Self-Pair: Synthesizing Changes from Single Source for Object Change Detection in Remote Sensing Imagery

Minseok Seo, Hakjin Lee, Yongjin Jeon, Junghoon Seo

SI Analytics

70, Yuseong-daero 1689beon-gil, Yuseong-gu, Daejeon, Republic of Korea

{minseok.seo, hakjinlee, yongjin117, jhseo}@si-analytics.ai

Abstract

For change detection in remote sensing, constructing a training dataset for deep learning models is difficult due to the requirements of bi-temporal supervision. To overcome this issue, single-temporal supervision which treats change labels as the difference of two semantic masks has been proposed. This novel method trains a change detector using two spatially unrelated images with corresponding semantic labels such as building. However, training on unpaired datasets could confuse the change detector in the case of pixels that are labeled unchanged but are visually significantly different. In order to maintain the visual similarity in unchanged area, in this paper, we emphasize that the change originates from the source image and show that manipulating the source image as an after-image is crucial to the performance of change detection. Extensive experiments demonstrate the importance of maintaining visual information between pre- and post-event images, and our method outperforms existing methods based on single-temporal supervision.

1. Introduction

Change detection aims to detect the location of interest regions among semantically changed areas. Generally, this change of interest (CoI) between multi-temporal high spatial resolution (HSR) remote sensing images is defined in the same area but at different times.

Recently, several supervised change detection methods [4, 9, 15, 3] have been proposed and showed promising results. Those methods are trained on the datasets consisting of pairs of bi-temporal images with change labels. However, these bi-temporal datasets require high cost to obtain compared to the other tasks such as segmentation [27] and detection [28]. It is due to several requirements for bi-temporal supervision. First of all, obtaining correctly registered bi-temporal pair images is difficult due to the physical
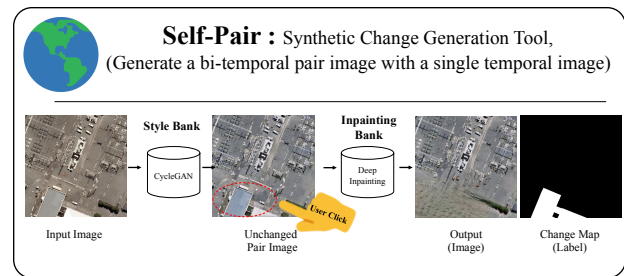

Figure 1. Synthetic change generation tool example.

limitations of satellites. Second, to decide whether the specific region is a change or not, observing both before- and after-image should be preceded. Lastly, change is rare even in the real world; which makes it hard to obtain bi-temporal pair images containing change of interests. For these reasons, publicly [1, 7, 5, 18] opened change detection datasets are small-scaled and imbalanced.

To solve the problem, Zheng *et al.* [33] proposed a method to train a change detection model using only single-temporal labeled unpair images. Instead of using bi-temporal labeled pair images, it trains the change detection model using a training dataset consisting of unpaired images which are randomly sampled from the training set. The change label between two unpaired images is defined as whether the semantics of the same pixel location in two images are different or not (i.e., xor operation). This approach enables training change detectors without high-cost bi-temporal pairwise datasets, however, it ignores the structure and style consistency in unchanged areas. In this setting, for example, buildings in paired images but with different color or texture are labeled unchanged. Also different semantics such as road and grassland are also labeled unchanged regions since they are not change of interest. This inconsistency makes the model confused to learn what the change is. Furthermore, unpaired setting ignores useful context information such as object sizes, common patterns, or styles of the area, which can be obtained from bi-

temporal paired images.

In this paper, we propose *Self-Pair*, a novel synthetic image generation method of constructing input pairs for change detection models. The key idea of *Self-Pair* is that more diverse and realistic pair images can be generated with single-temporal single images, retaining the characteristics of real-world settings. *Self-Pair* relieves complicated conditions of change detection datasets such as high cost labeling, registration, and preserves the characteristics of unchanged regions as shown in the Fig. 1. We verified the performance of the existing change detection methods and *Self-Pair* in both in-domain and cross-domain settings. On extensive benchmarks, our approach outperforms previous single-temporal supervision method [33], and even bi-temporal supervision methods in some cases. Moreover, by conducted experiments on SNUNet-CD [9], BIT-CD [4], and ChangeStar [33] architectures, our method shows applicability to the various change detection architectures.

2. Related Works

2.1. Object Segmentation in Remote Sensing

Semantic segmentation in remote sensing images is challenging because of significant scale variation, background complexity, and imbalance between background and foreground. To address these challenges, Zheng *et al.* [35] proposed *FarSeg* in the perspective of foreground modeling. They achieved high performance with a better trade-off between speed and accuracy compared with general semantic segmentation methods [6, 32]. Li *et al.* [17] achieved state-of-the-art through affinity context modeling, which focuses on solving the background complexity and background-foreground imbalance problem.

2.2. Object Change Detection in Remote Sensing

Change detection has been studied along with the rising need of utilizing remote sensing images to find meaningful changes [23, 14]. As deep learning progresses, the change detection methods based on deep learning also show promising performance [20, 34, 9, 4]. However, these methods require well-defined datasets with bi-temporal supervision, and most public datasets are small-scale, hence these methods show poor performance on the real-world cases [22]. The major reason for this phenomenon is that changes are rare compared to non-changes in bi-temporal paired images [33], and there is a difficulty in collecting bi-temporal paired images. For this reason, the progress of change detection was relatively slow compared with the other tasks.

Recently, various change detection benchmark datasets have been proposed to solve this problem, but they still suffer from a lack of data samples [7, 18, 5, 11]. To address this issue, Zheng *et al.* [33] recently proposed ChangeStar that detects changes using single temporal unpaired images with pseudo labels. ChangeStar significantly alleviated the training data collection problem in change detection. It uses `xor` operation on the semantic segmentation labels of two single temporal unpaired images to make pseudo change label, and use pseudo change label to train the change detection model. However, ChangeStar does not consider the style, texture and consistency information coming from the bi-temporal paired images, which significantly degrades performance compared with the bi-temporal paired change detection methods.

2.3. Data Augmentation in Remote Sensing Imagery

In recent years, the strong augmentation strategy such as copy-and-paste (CP) and inpainting is widely used to various deep learning tasks(e.g. classification [31, 30], object detection [8], and object segmentation [10]). Also in remote sensing domain, there had been trial to exploit strong augmentation strategy. Kumdakcı *et al.* [16] proposed inpainting method for augmenting the vehicle instances to solve the data shortage problem, however, it has limitations of generating instances with diversity.

3. Approach

3.1. Change Detection from Bi-temporal Supervision

The goal of change detection based on bi-temporal supervision with given the pre-event images $\{\mathbf{X}_1^{t_0}, \ldots, \mathbf{X}_N^{t_0} | \mathbf{X}_k^{t_0} \in \mathbb{R}^{C \times H \times W}\}$ and the corresponding post-event images $\{\mathbf{X}_1^{t_1}, \ldots, \mathbf{X}_N^{t_1} | \mathbf{X}_k^{t_1} \in \mathbb{R}^{C \times H \times W}\}$ can be formulated as follows:

$$\min_\theta \sum_{k=1}^N \mathcal{L}(\mathcal{F}_\theta(\mathbf{X}_k^{t_0}, \mathbf{X}_k^{t_1}), \mathbf{Y}_k^{t_0 \to t_1}), \qquad (1)$$

where $\mathcal{L}$ indicates the loss function between the predicted change map obtained by change detector $\mathcal{F}_\theta$ with paired bi-temporal images $(\mathbf{X}_k^{t_0}, \mathbf{X}_k^{t_1})$, which indicates the pre-event image and the post-event image of a specific area k. Change map, $\mathbf{Y}_k^{t_0 \to t_1} \in \{0, 1\}^{H \times W}$ represents the regions of change of interest between the pre- and post-event images.

Deep learning change detection models based on bi-temporal paired images(*Pair* training method), require the dataset to be a collection of image pairs. This requirement leads high cost of building process and insufficient samples problem because set of image pairs should be taken from the same region at the different time and also should contain the CoIs.

3.2. Disentangle the Change Map

Zheng *et al.* [33] proposed ChangeStar, which defines change as an area where the same pixel position in pre- and
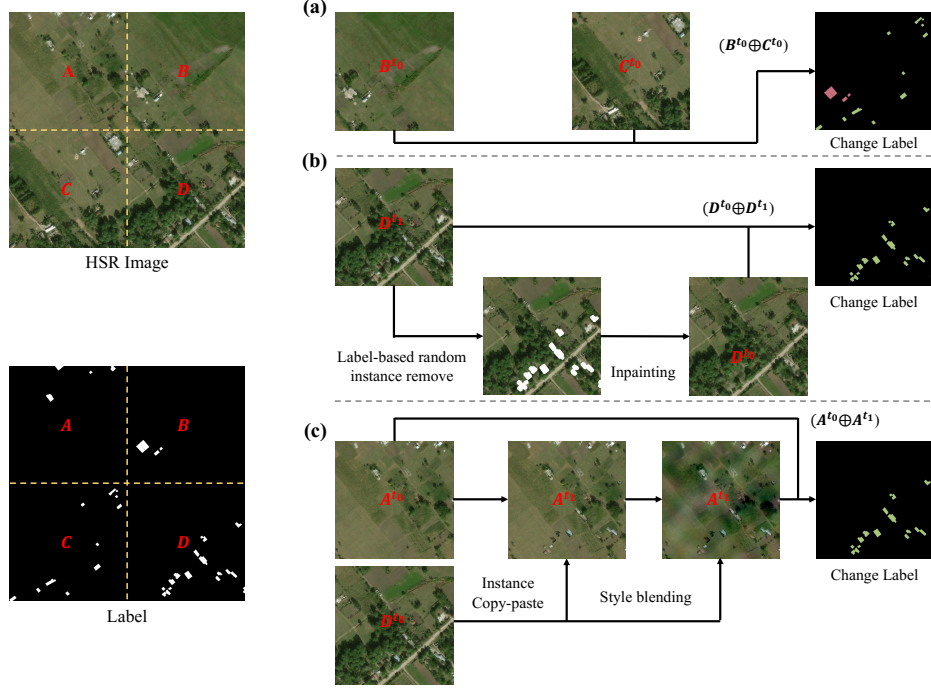
Figure 2. *Self-Pair*. Three ways of generating single temporal paired images with pseudo change label using a single image. (a) Randomly cropping two patches from a single image. (b) Semantic label based inpainting approach. (c) Semantic label based copy-and-paste with style blending approach.

post-event images has different semantic information. From the definition of change by Zheng *et al.*, we can disentangle the change map $\mathbf{Y}^{t_0 \to t_1}$ to the difference of the semantic information $\mathbf{Y}_k \oplus \mathbf{Y}_l$ from two different images $\mathbf{X}_k$ and $\mathbf{X}_l$. The relaxed formulation without bi-temporal information $\{\mathbf{X}_1^{t_1}, \ldots, \mathbf{X}_N^{t_1}\}$ can be expressed as follows:

$$\min_{\theta} \sum_{k=1}^{N} \sum_{l=1}^{N} \mathbb{1}(k \neq l) \mathcal{L}(\mathcal{F}_\theta(\mathbf{X}_k^{t_0}, \mathbf{X}_l^{t_0}), \mathbf{Y}_k^{t_0} \oplus \mathbf{Y}_l^{t_0}), \quad (2)$$

where $\mathbb{1}$ and $\oplus$ denotes the indicator function and the `xor` operation, respectively. Disentangle the change map alleviates the constraint in the original formulation by replacing paired images to unpaired images that having no geographical relationship. By disentangling the change map, change detection model can be trained with pseudo bi-temporal paired images randomly sampled from semantic segmentation datasets. We will call this method the *Unpair* training method.

Unpair training method achieved impressive performance with resolving the high cost problem of constructing bi-temporal paired dataset, however still having a significant performance gap compared with existing state-of-the-art methods that trained with bi-temporal supervision. As shown in Figure 3-(a), ChangeStar ignores meaningful characteristics of bi-temporal paired images: consistency in unchanged area, common patterns, and style similarity

across two images. In Figure 3-(c), areas in the green boxes which labeled as unchanged contain different semantic information (*Left*: Tree, *Right*: Road). Such inconsistencies of unchanged regions across input images could confuse the model to learn what the change is, which leads to degradation of the performance. This indicates that the ChangeStar's *Unpair* training formulation is missing essential properties of the change detection for the real-world bi-temporal setting.

3.3. General Formulation of Change Detection

To overcome the problems come from dataset and reflect the properties missed by *Unpair* training formulation, we rethink how the change creates from given two pre- and post-event images: **all changes originate from the source and arise from the manipulation of the source**. In order to represent this principle, we reformulate change detection as follows:

$$\min_{\theta} \sum_{k=1}^{N} \mathcal{L}(\mathcal{F}_\theta(\mathbf{X}_k^{t_0}, g(\mathbf{X}_k^{t_0})), \mathbf{Y}_k^{t_0} \oplus g(\mathbf{Y}_k^{t_0})), \quad (3)$$

where g is a function that maps an image or label of specific area at particular time to an image or label of arbitrary time step where the change occurs.

This formulation allows us to utilize single-temporal images $\mathbf{X}_k^{t_0}$ like ChangeStar, and to devise more plausible

(a) Bitemporal supervised learning   (b) Single-pair supervised learning   (c) Unpair supervised learning
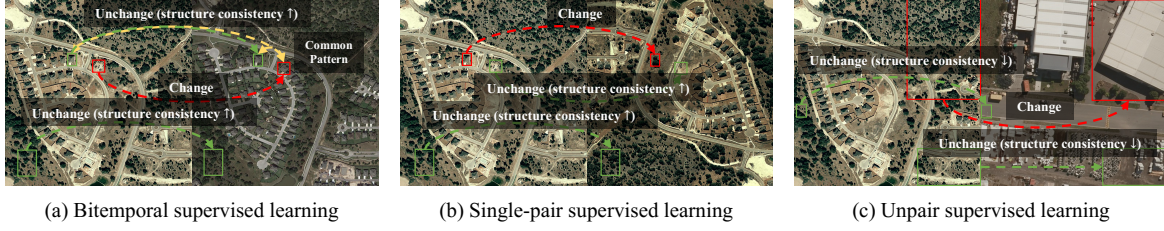
Figure 3. Qualitative comparison of structure similarity by bi-temporal, single-pair, and unpair scenarios.

mapping functions g that preserve meaning information in the real-world bi-temporal paired image setting. With this formulation, in this paper, we focus on the study of designing the function g and propose three simple and effective augmentation strategies.

3.4. Proposed Method

For the candidate of function g, photometric transformation and geometric transformation functions could be on the lists. However, only geometric transformation functions are discussed in this paper due to photometric transformation could not augment the new change instances.

3.4.1 Random Crop from Single Image

A very naive method for generating single-temporal paired images based on geometric transformation is random crop. Randomly cropped two patches with no overlapped regions from the source image are set as an input pair for change detection model. This approach is similar to the method proposed in [33] but different in generating a pair from a single source rather than randomly sampled two images. Although this strategy loses most of its structural consistency, it can still retain the style similarity or typical patterns which can be observed in the real-world bi-temporal paired images. Moreover, this aggressive strategy could behave as a strong augmentation considering the massive changes that can happen in the real world as well.

For the experiment of random crop strategy, rotation operation to the cropped patches is mixed and comprehensive random crop method for change instance augmentation can be expressed as follows:

$$\min_{\theta} \sum_{k=1}^{N} \mathcal{L}(\mathcal{F}_\theta(\mathrm{crop}_1(\mathbf{X}_k^{t_0}), r(\mathrm{crop}_2(\mathbf{X}_k^{t_0}))), \quad (4)$$
$$\mathrm{crop}_1(\mathbf{Y}_k^{t_0}) \oplus r(\mathrm{crop}_2(\mathbf{Y}_k^{t_0}))),$$

where r is a random rotation function, crop_1 and crop_2 are notation of random crop function. Each crop functions crop same location for the corresponding image and label while cropping different locations for single-temporal paired setting.

3.4.2 Inpainting based on Labels

One of the most common changes happens in the real world is the disappearance of objects. We implement this behavior by erasing randomly selected instances and inpaint the background based on surroundings. Since only minimal changes occur in the entire image, most structural consistency is maintained in the unchanged area. Unlike random crop strategy, inpainting strategy preserves the important informations, such as structural consistency, common patterns, and style similarities that can be observed in bi-temporal paired settings. Inpainting based change instance augmentation strategy can be expressed as below:

$$\min_{\theta} \sum_{k=1}^{N} \mathcal{L}(\mathcal{F}_\theta(\mathbf{X}_k^{t_0}, \mathrm{inpaint}((\mathbf{X}_k^{t_0} \times a), 1 - a), \quad (5)$$
$$\mathbf{Y}_k^{t_0} \oplus (\mathbf{Y}_k^{t_0} \times a)),$$

where a is binary mask of the objects to be erased, and these objects are randomly sampled at each time. For the implementation, we adopt method in *Telea et.al* [24] for inpainting, and the inpainted images are set to pre-event images.

3.4.3 Copy-and-Paste Instance Labels

The other changes that commonly happens in the real world is occurance of the objects. A simple method for adding objects to an image is copying objects from the source image using semantic masks and pasting it to the target image. Copy-and-Paste strategy can be used for easing the extreme imbalance between foreground and background that commonly observed in remote sensing images.

For the implementation of copy-and-paste strategy for augmenting the change instances, we copy objects from one of cropped patches and paste them to the other cropped patch. Even though the objects are extracted from a single source, there may be an artifact near boundary of the pasted object due to the randomness of the paste location. To eliminate the artifacts to make augmented sample more realistic, fast Fourier transform based blending method is used.

Let $\mathcal{F}^A$, $\mathcal{F}^P$ be the amplitude and phase components of

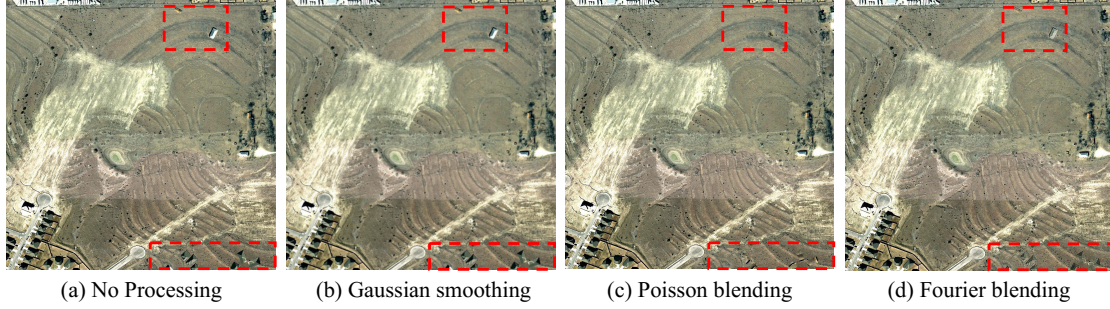| (a) No Processing | (b) Gaussian smoothing | (c) Poisson blending | (d) Fourier blending |

Figure 4. Qualitative comparison of blending methods used in Copy & Paste [8] and our Fourier blending.

the Fourier transform $\mathcal{F}$,

$$\mathcal{F}(x)(m, n) = \sum_{h,w} \mathbf{X}(h, w) e^{-j2\pi(\frac{h}{H}m + \frac{w}{W}n)}, j^2 = -1. \quad (6)$$

We denote M_β as a mask for blending the amplitudes from each original image and modified image by copy-and-paste method. M_β strictly followed the method proposed in [29]. Given two images (original image $\mathbf{X}_k^{t_0}$ and copy-and-pasted image $\mathbf{X}_k^{t_0\prime}$), realistically synthesized copy-and-pasted image using Fourier blending can be follw as:

$$\mathbf{X}_k^{t_1\prime} = \mathcal{F}^{-1}([M_\beta \circ \mathcal{F}^A(\mathbf{X}_k^{t_0}) + (1 - M_\beta) \circ \mathcal{F}^A(\mathbf{X}_k^{t_0\prime}),$$
$$\mathcal{F}^P(\mathbf{X}_k^{t_0\prime})]). \quad (7)$$

Figure 4 is a visual comparison between the conventional blending method and our Fourier blending method. Our Fourier blending method is inspired by [29, 13] that mix the styles of images from two different domains. However, unlike the those studies, we reduce the style gap at the modified regions by replacing the amplitude in the same image in the same domain.

4. Experiments

We evaluate our method in a cross-domain setting, that train the model with building segmentation dataset and validate under building change detection datasets which are constructed in a different purpose. For more fair comparison of *Unpair, Pair,* and *Self-Pair*, evaluation under in-domain setting is also conducted. Note that the LEVIR-CD dataset only offers change labels between t_0 and t_1 images, so in-domain experiments could not be performed.

4.1. Experimental Settings

Training Datasets. Three building segmentation datasets for remote sensing are used to train change detectors in the formulations of *Unpair* Eq.(2) and *Self-Pair* Eq.(3), which exploit only single temporal supervision:

- **xView2 pre-disaster** [33]. xView2 dataset is originally proposed for building damage assessment. The

pre-disaster dataset, which is a subset of the xView2 dataset contains 9,168 pre-disaster HSR images and 316,114 building polygons. We use subset from *train* and *tier3* split dataset. Each image has a size of 1,024 × 1,024 pixels.

- **SpaceNet2** [26]. SpaceNet2 dataset consists of 10,590 HSR images of size 650 × 650 pixels of 0.3 m GSD and 219,316 urban building instance annotations. Following ChangeStar, we also used only 3-bands pan-sharpened RGB images and their annotations.

- **WHU building change detection** [15]. WHU dataset is constructed with one pair of aerial images of size 15,354 × 32,507 pixels obtained in 2012 and 2016 of the same area. It provides 12,796 and 16,077 building instance labels respectively, and change labels across the pair. Train, validation, and test set are composed of 4,736, 1,036, and 2,416 tiles that extracted from original aerial image pair.

Evaluation Datasets. We evaluate *Unpair* and *Self-Pair* methods on the LEVIR-CD and WHU building change detection datasets, which are widely used in change detection evaluation.

- **LEVIR-CD** [5]. The LEVIR-CD dataset contains 637 bi-temporal pairs of HSR images and 31,333 change labels on building instances. Each image has a size of 1,024 × 1,024 pixels with 0.5 m GSD. The change label provides the information about occurance of new buildings and disappearance of existing buildings. Train, validation and test set are split into 445, 128, and 64 pairs. Evaluation of *Unpair* and *Self-Pair* method are conducted with the test set.

Implementation details. We experiment our augmentation strategies based on three state-of-the-art change detectorsSNUNet-CD [9], BIT-CD [4], and ChangeStar [33]. Since these studies were conducted on different settings of backbone, optimizer and training schedule, experiments of *Self-Pair* follow the most of hyperparameters

Model	Method	Train on xView2 pre-disaster				Train on SpaceNet2				Oracle			
		WHU		LEVIR-CD		WHU		LEVIR-CD		WHU		LEVIR-CD	
		IoU (%)	F1 (%)	IoU (%)	F1 (%)	IoU (%)	F1 (%)	IoU (%)	F1 (%)	IoU (%)	F1 (%)	IoU (%)	F1 (%)
SNUNet-CD	Pair									74.54	87.10	81.93	92.11
	Unpair	64.28	72.11	71.22	80.05	66.91	74.77	64.01	70.98				
	Self-Pair	**69.19**	**79.40**	**77.51**	**84.52**	**72.95**	**81.70**	**69.38**	**79.83**				
BIT-CD	Pair									74.48	86.07	81.51	90.86
	Unpair	60.15	70.22	63.29	73.43	66.12	73.85	63.10	69.23				
	Self-Pair	**68.37**	**78.76**	**72.91**	**82.54**	**71.81**	**80.74**	**67.04**	**77.29**				
ChangeStar	Pair									79.89	87.92	91.09	94.91
	Unpair	75.61	82.29	80.13	88.65	64.51	72.14	60.20	68.32				
	Self-Pair	**82.94**	**91.09**	**84.59**	**93.14**	**77.58**	**83.90**	**74.41**	**81.22**				

Table 1. Experimental results of Unpair and *Self-Pair* methods in xView2 pre-disaster → (WHU, LEVIR-CD) and SpaceNet2 → (WHU, LEVIR-CD) cross domain tasks. Oracle is a single domain bi-temporal supervised training setup.

of each studies' experimental details. For SNUNet-CD, 16-channel model is adopted, and for BIT-CD and ChangeStar, ResNet50 backbone [12] is adopted. Three ways of augmentation approaches of *Self-Pair* are applied with the same probability in the training stage.

4.2. Cross-domain Evaluation Results

For evaluating the *Self-Pair*, two formerly proposed methods(*Pair* and Unpair) are used for comparison. *Pair* method trains the model with bitemporal supervision and evaluated under in-domain setting which can be considered as an upper bound. Unlike *Pair* and *Unpair*, *Self-pair* trains the model with single-temporal supervision and evaluated under cross-domain settings to checkout the generalization performance. Here *Unpair* trains the models with Zheng *et al.* [33]'s method, and *Self-pair* indicates the models trained with our proposed method.

As shown in Table 1, all of change detectors trained with our augmentation method outperform the change detectors trained with *Unpair* method regardless of its architecture. Even for the ChangeStar model, the performance of the model trained with our method outperforms the model trained with *Pair* method. This implies that *Self-Pair* method can approximate the distribution of changes in the real-world better than given fixed dataset.

4.3. In-Domain Evaluation Results

We evaluate the performance of ChangeStar model trained with each Pair, *Unpair*, and *Self-Pair* on the WHU and LEVIR-CD dataset to compare the in-domain and cross-domain performance. The experimental results are in Tab 2. As shown in Tab 2, our *Self-Pair* method based ChangeStar achieved the best performance not only under in-domain experiment on the WHU dataset but also in the cross-domain experiment with LEVIR-CD dataset. In addition, the *Pair* method showed the lowest performance in the cross-domain experiment, and was not significantly different from the *Unpair* method even in the in-domain ex-

Model	Method	Train on WHU Trainset			
		WHU		LEVIR-CD	
		IoU (%)	F1 (%)	IoU (%)	F1 (%)
ChangeStar	Unpair	78.13	86.41	59.29	68.82
ChangeStar	Self-Pair	**83.57**	**90.77**	**66.79**	**78.41**
ChangeStar	Pair	79.89	87.92	51.23	55.11

Table 2. Evaluation result on in-domain (WHU Testset), cross-domain (LEVIR-CD Testset) performance according to each augmentation method(*Pair, Unpair* and *Self-Pair*).

periment. Those results indicate that paired images are not essential in both in-domain and cross-domain settings for change detection.

Table 3 is an experiment to analyze the effect of each of three components in *Self-Pair*. As shown in Tab 3, it can be seen as all components of *Self-Pair* have a complementary relationship. Also, when both of *Self-Pair* and *Pair* method are used for training, the model showed 3.34% higher performance than trained with *Self-Pair* method alone. Eventually, summarizing the results of Tab 2 and Tab 3, it shows that paired input setting is not essential for change detection, while can give advantages for performance improvement.

4.4. Qualitative Results

Figure 5 shows the results of qualitative analysis of camparison between *Pair, Unpair*, and *Self-Pair* on the WHU building change detection dataset. As shown in Fig. 5-(e), *Unpair* shows high TP (True Positive) score in both examples but also shows high FP (False Positive) score as well. Compare with *Pair*, *Self-Pair* shows lower FP and FN (False Negative) with showing higher TP. To summarize, results of comparison of the qualitative analysis show that for a task which only focusing the true positive score, using *Unpair* method is efficient, or for the task which requires lower false positives and higher true positives, using *Self-Pair* could be a best augmentation mthod.

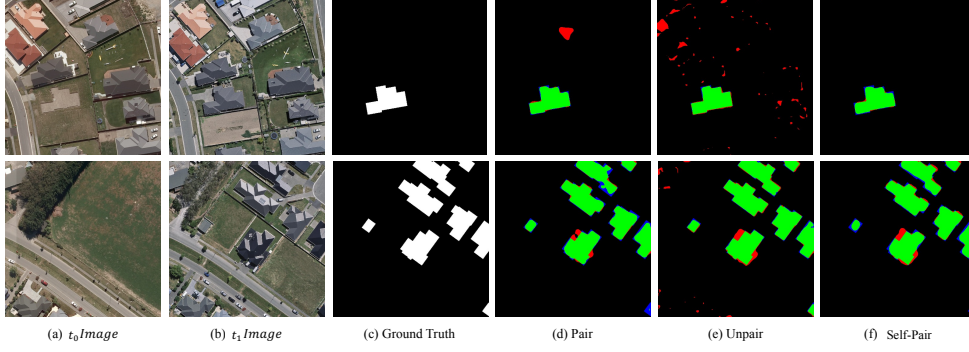| (a) $t_0 Image$ | (b) $t_1 Image$ | (c) Ground Truth | (d) Pair | (e) Unpair | (f) Self-Pair |

Figure 5. Qualitative analysis of pair, unpair, and Self-Pair. True positives (TP), false positives (FP), and false negatives (FN) are represented as green, red, and blue, respectively.

Method	Components					Metric		Gain	
	UN	CR	IN	CP	PA	IoU	F1	IoU	F1
Unpair	✓					78.13	86.41	0	0
Self-Pair		✓				79.23	87.10	+1.10	+0.69
Pair					✓	79.89	87.92	+1.76	+1.51
Self-Pair		✓	✓			82.96	90.21	+4.83	+3.80
Self-Pair		✓	✓	✓		83.57	90.77	+5.44	+4.36
Pair		✓	✓	✓	✓	86.91	94.02	+8.78	+7.61

Table 3. The result of comparing the effect of each component of *Self-Pair* (*UN*: Unpair, *CR*: Crop-and-Rotation, *IN*: Inpainting, *CP*: Copy-and-Paste with Blending, *PA*: Pair) .

Cost	EMD(Pair, Self-Pair)	EMD(Pair, Unpair)
(a) *Changed Areas*		
Cosine ($\times 10^{-2}$)	**4.3173**	10.6672
Euclidean ($\times 1$)	**2.3946**	3.7833
(b) *Unchanged Areas*		
Cosine ($\times 10^{-2}$)	**3.4424**	5.9130
Euclidean ($\times 1$)	**1.5830**	2.2914

Table 4. Earth mover's distance (EMD) between sets of intermediate features from {*Pair* and *Self-Pair*} or {*Pair* and *Unpair*} shown in Fig. 6. *Cost* indicates which type of distance metric is used.

Model	Method	IoU	Gain
FarSeg [35]	Baseline	71.50	0
FarSeg [35]	Naive-CP [10]	69.29	-2.21
Farseg [35]	CPwB	74.71	+3.21
PFNet [17]	Baseline	70.98	0
PFNet [17]	Navie-CP [10]	69.10	-1.88
PFNet [17]	CPwB	71.91	+0.93

Table 5. Evaluation result of building semantic segmentation according to each naive copy-and-paste and copy-and-paste with blending augmentation strategy.

4.5. Discussion and Ablation Study

Why *Self-Pair* works? *Self-Pair* is a strategy to generate a visually plausible realistic synthetic image to consist single-temporal paired images from a single image source. If *Self-Pair* could create the characteristics of the bi-temporal paired images, the domain gap between bi-temporal paired setting and the *Self-Pair* setting should be small [19].

Figure 6 shows the results of t-SNE embedding [25] of *Self-Pair*, bi-temporal paired setting (*Pair*), and single-temporal paired setting (*Unpair*). As show in Figure 6, samples from *Self-Pair* embed near to the embedded samples from *Pair* in both cases where the change occurred and no changes. This indicates that *Self-Pair* creates the synthetic images by utilizing characteristics of bi-temporal paired images. Note that *Self-Pair* can be embedded more widely than embedding result in Fig. 6, during training phase by randomness from inpainting method and copy-paste with blending method.

Table 4 shows Earth mover's distance (EMD) [2] between sets of intermediate features. Regardless of whether sample is labeled to change or non-change, EMD (*Self-Pair*, *Pair*) costs smaller distance than EMD (*Unpair*, *Pair*). This tendency is also maintained regardless of whether the cosine metric or euclidean metric

is used for compute the distance. Accordingly, the results quantitatively proves that our method effectively reduces the domain gap better than the existing single-temporal paired strategy[33].

Ablation study Table 5 shows the results of comparing naive copy-and-paste (Naive-CP) [10] and copy-and-paste with blending (CPwB) in building semantic segmentation dataset. As shown in the table, performance of Naive-CP is lower than baseline. For the semantic segmentation on the remote sensing domain, background context is highly important(e.g. no building on the sea). Also, scale of ob-
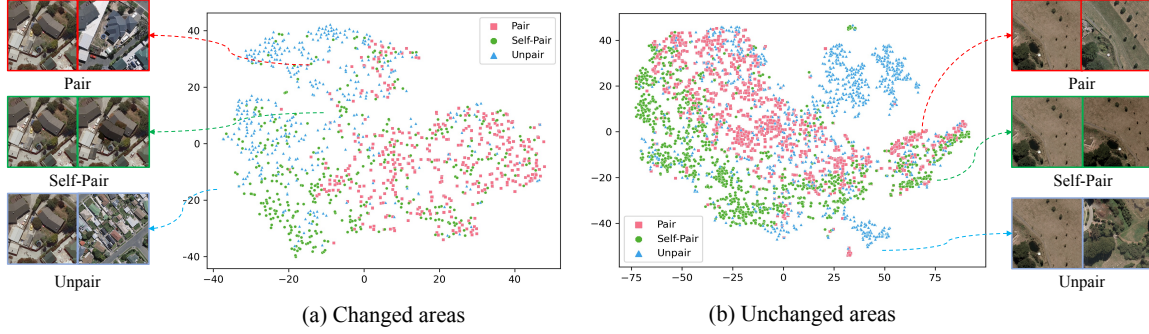
(a) Changed areas   (b) Unchanged areas

Figure 6. The result of t-SNE [25] by concatenating the intermediate features of pre and post-event images from ChangeStar which trained on the WHU change detection dataset. (a) is the result of the area where the change occurred, and (b) is the result of the area where having no changes.
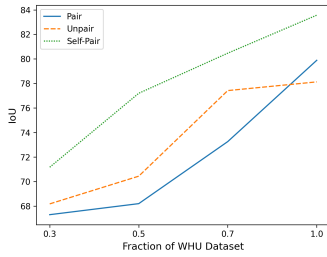

Figure 7. Data-efficiency on the WHU benchmark. Note that *Self-Pair* and *Unpair* requires three times as many epochs as pairs for convergence.

Model	Method	IoU	Gain
FarSeg [35]	Gaussian smoothing	74.54	-0.17
FarSeg [35]	Poisson blending [21]	68.90	-2.60
Farseg [35]	Fourier blending	76.19	+1.48
PFNet [17]	Gaussian smoothing	72.29	+0.38
PFNet [17]	Poisson blending [21]	69.78	-2.13
PFNet [17]	Fourier blending	73.41	+1.50

Table 6. The result of comparing the effect of each blending methods.

jects are having difference between different scenes. This is why the performance of Naive-CP that designed without considering characteristics of HSR remote sensing images is significantly reduced. On the other hand, CPwB is using only one single image for setting an input pair for change detection model which means that there is no scale variation and the context of background is fixed. For these reasons, CPwB could easily adapt to the characteristics of remote sensing imagery domain and showing improved results to the baseline while Naive-CP fails to adapt to it and shows degraded results.

Table 6 shows the result of comparing the performance of Gaussian smoothing, Poisson blending [21] which used in [8], and our Fourier blending. Gaussian smoothing has no improvements and Poisson blending performed poorly compared with copy-and-paste with no blending. To analyze this phenomenon, we visualize the augmented samples based on each blending methods in Fig. 4. In Fig. 4, the pasted buildings shown in the red boxes are very small and in the Fig. 4-(c), most of pasted buildings are erased after blended with poisson method. However, in Fig. 4-(d) every pasted buildings are left with changed texture when fourier method is used for blending. Consequently, qualitative and quantitative results shows that blending method can mitigate the artifacts which created by hard augmentation and

leads to the performance improvements.

Data-efficiency on the WHU dataset. *Self-Pair* augments data in various forms according to three strategies. Figure 7 shows the results of evaluating the data efficiency of *Pair*, *Unpair*, and *Self-Pair*. *Self-Pair* significantly improves data efficiency of the WHU building change detection dataset.

5. Conclusions

In this work, we redefine the change detection problem in way of how the change happens - *how to model the changes happen in the real-world*. We proposed a novel data augmentation method *Self-Pair*, which generates the synthetic image for constructing a input pair based on single-temporal single image and alleviates the problem of high cost of collecting pair set which contains changes for bi-temporal paired supervised learning. We hope our method reduces the time cost of data collection and makes object change detection research more accessible, scalable, and economical.

Acknowledgement

References

[1] Csaba Benedek and Tamás Szirányi. Change detection in optical aerial images by a multilayer conditional mixed markov model. *IEEE Transactions on Geoscience and Remote Sensing*, 47(10):3416–3430, 2009.

[2] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[3] Hongjia Chen, Fangling Pu, Rui Yang, Rui Tang, and Xin Xu. Rdp-net: Region detail preserving network for change detection. *arXiv preprint arXiv:2202.09745*, 2022.

[4] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[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.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[7] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. IEEE, 2018.

[8] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018.

[9] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 2021.

[10] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021.

[11] Maoguo Gong, Jiaojiao Zhao, Jia Liu, Qiguang Miao, and Licheng Jiao. Change detection in synthetic aperture radar images based on deep neural networks. *IEEE transactions on neural networks and learning systems*, 27(1):125–138, 2015.

[12] 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.

[13] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021.

[14] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of photogrammetry and remote sensing*, 80:91–106, 2013.

[15] 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.

[16] Hilmi Kumdakcı, Cihan Öngün, and Alptekin Temizel. Generative data augmentation for vehicle detection in aerial images. In *International Conference on Pattern Recognition*, pages 19–31. Springer, 2021.

[17] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2021.

[18] Jin Liu and Shunping Ji. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2020.

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[20] Lichao Mou, Xiaoxiang Zhu, Maria Vakalopoulou, Konstantinos Karantzalos, Nikos Paragios, Bertrand Le Saux, Gabriele Moser, and Devis Tuia. Multitemporal very high resolution from space: Outcome of the 2016 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3435–3447, 2017.

[21] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[22] 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.

[23] Ashbindu Singh. Review article digital change detection techniques using remotely-sensed data. *International journal of remote sensing*, 10(6):989–1003, 1989.

[24] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[26] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[27] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.

[28] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[29] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[33] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15193–15202, 2021.

[34] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022.

[35] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2020.