative adversarial network (GAN), to generate images from a small number of input examples. FIGR [18] and DAWSON [33,34] combine adversarial learning with meta-learning methods (i.e., reptile [35] and MAML [36]), but the quality of the generated images is low and not suitable for practical applications. Metric-based methods [37,38] involve learning a distance metric that can measure the similarity between images in a high-dimensional feature space. The generator can then generate new images by finding the closest image in the feature space to the input examples. Some notable metric-based methods include Siamese networks [39,40], which learn a distance metric between images in a feature space, and prototypical networks [41], which use

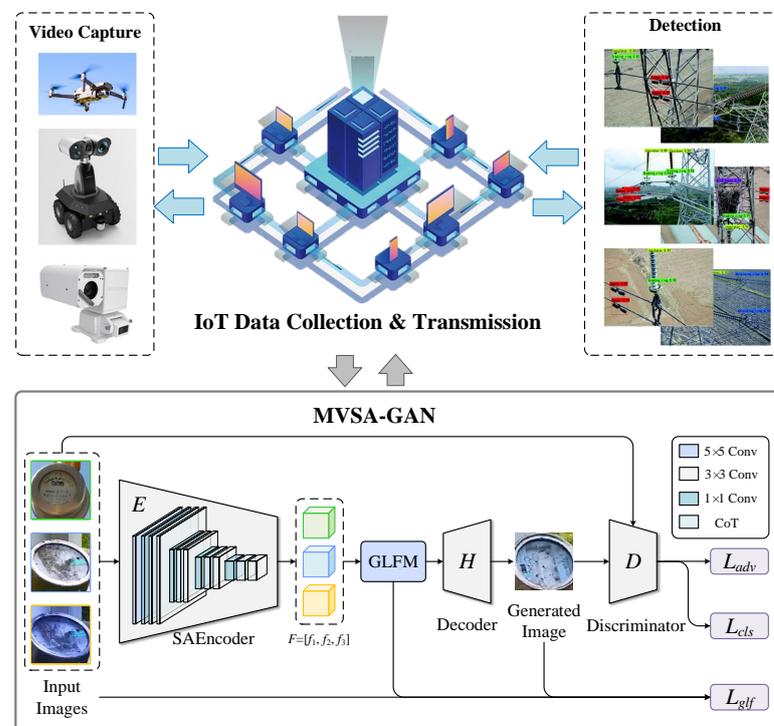
prototypes to represent classes in a feature space. Fusion-based methods [16,42,43] involve combining global and local features of the input images to generate high-quality images.

GMN [44] and MatchingGAN [45] use VAE and GAN to generalize the matching network from a small sample classification task (few-shot classification) to a small sample generation task. F2GAN [46] improves MatchingGAN by adding a non-local attention fusion module to fuse and fill features at different levels to generate images. LoFGAN fuses deep features at a finer level by selecting, matching, and replacing local representations, and uses a local-based reconstruction loss to reduce aliasing artifacts. However, the above methods only focus on local features, and the power scene is complex and dynamically changing. A single change in the local features is difficult to apply to the actual power equipment defect detection task.

### 3. MVSA-GAN

#### 3.1. Overall Framework

The overall framework of MVSA-GAN is shown in Figure 1, which includes four modules: IoT data collection, SAEncoder, GLFM, decode, and discriminator. First, sensors and devices in the IoT are used to collect various types of power scenario data, such as current, voltage, temperature, etc. These data can be combined with the subsequent MVSA-GAN to help generate synthetic images with features of electric scenes. Second, a small number of training images are input into the encoder, and operations, such as GLFM and encoder, are used to generate similar images. Then the input image and the generated image pass through a discriminator to improve the authenticity of the generated image. Class loss  $L_{cls}$  and the adversarial loss  $L_{adv}$  are used to update the parameters in the discriminator. Finally, the input image, the generated image, and the output features of GLFM undergo a global–local loss  $L_{glf}$  calculation to improve the robustness and authenticity of the generative model.



**Figure 1.** The overall structure of the MVSA-GAN.

The SAEncoder can extract more feature information from limited samples. By using the Transformer structure, it can capture contextual features, making the features extracted by the encoder more diverse and effective. This allows for better modeling and learning

of details and critical parts of electrical defect scenarios, resulting in more realistic and accurate images. Secondly, GLFM can fuse features from different scales, so that the generated images have more comprehensive and rich information. In the power defect scenario, defects are usually distributed in different locations and scales, so it is necessary to fully utilize and fuse features of different scales. Through the multi-view fusion module, we can better combine features of different scales, so that the generated images have more accurate and comprehensive defect information.

### 3.2. SAEncoder

A context encoder based on the self-attention mechanism is proposed, named SAEncoder, which uses the ability of the Transformer structure to capture context features to enrich the features extracted by the encoder. SAEncoder can improve the authenticity of subsequently generated images by learning image details. It consists of five stages, and each stage consists of a context-based Transformer (CoT) module [47]. For the self-attention mechanism of the Transformer, each input unit can calculate a weight vector based on its relative position with other input units, which can be used to weigh the feature representation of the input unit. Therefore, each unit can take into account the contextual information of the entire input sequence, not just the part before or after it. This allows the encoder to extract more information from limited samples and generate more realistic images. In addition, since each input unit can take into account the contextual information of the entire input sequence, it can better capture the long-term dependencies, and better capture the correlation and continuity between images, making the generated images more natural.

CoT is a Transformer module, and we use the Transformer structure to improve the ability of the encoder (SAEncoder) to capture contextual features. The self-attention mechanism plays a key role in CoT. The self-attentiveness mechanism works as follows:

$$Att(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where  $Q$  stands for query, which represents the query features of the  $i$ -th patch learned by the neural network,  $K$  stands for key, which represents the queried features of the  $j$ -th patch learned by the neural network, and  $V$  stands for value, which represents the semantic features learned by the neural network from the  $j$ -th patch. The self-attentive mechanism adaptively learns and matches the features learned by the neural network by such words.

The structure of CoT is shown in Figure 2. First, a  $3 \times 3$  convolution performs contextual encoding on the input  $K$  to obtain a static contextual representation of the input. Further, the concat operation is implemented on the encoded  $K$  and the input  $Q$ , and the dynamic multi-head attention matrix is learned through two consecutive  $1 \times 1$  convolutions. The learned attention matrix is multiplied by the input values to achieve a dynamic contextual representation of the input. Finally, we take the results of the static and dynamic contextual representations as output. The input  $K$  in the figure contains a large amount of contextual information, through which the learning of the dynamic attention matrix can be optimized, and the feature representation ability of the network can be further enhanced.

The hierarchical structure of the SAEncoder is shown in Table 1. In the first stage, the dimension of the input is first increased and the scale of the image is kept unchanged, and then the scale of the input image is reduced, step by step, through four stages, reducing the calculation pressure, and extracting high-dimensional features. In the last four stages, this paper constructs a block structure, which contains a  $1 \times 1$  convolution, a CoT module, and a  $3 \times 3$  convolution in the block. The former two keep the dimension of the feature unchanged to extract deep features in the feature map, and the latter uses a  $3 \times 3$  convolution to increase the dimension of the feature map. Such a structure can improve the feature extraction ability of the encoder, and further enable the generator to generate images with real image details.

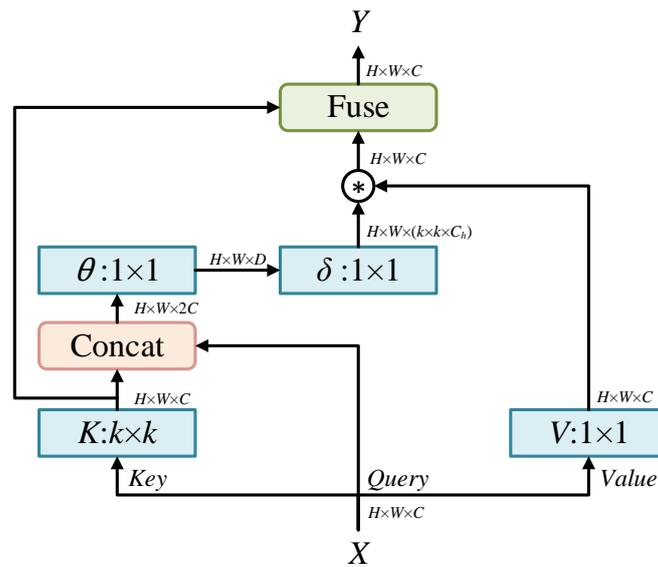


Figure 2. Context-based Transformer module structure.

Table 1. Hierarchy of the SAEncoder.

	Encoder <sup>1</sup>	Output Size
Stage 1	$5 \times 5, 32$	$128 \times 128$
Stage 2	$\begin{pmatrix} 1 \times 1, 32 \\ CoT, 32 \\ 3 \times 3, 64 \end{pmatrix}$	$64 \times 64$
Stage 3	$\begin{pmatrix} 1 \times 1, 64 \\ CoT, 64 \\ 3 \times 3, 128 \end{pmatrix}$	$32 \times 32$
Stage 4	$\begin{pmatrix} 1 \times 1, 128 \\ CoT, 128 \\ 3 \times 3, 128 \end{pmatrix}$	$16 \times 16$
Stage 5	$\begin{pmatrix} 1 \times 1, 128 \\ CoT, 128 \\ 3 \times 3, 128 \end{pmatrix}$	$8 \times 8$

<sup>1</sup> This column presents information about the layers contained in the stage, and the cell provides the convolution kernel size and the number of channel layers of the convolution.

### 3.3. Global and Local Fusion Module

The existing local fusion module (LFM) [16] randomly chooses one of the encoded features as the base and the rest of the features as references, and fuses them by local selection, local matching, and local replacement. The LFM module takes the output feature map of the encoder as input, randomly selects one of the feature maps as the base feature  $f_{base}$ , and uses the other feature maps as the reference features  $F_{ref}$ . For example, when the LFM module performs a five-shot image generation task, the LFM will use the remaining four feature maps as  $F_{ref}$ . The local fusion module will take the select  $f_{base}$  as the basis and the rest  $F_{ref}$  as a bank of local features to produce a fused feature. The entire fusion process of LFM can be divided into three steps, including local selection, local matching, and local replacement.

In the local selection stage,  $n$  local representations in  $f_{base}$  should be replaced randomly. After the local selection ends,  $n$  c-dimensional local representations  $\Phi_{base}$  from the base feature  $f_{base}$  will be obtained. In the local matching stage, LFM will look for matching local representations in  $F_{ref}$ , which can replace  $\Phi_{base}$  semantically. We calculate the similarity

matrix  $M$  of  $\Phi_{base}$  and  $f_{ref}$  by Equation (3). The feature map  $\Phi_{ref}$  that best matches  $\Phi_{base}$  can be obtained through the similarity matrix  $M$ .

$$M^{(i,j)} = g\left(\phi_{base}^{(i)}, f_{ref}^{(j)}\right) \tag{3}$$

where  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, h \times w\}$ , and  $g$  is a similarity metric,  $f_{ref} \in F_{ref}$ .

In the local replacement stage, the  $\Phi_{ref}$  and  $\Phi_{base}$  obtained in the local matching stage are weighted and summed to obtain the final fused local representation. Finally, these fused local representations are replaced on  $f_{base}$ , and the fused feature map  $F_{fuse}$  is obtained as the output of LFM.

For each  $c$ -dimensional local representation in  $\phi_{base}$ , we now have  $k - 1$  candidate local representations. For example,  $\phi_{ref}^{(1)} \in \mathbb{R}^{(k-1) \times c}$  contains the most similar local representations with the first local representation  $\phi_{base}^{(1)} \in \mathbb{R}^c$ , which we can find in every  $f_{ref}$  (see the dotted lines in the LFM module in Figure 3). We fuse all of these local representations together and replace them with the corresponding positions in  $f_{base}$ . We use a random coefficient vector  $a = [a_1, \dots, a_k]$  to fuse the features for all the positions selected,

$$\phi_{fuse}^{(t)} = a_{base} \cdot \phi_{base}^{(t)} + \sum_{i=1, \dots, k, i \neq base} a_i \cdot \phi_{ref}^{(i)}(t) \tag{4}$$

where  $\sum_{i=1}^k a_i = 1$ ,  $a_i \geq 0$ , and  $t = 1, \dots, n$ . We retain original local representation with a ratio,  $a_{base}$ . Then, we replace all the  $n$  fused local representations  $\phi_{fuse}$  with the corresponding positions in  $f_{base}$ . This produces a fused feature map  $F$  as the output of the LFM module.

LFM can well-fuse the local features of each feature map, generate a representative fusion feature map, and enhance the diversity of the generated network. However, LFM only considers the local features of the feature map, the fused feature map lacks global representation, and the semantic information of the generated fused feature map still does not change significantly, which will cause the model to be prone to overfitting.

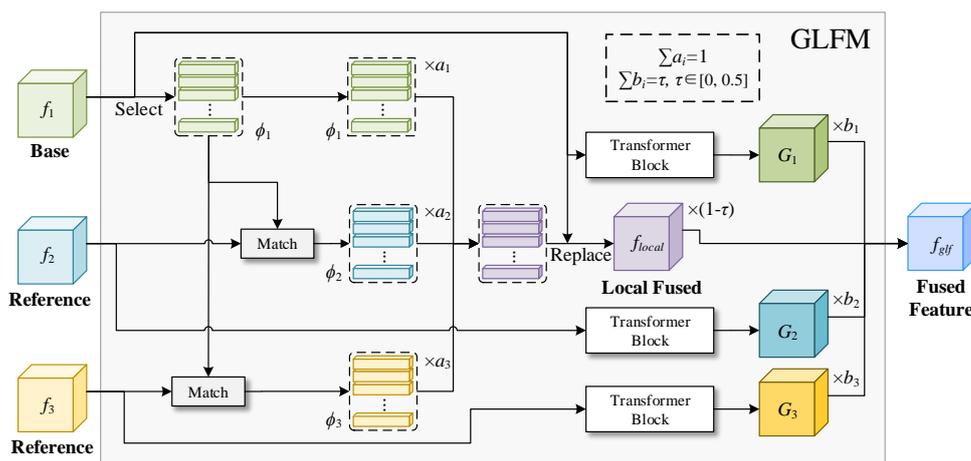


Figure 3. The overall structure of GLFM.

In response to this problem, a global and local fusion module (GLFM) is proposed in this paper, which incorporates a global representation into the fusion of local representations. It greatly integrates the global semantic information of  $F_{ref}$ , making the fused feature map more representative, and further training a more robust small-sample generation model. As shown in Figure 3, GLFM adds a Transformer block to extract the global representation of the input feature map on the basis of the LFM module. This global repre-

sensation represents the semantic information of  $f_{ref}$ . We set a random coefficient vector  $b = [b_1, \dots, b_k]$  to deeply fuse the generated global representation  $G = [G_1, \dots, G_k]$  with  $f_{local}$ , where  $k$  stands for  $k$ -shot image generation. Also, we set a threshold parameter  $\tau$  to avoid the impact of unimportant global representation on the fusion of local representations. The calculation for  $f_{glf}$  is as follows:

$$f_{glf} = (1 - \tau) \cdot f_{local} + \sum_i^k b_i \cdot G_i \quad (5)$$

where  $\tau$  and  $b$  are variable parameters, and the parameters are adaptively adjusted according to the feature map input by GLFM.

## 4. Experiments

### 4.1. Datasets

In this paper, a total of 59,169 images of defects in various scenes of substations were collected. The defect categories include blurred dials, damaged dials, cracked insulators, and metal corrosion. The collected defect images are divided into 12 categories, and the specific categories and image data information are shown in Table 2.

To demonstrate that the MVSA-GAN can generate realistic and rich images in the case of limited samples, this paper randomly selected 400 images from each category, and a total of 4800 images constitute the training dataset. At the same time, in the training process, 10 image categories were used as visible images during training, and the other 2 categories were used as invisible images.

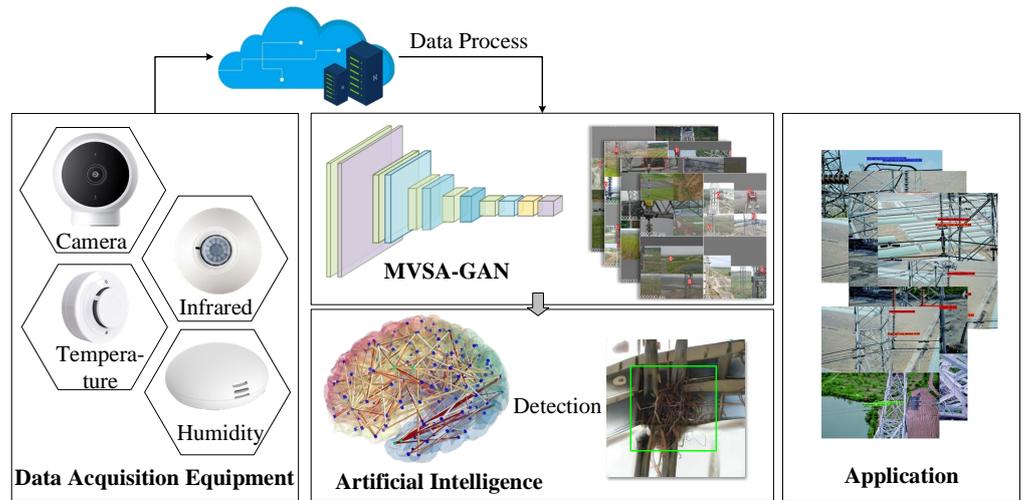
**Table 2.** Power scene defect data category information.

Category ID	Category Name	Number
1	Blurred Dial	1756
2	Broken Dial	533
3	Insulator Rupture	953
4	Insulator Crack	678
5	Oily Dirt on the Surface of Oil Leaking Parts	5922
6	Oil on the Ground	1494
7	Damaged Respirator Silicone Cartridge	527
8	Discoloration Silicone Respirator	3225
9	The Box Door is Closed Abnormally	1405
10	Suspended Solids	2006
11	The Bird's Nest	1316
12	Damaged Cover	556

### 4.2. Experiment Details

The experimental environment configuration for the proposed method in this paper is based on Ubuntu16.04.1 (Linux 4.15.0-142-generic). We used an NVIDIA GPU (GeForce RTX 3090) with 24 GB of video memory and 256 GB of system memory. The batch size for training was set to 16, with a total of 100,000 iterations.

An IoT-based platform was established to evaluate the MVSA-GAN applied to defect detection in power scenarios, as shown in Figure 4. The platform integrates various sensor devices to monitor the operating status of the power equipment and systems in real time. By collecting data from different sensors, the platform is able to generate image samples related to power scenes. Real-time sensor data collected through the IoT platform provide key information for small-sample image generation. These data can be combined with existing electrical equipment defect image samples to generate additional defect image samples through a generative algorithm. Therefore, a large number of power equipment defect images can be simulated and generated without actual defects, facilitating the training and testing of defect detection algorithms.



**Figure 4.** The IoT-based test platform for the evaluation of the MVSA-GAN.

#### 4.3. Evaluation Metrics

In the experimental part, four image quality evaluation metrics are used to quantitatively evaluate the proposed method, as follows:

The Frechet inception distance (FID) score [48] is used to calculate the distance between the real data  $S_{real}$  and the generated data  $S_{gen}$ . The similarity between the two sets of images is measured by the statistical similarity of the visual features of the original image. In general, the mean and covariance matrices are used to calculate the distance between two distributions, as follows:

$$FID(x, g) = \|\mu_x - \mu_g\|_2^2 + Tr\left(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}\right) \quad (6)$$

Learned perceptual image patch similarity (LPIPS) [49] is the similarity of learning perceptual image patches, which can measure the difference between two images. The lower the value, the more similar the two images are. The calculation formula is as follows:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l)\|_2^2 \quad (7)$$

where  $d$  is the distance between  $x_0$  and  $x$ . Extract the feature stack from the  $L$  layer and perform unit normalization in the channel dimension (unit-normalize). Use the vector  $W_L$  to scale the number of activated channels and calculate the  $L_2$  distance. Finally, we average over the space and sum over the channels.

The inception score (IS) [50] is used to evaluate the difference between two distributions, and the calculation formula is as follows:

$$IS(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) || p(y))) \quad (8)$$

where  $\mathbb{E}_{x \sim p_g}$  means to traverse all generated samples and find their average.  $D_{KL}$  represents the KL-divergence, and  $D_{KL}(p(y|x))$  is used to measure the degree of approximation between the distribution  $P$  and  $Q$ .  $p(y|x)$  represents the probability distribution of all categories for the picture  $x$ .  $p(y)$  is marginal probability.

The peak signal-to-noise ratio (PSNR) [51] is used to evaluate the noise contained in the image; the calculation formula is as follows:

$$PSNR = 10 \log_{10} \left( \frac{MAX_I^2}{MSE} \right) \quad (9)$$

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (10)$$

The size of the original image  $I$  is  $mn$ , and the image  $K$  is generated after adding noise to it.  $MAX_I$  represents the maximum pixel value of the image.

The  $F_1$  score [52] is introduced as an indicator for evaluating downstream tasks, and its calculation formula is as follows:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

#### 4.4. Evaluation Results

We compare MVSA-GAN with existing GAN-based methods, and the visualization results are shown in Figure 5. This method first judges the defect category of the input image to learn defect features, and then combines global and local features to generate corresponding virtual images. It can be seen from Figure 5 that the generated images are real and diverse, and have good processing effects on defect images in different scenarios. Furthermore, the model shows strong robustness in preserving the main features of original defects while incorporating features learned from images of other defect categories. The generated images are faithful to the original texture clarity and fidelity of the input images. By combining global and local features, generative models can capture complex and diverse textures, resulting in high-quality images that closely resemble the input data. The method's ability to learn and incorporate features from images of other defect categories enhances its robustness and adaptability to various real-world scenarios.



Figure 5. Generation samples of defect data in power scenarios by different methods.

Table 3 presents the quantitative comparative analysis between the MVSA-GAN and state-of-the-art (SOTA) methods, including FIGR [18], MatchingGAN [45], DAWSON [34], LoFGAN [16], GMN [44], and DAGAN [53]. Our method outperforms other methods in four key metrics. The proposed method achieves 67.87 and 0.179 on the FID and LPIPS, and 164.23 and 23.41 on the IS and PSNR. The experimental results show that the proposed method can capture complex textures and generate high-quality images that are very similar to the input data. In terms of IS and PSNR, it is 11.53 and 4.1 higher than the

next highest method, respectively. It can be seen that the proposed method is robust and adaptable to complex power scenarios.

**Table 3.** Comparative experiment.

Method	The Smaller the Better		The Larger the Better	
	FID	LPIPS	IS	PSNR
FIGR	83.09	0.320	137.42	12.84
MatchingGAN	73.85	0.241	148.35	19.31
DAWSON	82.49	0.337	140.14	14.32
LoFGAN	86.98	0.362	133.95	12.47
GMN	79.14	0.292	141.33	14.85
DAGAN	75.75	0.265	152.70	18.46
Ours	<b>67.87</b>	<b>0.179</b>	<b>164.23</b>	<b>23.41</b>

#### 4.5. Ablation Experiment

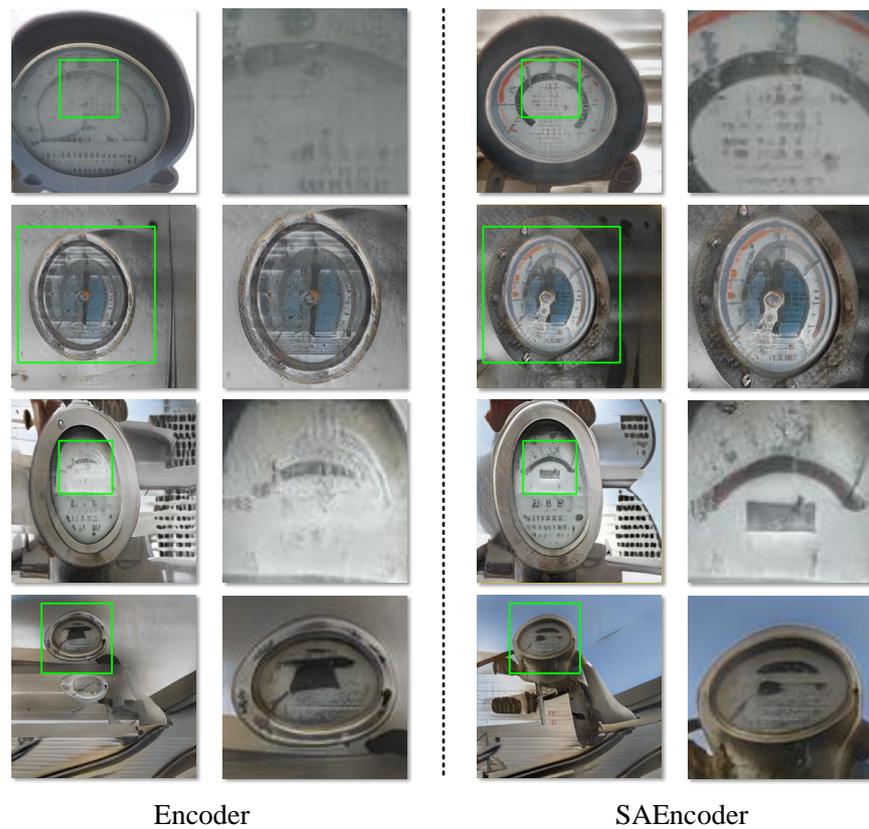
In this section, we evaluate the effectiveness of the proposed GLFM and SAEncoder in the generation of the electrical scene defect images through ablation experiments, as shown in Table 4. Through the quality evaluation of the image generated by a single module and the joint action of the two, it is obvious that the authenticity of the image generated by the proposed method can be seen. In particular, after adding GLFM, the three indicators used for measuring image quality are significantly optimized. Under the joint action of GLFM and SAEncoder, the FID, IS, and PSNR of the final image generated by the model are 67.87, 164.23, and 23.41, respectively. It can be seen from the changes in the data that the proposed method can well adapt to complex power scenes, demonstrating its ability to capture and learn subtle defect features in the scene.

**Table 4.** Ablation experiment.

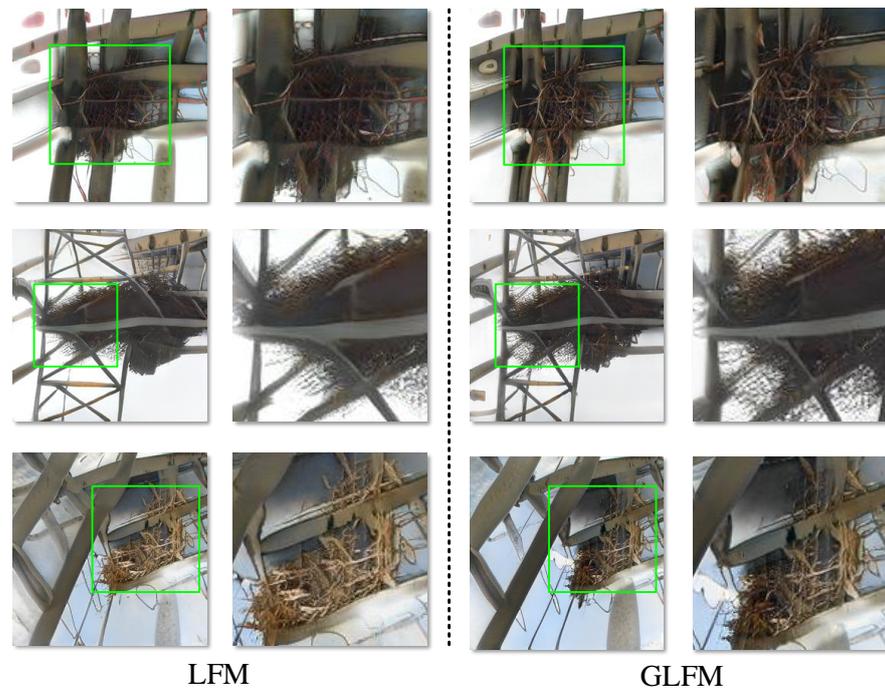
	+GLFM	+SAEncoder	FID	IS	PSNR
Net.1			86.98	133.95	12.47
Net.2	✓		73.50	154.76	18.58
Net.3		✓	72.28	150.35	17.16
Net.4	✓	✓	<b>67.87</b>	<b>164.23</b>	<b>23.41</b>

Figure 6 specifically shows the feature learning ability of the SAEncoder for the details in complex power defect scenes. For the encoder without the self-attention mechanism, it cannot capture comprehensive contextual information due to the limitation of the receptive field. This results in an incoherent generated image with blurred, distorted details. The SAEncoder establishes long-distance connections to global features through a self-attention mechanism, making image generation more natural.

Figure 7 provides a visualization of the effectiveness of GLFM in generating models. From Figure 7, the inclusion of GLFM enhances the generated image by making it clearer and more detailed. In the power scene defect image generation task, since there are many details and texture information in the image, the fusion of global features and local features is particularly important for generating high-quality images. Global features and local features have different advantages in extracting image information. Global features can capture the global structure and background information of the entire image, while local features can capture the detailed information of specific regions in the image. GLFM can lead to the interaction and integration of local features and global features, to better capture the structure and detailed information of the image and improve the realism and clarity of the generated image.



**Figure 6.** On the left, we can see the synthesis result obtained using the baseline encoder. On the right side, we can observe the synthesis result obtained using SAEncoder.



**Figure 7.** The left image shows the network synthesis result without GLFM, and the right image shows the network synthesis result using GLFM.

#### 4.6. Downstream Tasks

To demonstrate that the MVSA-GAN has an enhanced effect on downstream tasks, in this paper, ResNet50 [54] and VGG16 [55] are used as classification networks, and training and evaluation are performed on datasets with and without synthetic data. The evaluation results are shown in Table 5. “Real” represents a small number of real samples, and “Sample” represents the data after image generation using the proposed method. The results show that the images generated by the generative model have a large improvement in the accuracy of the classification task. This is because the generative model can generate a large amount of virtual data, which enriches the sample diversity of the training dataset, helps the model to better learn the characteristics of the data, and improves the generalization ability of the model.

Since the power scene defect dataset is usually small in size, using real datasets for training may lead to overfitting of the model, which cannot effectively capture the characteristics of the data. The method proposed in this paper can increase the diversity of data by generating virtual data, which can help improve the generalization ability of the model and the accuracy of defect detection.

**Table 5.** Optimization of model performance for downstream tasks with generated data.

Network	Dataset	F <sub>1</sub> Score			Total
		Meter Breakage	Metal Corrosion	Broken Insulator	
ResNet50	Real	89.0%	79.7%	90.4%	87.1%
ResNet50	Real+Sample	92.8%	84.4%	94.2%	90.8%
VGG16	Real	86.3%	78.3%	89.5%	85.6%
VGG16	Real+Sample	88.1%	85.0%	94.1%	89.4%

## 5. Conclusions

To address the scarcity of datasets for defect detection in electric power scenes, a generative adversarial network based on self-attention multi-view fusion (MVSA-GAN) is proposed for few-shot image generation. First, the power scene is complex and dynamic, and complex background features will cause artifacts in the generated image. SAEncoder is proposed to face the above challenge. By focusing on the relevant features of each part to capture contextual information, the generated image is more coherent and global, and effectively alleviates the artifacts and confusion generated by complex backgrounds. Secondly, to overcome the problem of existing small sample generation networks in generating images with both authenticity and diversity, GLFM is proposed. This module is used to selectively fuse global and local features, not only capturing background features but also extracting structure and texture details of defect parts, so as to achieve the authenticity and diversity of generated images. The proposed method provides high-quality data for electrical scene defect detection, achieving 67.87 and 0.179 on FID and LPIPS scores.

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

**Funding:** This research was funded in part by the National Natural Science Foundation of China (grant nos. U1909201, 62101490, 6212780029, and 2022C01056), in part by the Research Startup Funding from Hainan Institute of Zhejiang University (grant no. 0210-6602-A12203), in part by the State Grid Corporation of China Technology Project (grant no. 5700-202019487A-0-0-00), and in part by the National Natural Science Foundation of Zhejiang Province (grant no. LQ21F030017).

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data generated during this study are currently private due to our team is about to have a new breakthrough in this study. The data will be made public soon.

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

## References

1. Mohamad Noor, M.B.; Hassan, W.H. Current research on Internet of Things (IoT) security: A survey. *Comput. Netw.* **2019**, *148*, 283–294. [[CrossRef](#)]
2. Liu, Z.; Wu, G.; He, W.; Fan, F.; Ye, X. Key target and defect detection of high-voltage power transmission lines with deep learning. *Int. J. Electr. Power Energy Syst.* **2022**, *142*, 108277. [[CrossRef](#)]
3. Tao, X.; Zhang, D.; Wang, Z.; Liu, X.; Zhang, H.; Xu, D. Detection of Power Line Insulator Defects Using Aerial Images Analyzed With Convolutional Neural Networks. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *50*, 1486–1498. [[CrossRef](#)]
4. Zhao, H.; Dai, R.; Xiao, C. A Machine Vision System for Stacked Substrates Counting With a Robust Stripe Detection Algorithm. *IEEE Trans. Syst. Man, Cybern. Syst.* **2019**, *49*, 2352–2361. [[CrossRef](#)]
5. Tsai, D.M.; Wu, S.C.; Chiu, W.Y. Defect Detection in Solar Modules Using ICA Basis Images. *IEEE Trans. Ind. Inform.* **2013**, *9*, 122–131. [[CrossRef](#)]
6. Velichko, A.; Korzun, D.; Meigal, A. Artificial Neural Networks for IoT-Enabled Smart Applications: Recent Trends. *Sensors* **2023**, *23*, 4853. [[CrossRef](#)]
7. Weerasinghe, S.; Zaslavsky, A.; Loke, S.W.; Hassani, A.; Medvedev, A.; Abken, A. Adaptive Context Caching for IoT-Based Applications: A Reinforcement Learning Approach. *Sensors* **2023**, *23*, 4767. [[CrossRef](#)]
8. Xiao, L.; Wan, X.; Lu, X.; Zhang, Y.; Wu, D. IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security? *IEEE Signal Process. Mag.* **2018**, *35*, 41–49. [[CrossRef](#)]
9. Dong, Z.; He, Y.; Hu, X.; Qi, D.; Duan, S. Flexible memristor-based LUC and its network integration for Boolean logic implementation. *IET Nanodielectr.* **2019**, *2*, 61–69. [[CrossRef](#)]
10. Zhang, Z.; Dong, Z.; Lin, H.; He, Z.; Wang, M.; He, Y.; Gao, X.; Gao, M. An improved bidirectional gated recurrent unit method for accurate state-of-charge estimation. *IEEE Access* **2021**, *9*, 11252–11263. [[CrossRef](#)]
11. Davari, N.; Akbarizadeh, G.; Mashhour, E. Corona Detection and Power Equipment Classification Based on GoogleNet-AlexNet: An Accurate and Intelligent Defect Detection Model Based on Deep Learning for Power Distribution Lines. *IEEE Trans. Power Deliv.* **2022**, *37*, 2766–2774. [[CrossRef](#)]
12. Geng, Z.; Shi, C.; Han, Y. Intelligent Small Sample Defect Detection of Water Walls in Power Plants Using Novel Deep Learning Integrating Deep Convolutional GAN. *IEEE Trans. Ind. Inform.* **2022**, *19*, 7489–7497. [[CrossRef](#)]
13. Ojha, U.; Li, Y.; Lu, J.; Efros, A.A.; Lee, Y.J.; Shechtman, E.; Zhang, R. Few-shot image generation via cross-domain correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10743–10752.
14. Zhao, Y.; Ding, H.; Huang, H.; Cheung, N.-M. A closer look at few-shot image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9140–9150.
15. Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10551–10560.
16. Gu, Z.; Li, W.; Huo, J.; Wang, L.; Gao, Y. Lofgan: Fusing local representations for few-shot image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 8463–8471.
17. Dong, Z.; Ji, X.; Zhou, G.; Gao, M.; Qi, D. Multimodal Neuromorphic Sensory-Processing System with Memristor Circuits for Smart Home Applications. *IEEE Trans. Ind. Appl.* **2022**, *59*, 47–58. [[CrossRef](#)]
18. Clouâtre, L.; Demers, M. Figr: Few-shot image generation with reptile. *arXiv* **2019**, arXiv:1901.02199.
19. Dong, Z.; Li, C.; Qi, D.; Luo, L.; Duan, S. Multiple memristor circuit parametric fault diagnosis using feedback-control doublet generator. *IEEE Access* **2016**, *4*, 2604–2614. [[CrossRef](#)]
20. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
21. 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. Proces. Syst.* **2017**, 6000–6010.
22. 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](#)]
23. Ji, X.; Dong, Z.; Lai, C.S.; Qi, D. A physics-oriented memristor model with the coexistence of NDR effect and RS memory behavior for bio-inspired computing. *Mater. Today Adv.* **2022**, *16*, 100293. [[CrossRef](#)]
24. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3313–3332. [[CrossRef](#)]

25. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
26. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Langs, G.; Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* **2019**, *54*, 30–44. [[CrossRef](#)] [[PubMed](#)]
27. Ji, X.; Lai, C.S.; Zhou, G.; Dong, Z.; Qi, D.; Lai, L.L. A Flexible Memristor Model with Electronic Resistive Switching Memory Behavior and its Application in Spiking Neural Network. *IEEE Trans. Nanobiosci.* **2022**, *22*, 52–62. [[CrossRef](#)] [[PubMed](#)]
28. 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, 18–23 June 2018; pp. 1316–1324.
29. Wang, J.; Chen, Y.; Dong, Z. SABV-Depth: A biologically inspired deep learning network for monocular depth estimation. *Knowl.-Based Syst.* **2023**, *263*, 110301. [[CrossRef](#)]
30. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
31. Wang, D.; Cheng, Y.; Yu, M.; Guo, X.; Zhang, T. A hybrid approach with optimization-based and metric-based meta-learner for few-shot learning. *Neurocomputing* **2019**, *349*, 202–211. [[CrossRef](#)]
32. Ding, G.; Han, X.; Wang, S.; Wu, S.; Jin, X.; Tu, D.; Huang, Q. Attribute Group Editing for Reliable Few-shot Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11194–11203.
33. Ji, X.; Qi, D.; Dong, Z.; Lai, C.S.; Zhou, G.; Hu, X. TSSM: Three-State Switchable Memristor Model Based on Ag/TiO<sub>x</sub> Nanobelt/Ti Configuration. *Int. J. Bifurc. Chaos* **2021**, *31*, 2130020. [[CrossRef](#)]
34. Liang, W.; Liu, Z.; Liu, C. Dawson: A domain adaptive few shot generation framework. *arXiv* **2020**, arXiv:2001.00576.
35. Nichol, A.; Schulman, J. Reptile: A scalable metalearning algorithm. *arXiv* **2018**, arXiv:1803.02999.
36. Antoniou, A.; Edwards, H.; Storkey, A. How to train your MAML. *arXiv* **2018**, arXiv:1810.09502.
37. Tseng, H.-Y.; Lee, H.-Y.; Huang, J.-B.; Yang, M.-H. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv* **2020**, arXiv:2001.08735.
38. Li, X.; Yu, L.; Fu, C.-W.; Fang, M.; Heng, P.-A. Revisiting metric learning for few-shot image classification. *Neurocomputing* **2020**, *406*, 49–58. [[CrossRef](#)]
39. Dong, Z.; Ji, X.; Lai, C.S.; Qi, D. Design and Implementation of a Flexible Neuromorphic Computing System for Affective Communication via Memristive Circuits. *IEEE Commun. Mag.* **2022**, *61*, 74–80. [[CrossRef](#)]
40. Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Bordes, F.; Vincent, P.; Joulin, A.; Rabbat, M.; Ballas, N. Masked siamese networks for label-efficient learning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 456–473.
41. Hong, Y.; Niu, L.; Zhang, J.; Zhang, L. Matchinggan: Matching-Based Few-Shot Image Generation. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
42. Wang, J.; Chen, Y.; Dong, Z.; Gao, M. Improved YOLOv5 network for real-time multi-scale traffic sign detection. *Neural Comput. Appl.* **2023**, *35*, 7853–7865. [[CrossRef](#)]
43. Ji, X.; Dong, Z.; Han, Y.; Lai, C.S.; Zhou, G.; Qi, D. EMSN: An Energy-Efficient Memristive Sequencer Network for Human Emotion Classification in Mental Health Monitoring. *IEEE Trans. Consum. Electron.* **2023**. [[CrossRef](#)]
44. Bartunov, S.; Vetrov, D. Few-shot generative modelling with generative matching networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Playa Blanca, Lanzarote, Spain, 9–11 April 2018; pp. 670–678.
45. Li, X.; Yang, X.; Ma, Z.; Xue, J.H. Deep metric learning for few-shot image classification: A Review of recent developments. *Pattern Recognit.* **2023**, 109381. [[CrossRef](#)]
46. Hong, Y.; Niu, L.; Zhang, J.; Zhao, W.; Fu, C.; Zhang, L. F2gan: Fusing-and-filling gan for few-shot image generation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle WA, USA, 12–16 October 2020; pp. 2535–2543.
47. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [[CrossRef](#)]
48. 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, 13–19 June 2020; pp. 6070–6079.
49. Talebi, H.; Milanfar, P. Learned perceptual image enhancement. In Proceedings of the 2018 IEEE International Conference on Computational Photography (ICCP), Pittsburgh, PA, USA, 4–6 May 2018; pp. 1–13.
50. Barratt, S.; Sharma, R. A note on the inception score. *arXiv* **2018**, arXiv:1801.01973.
51. Wang, Y.; Li, J.; Lu, Y.; Fu, Y.; Jiang, Q. Image quality evaluation based on image weighted separating block peak signal to noise ratio. In Proceedings of the International Conference on Neural Networks and Signal Processing, Nanjing, China, 14–17 December 2003; pp. 994–997.
52. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [[CrossRef](#)]
53. Yang, G.; Yu, S.; Dong, H.; Slabaugh, G.; Dragotti, P.L.; Ye, X.; Liu, F.; Arridge, S.; Keegan, J.; Guo, Y. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imaging* **2017**, *37*, 1310–1321. [[CrossRef](#)]

54. de la Rosa, F.L.; Gómez-Sirvent, J.L.; Sánchez-Reolid, R.; Morales, R.; Fernández-Caballero, A. Geometric transformation-based data augmentation on defect classification of segmented images of semiconductor materials using a ResNet50 convolutional neural network. *Expert Syst. Appl.* **2022**, *206*, 117731. [[CrossRef](#)]
55. Duan, J.; Liu, X. Online monitoring of green pellet size distribution in haze-degraded images based on VGG16-LU-Net and haze judgment. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5006316.

**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.