GeoPalette: Road Segmentation with Limited Satellite Imagery

Wenmiao Hu
hu.wenmiao@u.nus.edu
Grab-NUS AI Lab, Singapore

Yifang Yin
idsyin@nus.edu.sg
Grab-NUS AI Lab, Singapore

Ying Kiat Tan
tanyk@comp.nus.edu.sg
National University of Singapore

An Tran
an.tran@grabtaxi.com
Grabtaxi Holdings Pte Ltd, Singapore

Hannes Kruppa
hannes.kruppa@grabtaxi.com
Grabtaxi Holdings Pte Ltd, Singapore

Roger Zimmermann
rogerz@comp.nus.edu.sg
National University of Singapore

Figure 1: When the real training dataset is limited, synthetic training pairs can be generated to assist model training. (a) Synthetic images with different appearances generated from GeoPalette using a real-world road mask. (b) Synthetic images from cSinGAN generators using a synthetic mask from OpenStreetMap (OSM). (c) Workflow of generating synthetic images.

ABSTRACT

In recent years, Geo-information extraction from high-resolution satellite imagery has attracted a lot of attention. However, because of the high cost of image acquisition and annotation, there are limited datasets available. Compared to close-range imagery datasets, existing satellite datasets have a much lower number of images and cover only a few scenarios (cities, background environments, *etc.*). They may not be sufficient for training robust learning models that fit all environmental conditions or be representative enough for training regional models that optimize for local scenarios. In this study, we propose GeoPalette, a Generative Adversarial Network (GAN) based tool to generate additional synthetic training samples for boosting model performance when the training dataset is limited. Our experiments on road segmentation show that using additional synthetic data can improve the model performance mean Intersection over Union (mIoU) from 60.92% to 64.44%, when 1,000 real training pairs are available for learning, which reaches a similar level of performance as a model is standard-trained on 4,000 real pairs (64.59%), *i.e.*, a 4-fold reduction in real dataset size.

CCS CONCEPTS

- **Information systems** → Content analysis and feature selection;
- **Computing methodologies** → Image segmentation; Transfer learning.

KEYWORDS

Assisted-learning, GAN, road segmentation, satellite imagery

ACM Reference Format:

Wenmiao Hu, Yifang Yin, Ying Kiat Tan, An Tran, Hannes Kruppa, and Roger Zimmermann. 2021. GeoPalette: Road Segmentation with Limited Satellite Imagery. In *29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21)*, November 2–5, 2021, Beijing, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3474717.3483914>

1 INTRODUCTION

With the increasing image acquisition capacities of high-resolution commercial satellite constellations, information extraction from satellite imagery in near real-time has become possible in recent years [3, 12]. However, existing high-resolution satellite imagery datasets cover only a few cities and are for specific tasks. Compared to close-range imagery datasets, satellite datasets are still very limited in the number of images and their diversity. To expand a dataset or to create a small dataset for transfer learning, researchers are faced with enormous financial costs in purchasing the imagery and conducting precise annotation. Hence, a *synthetic training dataset* is potentially an alternative source to assist model training, which

This work is licensed under a Creative Commons Attribution International 4.0 License.
SIGSPATIAL '21, November 2–5, 2021, Beijing, China
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8664-7/21/11.
<https://doi.org/10.1145/3474717.3483914>

would bring great value in avoiding large-scale data collection especially when acquisition and/or annotation conditions are difficult to be met.

Among different image generation techniques, Generative Adversarial Networks (GAN) [5] show their capability in generating visually realistic images at a relatively low cost. It has been used in many computer vision tasks, such as generating image/videos for visual entertainment, creating synthetic data for close-range dataset [1, 4, 8, 10, 11]. [1, 4, 10, 15] use GAN to generate new samples or translate images into another domain to avoid annotation or extend limited datasets. However, those studies were either implemented for classifications, which do not require the generated images have pixel-wise correspondence with the ground truth, or only transfer images to a specific domain, which limits the number of generated samples and styles. Hence, to generate meaningful datasets for improving dense remote sensing model (*i.e.*, road segmentation, building detection), we still face four critical challenges:

(1) For dense remote sensing model training, a generated *high-resolution* synthetic image needs to have strong *pixel-level correspondences* to its ground-truth mask. Methods [1, 4, 8, 11] for (i) low-resolution, and/or (iii) unsupervised or coarse-supervised augmentation, and/or (ii) classification dataset do not suit for this task.

(2) Satellite image usually contains enormous small scattered objects. A sparse ground-truth mask is labeled to indicate the object of interest. Creating proper content and minimizing artifacts for the majority of unlabeled sparse areas is a unique challenge for remote sensing dataset synthesis.

(3) A dataset can be enhanced at appearance style and scene structure two levels. We are missing an image generating tool that contains two orthogonal functionalities that enhance the appearance coverage of the existing dataset and create novel training pairs with unseen scene structures.

(4) Because of high ground-area coverage, satellite image usually contains a mixture of different types of ground objects (*i.e.*, Forest *etc.*). The majority of images cannot be labeled for one appearance category and it is difficult to group images into sub-categories for training. Hence we need a generating tool, which only learns from few images and is able to generate images for a specific style or mixture of representative styles.

In this paper, we propose conditional-SinGAN (cSinGAN) and GeoPalette for generating synthetic remote sensing training datasets by learning one or a few samples. Our experiments show that a synthetic image-assisted road segmentation model with 1,000 real samples can achieve comparable performance as a standard-trained model using 4,000 real samples.

2 APPROACH

Proposed Image Generators: Inspired by SinGAN [11], we design *conditional-SinGAN* (cSinGAN) to restrict the synthesized images to follow the desired scene structures described by the mask. Given a training pair, image (x) and mask (y), cSinGAN inherits the multi-scale generator-discriminator scheme from SinGAN to learn the training pair from a lower resolution gradually to the full resolution. The base scale generator of cSinGAN takes a noise input, a resized mask and an empty reference image to generate a low-resolution seed image. The upper-scale generators take a noise input, a rescaled

Figure 2: (a) cSinGAN and GeoPalette generator structure. (b) latent space division of categorical noise for GeoPalette.

mask and the image generated from the previous scale to generate a larger image with more details. In training, the generator (1) reconstructs the target image when a specific anchor noise z_{rec} is given and (2) generates a synthetic image when another noise z_{rand} is given.

Since only one training pair is used in training, the generator easily overfits to this mask and memorizes its image. As a result, at inference time, it generates high-fidelity but nearly identical images as the training image for the training mask and performs poorly for unseen masks that could not fill the sparse areas between road branches in the synthetic images. A diversity-sensitive (ds) loss [13] between the reconstructed image and generated image is added during training to force the generator to produce different synthetic images when the latent noise is disparate from the anchor noise. It significantly alleviates the overfitting problem. The generator structure of cSinGAN at each scale is shown in Figure 2(a). The discriminator uses the Wasserstein GAN with gradient penalty (wGAN-GP) [6]. The objective function of cSinGAN learning from a given training image (x) and mask (y) at each scale with generator (G) and wGAN-GP critic (f) is shown as:

$$\begin{aligned} \min_G \max_f L_{cSinGAN}(G, f) &= L_{GAN}(G, f) + \alpha L_2(x, G(y, z_{rec})) - L_{ds}(G, y, z) \\ \min_G \max_f L_{GAN}(G, f) &= f(x) - f(G(y, z_{rand})) + (\|\nabla_{\hat{x}} f(\hat{x})\|_2 - 1)^2 \\ \hat{x} &= \epsilon x + (1 - \epsilon) G(y, z_{rand}), \quad \epsilon \in (0, 1) \\ L_{ds}(G, y, z) &= clip \left(\beta \frac{L_2(G(y, z_{rec}), G(y, z_{rand}))}{\|z_{rec} - z_{rand}\|_2^2}, \lambda_{ds} \right) \end{aligned}$$

where $\beta = 1$, $\lambda_{ds} = 0.1$; $\alpha = 10$ as default, but may set to other values depending on the training stability.

cSinGAN is designed to generate synthetic images mimicking a specific appearance style. However, when multiple appearance style is required (*e.g.*, different background environments for road segmentation, Figure 1), multiple cSinGAN generators are required, one for each category. It can take tremendous effort to train and generate images from each generator, especially when many target appearance styles are needed. Additionally, as the majority of satellite imagery covers a mixture of ground objects, each image can form a unique style depends on its combination of objects. It is impossible to sample all possible appearance styles exhaustively. For certain tasks, selecting only a few representative styles and create a blended style is preferred. With this motivation, we created Multi-Category conditional SinGAN (MCat-cSinGAN), also

named as *GeoPalette*. It learns multiple appearance styles in one generator simultaneously and generates images with different appearance styles at inference time. To achieve this goal, GeoPalette disentangles the latent noise space into different sub-regions, which allows the generator to learn different representative appearance styles in their assigned noise regions, see Figure 2(b). For training, one training image-mask pair is required for each category. When learning the sample from category A, GeoPalette samples an input noise from the latent noise region assigned for category A. At inference time, based on the required appearance category, GeoPalette samples the latent noise from the desired region to guide the appearance of the synthetic image. Moreover, the latent space can be rearranged depending on the use cases.

Enhancement Strategy: To enhance the coverage of a dataset with respect to appearance and scene structure levels, we defined two enhancement strategies. *Self-augmentation*: given a ground truth mask from the existing real dataset, it generates synthetic images with different appearances while maintaining the scene structure of the ground truth; the synthetic images are paired with the real masks to form new training pairs; *Scene-creation*: given a ground truth outside the existing real dataset, it produces synthetic images, mimicking the learned appearance of the real images and representing the scene structure of the new ground truth; the synthetic images and the new masks are paired to create novel training pairs (Figure 1). We utilize the cGAN technique [9] and propose the abovementioned models. Independent of the choice of cGAN model, the workflow for training and inference is shown in Figure 1(c).

3 EXPERIMENTS

Dataset and Settings: The road segmentation datasets from the 2018 CVPR DeepGlobe Challenge [3] contains 6,226 annotated high-resolution satellite images ($1,024 \times 1,024$ pixels, 0.5 meter/pixel). Following [2], 1,530 samples are used for testing. To simulate the data starvation scenario, the rest are randomly split into: an “available” training set (*Train*) with 1,000 samples; an additional holdout set (*Add*) with 3,000 samples, from which the masks are used for scene-creation; and a validation set with 696 samples for segmentation model training. From the Train set, 7 representative style samples are selected: brownish city, whitish city, forest, green field, reddish field, waterbody and yellowish rural area (desert). For road segmentation, we use the DeepGlobe challenge winner D-linkNet [14] and trained two baseline B1 and B2. The bottom baseline (B1), trained with 1,000 real samples (Train), indicates the minimal performance of an assisted trained model that does not degrade the overall performance. The top baseline (B2), trained on 4,000 real samples (Train + Add), is taken as the highest performance can be reached by using 4,000 training samples. Following [3], mean Intersection over Unit (mIoU) is set as the primary metric. To compare with our generators, our enhanced pix2pix [8], namely pix2pix-wGAN is trained using the 1,000 real image pairs (Train). For this model, we add a noise input to generate different images for the same mask and replace the discriminator with wGAN-GP critic [6] to improve the training stability and the image quality for high-resolution image generation. We set the following training options for the experiment. *Enhancement strategy*: self-augmentation or scene-creation. *Real-to-fake-image ratio*: R1: 1:1 1,000 real, 1,000 fake, from the

Figure 3: Categorical images generated from cSinGAN and GeoPalette along with the training images.

same style; R2: 1:3 1,000 real, 3,000 fake from the same style; R3: 1:3 1,000 real, 3,000 fake image 1,000 each from 3 style (different noises for pix2pix-wGAN). *Generator*: a set of 7 cSinGANs; GeoPalette; pix2pix-wGAN. *Mixed training strategy*: standard training using mixed dataset; three-phase training adapting from [7].

Results and Discussion: Compared to pix2pix-variants, cSinGAN and GeoPalette generate images without strong artifacts in sparse areas (Figure 3). As cSinGAN generators are trained separately, the generated images for the 7 categories are strictly distinguishable. For some road masks, GeoPalette shows blended styles among the brownish city, whitish city and forest. This may be caused by the similarity among the 3 styles and the noises sampled near the boundary of latent space sub-regions. Although GeoPalette does not strictly follow the target style for some categories, we still can see significant variations among the synthetic samples from these categories for the same road mask. Depends on the specifications of the main task, readers may choose cSinGAN when few clear-cut styles are required, such as local transfer learning or imbalanced-style toning; or choose GeoPalette when the real dataset does not have clear style labels or contains a wide variety of mixture styles, which exhaustive sampling of all styles is not practical.

Table 1: Synthetic data assisted trained segmentation models performance compared to baselines.

Generator	Test mIoU	To B1	Test mIoU	To B1
Baseline B1	60.920%	-	-	-
Baseline B2	64.590%	3.670%	64.590%	3.670%
Strategy	Self-augmentation	Scene-creation		
Real:Fake Ratio 1:3 R3				
cSinGAN set	64.440%	3.520%	64.420%	3.500%
GeoPalette	64.320%	3.400%	64.160%	3.240%
pix2pix-wGAN	63.950%	3.030%	-	-
Real:Fake Ratio 1:3 R2				
cSinGAN set	-	-	64.060%	3.140%
GeoPalette	-	-	64.020%	3.100%
pix2pix-wGAN	-	-	63.900%	2.980%
Real:Fake Ratio 1:1 R1				
cSinGAN set	64.090%	3.170%	64.220%	3.300%
GeoPalette	63.940%	3.020%	64.380%	3.460%
pix2pix-wGAN	63.870%	2.950%	63.820%	2.900%

Baseline B1 and B2 obtained mIoU of 60.92% and 64.59%, which shows a 3.670% performance uplift brought by the additional 3,000 real training pairs. All of the standard-assisted-trained models show a decreased performance compared to B1. However, if the same synthetic dataset is trained with three-phase training, all models show improvements by at least 2.740% on mIoU. The best assisted trained model shows an improvement of 3.520%, which almost reaches B2. This indicates if the quality of the synthetic data does not reach the level of real images, it still can improve the performance of a model, when a proper mix training method is used.

For the self-augmentation, models trained with cSinGAN images provide the highest improvements, followed by GeoPalette, then pix2pix-wGAN (Table 1). By increasing the fake-image ratio, all assisted trained models obtain additional improvements, which are more significant for cSinGAN and GeoPalette, but marginal for pix2pix-wGAN. This is expected, as pix2pix-wGAN learns an averaged appearance style from the full dataset and does not give categorical variation. For the scene-creation strategy (Table 1), we observe a similar trend, with one exception of GeoPalette 1:1 outperforming the cSinGAN 1:1 assisted trained model. This exception might be a result of multi-categorical learning, which leads the GeoPalette to have a better and unbiased understanding between roads and background regardless of the appearance styles. It can be beneficial for assisted training when the generated images are not as real as the real images and the corresponding real images are not available for the model training (*i.e.*, scene-creation). When increasing the fake image ratio, R2 strategy gives decreased improvements for cSinGAN and GeoPalette and marginal additional improvements for pix2pix-wGAN. The decreased improvements might be caused by adding a lot of synthetic images in the same style, which misleads the model towards a specific background environment. On the other hand, R3 strategy, which adds 3 categorical variants for each unseen mask, has a better performance than R2 strategy. Both enhancement strategies improve the performance of the assisted trained model when the training dataset is limited. In general, we believe the self-augmentation strategy is easier to learn because

of the presence of the corresponding real images, especially when the synthetic images are not very real. Although scene-creation gives more freedom of adding additional road structures, depending on the use cases, readers need to balance the number of synthetic images added for each style.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we present GeoPalette and cSinGAN for generating synthetic remote sensing training dataset and demonstrate using synthetic data can significantly boost the model performance when the real dataset is limited. We believe our generators and this assisted training method can be applied to a broad range of segmentation-based tasks. For the future work, we work on synthetic image selection metrics to shortlist useful data for efficient assisted training.

ACKNOWLEDGMENTS

This work was funded by the Grab-NUS AI Lab, a joint collaboration between GrabTaxi Holdings Pte. Ltd. and National University of Singapore, and the Industrial Postgraduate Program (Grant: S18-1198-IPP-II) funded by the Economic Development Board of Singapore.

REFERENCES

- [1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* (2017).
- [2] Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, CV Jawahar, and Manohar Paluri. 2019. Improved road connectivity by joint learning of orientation and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10385–10393.
- [3] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 172–181.
- [4] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* 321 (2018), 321–331.
- [5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
- [6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5769–5779.
- [7] Hiroshi Inoue. 2018. Data Augmentation by Pairing Samples for Images Classification. *arXiv:1801.02929 [cs, stat]* (2018).
- [8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.
- [9] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
- [10] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific reports* 9, 1 (2019), 1–9.
- [11] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4570–4580.
- [12] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. 2019. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv:1807.01232* (2019).
- [13] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024* (2019).
- [14] Lichen Zhou, Chuang Zhang, and Ming Wu. 2018. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 182–186.
- [15] Zhengxia Zou, Tianyang Shi, Wenyuan Li, Zhou Zhang, and Zhenwei Shi. 2020. Do game data generalize well for remote sensing image segmentation? *Remote Sensing* 12, 2 (2020), 275.