This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SyntheWorld: A Large-Scale Synthetic Dataset for Land Cover Mapping and Building Change Detection

Jian Song[1,2], Hongruixuan Chen[1,2], and Naoto Yokoya[1,2]
[1]The University of Tokyo, Japan
[2]RIKEN AIP, Japan
song@ms.k.u-tokyo.ac.jp, Qschrx@gmail.com, yokoya@k.u-tokyo.ac.jp

Abstract

Synthetic datasets, recognized for their cost effectiveness, play a pivotal role in advancing computer vision tasks and techniques. However, when it comes to remote sensing image processing, the creation of synthetic datasets becomes challenging due to the demand for larger-scale and more diverse 3D models. This complexity is compounded by the difficulties associated with real remote sensing datasets, including limited data acquisition and high annotation costs, which amplifies the need for high-quality synthetic alternatives. To address this, we present SyntheWorld, a synthetic dataset unparalleled in quality, diversity, and scale. It includes 40,000 images with submeter-level pixels and fine-grained land cover annotations of eight categories, and it also provides 40,000 pairs of bitemporal image pairs with building change annotations for building change detection. We conduct experiments on multiple benchmark remote sensing datasets to verify the effectiveness of SyntheWorld and to investigate the conditions under which our synthetic data yield advantages. The dataset is available at https://github.com/JTRNEO/SyntheWorld.

1. Introduction

High-resolution remote sensing image processing is vital for urban planning, disaster response, and environmental monitoring. Although advances in deep neural networks and the emergence of various benchmark datasets have led to significant progress in these research areas, the unique aspects of remote sensing image processing tasks still present many challenges.

First, acquiring large-scale datasets that compare with those in computer vision and natural language processing is difficult due to the sensitivity, privacy, and commercial considerations of remote sensing data. As a result, remote sensing datasets tend to be significantly smaller. Second, compared to fields like computer vision or natural language processing, remote sensing data annotation is both more costly and time-intensive. For example, annotating a 1024×1024 image from a large land cover mapping dataset such as [40] usually takes more than two hours. Finally, variations in image capture conditions such as sensor type, image acquisition season, and geographical location introduce a severe domain shift problem in remote sensing image processing.

Synthetic datasets, with their low-cost acquisition, high fidelity, and diversity, present a viable solution to these challenges. In the field of computer vision, numerous high-quality synthetic datasets [4, 13, 22, 29, 33, 39] have already emerged, primarily serving tasks such as semantic segmentation, depth estimation, optical flow estimation, and 3D reconstruction of street-view and indoor-view scenario. However, high-quality synthetic datasets for remote sensing are scarce in comparison. The most important reason is, as described in [18], in a virtual world constructed for street-view or indoor-view scenes, the distance between the sensor and the target location is relatively small (a few or tens of meters), with the main focus being on pedestrians, vehicles, road signs, or various furniture, resulting in a relatively small virtual world size. In contrast, in remote sensing scenarios, sensors are often located tens of thousands of meters away from the target virtual world, making even a relatively small virtual world extend over several square kilometers, while maintaining a multitude of diverse targets, such as thousands of trees in different poses and hundreds of buildings of different styles. This makes the construction of large-scale synthetic remote sensing datasets exceptionally challenging.

Upon a thorough survey of the available synthetic remote sensing datasets [3, 18, 28, 35, 41, 43], we discern that each of them has specific limitations. First, most existing works focus on a single task, such as building segmentation [18, 43] or object detection [35, 41]. However, there is a notable lack of effective synthetic datasets for critical tasks like multi-class land cover mapping and building change detection. Furthermore, these datasets exhibit lim-

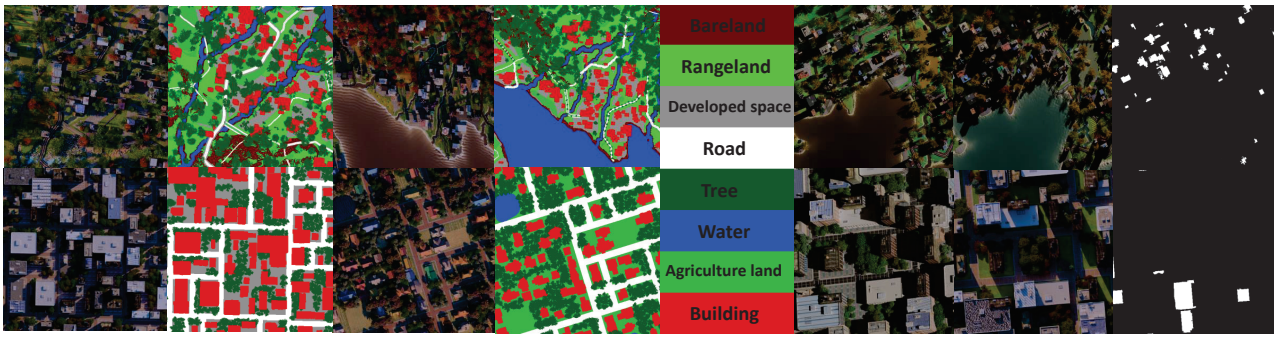


Figure 1. Examples of SyntheWorld dataset.

ited diversity due to constraints associated with the size of the virtual world and the tools used. They either emulate real-world cities to create a limited number of virtual environments or use real remote sensing images as the background. Furthermore, when it comes to 3D models in the virtual world, existing methodologies consistently rely on predefined textures, layouts, and geometries, resulting in a restrictive range of styles for buildings, trees, and other land objects.

In this work, we use the freely available open-source 3D modeling software Blender [7], along with various plugins from the Blender community, GPT-4 [26], and the Stable Diffusion model [31], to develop a procedural modeling system specifically for generating high-resolution remote sensing datasets. We present SyntheWorld, the largest high-resolution remote sensing image dataset for land cover mapping and building change detection tasks. Fig. 1 displays some examples from the proposed SyntheWorld dataset.

The main contributions of this work are:

- We introduce SyntheWorld, the first fully synthetic high-resolution remote sensing dataset, which integrates procedural 3D modeling techniques with Artificial Intelligence Generated Content (AIGC).
- We use SyntheWorld as the first synthetic dataset specifically designed to improve performance in two crucial tasks: multi-class land cover mapping and building change detection.
- We propose the Relative Distance Ratio (RDR), a new metric designed to quantify the conditions under which the synthetic dataset can drive performance improvements.
- Through comprehensive experiments on various remote sensing benchmark datasets, we demonstrate the utility and effectiveness of our dataset.

2. Related Works

2.1. Remote Sensing Image Processing Tasks

2.1.1 Land Cover Mapping

The discipline of land cover mapping is a crucial component of remote sensing image processing, where the goal is to categorize and depict physical features on Earth's surface, such as grass, trees, water bodies, bareland, buildings, etc. This task resembles semantic segmentation in traditional computer vision. Although the introduction of benchmark datasets for real-world scenarios, such as DeepGlobe [11], LoveDA [38], and OpenEarthMap (OEM) [40], has made significant advances in associated research, there is still a clear need for high-quality synthetic datasets. This is an area where the field of computer vision has made significant progress. Recognizing this gap, we were motivated to create SyntheWorld, a synthetic dataset crafted to improve performance in land cover mapping tasks.

2.1.2 Building Change Detection

The task of building change detection forms another crucial component within the realm of remote sensing image processing. It involves the identification and localization of modifications in man-made structures, especially buildings, over time, achieved through the analysis of images of the same area captured at different intervals. It is an indispensable technique for assessing damage in scenarios such as earthquakes, hurricanes, or floods, and for monitoring urban development and expansion over time. Typical annotations for this task involve binary masks, with networks trained to predict areas of building change based on input image pairs from two time points. While the emergence of benchmark real-world datasets such as WHU-CD [16], LEVIR-CD+ [5], and SECOND [42] have provided the field with valuable data resources, the lack of high-quality synthetic datasets has hindered the pace of related research.

2.2. Existing Synthetic Datasets

2.2.1 Street-view & Indoor-view

As we mentioned, the availability of large, high-quality synthetic datasets for street-view and indoor-view has driven the development of related techniques in traditional computer vision. The MPI Sintel Dataset [4] is widely used for training and evaluating optical flow algorithms, capturing natural scenes and motions in its synthetic dataset derived from an animated film. SceneFlow [22], with more than 35,000 synthetic stereo video sequences, is designed for the evaluation of optical flow, disparity, and scene flow algorithms. SYNTHIA [33], a dataset composed of 9,400 multi-viewpoint frames from a virtual city, targets urban scene understanding tasks with its pixel-level semantic annotations. The GTA5 dataset [29], comprising 24,966 synthetic images from the perspective of a car in virtual cities, is tailored to the understanding of urban scenes with its pixel-level semantic annotations compatible with the Cityscapes dataset [9]. Synscapes [39], featuring 25,000 photorealistic street scenes, aims to improve the performance of computer vision models in outdoor scenes with its precise semantic labels. Finally, SceneNet [13], a diverse synthetic dataset of over 5 million indoor scenes with RGB-D images and semantic labels, is designed for indoor scene understanding tasks.

2.2.2 Overhead-view

The AICD dataset [3], one of the earliest datasets with an overhead view, uses the Virtual Battle Station 2 game engine to simulate building alterations. Despite its 1,000 pairs of 800×600 RGB image pairs with building change masks, its 500 change instances are limited compared to the tens of thousands found in real-world datasets. The GTA-V-SID dataset [43], extracted from the GTA-V game, covers a $100km^2$ area with 121 500×500 aerial RGB images. Although it is useful for building segmentation tasks, its 1m GSD limits performance in high-resolution remote sensing datasets. Syntinel-1 [18], the first high-resolution synthetic remote sensing dataset for building segmentation, is based on CityEngine and offers a variety of urban styles. The Syntcities dataset [28] is for disparity estimation in remote sensing images, featuring three virtual cities and 8,100 pairs of high-resolution images. RarePlanes [35], a semi-synthetic dataset for aircraft object detection, combines real WorldView-3 satellite imagery and 3D models.

3. Dataset Generation and Description

Constructing a virtual city manually is time-consuming. Comparatively, SyntheWorld differs from existing overhead-view synthetic datasets by using procedural modeling. Previous studies in computer graphics have explored procedural modeling for cities and buildings [17, 24, 25], but none have utilized these techniques for the creation of overhead view datasets. We create our own procedural rules to create 3D geometries and apply textures derived from generative models, which minimize labor costs and enrich diversity.

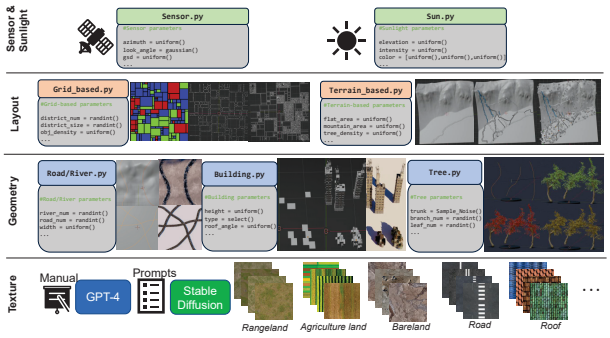


Figure 2. The essential components for building SyntheWorld dataset.

3.1. Generation Workflow

Layout. We adopt grid-based and terrain-based methods for the virtual world, as illustrated in Fig. 2. For the grid-based method, we randomly slice a grid of 0.25-0.36km^2 into several blocks of varying numbers and sizes, placing different types of buildings and trees in each block, and the boundaries between the blocks serve as our road system. It mainly simulates the more regular city and suburban layouts, and also contributes to the production of 0.3-0.6m GSD synthetic remote sensing images. For the terrain-based method, we use random noise textures to generate terrains such as mountains, plains, and oceans with ranges of 1-2km^2. Placing rivers, roads, buildings, and trees according to carefully designed rules based on *Geometry Nodes* in Blender [7], this method mimics irregular layouts in developing regions. It mainly contributes to the production of 0.6-1.0m GSD synthetic remote sensing images.

Geometry. The geometry row in Fig. 2 demonstrates our approach to procedurally model trees and buildings. For buildings, we use random noise to cut out differently shaped grids on a flat plane, which we then extrude into 3D geometries following pre-set rules. Users can control predefined parameters to generate an infinite number of different geometric styles. We distribute predefined asset components (walls, roofs, windows, etc.) to the geometry and finally map the texture generated by AIGC to the building. For trees, we use random-shaped curves as trunks and distribute different styles of tree components to the curve following certain rules.

Texture. The last row in Fig. 2 shows examples of our process for generating corresponding texture assets using

RS Synthetic Datasets	Features and Composition						
	GSD (m)	Task	# of images	Image Size	Automatic Labeling	Fully Synthetic	Procedural Modeling
AICD [3]	–	BCD	1,000 pairs	800×600	√	√	×
GTA-V-SID [43]	1	BS	121	500×500	×	√	×
Synthinel-1 [18]	0.3	BS	1,054	572×572	√	×	×
RarePlanes [35]	$0.31 \sim 0.39$	OD	50,000	512×512	√	×	×
SyntCities [28]	0.1, 0.3, 1.0	DE	8,100 pairs	1024×1024	√	×	×
SyntheWorld (Ours)	$0.3 \sim 0.6$ $0.6 \sim 1.0$	BS/LC/BCD	30,000 pairs 10,000 pairs	512×512 1024×1024	√	√	√

Table 1. Features and composition comparison among remote sensing synthetic datasets. LC: land cover mapping. BCD: building change detection. BS: building segmentation. OD: object detection. DE: disparity estimation.

AIGC. In terms of operational specifics, we first make a Stable Diffusion usage guide as a prompt to help GPT-4 understand its workings and prompt forms. We then provide excellent prompts as examples and ask GPT-4 to generate different themed prompts for different types of textures. In total, we generated around 140,000 seamless textures for different geometry to build SyntheWorld, far exceeding the number of textures used by existing overhead-view datasets. See the supplementary material for detailed prompts and generated images.

3.2. Structure of Dataset

As shown in Tab. 1, SyntheWorld is a comprehensive image dataset, consisting of 40,000 pairs of images. Of these, 30,000 pairs have a GSD ranging from 0.3 to 0.6 m, with each image having size 512×512. The remaining 10,000 pairs have a GSD of 0.6 to 1.0 m and a larger image size of 1024×1024.

Each pair in the dataset contains a post-event image, which is utilized for the land cover mapping task. These post-event images are accompanied by semantic labels of eight categories, as shown in Fig. 1. These categories are consistent with those of the OEM [40] dataset. Correspondingly, the pre-event images are derived by introducing variability in each scene. This involves different textures, lighting parameters, and camera settings. Additionally, there is a 10% to 50% chance that any given building in the scene might be removed.

Both pre-event and post-event images from each pair are used collectively for the building change detection task. Accordingly, the dataset comes with 40,000 binary classification masks corresponding to this task.

The off-nadir angle of all images ranges from $-25°$ to $25°$ and follows a Gaussian distribution with a certain mean $0°$ and variance $2.3°$. Similarly, we simulate the sun's position during the day in most countries by adjusting the zenith (ranging between $25°$ to $35°$) and the elevation parameters (ranging between $45°$ to $135°$), as guided by the documentation of the Pro Atom [8] addon in the Blender community, both parameters following a uniform distribution. This inclusion of various viewing angles and sun elevation enhances the robustness of SyntheWorld and ensures its applicability to a wide range of real-world datasets.

3.3. Comparison with Existing Synthetic Datasets

As depicted in Tab. 1, we provide a comparative analysis of SyntheWorld and existing synthetic remote sensing datasets, in terms of their features and composition. The Task column presents the primary tasks illustrated in the corresponding dataset's literature.

Regarding label generation, the GTA-V-SID dataset [43] consists of screenshots of the GTA-5 commercial video game, with buildings manually annotated. On the contrary, the remaining datasets are capable of automatically generating annotations via the corresponding 3D software.

In terms of complete synthesis, only SyntheWorld achieves this feat. The other datasets have adopted real remote sensing images to some extent as texture or as part of the dataset during their construction.

Finally, in SyntheWorld, most 3D models are generated using procedural modeling, while in other synthetic datasets, the geometric structure and texture of the models are either predefined or meticulously designed by 3D artists. This unique characteristic of SyntheWorld significantly enhances its diversity.

4. Experiments

4.1. Real-world Benchmark Datasets

To validate the versatility and effectiveness of SyntheWorld, we performed experiments using several high-resolution remote sensing datasets from various real-world scenarios. In the subsequent discussion, we present an in-depth overview of these datasets. In the experiments showcased in this section, we employ "w" to signify the utilization of the SyntheWorld dataset and "w/o" to indicate its non-use.

For the building segmentation task, we relied on OEM [40] and LoveDA [38] datasets, as well as INRIA [21] and BANDON [27] datasets. The INRIA dataset, which

Train on \ Test on	OEM*	LoveDA*	INRIA	BANDON
GTA-V-SID [43]	2.43	0.88	1.74	1.64
Synthinel-1 [18]	35.37	14.13	39.89	28.19
SyntCities [28]	23.61	21.39	30.39	30.01
SyntheWorld	**49.26**	**37.28**	**45.76**	**34.01**
OEM* [40]	80.48	55.35	75.61	64.19

Table 2. mIoU(%) results of the building segmentation task using DeepLabv3+. * means to use only the part of the building label in the dataset.

Datasets	w/o	w/
OEM [40]	**66.96**	66.84
LoveDA [38]	51.14	**53.32**
O→L	**35.28**	34.83
L→O	21.95	**25.24**

Table 3. Land cover mapping mIoU(%) outcomes from intra-dataset and cross-dataset evaluations, utilizing the DeepLabv3+ model for all experiments. O→L denotes training on the OEM training set and testing on the LoveDA validation set, while L→O represents the converse.

targets building footprint segmentation, incorporates aerial images from ten cities in the United States and Europe at a resolution of 0.3 m. The BANDON dataset stands out with significant off-nadir angles and focuses on urban areas with skyscrapers. It offers high-resolution 0.6m remote sensing images from Beijing and Shanghai.

We turned to OEM and LoveDA datasets again for the multi-class land cover mapping task. The OEM dataset, encompassing 97 regions across 44 countries worldwide, provides high-resolution images with detailed eight-class land cover annotations. The LoveDA dataset offers 0.3m GSD remote sensing images from three diverse regions in China, labeled with seven land cover categories.

In the building change detection task, we harnessed the WHU-CD [16], LEVIR-CD+ [5], and SECOND [42] datasets. The LEVIR-CD+ dataset consists of 987 image pairs, with 637 pairs in the training set and 348 pairs in the test set. SECOND, a semantic change detection dataset, collects 4662 pairs of aerial images from various platforms and sensors across cities like Hangzhou, Chengdu, and Shanghai. The WHU-CD dataset consists of two pairs of super-high-resolution (0.075m) aerial images. We cropped these large training (21243×15354) and testing (11265×15354) images into non-overlapping 512×512 patches for our experiments.

4.2. Building Segmentation

To compare with existing overhead-view synthetic datasets, which mainly include semantic labels of buildings, we performed building segmentation experiments. We use the DeepLabv3 + [6] network equipped with ResNet-50 [14] backbone. We adopted the SGD optimizer [30] for all synthetic datasets, employing a learning rate of 1e-3, a weight decay of 5e-4, and a momentum of 0.9; for the OEM dataset, we opted for a higher learning rate of 1e-2.

The results are presented in Tab. 2. The GTA-V-SID [43] dataset underperforms on various high-resolution real-world datasets due to its smaller quantity and 1m GSD. The model trained on the SyntheWorld dataset outperforms other datasets on four real-world datasets, especially on the OEM and LoveDA datasets. These two datasets include a considerable number of buildings in developing or developed areas. Thus, the performance of SyntheWorld far exceeds that of other competitors in these two datasets. As the buildings in the INRIA [21] and BANDON [27] datasets are predominantly high-rises in urban areas or well-organized detached houses in suburban areas, the advantage of the SyntheWorld dataset is not as evident as in the other two datasets, but still shows the best performance. Furthermore, the last column of Tab. 2 shows the performance of the model trained on the OEM dataset and tested on other datasets. Although SyntheWorld significantly outperforms other synthetic datasets, there is still a gap compared to real-world datasets.

In Fig. 3 (a), we also visualized the feature extract of the well-trained ResNet-50 of all synthetic and real datasets using UMAP [23]. In terms of feature space, SyntheWorld is closer to real-world datasets than any existing synthetic datasets.

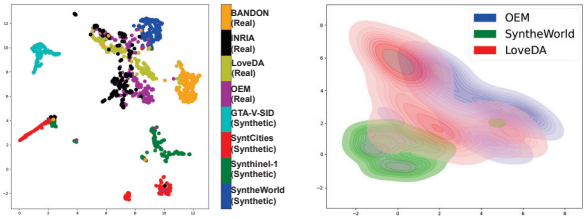
4.3. Land Cover Mapping

SyntheWorld is the first synthetic dataset that offers consistent annotations compatible with high-resolution real-world benchmarks. In this section, we primarily discuss the performance of SyntheWorld in the land cover mapping task.

4.3.1 Cross-dataset Experiments

To evaluate the enhancements brought about by using SyntheWorld, we adopted the mixed training strategy [29] often used with synthetic datasets, a batch size of 8, including 7 real images and 1 synthetic image per batch. The model was trained using DeepLabv3+ with the SGD optimizer and an initial learning rate of 1e-2, accompanied by a weight decay of 5e-4, and a momentum of 0.9. All experiments were trained for 100 epochs on a Tesla A100 GPU. Specifically, we map the rangeland class and the developed space class in OEM and SyntheWorld to the background class in LoveDA to keep the classes consistent.

Tab. 3 outlines the results obtained by integrating training images from a real-world dataset with SyntheWorld and the results of cross-dataset tests using the SyntheWorld dataset. Incorporating SyntheWorld with the entire



(a) (b)

Figure 3. (a) 2D UMAP visualization of synthetic and real datasets. We use ResNet-50 pre-trained on the OEM dataset as the feature extractor; (b) Colormap of density estimation for SyntheWorld, OEM, and LoveDA dataset.

Datasets	1%		5%		10%	
	w/o	w/	w/o	w/	w/o	w/
OEM [40]	40.9	**45.01**	52.21	**54.0**	58.40	**59.31**
LoveDA [38]	34.59	**36.75**	42.38	**44.58**	45.27	**48.12**

Table 4. mIoU(%) results from the DeepLabv3+ model, trained both with and without SyntheWorld, and deployed on two real-world land cover mapping datasets at various proportions of real image utilization.

OEM [40] training set does not result in performance enhancements. Similarly, combining SyntheWorld with the OEM training set and subsequently testing on LoveDA [38] slightly reduces model efficacy. However, when we merge SyntheWorld with the LoveDA and test on the same, the model's mIoU increases by 2.18 points. In addition, a 3.29-point improvement in mIoU is observed when testing the OEM test set after integrating SyntheWorld and LoveDA.

To investigate the observed phenomenon, we made density estimation maps for the three datasets as displayed in Fig. 3 (b). This reveals a notable overlap between SyntheWorld and OEM, with a lesser overlap in relation to LoveDA. The expansive coverage of the OEM dataset surpasses that of LoveDA and SyntheWorld. This finding sheds light on the patterns observed in Tab. 4. The vast diversity of the OEM dataset effectively captures the most data diversity inherent in SyntheWorld. Therefore, no performance enhancement results from integrating SyntheWorld. Nevertheless, the substantial overlap between SyntheWorld and OEM enables a performance boost when SyntheWorld is merged with LoveDA and tested on OEM. Conversely, the lesser overlap between SyntheWorld and LoveDA means that integrating SyntheWorld during OEM training does not lead to improvements in the LoveDA test set.

Subsequently, we assessed performance when integrating SyntheWorld with varying proportions of real-world datasets. Tab. 4 presents the findings. Irrespective of the real-world dataset being OEM or LoveDA, the integration

Train on \ Test on	Urban		Rural	
	w/o	w	w/o	w
Urban	47.00	**50.32**	33.44	**37.95**
Rural	36.86	**38.17**	48.64	**51.66**

Table 5. Land cover mapping results, measured in mIoU(%), from cross-domain experiments involving urban and rural areas of the LoveDA dataset.

of SyntheWorld consistently enhances model performance when the quantity of training data is limited.

4.3.2 Cross-domain Experiments

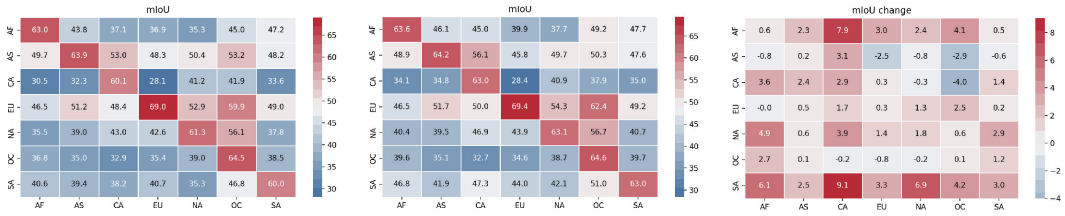
In order to examine the performance of SyntheWorld in out-of-domain test scenarios, we partition the OEM [40] dataset into seven distinct continents. Africa, Asia, Europe, Central America, North America, South America, and Oceania. Simultaneously, for the LoveDA [38] dataset, we conducted experiments using urban and rural areas as separate domains. We conduct experiments with various decoders and encoders; in this section, we show the results of one model. See supplementary material for more results from different models, dataset division, and experimental setup.

Continent-wise experimental results. Fig. 4 displays the results of cross-continent experiments in the OEM dataset using the U-Net [32] architecture with the EfficientNet-B4 [37] encoder. We can observe that our SyntheWorld dataset can significantly enhance performance across most dataset pairs. Also, we show in Fig. 5 the qualitative results when synthetic data can lead to a boost. More results can be found in the supplementary material. However, in some cases, the synthetic dataset does not yield a substantial improvement and could even degrade the model performance. It is crucial to investigate the reasons for such enhancement and impairment for the use of synthetic datasets. Therefore, we have conducted a further analysis of these results in Sec. 4.3.3.

Urban-Rural experimental results. We conducted similar cross-domain experiments on the LoveDA dataset, which includes two domains, rural and urban. The results are illustrated in Tab. 5. We found that the SyntheWorld dataset enhances model performance in both in-domain and out-of-domain tests.

4.3.3 Relative Distance Ratio

The cross-domain experiments discussed in Sec. 4.3.1 and Sec. 4.3.2 show that the SyntheWorld dataset does not always yield significant improvements. This highlights the need to understand the underlying causes. We introduce a metric, the Relative Distance Ratio (RDR), aiming to quantify the relationship between source, target, and synthetic



(a) (b) (c)

Figure 4. Results of continent-wise in-domain and out-of-domain land cover mapping experiments of OEM dataset. The x-axis represents the target domain and the y-axis represents the source domain. U-Net with EfficientNet-B4 encoder is used for all experiments. (a) The mIoU results of without using SyntheWorld; (b) The mIoU results of mixed training with SyntheWorld; (c) Changes in mIoU.

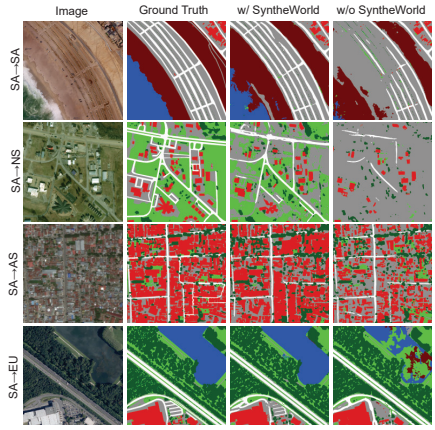


Figure 5. Qualitative results by U-Net model of continent-wise land cover mapping task.

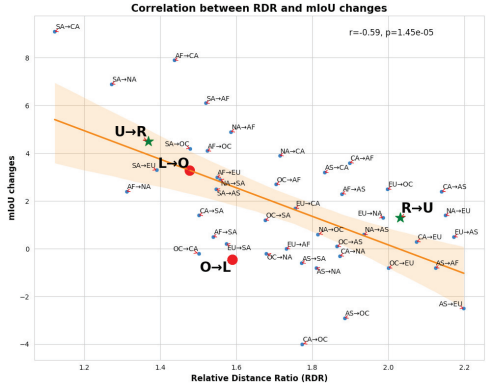


Figure 6. Scatter diagram with correlation between mIoU changes and proposed Relative Distance Ratio (RDR).

datasets and clarify when synthetic data can bring improvements.

For measuring the distance between datasets, various methods have been discussed in the literature [1, 15, 34]. The most commonly used measure of the distance between synthetic and real datasets is the FID score [15]. Here we adopt the Fréchet Distance, as the measure of distance between different datasets. Since the Inception model [36] pre-trained on ImageNet [12] is not suitable for remote sensing datasets, we use ResNet-50 [14] pre-trained on the OEM [40] dataset. The formula to compute the FD between any dataset pair is as follows:

$$FD(x,y) = ||\mu_x - \mu_y||^2 + \mathrm{Tr}(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{\frac{1}{2}}) \quad (1)$$

where μ_x and μ_y denote the mean feature vectors of datasets x and y, respectively, and Σ_x and Σ_y represent the covariance matrices of the corresponding feature vectors.

Then we denote the source domain dataset as S, the target domain dataset as T, SyntheWorld as G, and the FD between any two datasets as $\delta(.,.)$. Afterwards, the distance between the source domain dataset S and the target domain dataset T can be expressed as:

$$\delta(f_S, f_T) = FD(f_S, f_T) \quad (2)$$

Similarly, the distance between the target domain dataset T and the synthetic dataset G can be represented as:

$$\delta(f_T, f_G) = FD(f_T, f_G) \quad (3)$$

These f_T, f_S and f_G are obtained by applying a ResNet-50 model, pre-trained on the OEM dataset.

Subsequently, we can define the Relative Distance Ratio (RDR), denoted as $\mathcal{R}(f_S, f_T, f_G)$, to be calculated using the following formula:

$$\mathcal{R}(f_S, f_T, f_G) = \frac{\delta(f_T, f_G)}{\delta(f_S, f_T)} \quad (4)$$

Intuitively, a smaller $\mathcal{R}$ indicates a greater capacity of the model to integrate knowledge from the synthetic data and transfer it to the target domain. To validate this, we presented a correlation scatter plot in Fig. 6, which reveals a negative correlation between $\mathcal{R}$ and the improvement in mIoU. This observation aligns with our initial conception of designing the RDR metric. Therefore, the proposed RDR

Datasets	STANet-PAM		DTCDSCN		ChangeFormer	
	w/o	w/	w/o	w/	w/o	w/
LEVIR-CD+ [5]	0.752	**0.782**	0.793	**0.812**	0.784	**0.835**
SECOND* [42]	0.713	**0.733**	0.712	**0.727**	0.723	**0.734**
WHU-CD [16]	0.707	**0.802**	0.769	**0.862**	0.783	**0.836**

Table 6. F1 score resulting from the use or non-use of SyntheWorld across three building change detection benchmark datasets, assessed with three different models.

Datasets	1%		5%		10%	
	w/o	w/	w/o	w/	w/o	w/
LEVIR-CD+ [5]	0.517	**0.646**	0.636	**0.731**	0.726	**0.764**
SECOND* [42]	0.401	**0.435**	0.546	**0.622**	0.583	**0.631**
WHU-CD [16]	0.242	**0.312**	0.433	**0.638**	0.510	**0.705**

Table 7. Comparison of F1 scores from the DTCDSCN model trained with and without SyntheWorld, applied on three different real-world datasets at varying ratios of real image use.

metric effectively serves as a quantitative conditional criterion for employing synthetic data, that is, when $\mathcal{R}$ is large, there is a risk of using synthetic data and vice versa.

4.4. Building Change Detection

In this section, we demonstrate the effectiveness of SyntheWorld on the building change detection task. We employ four prevalent building change detection networks, FC-siam-Diff [10], STANet-PAM [5], DTCDSCN [19], and ChangeFormer [2]. We adhere to a mixed training strategy that includes a 7:1 real-to-synthetic image ratio. For ChangeFormer and DTCDSCN we use AdamW [20] optimizer with learning rate 1e-4, for the other two models we use Adam optimizer with learning rate 1e-3. Each mixed training experiment is trained for 100 epochs on the Tesla A100 GPU.

Tab. 6 presents the F1 score of three different models applied to three different datasets in the real world. Evidently, for each real-world dataset and each model type, integrating the SyntheWorld dataset induces an improvement, notably for the WHU-CD dataset where it can induce almost a 10-point increase in the F1 score when using the STANet-PAM and DTCDSCN models. Also, we display in Fig. 7 the qualitative results when using SyntheWorld can lead to enhancements. More results can be found in the supplementary material.

Tab. 7 reveals the F1 score of the DTCDSCN model with different proportions of the real-world training set, with and without the incorporation of SyntheWorld. Across all real-world datasets, SyntheWorld invariably provides substantial performance improvement when training data is scarce.

Tab. 8 illustrates the generalizability of the SyntheWorld dataset with the FC-siam-Diff model. We draw comparisons with three real datasets and the AICD synthetic dataset. The results show that by using only the SyntheWorld dataset for training, we can achieve acceptable results on all three datasets. Specifically, compared to the AICD [3] dataset, ours has a significant performance and generalization advantage.

Train on \ Test on	LEVIR-CD+	SECOND*	WHU-CD
LEVIR-CD+ (Real)	0.751	0.180	0.614
SECOND* (Real)	0.405	0.614	0.522
WHU-CD (Real)	0.222	0.248	0.812
AICD (Synthetic)	0.094	0.267	0.092
SyntheWorld	**0.419**	**0.386**	**0.457**

Table 8. Evaluation of generalizability across multiple building change detection datasets. The table shows the F1 scores. * means to use the part of building change label in SECOND.

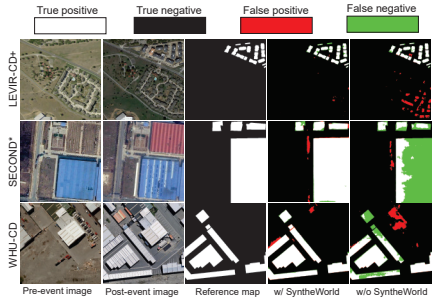


Figure 7. Qualitative results by DTCDSCN model of building change detection task on three datasets.

5. Discussion and Societal Impacts

We introduced SyntheWorld, the most extensive synthetic remote sensing dataset, used for land cover mapping and building change detection. Its diversity, enhanced by procedural modeling and AIGC, sets it apart from other datasets. Comprehensive experiments validate SyntheWorld's utility and flexibility. Furthermore, we investigate scenarios where SyntheWorld does not enhance performance, proposing the RDR metric for initial exploration of when SyntheWorld can deliver lift.

Notably, SyntheWorld has a significant gap compared to real datasets. This stems from some modeling rules mismatching real-world distributions, a challenge we aim to address in future work. Additional future work involves leveraging SyntheWorld to explore domain adaptation and generalization techniques in remote sensing.

Acknowledgements

This work was supported in part by JST FOREST Grant Number JPMJFR206S; Microsoft Research Asia; and the GSFS Challenging New Area Doctoral Research Grant (Project No. C2303).

References

[1] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020. 7

[2] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 8

[3] Nicolas Bourdis, Denis Marraud, and Hichem Sahbi. Constrained optical flow for aerial image change detection. In *2011 IEEE international geoscience and remote sensing symposium*, pages 4176–4179. IEEE, 2011. 1, 3, 4, 8

[4] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 1, 3

[5] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. 2, 5, 8

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5

[7] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 3

[8] Contrastrender. Pro atmo: User documentation. Blender Addons, 2022. Accessed: 2023-03-24. 4

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[10] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018. 8

[11] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[13] Ankur Handa, Viorica Pătrăucean, Simon Stent, and Roberto Cipolla. Scenenet: An annotated model generator for indoor scene understanding. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5737–5743. IEEE, 2016. 1, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 7

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[16] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018. 2, 5, 8

[17] Joon-Seok Kim, Hamdi Kavak, and Andrew Crooks. Procedural city generation beyond game development. *SIGSPATIAL Special*, 10(2):34–41, 2018. 3

[18] Fanjie Kong, Bohao Huang, Kyle Bradbury, and Jordan Malof. The synthinel-1 dataset: A collection of high resolution synthetic overhead imagery for building segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1814–1823, 2020. 1, 3, 4, 5

[19] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2020. 8

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8

[21] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 4, 5

[22] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1, 3

[23] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 5

[24] Pascal Müller, Peter Wonka, Simon Haegler, Andreas Ulmer, and Luc Van Gool. Procedural modeling of buildings. In *ACM SIGGRAPH 2006 Papers*, pages 614–623. 2006. 3

[25] F Kenton Musgrave, Craig E Kolb, and Robert S Mace. The synthesis and rendering of eroded fractal terrains. *ACM Siggraph Computer Graphics*, 23(3):41–50, 1989. 3

[26] OpenAI. Gpt-4 technical report, 2023. 2

[27] Chao Pang, Jiang Wu, Jian Ding, Can Song, and Gui-Song Xia. Detecting building changes with off-nadir aerial images. *Science China Information Sciences*, 66(4):1–15, 2023. 4, 5

[28] Mario Fuentes Reyes, Pablo d'Angelo, and Friedrich Fraundorfer. Syntcities: A large synthetic remote sensing dataset for disparity estimation. *IEEE Journal of Selected Topics in*

Applied Earth Observations and Remote Sensing, 15:10087–10098, 2022. 1, 3, 4, 5

[29] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 1, 3, 5

[30] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 5

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[33] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1, 3

[34] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 7

[35] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 207–217, 2021. 1, 3, 4

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[38] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2, 4, 5, 6

[39] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 1, 3

[40] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 1, 2, 4, 5, 6, 7

[41] Yang Xu, Bohao Huang, Xiong Luo, Kyle Bradbury, and Jordan M Malof. Simpl: Generating synthetic overhead imagery to address custom zero-shot and few-shot detection problems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4386–4396, 2022. 1

[42] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks. *arXiv preprint arXiv:2010.05687*, 2020. 2, 5, 8

[43] Zhengxia Zou, Tianyang Shi, Wenyuan Li, Zhou Zhang, and Zhenwei Shi. Do game data generalize well for remote sensing image segmentation? *Remote Sensing*, 12(2):275, 2020. 1, 3, 4, 5