remote sensing

MDPI

Article

Remote Sensing Image Augmentation Based on Text Description for Waterside Change Detection

Chen Chen [1], Hongxiang Ma [1], Guorun Yao [1], Ning Lv [2], Hua Yang [3], Cong Li [4] and Shaohua Wan [5,*]

[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China; cc2000@mail.xidian.edu.cn (C.C.); mhongxiang@163.com (H.M.); gr_yao@126.com (G.Y.)
[2] School of Electronic Engineering, Xidian University, Xi'an 710071, China; nlv@mail.xidian.edu.cn
[3] School of Economics and Management, Northwest University, Xi'an 710127, China; yang.flower@163.com
[4] State Grid JiLin Province Electric Power Company Limited Information Communication Company, Changchun 130000, China; congli8462@163.com
[5] The School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China
* Correspondence: shaohua.wan@ieee.org

Abstract: Since remote sensing images are difficult to obtain and need to go through a complicated administrative procedure for use in China, it cannot meet the requirement of huge training samples for Waterside Change Detection based on deep learning. Recently, data augmentation has become an effective method to address the issue of an absence of training samples. Therefore, an improved Generative Adversarial Network (GAN), i.e., BTD-sGAN (Text-based Deeply-supervised GAN), is proposed to generate training samples for remote sensing images of Anhui Province, China. The principal structure of our model is based on Deeply-supervised GAN(D-sGAN), and D-sGAN is improved from the point of the diversity of the generated samples. First, the network takes Perlin Noise, image segmentation graph, and encoded text vector as input, in which the size of image segmentation graph is adjusted to 128 × 128 to facilitate fusion with the text vector. Then, to improve the diversity of the generated images, the text vector is used to modify the semantic loss of the downsampled text. Finally, to balance the time and quality of image generation, only a two-layer Unet++ structure is used to generate the image. Herein, "Inception Score", "Human Rank", and "Inference Time" are used to evaluate the performance of BTD-sGAN, StackGAN++, and GAN-INT-CLS. At the same time, to verify the diversity of the remote sensing images generated by BTD-sGAN, this paper compares the results when the generated images are sent to the remote sensing interpretation network and when the generated images are not added; the results show that the generated image can improve the precision of soil-moving detection by 5%, which proves the effectiveness of the proposed model.

Keywords: data augmentation; deeply monitoring; GAN; remote sensing image; text description

check for updates

Citation: Chen, C.; Ma, H.; Yao, G.; Lv, N.; Yang, H.; Li, Cong; Wan, S. Remote Sensing Image Augmentation Based on Text Description for Waterside Change Detection. *Remote Sens.* **2021**, *13*, 1894. https://doi.org/10.3390/rs13101894

Academic Editors: Priyan Malarvizhi Kumar, Hari Mohan Pandey and Gautam Srivastava

Received: 29 March 2021
Accepted: 5 May 2021
Published: 12 May 2021

1. Introduction

With the rapid development of remote sensing technology [1], it is relatively easy to acquire a remote sensing image, but there are still problems: the acquired image cannot be used immediately and often requires a cumbersome processing process. Among them, the obtained samples lack the corresponding label, which requires a high sample label for the research of deep learning. Researchers need to spend a great deal of energy to annotate the existing image, and this has greatly hindered the widespread use of remote sensing images. How to save time and labor costs with the labeling of high-quality samples has become an urgent problem to be solved. As an effective means to solve this problem, data augmentation has become a hot research topic.

As an important branch in remote sensing, remote sensing dynamic soil detection has a high demand for remote sensing images. However, there is a lack of remote sensing data, and the diversity of samples is not enough to improve the generalization ability of the network. Taking the research on change detection (including dynamic soil detection) as an example, some studies ignore the problem of the lack of images [2] and the security reason to share images [3,4], but others pay attention to this problem and propose various data augmentation strategies to solve it [5,6]. Why is data augmentation strategy needed? The reasons are as follows. The current training flow commonly used by remote sensing interpretation networks (i.e., the detection network in the change detection task) is shown in Figure 1.

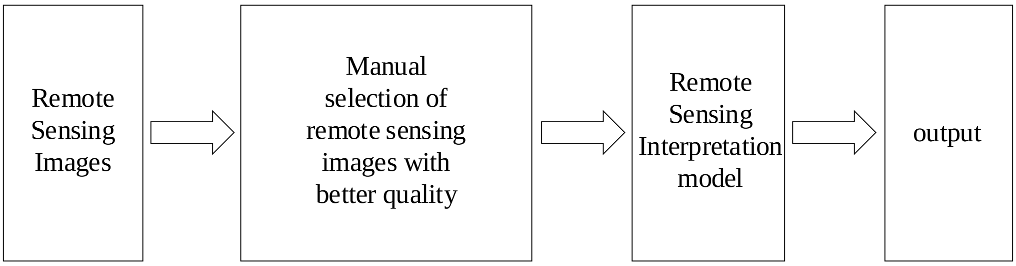
Remote Sensing Images
Manual selection of remote sensing images with better quality
Remote Sensing Interpretation model
output

Figure 1. Remote sensing interpretation network training flow chart.

As can be seen from Figure 1, the staff need to select the better-quality remote sensing image for the interpretation task, but the time cost of this process is huge. This problem is caused by the low quantity and poor quality of remote sensing images. With the development of artificial intelligence, data augmentation is an effective method to solve this problem. It can enlarge the sample in a small amount of data and satisfy the requirement of deep learning. Therefore, data augmentation is used to expand the remote sensing image data, and the accuracy of the remote sensing interpretation network is improved. Data augmentation steps are added to the training flow of remote sensing interpretation, as shown in Figure 2.

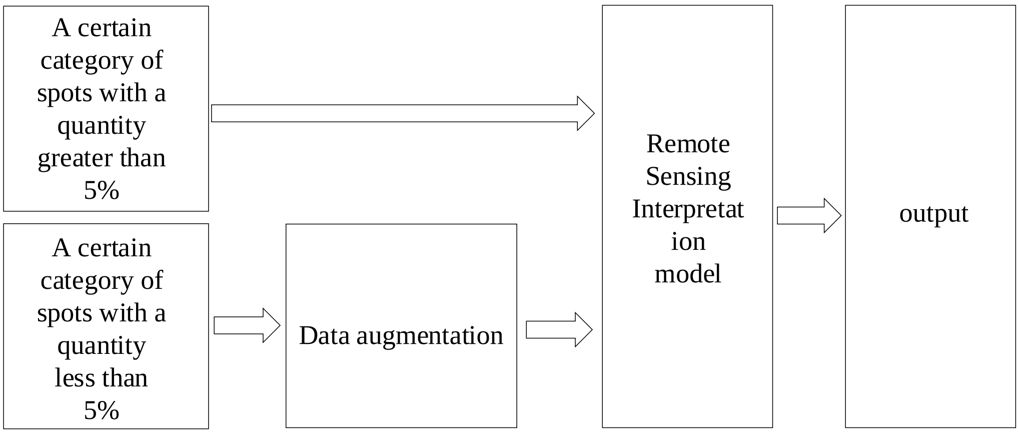
A certain category of spots with a quantity greater than 5%
A certain category of spots with a quantity less than 5%
Data augmentation
Remote Sensing Interpretat ion model
output

Figure 2. Remote sensing interpretation network with improved training flow chart.

Data augmentation generally includes traditional data augmentation algorithms and data augmentation algorithms based on deep learning [7]. The former includes rollover, scaling, cropping, and rotation [8]. These algorithms perform geometric transformations on existing images to increase the number of images. The latter includes variational autoencoder VAE [9] and generative adversarial network GAN [10], both are based on multilayer neural networks. VAE can map low-dimensional inputs to high-dimensional data, but they need prior knowledge; it is more convenient to use GAN for data augmentation without knowing the complicated reasoning process in advance. The training process for data augmentation of GAN is shown in Figure 3.

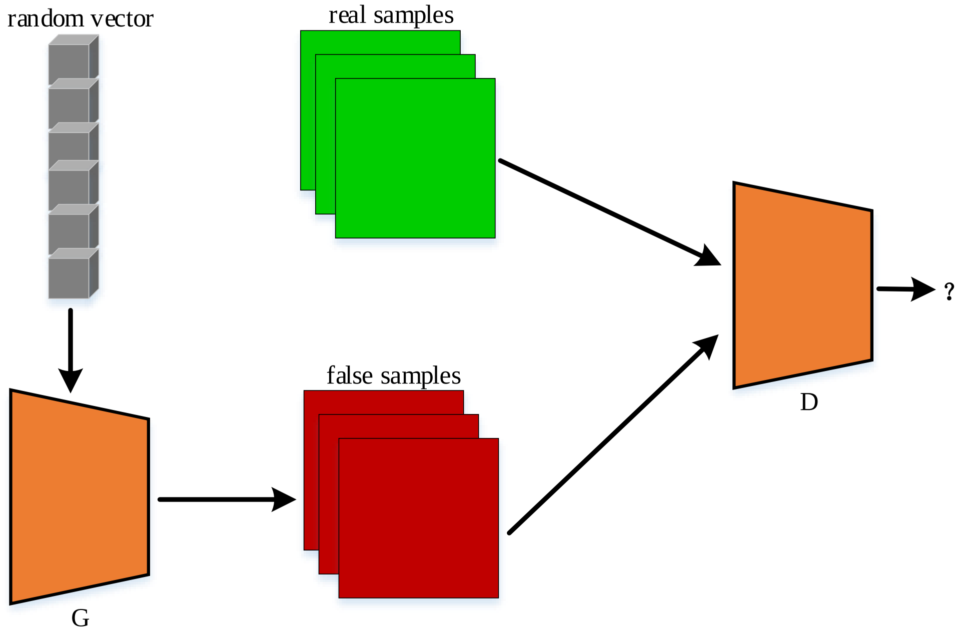

Figure 3. GAN training flow chart. G represents the generator of GAN and D represents the discriminator of GAN. The function of G is to learn the mapping rules of the random noise to the generated data and then obtain the generated image (the false sample). D is used to determine whether a sample is a real sample or a false sample.

In recent years, good progress has been made in image data augmentation. To facilitate the work, the related research is introduced from these directions: conditional generative adversarial network (cGAN), image generation, and image semantics and text semantic loss.

1.1. Conditional Generative Adversarial Network

Compared with the original generative adversarial network, the conditional generative adversarial network adds the constraint information at network's input. Still, it has made great progress in image generation. P. et al. regarded the conditional generation antagonism network as a general solution for image generation [11]. The network proposed by P. takes the sketch of the image as the conditional constraint information and generates the image from the sketch [12]. The generation of remote sensing data also belongs to the field of image generation. Herein, the research is based on the generative adversarial network.

1.2. Image Generation

At present, image generation based on GAN can be divided into two categories: the first is to generate the image of the specified category; the second is to generate the image matching the text description.

In 2014, Based on cGAN, J. et al. used random noise and specific attribute information as input, and randomly used conditional data sampling in the training process to generate a good face image [13]. In the framework of the Pierre-Simon Laplace pyramid, E. and his colleagues constructed a cascade generation confrontation network in 2015, which can generate high-quality natural images from coarse to fine [14]. In 2016, C.K. et al. applied the GAN to the image super-resolution problem. In the process of training the network, backpropagation of the gradient estimation after deagitation was performed. Good results were achieved in natural image generation in the ImageNET dataset [15]. A.M. et al. proposed a new method of image generation, DGN-AM, which is based on a prior DGN (deep generator network) and combined with the AM (activation maximization) method. By maximizing the activation functions of one or more neurons in the classifier, a realistic image is synthesized [16]. In 2017, A. et al. proposed PPGN based on DGN-AM, consisting of a generator G and a conditional network C that tells the generator to generate classes; it generated high-quality images and performed well in image repair tasks [17]. W.R. et al. proposed an ArtGAN to generate natural images such as birds, flowers, faces, and rooms [18].

In 2016, S. et al. encoded the text description into character vector as part of the input of generator and discriminator, respectively, based on the conditional generative adversarial network, the assumption that text descriptions can be used to generate images was validated on general datasets such as MS COCO [19–21]. S. et al. proposed a GAWWN network, in which a constraint box is added to guide the network to generate a certain attitude image at a given position [22]. In 2017, H. and others applied the idea of distributed generation to the generation of confrontation network and proposed a StackGAN model [23,24]. The first step is to generate a relatively fuzzy image, mainly the background, contour, etc. The second step is to take the image generated in the first step as the input; at the same time, text features are fused to correct the loss of the first stage, resulting in a high-definition image. In 2018, H. et al. improved the StackGAN model by using different group generators and discriminators to train at the same time. Images with different accuracy were generated. The low-accuracy images were trained in the high-accuracy generators, different group generators and discriminators use the same text features as constraints, resulting in better results than other generation models [25]. T. and others improved the StackGAN model using the attention mechanism, proposed the ATTNGAN model, paid more attention to the related words in the text description in the process of the phased generation, and generated more detailed information in different subregions of the image [26,27]. S. and others put forward a model of image generation based on semantic layout. Firstly, the corresponding semantic layout of the text is obtained by a layout generator, then the corresponding images are generated by an image generator. Finally, the validity of the model is verified on the MS-COCO dataset, and a natural image of diversity is generated [28,29]. Although the abovementioned GANs have achieved good results in the field of image generation, most of these were generated for natural images. Remote sensing images are different from natural images because of their unique spectral characteristics and huge amount of data, requiring high quality, speed, and diversity. The proposed model (BTD-sGAN) is suitable for remote sensing image generation to solve these problems.

In addition, generating the corresponding image from the text description involves the knowledge of multimodal representation learning. In 2021, F. et al. proposed a network named EAAN that can correlate visual and textual content [30], and also performed research on natural images. This paper attempts to study remote sensing images.

1.3. Image Semantics and Text Semantic Loss

In image processing, semantic loss is inevitable in the process of image convolution or downsampling. To avoid semantic loss of the image, T. et al. [31] proposed a new conditional normalization method, called SPADE, which solves the problem of semantic loss in batch normalization, but does not pay attention to the semantic loss of text. Therefore, this paper improves the downsampling process of the generator, adds the text feature to constrain, reduces the semantic loss of the text, and improves the diversity of the generated images.

Herein, the work is based on the structure of GAN because of the excellent effect of GAN on several datasets [32,33]. The task of target detection and image segmentation based on remote sensing image needs not only the generated image, but also the corresponding label of the image. Although GAN has achieved good results in natural image generation, there is little research on remote sensing image generation in GAN. Herein, the following problems will be solved: (1) the number of tagged remote sensing images is little; (2) the diversity of remote sensing image samples is insufficient.

Herein, an improved model named BTD-sGAN (Text-based Deeply-supervised GAN) is proposed. To solve the problem of insufficient samples with labels, we use the network segmentation graph as input in the input of the BTD-sGAN, which can restrict the process of image generation to avoid the final image of the secondary annotation. To solve the problem of insufficient sample diversity, the main body of BTD-sGAN is the deeply-supervised generation network, D-sGAN (Deeply-supervised GAN) [34], the generator structure is still Unet++ network and the discriminator structure is FCN network. BTD-sGAN takes the image segmentation graph, Perlin and text vector, which are fused as input. At the same time, to reduce the semantic loss of the text, the text vector is always used as a supervisor to correct the loss during the downsampling process. The experimental results for BTD-sGAN show that the improved network can not only increase the number of generated samples with tags, but also increase the diversity of generated samples.

2. Materials and Methods

2.1. Methods

Herein, the practical application direction is as a remote sensing dynamic soil detection project data generation module, mainly for China remote sensing data for the experiment. Remote sensing dynamic soil detection is used to identify and label some types of buildings that violate regulations through the image segmentation network, but due to the lack of remote sensing data, interpretation accuracy is faced with a breakthrough bottleneck. Therefore, this paper is based on the above remote sensing data for the study of data augmentation.

The improved model (BTD-sGAN) is based on D-sGAN, and the training process is similar. It should be noted that the Gaussian noise at the input of the generator is replaced by Perlin noise, and the segmented image and encoded text vector are fused. The discriminator also adds a text vector as a constraint. The improved generative adversarial network learns the mapping of segmentation graph x, image z, and text vector v to real image y. The image z follows the Perlin distribution. The training flow for the entire network is shown in Figure 4.

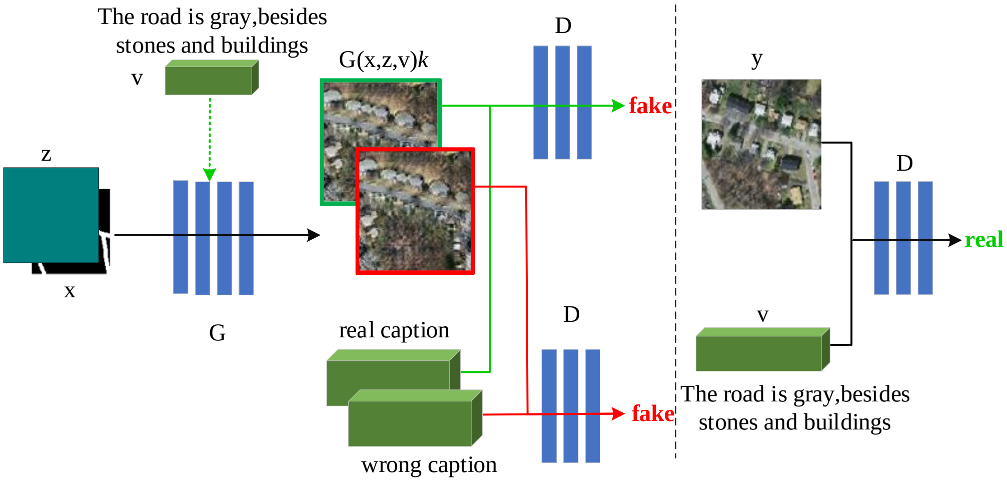
The road is gray,besides stones and buildings
v
z
x
G
G(x,z,v)*k*
D
fake
real caption
wrong caption
D
fake
y
D
real
v
The road is gray,besides stones and buildings

Figure 4. BTD-sGAN training process.

In Figure 4, an image segmentation graph x is added to the input to solve the problem that the GAN-INT-CLS [19] model cannot capture localization constraints in the image. Herein, the experiment verifies the effectiveness of adding a segmentation graph at the input end.

2.1.1. Lower Sampling Procedure

Different from the downsampling module in the D-sGAN model, to improve the diversity of the generated samples and reduce the semantic loss of the text, the method of using segmentation graph to monitor was not used, only the real text feature vector was used to supervise the sampling process. It is important to note that this subsampling procedure was applied to generators and discriminators. The down-sampling module of BTD-sGAN is shown in Figure 5.

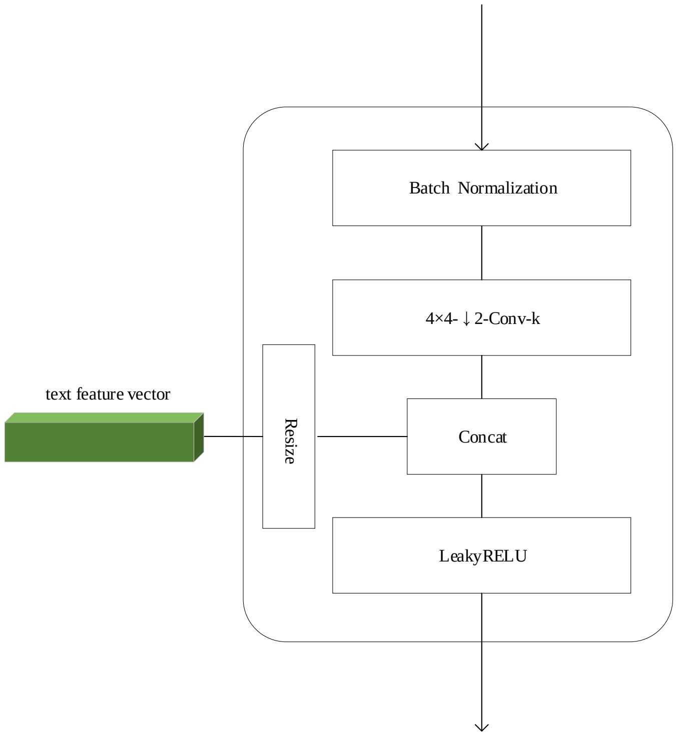

Figure 5. The down-sampling module of BTD-sGAN.

2.1.2. BTD-sGAN Structure

The Unet++ network uses a "dense link" network structure [35], which can effectively combine the features from the encoder and the decoder to reduce the semantic loss of the image, so the model of BTD-sGAN based on Unet++ is improved. In the D-sGAN, the idea of using multiple discriminators to supervise the generator was put forward, which can improve the quality of image generation and reduce the generation of image at the same time. Although the main structure of the generator was based on Unet++, discriminators (the first and second discriminators of BTD-sGAN L_4 in Figure 6) were only used to monitor the output of the second and fourth layers. The down-sampling module mentioned in Section 2.1.1 was used for both the generator and discriminator. A schematic of the entire network structure is shown in Figure 6.

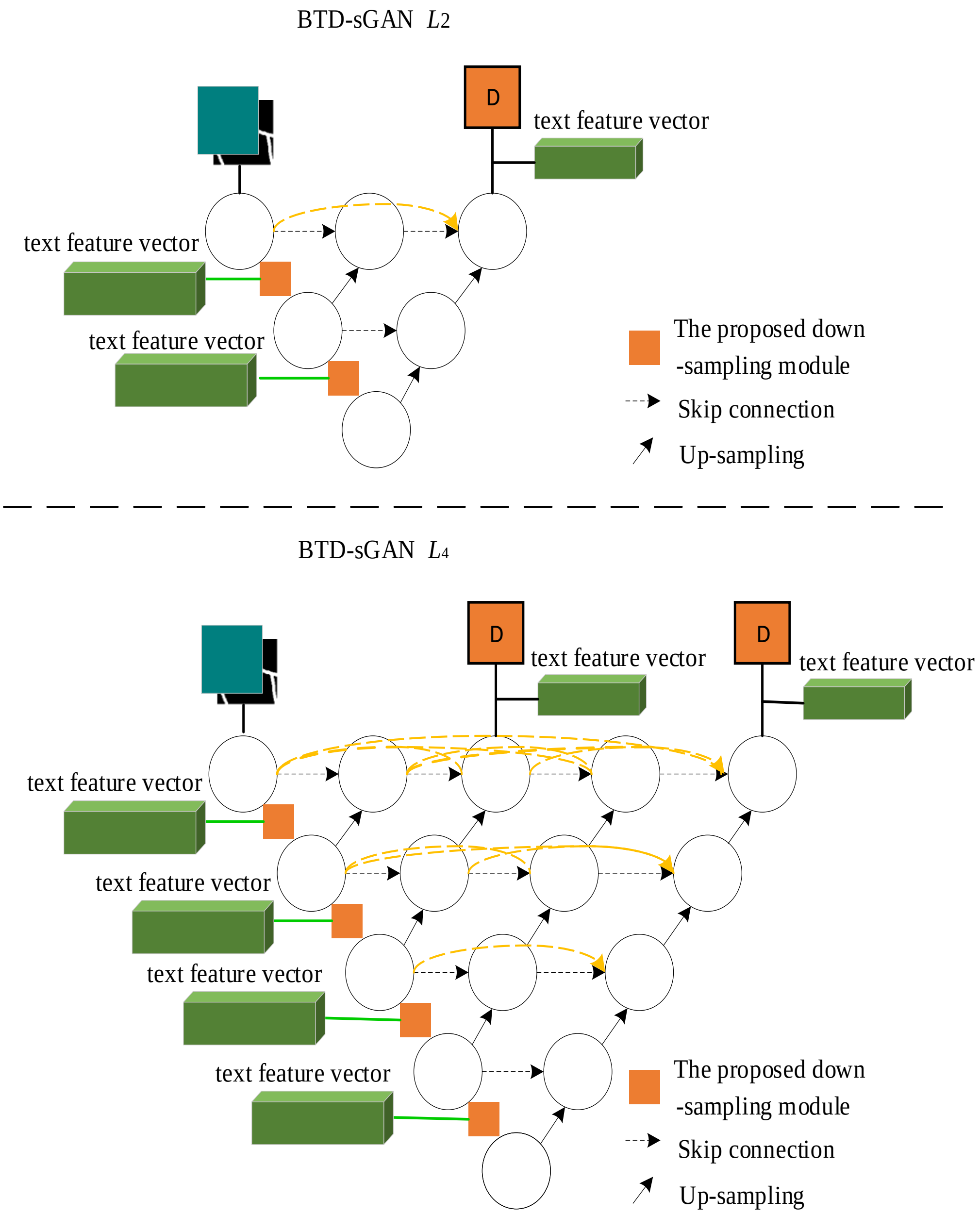

Figure 6. The structure of BTD-sGAN.

2.1.3. Loss Function

The BTD-sGAN loss function consists of two parts, the generator part and the discriminator part, which can be expressed as

$$\begin{aligned}\mathcal{L}_{BTD-sGAN}(G,D) = {} & \mathbb{E}_{x,y}[\log D(v,y)] \\ & + \mathbb{E}_{x,z,v}[\log(1 - D(v, G^*(x,z,v)) - D(v^*,y))].\end{aligned} \tag{1}$$

The matching text feature vector v, true image y, and mismatched text feature vector v^* are represented. The discriminator only detects true when the real image and text match, false when the real image and text do not match, and false when the generated image and text match.

In particular, the discriminator is used to monitor the two-layer and four-layer outputs of Unet++, so it can be expressed as

$$D(v, G^*(x, z, v)) = \sum_{k=1,2} \lambda_k D(v, G(x, z, v)). \tag{2}$$

The generator tries to minimize the loss, and the discriminator tries to maximize the loss. Herein, we used $\lambda_k (k = 1, 2)$ to represent the subnet's weight, and the parameters satisfy the relation $\lambda_1 + \lambda_2 = 1$ and $\lambda_1 < \lambda_2$.

2.2. Datasets

Existing generation models based on text description (such as GAN-INT-CLS, StackGAN++) are mostly studied on the basis of natural images. For fairness, the natural image dataset Oxford-102 [36,37] was used to compare the effects of BTD-sGAN model and other models. At the same time, in order to observe the performance of BTD-sGAN model in the actual remote sensing image generation task, remote sensing datasets from the Jiangxi and Anhui provinces in China were used for training and testing.

2.2.1. Oxford-102 Dataset

Oxford-102 belongs to the natural image dataset, which contains images of flowers, including 102 different flower species and a total of 8189 images. Some images of the Oxford-102 dataset are shown in Figure 7.

Figure 7. The Oxford-102 dataset.

2.2.2. Remote Sensing Datasets of Jiangxi and Anhui Provinces, China

The remote sensing datasets of the Jiangxi and Anhui provinces in China were shot by China Gaofen Satellite with a ground resolution of 2 m and the original remote sensing image resolution of 13,989 × 9359. In this paper, the image was cropped to 128 × 128 size. A partial image of the remote sensing dataset is shown in Figure 8.

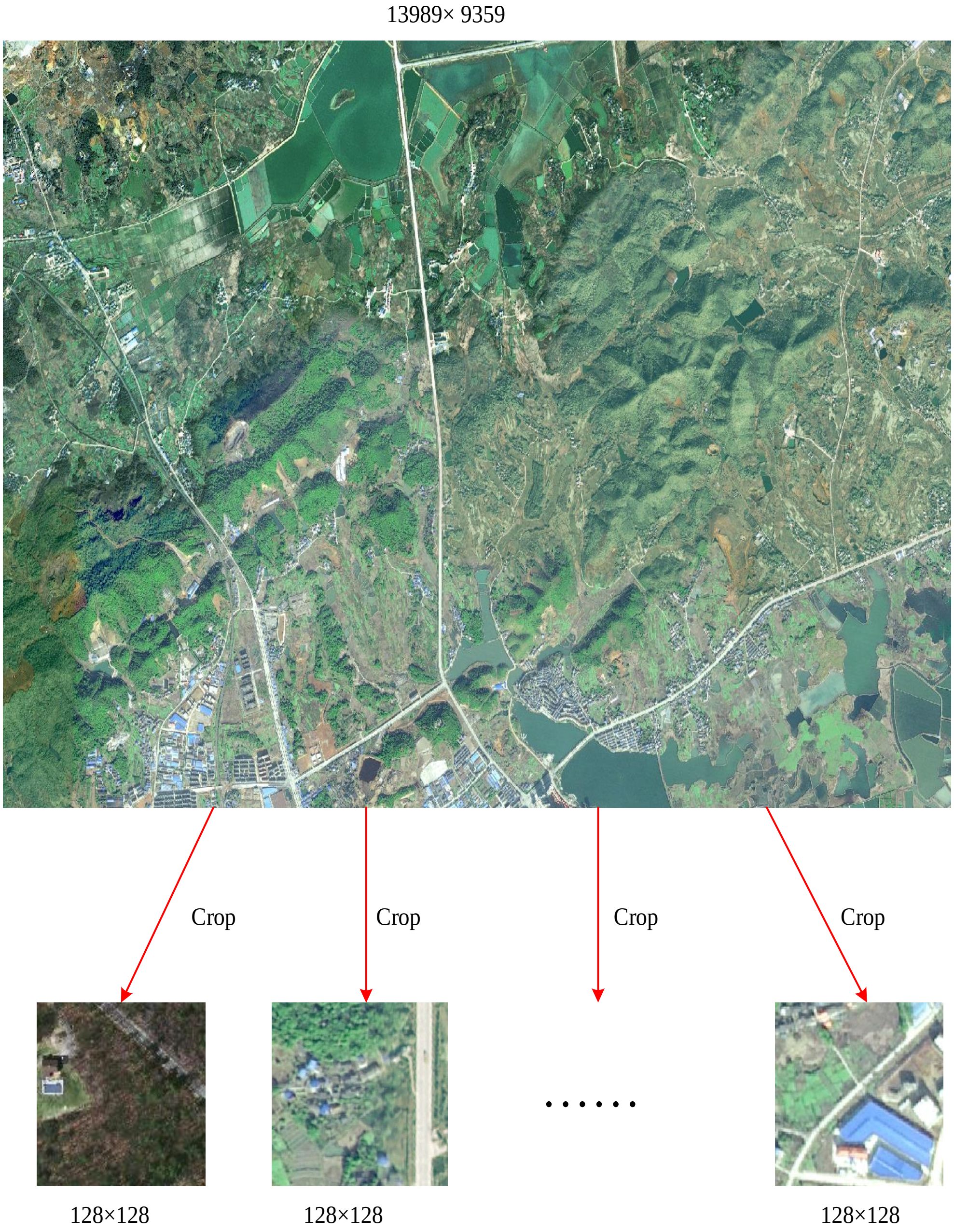

Figure 8. Remote sensing datasets of the Jiangxi and Anhui provinces in China.

2.3. *Evaluation Metrics*

The proposed model (BTD-sGAN) focuses on the diversity of generated images. To evaluate the quality and diversity of the generated images, the recently proposed evaluation metric—Inception Score (abbreviated as IS) [38]—was selected. At the same time, to evaluate whether the generated sample matches the given text description, an artificial evaluation method called "Human Rank" was adopted. For the generation time of BTD-sGAN, the evaluation metric called "Inference Time" was proposed. To evaluate the effect of the proposed model on the actual remote sensing dataset, the generated image was sent into the training set of the remote sensing interpretation model, and the effect of the proposed model was reflected through the interpretation accuracy, which is called "Interpretation Score". These evaluation metrics are detailed as follows.

2.3.1. Inception Score

The IS (Inception Score) evaluation index can comprehensively consider the quality and diversity of the generated images. The evaluation equation can be expressed as

$$Inception\ Score = \exp(\mathbb{E}_x D_{KL}(p(y|x)||p(y))), \tag{3}$$

where x represents the generated image and y represents the prediction label of x for Inception model [39,40]. For a good generation model, it is expected that the model can generate images of high quality and diversity. Therefore, the KL divergence between edge distribution $p(y)$ and conditional distribution $p(y|x)$ should be as large as possible.

2.3.2. Interpretation Score

This index is proposed according to the actual remote sensing interpretation task. It is assumed that there are n remote sensing images in the dataset used by the interpretation model, including kn remote sensing images generated by the generation model. There are $(1-k)n$ remote sensing images in the actual remote sensing dataset (such as remote sensing images of the Jiangxi and Anhui provinces in China), where k is the mixing coefficient and the value range is [0, 1]. Two thirds of this dataset was used as the training set and 1/3 as the test set. Then, remote sensing interpretation models (such as Unet and FCN) were trained and tested on the n remote sensing data images, and the interpretation accuracy of the interpretation model is called "Interpretation Score". Herein, the interpretation types of remote sensing images only include map spots (illegal ground object targets) and nonmap spots (ground object targets other than map spots). If the "overlap ratio" of interpretation results is used to represent interpretation accuracy, the expression of "Interpretation Score" is shown as

$$Interpretation\ Score = \left(\frac{P_{11}}{P_{11}+P_{12}+P_{21}} + \frac{P_{22}}{P_{22}+P_{21}+P_{12}}\right)\Big/2, \tag{4}$$

where P_{11} represents the number of spot pixels interpreted as spot pixels, P_{12} represents the number of spot pixels interpreted as nonspot pixels, P_{21} represents the number of nonspot pixels interpreted as spot pixels, and P_{22} represents the number of nonspot pixels interpreted as nonspot pixels.

2.3.3. Human Rank

IS (Inception Score) cannot reflect the matching degree between the generated image and the text description, so the artificial evaluation method was used. The specific evaluation methods are as follows: 30 text descriptions are randomly selected from the dataset, 3 images are generated for each model, 10 evaluators are selected to rank the results of each model, and the average value of the ranking is taken as the artificial evaluation score of the model. The smaller the ranking is, the better the model effect is. This artificial evaluation method is called "Human Rank". Suppose that the score given by the ith person for the ranking of a model is R_i, then, the score of the model can be expressed as

$$HumanRank = \sum_{i=1}^{10} R_i \Big/ 10, \tag{5}$$

where i represents the serial number of people who rank the model.

2.3.4. Inference Time

"Inference Time" refers to the time of image generation, i.e., the time taken by the generation model to generate multiple remote sensing images. It usually means the time taken to generate mKB remote sensing images, where m represents the amount of memory occupied by the generated image. The unit of "Inference Time" is second.

3. Results

To evaluate the effectiveness of the proposed algorithm in different scenarios, two evaluation experiments are carried out. In the first part, the natural image dataset Oxford-102 (universal dataset) is used as the training set, and the effects of BTD-sGAN, GAN-INT-CLS [19], and StackGAN++ [25] are compared. In the second part, to verify the diversity of the generated remote sensing images, remote sensing images of the Jiangxi and Anhui provinces in China are used as training sets to test the performance of BTD-sGAN on the actual remote sensing datasets. At the same time, BTD-sGAN is compared with GAN-INT-CLS and StackGAN++ in the second experiment.

3.1. Experiment 1

In this experiment, Oxford-102 flower dataset is used, and images in the whole dataset are described manually to form an "image–text description" data pair. Two thirds of the data pairs in the dataset are taken as the training set, and 1/3 of the data pairs are taken as the test set. BTD-sGAN, GAN-INT-CLS, and StackGAN++ are trained and tested. Finally, the generated results of several models are obtained. During training, the three models use the same data pair. The experimental parameters are 50 epochs, each epoch iterates 150 times, and each time 64 samples are trained. During testing, the three models obtain the generated results and evaluation scores according to the same text description. The experimental process of model comparison is shown in Figure 9.

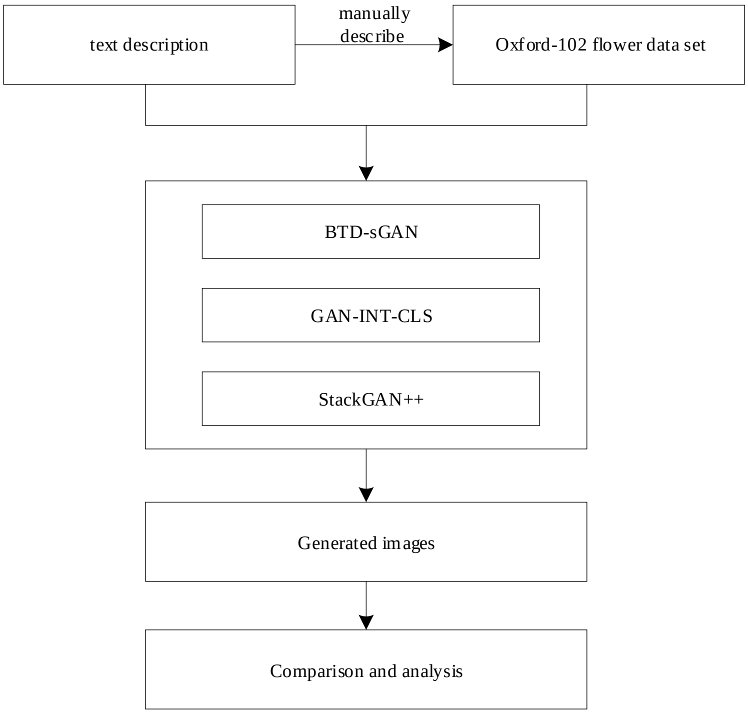

Figure 9. Experimental process of model comparison.

The 3KB text descriptions in the test set are randomly selected for testing. The generated results of the different models are shown in Figure 10.

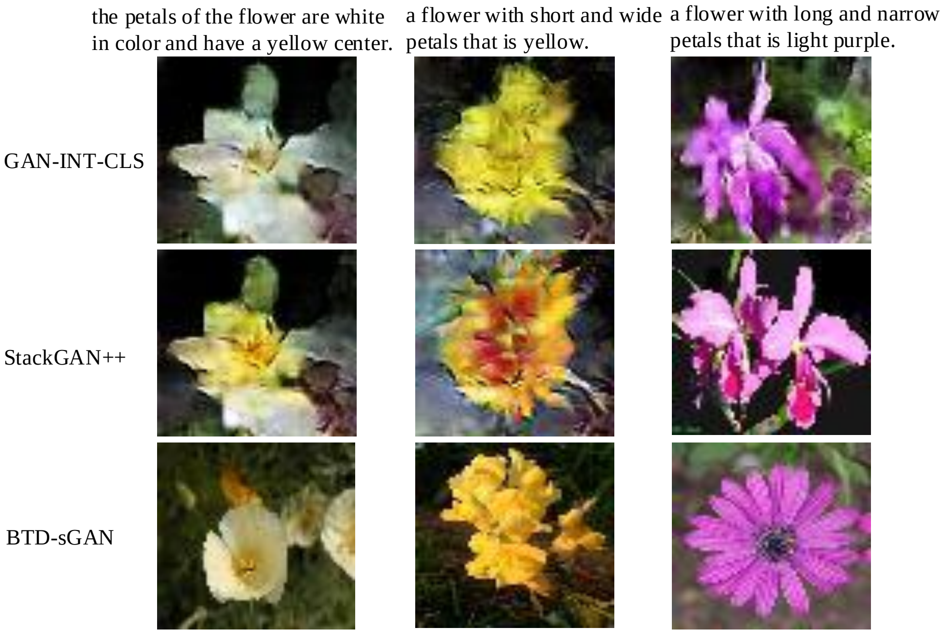
the petals of the flower are white in color and have a yellow center.
a flower with short and wide petals that is yellow.
a flower with long and narrow petals that is light purple.
GAN-INT-CLS
StackGAN++
BTD-sGAN

Figure 10. Three GAN generation results on the Oxford-102 flower dataset.

At the same time, in numerical terms, "Inception Score", "Human Rank", and "Inference Time" are used to compare the effects of different models. The performance comparison of different models is shown in Table 1.

Table 1. Performance comparison of different models on the Oxford-102 dataset.

Model	Inception Score	Human Rank	Inference Time (s)
GAN-INT-CLS	2.56 ± 0.03	1.98 ± 0.04	54
StackGAN++	3.52 ± 0.02	1.75 ± 0.03	62
BTD-sGAN	3.66 ± 0.03	1.18 ± 0.02	40

To more intuitively show the generation performance differences of different models, the scores are also shown in Figure 11.

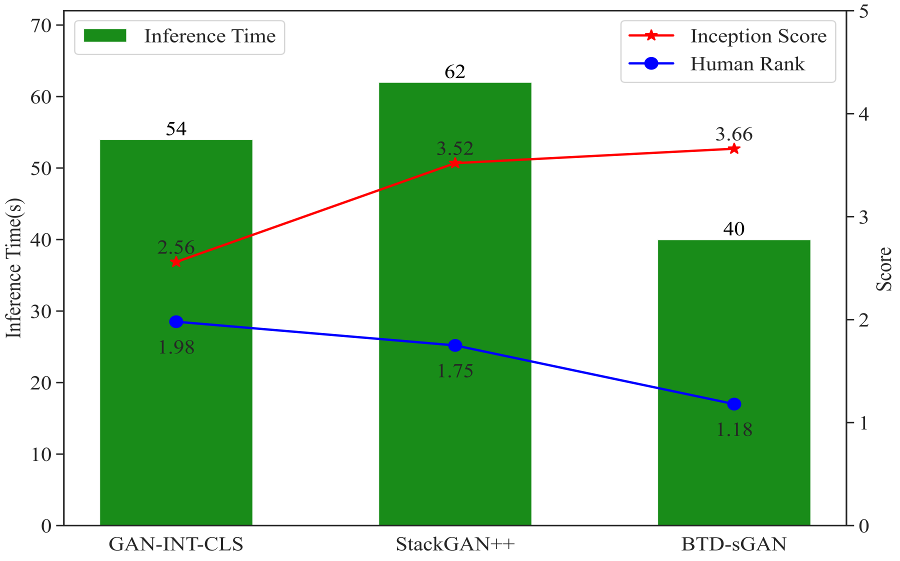

Figure 11. The performance comparison of different models.

The results in Table 1 show that the BTD-sGAN is higher in “Human Rank” than GAN-INT-CLS and StackGAN++, increasing by 0.8 (from 1.98 to 1.18) and 0.57 (from 1.75 to 1.18), respectively. Compared with GAN-INT-CLS and StackGAN++, BTD-sGAN has an increase of 1.10 (from 2.56 to 3.66) and 0.14 (from 3.52 to 3.66) in “Inception Score”, respectively. In addition, 14 s (from 54 s to 40 s) and 22 s (from 62 s to 40 s) are reduced in the “Inference Time”, respectively. Figure 11 more intuitively shows that BTD-sGAN has a shorter generation time, smaller ranking score, and larger IS score compared to the other two models.

3.2. Experiment 2

The ultimate purpose of constructing BTD-sGAN is to enhance the data of remote sensing images and serve those researches based on remote sensing images, such as remote sensing interpretation tasks. Therefore, remote sensing images from the Jiangxi and Anhui provinces of China are used as datasets to train and test the effect of BTD-sGAN. In particular, the image is a multispectral remote sensing image. The experiment only uses the data of RGB channels, and the final image results from the fusion of RGB channels.

Similar to Experiment 1, the remote sensing dataset is described manually to form a “remote sensing image–text description” data pair. Among them, 2/3 of the data pairs are used as the training set and 1/3 of the data pairs are used as the test set. Text description is randomly selected for testing, and the model generates 3 remote sensing images according to each text description. The generation effect of BTD-sGAN on the actual remote sensing dataset is shown in Figure 12.

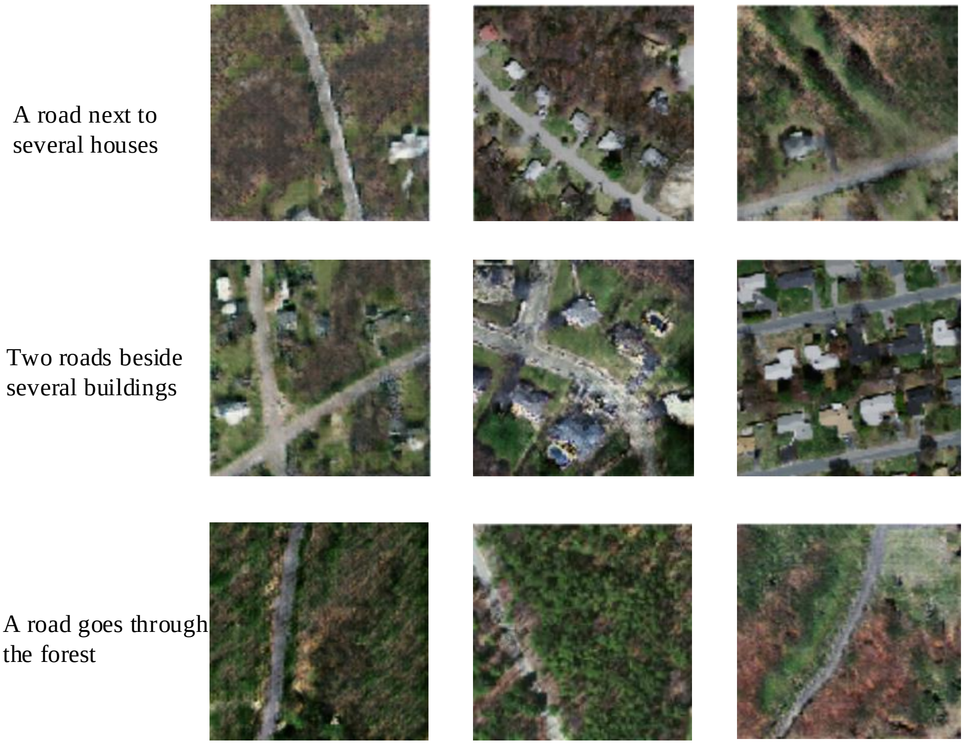
A road next to several houses
Two roads beside several buildings
A road goes through the forest

Figure 12. Generation results of BTD-sGAN on China remote sensing datasets.

As can be seen from Figure 12, BTD-sGAN can generate various remote sensing images according to text description. Taking the text description, “a road next to several houses”, as an example, BTD-sGAN generates three different shapes of roads according to this description that all meet the requirements of this text description. The above results show that BTD-sGAN can generate diverse images and meet the needs of image diversity in the remote sensing image generation task.

On the basis of China remote sensing datasets, the generation results of BTD-sGAN, GAN-INT-CLS, and StackGAN++ are also compared. The results are shown in Figure 13.

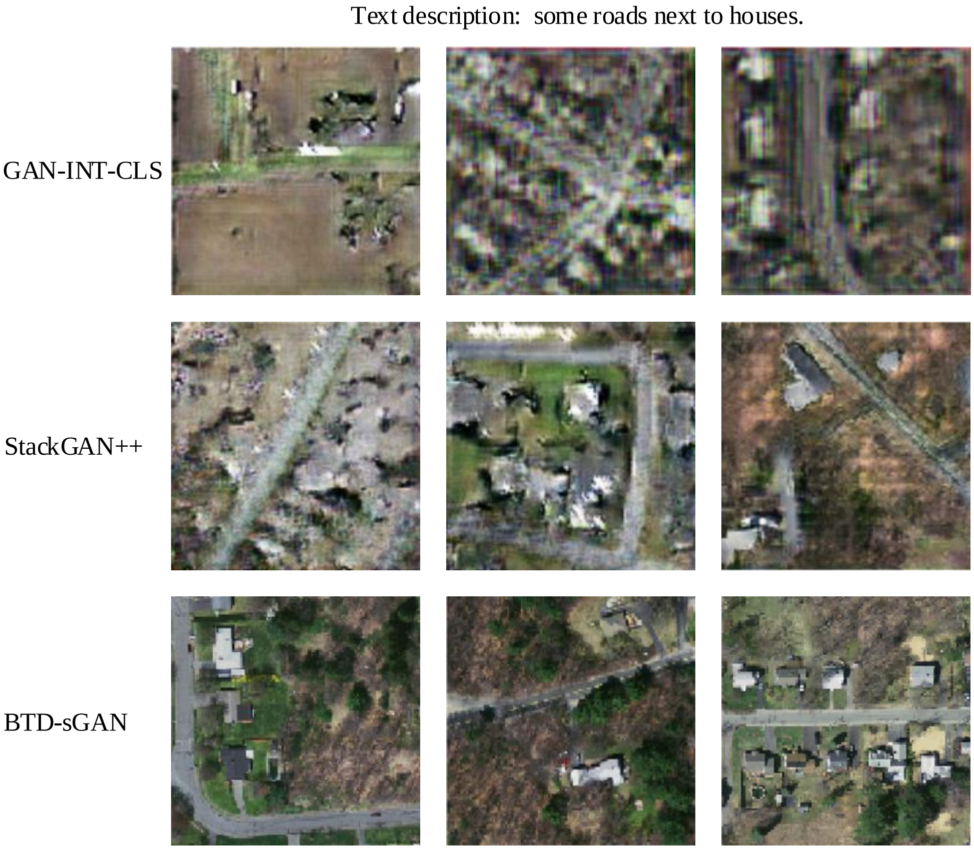

Figure 13. The results of different models on actual remote sensing dataset.

In Figure 13, compared with GAN-INT-CLS and StackGAN++, the remote sensing image generated by BTD-sGAN is clearer and matches the text description.

Furthermore, the performance of BTD-sGAN is evaluated numerically. A new method is used to evaluate BTD-sGAN, namely, "Interpretation Score". The idea of this method is as follows: The generated data is sent to the remote sensing interpretation network to see if the generated image is helpful to improve the accuracy (equivalent to the "Interpretation Score") of the interpretation network. The higher the value of "Interpretation Score", the better the effect of model generation. A flow chart of the experiment is shown in Figure 14.

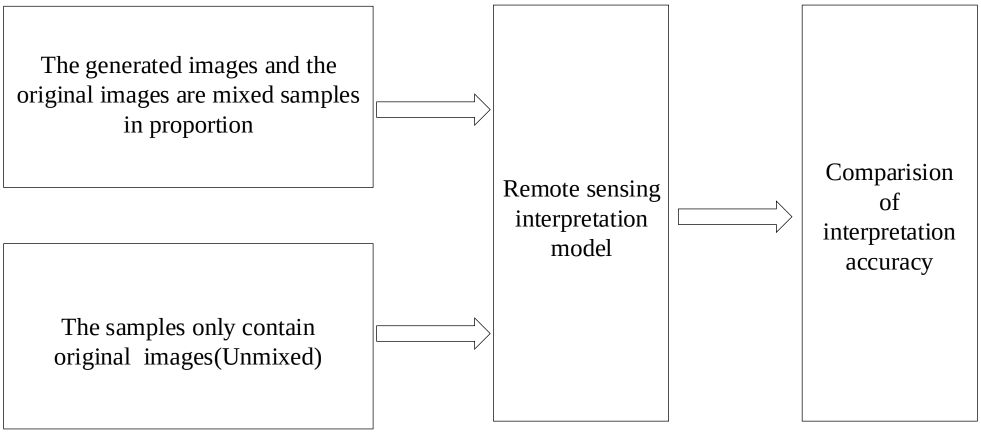
The generated images and the original images are mixed samples in proportion
The samples only contain original images(Unmixed)
Remote sensing interpretation model
Comparision of interpretation accuracy

Figure 14. Flow chart of the experiment.

After mixing different proportions of the generated images in the dataset, the change of “Interpretation Score” with mixed proportions is shown in Figure 15.

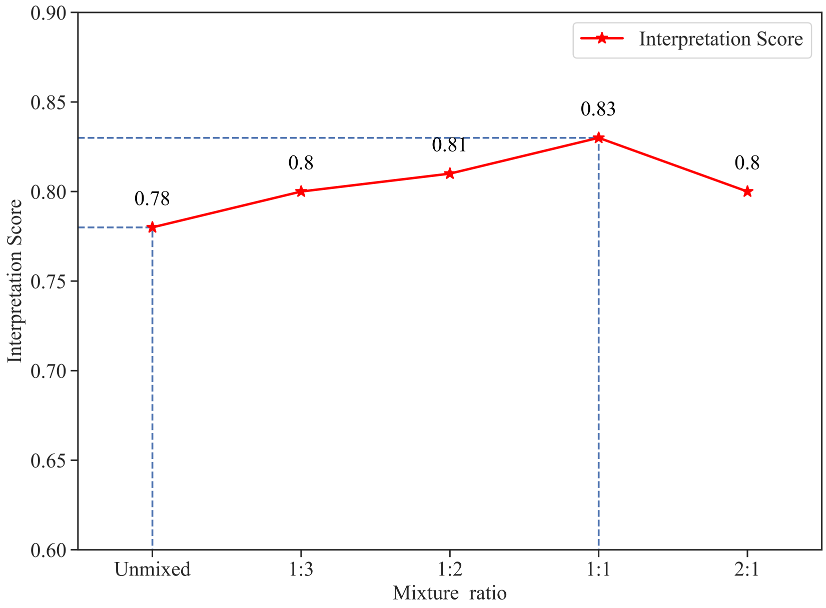

Figure 15. Diagram of the change of “Interpretation Score” with mixture ratio.

In Figure 15, mixture ratio represents the ratio of the generated images to the original images in the dataset.

4. Discussion

In the results section, two experiments were used to verify the effectiveness of BTD-sGAN. Experiment 1 is based on the universal dataset (Oxford-102 flower dataset), which can ensure the fairness of all models in the comparative experiment. For the dataset, Figure 10 shows the generated results of different models. Two conclusions can be drawn from the generated results. 1) BTD-sGAN can generate images according to text description, which proves the rationality of the model. 2) Visually, compared with GAN-INT-CLS and StackGAN++, the generation results of BTD-sGAN are clearer and of better quality; the performance of BTD-sGAN was evaluated quantitatively. Table 1 and Figure 11 show that BTD-sGAN is superior to other models in the three indexes of “Inception Score”, “Human Rank”, and “Inference Time”, which indicates that BTD-sGAN can generate clearer and more diverse images according to text description, and shorten the time of image generation to meet the needs of the actual generation task.

Experiment 2 is based on remote sensing datasets of the Jiangxi and Anhui provinces, China. This experiment is used to test BTD-sGAN’s performance on the actual remote sensing dataset. First, whether BTD-sGAN can generate a variety of remote sensing images according to the text description is tested. Figure 12 shows that BTD-sGAN can generate various remote sensing images, which proves that BTD-sGAN can be used in the actual remote sensing generation task. Then, the performance of different generation models is compared. In Figure 13, BTD-sGAN generates clearer images than others. The previous part evaluates BTD-sGAN according to vision. Numerically, the metrics “Interpretation Score” is used. Figure 15 shows that the scores of remote sensing interpretation after mixing can be improved compared with that of unmixed samples, and when the mixing ratio is 1:1, the precision can be improved by 5%. This is because the diversity of the generated samples is higher than that of the original images and the generalization ability of the network is improved. However, when the mixture ratio reaches 2:1, the interpretation accuracy will decrease. Due to the large proportion of generated samples, the network learns the features of the generated samples and the insufficient learning of the features of the original remote sensing images.

5. Conclusions

Aiming at the lack of samples in the deep learning-based remote sensing image detection project, a new text-based generative adversarial network called BTD-sGAN is proposed for the data augmentation of remote sensing image. Two experiments were used to verify the effect of BTD-sGAN. The first experiment was used to test the performance of BTD-sGAN on the universal dataset, and the second experiment was used to test the performance of BTD-sGAN on the actual remote sensing dataset. In Experiment 1, BTD-sGAN generated higher quality images than other models. Compared with GAN-INT-CLS and StackGAN++, BTD-sGAN increased by 1.10 and 0.14 in "Inception Score" and 0.8 and 0.57 in "Human Rank", and decreased by 14 s and 22 s in "Inference Time", respectively. In Experiment 2, BTD-sGAN produced clearer and more varied remote sensing images than GAN-INT-CLS and StackGAN++. The results show that the remote sensing image generated by BTD-sGAN can help improve the accuracy of remote sensing interpretation network by 5%. In general, BTD-sGAN can be applied to the actual remote sensing generation tasks, and can also provide the data support for remote sensing interpretation (e.g., soil-moving detection) and other tasks.

However, BTD-sGAN still has some limitations. The text vectors are used to correct text semantic loss during downsampling, which leads to image semantic loss to a certain extent. In other words, the quality of the generated image is sacrificed. In contrast, the diversity of the generated image is gained. The results presented herein were limited to only RGB bands. The effectiveness of the method for other spectral bands, such as Near-Infrared and Red Edge that are used for various purposes, requires further investigation and is subject to future work. In addition, there are many related types of research in the field of remote sensing based on deep learning, and the demand will be different. The future direction is to improve the model to meet the need of remote sensing generation. This paper will also try to apply the model to some other fields (such as Internet of Vehicles [41]) for data augmentation, so as to further test the practical applicability of the model.

Author Contributions: All authors contributed to the study conception and design. Conceptualization, C.C. and N.L.; methodology, H.M., H.Y., G.Y., S.W.; software, C.C., H.M., C.L., S.W.; validation, N.L., G.Y.; formal analysis, S.W.; investigation, C.C.; resources,C.C., H.M.; data curation, C.C., S.W.; writing—original draft preparation, C.C.; writing—review and editing, H.M., N.L., S.W., G.Y.; visualization, C.L., H.Y.; supervision, C.C., S.W.; project administration, C.C.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2020YFB1807500), the National Natural Science Foundation of China (62072360, 62001357, 61672131, 61901367), the key research and development plan of Shaanxi province (2021ZDLGY02-09, 2020JQ-844), the fundamental research funds for the central universities (31412111303, 31512111310), the key laboratory of embedded system and service computing (Tongji University) (ESSC KF2019-05), Ministry of Education, Xi'an Science and Technology Plan (20RGZN0005) and the Xi'an Key Laboratory of Mobile Edge Computing and Security (201805052-ZD3CG36).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from the local water utilities and are available from the authors with the permission of the City.

Acknowledgments: The City provided GIS and remote sensing data that allowed us to complete the data augmentation and our analyses. Yang Zhou helped solidify the data agreement between the City and the University. Ji Li provided feedback on methodology and data analysis. We also acknowledge the use of imagery from the Ministry of Water Resources of China. Finally, we'd like to thank Fenglin Xiao for his assistance in data processing and visualization.

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

Abbreviations

The following abbreviations are used in this manuscript:

VAE	Variational Auto Encoder
GAN	Generative Adversarial Network
cGAN	conditional Generative Adversarial Network
DGN	Deep Generator Network
AM	Activation Maximization
PPGN	Plug & Play Generative Network
ArtGAN	Artwork Synthesis with Conditional Categorical GAN
StackGAN	Stacked Generative Adversarial Network
ATTNGAN	Attentional Generative Adversarial Network
IS	Inception Score

References

1. Lv, N.; Chen, C.; Qiu, T.; Sangaiah, A.K. Deep Learning and Superpixel Feature Extraction Based on Contractive Autoencoder for Change Detection in SAR Images. *IEEE Trans. Ind. Inform.* **2018**, *14*, 5530–5538. [CrossRef]
2. Ghaderpour, E.; Vujadinovic, T. Change Detection within Remotely Sensed Satellite Image Time Series via Spectral Analysis. *Remote Sens.* **2020**, *12*, 4001. [CrossRef]
3. Srivastava, G.; Kumar, C.N.S.V.; Kavitha, V.; Parthiban, N.; Venkataraman, R. Two-stage data encryption using chaotic neural networks. *J. Intell. Fuzzy Syst.* **2020**, *38*, 2561–2568. [CrossRef]
4. Shivani, S.; Patel, S.C.; Arora, V.; Sharma, B.; Jolfaei, A.; Srivastava, G. Real-time cheating immune secret sharing for remote sensing images. *J. Real-Time Image Process.* **2020**. [CrossRef]
5. Zhu, Z.; Woodcock, C.E.; Holden, C.; Yang, Z. Generating synthetic Landsat images based on all available Landsat data: Predicting Landsat surface reflectance at any given time. *Remote Sens. Environ.* **2015**, *162*, 67–83. [CrossRef]
6. Yan, Y.; Tan, Z.; Su, N. A data augmentation strategy based on simulated samples for ship detection in rgb remote sensing images. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 276. [CrossRef]
7. Chen, C.; Liu, B.; Wan, S.; Qiao, P.; Pei, Q. An Edge Traffic Flow Detection Scheme Based on Deep Learning in an Intelligent Transportation System. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1840–1852. [CrossRef]
8. Jiang, J.; Ma, J.; Chen, C.; Wang, Z.; Cai, Z.; Wang, L. SuperPCA: A Superpixelwise PCA Approach for Unsupervised Feature Extraction of Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4581–4593. [CrossRef]
9. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the ICLR 2014: International Conference on Learning Representations (ICLR) 2014, Banff, AB, Canada, 14–16 April 2014.
10. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27; pp. 2672–2680.
11. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
12. Sangkloy, P.; Lu, J.; Fang, C.; Yu, F.; Hays, J. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6836–6845.
13. Gauthier, J. Conditional generative adversarial nets for convolutional face generation. In Proceedings of the Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter Semester, Toulon, France, 24–26 April 2017; Volume 2014, p. 2.
14. Denton, E.; Chintala, S.; Szlam, A.; Fergus, R. Deep generative image models using a Laplacian pyramid of adversarial networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28, pp. 1486–1494.
15. Sønderby, C.K.; Caballero, J.; Theis, L.; Shi, W.; Huszár, F. Amortised MAP Inference for Image Super-resolution. In Proceedings of the International Conference on Learning Representations (ICLR) 2016, San Juan, Puerto Rico, 2–4 May 2016.
16. Nguyen, A.M.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, 9 December 2016; Volume 29, pp. 3387–3395.
17. Nguyen, A.; Clune, J.; Bengio, Y.; Dosovitskiy, A.; Yosinski, J. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3510–3520.
18. Tan, W.R.; Chan, C.S.; Aguirre, H.E.; Tanaka, K. ArtGAN: Artwork synthesis with conditional categorical GANs. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3760–3764.

19. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning—Volume 48, New York, NY, USA, 20–22 June 2016; pp. 1060–1069.
20. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
21. Ding, S.; Qu, S.; Xi, Y.; Wan, S. Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing* **2020**, *398*, 520–530. [CrossRef]
22. Reed, S.; Akata, Z.; Mohan, S.; Tenka, S.; Schiele, B.; Lee, H. Learning what and where to draw. In Proceedings of the NIPS'16 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Voume 29, pp. 217–225.
23. Zhang, H.; Xu, T.; Li, H. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5908–5916.
24. Zhao, Y.; Li, H.; Wan, S.; Sekuboyina, A.; Hu, X.; Tetteh, G.; Piraud, M.; Menze, B. Knowledge-aided convolutional neural network for small organ segmentation. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 1363–1373. [CrossRef] [PubMed]
25. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1947–1962. [CrossRef] [PubMed]
26. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1316–1324.
27. Wan, S.; Xia, Y.; Qi, L.; Yang, Y.H.; Atiquzzaman, M. Automated colorization of a grayscale image with seed points propagation. *IEEE Trans. Multimed.* **2020**, *22*, 1756–1768. [CrossRef]
28. Hong, S.; Yang, D.; Choi, J.; Lee, H. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7986–7994.
29. Gao, Z.; Li, Y.; Wan, S. Exploring deep learning for view-based 3D model retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–21. [CrossRef]
30. Huang, F.; Jolfaei, A.; Bashir, A.K. Robust Multimodal Representation Learning with Evolutionary Adversarial Attention Networks. *IEEE Trans. Evol. Comput.* **2021**. [CrossRef]
31. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic Image Synthesis With Spatially-Adaptive Normalization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
32. Tang, B.; Tu, Y.; Zhang, Z.; Lin, Y. Digital Signal Modulation Classification With Data Augmentation Using Generative Adversarial Nets in Cognitive Radio Networks. *IEEE Access* **2018**, *6*, 15713–15722. [CrossRef]
33. Yang, J.; Zhao, Z.; Zhang, H.; Shi, Y. Data Augmentation for X-Ray Prohibited Item Images Using Generative Adversarial Networks. *IEEE Access* **2019**, *7*, 28894–28902. [CrossRef]
34. Lv, N.; Ma, H.; Chen, C.; Pei, Q.; Zhou, Y.; Xiao, F.; Li, J. Remote Sensing Data Augmentation Through Adversarial Training. In Proceedings of the IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2511–2514.
35. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef] [PubMed]
36. Nilsback, M.E.; Zisserman, A. Automated Flower Classification over a Large Number of Classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
37. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 806–813.
38. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. In Proceedings of the NIPS'16 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 2234–2242.
39. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, 27–30 June 2016; pp. 2818–2826.
40. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
41. Chen, C.; Wang, C.; Qiu, T.; Atiquzzaman, M.; Wu, D.O. Caching in Vehicular Named Data Networking: Architecture, Schemes and Future Directions. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2378–2407. [CrossRef]