IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 6, 2020, accepted April 18, 2020, date of publication April 30, 2020, date of current version May 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2989458

Target Detection Based on Simulated Image Domain Migration

YAOLING WANG[1,2,3], JUN GU[1,2,3], LIANGJIN ZHAO[1,3], YUE ZHANG[1,3], (Member, IEEE), AND HONGQI WANG[1,3]

[1]Aerospace Information Institute, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Network Information System technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Yue Zhang (zhangyue@aircas.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant 41801349.

ABSTRACT Annotating a large amount of data manually for supervised learning is an indispensable and expensive part. A novel system using the simulation dataset is proposed in this paper. This framework can train the neural networks for remote sensing object detection without any manually labeled dataset. The whole system can be divided into three parts. The first part is the dataset simulator. The simulator synthesizes remote sensing images with the aircraft targets based on real remote sensing images (without any aircraft targets). In the process of data generation, the simulator automatically marks the position information of the aircraft. The second part is the image dataset domain adaptation work. We introduce the work of Cycle-GAN into this part to bridge the perceptual gap between the simulation dataset and reality dataset. Specially, we propose a multi-scale generator into the original Cycle-GAN model to achieve better domain adaptation performance. The final part is the object detection neural network. The domain adaptation quality of the remote sensing images reconstructed by our novel cycle-gan network achieves better performance both in the structural similarity measure and visual appearance. The object detection model trained with the dataset processed by our novel system can get better detection precision. The analytic experiments on the test dataset demonstrate that the object detection model trained with the dataset processed by our novel system can get better detection precision.

INDEX TERMS Remote sensing image, generative and adversarial network, object detection, data simulation.

I. INTRODUCTION

Neural network training and testing rely on a large amount of annotated data for supervised learning. But manually annotating datasets is a very time consuming and laborious task. Especially when tasks require a lot of expertise or manual tagging is too difficult, we won't be able to get a lot of data. A small amount of data does not well drive the training of the entire neural network, so the performance of the neural network will not achieve the desired goal. For example, in the field of remote sensing image object detection, it takes a lot of time to filter a remote sensing image containing a target and mark the target in the image.

A promising way to overcome data limitations is to use a data simulation platform to generate auto-annotating data in recent years, several such annotated dataset has been created for geometric problems such as optical flow, scene flow, stereo disparity estimation, and camera pose estimation. Handa *et al.* [1] focus their attention on depth-based semantic per-pixel labeling as a scene understanding problem and show the potential of computer graphics to generate virtually unlimited labeled data from synthetic 3D scenes. Butler *et al.* [2] first introduce a new optical flow dataset derived from the open-source 3D animated short film Sintel. Later, Dosovitskiy *et al.* [3] generated a large synthetic flight chair dataset to solve the problem of using deep learning to study the lack of datasets for optical flow estimation. Mayer *et al.* [4] propose three synthetic stereo video datasets with sufficient realism, variation, and size to successfully train large networks for disparity, optical flow, and scene flow estimation.

These methods are capable of generating high-quality simulation data. However, the synthesis of the real quality of photographs in the above methods requires the researchers

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

to model the specific environment and application in detail. So the cost of data simulation is very large. This is in contrast to the use of a simulation dataset to reduce the cost and marking difficulty of marking neural network training datasets. At the same time, the simulated composite image is somewhat different from the actual image. At present, many methods do not pay attention to the differences between these different domain images. These methods train the neural network directly with the simulation dataset without thinking that the trained networks may not suitable for the actual dataset and may not achieve a promising result. Howe *et al.* [5] proposed a novel data augmentation strategy based on simulated samples object detection in remote sensing images. These methods consider the problem of how to match the target object with the background image in terms of size, tilt angle, and image resolution. All these strategies alleviate the problem of distortion of analog image synthesis results to some extent and achieve a better result. However, from the results of the simulation dataset, it can be found that the texture of the actual background remote sensing image is different from that of the surface texture of the simulation aircraft. From the perspective of human vision, simulation data and real image data can be clearly identified. If the image distribution difference between the simulated image and the real image can be further processed, the simulated image can better fit the distribution of the actual image data, and the network's adaptability to the actual data set can be improved to a certain extent.

With the application of deep convolution neural networks (CNNs) [6] in object detection, more and more efficient detection algorithms have been proposed, such as region proposals with convolution neural networks (RCNN) [7], Spatial Pyramid Pooling Network (SPP-Net) [8], and Fast-RCNN [9]. Faster-RCNN [10] proposes a Region Proposal Network (RPN) structure and improves the detection efficiency while achieving end-to-end training. Instead of relying on regional proposals, You Only Look Once (YOLO) [11] and Single Shot MultiBox Detector (SSD) [12] directly estimate the object region and truly enable real-time detection. Feature Pyramid Network (FPN) [13] adopts the multi-scale feature pyramid form and makes full use of the feature map to achieve better detection results. Region-based Fully Convolutional Networks (R-FCN) [14] builds a fully convolution network, which greatly reduces the number of parameters, improves the detection speed, and has a good detection effect. [15]–[17] propose a set of remote sensing images object detection methods based on deep neural networks. These methods are constantly improving the detection accuracy of the target detection or the detection speed. However, in many practical situations, the performance of the target detection network is sufficient for everyday applications. Researchers rarely pay attention to the importance of labeling data.

In this paper, we design a method of remote sensing image object detection training with simulation data. First, we propose a remote sensing dataset simulator-based on the engine of Unity 3D. Our simulator synthesizes the dataset based on the given target aircraft models and actual remote sensing images. At the same time, the position annotation of the object targets will be generated automatically. Our simulation system doesn't need to model specific scenes, nor do we need to pay attention to too much detail and have fewer restrictions. Then, we innovatively use the domain migration idea to shorten the distribution gap between the simulated image domain and the actual image domain. The simulation image composited by our simulator can better fit the real remote sensing image data with the domain adaptation process. CycleGAN [18] uses an unsupervised way to achieve data conversion tasks between different image style domains and achieves good results. Based on CycleGAN, we have innovatively proposed a multi-scale generator network. Compared with the original generator in CycleGAN, the multi-scale generator network can better capture the data distribution of the target domain image, so as to achieve better domain conversion effect. Finally, we use the simulated image data and automatically generated tag data after the domain migration to directly train the target detection network, and test the trained network directly on the real remote sensing image data. In summary, the main contributions of this work are as follows:

1) We introduce the self-adaption step between the environment and the aircraft objects by Cycle-GAN framework into the generation of the remote sensing image simulation.
2) We propose a multi-scale generator, which can better adapt to remote sensing images with different resolutions, and thus obtain better results.
3) We propose a data generation framework with better fusion between the environment and the background of the target object.

The remainder of this paper is organized as follows. Section II introduces the details of our proposed method. Section III verifies the effectiveness of our framework by performing comparisons with the state-of-the-art methods. In Section IV, we discuss the issues of our method according to the experimental results. Section V concludes the discussions of the study.

II. PROPOSED METHOD

A. FRAMEWORK

The entire framework, as illustrated in Figure 2, can be divided into training and testing progress. The training progress includes a dataset simulation part as shown in Figure 1, a domain adaptation part as illustrated in Figure 3, and a target detection network training part. We use aircraft models and common remote sensing images to obtain raw simulation data. In the process of generating the simulation dataset, the coordinates of the aircraft and other related target detection parameter files are automatically generated by our simulator program. There is a certain difference between the simulated images and the real remote sensing images

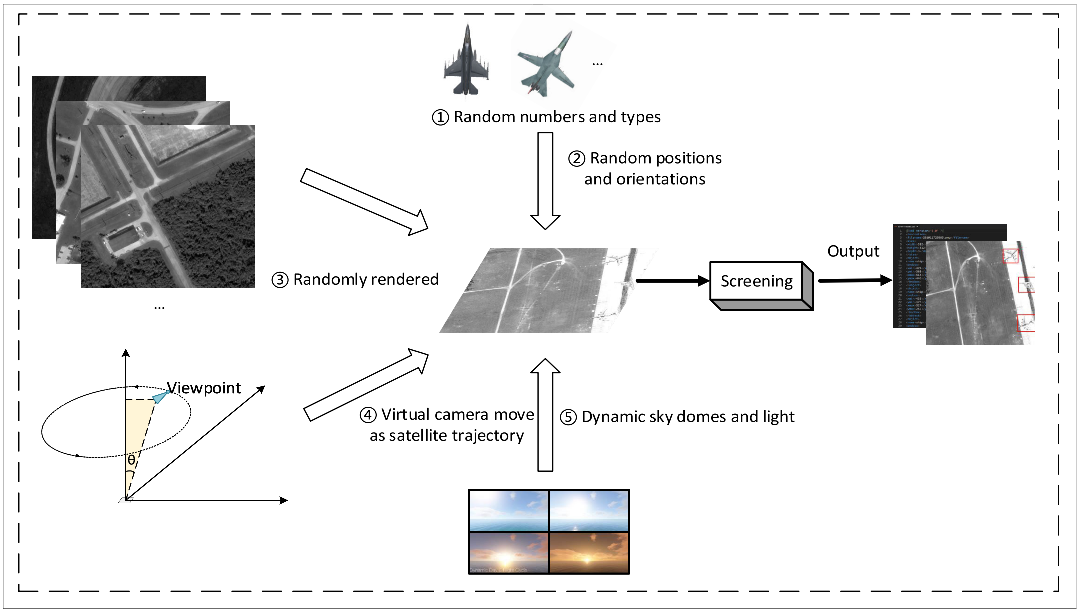

...
① Random numbers and types
② Random positions and orientations
③ Randomly rendered
...
Screening
Output
Viewpoint
θ
④ Virtual camera move as satellite trajectory
⑤ Dynamic sky domes and light

FIGURE 1. The framework of the dataset simulator.

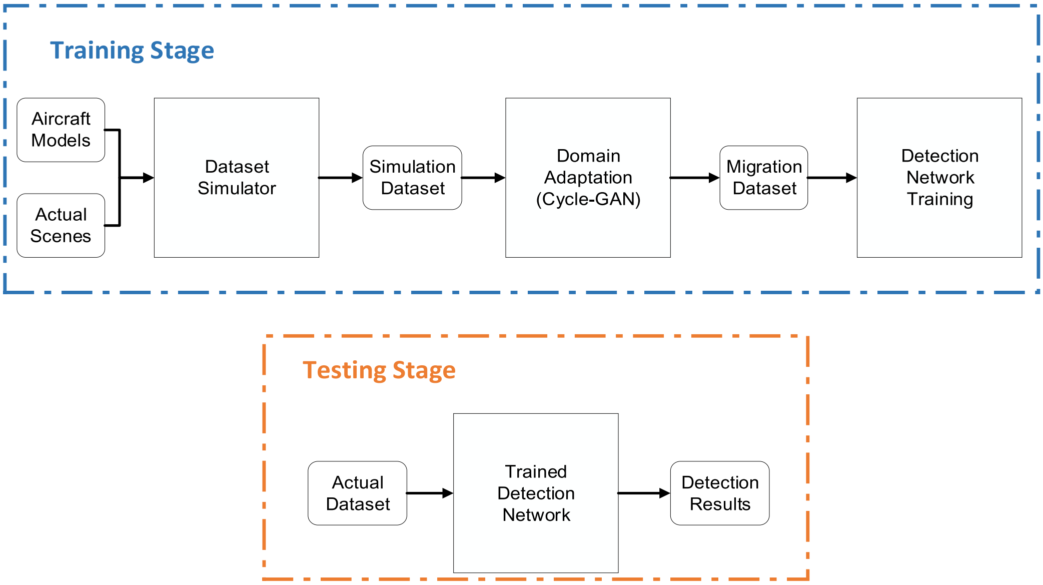


FIGURE 2. The entire framework of our system.

when comparing the scenes and the aircraft targets. At the same time, the detection network directly trained with the simulated images can not achieve a good detection precision on the actual remote sensing image dataset. So we propose a domain adaptation part to improve the problem. The domain adaptation part makes the image generated by the simulation more consistent with the image data distribution of the real scene, and can better improve the accuracy of the aircraft target detection in the real remote sensing scene. Due to the lack of paired simulation images and real images, we built our domain adaptation module using the unsupervised Cycle-GAN model. Finally, we use the transformed analog image as the training set for the aircraft target detection network. The training process for the entire framework does not rely on any manually tagged data.

In the testing phase, we use the object detection model trained with the remote sensing images processed by our simulator and domain adaptation part to directly test the data of the real remote sensing image dataset.

B. DATASET SIMULATION

We build a remote-sensing image simulation platform by the Unity3D [19] game engine. The whole architecture is illustrated in Figure 1. A random number of aircraft are placed in a 3D scene at random positions and orientations. The scene is rendered from a changing viewpoint from −20 degrees

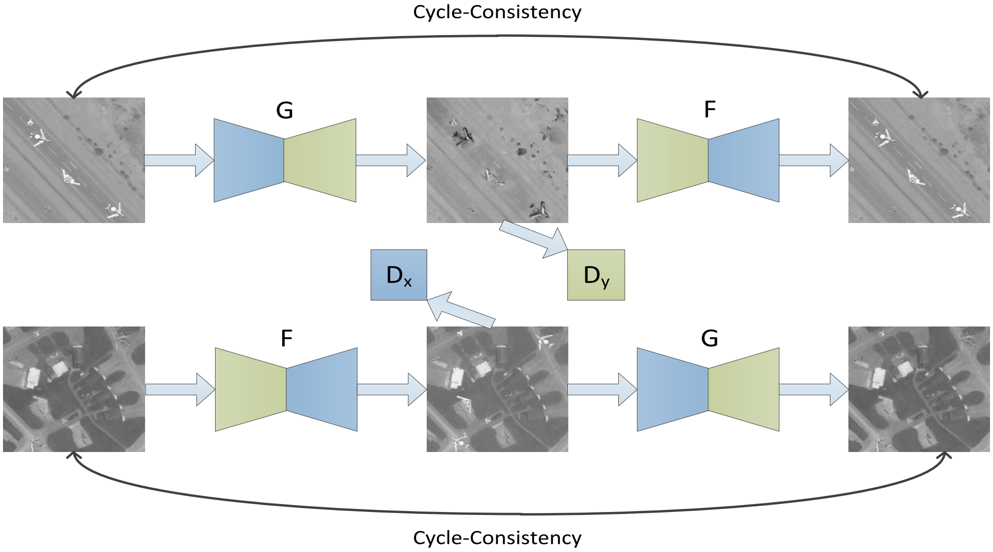


FIGURE 3. The framework of CycleGAN.

to 20 degrees, after which the result is composed over a random background remote-sensing image. It is worth noting that the poor-quality images are removed from results by sum modulus difference (SMD). The resulting images, with automatically generated xml files with ground truth labels, are used for training the neural network. More specifically, images were generated by randomly varying the following aspects of the scene:

- Number and types of aircraft. Design a set of aircraft 3D models, including fighter-planes, bombers, early warning aircraft and transport aircraft.
- Location and direction of aircraft. Train the SVM classifier to find a location that is suitable for parking the aircraft.
- Terrain texture rendering. 44k remote-sensing images are randomly rendered by terrain engine and baked the illumination using light-mapper.
- Location of the virtual camera with respect to the scene. Set the height and change the position of the virtual camera based on the satellite trajectory.
- Dynamic sky domes with day and night cycle, dynamic clouds and physically based atmospheric scattering. Our pipeline uses an internally created plug-in to the Unity3D that is capable of outputting 512 pixels × 512 pixels images with annotations at 10 Hz.

C. DOMAIN ADAPTATION

1) CycleGAN FRAMEWORK

Our domains adaption framework is an enhanced version of the CycleGAN architecture. One domain is the distribution of our simulation dataset and the other is the real remote sensing images. As shown in Figure 3, the whole network architecture is composed of two generators and two discriminators. One of the generators is forward mapping G, which converts the image of the x domain into an image that matches the probability distribution of the y domain. Another generator F is the inverse of generator G. The two discriminators are D_x and D_y. The D_x aims to distinguish the between images x and translated image $F(y)$. In the same way, D_y aims to discriminate between y and $G(x)$.

2) MULTI-SCALE GENERATOR

The generator is decomposed into two sub-networks as shown in Figure 4: G_{global} and G_{local}. G_{global} is designed as a global generator operator at a low resolution and G_{local} outputs an image with a high resolution. The G_{global} is proposed as the main part to generator the basic structure of the images and the G_{local} can be regarded as a local enhancer network to capture fine structure of the inputs.

The CNN architecture of G_{global} is similar to the one proposed by Johnson *et al.* [20], which has been proven successful for style transfer task. It contains two stride convolution blocks with stride $\frac{1}{2}$, nine residual blocks and two transposed convolution blocks. Each residual block consists of a convolutional layer, instance normalization layer [21] and ReLU activation.

The G_{local} consists of two convolutional blocks, a set of residual blocks and a transposed convolutional blocks. According to Figure 4, the input to the residual blocks in G_{local} is the element-wise sum of the output feature map of the two convolutional layers in G_{local}, and the feature map of G_{global}.

3) LOSS FUNCTION

The whole architecture profits from the combination of cycle-consistency loss besides the regular discriminator and generator losses. The cycle-consistency loss calculates $L1-$ *norm* between the original and cyclic image for unpaired

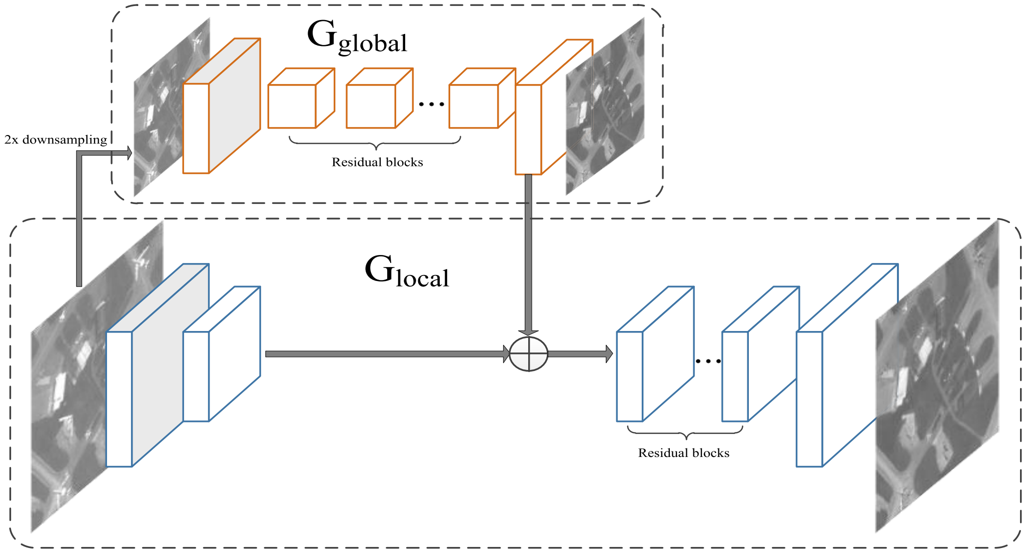


FIGURE 4. The network architecture of our multi-scale generator.

domains translation task. The cycle consistency loss is the same as utilized in CycleGAN.

$$L_{CYC}(F, G) = E_{x\sim p_{data}(x)}[||F(G(x))-x||_1] + E_{y\sim p_{data}(y)}[||G(F(y))-y||_1] \quad (1)$$

where we denote the data distribution as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$.

The original CycleGAN uses vanilla GAN objective as the adversarial loss function for both mapping functions. However, the original GAN loss function will result in model collapse and unstable training. Recently, the least square GAN [22] is proposed as a more stable alternative method, which can generate higher quality results. We use the least square adversarial loss as our model critic function. The objective function $L_{GAN}(G, D)$ is calculated as follows:

$$L_{GAN}(G) = \frac{1}{2}E_{x\sim p_{data}(x)}[(D_y(G(x)))^2] \quad (2)$$

$$L_{GAN}(D) = \frac{1}{2}E_{y\sim p_{data}(y)}[(D(y)-1)^2] + \frac{1}{2}E_{x\sim p_x(x)}[(D(G(x))+1)^2] \quad (3)$$

The full obeject of loss function are formulates as a combination of adversarial and cycle-consistency loss:

$$L(F, G, D_x, D_y) = L_{GAN}(G, D_y, X, Y) + L_{GAN}(F, D_x, Y, X) + \lambda L_{CYC}(F, G) \quad (4)$$

where λ controls the relative importance of the two objectives.We aim to solve:

$$G^*, F^* = argmin_{G,F} max_{D_x,D_y} L(F, G, D_x, D_y) \quad (5)$$

III. EXPERIMENT

A. SETTINGS

1) DATASETS

To promote our research in remote sensing images, we construct a simulation dataset. The aircraft model in the synthetic image mainly includes more than 30 kinds of aircraft including fighter planes, bombers, early warning aircraft and transport aircraft. The dataset contains a total of 9408 synthetic remote sensing images. The training set contains 7526 remote sensing images, and the test set contains 1882 remote sensing images. The size of the image is 512 × 512.

We created a real-time remote sensing image dataset from Google Earth, with data from Quickbird, WorldView, Landsat, and more. The actual remote sensing image contains a total of 9044 images, where the training set contains 7668 images and the test set contains 1358 images.

2) TRAINING SETTINGS

For the domain adaption experiments, we use the Adam [23] solver by setting $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\epsilon = 10^{-8}$. The batch-size is 1. All networks were trained from scratch with a learning rate of 0.0002. We keep the same learning rate for the first 100 epochs and linearly decay the rate to zero over the next 100 epochs. All convolutional filters are initialized by the method of He *et al.*'s initialization [24].

For object detection experiments, the feature extraction layer of the Faster R-CNN model uses the residual network ResNet-50 [25] and is initialized using ImageNet trained network weights. The optimization algorithm is a stochastic gradient descent method. The network learning rate is 0.0005, the maximum number of iterations is 70,000 rounds, and one training image is input per round. The non-maximum value suppression IOU threshold in RPN is set to 0.7, the weight attenuation coefficient is 0.0001, and the

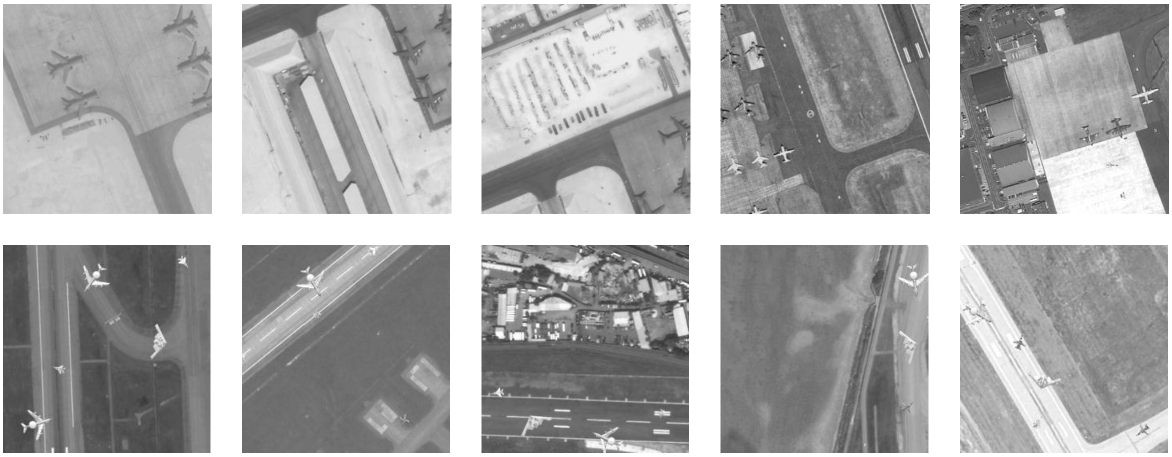


FIGURE 5. Examples of our dataset. The first row is the actually remote sensing images, and the second row is the simulation dataset.The network architecture of our multi-scale generator.

TABLE 1. Comparative experiments of our model on UC Merced for ×2 SR. Removing each component will degrade the final performance.

	Original Generator	Multiscale Generator
PSNR	27.839	29.404
SSIM	0.9169	0.9303

TABLE 2. Target detection mean accuracy of comparative experiments.

Type	AP
Simulation Dataset	0.1144
Original CycleGAN	0.1545
Our CycleGAN	**0.3887**

momentum coefficient is 0.9. Since the edge of the aircraft bounding box in the dataset is between 20 and 120 pixels, the initial side length of the multi-scale anchor is set to 16, 32, 64, 128, and the aspect ratio is set to 0.5, 1.0, 2.0, so each the convolutional sliding window position defaults to 12 anchors. All the processes are computed with an NVIDIA Tesla P100 GPU.

During the training process, the domain adaptive and object detection parts respectively occupy about 4G graphics memory. The multi-scale Cycle-GAN takes the same graphic memory as the training process in the testing phase. The memory usage of the target detection part during the test depends on the input image size.

B. DOMAIN ADAPTION ANALYSIS

The domain adaptation process is designed to make the simulated dataset tend to be a distribution of real remote sensing images. In order to obtain more realistic results, we propose a multi-scale generator based on the original generator of the CycleGAN framework. To demonstrate the effectiveness of our new generators and to verify the effectiveness of this multi-scale structure, we conducted experiments in actual remote sensing images. Since we do not have paired simulation datasets and actual remote sensing images, we use actual captured remote sensing images and network reconstructed images to assess the effectiveness of the network.

We use the peak signal-to-noise ratio (PSNR) [dB] and structural similarity index measure (SSIM) [26] as criteria to evaluate the performance of our proposed model. The experiment results show in Table 1, which are measured by the mean value of PSNR and SSIM on the testing dataset. From the quantitative results, our multi-scale generator obtains higher indicators. The PSNR and SSIM of our generator are 1.565dB and 0.0134 higher than the original one respectively.

In order to more fully demonstrate the effectiveness of our approach, we also show some of the visual comparisons, as shown in Figure 6. We observe that our proposed method can achieve better image reconstruction performance. The structure and texture of the target objects after domain transfer are more clear with our multi-scale generator. As shown in the first row of Figure 6, the head part of the airplane fused with the background after processed by the original generator, and an object that did not exist originally appeared next to the aircraft. Our designed generator can keep the basic structure of the target object and do not create new objects. According to the second row, when the target object in the scene is small, our improved domain migration method can better reconstruct small targets in the scene compared to the single-scale generator. Therefore, in general, our method has a more prominent effect on visual evaluation.

C. OBJECT DETECTION ANALYSIS

Our ultimate goal is to improve the accuracy of target detection, so we designed a set of experiments to validate our approach. We use three different types of data: raw simulation data, original CycleGAN migration data, and improved CycleGAN migration data as a training set for the target detection network. To ensure the fairness of the comparison experiment, we used the same parameter settings to train the three target detection networks.

We use the actual remote sensing image data directly as the test set of the three optimized target detection

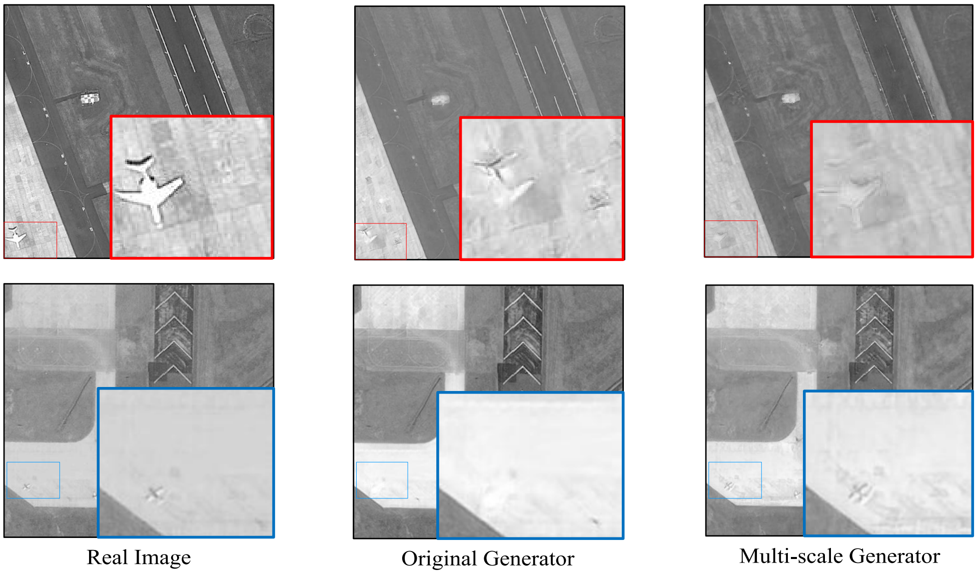


FIGURE 6. The comparison of the image conversion visual quality.

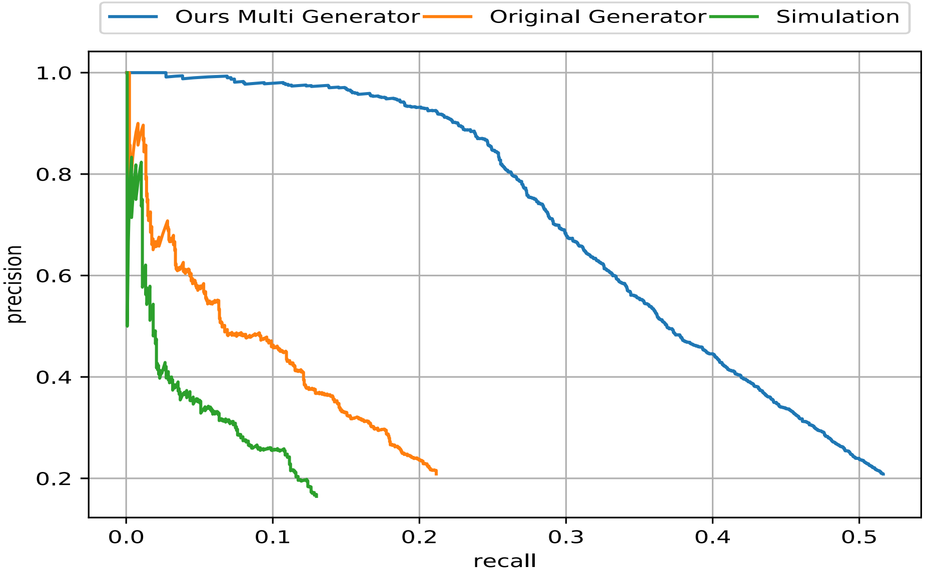


FIGURE 7. The PRC results on the actual remote sensing images.

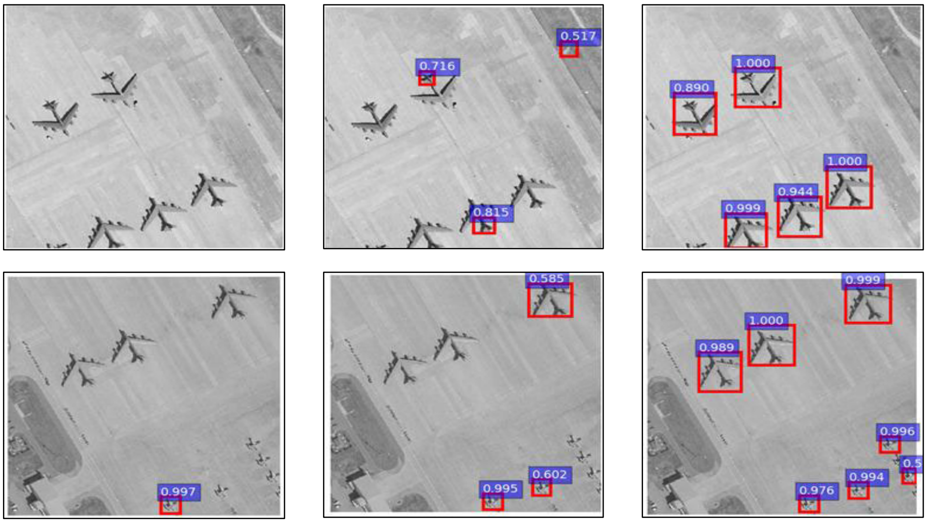
networks without performing any fine-tuning operations. The precision-recall curve (PRC) and average precision (AP) are adapted to quantitatively evaluate the performance of the object detection method. The indexes of precision and recall contain true positive (TP), false positive (FP) and false negative (FN). TP denotes the number of correct detections, FP denotes the number of false detection, and FN denotes the number of missing detections. The precision metric measures the fraction of detections that are TP. The recall metric measures the fraction of positives that are correctly retrieved. The precision and recall metrics are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Figure 7 shows the quantitative comparison results of the PRC. From this result, we can clearly see that our optimized method can have a very outstanding advantage compared to other methods.

The AP metric computes the area under the PRC. A higher AP value indicates better performance and vice versa. The AP trained in the original simulation dataset has a target detection value of 11.44% on the real remote sensing dataset, which is far from the requirements of aircraft detection. Comparing the results of Table 1, the dataset after processed by the basic CycleGAN can obtain a litter higher target detection precision compared with the raw simulation remote sensing dataset. However, the magnitude of accuracy improvement is very small. Our improved method achieved 38.87% AP value in the comparative test. Our approach has more than tripled



Simulation   Original Generator   Multi-scale Generator

FIGURE 8. The left side is the cross-domain detection result of the original composite image, the middle is the detection result after using the CycleGAN style migration, and the right side is the multi-scale CycleGAN for the style migration result.

the accuracy of the row simulation remote sensing dataset and more than twice the accuracy of the original CycleGAN. According to the analytic experiments, our improved method can better fit the image distribution of the actual test remote sensing image dataset and achieve better object detection accuracy.

It can be seen from Figure 8 of the aircraft target detection result that the simulated image has a large number of missed detections on the test set. From the second column of Figure 3, it can be concluded that the image after the original CycleGAN domain migration process can not capture the aircraft target on the test set better, and more cases capture the tail of the aircraft. From the target detection result graph of our multi-scale CycleGAN network, it can be analyzed, which can give greater confidence to the large target network on the test set, and can also be detected for small targets in the scene. Through the above analysis, our method has certain advantages.

D. COMPARISON WITH OTHER METHODS

To verify the effectiveness of our method, we compared it to Yan's [27] method currently published. Yan *et al.* first proposed the importance of a simulation dataset in remote sensing object detection. However, the code of the relevant comparison method is not open source, we implemented the algorithm according to the strategies in the author's paper by ourselves. According to the content of section II-A, we use Yan's method to generate simulation images and use them to train the target detection network. Finally, the trained model is used to detect real remote sensing images. The comparison result is shown in Table 3.

TABLE 3. Comparison result with other methods.

Type	AP
Yan et al. [27]	0.1074
Ours Simulation	0.1144
Ours System	0.3887

According to the first and the second rows of the result, we can find that the object detection accuracy of our method is only a little better than Yan's. However, our dataset simulation takes more computing sources for introducing SVM algorithms, dynamic skydomes, and other factors, to obtain only a little promotion. So Yan's algorithm is a simple but effective method to simulate more remote sensing images.

When comparing the final results of our system and Yan's, we can conclude that our whole simulation system has a more prominent performance in the final target detection accuracy, which also proves the power of the domain adaptation in our method.

IV. DISCUSSION

We propose a target detection method that does not require manual marking, and verify the effectiveness of the method through several experiments. Although our method has achieved good results, there is still much room for improvement in aircraft target detection accuracy on test dataset. By observing the results of domain migration processing and target detection, we analyze the effect of domain migration processing on the accuracy of subsequent target detection.

In the domain migration experiment, it is important that the converted image fits the distribution of the actual test dataset well. It can be seen intuitively from Figure 9 that the original CycleGAN processed data is not visually different from the original dataset. In the improved CycleGAN network, the overall distribution of images is closer to the distribution of test data. However, the aircraft structure in the processed image data is partially missing. On the test dataset, the network trained by our domain migration dataset

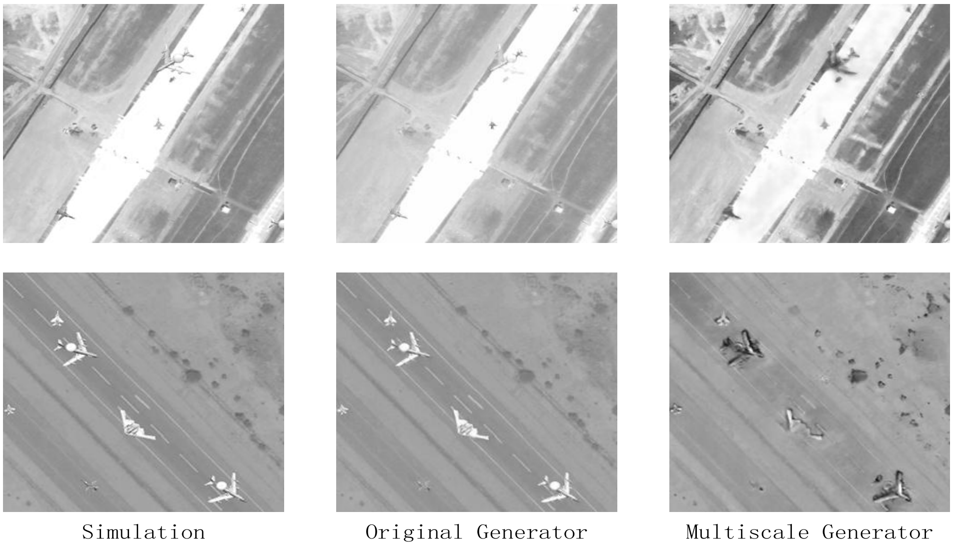


FIGURE 9. Domain transfer image comparison.

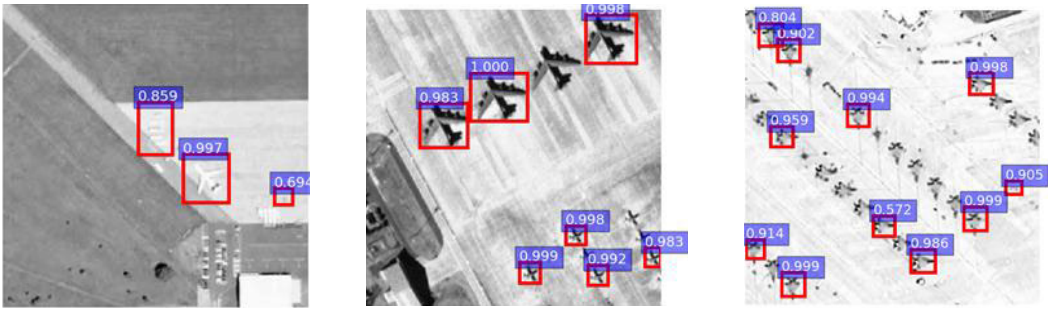


FIGURE 10. The object detection results with lower accuracy.

has greatly improved the accuracy of aircraft target detection compared to the network trained by the original simulation dataset. However, from Figure 10, there are many false alarms and missed detections on the test dataset. Therefore, a higher quality completion domain migration task is a problem worthy of further study.

V. CONCLUSION

In this article, we present a new framework. The framework automatically generates image datasets through data synthesis tools and automatically generates relevant annotation information. In order to establish a data domain distribution relationship between the composite image and the actual image, we propose an improved multi-scale cycle-gan network to bridge the reality gap. Our framework is tested in the real remote sensing dataset. Experiments have shown that our framework can achieve certain detection results in real remote sensing image datasets without any annotation data. In the future, we will continue to focus on image synthesis technology, automatic markup technology and domain adaptive technology. By combining these techniques, we will further improve the accuracy of target detection without manual labeling of data.

REFERENCES

[1] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "Understanding RealWorld indoor scenes with synthetic data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4077–4085.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 611–625.

[3] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[4] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[5] J. Howe, K. Pula, and A. A. Reite, "Conditional generative adversarial networks for data augmentation and adaptation in remotely sensed imagery," *Appl. Mach. Learn.*, vol. 11139, Sep. 2019, Art. no. 111390G.

[6] S.-C.-B. Lo, S.-L.-A. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE Trans. Med. Imag.*, vol. 14, no. 4, pp. 711–718, Dec. 1995.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[13] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[14] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[15] Y. Wang, Y. Zhang, Y. Zhang, L. Zhao, X. Sun, and Z. Guo, "SARD: Towards scale-aware rotated object detection in aerial imagery," *IEEE Access*, vol. 7, pp. 173855–173865, 2019.

[16] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[17] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS Int. J. Geo-Inf.*, vol. 161, pp. 294–308, Feb. 2020.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[19] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: Research challenges," *Ad Hoc Netw.*, vol. 3, no. 3, pp. 257–279, May 2005.

[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[21] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[22] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

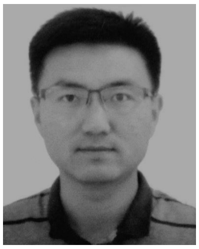
[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

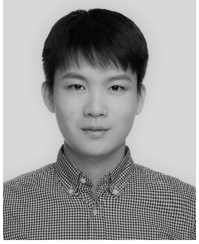
[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[27] Y. Yan, Z. Tan, and N. Su, "A data augmentation strategy based on simulated samples for ship detection in RGB remote sensing images," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 6, p. 276, 2019.



YAOLING WANG is currently pursuing the Ph.D. degree with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and remote sensing image processing.



JUN GU received the B.Sc. degree from Xidian University, Xi'an, China, in 2017. He is currently pursuing the M.Sc. degree with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and remote sensing image processing, especially on low-level tasks.

LIANGJIN ZHAO received the B.E. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2015, and the master's degree from the Beijing Institute of Technology, Beijing, China, in 2018.

He is currently a Research Assistant with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include the target detection and recognition in unmanned aerial vehicle remote sensing images, and simultaneous localization and mapping.

YUE ZHANG (Member, IEEE) received the B.E. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 2012, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interest includes the analysis of optical and synthetic aperture radar remote sensing images.

HONGQI WANG is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote-sensing image understanding.