Real-World Knowledge-Guided Change Data Synthesis for Remote Sensing

Yaoyi Qi[1,*], **Xingxing Weng**[1,*], **Chao Pang**[1,†], **Yongkang Cui**[1]
XiangYu Hao[1], **Xiaokang Zhang**[1], **Guibo Zhu**[2,3], **Gui-Song Xia**[1,4]

[1]*School of Artificial Intelligence, Wuhan University* [2]*Wuhan AI Research*
[3]*Institute of Automation, University of Chinese Academy of Sciences* [4]*Institute for Math & AI, Wuhan*

pangchao@whu.edu.cn

Abstract

Change data synthesis provides a cost-effective solution for expanding training data and improving the performance of change detection models. However, existing synthesis methods typically rely on handcrafted rules to simulate changes, where limited coverage of class transitions restricts the diversity of synthesized data, while predefined transition designs limit their flexibility in accommodating varied change types. In this work, we introduce KnowChange, a knowledge-guided change data synthesis framework that leverages pretrained vision-language models as knowledge sources to reason about plausible change locations and class transitions from pre-change scenes and desired change types. By integrating knowledge-guided change simulation with generalizable synthesis models, KnowChange enables flexible synthesis of diverse change types within a unified framework. Extensive experiments demonstrate that KnowChange-generated data consistently outperforms existing synthetic datasets in both synthetic-to-real transfer and synthetic data augmentation, despite being generated at a compact scale. Further analyses show that the knowledge-guided change simulation can be seamlessly integrated into existing synthesis pipelines and enhance the downstream utility of synthesized data.

Keywords: Remote sensing, Change detection, Synthetic data generation, Knowledge-guided synthesis
Web: https://knowchange.vercel.app/
Code: https://github.com/LINGQI711/KnowChange

1 Introduction

Change data synthesis aims to automatically generate bi-temporal images with pixel-level change masks, and optionally semantic masks for each timestamp to characterize class transitions between the two images. By reducing reliance on expensive manual annotation, change data synthesis provides a scalable solution for increasing training data diversity and has attracted growing attention in remote sensing [1, 2, 3].

Existing change data synthesis methods [1, 3, 4] typically start from single-temporal images with semantic masks and simulate future changes to obtain post-change semantic masks, which are then used to guide the generation of post-change images. In this pipeline, change simulation is crucial, as it determines where

*Equal contribution †Corresponding author

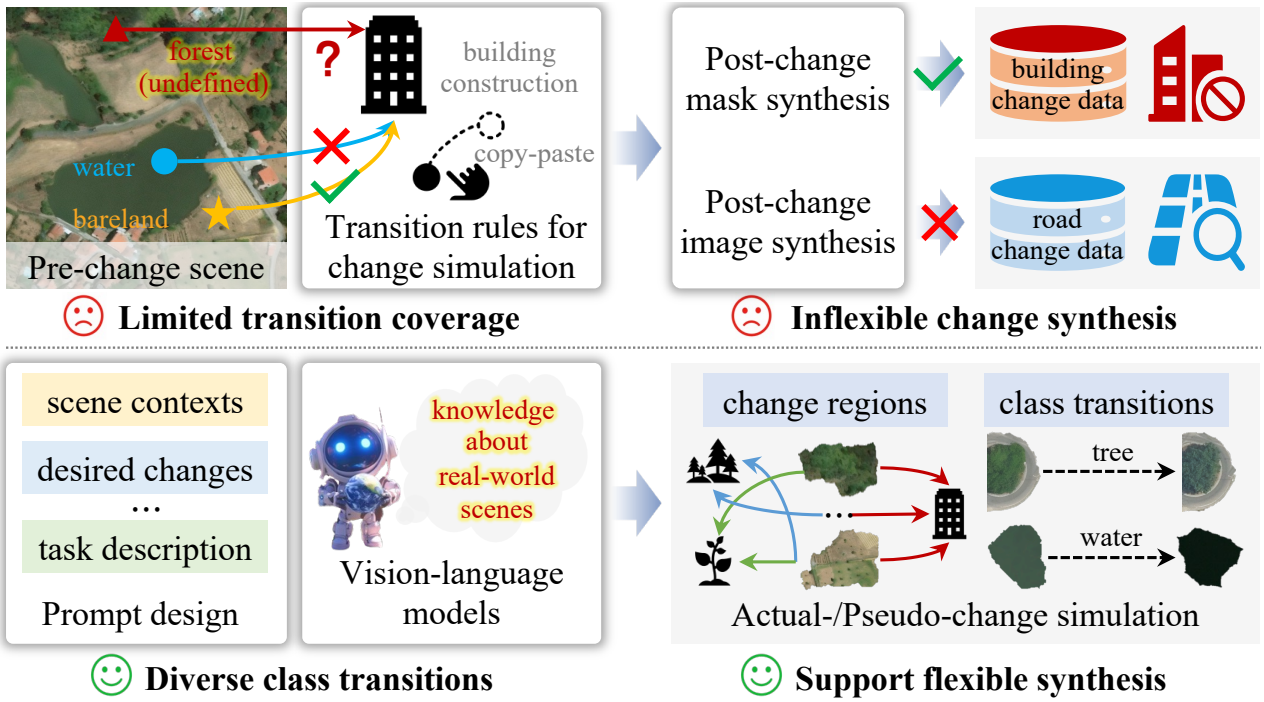


Figure 1 **Top**: Existing synthesis methods rely on handcrafted rules to simulate changes. Limited coverage of class transitions restricts the diversity of synthesized data, while predefined transition designs limit flexibility in accommodating varied change types. **Bottom**: KnowChange leverages pretrained vision-language models as knowledge sources to reason about plausible change regions and class transitions given pre-change scene contexts and user-specified change types, enabling flexible synthesis of diverse change data.

changes occur and what class transitions take place. Many methods implement this simulation through handcrafted rules that specify class transitions and guide region-level manipulations (e.g., copy-paste). Although rule-based change simulation has enabled the construction of large-scale synthetic datasets, it suffers from two fundamental limitations, as illustrated in Fig. 1.

First, handcrafted rules usually cover only a limited set of class transitions. Due to the large field of view and diverse land-cover categories in remote sensing images, real-world changes often involve a broader transition space than predefined rules can capture. For example, existing rules allow only a few land-cover categories, such as bareland, rangeland, and developed land, to transition into buildings, whereas real-world urban expansion can also involve transitions from forests or other land-cover categories into buildings. Such limited transition coverage restricts the diversity of synthesized change data, potentially constraining the performance of change detection models trained on synthetic data.

Second, desired change types vary across application scenarios. For example, urban development monitoring involves building construction or demolition, whereas transportation monitoring concerns road-related changes. However, rule-based change simulation relies on predefined transitions with fixed patterns, limiting its flexibility in accommodating new change types. Supporting new change types requires redesigning the transition rules and adapting the image synthesis models accordingly.

Existing change simulation largely relies on human knowledge about real-world changes encoded in handcrafted rules. However, manually enumerating and encoding such knowledge into rules for diverse and evolving change scenarios is inherently difficult. Recent vision-language models (VLMs) pretrained on large-scale vision-language corpora capture rich visual-semantic knowledge about real-world scenes, including object categories, their relationships, and scene contexts. This motivates us to leverage VLMs as

a knowledge source to infer where changes can occur and what class transitions are plausible, enabling real-world knowledge-guided change simulation.

To instantiate this idea, we present a knowledge-guided change data synthesis framework for remote sensing, named KnowChange (Fig. 3). Given pre-change images with rich semantic annotations and user-specified change types, KnowChange first prompts a pretrained VLM to infer plausible change locations and class transitions, producing a global layout of the post-change semantic mask. This change simulation eliminates the need for manually predefined transition rules, enabling a more diverse range of class transitions by reasoning over scene contexts. Moreover, desired change types are directly incorporated into the simulation process through prompts, avoiding repeated customization of transition rules across different application scenarios.

The inferred layout is then instantiated by a layout-to-mask model, which refines the shapes of changed object and improves local object-context compatibility to produce pixel-level post-change semantic masks. Following existing synthesis pipelines, a mask-to-image model generates post-change images conditioned on the synthesized semantic masks. To support flexible synthesis of evolving change types, we curate a large-scale multi-category semantic segmentation dataset and develop effective training strategies for both models, allowing them to capture rich visual priors of object appearance. Consequently, KnowChange can generate valid object shapes and high-fidelity appearances for diverse change types without additional retraining.

Using KnowChange, we synthesize three datasets, including Know-BCD, Know-SEC, and Know-HR for building and semantic change detection. Extensive experiments demonstrate that models trained on our synthetic datasets outperform those trained on existing ones (Fig. 2), achieving an average IoU gain of **6.64** on four building change detection benchmarks and an average F1 gain of **6.78** on two semantic change detection benchmarks. Furthermore, ablation studies show that knowledge-guided change simulation can be readily integrated into existing methods, such as HySCDG [4] and Changen2 [1], substantially improving the effectiveness of synthesized data for downstream change detection (Fig. 6).

Our contributions are summarized as follows:

- We introduce knowledge-guided change simulation that exploits pretrained VLMs to infer plausible change locations and class transitions, addressing the limited transition coverage of handcrafted rules and enhancing the diversity of synthesized change data.
- We present KnowChange, a flexible change data synthesis framework that combines VLM-based change reasoning with generative models, enabling adaptive synthesis of user-specified change types without repeated customization of the synthesis pipeline.
- We create three synthetic datasets for building and semantic change detection and demonstrate that training with our datasets improves detection accuracy and generalization over existing synthesis datasets.

2 KnowChange

2.1 Framework Overview

Let $I_{\text{pre}} \in \mathbb{R}^{H\times W\times 3}$ and $S_{\text{pre}} \in \mathbb{R}^{H\times W}$ denote the pre-change image and its semantic mask, respectively, where each pixel in S_{pre} indicates its semantic category. Given pre-change scene information (I_{pre}, S_{pre}) and

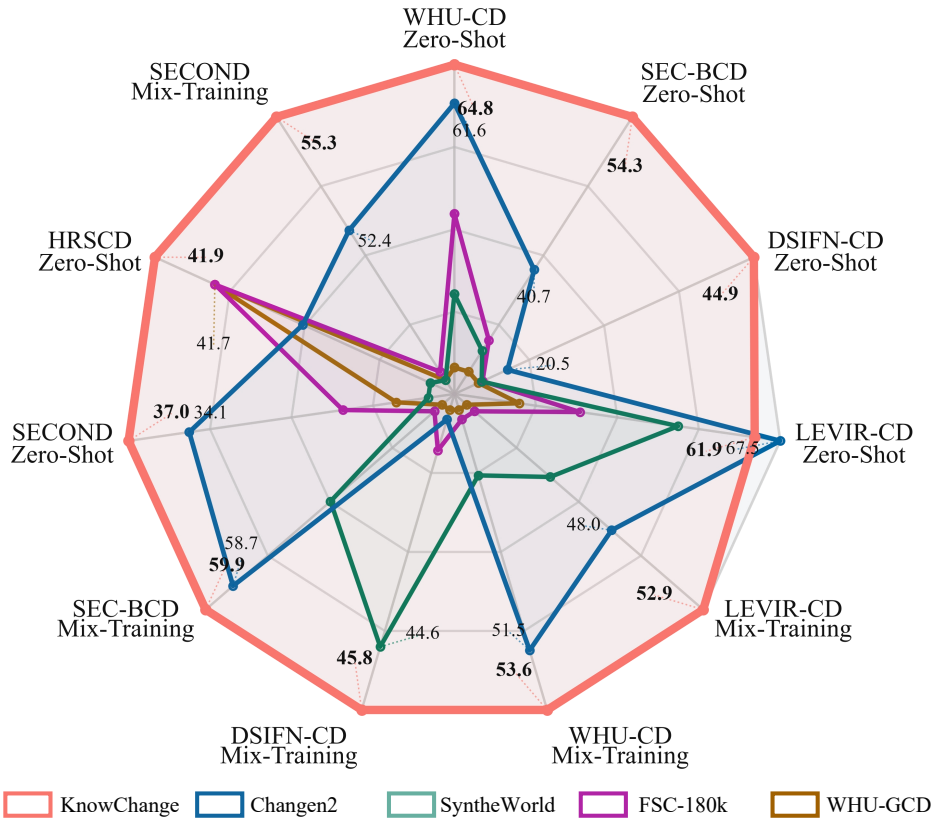


Figure 2 Downstream performance comparison of models trained with change data synthesized by different methods.

a set of user-specified change types $C = \{c_i\}_{i=1}^{N_c}$, KnowChange aims to synthesize the corresponding post-change image $I_{\text{post}} \in \mathbb{R}^{H\times W\times 3}$ and semantic mask $S_{\text{post}} \in \mathbb{R}^{H\times W}$. The binary change mask $M \in \{0,1\}^{H\times W}$ is obtained by comparing S_{pre} and S_{post}. As illustrated in Fig. 3, KnowChange consists of two key components: knowledge-guided change simulation and generalizable semantic-guided synthesis. The change simulation is performed by prompting a pretrained VLM to infer plausible change regions r_i and corresponding post-change categories c_i, producing a global layout $\mathcal{G}_{\text{chg}} = \{(r_i, c_i)\}_{i=1}^{N_r}$. The semantic-guided synthesis component consists of a layout-to-mask (L2M) model and a mask-to-image(M2I) model. Given S_{pre} and $\mathcal{G}_{\text{chg}}$, the L2M model generates a pixel-level post-change semantic mask S_{post} by refining object shapes and ensuring local object-context compatibility. The M2I model then synthesizes the post-change image I_{post} conditioned on S_{post} and I_{pre}.

2.2 Knowledge-Guided Change Simulation

Existing rule-based simulation requires predefined class transitions, limiting the diversity of synthesized changes. We instead formulate change simulation as knowledge-guided reasoning, where a pretrained VLM infers plausible changes from scene contexts. Specifically, the VLM takes the pre-change image I_{pre} and semantic mask S_{pre} as inputs, along with a carefully designed prompt containing task description, desired change categories $\{c_i\}_{i=1}^{N_c}$, and the mapping between semantic classes and colors in S_{pre} (detailed prompt design is provided in the Appendix). Guided by the prompt, the VLM conducts actual- and pseudo-change reasoning.

Actual-Change Simulation: Actual-change reasoning involves determining changed regions and their

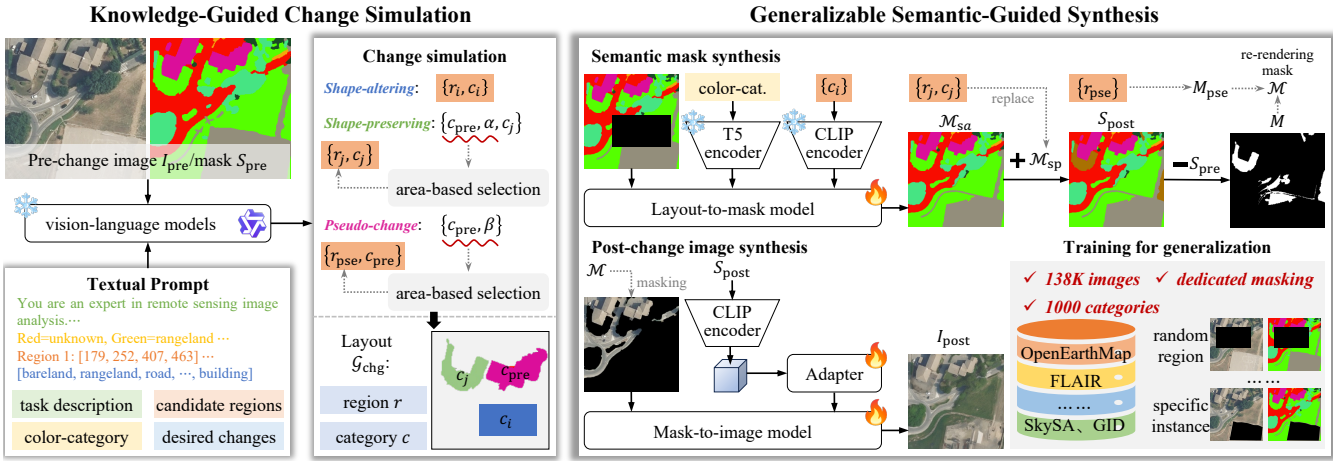


Figure 3 Overview of KnowChange, a knowledge-guided framework for change data synthesis. Given a pre-change image I_{pre}, its semantic mask S_{pre}, and a textual prompt P, Knowledge-Guided Change Simulation reasons about actual and pseudo changes to construct the change layout $\mathcal{G}_{chg}$. Generalizable Semantic-Guided Synthesis generates $\mathcal{M}_{sa}$, combines it with $\mathcal{M}_{sp}$ to obtain S_{post} and $\mathcal{M}$, and uses them to guide the synthesis of the post-change image I_{post}.

corresponding post-change categories ($\{(r_i, c_i)\}_{i=1}^{N_r}$). Depending on whether the post-change region preserves the shape of the pre-change region, we categorize actual changes into two complementary transition modes: shape-preserving and shape-altering transitions. In shape-preserving transitions, the shape of the original region is preserved, while its semantic category evolves from c_{pre} to c_i. For example, structured farmland becoming abandoned may transition into grassland, where the original parcel boundary can be preserved. In this mode, the VLM predicts c_i, together with the corresponding pre-change category c_{pre} and a region selection ratio $\alpha \in (0, 1)$, from which r_i is derived. Formally, let $\mathcal{K} = \{k_j\}_{j=1}^{N_k}$ denote the connected components of category c_{pre} in S_{pre}, sorted in descending order of area. The number of selected components is determined as $n = \lceil \alpha N_k \rceil$, and the changed region is obtained by taking the union of the first n components, i.e., $r_i = \bigcup_{j=1}^{n} k_j$.

In contrast, shape-altering transitions require newly instantiated regions according to post-change categories. Typical examples include newly constructed buildings emerging on grassland, where no corresponding regions exist in the pre-change scene. To enable this transition, we provide the VLM with randomly generated candidate rectangular regions $\mathcal{B} = \{b_j\}_{j=1}^{N_b}$ in the prompt. The VLM then selects appropriate boxes and assigns post-change categories, producing (r_i, c_i). Combining both transition modes yields the final change layout for post-change semantic mask synthesis.

Pseudo-Change Simulation: In real-world datasets, variations in imaging conditions may cause unchanged regions to exhibit visual differences across time while preserving their semantic categories (e.g., grass appearing denser or sparser), a phenomenon known as pseudo-change. To improve synthesis realism, we simulate pseudo changes alongside actual changes. Specifically, the VLM identifies pre-change categories c_{pre} that may exhibit pseudo changes and estimates their region perturbation ratios $\beta \in (0, 1)$. Using the same region selection strategy as shape-preserving transitions, pseudo-change regions r_{pse} are derived based on β. The obtained pairs (r_{pse}, c_{pre}) are used to guide the M2I model to perform category-preserving appearance reconstruction during post-change image synthesis.

2.3 Generalizable Semantic-Guided Synthesis

The change simulation provides only a change layout specifying where and what changes occur, rather than the complete pixel-level post-change semantic mask required for image synthesis. The recent method [2] generates S_{post} by sampling object shapes of post-change categories c_i from manually maintained mask libraries and pasting them into changed regions r_i of S_{pre}. However, such strategies are limited by shape diversity and cannot guarantee spatial coherence with surrounding areas, resulting in unrealistic scenes, e.g., newly constructed buildings may not seamlessly blend with adjacent land covers. We therefore introduce a semantic mask synthesis model to expand the sparse change layout into a complete post-change semantic mask, which subsequently guides image synthesis.

Semantic Mask Synthesis: The change layout is generated through two transition modes: shape-preserving and shape-altering transitions. Accordingly, the post-change semantic mask is obtained by integrating the masks derived from the two transition layouts. Let $\mathcal{M}_{\text{sp}}$ and $\mathcal{M}_{\text{sa}}$ denote the masks derived from the shape-preserving and shape-altering layouts, respectively. The shape-preserving mask $\mathcal{M}_{\text{sp}}$ is directly obtained by replacing the semantic labels of regions r_i in S_{pre} with their corresponding post-change categories c_i. In contrast, the shape-altering layout only provides coarse region constraints (i.e., candidate boxes). Therefore, we employ a layout-to-mask (L2M) model, which takes S_{pre} with changed regions masked out as input, to instantiate these regions and generate $\mathcal{M}_{\text{sa}}$. Specifically, L2M model adopts the FLUX.1 [5] architecture, which is equipped with two text encoders. The T5 encoder [6] provides rich representations for complex textual descriptions, while the CLIP text encoder [7] provides category-level representations aligned with visual concepts. Accordingly, the category-color mapping in S_{pre} is encoded by the T5 encoder, while the desired post-change categories are encoded by the CLIP text encoder. The resulting textual representations are used as conditioning signals to generate $\mathcal{M}_{\text{sa}}$. Finally, S_{post} is obtained by replacing the corresponding regions in $\mathcal{M}_{\text{sa}}$ with $\mathcal{M}_{\text{sp}}$.

Post-Change Image Synthesis: The change mask M is computed from the semantic difference between S_{pre} and S_{post}, while the pseudo-change mask M_{pse} is derived from the VLM-inferred pseudo-change regions r_{pse}. The two masks are combined to obtain the re-rendering mask $\mathcal{M} = M \bigcup M_{\text{pse}}$. The regions indicated by $\mathcal{M}$ are marked out from I_{pre}, which is then fed into a mask-to-image (M2I) diffusion model to synthesize I_{post} conditioned on S_{post}. Since pseudo-change regions preserve semantic categories between S_{pre} and S_{post}, the M2I model re-renders objects with unchanged categories while introducing subtle appearance variations. In contrast, actual change regions involve class transitions and require appearances consistent with the post-change categories. Existing change data synthesis methods [8] use RGB-encoded semantic masks as conditions for image synthesis. However, such designs require a fixed color assignment for each category, making them difficult to scale to scenarios with diverse categories, where assigning visually distinguishable colors to a large number of categories becomes impractical. To address this limitation, we employ a CLIP text encoder to generate spatially dense embeddings from semantic categories in S_{post}. These embeddings are further transformed by an adapter into control features for the diffusion model. The adapter consists of a 1×1 channel projection, three stride-2 convolutional blocks, and a zero-initialized output convolution.

Training for Generalization: Although semantic-guided synthesis enables flexible change data generation, its generalization ability is constrained by the diversity of training data. Existing methods [1, 4, 8] typically train synthesis models on datasets with limited category coverage, restricting their ability to synthesize varied change categories. To enhance category-level generalization, we curate a large-scale semantic segmentation corpus by consolidating existing datasets, including OpenEarthMap [9], FLAIR [10], Vaihingen [11], Potsdam [11], GID [12], and SkySA [13], forming a corpus of 138K remote sensing images. Dataset statistics are provided in Table 1. These datasets provide over 1,000 object categories, enabling the

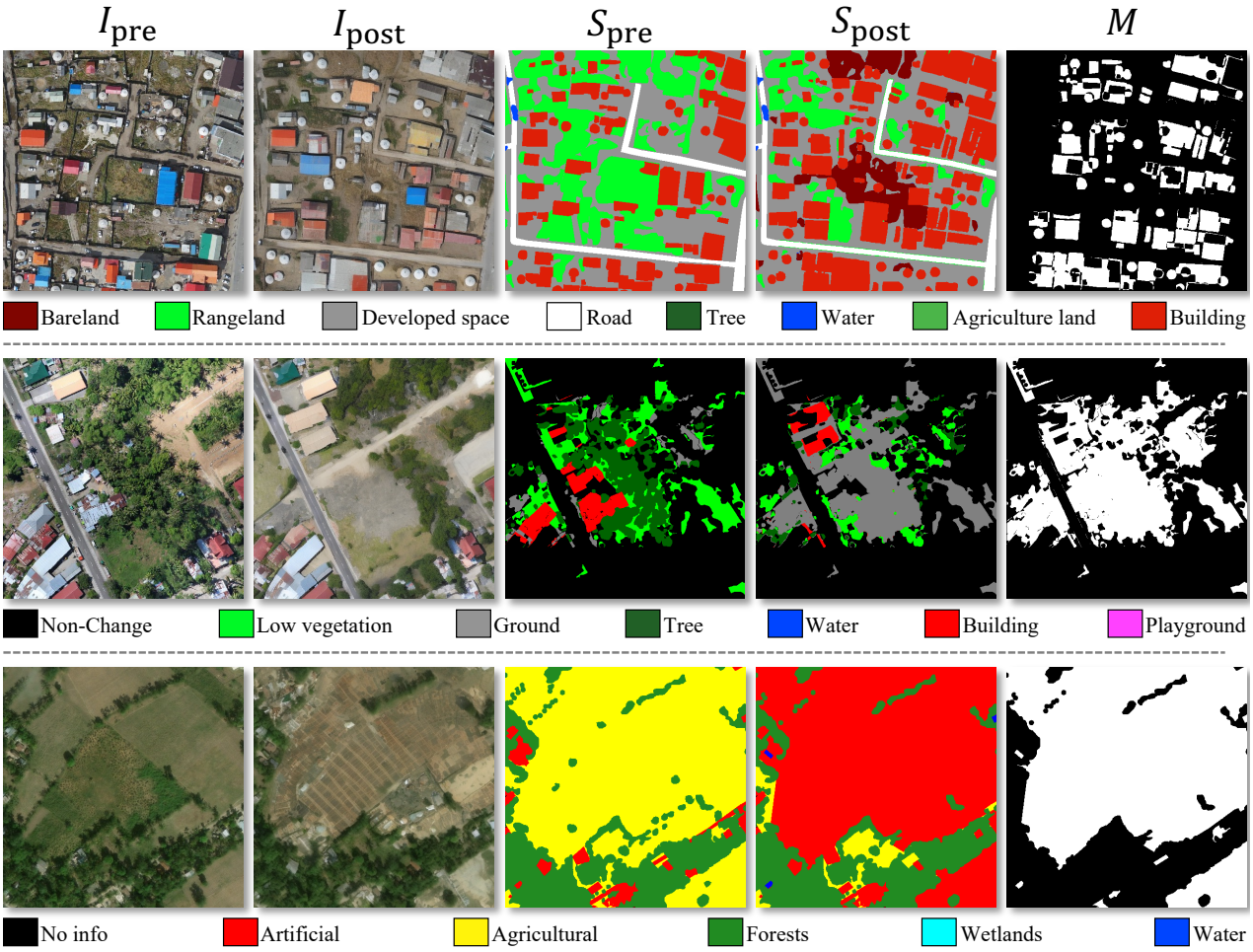


Figure 4 Examples from Know-BCD, Know-SEC, and Know-HR (top to bottom).

L2M and M2I models to learn rich object priors.

During training, masked inputs are prepared to match the inference inputs of the L2M and M2I models. Inspired by image editing practices [14, 15], we design dedicated masking procedures for the two models. For L2M training, masked semantic masks are generated using category-aware instance masking and random region masking. The text conditions describe either the masked category name or the dominant categories within the masked region, enabling the model to learn category-specific shape priors while maintaining spatial coherence across different categories. For M2I training, masked images are generated with three masking granularities: instance-level masking for category-specific appearance learning, region-level masking for cross-category boundary synthesis, and global-level masking for scene-level reconstruction.

For $k \in \{S, I\}$, L2M and M2I are trained under a unified conditional regression objective:

$$\mathcal{L}_k = \mathbb{E}_{(z_0^k, C_k)\sim\mathcal{D}_k, t, \epsilon}\left[\left\| f_{\theta_k}(z_t^k, t, C_k) - y_k \right\|_2^2\right], \tag{1}$$

where $\mathcal{D}_k$ denotes the corresponding training distribution. For L2M, z_0^s denotes the latent representation of the complete semantic mask S_{pre}, and $z_t^s = (1-t)z_0^s + t\epsilon$, $f_{\theta_s} = v_{\theta_s}$, $y_s = \epsilon - z_0^s$, and $C_s = (S_{\text{pre}}, \mathcal{M}, P)$, yielding a flow-matching objective. For M2I, z_0^I is the latent of I_{pre}, z_t^I is its noisy latent, $f_{\theta_I} = \epsilon_{\theta_I}$, $y_I = \epsilon$, and $C_I = (z_0^I, \mathcal{M}, S_{\text{post}})$, yielding a noise-prediction objective. The M2I objective jointly supervises

Table 1 Summary of remote sensing datasets used for training the layout-to-mask and mask-to-image models.

Dataset	Resolution (m)	Bands	# Classes	# Samples
OpenEarthMap[9]	0.25 ~ 0.5	RGB	8	11,662
FLAIR[10]	0.2	RGB + NIR	19	61,712
Vaihingen[11]	0.09	IR-R-G	6	864
Potsdam[11]	0.05	RGB + IR	6	7,406
GID[12]	1	RGB + NIR	15	23,100
SkySA[13]	Variable	RGB	1,763	33,775

ControlNet and the adapter to learn spatial and semantic guidance, respectively. Here, $\mathcal{M}$ is sampled using the three masking procedures described above, t denotes the normalized timestep, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

2.4 Flexible Change Data Synthesis

Unlike existing methods that require manually designed transition rules and repeated retraining of synthesis models for different change types, a single trained KnowChange framework can flexibly synthesize change data with diverse change types. We combine OpenEarthMap and FLAIR as source datasets for change data synthesis. These datasets contain heterogeneous label taxonomies and provide 26 semantic categories in total. Based on this rich semantic annotation, KnowChange synthesizes three datasets for building and semantic change detection, namely Know-BCD, Know-SEC, and Know-HR. Know-SEC follows the change categories defined in SECOND [16], while Know-HR follows those defined in HRSCD [17]. Each dataset contains 10K samples, with examples shown in Fig. 4. Detailed statistics, comparisons with existing synthetic datasets, additional examples, and details of the synthesis process are provided in Appendix.

3 Experiments

3.1 Experimental Setup

Datasets and Evaluation Metrics: We evaluate the utility of synthesized data on six widely used change detection benchmarks, including four building change detection (BCD) datasets (LEVIR-CD [18], WHU-CD [19], DSIFN-CD [20], and SEC-BCD [16]) and two semantic change detection (SCD) datasets (SECOND and HRSCD). We report F1-score and IoU for BCD, and F1-score, mIoU, SCS [21] and SeK [16] for SCD.

Downstream Models: For downstream evaluation, we train representative change detection models on synthesized datasets and test them on real-world benchmarks. Specifically, we adopt ChangeFormer [22] for BCD and Change3D [23] for SCD, trained for 42K/30K iterations with batch sizes of 24/8, respectively.

Synthesis Model Details: The L2M adopts FLUX.1-Fill as the backbone and is fine-tuned with LoRA [24] (rank 32) using AdamW for 30 epochs with a batch size of 16 and a learning rate of 5×10^{-5}. For the M2I model, we adopt an SD-v1.5 [25] model fine-tuned on remote sensing images [4], and replace the image-conditioned encoder with a CLIP text encoder. The U-Net and ControlNet [26] are optimized with learning rates of 2×10^{-5} and 5×10^{-6}, respectively. The M2I model is trained for 20 epochs with a batch size of 32. Both models are trained on the collected 138K samples with 512×512 inputs using two NVIDIA 96G H20 GPUs. More training details are provided in Appendix.

3.2 Downstream Utility of Synthesized Data

Table 2 Synthetic-to-real transfer results on building change detection benchmarks. *Average* indicates the mean performance over four benchmarks. The best and second-best results are shown in bold and underline, respectively.

Dataset	Source Task	# Samples	LEVIR-CD		WHU-CD		DSIFN-CD		SEC-BCD		Average	
			IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
WHU-GCD	SCD	25K	1.26	2.48	1.67	3.30	2.91	5.66	5.36	10.19	2.80	5.40
Changen2-S9	SCD	27K	1.99	3.90	3.80	7.32	4.73	9.04	6.69	12.54	4.30	8.20
FSC-180k	SCD	180K	11.19	20.13	33.50	50.19	5.49	10.40	16.40	28.18	16.64	27.22
Know-SEC	SCD	**10K**	30.81	47.10	34.95	51.79	**32.38**	**48.92**	29.54	45.61	31.92	48.36
SyntheWorld	BCD	40K	28.66	44.55	23.11	37.54	5.08	9.66	14.23	24.91	17.77	29.17
Changen2-S1	BCD	15K	**50.89**	**67.45**	44.48	61.57	11.40	20.48	25.52	40.67	33.07	47.54
Know-BCD	BCD	**10K**	44.81	61.89	**47.87**	**64.75**	28.93	44.88	**37.24**	**54.27**	**39.71**	**56.44**

Table 3 Synthetic-to-real transfer results on semantic change detection benchmarks. *Average* indicates the mean performance over two benchmarks. The best and second-best results are shown in bold and underline, respectively.

Dataset	#Samples	SECOND			HRSCD			Average			Δ F1 (vs. ours)
		F1	mIoU	SCS	F1	mIoU	SCS	F1	mIoU	SCS	
WHU-GCD	25K	15.32	19.85	18.06	41.71	39.10	**36.12**	28.52	29.48	27.09	-10.97
FSC-180k	180K	22.95	50.40	23.84	39.36	**53.98**	31.46	31.16	52.19	27.50	-8.33
Changen2-S9	27K	34.06	**57.75**	26.70	31.35	41.59	17.81	32.71	49.67	22.26	-6.78
Know-SEC/HR	**10K**	**37.03**	55.15	**28.31**	**41.94**	50.78	32.16	**39.49**	**52.97**	**30.24**	-

Table 4 Performance comparison of synthetic data augmentation with 5% real training data from different benchmarks. For BCD tasks, models are evaluated on four BCD benchmarks, and results are reported as average IoU and F1. For SECOND, models are evaluated on the SECOND test set. The best and second-best results are shown in bold and underline, respectively.

Dataset	5% LEVIR-CD		5% WHU-CD		5% DSIFN-CD		5% SEC-BCD		Dataset	5% SECOND			
	IoU	F1	IoU	F1	IoU	F1	IoU	F1		F1	mIoU	SeK	SCS
FSC-180k	15.80	26.32	17.81	29.77	21.96	35.84	21.96	33.56	No Syn.	50.84	65.05	10.53	38.81
SyntheWorld	28.97	43.27	29.02	41.75	28.95	44.62	36.09	52.79	FSC-180k	41.34	60.00	4.79	26.95
Changen2-S1	32.61	48.01	36.62	51.54	17.49	29.74	42.56	58.66	Changen2-S9	52.37	66.21	11.85	**39.93**
Know-BCD	**36.28**	**52.88**	**37.13**	**53.59**	**29.74**	**45.77**	**43.20**	**59.94**	**Know-SEC**	**55.27**	**67.24**	**13.89**	39.87

Synthetic-to-Real Transfer: We train change detection models solely on synthesized datasets and evaluate their generalization on real-world benchmarks. Table 2 reports synthetic-to-real transfer results on BCD benchmarks. Following prior practices [27], we filter building-related samples from synthesized datasets originally designed for SCD and use them for BCD model training. Among the four SCD-oriented synthesized datasets, the Know-SEC-trained model generalizes best to real-world BCD benchmarks despite using the fewest training samples, outperforming models trained with other SCD-oriented datasets by over 15/21 points in average IoU/F1. Among the three BCD-oriented synthesized datasets, training with Know-BCD achieves the best transfer results with only 10K samples. Although lower than Changen2-S1 [1] on LEVIR-CD, it substantially outperforms Changen2-S1 on the other three benchmarks, yielding an average IoU gain of 6.64 points. We further extend the evaluation to SCD benchmarks. Table 3 reports the results. Existing methods rely on handcrafted transition rules, limiting change diversity and introducing benchmark-specific biases. Consequently, models trained on synthetic data may generalize well to one benchmark but fail on another, as observed on WHU-GCD [2]. In contrast, KnowChange enables flexible

benchmark-specific synthesis, yielding customized training data for SECOND and HRSCD. Models trained on our synthesized datasets outperform those trained on existing ones, achieving an average F1 improvement of 6.78 points.

Synthetic Data Augmentation: Beyond direct synthetic-to-real transfer, we evaluate synthesized data augmentation under limited real-data regimes. Each synthetic dataset is combined with 5% real training data from one BCD benchmark at a time, and the resulting models are evaluated on all four BCD benchmarks. As shown in Table 4, models trained with Know-BCD consistently outperform those trained with other synthetic datasets across all four real-data settings. The largest gains reach 3.67 and 4.87 points in average IoU and F1, respectively. We further conduct this augmentation experiment on the SECOND benchmark. As shown in Table 4, Know-SEC achieves the best augmentation performance among existing synthetic datasets. Overall, the improvements in both transfer and augmentation experiments validate the effectiveness of KnowChange for generating high-quality synthetic change data.

Synthetic Data Quality Analysis: To understand the performance gains, we assess synthesized data quality using FID and KID metrics, with SECOND as the reference distribution. As shown in Fig. 5, KnowChange-generated data achieves lower FID and KID scores than existing synthetic data, indicating better alignment with real-world change data.

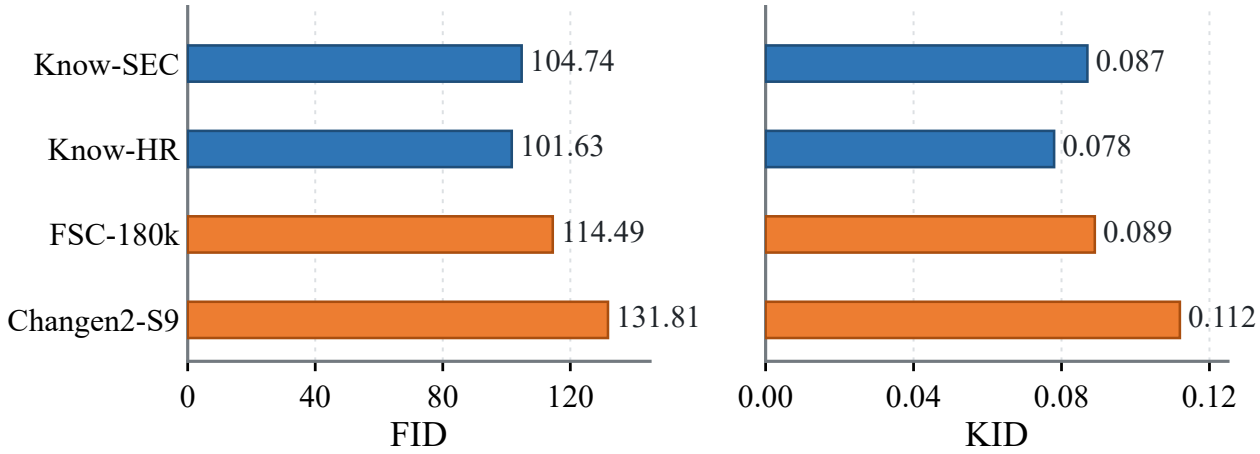


Figure 5 FID and KID comparison of synthetic datasets.

3.3 Ablation Studies

Effect of Knowledge-Guided Change Simulation: Since generative models are essential components of the synthesis pipeline but are not the focus of our contribution, we keep them fixed to isolate the effect of knowledge-guided change simulation (Table 5). For building change synthesis, we first establish a baseline using random copy-paste to generate post-change semantic masks. This strategy often produces unrealistic changes, e.g., pasting buildings onto water, leading to an average IoU of only 3.96. We then replace copy-paste with shape-altering transition (*SAT*). Although *SAT* cannot fully eliminate change location errors, the L2M model improves object-context compatibility, e.g., generating buildings near water boundaries rather than directly on water surfaces, increasing average IoU to 37.58. Adding shape-preserving transition (*SPT*) provides marginal gains, mainly because its generated building-to-building transitions are rarely considered in existing BCD benchmarks. Incorporating pseudo-change simulation (*PCS*) further improves average IoU by over 6.5 points. Finally, pretrained VLMs for knowledge-guided reasoning further boost the average IoU to 46.34. On SCD, removing VLM guidance causes a 9.34-point

F1 drop, validating the importance of knowledge-guided reasoning for realistic and diverse change simulation.

Table 5 Ablation study of knowledge-guided change simulation. Generative models are fixed, and different components are progressively added to evaluate their contributions.

Method	LEVIR-CD (IoU)	WHU-CD (IoU)	Average
Copy-Paste	4.10	3.82	3.96
SAT	33.95	41.20	37.58
+ *SPT*	35.33	40.31	37.82
+ *PCS*	44.07	44.62	44.35
+ *VLM*	**44.81**	**47.87**	**46.34**
SECOND (F1/mIoU/SCS)			
SAT+SPT+PCS	24.07	49.93	22.97
+VLM	**33.41**	**50.39**	**29.50**

Impact of VLMs: To investigate the impact of different VLMs on change simulation, we replace the reasoning model with different pretrained VLMs and evaluate the resulting synthesized data on the SECOND benchmark. As shown in Table 6, all VLM-guided strategies consistently outperform the w/o VLM baseline. Among different VLMs, Qwen3-VL [28] achieves the best F1 (33.41) and SCS (29.50), while DouBao-2.0-mini [29] and GLM-4.6V [30] obtain slightly higher mIoU. These results show that pretrained VLMs serve as effective knowledge providers for change simulation, and VLM selection influences the quality of synthesized data. We adopt Qwen3-VL as the default reasoning model in all experiments.

Table 6 Impact of VLMs on knowledge-guided change simulation evaluated on the SECOND benchmark.

VLM Guidance	F1	mIoU	SCS	ΔF1 (w/o VLM)
w/o VLM	24.07	49.93	22.97	-
GLM-4.6V	29.84	<u>53.16</u>	25.90	5.77
DouBao-2.0-mini	<u>31.72</u>	**53.66**	<u>25.98</u>	<u>7.65</u>
Qwen3-VL	**33.41**	50.39	**29.50**	**9.34**

Plug-and-Play Change Simulation: We integrate our knowledge-guided change simulation with existing synthesis methods, including HySCDG and Changen2, by replacing their change mask generation while keeping image generation models unchanged. As shown in Fig. 6, this replacement improves the performance of models trained on synthesized data across both BCD and SCD tasks. The improvement is particularly significant when combined with Changen2, achieving a 23-point mIoU gain on SECOND and over 21-point IoU gains on both LEVIR-CD and WHU-CD. These results demonstrate the potential of our change simulation as a plug-and-play component for existing synthesis methods.

Scaling Analysis of Synthetic Data: We train models with different amounts of Know-BCD, ranging from 1% to 100%, and evaluate them on four BCD benchmarks. As shown in Fig. 7, increasing synthesized data from 1% to 5% brings the most significant performance gains across four BCD benchmarks, suggesting that even a small amount of synthesized data provides valuable information for learning change patterns. Further increasing synthesized data consistently improves performance on three benchmarks, demonstrating the effectiveness of synthetic data scaling. On LEVIR-CD, although a performance drop is observed with 75% data, the performance recovers with the full dataset. Overall, these results demonstrate

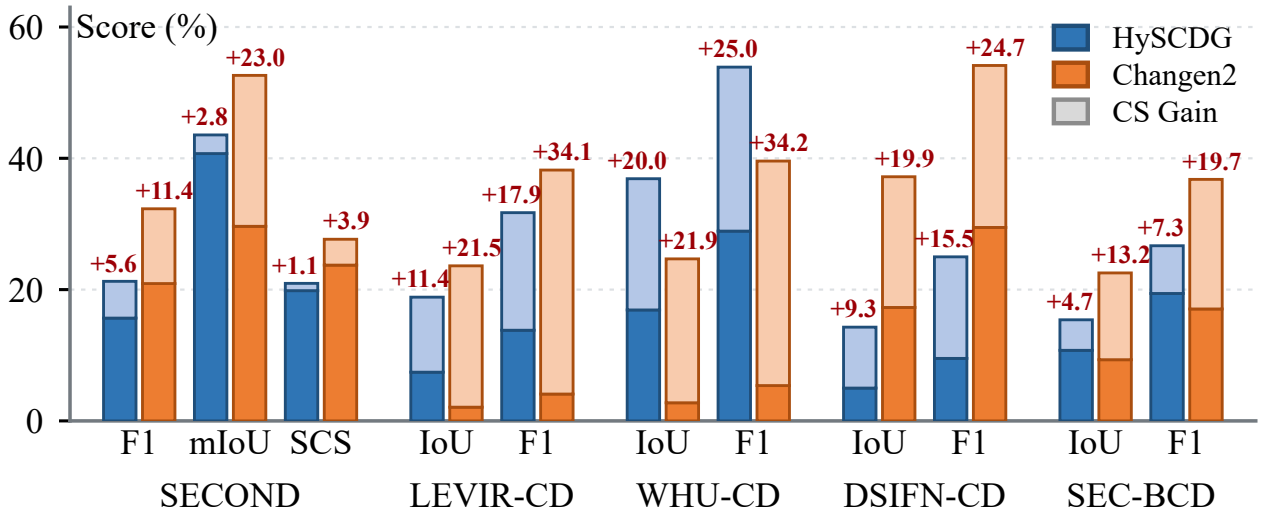


Figure 6 Performance comparison of existing synthesis methods with and without our change simulation.

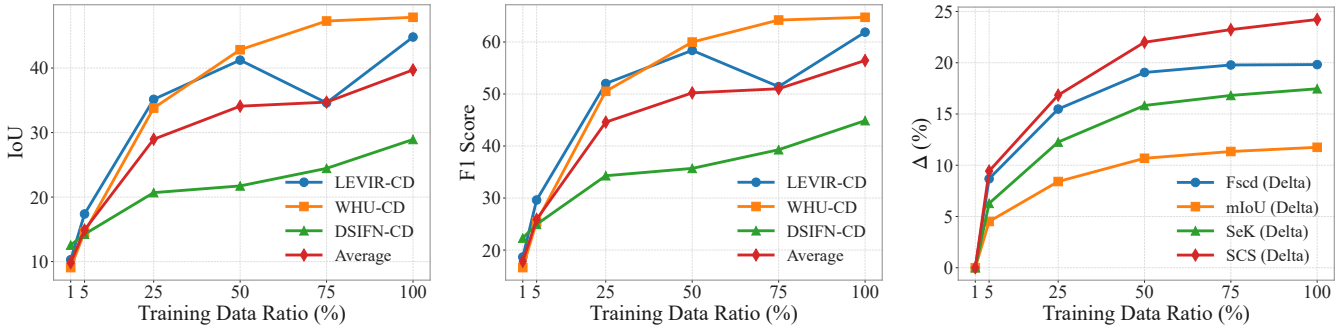


Figure 7 Scaling analysis of synthetic training data. Left and middle: Know-BCD scaling results on binary change detection benchmarks. Right: relative improvements on SECOND under varying Know-SEC training data ratios.

that Know-BCD provides increasingly effective training signals as more synthesized data becomes available. The SECOND scaling results show similar trends: increasing Know-SEC consistently improves semantic change detection metrics, indicating that the scaling benefit extends from binary to semantic change detection.

4 Conclusion

In this work, we introduce KnowChange, a knowledge-guided change data synthesis framework for remote sensing. Instead of using rule-based change simulation, KnowChange leverages pretrained vision-language models as knowledge sources to reason about plausible change regions and class transitions from scene contexts and desired changes. KnowChange combines knowledge-guided change simulation with generalizable layout-to-mask and mask-to-image models, enabling flexible synthesis of diverse change types. This process results in three synthetic datasets for building and semantic change detection. Extensive experiments demonstrate that KnowChange-generated data consistently improves synthetic-to-real transfer and synthetic data augmentation over existing synthetic datasets. Ablation studies highlight the importance of knowledge-guided reasoning and show that the proposed simulation can be seamlessly integrated into existing synthesis methods in a plug-and-play manner.

References

[1] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. ChangeN2: Multi-temporal remote sensing generative change foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(2):725–741, 2024.

[2] Yujie Zan, Shunping Ji, Songtao Chao, and Muying Luo. Open-vocabulary generative vision-language models for creating a large-scale remote sensing change detection dataset. ISPRS Journal of Photogrammetry and Remote Sensing, 225:275–290, 2025.

[3] Qiang Liu, Yang Kuang, Jun Yue, Pedram Ghamisi, Weiying Xie, and Leyuan Fang. Generating any changes in the noise domain. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(3):3698–3713, 2025.

[4] Yanis Benidir, Nicolas Gonthier, and Clément Mallet. The change you want to detect: Semantic change detection in earth observation with hybrid data generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2204–2214, 2025.

[5] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. arXiv:2506.15742, 2025.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, pages 8748–8763, 2021.

[8] Qi Zang, Jiayi Yang, Shuang Wang, Dong Zhao, Wenjun Yi, and Zhun Zhong. ChangeDiff: A multi-temporal change detection data generator with flexible text prompts via diffusion model. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 9763–9771, 2025.

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

[10] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. In Advances in Neural Information Processing Systems, volume 36, pages 16456–16482, 2023.

[11] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, I-3:293–298, 2012.

[12] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sensing of Environment, 237:111322, 2020.

[13] Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. SkySense-O: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14733–14744, 2025.

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv:2112.10741, 2021.

[15] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2149–2159, 2022.

[16] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. IEEE Transactions on Geoscience and Remote Sensing, 60:1–18, 2021.

[17] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. Computer Vision and Image Understanding, 187:102783, 2019.

[18] 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.

[19] 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.

[20] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 166:183–200, 2020.

[21] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21158–21167, 2022.

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

[23] Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3D: Revisiting change detection and captioning from a video modeling perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24011–24022, 2025.

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023.

[27] Jian Song, Hongruixuan Chen, and Naoto Yokoya. Syntheworld: A large-scale synthetic dataset for land cover mapping and building change detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 8287–8296, 2024.

[28] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. arXiv:2511.21631, 2025.

[29] ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. arXiv:2607.00248, 2026.

[30] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv:2507.01006, 2025.

[31] Manqi Zhao, Zifei Zhao, Shuai Gong, Yunfei Liu, Jian Yang, Xiong Xiong, and Shengyang Li. Spatially and semantically enhanced siamese network for semantic change detection in high-resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15:2563–2573, 2022.

[32] Yang Zhan, Kun Fu, Menglong Yan, Xian Sun, Hongqi Wang, and Xiaosong Qiu. Change detection based on deep siamese convolutional network for optical aerial images. IEEE Geoscience and Remote Sensing Letters, 14(10):1845–1849, 2017.

[33] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In Proceedings of the IEEE International Conference on Image Processing, pages 4063–4067, 2018.

[34] 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, 19:1–5, 2021.

[35] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2021.

[36] Shiqi Tian, Yanfei Zhong, Zhuo Zheng, Ailong Ma, Xicheng Tan, and Liangpei Zhang. Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application. ISPRS Journal of Photogrammetry and Remote Sensing, 193:164–186, 2022.

[37] Yinxia Cao and Xin Huang. A full-level fused cross-task transfer learning method for building change detection using noise-robust pretrained networks on crowdsourced labels. Remote Sensing of Environment, 284:113371, 2023.

[38] 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.

[39] Minseok Seo, Hakjin Lee, Yongjin Jeon, and Junghoon Seo. 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, pages 6374–6383, 2023.

[40] Hao Chen, Wenyuan Li, and Zhenwei Shi. Adversarial instance augmentation for building change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60:1–16, 2021.

[41] Zhuo Zheng, Shiqi Tian, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21818–21827, 2023.

[42] Ahmad Sebaq and Mohamed ElHelw. RSDiff: Remote sensing image generation from text using diffusion model. Neural Computing and Applications, 36(36):23103–23111, 2024.

[43] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. DiffusionSat: A generative foundation model for satellite imagery. arXiv:2312.03606, 2023.

[44] Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, and Deyu Meng. Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3614–3624, 2025.

[45] Datao Tang, Hao Wang, Yudeng Xin, Hui Qiao, Dongsheng Jiang, Yin Li, Zhiheng Yu, and Xiangyong Cao. TerraGen: A unified multi-task layout generation framework for remote sensing data augmentation. arXiv:2510.21391, 2025.

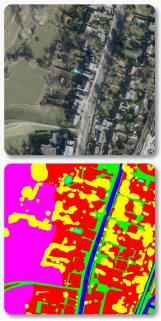
Contents

A Change Data Synthesis Details

A.1 Prompt Design for Global Layout Reasoning

The knowledge-guided change simulation module leverages Qwen3-VL [28] as a real-world knowledge source to infer plausible changes from pre-change scene contexts and user-specified change types. Given the pre-change image I_{pre}, semantic mask S_{pre}, and desired change categories $\mathcal{C} = \{c_i\}_{i=1}^{N_c}$, the VLM reasons about changed regions and their corresponding post-change categories, producing a change layout $\{(r_i, c_i)\}_{i=1}^{N_r}$. To support different change behaviors in remote sensing scenes, we design structured prompts for two actual-change transition modes, namely shape-preserving and shape-altering transitions, as well as pseudo-change simulation. All prompts use a unified JSON output format to facilitate reliable parsing and subsequent synthesis.

(1) For shape-preserving transitions, the shape of the original region is preserved, while its semantic category evolves from c_{pre} to c_i. The prompt takes as input I_{pre}, S_{pre}, the semantic category-color mapping, and desired change categories $\mathcal{C}$. The VLM predicts the post-change category c_i, together with the corresponding pre-change category c_{pre} and a region selection ratio $\alpha \in (0, 1)$, from which the changed region r_i is derived. Formally, let $\mathcal{K} = \{k_j\}_{j=1}^{N_k}$ denote the connected components of category c_{pre} in S_{pre}, sorted in descending order of area. The number of selected components is determined as $n = \lceil \alpha N_k \rceil$, and the changed region is obtained by taking the union of the first n components, i.e., $r_i = \bigcup_{j=1}^{n} k_j$. An example of the prompt and output for shape-preserving transitions is shown below.



Prompt: You are a remote sensing expert planning realistic ecological transitions. Context: You are provided with two images: 1. Satellite Image: The actual optical image of the terrain. 2. Semantic Segmentation Map: A color-coded map representing land cover types. Semantic Legend (Color Mapping): Red=rangeland, Green=developed space, Blue=road, Yellow=tree, Cyan=water, Magenta=agriculture land, Red=building Current Scene Context: - Potential Source Elements (based on detection): [rangeland, developed space, tree, water, agriculture land] - Available Target Textures for Generation: [bareland, rangeland, developed space, tree, water, agriculture land] Goal: Suggest 2 to 4 logical ecological transitions based on the instruction and the provided visual evidence. Rules: 1. PROTECT STRUCTURES: NEVER change buildings, roads, or man-made structures detected in the Semantic Map. 2. QUANTITY: Propose 2 to 4 distinct transitions. It is okay to suggest just 1 if that is the most logical. Do not exceed 4. 3. MODERATE IMPACT: Suggest a MODERATE RATIO (0.5 - 0.7). The change should be visible but preserve the underlying structure. 4. LOGICAL MAPPING: - Use the Semantic Map to confirm what the source material actually is (e.g., verify if a green area is Forest or Grass). - Good: Grass -> Soil (Drought), Forest -> Grass (Deforestation). - Avoid: Water -> Forest (Unlikely without specific instruction). Instruction: "Simulate a natural seasonal or ecological change." Output JSON Example: { "plan": [{"source": "Source_Cat_A", "target": "Target_Cat_B", "ratio": 0.6}, {"source": "Source_Cat_C", "target": "Target_Cat_D", "ratio": 0.5}] }

Answer: {"plan": [{"source": "rangeland", "target": "bareland", "ratio": 0.6}, {"source": "agriculture land", "target": "rangeland", "ratio": 0.5}, {"source": "tree", "target": "rangeland", "ratio": 0.6}]}

Figure 8 Example prompt and VLM output for shape-preserving transitions.

(2) For shape-altering transitions, newly instantiated regions are generated according to post-change categories, and their shapes may differ from existing pre-change regions. The prompt takes as input I_{pre}, S_{pre}, the semantic category-color mapping, randomly generated candidate rectangular regions $\mathcal{B} = \{b_j\}_{j=1}^{N_b}$, and desired change categories $\mathcal{C}$. The VLM selects appropriate candidate boxes and assigns post-change

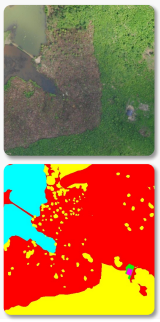
categories, producing (r_i, c_i). These coarse regions are subsequently refined by the layout-to-mask model into pixel-level object shapes that are compatible with the surrounding scene context. An example of the prompt and output for shape-altering transitions is shown below.

Prompt: You are an expert in remote sensing image analysis. Context: You will receive a satellite image and a corresponding semantic map, both containing marked bounding boxes with specific Region IDs and coordinates. You are also provided with the semantic color-to-category mapping.Your Task: 1. Visual Analysis & Filtering: Comprehensively analyze the visual features of each region using both images. Determine if the region is suitable for modification or object generation. Filter and retain only the suitable regions (identify them by their Region ID). 2. Categorization: For each suitable region, infer the most logical category for the change/object placement. - Constraint: The assigned category MUST be selected strictly from the provided Candidate Categories list. - Constraint: For boxes in the background category, priority can be given to foreground categories (such as buildings, trees, water bodies, relative to farmland, grassland, and similar foreground background relationships) - Constraint: The number of categories should not exceed two, and the number of boxes can exclude areas that you know are not suitable for change. 3. Please note that I do not want too many changes to appear on a single image, so you can only select two categories at most, allowing all boxes that are suitable for changes to be retained, but still retain boxes that are relatively more suitable for changes Output Format: Return the result in the following JSON format. - "Regions": A list of the Region IDs you selected (e.g., ["Region1", "Region3"]). - "Categories": A list of the corresponding categories (e.g., ["Category_A", "Category_B"]). - Important: Ensure the order in "Categories" matches the order in "Regions". Example: { "Regions": ["Region1", "Region3"], "Categories": ["Category_1", "Category_2"] } { "Regions": ["Region1","Region2","Region3"], "Categories": ["Category_1", "Category_2"] } Input Data: - Candidate Categories: [bareland, rangeland, developed space, road, tree, water, agriculture land, building] - Semantic Legend (Color Mapping): Red=unknown, Green=rangeland, Blue=developed space, Yellow=road, Cyan=tree, Magenta=building - Region IDs and Coordinates: - Region 1: [179, 252, 407, 463] - Region 2: [87, 183, 277, 340] - Region 3: [85, 126, 318, 306] Visual Inputs: 1. Image 1: Satellite image with 3 regions marked with RED BOXES. 2. Image 2: Semantic layout map with the same BOXES.

Answer: {"Regions": ["Region1", "Region2"], "Categories": ["developed space", "tree"]}

Figure 9 Example prompt and VLM output for shape-altering change simulation.

(3) For pseudo-change simulation, the semantic category remains unchanged while the visual appearance varies due to imaging-condition differences such as illumination, season, atmosphere, or sensor variation. The prompt takes as input I_{pre}, S_{pre}, the semantic category-color mapping, and the categories present in the image. The VLM identifies pre-change categories c_{pre} that may exhibit pseudo changes and estimates their region perturbation ratios $\beta \in (0, 1)$. For each predicted category c_{pre}, let $\mathcal{K} = \{k_j\}_{j=1}^{N_k}$ denote its connected components in S_{pre}, sorted in descending order of area. The number of selected pseudo-change components is determined as $n = \lceil \beta N_k \rceil$, and the pseudo-change region is obtained by taking the union of the first n components, i.e., $r_{\text{pse}} = \bigcup_{j=1}^{n} k_j$. The obtained pairs $(r_{\text{pse}}, c_{\text{pre}})$ are used to guide the M2I model to perform category-preserving appearance reconstruction during post-change image synthesis. An example of the prompt and output for pseudo-change simulation is shown below.



Prompt: You are an expert in Remote Sensing Data Augmentation. Context: You are provided with two images: 1. Satellite Image: The actual optical image of the terrain. 2. Semantic Segmentation Map: A color-coded map representing land cover types. Semantic Legend (Color Mapping): Red=rangeland, Green=developed space, Blue=road, Yellow=tree, Cyan=water, Magenta=building Available Natural Elements: Based on the data, the following natural categories are available for modification: [rangeland, developed space, tree, water] Task: Select 2 to 4 categories from the list above to apply inpainting/augmentation. Rules: 1. Strictly Natural Only: Do NOT select buildings, roads, or man-made structures. Only pick from the provided natural list. 2. Verify with Semantic Map: Use the Semantic Map to confirm the location and extent of these natural categories (e.g., ensure you are selecting actual Forest or Water areas). 3. High Ratio: Since these are natural textures (forest, grass, water, soil), use a High Ratio (0.7 - 0.95) to create diverse variations. Output JSON Format: { "mask_plan": [{"category": "Category_A", "ratio": 0.85}, {"category": "Category_B", "ratio": 0.90}] }

Answer: {"category": "rangeland", "ratio": 0.9}, {"category": "tree", "ratio": 0.85}

Figure 10 Example prompt and output for pseudo-change simulation.

A.2 More Examples

We present additional examples from our three generated datasets: Know-BCD (building change detection), Know-SEC (semantic change detection based on SECOND categories), and Know-HR (semantic change

detection based on HRSCD categories). All datasets are synthesized from two source benchmarks: OpenEarthMap [9] and FLAIR [10]. The examples organized by source dataset are shown in Fig. 11.

B Additional Experimental Details

B.1 Data Source

Table 7 Detailed statistics of the datasets used for binary change detection (BCD) and semantic change detection (SCD). Train/Val/Test denotes the number of samples after preprocessing.

Dataset	Resolution (m)	# Samples	Image Size	Train/Val/Test	Task
LEVIR-CD	0.5	637	1024×1024	2,548/–/1,392	BCD
WHU-CD	0.075	1	32507×15354	1,260/–/690	BCD
DSIFN-CD	0.03–1	3,940	512×512	3,600/340/48	BCD
SECOND	0.5–3	4,662	512×512	2,968/–/1,694	BCD & SCD
HRSCD	0.5	291	10000×10000	905/116/123	SCD

Evaluation Datasets for Change Detection: To comprehensively assess the utility of synthetic data, we evaluate on five widely adopted change detection benchmarks, covering both Binary Change Detection (BCD) and Semantic Change Detection (SCD) tasks. Detailed statistics of the original datasets are provided in Table 7. These datasets collectively span diverse geographic regions, seasonal variations, and change patterns, enabling rigorous evaluation of model generalization. To ensure consistent evaluation across heterogeneous benchmarks, we standardize all input images to 512×512 patches. For datasets with original resolutions larger than 512×512 (e.g., WHU-CD[19], HRSCD[17]), we partition them into non-overlapping 512×512 patches (e.g., intervals [0, 512], [512, 1024]), allowing overlap only for the boundary patches to ensure full coverage.

Specifically, the dataset splits are organized as follows: (1) LEVIR-CD [18] focuses on building changes in urban/suburban areas at 0.5m resolution. After cropping to 512×512, the dataset contains 2,548 training and 1,392 testing image pairs. (2) WHU-CD [19] provides a single large-scale aerial image (32,507×15,354) with extensive building construction/demolition annotations at 0.075m resolution. Following the official protocol, we partition it into 512×512 patches, yielding 1,260 training and 690 testing samples. (3) DSIFN-CD [20] covers multi-scale changes across diverse scenes at ~0.03-1m resolution, emphasizing complex rural-urban transitions. The official split provides 3,600 training, 340 validation, and 48 testing image pairs, all at 512×512 resolution. (4) SECOND [16] supports both BCD and SCD tasks with fine-grained semantic annotations for 7 land-cover categories. After standardizing to 512×512, we adopt the official split of 2,968 training and 1,694 testing pairs. (5) HRSCD [17] offers very-high-resolution (0.5 m) imagery with dense pixel-wise change labels. The original 291 large images ($10{,}000 \times 10{,}000$ pixels) are first cropped into 116,400 patches of 512×512 pixels. HRSCD exhibits substantial class imbalance, with most labeled regions remaining unchanged across time; for example, changes involving artificial surfaces and agricultural land account for only 0.6% of all change instances [31]. Following [23], we therefore retain only image pairs in which changed areas occupy at least 20% of the patch, resulting in a curated split of 905 training, 116 validation, and 123 testing samples. For zero-shot evaluation, models are trained exclusively on each synthetic dataset and directly evaluated on these real-world benchmarks without fine-tuning, ensuring a strict assessment of cross-domain generalization.

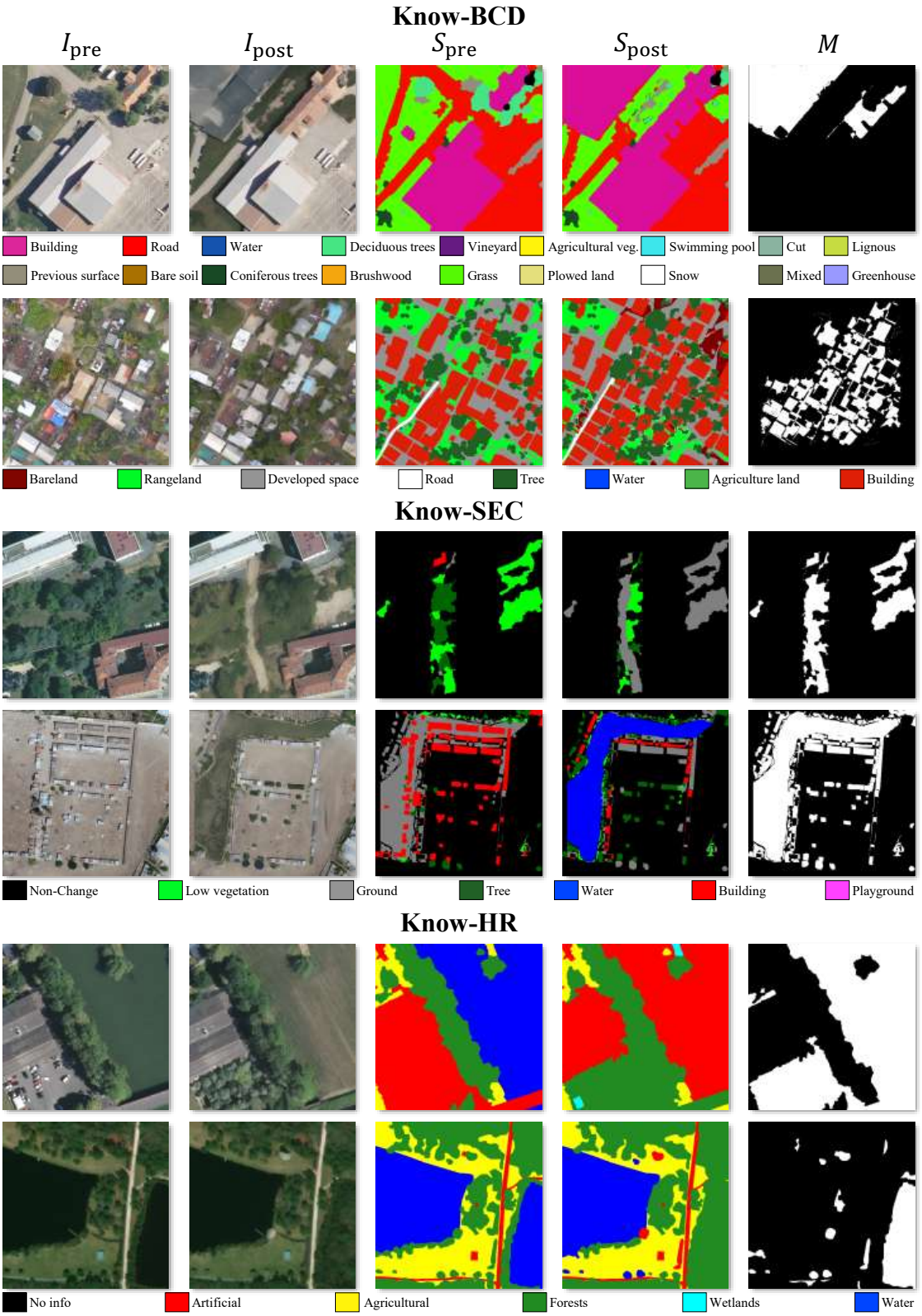


Figure 11 Visualization of generated samples from our three synthetic datasets: Know-BCD, Know-SEC, and Know-HR. In each dataset, the first row uses source images from FLAIR and the second row uses source images from OpenEarthMap.

B.2 Synthetic Datasets for Comparison

To evaluate the downstream utility of KnowChange-generated data, we compare our synthetic datasets with representative change data synthesis datasets used for binary and semantic change detection. These datasets differ in their source data, synthesis strategy, scale, and supported task types. Table 8 summarizes the synthetic datasets involved in our experiments, including existing datasets and the three datasets generated by KnowChange, namely Know-BCD, Know-SEC, and Know-HR.

Table 8 Overview of synthetic change detection datasets. BCD and SCD denote binary and semantic change detection.

Synthetic Dataset	Base Dataset	#Samples	Task
WHU-GCD	LoveDA, Evlab-SS, LandCover.ai, Google Earth	25K	SCD
SyntheWorld	Fully Synthetic, OpenEarthMap	40K	BCD
FSC-180k	FLAIR	180K	SCD
Changen2-S1/S9	xView2/OpenEarthMap	15K/27K	BCD/SCD
Know-BCD/SEC/HR	OpenEarthMap, FLAIR	10K/10K/10K	BCD/SCD/SCD

B.3 Layout-to-Mask Model Training

The primary objective of training the Layout-to-Mask (L2M) model is to enable the acquisition of rich category-shape associations and ensure consistency between object shapes and their surrounding context. The model takes as input the original semantic map S_{pre}, a semantic mask M_{sa} indicating regions to be inpainted, and textual prompts describing the target categories. To enhance the model's generalization capability across diverse scenarios, we design two complementary training strategies:

Category-Aware Instance Masking: We select specific semantic categories (either single or multiple) and identify their connected components. This strategy is implemented through two masking schemes:

- *Bounding Box Masking:* Connected components are masked using their minimum bounding boxes. This forces the model to infer plausible shape distributions of specified categories within rectangular constraints.
- *Connected Component Masking:* Connected components are masked using their exact shapes rather than bounding boxes. This focuses on fine-grained shape fidelity and precise boundary reconstruction.

For both schemes, the text prompt corresponds directly to the masked category names (e.g., "Add building and road in the masked area"), enabling the model to learn specific category-shape priors with explicit semantic guidance.

Random Region Masking: We apply random rectangular masks that may span multiple semantic categories without prior category selection. In this case, the text prompt is defined as the dominant categories (top two by pixel count) within the masked region. This strategy encourages the model to maintain smooth spatial transitions and handle complex contextual relationships across different semantic boundaries without explicit category guidance.

For all strategies, we employ a unified text prompt template that specifies the color legend for semantic categories and instructs the model to generate content within the masked region. The key difference lies

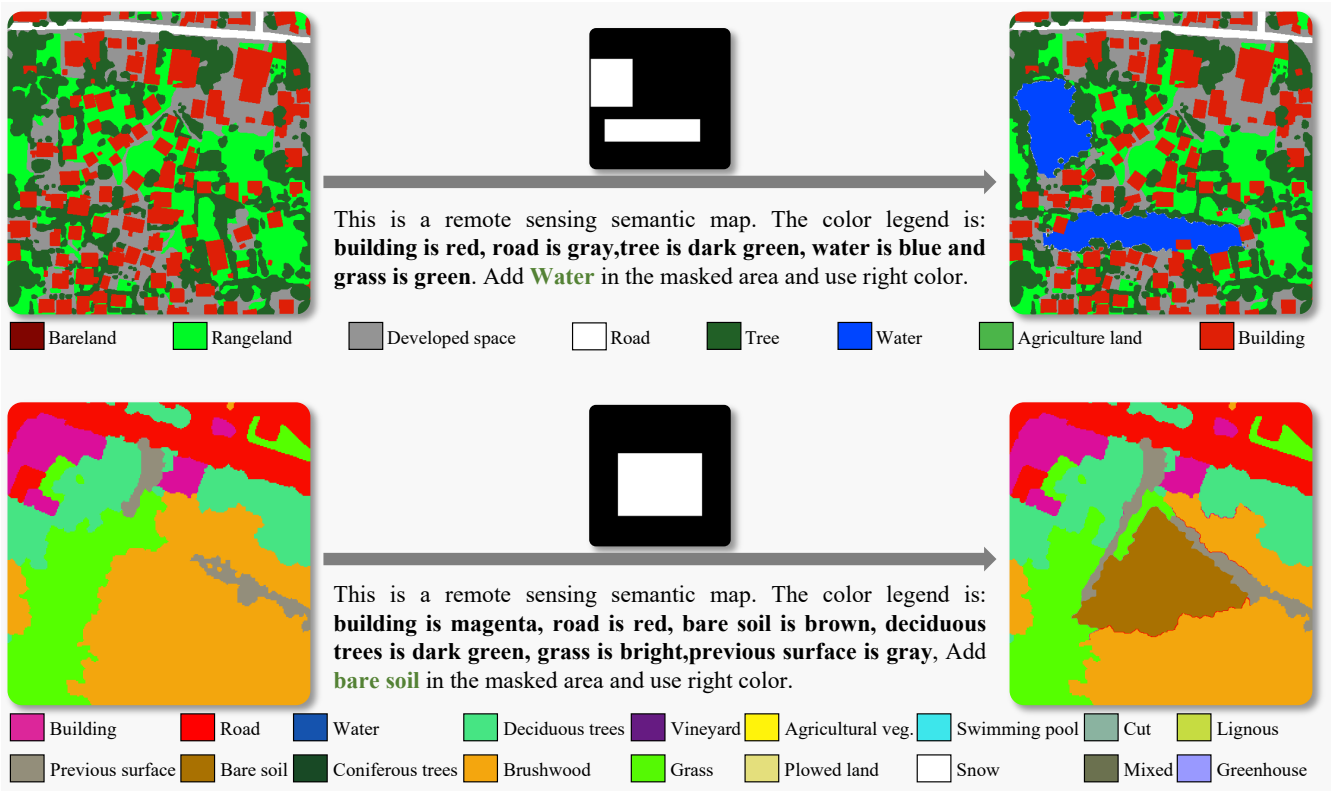


Figure 12 Examples of shape-altering transitions generated by the Layout-to-Mask model. The model fills masked regions with specified categories while maintaining shape plausibility and contextual consistency.

in how category names are specified in the prompt: For category-aware instance masking, the prompt includes the specific masked category names, providing explicit semantic guidance for shape generation. For random region masking, the prompt includes the top two dominant categories within the masked region, requiring the model to infer appropriate category placement based on context.

For example, a typical prompt follows the format:

> "This is a remote sensing semantic map. The color legend is: building is red, road is gray, water is blue and grass is green. Add building and road in the masked area and use right color."

To prevent the model from memorizing fixed color-category bindings, we randomly assign colors to semantic categories across training samples. This design encourages the model to learn semantic category-shape associations through textual prompts rather than relying on color cues alone. Conditioned on the prompt, mask, and visible context, the model completes the masked regions. Having been trained on large-scale datasets, the model learns category-specific contour information and spatial distribution patterns. This enables the model to controllably fill specified categories within the target region M_{sa} with plausible shapes, while maintaining consistency with surrounding objects. Specific examples illustrating this process are shown in Fig. 12.

B.4 Mask-to-Image Model Training

Motivated by masked-image reconstruction strategies widely used in image editing [14, 15], we train the Mask-to-Image (M2I) model to reconstruct remote sensing images from masked inputs under semantic

Table 9 Label mapping from the source datasets (OpenEarthMap and FLAIR) to the target categories of SECOND.

OpenEarthMap Categories	FLAIR Categories	Target (SECOND)	Target ID
Building	Building	Building	5
Agriculture land, Rangeland	Agri. veg., Plowed land, Vineyard, Grass, Brushwood	Low vegetation	1
Road, Developed space, Bareland	Road, Pervious, Bare soil, Cut, Mixed	Non-vegetated ground	2
Tree	Coniferous, Deciduous trees, Lignous	Tree	3
Water	Water, Swimming pool	Water	4
Background	Greenhouse, Snow	Non-change	0

guidance. Similar reconstruction-based training schemes have also been adopted in remote sensing change synthesis, such as WHU-GCD [2] and HySCDG [4]. Specifically, during training, the model takes a masked image $\widetilde{I}_{\text{pre}}$, a re-rendering mask M, and the corresponding semantic mask S_{pre} as inputs, and reconstructs the original image I_{pre}. This objective enables the model to synthesize category-consistent appearances within masked regions while preserving the visible surrounding context. During inference, S_{pre} is replaced with the synthesized post-change semantic mask S_{post} to generate I_{post}. Following these established practices, we employ instance-, regional-, and global-level masking, together with semantic vector dropout, to cover diverse spatial contexts and semantic conditions.

Instance-Level Masking: We identify connected components in the semantic mask and randomly select multiple components during each training iteration. The corresponding image regions are masked using their exact component boundaries rather than bounding boxes, while their original semantic categories are retained as conditioning signals. The model reconstructs the masked image regions based on the surrounding image context and the pixel-wise semantic embeddings derived from the semantic mask. By exposing the model to objects with diverse shapes, sizes, and textures, this strategy facilitates category-specific appearance learning. Extremely small components are filtered out to reduce the influence of noisy artifacts.

Regional-Level Masking: We apply random rectangular masks that may cover multiple semantic categories. The size and aspect ratio of each mask are randomly sampled to provide diverse regional contexts. The model reconstructs the masked image region based on the visible surrounding context and the semantic embeddings within the masked region. This strategy encourages the model to generate coherent appearances across semantic boundaries, such as building–road and vegetation–water interfaces.

Global-Level Masking: The entire image is occasionally masked, leaving no visible image context. The model is therefore required to reconstruct the scene using only the semantic condition. This strategy strengthens holistic scene modeling and encourages the model to preserve spatial relationships among semantic categories, such as roads connected to buildings and water bodies adjacent to vegetation.

Semantic Vector Dropout: In addition to spatial masking, we apply semantic vector dropout to improve robustness to incomplete or noisy semantic conditions. During training, the semantic embeddings of randomly selected regions are replaced with zero vectors, requiring the model to infer plausible visual content from the visible image context and the semantic information of surrounding regions. This strategy reduces the model's sensitivity to missing or inaccurate semantic annotations.

For all training strategies, the semantic mask is converted into pixel-wise CLIP semantic embeddings, which are further transformed by the adapter described in the main paper into control features for the M2I model. The model is optimized to reconstruct the original image from the masked input under semantic conditioning. By combining instance-, regional-, and global-level masking with semantic vector dropout, M2I learns category-consistent appearances while preserving local boundary coherence and

global contextual consistency.

B.5 Training Details for Change Detection Models

Binary Change Detection Training: For BCD evaluation on datasets including LEVIR-CD, WHU-CD, DSIFN-CD, and SEC-BCD, we adopt ChangeFormer [22] as the primary evaluation model. Since several synthetic datasets are derived from multi-category semantic segmentation sources, we extract binary change masks by considering only building-related transitions. Specifically, for our Know-BCD and Know-SEC, as well as other semantic-change-based synthetic datasets such as WHU-GCD [2], FSC-180k [4], and Changen2-S9 [1], a pixel is labeled as "changed" if its semantic category transitions involve building emergence, demolition, or expansion; all other semantic transitions are treated as unchanged. In contrast, datasets that are inherently building-focused, such as SyntheWorld [27] and Changen2-S1 [1], already provide binary change masks and require no additional processing. This unified protocol ensures consistency across all training data while leveraging the rich semantic information from source datasets.

Table 10 Label mapping from the source datasets (OpenEarthMap and FLAIR) to the target categories of HRSCD.

OpenEarthMap Categories	FLAIR Categories	Target (HRSCD)	Target ID
Building, Road, Developed space	Building, Road, Pervious surface	Artificial surface	1
Agriculture land	Agri. veg., Plowed land, Vineyard		
Rangeland, Bareland	Grass, Brushwood, Bare soil, Cut, Mixed	Agricultural land	2
Tree	Coniferous, Deciduous trees, Lignous	Forest	3
–	Snow	Wetland	4
Water	Water, Swimming pool	Water	5
Background	Greenhouse, Unknown	No info	0

Semantic Change Detection Training: For SCD evaluation on SECOND and HRSCD datasets, we use Change3D [23] for pixel-wise land-cover transition prediction. Since different datasets follow different class definitions, we apply label mapping rules to align the synthetic data with target benchmarks. The mapping tables from OpenEarthMap and FLAIR to SECOND and HRSCD categories are provided in Table 9 and Table 10, respectively. Specifically for the SECOND dataset, in addition to label mapping, we compare pre- and post-change semantic masks to identify unchanged regions, which are explicitly set to the non-change category to align with the benchmark protocol.

C Additional Experimental Results and Analysis

C.1 Analysis of Change Proportions and Category Transition Matrices

To quantitatively evaluate the semantic diversity and realism of the synthesized change data, we conduct a comprehensive analysis involving category transition matrices, overall change proportions, and the Jensen-Shannon (JS) divergence. Since previous studies have designed specific rules for category transitions[1, 2], this analysis aims to compare the differences in category changes between various synthetic datasets and the real dataset. All datasets are mapped to the unified SECOND taxonomy for a fair comparison. The visualization results comparing the real SECOND dataset with four synthetic datasets (Changen2-S9, WHU-GCD, FSC-180k, and our Know-SEC) are shown in Fig. 13.

The SECOND dataset serves as the real-world baseline with a change proportion of 19.94%. Among the synthetic datasets, Changen2-S9 exhibits the highest proportion at 25.52%, whereas FSC-180k records

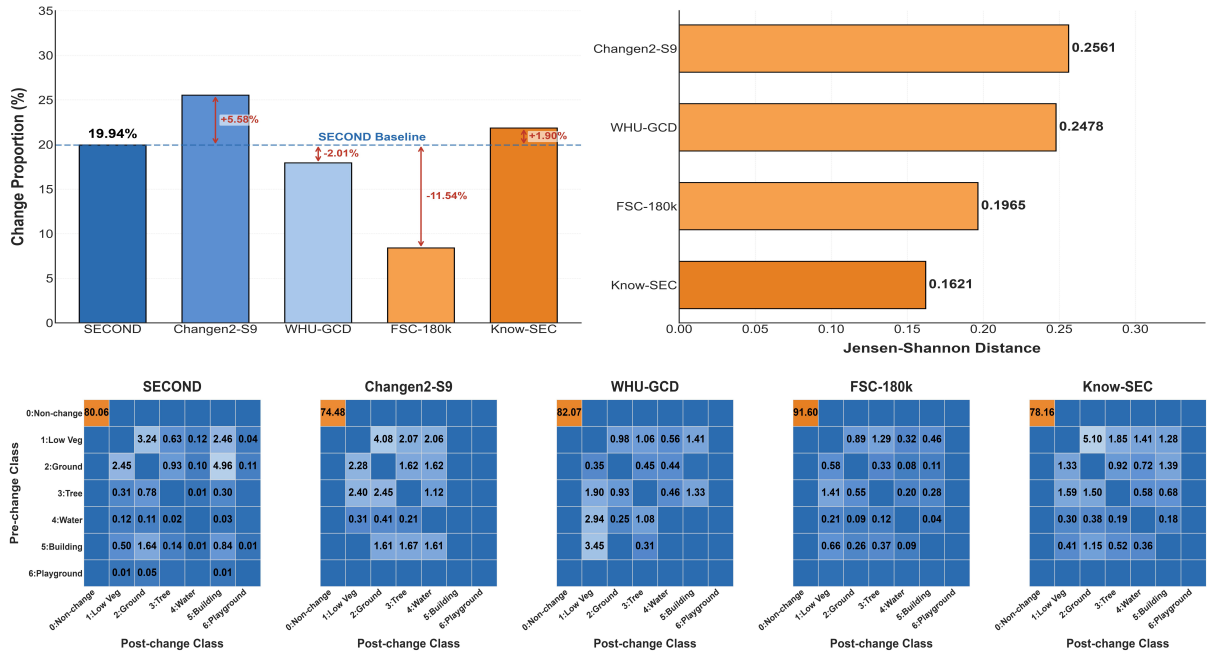


Figure 13 Overview of the statistical comparison between the real SECOND dataset and synthetic datasets regarding change proportions and category transition differences.

the lowest at approximately 8.4%. Both WHU-GCD and our Know-SEC align most closely with the baseline, showing marginal differences of 2.01% and 1.90%, respectively. This proximity suggests that our synthesized data closely approximates the change density found in real-world scenarios. Additionally, regarding distributional similarity, our Know-SEC achieves the lowest JS distance of 0.1621 compared to the SECOND dataset.

The class transition matrices further reveal clear differences in transition coverage among the synthetic datasets. SECOND contains diverse inter-class and intra-class transitions, reflecting the complexity of real-world changes. Although Changen2-S9 and WHU-GCD achieve change proportions of 25.52% and 17.93%, respectively, their transition matrices remain relatively sparse, with several class transitions observed in SECOND being absent. This is mainly because their change simulation relies on predefined transition matrices or handcrafted rules under fixed category settings. Consequently, transitions not explicitly covered by these designs cannot be adequately synthesized, reflecting the limited transition coverage and inflexibility discussed in the main paper. FSC-180k covers a broader range of inter-class transitions, but its change proportion is only 8.40%, substantially lower than the 19.94% observed in SECOND. Its transition frequencies are also highly imbalanced, indicating that expanding the predefined transition set alone does not ensure a realistic transition distribution.

In contrast, Know-SEC does not restrict change simulation to a predefined transition matrix. Instead, the VLM infers plausible class transitions from the scene context and textual prompt, resulting in broader transition coverage and a change proportion of 21.84%, which differs from SECOND by only 1.90 percentage points. Know-SEC consequently achieves the lowest JS distance of 0.1621 among the compared synthetic datasets, indicating the closest overall transition distribution to the real dataset. Nevertheless, some distributional imbalance remains. For example, the Low Vegetation-to-Ground transition accounts for 5.10% in Know-SEC, compared with 3.24% in SECOND, and its intra-class transitions remain less diverse. This discrepancy may arise because KnowChange focuses on the contextual plausibility of individual changes rather than explicitly matching empirical transition frequencies, while the class distribution of

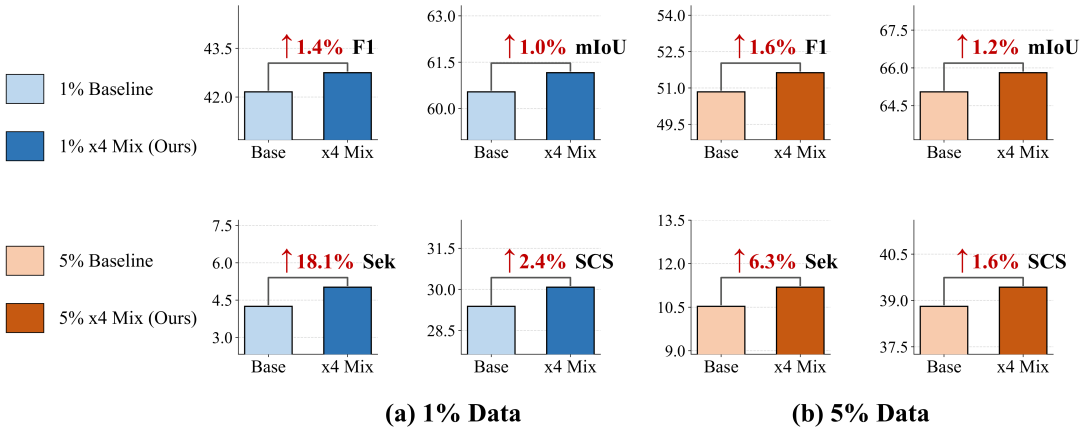


Figure 14 Impact of 4× in-domain augmentation under different labeled data settings.

the source semantic data also affects the generated distribution.

C.2 Effectiveness of In-Domain Data Augmentation:

Different from methods such as Changen2 and HySCDG, KnowChange leverages VLMs as its knowledge source and is trained over a broad semantic vocabulary, enabling it to perform in-domain augmentation on previously unseen datasets. To evaluate this capability, we randomly sample 1% and 5% of the labeled training samples from SECOND, a dataset not seen during the training of KnowChange. For each setting, KnowChange uses the sampled images and their labels to synthesize target-domain change samples at four times the size of the corresponding subset. The generated samples are then combined with the sampled real data to train the downstream model. As shown in Fig. 14, incorporating the synthesized data improves all four metrics under both labeled-data settings. SeK exhibits the most pronounced gains, increasing by 18.1% and 6.3% under the 1% and 5% settings, respectively. These results suggest that few-shot in-domain augmentation with KnowChange can partially alleviate category imbalance and limited supervision on unseen datasets, resulting in modest improvements in semantic change detection performance.

C.3 Generalization to Additional Remote Sensing Tasks

The main experiments focus on building and semantic change detection, leaving the applicability of the generalizable semantic-guided synthesis component beyond change data construction less explored. To complement these evaluations, we further examine whether KnowChange can support other remote sensing tasks through semantic-map-conditioned image synthesis. Without requiring real images as references, the framework can generate remote sensing images from semantic maps while preserving pixel-level correspondence between the generated images and input annotations. This property enables the direct construction of image–annotation pairs for semantic segmentation and category-specific land-cover extraction, such as buildings, roads, and water bodies. As illustrated in Fig. 15, KnowChange produces visually plausible images that remain consistent with the provided semantic layouts across these tasks. These qualitative results complement the main experiments by demonstrating that the proposed semantic-guided synthesis framework is not restricted to change detection and can potentially support data construction for a broader range of remote sensing tasks.

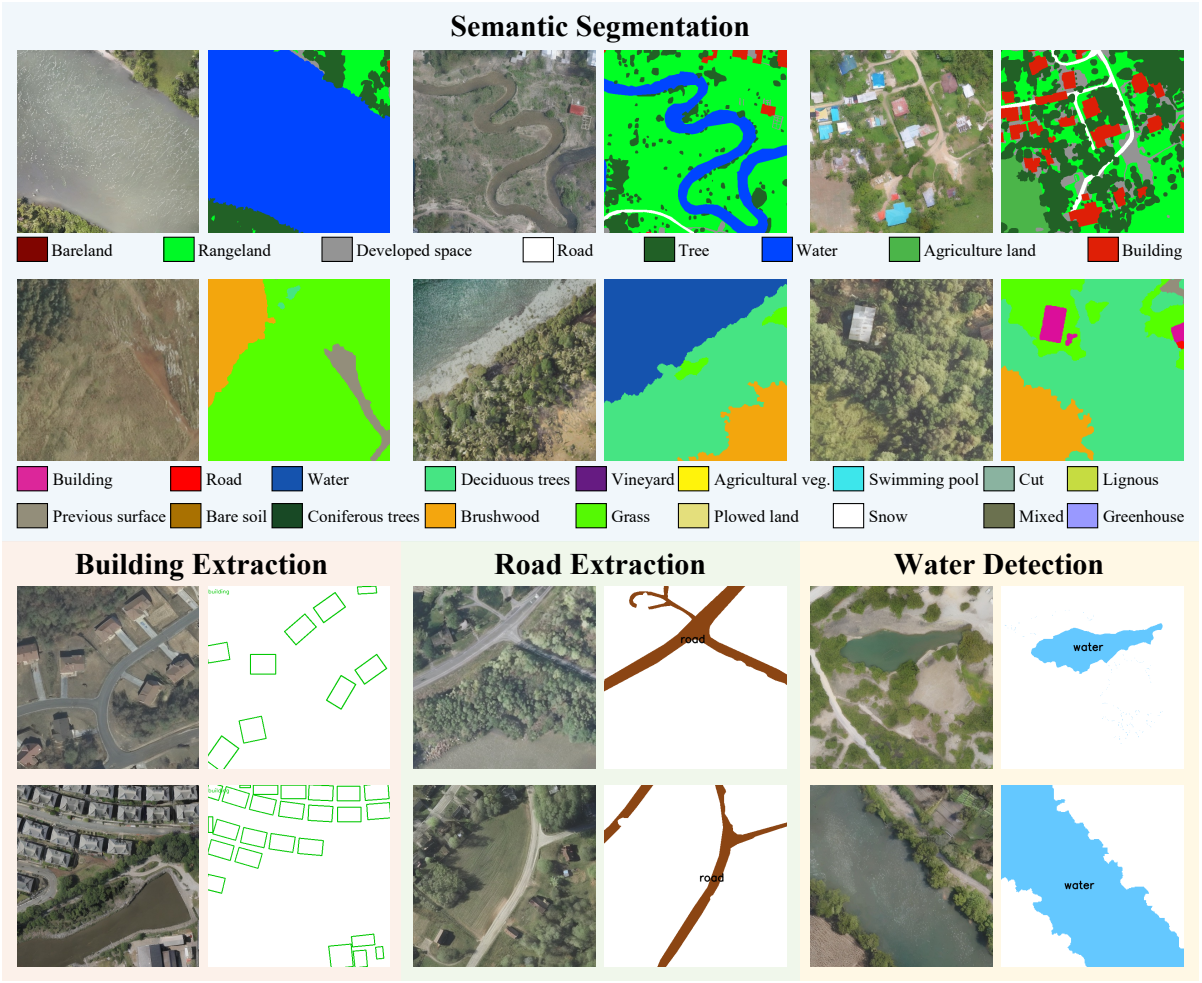


Figure 15 Visualization of multi-task capabilities using our framework. These samples are generated directly from semantic maps without pre-temporal image inputs (I_{pre}). The visualization demonstrates applicability to downstream tasks including semantic segmentation, building extraction, road extraction, and water detection.

C.4 Image Reconstruction Quality Analysis

The M2I model is responsible for rendering post-change images according to semantic masks while preserving the content of unchanged regions. To evaluate its image synthesis capability, we use M2I, HySCDG, and Changen2 to reconstruct images from the SECOND dataset under the same evaluation setting. The reconstructed images are compared with their corresponding real images using PSNR, SSIM, and LPIPS.

Table 11 Image reconstruction quality comparison on the SECOND dataset.

Model	PSNR ↑	SSIM ↑	LPIPS ↓
M2I	**22.3340**	**0.7042**	**0.1877**
HySCDG	21.3064	0.6900	0.1960
Changen2	22.1237	0.6985	0.1968

As shown in Tab. 11, M2I consistently outperforms HySCDG and Changen2 across all three metrics. Compared with the strongest competing result for each metric, M2I improves PSNR and SSIM by 0.2103

Figure 16 Long-term urban evolution synthesized from a single-temporal scene. KnowChange iteratively generates semantic maps (top) and their corresponding remote sensing images (bottom), producing multi-category changes while maintaining semantic and spatial consistency across time.

and 0.0057, respectively, while reducing LPIPS by 0.0083. The higher PSNR and SSIM indicate better preservation of image content and spatial structure, whereas the lower LPIPS reflects improved perceptual similarity to the real images. These results demonstrate that M2I can more accurately reconstruct remote sensing images from semantic conditions, supporting its ability to generate structurally consistent and visually realistic change data.

C.5 Long-Term Urban Evolution Synthesis

Beyond bi-temporal change data synthesis, KnowChange can be iteratively applied to transform a single-temporal scene into a long-term multi-temporal sequence. Specifically, the image and semantic map generated at each timestamp are used as the input for the subsequent timestamp, allowing the scene to evolve progressively over time. As shown in Fig. 16, the synthesized sequence presents complex urban evolution involving multiple land-cover categories, including buildings, vegetation, trees, roads, and water bodies. Across successive timestamps, the spatial distribution of these categories changes gradually, while the overall road structure and unchanged regions remain coherent. The generated images also remain well aligned with their corresponding semantic maps throughout the sequence. These results demonstrate the capability of KnowChange to model long-horizon, multi-category urban evolution while preserving temporal, spatial, and semantic consistency, extending its applicability from bi-temporal change synthesis to multi-temporal change simulation.

D Related Work

Binary & Semantic Change Detection: Change detection in remote sensing imagery identifies differences across temporal observations, evolving from traditional algebraic methods to deep learning-based approaches. Binary change detection (BCD) methods initially adopted Siamese CNNs [17, 32, 33] and were subsequently enhanced with attention mechanisms [18, 34] and Transformers [22, 35]. These methods focus on locating changed regions but do not characterize the associated semantic transitions. Semantic change detection (SCD) extends BCD by jointly identifying change locations and land-cover categories [16, 17]. However, SCD remains challenged by severe class imbalance, scarce annotations for rare categories, and the complex transition space between land-cover classes [2, 36]. Transfer learning partially alleviates annotation scarcity by exploiting models pretrained on large-scale source domains [36, 37], but both BCD and SCD still rely heavily on pixel-level annotations. In particular, rare and diverse change transitions are

difficult to cover with real training samples, motivating change data generation as a scalable means of expanding training distributions [1, 2, 4, 8].

Data Generation for Change Detection: Data generation alleviates the scarcity of annotated datasets. Fully synthetic methods use 3D rendering engines such as SyntheWorld [27] to create scenes from scratch, offering explicit parameter control but often lacking real-world texture diversity. Hybrid approaches instead insert synthetic changes into real images. Early methods employed copy-paste operations, such as ChangeStar [38] and Self-Pair [39], which may introduce visible artifacts and inconsistent object-context relationships. IAug [40] utilizes GANs to simulate building changes but has limited applicability to diverse land-cover transitions. Recent approaches leverage diffusion models to improve synthesis quality. Changen [41] and Changen2 [1] synthesize bi-temporal images conditioned on semantic layouts, while FSC-180k [4] combines Stable Diffusion with ControlNet [26] for semantic-guided inpainting. SD-Inpainting-RS [2] further employs predefined semantic rules and geometric constraints to guide change generation. Text-guided methods such as RSDiff [42] and DiffusionSat [43], as well as single-temporal generators such as AeroGen [44] and TerraGen [45], can synthesize high-quality remote sensing images but are not designed to construct spatially aligned bi-temporal change pairs. Despite these advances, existing change synthesis methods commonly rely on handcrafted transition rules with limited coverage and require repeated customization for different change types. KnowChange instead leverages pretrained VLMs as knowledge sources to infer plausible change locations and class transitions from scene contexts and desired change types. Combined with generalizable semantic-guided synthesis models, this design enables flexible generation of diverse change data without repeatedly redesigning the synthesis pipeline.