# ChangeDiff: A Multi-Temporal Change Detection Data Generator with Flexible Text Prompts via Diffusion Model

**Qi Zang[1], Jiayi Yang[1], Shuang Wang[1]*, Dong Zhao[1], Wenjun Yi[1], Zhun Zhong[2]**

[1]School of Artificial Intelligence, Xidian University, China
[2]School of Computer Science and Information Engineering, Hefei University of Technology, China
qzang@stu.xidian.edu.cn; shwang@mail.xidian.edu.cn

## Abstract

Data-driven deep learning models have enabled tremendous progress in change detection (CD) with the support of pixel-level annotations. However, collecting diverse data and manually annotating them is costly, laborious, and knowledge-intensive. Existing generative methods for CD data synthesis show competitive potential in addressing this issue but still face the following limitations: 1) difficulty in flexibly controlling change events, 2) dependence on additional data to train the data generators, 3) focus on specific change detection tasks. To this end, this paper focuses on the semantic CD (SCD) task and develops a multi-temporal SCD data generator ChangeDiff by exploring powerful diffusion models. ChangeDiff innovatively generates change data in two steps: first, it uses text prompts and a text-to-layout (T2L) model to create continuous layouts, and then it employs layout-to-image (L2I) to convert these layouts into images. Specifically, we propose multi-class distribution-guided text prompts (MCDG-TP), allowing for layouts to be generated flexibly through controllable classes and their corresponding ratios. Subsequently, to generalize the T2L model to the proposed MCDG-TP, a class distribution refinement loss is further designed as training supervision. Our generated data shows significant progress in temporal continuity, spatial diversity, and quality realism, empowering change detectors with accuracy and transferability.

## Introduction

Change detection (CD), a key Earth observation task, employs bitemporal remote sensing data to gain a dynamic understanding of the Earth's surface, producing pixel-wise change maps for ground objects (Feranec et al. 2007; Chen et al. 2013; Kadhim, Mourshed, and Bray 2016). In recent years, data-driven deep learning models have provided promising tools for CD and achieved remarkable progress (Lei et al. 2019; Arabi, Karoui, and Djerriri 2018; Dong et al. 2018). These advancements rely on large-scale, high-quality pixel-level annotations. However, building such a dataset poses a significant challenge because collecting diverse data and manually annotating them is costly, labor-intensive, and requires expert intervention. As a result, these challenges unsurprisingly restrict the size of existing public CD datasets, compared to general-purpose vision datasets such as ImageNet (Deng et al. 2009).

| Methods | Text Control | Layout Diversity | Extra Seg. Data | SCD |
|---|---|---|---|---|
| IAug | ✗ | ✗ | ✓ | ✗ |
| ChangeStar | ✗ | ✗ | ✗ | ✗ |
| Self-Pair | ✗ | ✗ | ✗ | ✗ |
| Changen | ✗ | ✗ | ✗ | ✓ |
| Ours | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of data synthesis method in change detection regarding functionality, data support, and application tasks. Our ChangeDiff shows more practical and strong functions and application scenarios.

To alleviate high demand for data annotation, data synthesis has emerged as an alternative solution with promising application potential. Currently, a few synthesis techniques for binary CD (*e.g., building variations*) have been studied, categorized into two mainstreams: data augmentation-based and data generation-based methods. In the former, IAug (Zheng et al. 2021) and Self-Pair (Seo et al. 2023) use copy-paste and image inpainting techniques, pasting instances or patches from other regions onto target images to simulate building changes. However, the inconsistency between pasted areas and backgrounds makes it challenging to create realistic scene changes. In the latter, Changen (Zheng et al. 2023) introduces a generic probabilistic graphical model to generate continuous change pairs, improving the realism of synthetic images. However, Changen still relies on copy-paste operation of the image mask (semantic layout) to create changes, making it difficult to flexibly control change events. Additionally, the mask-based copy-paste is not easily applicable to semantic CD (SCD) task due to the lack of complete masks. Moreover, it requires additional segmentation data to train the probabilistic model, limiting transferability to specific target data. A detailed comparison of these methods is provided in Table 1.

Recently, driven by latent diffusion models (Rombach et al. 2022), generative models have reached a new milestone(Khanna et al. 2023). Stable Diffusion (Podell et al. 2023) and DALL-E2 (Ramesh et al. 2022) introduce large-scale pretrained text-to-image (T2I) diffusion models that can generate high-quality images matching textual descriptions. Furthermore, advanced work, *e.g.*, ControlNet (Zhang

*Corresponding author

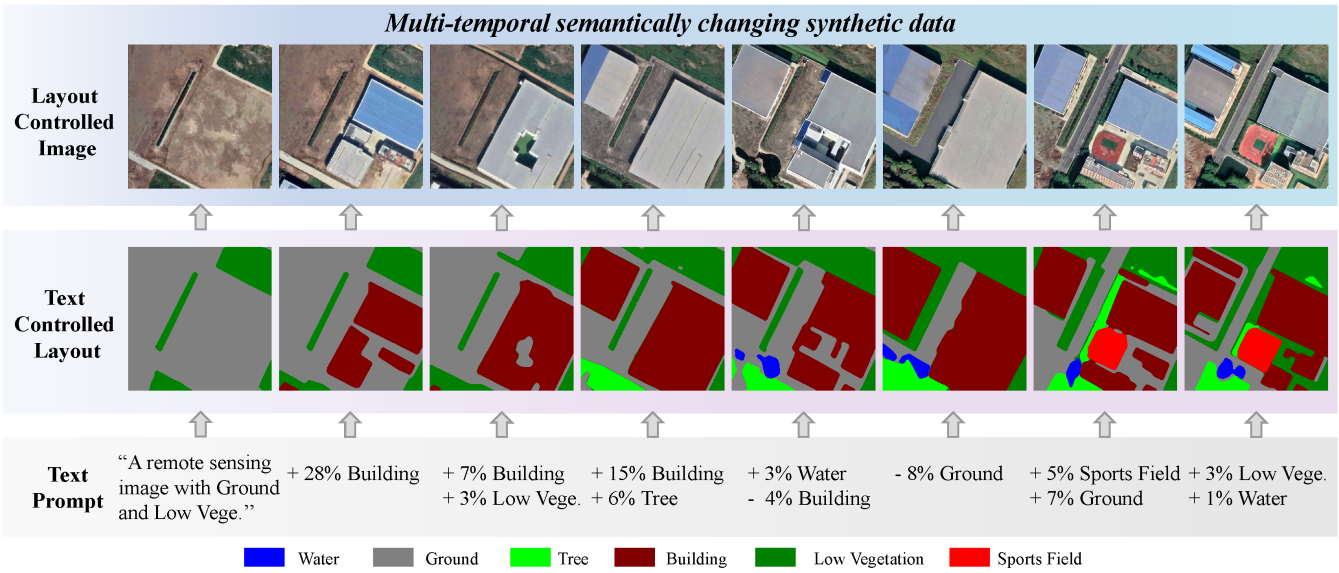



Figure 1: Multi-temporal semantic change synthetic data synthesized by our ChangeDiff, which is trained on the sparsely labeled semantic change detection *SECOND* (Yang et al. 2021) dataset. It describes an *area under construction, where man-made facilities are gradually being completed.* ChangeDiff takes text prompts as input, generates semantic events, and specifies changes in a controllable manner by modifying the text prompts.

2023), has shown that by incorporating fine-grained controls, such as semantic layouts, textures, or depth, T2I models can be adapted into layout-to-image (L2I) models, allowing for more flexible generation of images matching input layouts. *This inspires us to question whether advanced T2I and L2I models can be applied to CD data synthesis to enhance CD tasks.*

Through our analysis, we identify the core challenge as: *how to construct continuous change events using input text*, which arises from the following aspects: ① In CD tasks, especially SCD, semantic annotations are incomplete (sparse), with only the semantic classes of changed areas being labeled. This makes it difficult to create precise text-to-image mappings and train T2I or L2I models. ② Common used text prompts, such as "`A remote sensing image of {classname}`", are inadequate for present continuous change events. We need to develop suitable text prompts tailored for CD tasks. ③ Text-to-image mapping does not provide spatial semantics of the generated images, rendering the synthesized image pairs for downstream supervised training.

In this paper, we explore the potential of T2I and L2I models for SCD tasks and develop a novel multi-temporal SCD data generator without requiring paired images or external datasets, coined as **ChangeDiff**. To address the core challenges, ChangeDiff innovatively divides the change data generation into two steps: 1) it uses carefully designed text prompts and text-to-layout (T2L) diffusion models to generate continuous layouts; 2) it employs layout-to-image (L2I) diffusion model to transform these layouts into continuous time-varying images, as illustrated in Fig. 1. Specifically, we innovatively develop a text-to-layout (T2L) generation model using multi-class distribution-guided text prompt (MCDG-TP) as input to generate layout flexibly. Our MCDG-TP translates the layout into semantic class distributions via text, *i.e.*, each class with a class ratio, which offers a simple yet powerful control over the scene composition. Meanwhile, the ingredients of MCDG-TP differ from the text prompts used for pre-training the T2I model. This difference prevents the T2I model from generalizing to texts with arbitrary compositions, as text is the sole driver of optimization during noise prediction. To address this, we design a class distribution refinement loss to train our T2L model. With the trained T2L model, sparse layouts enable completion by inputting text with amplified class ratios; then, taking the completed layout as a reference, time-varying events can be simulated via MCDG-TP in three modes: ratio reshaper, class expander, and class reducer. Subsequently, the fine-tuned L2I model synthesizes new images aligned with the simulated layout masks. The data generated by our ChangeDiff shows significant progress in temporal continuity, spatial diversity, and quality realism. The main contributions of this paper are five-fold:

- To the best of our knowledge, we are the first to explore the potential of diffusion models for the SCD task and develope the first SCD data generator ChangeDiff.
- We propose a multi-class distribution-guided text prompt (MCDG-TP), using controllable classes and ratios, to complement sparse layout masks.
- We propose a class distribution refinement loss as training supervision to generalize the text-to-image diffusion model to the proposed MCDG-TP.
- We propose MCDG-TP in three modes to simulate time-

varying events for synthesizing new layout masks and corresponding images.

- Extensive experiments demonstrate that our high-quality generated data is beneficial to boost the performance of existing change detectors and has better transferability.

## Related Work

**Binary & Semantic Change Detection & Text-guided Diffusion Model.** Please refer to the Appendix A.

**Data Synthesis in Change Detection.** Currently, there are several advanced data synthesis methods for the change detection task. ChangeStar (Zheng et al. 2021) and Self-Pair (Seo et al. 2023) employ simple copy-paste operations, pasting patches from other regions onto the target image to simulate changes. However, artifacts introduced by the paste operation and the inconsistency between foreground and background make it challenging to create realistic scene changes. IAug (Chen, Li, and Shi 2021) uses a generative model (GAN) to synthesize changed objects, but its building-specific modeling approach limits its generalization to diverse scenes. Changen (Zheng et al. 2023) proposes a generic probabilistic graphical model to represent continuous change events, enhancing the realism of synthetic images. Its latest version, Changen2 (Zheng et al. 2024), introduces a diffusion transformer model, further improving generation quality. However, it relies on additional segmentation data to train the probabilistic model, making it unsuitable for direct data augmentation in change detection tasks.

Our ChangeDiff does not rely on additional segmentation data, simplifying its integration into existing workflows. Besides, ChangeDiff supports text control, enabling users to specify the generated changes precisely. Furthermore, ChangeDiff can synthesize diverse and continuous layouts, which is crucial for improving the transferability of synthetic data.

## Method

Given a single-temporal image $x \in \mathbb{R}^{H\times W\times 3}$ and its sparsely-labeled (only the change area) semantic layout $y \in \mathbb{R}^{H\times W}$ in the semantic change detection (SCD) dataset, our SCD data generator ChangeDiff aims to simulate temporal changes via diffusion models conditioned on diverse completed semantic layouts. The pipeline of ChangeDiff is shown in Fig. 2. Overall, ChangeDiff consists of the text-to-layout (T2L) diffusion model for changing layout synthesis and the layout-to-image (L2I) diffusion model for changing image synthesis. In the next part, we first introduce the detailed design of T2L and L2I diffusion models. Then, we discuss how to flexibly encode semantic layouts into the text prompt. Lastly, we introduce how to synthesize complete and diverse layouts via text prompts.

### Preliminary

Both T2L and L2I models are built on the latent diffusion model (LDM) (Rombach et al. 2022), widely used in conditional generation tasks for its powerful capabilities. LDM learns the data distribution by constructing a $T$-step denoising process in the latent space for the normally distributed variables added to the image $x \in \mathbb{R}^{H\times W\times 3}$. Given a noisy image $x_t$ at time step $t \in \{1,\dots,T\}$, the denoising function $\epsilon_\theta$ parameterized by a U-Net (Ronneberger, Fischer, and Brox 2015) learns to predict the noise $\epsilon$ to recover $x$.

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,t,\epsilon\sim\mathcal{N}(0,1)}\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(\mathbf{T}))\|^2, \quad (1)$$

where $z = \varepsilon_\theta(x)$ is the encoding of the image in latent space. $\tau_\theta$ is a pre-trained text encoder (Radford et al. 2021) that enables cross-attention between the text $\mathbf{T}$ and $z_t$ to steer the diffusion process.

**T2L Model.** Given a semantic layout $y$ where pixel values are category IDs, we encode $y$ as a three-channel color map $\mathbf{C}_y$ in RGB style, and each color indicates a category. This enables leveraging the pre-trained LDM model to generate semantic layouts from input text prompts $\mathbf{T}$.

**L2I Model.** Recent works like ControlNet (Zhang 2023) propose conditioning on both semantic layout $y$ and text prompt $\mathbf{T}$ to synthesize images aligned well with specific semantic layouts. Following this, we adopt the ControlNet structure, which adds a trainable side network to the LDM model for encoding semantic layout $y$.

### Flexible Text Prompt as Condition

We explore encoding semantic layouts via text prompts to utilize the T2L model for completing sparse layouts and generating diverse layouts.

**Semantic Layout Translation.** A semantic layout $y$ can be decomposed into two components: category names $\{n_j\}$ and their corresponding connected areas $\{a_j\}$. The names can be naturally encoded by filling in the corresponding textual interpretation into $\mathbf{T}$, but connected areas can not since the pixels within them are arranged consecutively. To discretize, we partition each pixel into cells at distinct coordinate points. Cells with the same category ID can be count-aggregated into a unique class ratio, quantitatively characterizing the corresponding $\{a_j\}$ for insertion into the encoding vocabulary of $\mathbf{T}$. Formally, given a semantic layout $y$, the connected areas $a_j$ of the $j$-th class are represented by the class ratio $R_j$ as,

$$R_j = \sum_{hw}^{HW} [\mathbb{1}(y_{hw} = j)]/HW. \quad (2)$$

$\mathbb{1}(\cdot)$ is an indicator function, which is 1 if the category ID is $j$, otherwise it is 0. Then, a certain category $j$ in the semantic layout $y$ is encoded as a phrase $(n_j, R_j)$ with two tokens.

**Text Prompt Construction.** With the phrases encoding multiple categories, we serialize them into a single sequence to generate text prompts. All phrases will be sorted randomly. Specifically, we adopt a template to construct the text prompt, "`A remote sensing photo with {class distributions}`", where class distributions = "$\dots(n_{j-1}, R_{j-1})(n_j, R_j)(n_{j+1}, R_{j+1})\dots$". We term this as multi-class distribution-guided text prompt (MCDG-TP).

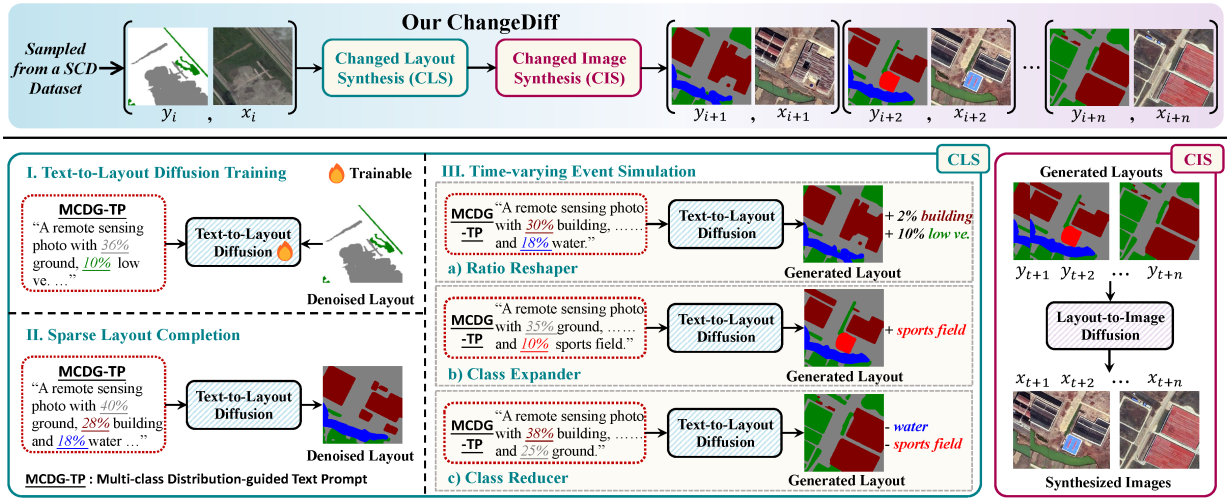



Figure 2: An overview of the proposed ChangeDiff, which consists of two components, *i.e.*, changing layout synthesis (CLS) and changing image synthesis (CIS). In the CLS: I. The text-to-layout (T2L) model is fine-tuned on the target data via MCDG-TP; II. Text with amplified class ratios is fed into the fine-tuned T2L model to generate the completed layouts; III. Taking the completed layout as a reference, texts with different class compositions are fed into the fine-tuned T2L model to synthesize temporally changed layouts. In the CIS: New images aligned with the changed layouts are synthesized via a layout-to-image model fine-tuned to the target data.

## Class Distribution Refinement Loss

The loss function $\mathcal{L}_{LDM}$ used to pre-train the LDM model only optimizes the denoising function $\epsilon_\theta$ (See Eq. (1)), enabling it to predict noise and thus recover the clean image. There is no explicit constraint between the embeddings $\tau_\theta(\mathbf{T})$ of text prompts and noisy images $z_t$ in the latent space. As a result, when provided with refined text prompts during inference, the learned denoising function $\epsilon_\theta$ may fail to generate the corresponding object composition. To better understand the text prompt $\mathbf{T}$, we design a class distribution refinement loss $\mathcal{L}_{CDR}$ that explicitly supervises the cross-attention map between $\tau_\theta(\mathbf{T})$ and $z_t$ during training. Our $\mathcal{L}_{CDR}$ assists the model in capturing information from the budget ratio and spatial location. For a cross-attention map at any layer $m \in M$, the $\mathcal{L}_{CDR}$ is defined as follows,

$$\mathcal{L}_{CDR} = \underbrace{\mathcal{L}_{RAT}}_{\text{ratio}} + \underbrace{\mathcal{L}_{SPA}}_{\text{spatial location}} . \tag{3}$$

For the ratio, the cross-attention maps $\mathcal{A}_{class}$ of objects are weighted by that of their corresponding ratios $\mathcal{A}_{ratio}$ to define a combined map $\mathcal{A}_{com} = \mathcal{A}_{class} \cdot \mathcal{A}_{ratio}$. $\mathcal{A}_{com}$ is taken as the constraint target in $\mathcal{L}_{RAT}$,

$$\mathcal{L}_{RAT} = \frac{1}{J}\sum_{j}^{J}\left|\frac{\sum_{hw}^{H_m W_m}[\mathbb{1}(\mathcal{A}^{hw}_{(com,j)}>0)]}{H_m W_m} - R_j\right|. \tag{4}$$

$\mathcal{A}^{hw}_{(com,j)}$ is the intersection of class activations in the generated features and the GT. $H_m \times W_m$ is the map size of the $m$-th layer. This formula enforces that the cross-attention activations from the diffusion model align with the true class ratio $R_j$ for any class $j$.

For the spatial location, we acquire a binary segmentation map $\mathcal{M}_j$ for each object in $\mathbf{T}$ from its respected semantic layout $y$, providing the ground truth distribution. $\mathcal{L}_{SPA}$ aggregates pixel-level activation values of objects and encourages their even distribution,

$$\mathcal{L}_{SPA} = \frac{1}{JH_mW_m}\sum_{j}^{J}\sum_{hw}^{H_mW_m}(\mathcal{A}^{hw}_{(class,j)} - \mathcal{M}^m_j)^2. \tag{5}$$

$\mathcal{A}^{hw}_{(class,j)}$ is the cross-attention map from the diffusion model, which means activations of category words on generated image features. $\mathcal{M}^m_j$ is formed through bilinear interpolation followed by binarization to match the resolution of the map of the $m$ -th layer. This formula enforces spatial alignment of activations with the true response $\mathcal{M}^m_j$.

## Changing Layout Synthesis

To synthesize the changed images, reasonable and diverse layout synthesis in the temporal dimension is required as a semantic guide.

**T2L Model Training.** Given the target SCD dataset, we use the [text prompt $\mathbf{T}_y$, color map $\mathbf{C}_y$] training pairs to fine-tune the T2L model, where $\mathbf{T}_y$ is the corresponding text generated via our MCDG-TP for the color map. With the proposed loss $\mathcal{L}_{CDR}$ and the original loss $\mathcal{L}_{LDM}$, the T2L model is supervised during fine-tuning as follows,

$$\mathcal{L}_{\text{ChangeDiff}} = \mathcal{L}_{LDM} + \sum_{m}^{M}\mathcal{L}^m_{CDR}. \tag{6}$$

**Sparse Layout Completion.** Since generating a changed image requires a complete layout, it is necessary to complete

the sparse layout to serve as a reference for synthesizing the changed layout. To obtain the completed layout, we input text prompt $\mathbf{T}_c$ with amplified class ratios and random noise $z_c$ sampled from $z \sim \mathcal{N}(0,1)$ into the fine-tuned T2L model. By varying $\mathbf{T}_c$ and $z_c$, various reference color maps $\mathbf{C}_i$ with different object compositions can be obtained.

**Time-varying Event Simulation.** With an arbitrary reference layout and its corresponding $[\mathbf{T}_c, z_c]$, we input noise sampled following $z_c$ and varied text into the fine-tuned T2L model to simulate real-world time-varying events. For the varied text, we construct MCDG-TP in three modes, a) ratio reshaper $\mathbf{T}_s$: randomly change the ratio of each class in $\mathbf{T}_c$; b) class expander $\mathbf{T}_e$: randomly create new classes into $\mathbf{T}_c$; c) class reducer $\mathbf{T}_d$: randomly remove certain classes from $\mathbf{T}_c$. Diverse changed color maps $\{\mathbf{C}_{i+1}, \ldots, \mathbf{C}_{i+n}\}$ in spatiality can be obtained, and then we can get the changed layout masks $\{y_{i+1}, \ldots, y_{i+n}\}$ via a learning-free projection function $f_{color \to mask} : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{H \times W \times J}$, *i.e.*, maps RGB values to mask space by simple color matching, *e.g.*, mapping red pixel ([255, 0, 0]) to class id 1.

## Changing Image Synthesis

Conditioned on the synthesized color maps $\{\mathbf{C}_i, \ldots, \mathbf{C}_{i+n}\}$ and texts $\{\mathbf{T}_i, \ldots, \mathbf{T}_{i+n}\}$, we use the fine-tuned L2I model to synthesize images $\{x_i, \ldots, x_{i+n}\}$ aligned with the given layouts. During synthesis, we randomly sample a noise $z_i$ from $z \sim \mathcal{N}(0,1)$ to obtain a reference image $x_i$. Starting from $x_i$, the semantic content of images over time should remain within a certain similarity range. To this end, the input noises $\{z_{i+1}, \ldots, z_{i+n}\}$ for the changed images is sampled via a stitching mechanism, which is formulated by,

$$z_{i+1} = \alpha z_i + (1-\alpha) z_{r(r \neq i)}. \tag{7}$$

$z_{r(r \neq i)}$ is arbitrary noise sampled from $z \sim \mathcal{N}(0,1)$. Proportional injection of $z_{r(r \neq i)}$ ensures the semantic content of the synthesized image conforms to realistic and reasonable changes, *i.e.*, be temporally continuous. At this point, the synthesized images can be paired with layout masks to form new large-scale training samples with dense annotations, $\{(x_i, y_i), \ldots, (x_{i+n}, y_{i+n})\}$.

# Experiments

## Datasets and Experimental Setup

**Setup.** We evaluate the effectiveness of ChangeDiff on semantic change detection tasks using the following two settings: ***1) Data Augmentation***: This setup aims to verify if synthetic data from ChangeDiff can enhance the model's discrimination capability on in-domain samples. We use three commonly used semantic change detection datasets, including SECOND (Yang et al. 2021), Landsat-SCD (Yuan et al. 2022), and HRSCD (Daudt et al. 2019). We train our ChangeDiff on these datasets, respectively. ***2) Pre-training Transfer***: This setup aims to verify if ChangeDiff can leverage publicly available semantic segmentation data to synthesize extensive data for pre-training, benefiting downstream change detection tasks. We use additional semantic segmentation data from LoveDA (Wang et al. 2021) as the training source for ChangeDiff and perform pre-training with the synthetic data. We then validate its effectiveness on the SECOND and HRSCD target datasets using two transfer ways, including "zero-shot transfer" and "fine-tuning transfer".

**Datasets.** Please refer to the Appendix B.

**Implementation Details.** Please refer to the Appendix C.

| Train on 5% SECOND Train Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| SSCDL | **83.9** | 65.4 | 10.5 | 50.8 | 54.9 | 46.4 | 63.4 | 61.8 | **65.0** |
| + Ours | 83.5 | **66.3** | **13.0** | **53.0** | **56.8** | **48.7** | **65.5** | **69.1** | 62.3 |
| BiSRNet | 83.6 | 64.8 | 9.6 | 49.5 | 53.9 | 45.4 | 62.5 | 60.7 | **64.3** |
| + Ours | **83.8** | **66.3** | **12.3** | **52.4** | **56.7** | **48.3** | **65.2** | **66.5** | 63.9 |
| TED | 83.9 | 65.7 | 10.5 | 50.3 | 55.4 | 46.8 | 63.8 | 62.1 | 65.5 |
| + Ours | **84.7** | **66.6** | **12.4** | **53.0** | **56.8** | **48.0** | **64.9** | **62.9** | **66.9** |
| A2Net | **83.8** | 66.1 | 10.4 | 49.7 | 56.1 | 47.5 | 64.4 | 63.3 | **65.5** |
| + Ours | 83.7 | **66.2** | **11.6** | **51.4** | **56.5** | **48.1** | **65.0** | **66.0** | 63.9 |
| SCanNet | 82.8 | 65.1 | 11.3 | 51.1 | 54.8 | 47.0 | 64.0 | **67.6** | 60.7 |
| + Ours | **85.4** | **67.1** | **13.6** | **54.2** | **57.6** | **48.5** | **65.3** | 61.9 | **69.1** |
| Train on 20% SECOND Train Set | | | | | | | | | |
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| A2Net | 84.9 | 68.9 | 15.8 | 55.9 | 60.9 | 52.3 | 68.7 | **71.6** | 66.0 |
| + Ours | **86.1** | **69.6** | **16.6** | **57.3** | **61.9** | **52.8** | **69.1** | 68.1 | **70.1** |
| SCanNet | 84.8 | 68.6 | 16.2 | 56.5 | 60.4 | 51.9 | 68.4 | **72.0** | 65.1 |
| + Ours | **86.9** | **70.1** | **18.2** | **59.1** | **62.6** | **53.2** | **69.4** | 66.5 | **72.6** |
| Train on 100% SECOND Train Set | | | | | | | | | |
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| SSCDL | 87.3 | 71.9 | 20.6 | 60.2 | 65.4 | 56.1 | 71.9 | 68.3 | **75.8** |
| + Ours | **88.2** | **73.3** | **23.0** | **63.3** | **67.4** | **58.1** | **73.5** | **71.6** | 75.5 |
| BiSRNet | 87.4 | 71.9 | 20.9 | 60.8 | 65.3 | 56.0 | 71.8 | 68.2 | 75.8 |
| + Ours | **88.4** | **73.4** | **23.4** | **63.7** | **67.6** | **58.2** | **73.5** | **70.6** | **76.8** |
| TED | 87.4 | 72.4 | 21.3 | 61.1 | 66.1 | 56.9 | 72.5 | 69.9 | 75.4 |
| + Ours | **88.3** | **73.6** | **23.4** | **63.6** | **67.9** | **58.6** | **73.9** | **71.8** | **76.0** |
| A2Net | 87.8 | 72.8 | 22.3 | 61.9 | 66.7 | 57.4 | 72.9 | 69.3 | 76.9 |
| + Ours | **88.1** | **73.3** | **23.2** | **63.4** | **67.5** | **58.3** | **73.6** | **72.4** | **76.9** |
| SCanNet | 87.8 | 72.9 | 22.8 | 62.6 | 66.8 | 57.7 | 73.1 | 70.6 | 75.9 |
| + Ours | **89.0** | **73.7** | **24.4** | **65.1** | **67.9** | **58.3** | **74.1** | **72.6** | **76.0** |

Table 2: Performance of our ChangeDiff using as data augmentation on the *SECOND* dataset.

## Data Augmentation:

We validate ChangeDiff as data augmentation on three datasets: SECOND, Landsat-SCD (sparse annotations), and HRSCD (complete annotations). ChangeDiff is integrated with various methods, including CNN-based approaches (SSCDL, BiSRNet, TED, A2Net) and Transformer-based SCanNet. Results show significant improvement in addressing semantic imbalance (Sek metric) and enhancing binary change detection (F1-score).

***Augmentation for SECOND Dataset.*** As shown in Table 2, for models trained on 5% of the SECOND dataset, our method consistently improves performance, with average gains of 2.5% in SeK, 2.3% in IoU, 2.1% in F1 score, and notable improvements in recall and precision. As the training set increases to 20% and 100%, our method continues to deliver substantial gains, with 2.4% in SeK, 1.7% in IoU, and 1.4% in F1 score, along with enhanced recall and precision, proving its effectiveness in various scenarios.

***Augmentation for Landsat-SCD Dataset.*** The low resolution of the Landsat-SCD dataset challenges data synthesis.

| Train on 5% Landsat-SCD Train Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| SSCDL | 45.4 | 41.4 | 20.4 | 39.3 | 38.1 | 35.7 | 42.0 | 41.6 | 42.4 |
| + Ours | **46.6** | **44.2** | **22.5** | **41.5** | **40.3** | **38.8** | **45.0** | **43.3** | **44.1** |
| BiSRNet | **41.8** | 36.7 | 26.1 | 37.9 | 36.8 | **36.3** | 40.8 | 41.3 | 40.4 |
| + Ours | 41.5 | **39.1** | **29.4** | **42.0** | **38.6** | 33.1 | **43.0** | **44.1** | **42.1** |
| TED | 44.3 | **43.6** | 24.6 | 41.8 | **43.9** | **40.8** | **46.4** | **46.5** | 46.2 |
| + Ours | **49.1** | 43.4 | **26.4** | **43.7** | 39.8 | 36.9 | 46.2 | 45.8 | **46.6** |
| A2Net | 38.6 | 35.2 | 27.2 | **39.4** | 36.1 | 31.7 | 39.5 | 40.5 | 38.5 |
| + Ours | **44.3** | **39.3** | **28.2** | 39.3 | **39.6** | **37.9** | **43.4** | **43.1** | **43.6** |
| SCanNet | 47.7 | 45.6 | 30.5 | 45.2 | 45.2 | 43.0 | 45.2 | 45.5 | 45.0 |
| + Ours | **50.1** | **47.7** | **33.2** | **47.5** | **47.7** | **44.2** | **46.5** | **46.9** | **46.2** |
| Train on 100% Landsat-SCD Train Set | | | | | | | | | |
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| SSCDL | 94.4 | 84.2 | 46.3 | 82.7 | 82.2 | 74.5 | 85.4 | 85.8 | 85.0 |
| + Ours | **96.8** | **86.7** | **48.3** | **83.0** | **84.0** | **75.6** | **86.5** | **87.9** | **85.2** |
| BiSRNet | 94.5 | 84.3 | 54.2 | 83.4 | 82.4 | 74.7 | 85.5 | 86.1 | 84.9 |
| + Ours | **95.0** | **86.3** | **56.5** | **85.7** | **83.4** | **76.9** | **87.2** | **88.8** | **85.6** |
| TED | 95.9 | 88.2 | 57.2 | 87.5 | 87.2 | 80.9 | 89.5 | 89.5 | 89.4 |
| + Ours | **98.3** | **89.1** | **58.0** | **87.7** | **87.2** | **81.5** | **90.6** | **90.3** | **91.0** |
| A2Net | 94.4 | 84.1 | 46.7 | 83.1 | 82.2 | 74.4 | 85.3 | 86.0 | 84.7 |
| + Ours | **94.5** | **85.4** | **46.8** | **84.7** | **83.7** | **75.5** | **86.4** | **87.6** | **85.3** |
| SCanNet | 96.5 | 89.4 | 61.5 | 89.6 | 88.6 | 82.8 | **90.6** | 91.0 | 90.2 |
| + Ours | **97.6** | **90.4** | **63.6** | **91.2** | **90.2** | **84.2** | 90.1 | **92.0** | **91.3** |

Table 3: Performance of our ChangeDiff using as data augmentation on the *Landsat-SCD* dataset.

In Table 3, with 5% training samples, ChangeDiff improves SeK across all models: SSCDL's SeK rises from 20.4% to 22.5%, and BiSRNet's from 26.1% to 32.4%. F1-scores also improve, with SSCDL's from 42.0% to 45.0% and A2Net's from 39.5% to 43.4%. With 100% training, SeK improves further, with SSCDL's SeK increasing from 46.3% to 48.3%, and TED's F1 rising from 89.5% to 90.6%.

***Augmentation for HRSCD Dataset.*** The HRSCD dataset faces class imbalance and coarse annotations, but ChangeDiff still improves performance. With 5% training samples, it boosts SeK and F1 across all models, improving imbalance handling and detection accuracy. With 100% training, ChangeDiff further enhances SeK and F1, refining both imbalance management and overall performance.

***Comparison with Augmentation Competitors.*** Fig. 3 compares data augmentation methods on SECOND and HRSCD. ChangeStar and Self-Pair cause SeK drops of 1.4% on SECOND and 1.2% on HRSCD. IAug underperforms by 1.0% on SECOND and 3.6% on HRSCD. Changen improves slightly on HRSCD but underperforms on SECOND. Our method boosts SeK by 2.1% on SECOND and 2.2% on HRSCD, outperforming others.

## Pre-training Transfer

This section validates the benefit of using ChangeDiff for data synthesis on the out-of-domain LoveDA dataset in the SCD task, through experiments on *zero-shot transfer* and *fine-tuning transfer*.

***Zero-shot Transfer.*** Table 5 compares four methods: Copy-Paste, ControlNet + Copy-Paste, Changen, and ChangeDiff, each generating 10k images. ChangeDiff leads with mIoU5 (60.1%), F1 (55.5%), and SeK5 (13.6%). Copy-Paste has the lowest scores: mIoU5 (39.9%), F1 (42.7%), and SeK5 (4.7%). ControlNet + Copy-Paste improves to mIoU5 (55.1%), F1 (49.4%), and SeK5 (10.7%), while Changen scores mIoU5 (53.1%), F1 (47.2%), and SeK5 (7.9%). ChangeDiff provides higher-quality data with better transferability.

| Train on 5% HRSCD Train Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| BiSRNet | 41.1 | 36.3 | 12.2 | 30.9 | **33.5** | 28.1 | **37.3** | **37.3** | **37.3** |
| + Ours | **43.5** | **37.2** | **14.1** | **31.9** | 32.8 | **29.9** | 36.7 | 36.2 | 37.1 |
| TED | 41.3 | 35.5 | 12.1 | 28.8 | 32.0 | 28.2 | 35.1 | 34.3 | 35.9 |
| + Ours | **44.2** | **37.9** | **14.9** | **29.0** | **34.0** | **30.6** | **36.6** | **34.6** | **38.9** |
| A2Net | 41.2 | 35.4 | 10.9 | 28.5 | 31.2 | 27.9 | 34.2 | 33.6 | 34.9 |
| + Ours | **45.3** | **36.0** | **12.9** | **30.9** | **33.9** | **29.7** | **35.7** | **33.8** | **37.8** |
| SCanNet | 43.2 | 36.0 | 13.0 | 29.2 | 32.9 | 28.9 | 35.7 | 35.2 | 36.2 |
| + Ours | **47.4** | **36.4** | **14.3** | **29.4** | **34.2** | **29.2** | **36.7** | **36.6** | **36.7** |
| Train on 100% HRSCD Train Set | | | | | | | | | |
| Methods | OA | mIoU | Sek | $F_{scd}$ | Kappa | IoU | F1 | Rec. | Pre. |
| BiSRNet | 87.2 | 72.6 | 23.6 | 61.9 | 67.4 | 57.7 | 74.1 | 72.5 | 75.7 |
| + Ours | **89.8** | **75.4** | **23.9** | **64.0** | **68.9** | **60.4** | **76.5** | **75.7** | **77.4** |
| TED | 87.4 | 72.9 | 23.7 | 62.0 | 67.6 | 58.5 | 73.7 | 72.3 | 75.2 |
| + Ours | **88.8** | **74.1** | **25.9** | **64.0** | **68.4** | **59.8** | **75.2** | **74.4** | **76.1** |
| A2Net | 87.9 | 73.2 | 23.8 | 62.8 | 67.9 | 58.6 | 74.2 | 72.7 | 75.8 |
| + Ours | **89.0** | **74.7** | **25.1** | **64.2** | **69.1** | **60.6** | **75.5** | **74.1** | **77.0** |
| SCanNet | 89.1 | 74.2 | 25.1 | 65.3 | 69.3 | 61.2 | 76.1 | 74.1 | 78.2 |
| + Ours | **90.4** | **76.0** | **26.7** | **66.6** | **71.2** | **62.2** | **77.2** | **75.7** | **78.9** |

Table 4: Performance of our ChangeDiff using as data augmentation on the *HRSCD* dataset.

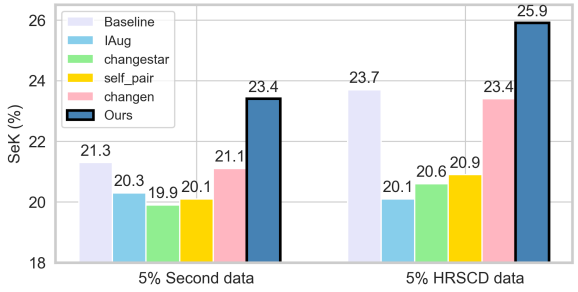
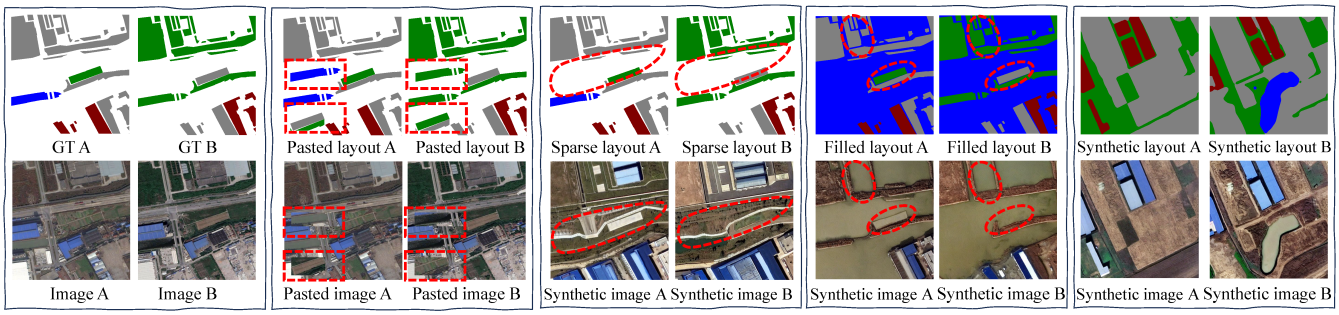



Figure 3: Comparison of data augmentation methods on two SCD datasets against competing approaches. We reproduce the results of Changen on the two SCD datasets, not using other segmentation datasets.

***Fine-tuning Transfer.*** In Fig. 5, we plot performance curves of various pre-trained models on the validation set, showing SCD (right) and BCD (left) metrics. Our ChangeDiff, pretrained on 10k synthetic change pairs with 8 semantic classes from LoveDA, achieves faster convergence and higher accuracy across different training sample ratios. Compared to ImageNet pretraining, ChangeDiff accelerates convergence and narrows the accuracy gap between models trained on 5% and 20% of the data. In contrast, Changen's binary pretraining underperforms on multiclass tasks despite




**a). Original data** | **b). Synthetic Data with copy-paste** | **c). Generative synthetic data with sparse layouts** | **d). Generative synthetic data with random filled layouts** | **e). Generative synthetic data with our synthetic layouts**

Figure 4: Visual comparison of augmentation methods for semantic change detection: a) Training data in SECOND dataset, with unlabeled white areas in GT. b) Self-Pair: Objects from other patches pasted onto the target image. c) Changen: Images synthesized from sparse-labeled layout. d) Random Layout: Images from randomly filled layout. e) Ours: Images from our synthetic layout. Red boxes in b), c), and d) highlight low image quality and unknown semantics due to poor layout.

| ***Zero-shot Transfer: Train on LoveDA and Test on SECOND$_{test}$*** | | | | | | |
|---|---|---|---|---|---|---|
| | SCD | | | BCD | | |
| Methods | $SeK_5$ | $Kappa_5$ | $mIoU_5$ | F1 | Pre. | Rec. |
| Copy-Paste (CP) | 4.7 | 39.6 | 39.9 | 42.7 | 46.1 | 39.7 |
| ControlNet + CP | 10.7 | 48.7 | 55.1 | 49.4 | 49.8 | 49.0 |
| Changen | 7.9 | 47.9 | 53.1 | 47.2 | 46.9 | 47.6 |
| ChangeDiff (Ours) | **13.6** | **55.9** | **60.1** | **55.5** | **55.1** | **55.9** |

Table 5: Comparison of zero-shot transfer setting. We evaluate the shared five semantic categories from LoveDA to SECOND: *Barren* → *ground*, *Forest* → *Tree*, *Agriculuture* → *Low Vegetation*, *Water* → *Water*, *Building* → *Building*. The SCD model used here is SCanNet.

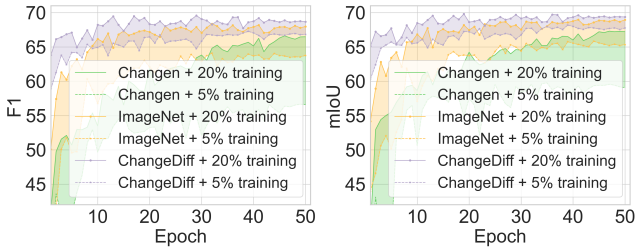



Figure 5: Impact of pre-training on model performance in SECOND dataset: ImageNet classification data. Changen binary synthetic change detection (90k pairs). Our multi-class synthetic change detection (10k pairs). All use SCanNet with ResNet-18 backbone.

using more data, as it focuses on building changes less suitable for diverse SECOND data.

**Ablation Studies.** We conduct an ablation study of MCDG-TP in three experiments: augmentation for SECOND and Landsat-SCD, and fine-tuning transfer on SECOND. We compare MCDG-TP with two variants: Copy-Paste and original T2I, as shown in Table 6. The results show MCDG-TP improves performance significantly. It boosts SeK by 1.6% and F1 by 1.1% in the Augment SECOND scenario, SeK by 2.7% and F1 by 1.3% in the Augment Landsat-SCD scenario, and SeK by 2.1% and F1 by 1.1% in Fine-tuning Transfer. These results prove MCDG-TP outperforms other methods in all experiments.

| Methods | Augment SECOND 100% | | Augment Landsat-SCD 100% | | Fine-tuning Transfer 100% | |
|---|---|---|---|---|---|---|
| | SeK | F1 | SeK | F1 | SeK | F1 |
| Baseline | 22.8 | 73.1 | 30.5 | 45.2 | 22.8 | 73.1 |
| Copy-Paste + L2I | 21.5 | 71.1 | 28.6 | 41.6 | 22.9 | 72.7 |
| Original T2L + L2I | 20.8 | 71.7 | 28.1 | 42.3 | 23.2 | 72.1 |
| MCDG-TP + T2L (Ours) + L2I | 24.4 (+1.6) | 74.1 (+1.1) | 33.2 (+2.7) | 46.5 (+1.3) | 24.9 (+2.1) | 74.3 (+1.1) |

Table 6: Ablation Studies on our MCDG-TP.

## Qualitative Analysis

***Comparison of Synthesis Quality.*** Fig. 4 compares semantic change detection data from different methods: b) Self-Pair: Object pasting causes foreground-background mismatches (red highlights). c) Changen (Sparse Layout): Sparse layout creates unclear regions (red highlights) with unknown semantics. d) Changen (Random Filled Layout): Random filling causes visible artifacts (red highlights). e) Ours: Our method produces high-quality images with better consistency and clarity.

## Conclusion

Change detection (CD) benefits from deep learning, but data collection and annotation remain costly. Existing generative methods for CD face issues with realism, scalability, and generalization. We introduce ChangeDiff, a new multi-temporal semantic CD data generator using diffusion models. ChangeDiff generates realistic images and simulates continuous changes without needing paired images or external datasets. It uses a text prompt for layout generation and a refinement loss to improve generalization. Future work could extend this approach to other CD tasks and enhance model scalability.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62271377, the National Key Research and Development Program of China under Grant Nos. 2021ZD0110400, 2021ZD0110404, the Key Research and Development Program of Shannxi (Program Nos. 2023YBGY244, 2023QCYLL28, 2024GX-ZDCYL-02-08, 2024GX-ZDCYL-02-17), the Key Scientific Technological Innovation Research Project by Ministry of Education, the Joint Funds of the National Natural Science Foundation of China (U22B2054).

## References

Arabi, M. E. A.; Karoui, M. S.; and Djerriri, K. 2018. Optical remote sensing change detection through deep siamese network. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 5041–5044. IEEE.

Chen, C.-F.; Son, N.-T.; Chang, N.-B.; Chen, C.-R.; Chang, L.-Y.; Valdez, M.; Centeno, G.; Thompson, C. A.; and Aceituno, J. L. 2013. Multi-decadal mangrove forest change detection and prediction in Honduras, Central America, with Landsat imagery and a Markov chain model. *Remote Sensing*, 5(12): 6408–6426.

Chen, H.; Li, W.; and Shi, Z. 2021. Adversarial instance augmentation for building change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Daudt, R. C.; Le Saux, B.; Boulch, A.; and Gousseau, Y. 2019. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187: 102783.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, H.; Ma, W.; Wu, Y.; Gong, M.; and Jiao, L. 2018. Local descriptor learning for change detection in synthetic aperture radar images via convolutional neural networks. *IEEE access*, 7: 15389–15403.

Feranec, J.; Hazeu, G.; Christensen, S.; and Jaffrain, G. 2007. Corine land cover change detection in Europe (case studies of the Netherlands and Slovakia). *Land use policy*, 24(1): 234–247.

Kadhim, N.; Mourshed, M.; and Bray, M. 2016. Advances in remote sensing applications for urban sustainability. *Euro-Mediterranean Journal for Environmental Integration*, 1(1): 1–22.

Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D.; and Ermon, S. 2023. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*.

Lei, Y.; Liu, X.; Shi, J.; Lei, C.; and Wang, J. 2019. Multiscale superpixel segmentation with deep features for change detection. *Ieee Access*, 7: 36600–36616.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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*, 234–241. Springer.

Seo, M.; Lee, H.; Jeon, Y.; and Seo, J. 2023. Self-pair: Synthesizing changes from single source for object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6374–6383.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*.

Yang, K.; Xia, G.-S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; and Zhang, L. 2021. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Yuan, P.; Zhao, Q.; Zhao, X.; Wang, X.; Long, X.; and Zheng, Y. 2022. A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images. *International Journal of Digital Earth*, 15(1): 1506–1525.

Zhang, L. e. a. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zheng, Z.; Ermon, S.; Kim, D.; Zhang, L.; and Zhong, Y. 2024. Changen2: Multi-Temporal Remote Sensing Generative Change Foundation Model. *arXiv preprint arXiv:2406.17998*.

Zheng, Z.; Ma, A.; Zhang, L.; and Zhong, Y. 2021. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15193–15202.

Zheng, Z.; Tian, S.; Ma, A.; Zhang, L.; and Zhong, Y. 2023. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In *Proceed-*

*ings of the IEEE/CVF International Conference on Computer Vision*, 21818–21827.