Scalable Multi-Temporal Remote Sensing Change Data Generation via Simulating Stochastic Change Process Supplementary Materials

Zhuo Zheng[1,2], Shiqi Tian[1], Ailong Ma[1], Liangpei Zhang[1], Yanfei Zhong[1,*]
[1]Wuhan University [2]Stanford University

https://github.com/Z-Zheng/Changen

A. Implementation details

A.1. Dataset description

xView2 dataset [5] is used to benchmark the models for one-to-many semantic change detection [10] in the context of the sudden-onset natural disasters. There are six diaster types of earthquake, wildfire, volcano, storm, flooding, and tsunami in the dataset. This dataset contains 9,168 image pairs of `train&tier3` split, 933 image pairs of `test` split, and 933 image pairs of `holdout` split, covering 45,361.79 km^2 areas. Each optical RGB image has a fixed size of 1,024×1,024 pixels. The images were collected from WorldView-2, WorldView-3, and GeoEye satellites, with varying sub-meter spatial resolutions. The total of building instances is 850,736.

xView2 pre-diaster, used in this paper, is the pre-disaster part of xView2 dataset.

LEVIR-CD dataset [2] consists of 637 bitemporal image pairs, which were collected from the Google Earth platform. Each image has a fixed size of 1,024×1,024 pixels, with a spatial resolution of 0.5 m. This dataset provides a total of 31,333 change (building appearing, building disappearing) labels of building instances, but without semantic segmentation masks. LEVIR-CD dataset is officially split into `train`, `val`, and `test`, three parts of which include 445, 64, and 128 pairs, respectively.

WHU-CD dataset [6] consists of two aerial images collected in 2012 and 2016, which contains 12,796 and 16,077 building instances respectively. Each image has a fixed size of 15,354×32,507 pixels with a spatial resolution of 0.2 m. The change type is mainly building construction. The experiment (Table 4 in the main text) requires an official train/val/test split of the dataset, whereas WHU-CD has no such one. Thus, we directly use the entire WHU-CD for zero-shot evaluation to avoid debate.

S2Looking dataset [9] contains 5,000 image pairs with spatial resolutions from 0.5 to 0.8 m and 65,920 change instances. The official `train`, `val`, and `test` splits include 3,500, 500, and 1,000 pairs, respectively. The images were collected from GaoFen, SuperView, and BeiJing-2 satellites of China, which mainly covered globally distributed rural areas. This dataset features side-looking satellite images, which pose a special yet important challenge that requires the change detector to have sufficient robustness to the registration error and the object geometric offset caused by off-nadir imaging angles. Each image of this dataset has a fixed size of 1,024×1,024 pixels.

A.2. Implementation details for fine-tuning

Fine-tuning on LEVIR-CD. Random flip, rotate, scale jitter, and cropping into 512×512 are used for training data augmentation. SGD is used as our optimizer, where the weight decay is 0.0001 and the momentum is 0.9. The total batch size is 16 and an initially learning rate is 0.03. We train for 200 epochs on `train` split, as common practices. A "poly" learning rate policy ($\gamma = 0.9$) is applied.

Fine-tuning on S2Looking. Random flip, rotate, scale jitter, and cropping into 512×512 are used for training data augmentation. SGD is used as our optimizer, where the weight decay is 0.0001 and the momentum is 0.9. The total batch size is 16 and an initially learning rate is 0.03. We train for 60k iterations on `train` split. A "poly" learning rate policy ($\gamma = 0.9$) is applied.

Evaluation metrics. F_1 score, precision rate (Prec.), and recall rate (Rec.) of change regions are used as evaluation metrics, where F_1 score is the main metric.

B. Scalability of Changen

B.1. Scaling up Resolution

Remote sensing images are always of big spatial resolutions beyond 256×256 due to the imaging with high altitudes, *e.g.*, satellite imaging. Therefore, there is a important requirement that the model trained with 256×256 images can be seamlessly applied to the larger image, *e.g.*, 1024×1024. It is easy for discriminative fully convolutional network, however, it is non-trivial for generative models to bridge the resolution gap. Fig. 1 shows the visual results. OASIS and Changen are both trained with 256×256 images. OASIS failed to generate realistic image when given

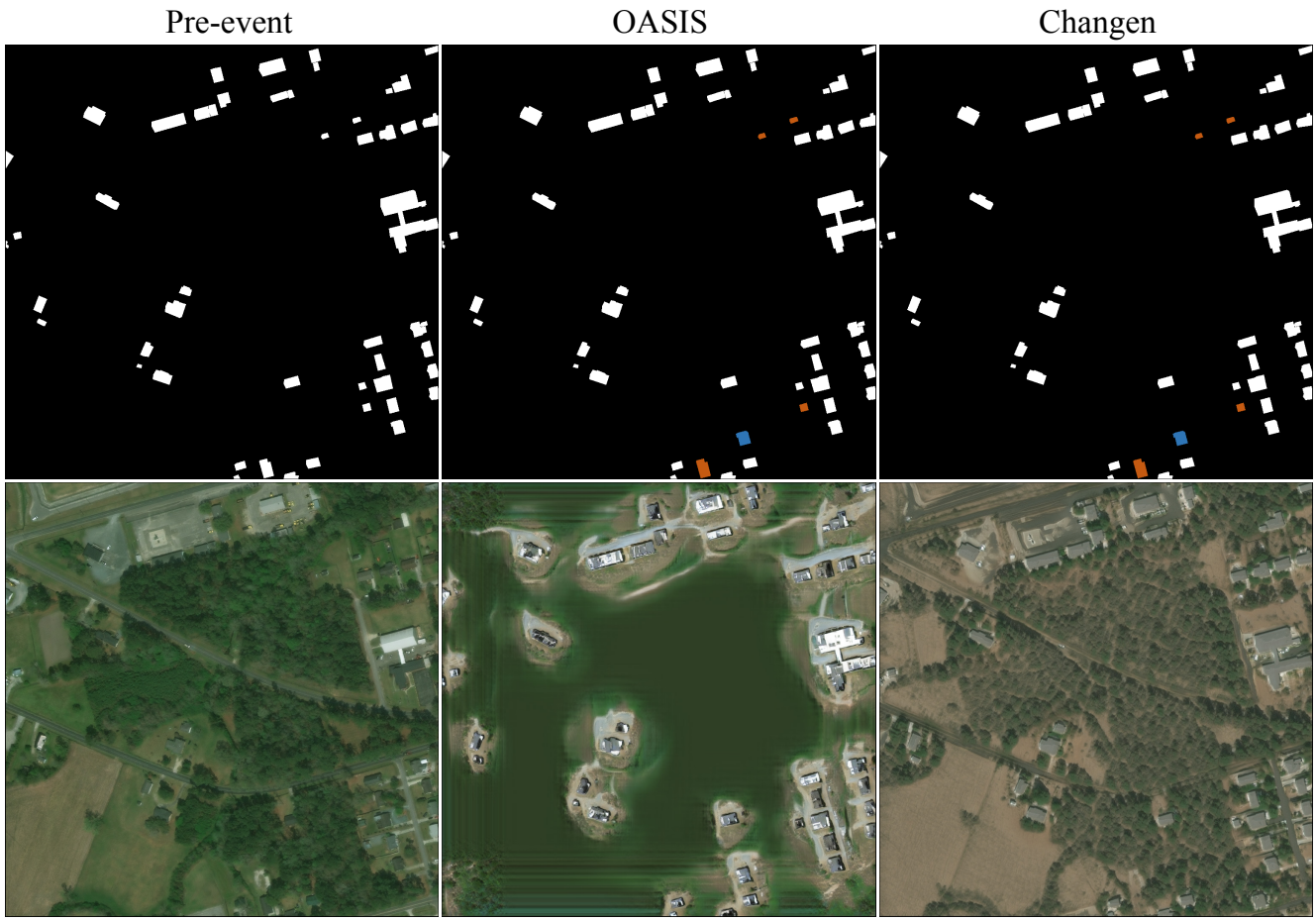

Figure 1. Scaling up the resolution to 1024×1024 pixels.

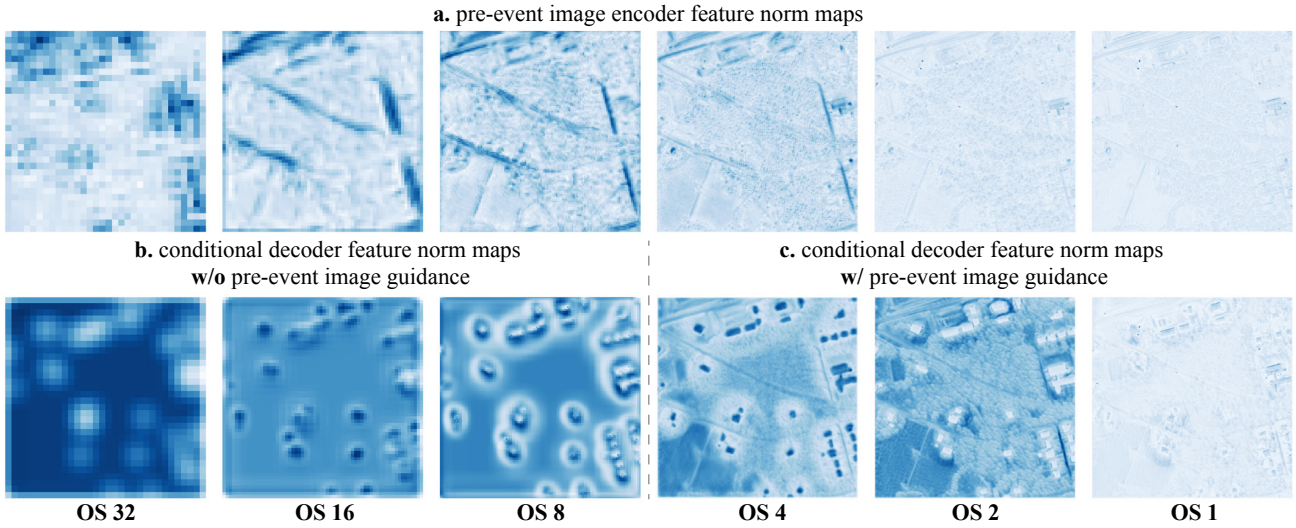

Figure 2. Feature ℓ_2 norm maps. **a**. the feature norm maps are computed over the pre-event image with Image Encoder of Changen. **b**. lower-resolution feature norm maps (OS 32, 16, 8) are computed without pre-event image guidance. **c**. higher-resolution feature norm maps (OS 4, 2, 1) are computed with pre-event image guidance. "OS": output stride.

a 1,024×1,024 semantic mask. Obvious artifacts are observed in background region. Changen still works with this 1024×1024 semantic mask, bridging the resolution gap.

We argue that the main reason why Changen works lies in the pre-event image guidance. To support our view, we visualize the feature map of each scale via ℓ_2 norm, as shown in Fig. 2. We train a variant of Changen, which removes pre-event image guidance in the first three scales (OS 32, 16, 8) and keep pre-event image guidance in the last three scales (OS 4, 2, 1), to investigate the impact of pre-event image guidance. From Fig. 2b, we can observe that the feature norm map of OS 8 is very similar to the image generated by OASIS in Fig. 1, from the perspective of the background smoothness. Once applying the pre-event image guidance, observed from Fig. 2c, the feature norm maps look to have more details. This visual evidence suggests that the pre-event image guidance is the key factor in bridging the resolution gap.

B.2. Scaling up Synthetic Data

We further scale up the synthetic data from 90k to 1.4M, namely Changen-1.4M, to verify whether data scaling can improve the model performance. As shown in Table 1, our synthetic data volume is at the leading edge. Scaling up Changen-90k to Changen-1.4M, ChangeStar with MiT-B1 further obtains 0.2% F_1 improvement, achieving 91.7% F_1 on LEVIR-CD. This suggests that data scaling can improve the model performance.

Table 1. Comparison with other synthetic change datasets

Dataset name	Image size (pixels)	#Image pairs
AICD [1]	800×600	1k
SynCW [7]	3,072×3,072	4
Changen-90k (ours)	256×256	90k
Changen-1.4M (ours)	256×256	1.4M

Table 2. Data scaling results on LEVIR-CD$^{\texttt{test}}$.

Method	Pre-train from	Backbone	F_1 ↑	#Params.	#Flops
ChangeStar (1×96)	ImageNet-1K	MiT-B1	90.0	18.4M	16.0G
Ours	`Changen-90k`	MiT-B1	91.5(↑1.5)		+0
Ours	`Changen-1.4M`	MiT-B1	91.7(↑1.7)		+0

C. Comparison with other Pre-training Methods

Our Changen is a generative model, which is capable of synthesizing multi-temporal change data from *single-temporal segmentation data*. With synthetic change data (e.g., `Changen-90k`) pre-training, the change detector gains more on the performance, compared to the commonly used ImageNet-1k supervised pre-training. Here we investigate the essential effect of Changen pre-training. The potential performance gain may come from (1) less domain gap between pre-train data and downstream data; (2) semantic segmentation supervision; (3) zero pretext task gap. (4) higher-quality synthetic change data. We discuss these factors next.

Table 3. Comparison with other pre-training methods on LEVIR-CD$^{\texttt{test}}$. All entries use ResNet-18 as the backbone. "xView2 pre.": xView2 pre-diaster dataset.

Method type	Pre-train method	Pre-train data	F_1 ↑
(a) ChangeStar (1×96)	classification	ImageNet-1K	90.5
(b) + self-supervised	SeCo [8]	SeCo-1M w/o label	89.9
(c) + self-supervised	MoCov2 [3]	xView2 pre.w/o label	90.4
(d) + seg. supervised	segmentation	xView2 pre.	90.6
(e) + synthetic data	change detection	`OASIS-90k`	90.6
(f) + Ours	change detection	`Changen-90k`	91.1

Factor 1: *less domain gap between pre-train data and downstream data.* The ImageNet-1k belongs to the natural scenario, which has a large domain gap with the Earth observation scenario. Can any Earth observation data reduce the domain gap? The result of Table 3(b) gives a negative answer. With a domain-specific self-supervised method (i.e., SeCo [8]) and 1 million Sentinel-2 images [4], the transferred change detection performance is instead reduced by 0.6%F_1, compared to ImageNet pre-training. This is because the spatial resolution of Sentinel-2 optical band is 10 m, while the images of LEVIR-CD has 0.5 m spatial resolution. The resolution gap is a massive barrier to transfer learning, although the scenario gap has been reduced.

xView2 pre-disaster dataset has sub-meter spatial resolutions close to LEVIR-CD. However, SeCo requires multi-temporal images as the pre-train data, while xView2 pre-disaster dataset can not meet that since it is single-temporal data. Thus, we use MoCo v2 [3] (the baseline of SeCo) to pre-train the backbone on the xView2 pre-disaster dataset, which yields 90.4% F_1, as Table 3(c) presents. This result somewhat bridges the resolution gap but is still inferior to ImageNet pre-training.

Overall, less domain (e.g., scenario, resolution) gap between pre-train data and downstream data is helpful to transfer the model to the downstream task. However, it is not a primary gain source of our Changen pre-training.

Factor 2: *semantic segmentation supervision.* In this case, Changen is trained using the xView2 pre-disaster dataset, which is a single-temporal building segmentation dataset. Therefore, the semantic segmentation supervision provided by this dataset may be a gain source. We use this segmentation dataset to pre-train the segmentation part of ChangeStar(1×96)[1]. This entry yields 90.6% F_1, outperforming ImageNet pre-training by 0.1% point, as Table 3(d)

[1]To whom is not familiar with ChangeStar, ChangeStar can be seen as a segmentation model with a simple change detection head.

presents. This result suggests that the semantic segmentation supervision is helpful but it is not a primary gain source of our Changen pre-training.

Factor 3: *zero pretext task gap.* Our Changen pre-training belongs to synthetic change data pre-training, and pretext task is exactly the change detection. Therefore, Changen pre-training has zero pretext task gap, which is fundamentally different from self-supervised pre-training, ImageNet pre-training, and segmentation pre-training. To investigate this factor, we use OASIS, our baseline of the generative model, to synthesize a multi-temporal change dataset (`OASIS-90k`) as Changen did. In this way, OASIS pre-training has also zero pretext task gap, which yields 90.6% F_1, as Table 3(e) presents. This result suggests that zero pretext task gap is helpful but it is not a primary gain source of our Changen pre-training.

Factor 4: *higher-quality synthetic change data.* After ablating other three factors, the last factor matters. Comparing Table 3(e) and (f), all variables are strictly controlled except the synthetic change data from two different generative models, i.e., OASIS and Changen. Therefore, we argue that the higher quality of the synthetic change data is a primary gain source of our Changen pre-training. Here, higher quality means better fidelity and diversity of generated images, which is measured by FID (45.13 *vs.* 34.74_{ours}, lower is better) and IS (4.95 *vs.* 5.41_{ours}, higher is better).

In summary, we find that synthetic change data pre-training is also a promising approach for remote sensing change detection. This pre-training features less domain gap and zero pretext task gap, where the transferability of pre-trained representation highly depends on the fidelity and diversity of generated images. Zero pretext task gap means that the model pre-trained in this way has zero-shot prediction capability, which other pre-training methods (above discussed) cannot achieve.

References

[1] 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.

[2] 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.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

[5] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

[6] 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.

[7] Maria Kolos, Anton Marin, Alexey Artemov, and Evgeny Burnaev. Procedural synthesis of remote sensing images for robust change detection with neural networks. In *International Symposium on Neural Networks*, pages 371–387. Springer, 2019.

[8] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, pages 9414–9423, 2021.

[9] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021.

[10] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265:112636, 2021.