







A Data Augmentation Strategy Combining a Modified pix2pix Model and the Copy-Paste Operator for Solid Waste Detection With Remote Sensing Images

Xiong Xu , Member, IEEE, Beibei Zhao, Xiaohua Tong , Senior Member, IEEE, Huan Xie , Senior Member, IEEE, Yongjiu Feng , Chao Wang , Changjiang Xiao , Xiaoxue Ke, and Jinhuan Du

Abstract—Solid waste detection is of great significance for environmental protection. In recent years, object detection methods based on deep learning have progressed rapidly. However, it is often extremely difficult to collect sufficient data to train a model with a good performance. In this article, a data augmentation strategy was introduced to generate sufficient synthetic high-quality images for solid waste detection. First, a modified pix2pix model was proposed, in which a local-global discriminator was designed to improve the detailed and global information of the generated images, which are commonly fuzzy with the original pix2pix model. Second, a copy-paste operator was utilized, which simply pastes the bounding box of the generated objects into different images to enhance the diversity of the samples. In this manner, the expanded dataset can be utilized to train different object detection models, for which FPN and Yolo-v4 were introduced as the validation models in this article. The experimental results show that the proposed strategy outperforms the traditional pix2pix method and the generated synthetic images can effectively improve the performance of object detection methods.

Index Terms—Copy-paste, data augmentation, local-global discriminator (LGD), object detection, pix2pix.

I. INTRODUCTION

IN RECENT decades, rapid economic development has caused severe environmental pollution problems, especially

Manuscript received 7 August 2022; revised 6 September 2022; accepted 20 September 2022. Date of publication 28 September 2022; date of current version 11 October 2022. This work was supported in part by the National key research and development program of China under Grant 2018YFB0505400, in part by the National Natural Science Foundation of China under Grant 41971299 and Grant 42221002, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, and in part by the Fundamental Research Funds for the Central Universities. (Corresponding author: Xiaohua Tong.)

Xiong Xu, Xiaohua Tong, Huan Xie, Yongjiu Feng, and Changjiang Xiao are with the College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China, also with the Shanghai Key Laboratory for Planetary Mapping and Remote Sensing for Deep Space Exploration, Shanghai 200092, China, and also with the Frontiers Science Center for Intelligent Autonomous Systems, Shanghai 201210, China (e-mail: vxixiong@tongji.edu.cn; xhtong@tongji.edu.cn; huanxie@tongji.edu.cn; yjfeng@tongji.edu.cn; cjxiao@tongji.edu.cn).

Beibei Zhao, Chao Wang, Xiaoxue Ke, and Jinhuan Du are with the College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China (e-mail: 1933666@tongji.edu.cn; wangchao2019@tongji.edu.cn; 1225187900@qq.com; 3248835284@qq.com).

Digital Object Identifier 10.1109/JSTARS.2022.3209967

in developing countries. The disposal of solid waste has always been an important subject and deserves more attention with the increasing environmental pressure. In China, ~1 billion tons of household or construction waste (CW) is produced annually [1]. Arbitrarily dumped solid waste may result in serious harm to the surrounding water, soil, and air. It is important to monitor the location and expansion of solid waste.

In-site investigation is commonly used to identify solid waste; however, it is time-consuming and inefficient [2]. Recently, with the rapid development of remote sensing imaging technology, high-resolution images have become commonly available with the advantages of large coverage and high efficiency. Many researchers have attempted to use high-resolution remote-sensing images to detect solid waste. Bagheri and Hordon [3] used aerial remote sensing images to visually interpret hazardous waste sites in Burlington City. Silverstri and Omri [4] obtained the distribution of illegal landfills (LFs) in Italy based on a supervised classification method that utilized the color, shape, and other features of a garbage dump with IKONOS images. Qin [5] extracted garbage dump sites from GaoFen-2 images using a decision tree classification method that combined the soil-regulated vegetation index calculation. Liu et al. [6] established rules of interpretation based on the color, shape, texture, and other shallow features of garbage, and performed garbage site identification combined with GIS spatial analysis. Zhang et al. [7] proposed a new object-based solid waste detection method that considers the spectral and textural features of urban solid waste in QuickBird images.

In recent years, convolutional neural networks (CNNs) have been widely used for object detection [8], [9], [10], [11]. CNN-based object detection methods can be divided into two types: anchor-based [12], [13] and anchor-free [14], [15] methods. The introduction of a CNN can significantly improve the accuracy of object detection; however, the prerequisite is that sufficient samples can be acquired to train the CNN model for considerable performance [16]. For real-life obstacles, such as solid waste images, the availability of sufficient data is not always possible [17]. Therefore, a data augmentation strategy is commonly used to address the problem of insufficient training samples in deep learning research [18]. Traditional data augmentation methods

are based on geometric transformations, such as reflection, flipping, scaling, translation, and rotation [19]. For example, a horizontal flip was used to expand the training data for faster R-CNN [12]. Random erasing [20] can reduce overfitting by introducing an occlusion for which a part of the image is erased randomly. Similarly, random cropping [21] was commonly used to extract image patches as new samples.

Recently, another type of data augmentation technique, generative adversarial network (GAN), has been proposed and widely used in various image recognition tasks [22], [23], [24]. Generally, a GAN consists of two components: a generator model that is trained to generate new samples and a discriminator model that tries to classify samples as either real or fake [22]. With the adversarial training procedure between the two models, the generator learns to generate plausible new data that are difficult to distinguish from the real samples by the discriminator. Based on the GAN model, Randford et al. [25] proposed the deep convolutional GAN, which incorporates the CNN into GAN to enable training to become more stable. Another Wasserstein GAN was introduced in [26] which utilizes the Wasserstein distance in the loss function of the generator. Abhishek proposed a novel spectral index GAN [27] for generating multispectral remote sensing images. This network is defined to effectively perform data augmentation, starting from a limited number of training samples in the spectral domain, to train deep learning models. Kim and Hwang [28] proposed a GAN framework where synthetic background images and infrared small targets are generated in two independent processes.

Furthermore, a pix2pix model was proposed [29] as a general framework for image translation. Huang et al. [30] proposed a data augmentation method utilizing a pix2pix GAN for the automatic generation of object images under various illumination effects. Grishin et al. [31] proposed a novel strategy for data preparation based on the fusion of objects and background images to form composite data. Rizwan et al. [32] further proposed a data augmentation scheme comprising pix2pix and a customized loss function to address the data sparsity challenge in the minimization of the drive test data.

In contrast to regular objects, solid waste targets commonly have arbitrary shapes, for which the boundaries and textures of the generated objects are always fuzzy. In this letter, a new data augmentation method is proposed for solid waste detection, where a modified pix2pix model and a copy-paste operator are integrated. For the modified pix2pix model, which was termed LGD-pix2pix, a local-global discriminator (LGD) was designed to ensure that the generated images had detailed information in the local region and maintained global color consistency. The copy-paste operator simply copies the bounding boxes of the objects in each image and pastes them with other images. Our primary contributions are summarized as follows.

- 1) A solid waste dataset was built manually in which four types of solid wastes could be distinguished, including CW, LF, municipal solid waste (MSW), and tailings pond (TP).
- 2) A data augmentation framework combining a modified pix2pix and copy-paste operator was proposed, which can

generate high-quality training images to improve the performance of solid waste detection using high-resolution remote sensing images.

- 3) An LGD was designed to improve the quality of the coarser-generated images from the pix2pix model and generate more realistic samples to increase the number and diversity of the augmented dataset.

The remainder of this letter is organized as follows. Section II provides the proposed data augmentation framework and describes the different components in detail. Section III introduces the built solid waste dataset. The experimental results are provided and discussed. Finally, conclusion is presented in Section IV.

II. METHODOLOGY

The framework of the proposed data augmentation method is presented in Fig. 1. The LGD-pix2pix model and copy-paste operator are integrated to generate fake samples with high-resolution properties. The detailed procedure is as follows:

A. Pix2pix

For GAN-based image translation applications, the GAN generator commonly relies on the input random noise to generate images, for which it is difficult to stabilize the output result [29]. For the pix2pix model, which belongs to the conditional GAN (cGAN) type, a pattern of the input image can be used as a constraint to generate the specified output fake image. The general training procedure for the pix2pix model is illustrated in Fig. 2. In our tasks, the generator aims to translate the semantic segmentation map to realistic remote sensing images, whereas the discriminator tries to distinguish the generated fake images from real ones. Pix2pix model requires pairs of images (x and y) during the training procedure, in which x is a segmentation map corresponding to the solid waste target and y is the real solid waste image. The generator was employed to learn a mapping from x without random noise to the corresponding y , whereas the discriminator determines whether the generated image $G(x)$ is real or fake.

For the training procedure, a segmentation map x is used to generate $G(x)$ with generator G . Both x and $G(x)$ values are input into discriminator D to obtain a probability value indicating whether the input x and $G(x)$ are a pair of real images. Therefore, the goal of discriminator D is to output a greater probability value when the input is a pair of real images (x and y) and a smaller value when the input (x and $G(x)$) is not. Generator G aims to maximize the output value of D with the pair of x and $G(x)$.

The commonly used objective function of the cGAN is as follows:

$$L_{cGAN}(G, D) = E_{x,y} [\log D(x, y)] + E_x [\log (1 - D(x, G(x, z)))] \quad (1)$$

where $G(x, z)$ is the generated image according to the random noise, $D(x, *)$ represents the result of the discriminator's judgment of the real image y or the generated image $G(x, z)$.

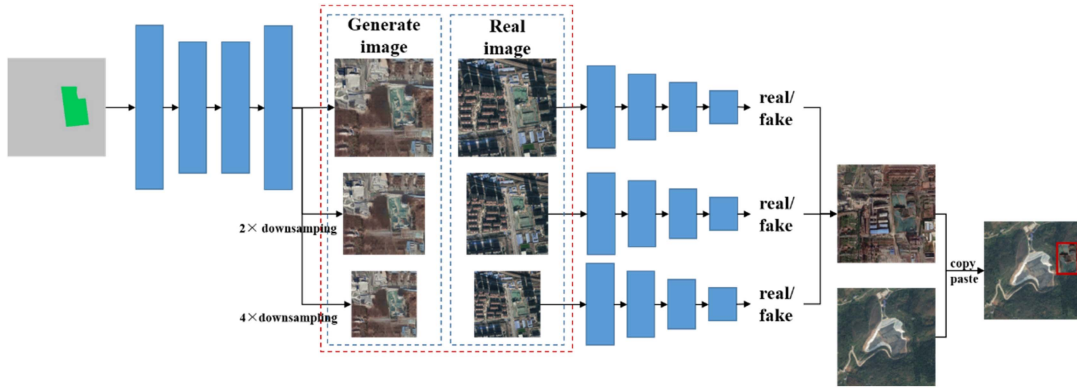


Fig. 1. Pipeline of the proposed data augmentation method.

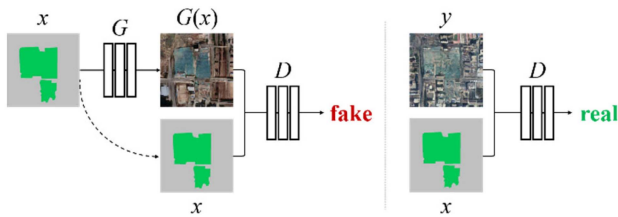


Fig. 2. Training of pix2pix algorithm.

For the pix2pix model, an $L1$ loss was further integrated to represent the difference between the generated image and the real image, defined as follows:

$$L_{L1}(G) = E_{x,y,z} [\|y - G(x, y)\|_1]. \quad (2)$$

Therefore, the objective function of the pix2pix model can be modeled as follows:

$$L(G, D) = L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (3)$$

where λ is a hyperparameter to balance the weight of the $L1$ loss, which is empirically set to 100, as suggested in [29]. The goal is to minimize the training error of the generator and maximize the training error of the discriminator, as expressed in

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G). \quad (4)$$

1) *Generator G*: In pix2pix, U-net is commonly used as generator G , which is a type of encoder-decoder CNN featuring a skip layer connection [33]. Its structure is shown in Fig. 3. The advantage of using the U-net network architecture is that images of the same size are connected between the encoding and decoding parts of the network, giving the generator the ability to skip some steps, also known as skip connections. The corresponding feature distributions in the U-net and decode processes are connected by skip connections so that low-level details are preserved under different resolution conditions. When the network is trained, part of the information can be transmitted directly over the connection.

2) *Discriminator D*: The discriminator judges the authenticity of the input image using a CNN structure, as shown in Fig. 4. Given an input image, a probability value can be obtained

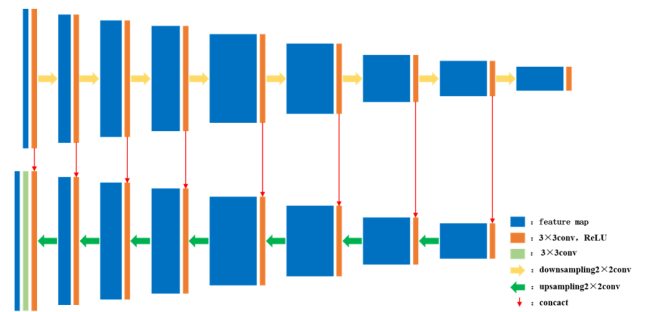


Fig. 3. Generator with U-net structure.

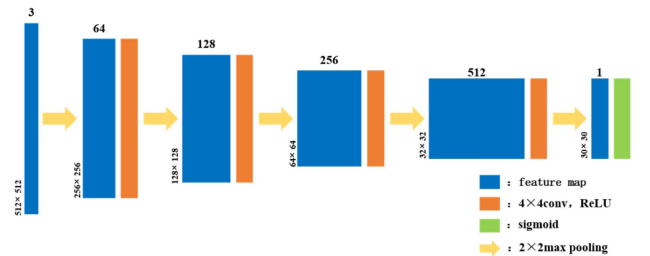


Fig. 4. Structure of the PatchGAN discriminator.

using the discriminator that indicates the authenticity of the image. If the probability value was close to zero, the input image was considered fake. Conversely, if the probability value is close to 1, the input image is identified as true. To simulate a high-frequency image, the structure of the local image patch needs to be considered. Therefore, the Markov discriminator PatchGAN [29] was used to judge the authenticity of each $N \times N$ patch and average the judgment results of all patches in an image as the output of the final discriminator. N is the hyperparameter of the network, which was set to 30 in this article. Dividing the image into multiple patches enables the model to process images of any size.

B. Local-Global Discriminator

The key issue with the original pix2pix model is that the generated images are mostly low-resolution and lack detailed information [29]. To generate solid waste images with higher

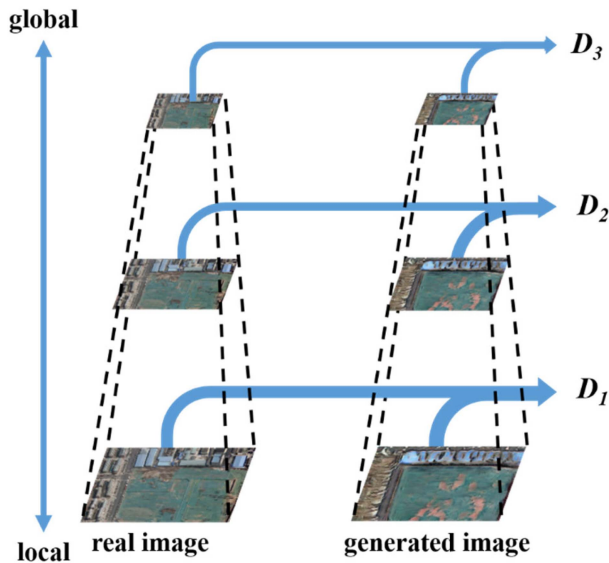


Fig. 5. Proposed local-global discriminators.

resolution to improve detection accuracy, the discriminator should have a larger receptive field by employing deeper network structures or larger convolutional kernels. However, this may result in an increase in the training time. In this letter, an LGD module was designed, which consists of multilayer discriminators D_i ($i \in [1, n]$) representing different scales, as shown in Fig. 5, where n is the number of discriminators. The image is discriminated at the local and global scales and the average value is calculated as the final discriminative output. This operator has the advantage of reserving information at both the local and global scales to guide the generator to generate images with rich details and global consistency. To balance the training cost and image generation quality, n is typically set to three.

In this case, the original semantic segmentation image and the corresponding solid waste image from the generator were downsampled at scales 2 and 4 to be the inputs of D_2 and D_3 , respectively. All discriminators of D_i have the same structure, for which D_1 with the finest scale encourages the generator to produce finer details, and D_3 has the largest receptive field with its coarsest scale. Thus, the proposed LGD operator can guide the generator to yield more realistic images. Therefore, the cost function can be rewritten as follows:

$$L(G, D) = \sum_{i=1,2,3} \lambda_{D_i} L_{cGAN}(G, D_i) + \lambda L_{L1}(G) \quad (5)$$

where λ_{D_i} is a regularization parameter used to control the weight of each scale discriminator, which is commonly set to one by default.

C. Copy-Paste Operator

It is straightforward to utilize the copy-paste strategy [34] to generate new images. Rather than pasting object instances into images, the bounding boxes of objects can be copied and pasted directly to different images to generate augmented images, as illustrated in Fig. 6.

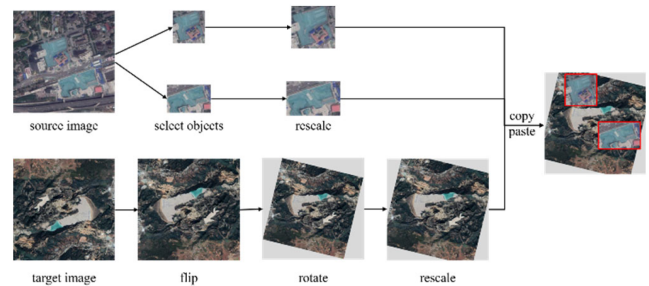


Fig. 6. Copy-paste operator.

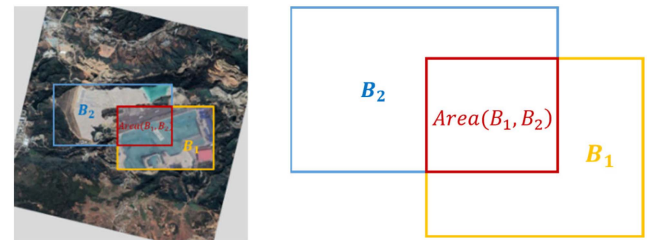


Fig. 7. Toy example of POI calculations.

First, two images were randomly selected, one of which was the source and the other was the target image. Subsequently, an arbitrary flipping operator, horizontal, vertical, or diagonal, was applied to the target image, and the flipped target image was further rotated and rescaled. Further, two objects from the source image were randomly selected and copied to the target image. It should be noted that the bounding boxes of these selected objects are also rescaled before pasting at an arbitrary position in the target image. It is evident that the pasted object may overlap with existing objects in the target image. Therefore, it is essential to determine whether the location of the pasted object should be adjusted. Given that B_1 and B_2 are the bounding boxes of the pasted and existing objects, respectively, the area of intersection between B_1 and B_2 can be expressed as $\text{Area}(B_1, B_2)$, as shown in Fig. 7.

Thus, the proportion of the interaction region (POI) in B_2 can be calculated as shown in Eq. (6). If the calculated POI value is greater than a given threshold, the location of the pasted object should be adjusted. To conclude, these pasted objects are recorded and the ground truth annotations are revised

$$\text{POI} = \frac{\text{Area}(B_1, B_2)}{\text{Area}(B_2)}. \quad (6)$$

III. EXPERIMENT

A. Dataset Description

Currently, there are no publicly available remote-sensing datasets for solid-waste detection applications. To evaluate the proposed data augmentation method, a solid waste dataset was first constructed, namely SWD, which can be further used as a benchmark dataset for solid waste detection. Based on the official MSW classification system and considering that the solid waste type can be recognized with remote sensing images, four

TABLE I
STATISTICS OF THE CONSTRUCTED SOLID WASTE DATASET

category	Train set		Validation set		Test set	
	#images	#instances	#images	#instances	#images	#instances
Construction waste	327	447	109	163	218	314
Landfill	369	387	123	134	246	264
Municipal solid waste	300	372	100	114	200	225
Tailings pond	379	424	127	137	254	282

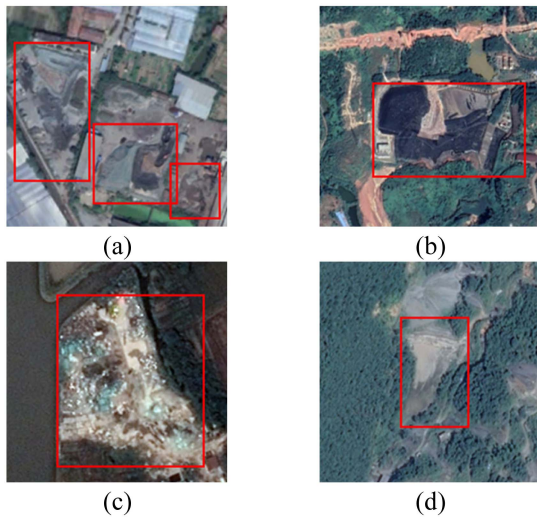


Fig. 8. Examples of four types of solid waste. (a) Construction waste. (b) Landfill. (c) Municipal solid waste. (d) Tailings pond.

types of solid waste were finally included in the SWD: CW, LF, MSW, and TP. Examples of the four different types of solid wastes are shown in Fig. 8. Subsequently, the procedure for image collection and annotation was elaborated as follows.

For CW and MSW, clues were found from a governmental investigation launched by the Ministry of Ecological Environment of China, in which the locations of solid waste were described in detail. For TPs, clues were obtained from a list published by the Ministry of Emergency Management of the People's Republic of China. For the LF, we searched for keywords related to this type of solid waste on the Baidu map and recorded the listed positions of the manually checked results. With these detailed descriptions of solid waste locations, the corresponding targets were further confirmed using Google Earth and the images were captured within a region of 512×512 .

Given the obtained images, the bounding box of each solid waste object (x_{\min} , y_{\min} , x_{\max} , y_{\max}) was annotated manually using LabelImg, which is a commonly used image annotation tool in object detection. In addition, a polygon label was also generated with Labelme, which is commonly used in image segmentation studies, as input for the modified pix2pix algorithm.

Finally, 2752 images were collected, of which the numbers of different solid wastes (CW, LF, MSW, and TP) were 654, 738, 600, and 760, respectively. These images were further randomly divided into training, validation, and test sets according to a 3:1:2 ratio, as given in Table I.

B. Parameters Setting

The Adam optimizer was adopted for the original pix2pix and LGD-pix2pix models. At least one image was included in each mini-batch. A total of 300 epochs were set for the training and the initial learning rate of 0.001 was reduced by a factor of 10 after the 200th epoch. The hyperparameters of the loss function were set as $\lambda = 100$ and $\lambda_{D_i} = 1$. Moreover, to validate the performance of the proposed data augmentation method, the expanded dataset was further tested using two typical object detection methods: feature pyramid networks (FPNs) [12] and Yolo-v4 [13]. The two models were trained with stochastic gradient descent for ten epochs and the initial learning rate was 0.0025, which was also reduced by a factor of 0.1 after the 16th epoch. The weight decay and momentum were set as 0.0001 and 0.9, respectively, while the other hyperparameters followed the default setting if not specifically noted, as suggested in the literature. The backbones used for FPN and Yolo-v4 are ResNet101 and Darknet53, respectively. All backbones were pretrained on the ImageNet dataset. The image size was set to 512×512 pixels in our experiments, which were conducted on a GRID P40-24Q GPU with CUDA 10.1.

C. Experimental Analysis

In this section, data augmentation methods are evaluated and tested using the previously built SWD dataset. The experiments show that the proposed method outperforms the traditional pix2pix method. Its effectiveness is further verified by ablation experiments. For clarity, the datasets generated using different data augmentation methods were distinguished by their specific abbreviations. Given the initial dataset SWD, the expanded datasets with the original pix2pix model, LGD-pix2pix model, and proposed strategy were termed SWD-P2P, SWD-LGD, and SWD-LPC (LGD-Pix2pix model incorporating the copy-paste operator), respectively. The final number of instances in the different datasets is given in Table II. The same number of training samples were obtained for both the SWD-P2P and SWD-LGD datasets to ensure fair competition. Identical test samples were used to evaluate the training samples' effectiveness from different datasets. The procedure for sample generation was as follows:

1) *Sample Generation With the Pix2pix Models*: With the built SWD dataset, the original pix2pix and LGD-pix2pix models were first trained. The test set of the SWD serves as the training set of the pix2pix models so that the training set of the SWD can be expanded with the fine-tuned pix2pix models,

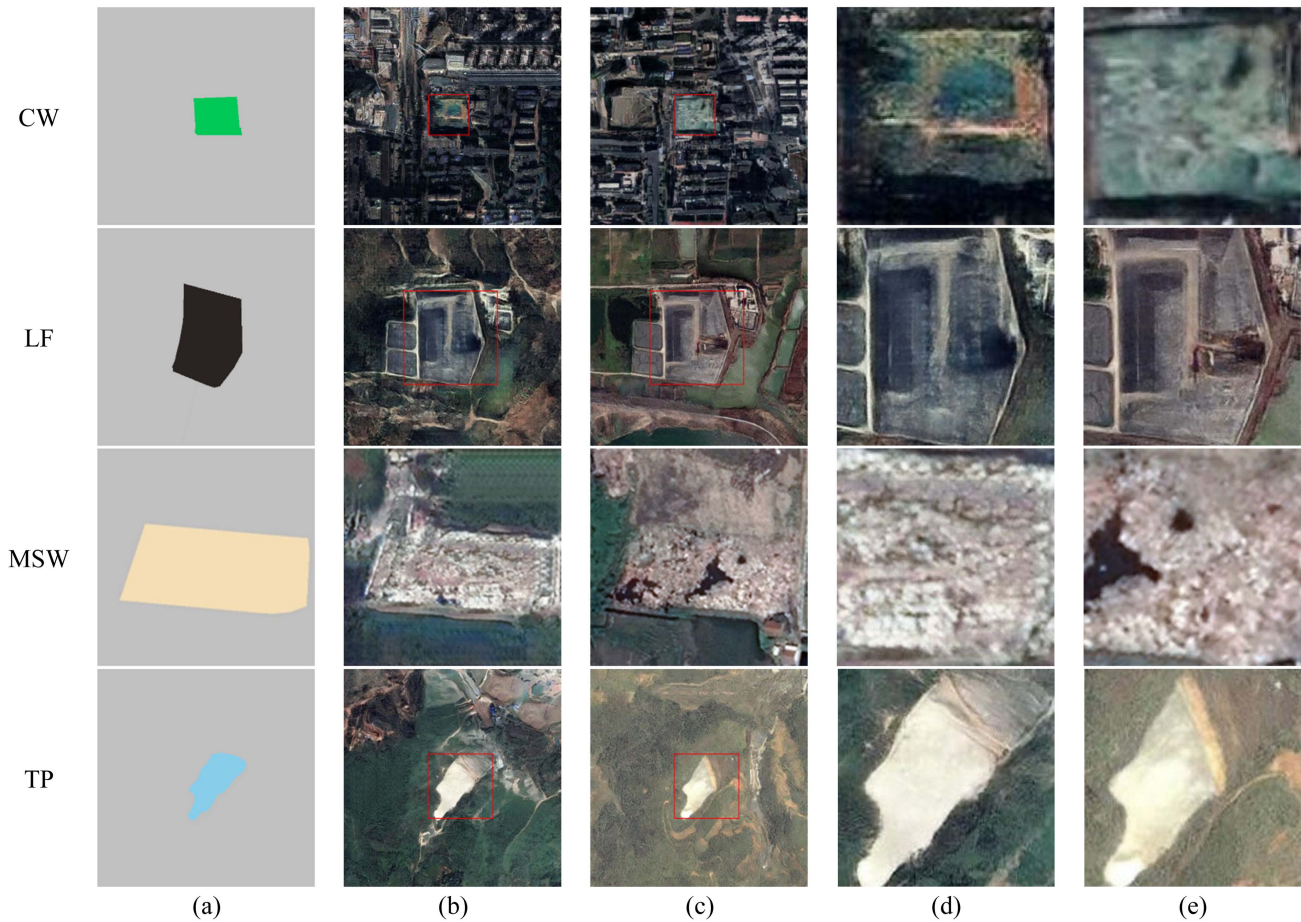


Fig. 9. Generated synthesis images with different pix2pix models. (a) Input segmentation map. (b) Original pix2pix. (c) Proposed LGD-pix2pix. (d) An enlarged part of (b). (e) Enlarged part of (c).

TABLE II
NUMBERS OF INSTANCES IN DIFFERENT DATASETS WITH DATA AUGMENTATION

	Training samples			Test samples
	SWD	SWD-P2P SWD-LGD	SWD-LPC	
CW	447	923	967	314
LF	387	751	792	264
MSW	372	695	756	225
TP	424	836	871	282
total	1630	3205	3386	1085



Fig. 10. Samples generated with the copy-paste operator.

which will be further used as a training set for the object detection methods. Both pix2pix models were trained on an identical test set of SWD under the same experimental conditions. The same SWD training set was utilized for expansion. The synthesized images with different pix2pix models are presented in Fig. 9. It can be observed that images generated with the original pix2pix model have clear texture information in the target area; however, the backgrounds are fuzzy. For the LGD-pix2pix model, the generated texture is more authentic in both the target area and background. The object boundary appears clearer than the results

from the pix2pix model. Therefore, different texture combinations can be incorporated to enhance sample diversity, for which the characteristics of each original image can be inherited and fused perfectly.

2) *Sample Generation With the Copy-Paste Operator*: After the SWD-LGD dataset was generated using the LGD-pix2pix model, a copy-paste operator was introduced to obtain the final SWD-LPC dataset. Fig. 10 shows a couple of generated samples with the copy-paste operator, where the red bounding boxes are pasted objects.

TABLE III
ACCURACY ASSESSMENT OF DIFFERENT DATA AUGMENTATION METHODS BY SWD DETECTION

Method	FPN					YOLO-v4				
Category	CW	LF	MSW	TP	mAP	CW	LF	MSW	TP	mAP
SWD	83.92%	84.79%	81.77%	86.03%	84.13%	82.63%	83.72%	80.07%	85.82%	83.06%
SWD-P2P	84.93%	85.68%	82.93%	87.21%	85.19%	83.88%	84.97%	81.76%	86.59%	84.30%
SWD-LPC	86.97%	87.83%	85.28%	88.79%	87.22%	84.98%	85.51%	83.80%	87.71%	85.50%

Horizontal, vertical, and diagonal flipping operators were randomly imposed on the target image, with a probability of 0.5° and a rotation angle of up to $\pm 10^\circ$. The flipped target image was further rescaled with a scale factor defined as

$$\text{scale factor} = \begin{cases} [0.50.8] & w > 256 \text{ and } h > 256 \\ [1.22.0] & w < 256 \text{ and } h < 256 \\ (0.81.2) & \text{otherwise} \end{cases} \quad (7)$$

where w and h are the width and height, respectively, of the pasted object. The threshold of the POI value was set to 0.1. The corresponding annotations of the augmented images were calculated and automatically recorded.

3) *Validation of Object Detection Application*: To validate the effectiveness of the data augmentation framework, two typical object detection methods, FPN and Yolo-v4, were employed. Different training sets were used for the training of the object detection methods and an identical test set was utilized for accuracy assessment, as given in Table II.

Table III presents the performance of the proposed framework compared with traditional augmentation methods. It can be seen that the highest accuracies of 87.22 and 85.50% are achieved with our proposed strategy on FPN and Yolo-v4, respectively. For the FPN detection model, improvements of 3.09 and 2.03% can be obtained with the generated SWD-LPC dataset, compared with the SWD and SWD-P2P datasets, respectively. Meanwhile, the proposed method achieved an improvement of 2.44% over the original SWD dataset with YOLO-v4 by increasing the training samples from 1630 to 3386, as given in Table II.

Moreover, it can also be noted that the best performance can be achieved with the proposed method for all four classes. The experimental results showed that the proposed strategy can effectively improve the dataset diversity and generalization performance of object-detection methods.

D. Discussions

1) *Impact of the LGD Structure*: The impact of the designed LGD structure was analyzed further. As mentioned previously, three discriminators were employed in this article, considering the balance between accuracy and efficiency. The detection results on the SWD dataset with different settings are given in Table IV, in which the utilization of three discriminators achieves the highest accuracy of 86.26%, based on the FPN. With an increase in the number of discriminators, the texture information of the generated solid waste images gradually enriched, and global consistency further improved. Hence, the richness

TABLE IV
COMPARISON WITH DIFFERENT NUMBER OF DISCRIMINATORS

Method	LGD	copy-paste	mAP
FPN	1	×	85.19%
FPN	2	×	85.75%
FPN	3	×	86.26%

TABLE V
COMPARISON WITH MULTIPLE COPY-PASTE OPERATORS

Method	LGD	copy-paste	mAP
FPN	3	1	87.22%
FPN	3	2	87.71%
FPN	3	3	87.79%
Yolo-v4	3	1	85.50%
Yolo-v4	3	2	85.77%
Yolo-v4	3	3	85.91%

of the augmented dataset also improved with the generated high-quality images.

2) *Impact of the Copy-Paste Operator*: In this experiment, we also discussed the effect of using multiple copy-paste operators to augment the SWD dataset. The multiplier was set to 1, 2, and 3, where 1, 2, and 3 indicate that the number of objects in the augmented training dataset is approximately double, triple, or quadruple of the original dataset after the copy-paste operation is used once, twice, or thrice, respectively. Table V gives a comparison of the accuracy with different numbers of copy-paste operators. It can be summarized that the more the copy-paste operator is repeated, the higher the accuracy obtained. The highest accuracies of 87.79 and 85.91% were achieved with triple copy-paste operators on FPN and Yolo-v4 respectively, an improvement of 0.57 and 0.41%, respectively, compared with the single copy-paste operator. The utilization of the copy-paste operator has proven to be effective; however, the performance cannot be improved significantly by multiple operators. Therefore, the copy-paste operator was used only once in this article. Accuracies of 87.22 and 85.50% were used, as given in Tables III and V.

IV. CONCLUSION

In this letter, a novel data augmentation method combined with an improved pix2pix model and copy-paste operator was proposed for solid waste detection applications. The samples of solid waste in remote sensing images are typically insufficient, for which object detection models are likely to overfit. Compared with the traditional pix2pix model, the proposed method can generate sufficient synthetic images with high resolution and clear texture advantages. To verify the effectiveness of the proposed method, a solid waste dataset was constructed and utilized for data augmentation and object detection applications. The experimental results show that the proposed method exhibits better performance than the original pix2pix model. Therefore, it can be used to improve the ability of object detection algorithms.

REFERENCES

- [1] National Bureau of Statistics of the People's Republic of China. *China Statistical Yearbook*. Beijing, China: China Statistics Press, 2011–2020.
- [2] S. Zhang et al., "A high resolution remote sensing detection method for intelligent solid waste yard based on mask R-CNN instance segmentation," in *Proc. Nat. Org. Solid Waste Treat. Resource Utilization Summit*, 2020, pp. 9–15.
- [3] S. Bagheri and R. M. Hordon, "Hazardous waste site identification using aerial photography: A pilot study in burlington county, New Jersey, USA," *Environ. Manage.*, vol. 12, no. 3, pp. 411–412, 1988.
- [4] S. Silvestri and M. Omri, "A method for the remote sensing identification of uncontrolled landfills: Formulation and validation," *Int. J. Remote Sens.*, vol. 29, no. 4, pp. 975–989, 2008.
- [5] H. C. Qin, "Research on urban domestic waste supervision method based on domestic high score remote sensing image," *Informatization China Construction*, no. 4, pp. 75–77, 2016.
- [6] Y. L. Liu et al., "Application research of Beijing 1 small satellite monitoring informal garbage dump," *J. Remote Sens.*, vol. 13, no. 2, pp. 320–326, 2009.
- [7] F. L. Zhang, S. H. Du, and Z. Guo, "Extraction of municipal solid waste using high-resolution image," *Spectrosc. Spectral Anal.*, vol. 33, no. 8, pp. 2024–2030, 2013.
- [8] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, 2019, Art. no. 765.
- [9] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Q. Liu, "A comparative study of state-of-the-art deep learning algorithms for vehicle detection," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 2, pp. 82–95, Mar. 2019.
- [10] U. Alganci, M. Soydas, and E. Sertel, "Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images," *Remote Sens.*, vol. 12, 2020, Art. no. 458.
- [11] H. Ljunggren, "Using deep learning for classifying ship trajectories," in *Proc. IEEE 21st Int. Conf. Inf. Fusion*, 2018, pp. 2158–2164, doi: [10.23919/ICIF.2018.8455776](https://doi.org/10.23919/ICIF.2018.8455776).
- [12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
- [13] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020.
- [14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
- [15] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
- [16] H. Wang, *Research On Data Enhancement Method Based On Generative Adversarial Network*. Nanjing, China: Nanjing Univ. Posts Telecommun., 2020.
- [17] P. Kaur, B. S. Khehra, and E. B. S. Mavi, "Data augmentation for object detection: A review," in *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, 2021, pp. 537–543.
- [18] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling CNNs with simple transformations," in *Proc. NIPS Workshop Mach. Learn. Comput. Secur.*, 2017, pp. 1–21.
- [19] E. D. Cubuk et al., "AutoAugment: Learning augmentation policies from data," 2018.
- [20] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13001–13008, 2020.
- [21] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 87.1–87.12.
- [22] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [23] J. Rabbi et al., "Small-Object detection in remote sensing images with End-to-End edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1432.
- [24] S. Y. Zhao and J. W. Li, "Low rank image generation method based on generative adversarial network," *Acta Automatica Sinica*, vol. 44, no. 5, pp. 829–839, 2018.
- [25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.
- [26] M. Arjovsky, S. Chintal, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
- [27] A. Singh and L. Bruzzone, "SIGAN: Spectral index generative adversarial network for data augmentation in multispectral remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6003305.
- [28] J. H. Kim and Y. Hwang, "GAN-Based synthetic data augmentation for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002512.
- [29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Imageto-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.
- [30] Y.-H. Huang, C.-H. G. Li, and Y.-M. Chang, "Illumination-Robust object coordinate detection by adopting pix2pix GAN for training image generation," in *Proc. IEEE Int. Conf. Technol. Appl. Artif. Intell.*, 2019, pp. 1–6.
- [31] N. Grishin, A. Lozhkina, K. Bukharov, D. Makhotkin, and V. Semenkin, "Composite data preparation algorithm for SAR imagery object recognition," in *Proc. Int. Conf. Eng. Telecommun.*, 2021, pp. 1–5.
- [32] A. Rizwan, A. Abu-Dayya, F. Filali, and A. Imran, "Addressing data sparsity with GANs for Multi-fault diagnosing in emerging cellular networks," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Commun.*, 2022, pp. 318–323.
- [33] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, 2019.
- [34] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," 2020.