1. Datasets

SECOND We utilized the SEmantic Change detectiON Dataset (SECOND) [52], a benchmark dataset for SCD. The SECOND dataset is created using high-resolution bi-temporal optical images captured by various aerial platforms and sensors. The dataset covers multiple cities in China, such as Hangzhou, Chengdu, and Shanghai. Each image measures 512 × 512 pixels, with spatial resolutions ranging from 0.5 to 3 meters per pixel. The dataset consists of a total of 4,662 RGB images, each paired with corresponding pre- and post-change maps. It has 6 main land-cover classes, i.e., non-vegetated ground surface, tree, low vegetation, water, buildings and playgrounds. To ensure compatibility with the CNAM-CD [58] dataset, we use only the post-change maps, which are focusing on growth changes unlike pre-change maps which focuses on extinction of areas. The dataset is divided into three parts: 2,968 images for training, 300 for validation, and 1,394 for testing.

CNAM-CD We utilized recently introduced CNAM-CD [58]. The dataset consists of images with 512x512 pixels capturing the changes occurred in the areas of State-level New Area (SLNA) from China between 2013 and 2022. Spatial resolution is 0.5 meters per pixel and change maps only shows growth changes. Dataset includes 5 main land-cover classes which are classified as bare land, vegetation, water, impervious surfaces, and others (clouds, hard shadows). Dataset is divided into three parts: 1750 images for training, 250 for validation and 278 for testing.

LEVIR-CD The LEVIR-CD dataset is sourced from Google Earth, comprises remote sensing images only focused on tracking the appearance and disappearance of buildings, hence includes a single semantic class. The original image resolution is 1024x1024 pixels. Following the approach in [2], we divided each image into 256x256 pixel sections, and split the dataset to 7,120 image pairs for training, 1,024 pairs for validation, and 2,048 pairs for testing.

WHU-CD The WHU-CD dataset [30] includes paired aerial images taken in 2012 and 2016, capturing an area impacted by an earthquake, where numerous changes are visible, it focuses on reconstructed and newly developed buildings, hence has a single change class. The training set contains 5,947 image pairs, with 743 pairs in the validation set and 744 pairs in the test set.

2. Comparison Metrics

Let $Q = q(i, j)$ denote confusion matrix where $q(i, j)$ represents the number of pixels that are classified as class i while ground truth class is j. Overall pixel accuracy is defined as:

$$OA = \sum_{i=0}^{N} q_{ii} / \sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij}. \tag{4}$$

Since non-change regions are dominating the pixel accuracy it is not best choice for the SCD task. Therefore, mIoU and SeK metrics are proposed along with the SECOND dataset as strong alternatives to mitigate class imbalance [52]. mIoU is the average of non-changed ($\mathrm{IoU_{nc}}$) and changed ($\mathrm{IoU_c}$) intersection of union values.

$$\mathrm{mIoU} = \left(\mathrm{IoU_{nc}} + \mathrm{IoU}_c\right)/2 \tag{5}$$

$$\mathrm{IoU_{nc}} = q_{00} / \left(\sum_{i=0}^{N} q_{i0} + \sum_{j=0}^{N} q_{0j} - q_{00} \right) \tag{6}$$

$$\mathrm{IoU}_c = \sum_{i=1}^{N} \sum_{j=1}^{N} q_{ij} / \left(\sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij} - q_{00} \right). \tag{7}$$

SeK is proposed along with mIoU to further evaluate the segmentation performance in pixels that are classified as changed regions. Let $\hat{Q} = \hat{q}_{ij}$ be the confusion matrix when true positive non-changed pixels ($q_{0,0}$) are discarded.

$$\rho = \sum_{i=0}^{N} \hat{q}_{ii} / \sum_{i=0}^{N} \sum_{j=0}^{N} \hat{q}_{ij} \tag{8}$$

And, SeK is defined as:

$$\eta = \sum_{i=0}^{N} \left(\sum_{j=0}^{N} \hat{q}_{ij} * \sum_{j=0}^{N} \hat{q}_{ji} \right) / \left(\sum_{i=0}^{N} \sum_{j=0}^{N} \hat{q}_{ij} \right)^2 \tag{9}$$

$$\mathrm{SeK} = e^{\mathrm{IOU}_c - 1} \cdot (\rho - \eta)/(1 - \eta). \tag{10}$$

Lastly, F_{scd} is proposed in [13] by adapting traditional F1 score for SCD to better evaluate segmentation performance in changed regions which is defined as:

$$\begin{aligned} P_{\mathrm{scd}} &= \sum_{i=1}^{N} q_{ii} / \sum_{i=1}^{N} \sum_{j=0}^{N} q_{ij} \\ R_{\mathrm{scd}} &= \sum_{i=1}^{N} q_{ii} / \sum_{i=0}^{N} \sum_{j=1}^{N} q_{ij} \\ F_{\mathrm{scd}} &= \frac{2 * P_{\mathrm{scd}} * R_{\mathrm{scd}}}{P_{\mathrm{scd}} + R_{\mathrm{scd}}} \end{aligned} \tag{11}$$

3. Synthetic Dataset Illustration

We illustrate the synthetic T2-images and corresponding change maps generated using real T1-images from SECOND in Figs. 1 and 2 and from CNAM-CD in Figs. 3 and 4, and we illustrate inpainted samples in Figs. 5 and 6.

4. Comparison Illustrations

We illustrate qualitative figures corresponding to cross-domain performance evaluation in LEVIR-CD and WHU-CD datasets in Figs. 7 and 8 respectively. We further illustrate more qualitative figures corresponding to within-domain performance evaluation in SECOND and CNAM-CD datasets in Figs. 9 and 10 respectively. We also illustrate the performance advantage of synthetic pretraining across different experiments Fig. 11.

5. Ablation Studies

We present a comprehensive comparison of performance across all ablation studies conducted over epochs, as described in the main document, on the SECOND validation set. We utilized a composite evaluation metric of $0.3 \times \text{mIoU} + 0.7 \times \text{SeK}$ following [58]. The encoder backbone ablation results are depicted in Fig. 12, the performance of CLIP variants in Fig. 13, the effects of different pretraining strategies in Fig. 14, the decoder backbone ablation in Fig. 15, and the impact of freezing CLIP weights for ImageNet pretraining and synthetic pretraining cases in Figs. 16 and 17 respectively.

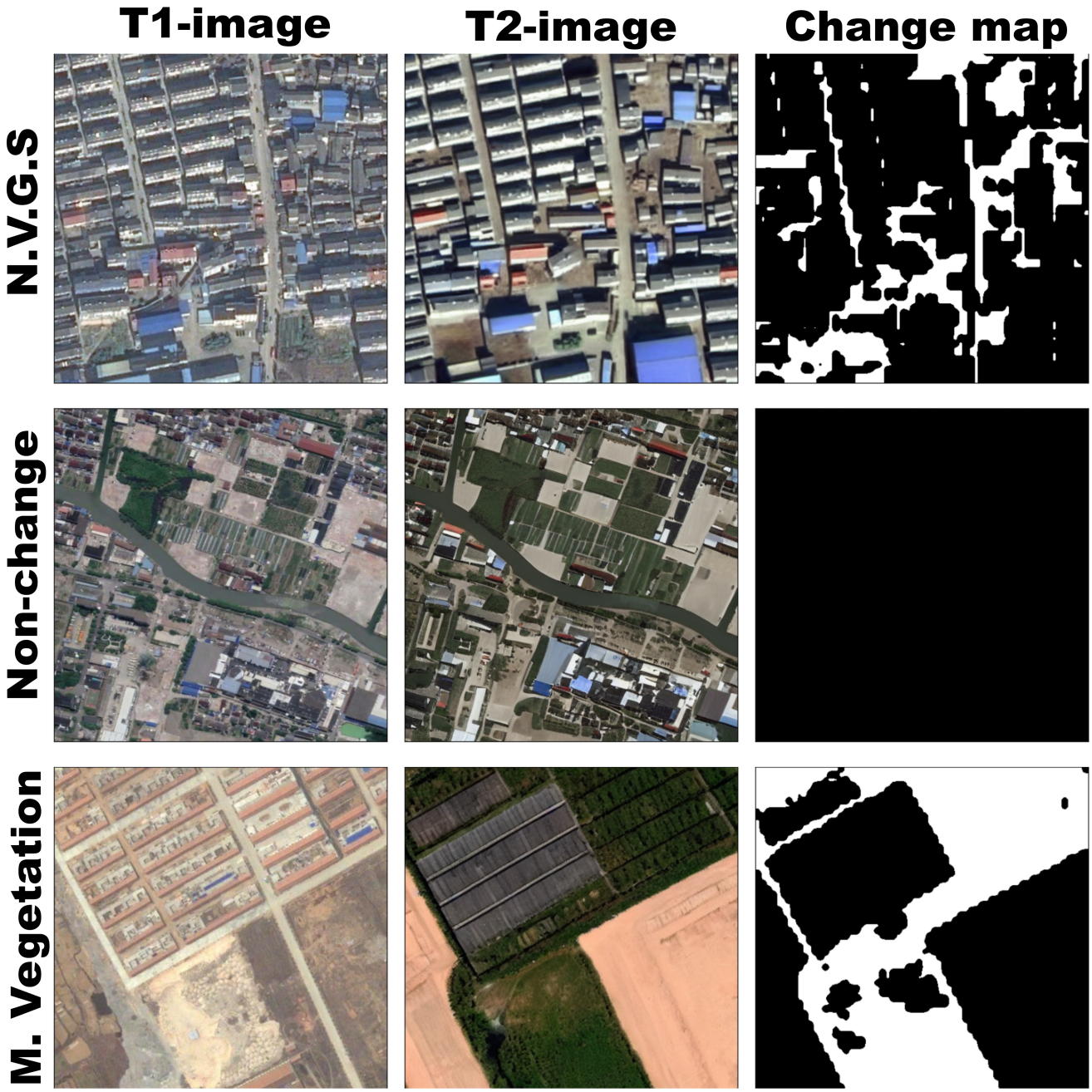

Figure 1. Synthetic samples generated using SECOND [52] are illustrated. ”N.V.G.S” refers to Non-Vegetated Ground Surface and ”M. Vegetation” refers to Medium Vegetation.

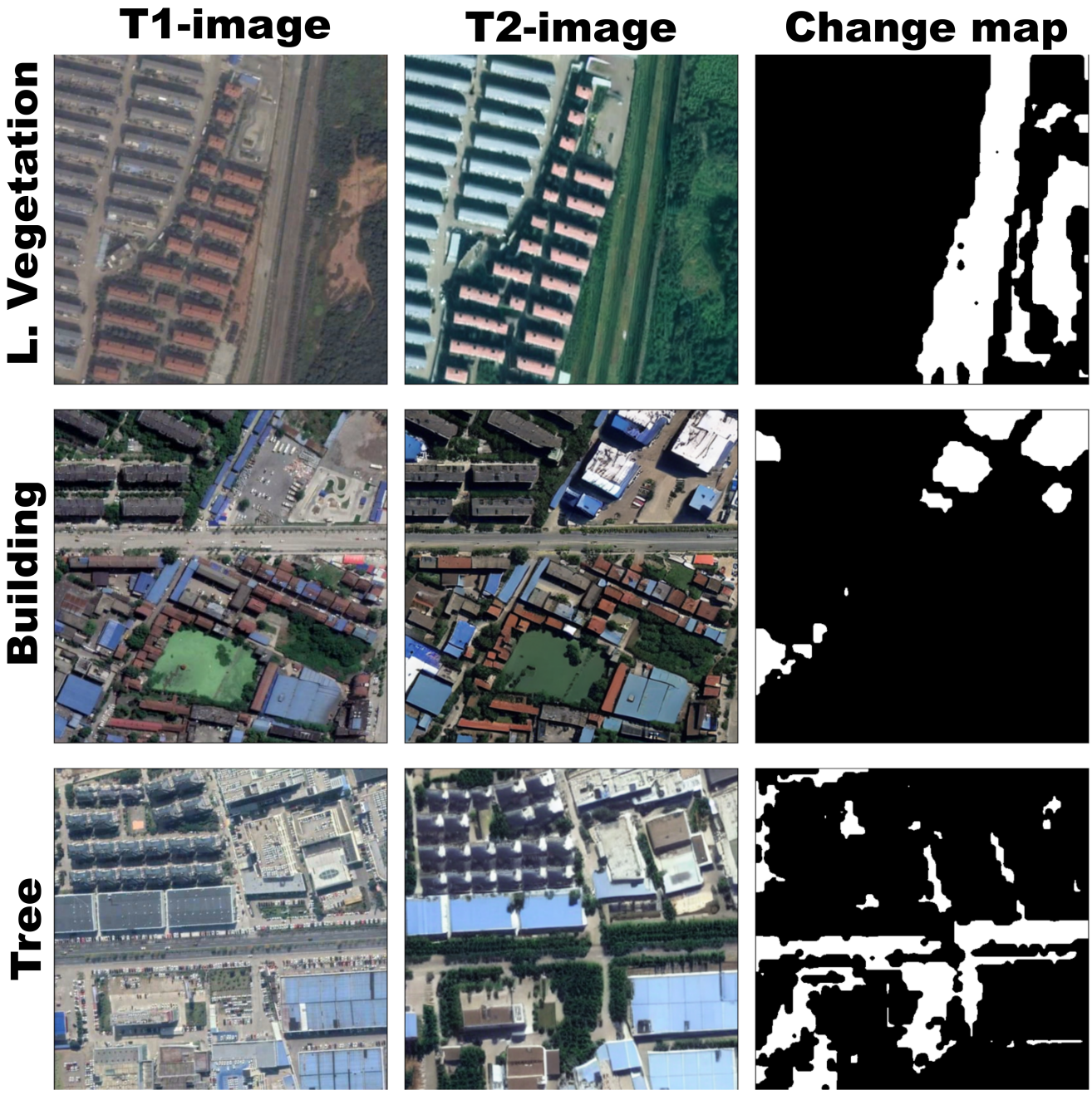

Figure 2. Synthetic samples generated using SECOND [52] are illustrated. ”L. Vegetation” refers to Low Vegetation.

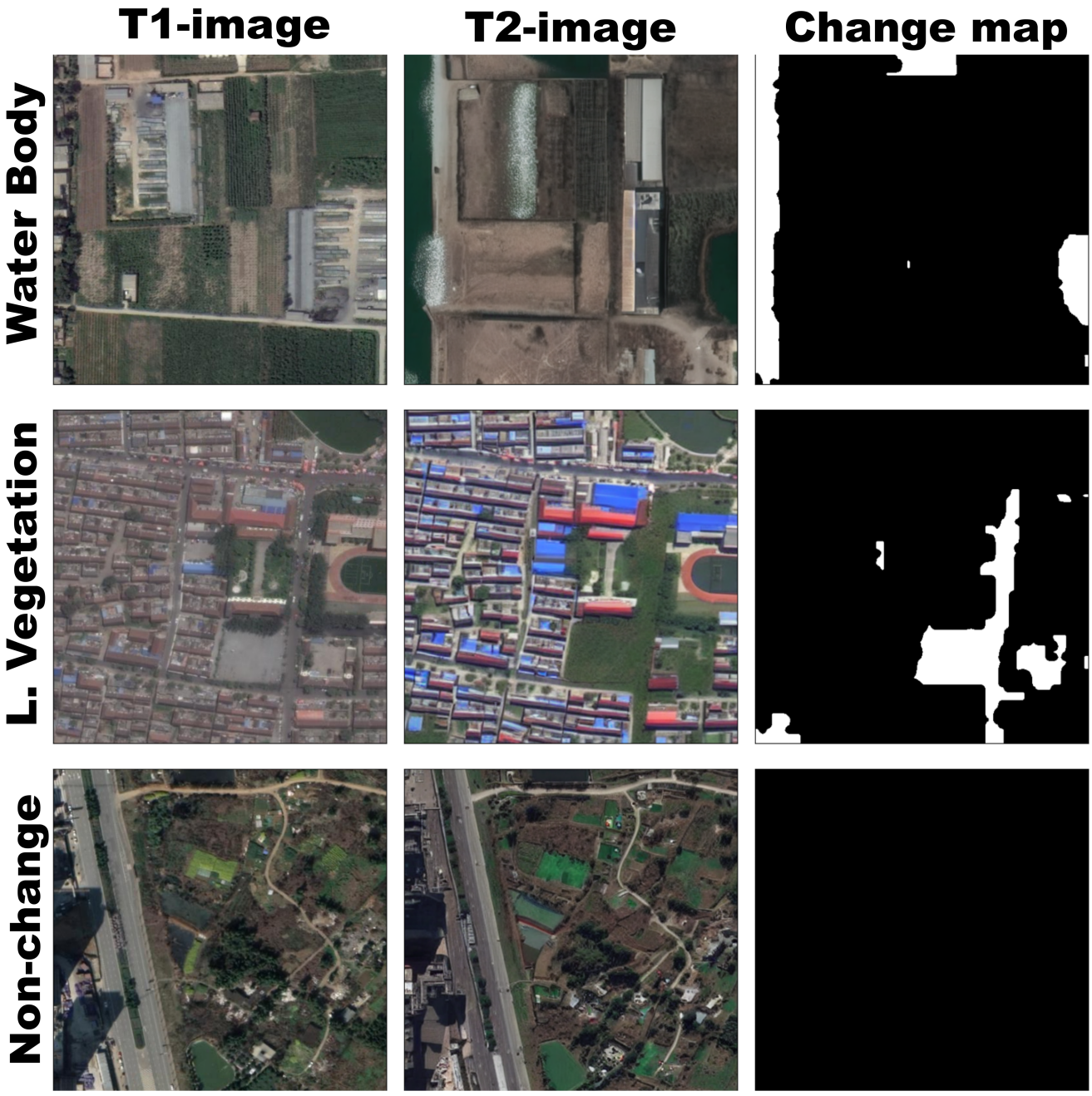

Figure 3. Synthetic samples generated using CNAM-CD [58] are illustrated. ”L. Vegetation” refers to Low Vegetation.

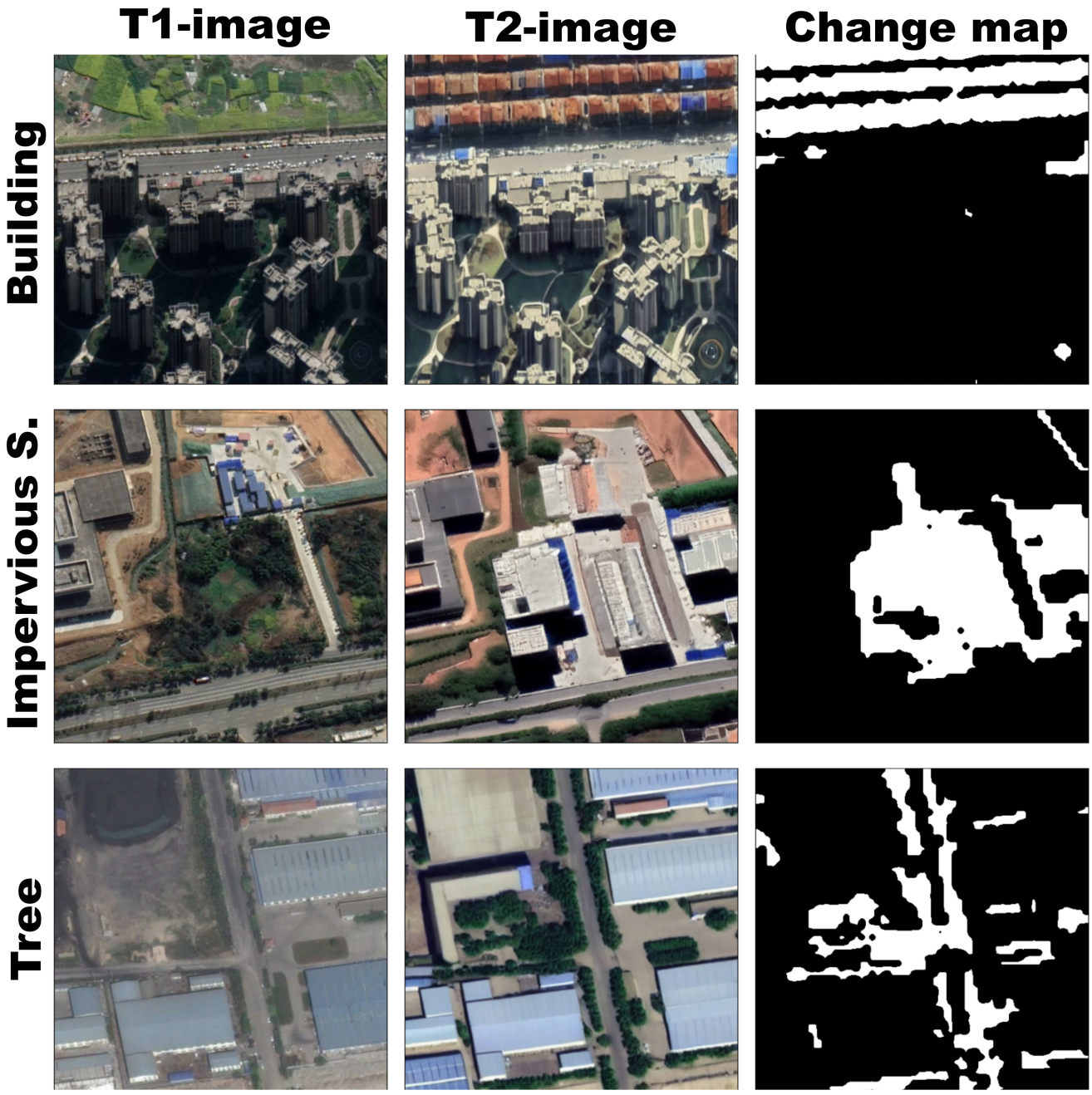

Figure 4. Synthetic samples generated using CNAM-CD [58] are illustrated. "Impervious S." refers to Impervious Surface.

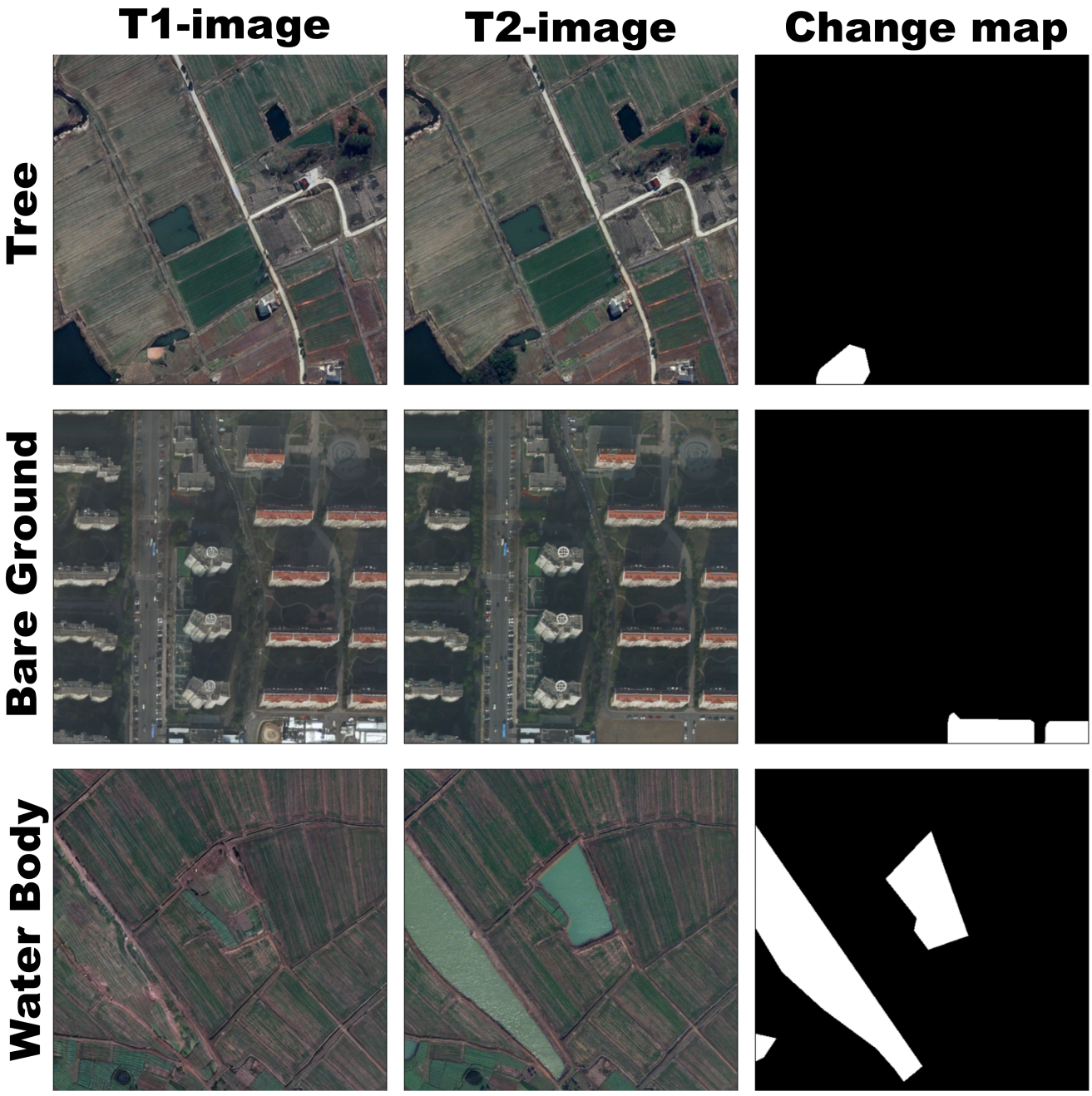

Figure 5. Inpainted synthetic samples are illustrated.

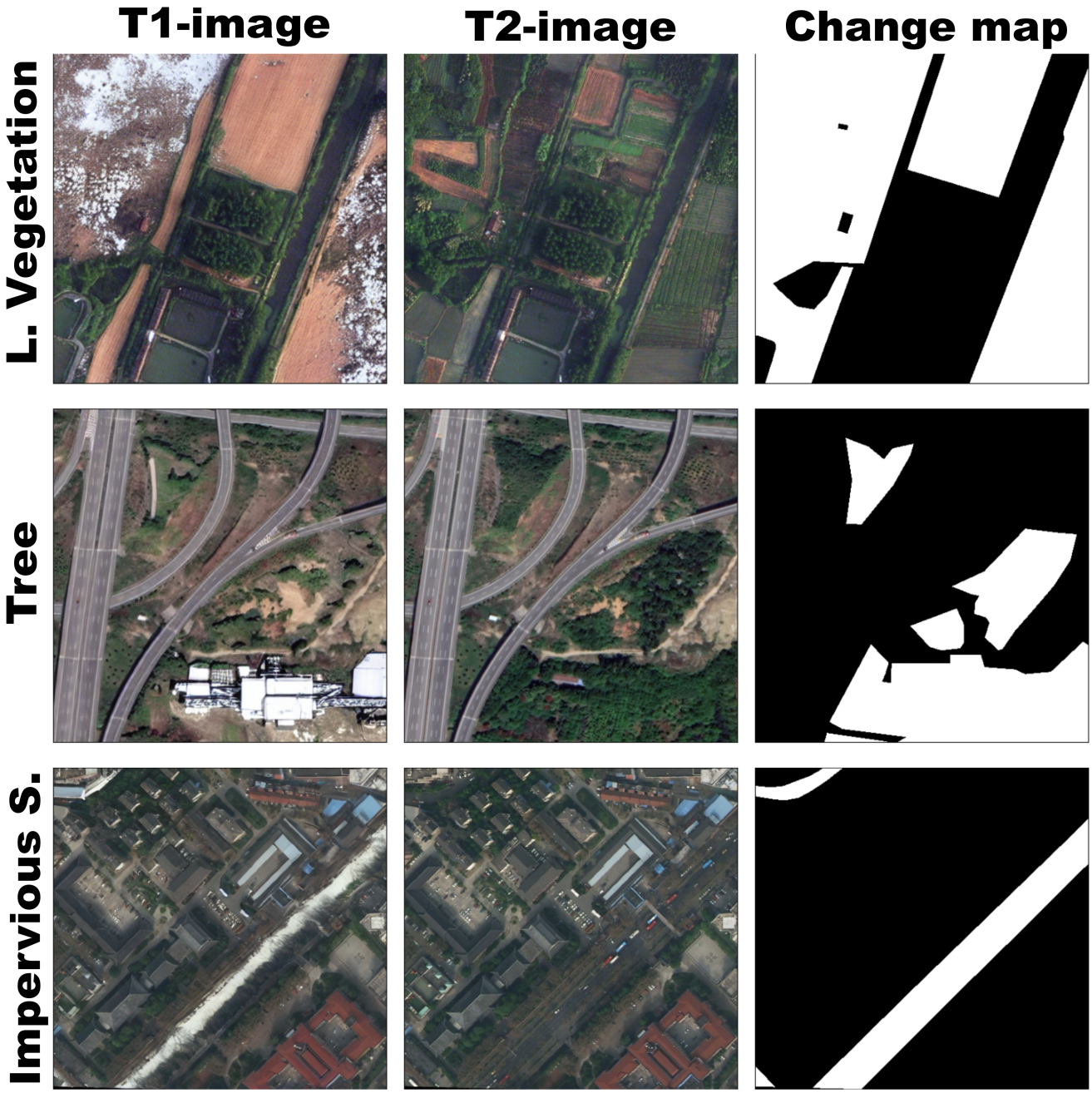

Figure 6. Inpainted synthetic samples are illustrated. ”L. Vegetation” refers to Low Vegetation and ”Impervious S.” refers to Impervious Surface.

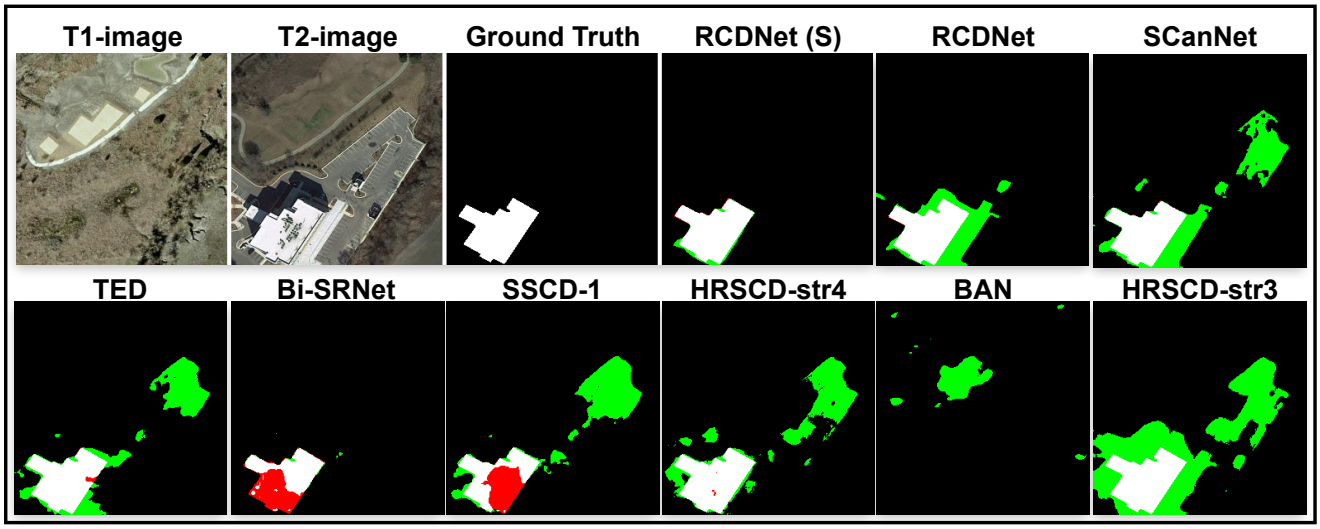

Figure 7. Cross-domain performance comparison in LEVIR-CD for ”building” category
. T1-images represent pre-change (earlier time point) and T2-images represent post-change (later time point). In the results, white pixels indicate true positives, black pixels indicate true negatives, green pixels denote false positives, and red pixels denote false negatives. The notation ”(S)” refers to synthetic pretraining applied to RCDNet.

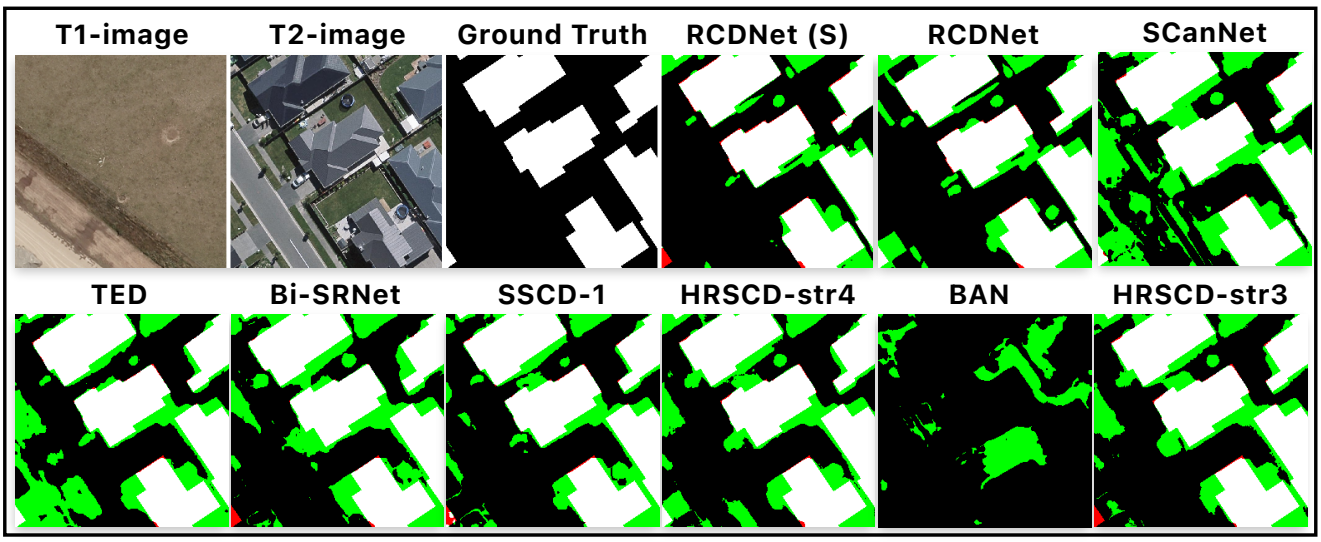

Figure 8. Cross-domain performance comparison in WHU-CD for ”building” category
. T1-images represent pre-change (earlier time point) and T2-images represent post-change (later time point). In the results, white pixels indicate true positives, black pixels indicate true negatives, green pixels denote false positives, and red pixels denote false negatives. The notation ”(S)” refers to synthetic pretraining applied to RCDNet.

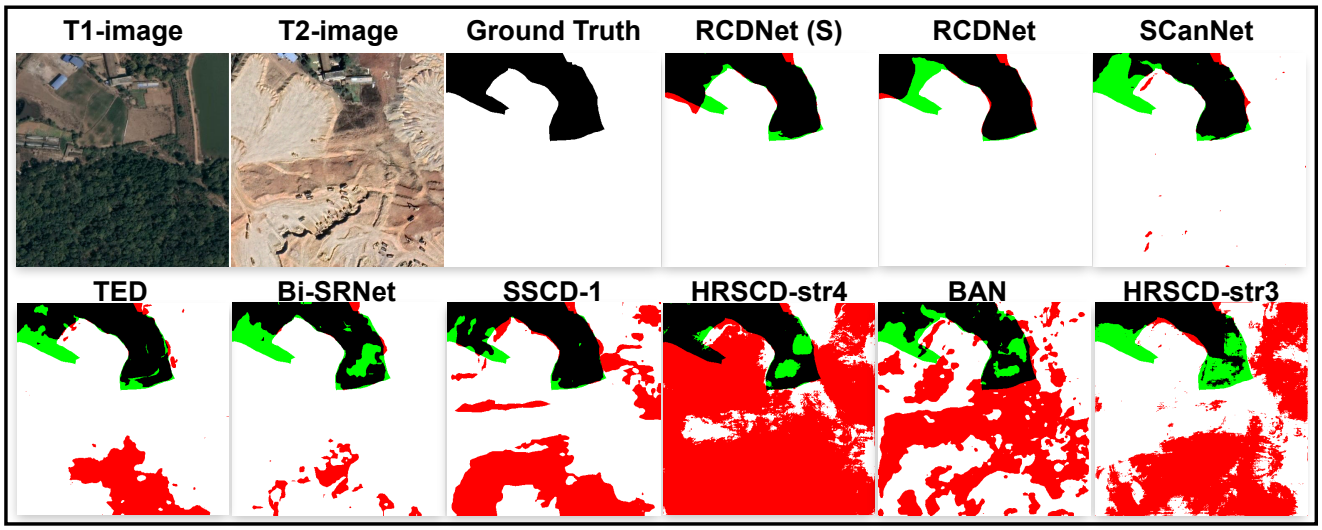

Figure 9. Within-domain performance comparison in CNAM-CD for ”Bare Ground” category
. T1-images represent pre-change (earlier time point) and T2-images represent post-change (later time point). In the results, white pixels indicate true positives, black pixels indicate true negatives, green pixels denote false positives, and red pixels denote false negatives. The notation ”(S)” refers to synthetic pretraining applied to RCDNet.

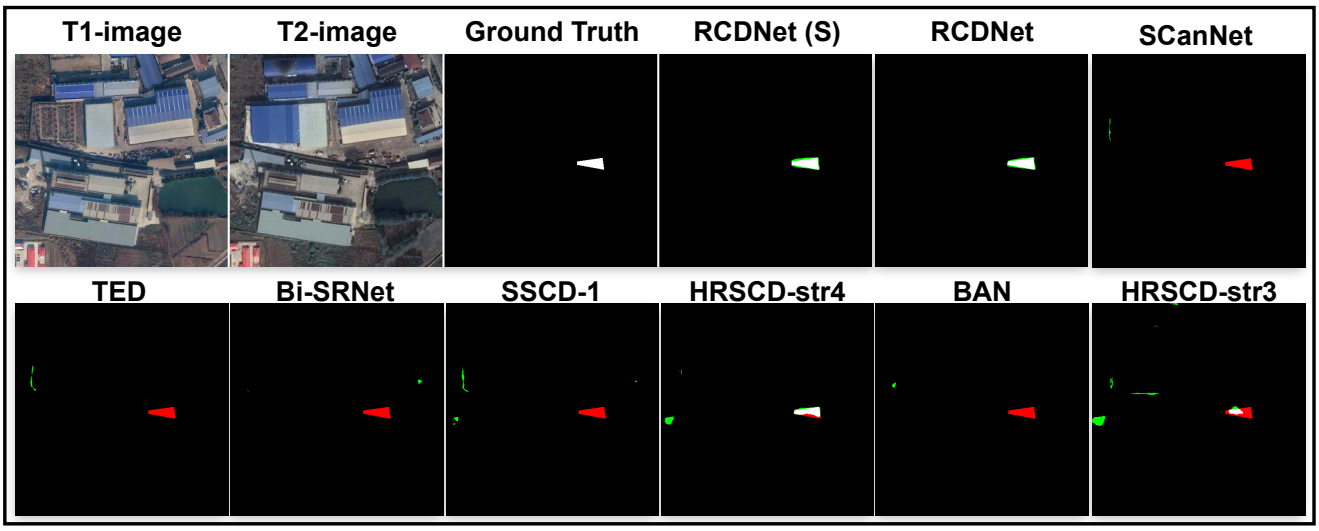

Figure 10. Within-domain performance comparison in SECOND for ”Non-Vegetated Ground Surface” category
. T1-images represent pre-change (earlier time point) and T2-images represent post-change (later time point). In the results, white pixels indicate true positives, black pixels indicate true negatives, green pixels denote false positives, and red pixels denote false negatives. The notation ”(S)” refers to synthetic pretraining applied to RCDNet.

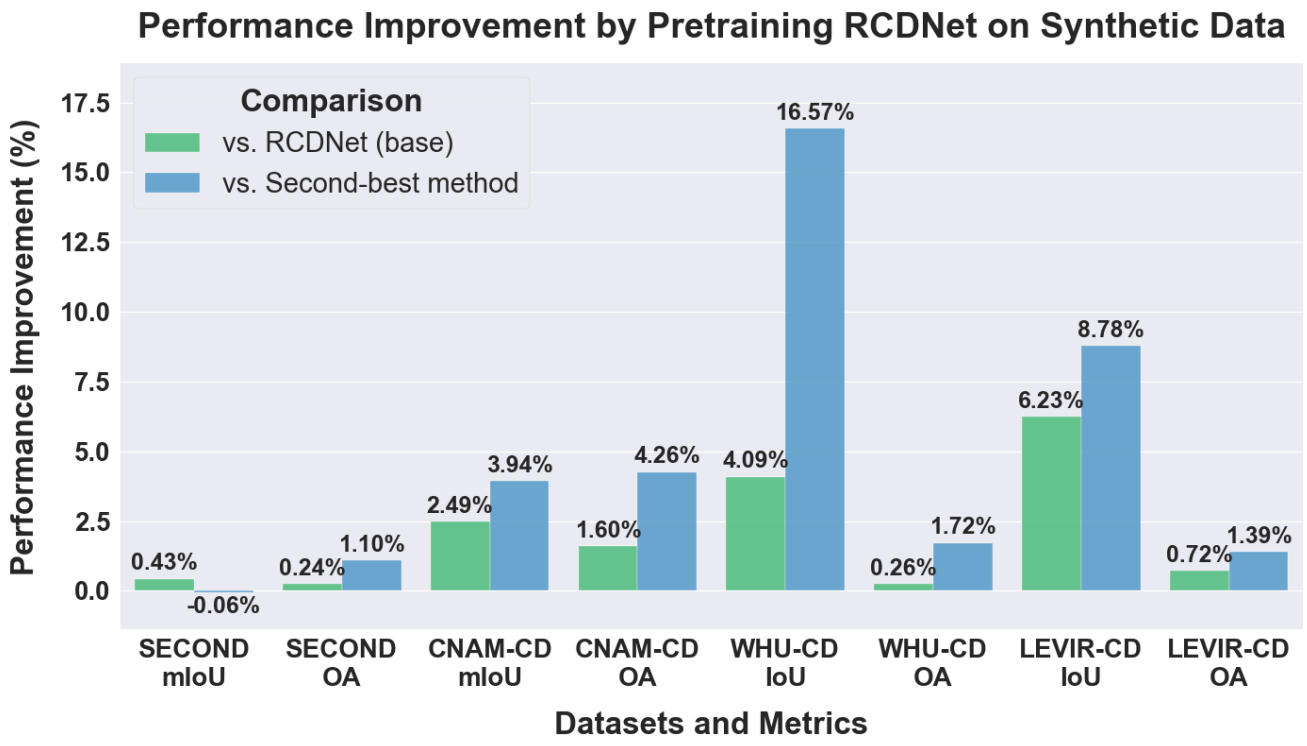

Figure 11. Performance improvements from synthetic data pretraining over baseline RCDNet and second-best methods across multiple datasets and metrics. Metrics for WHU-CD and LEVIR-CD represent cross-domain experiments, while SECOND and CNAM-CD metrics represent within-domain performance.

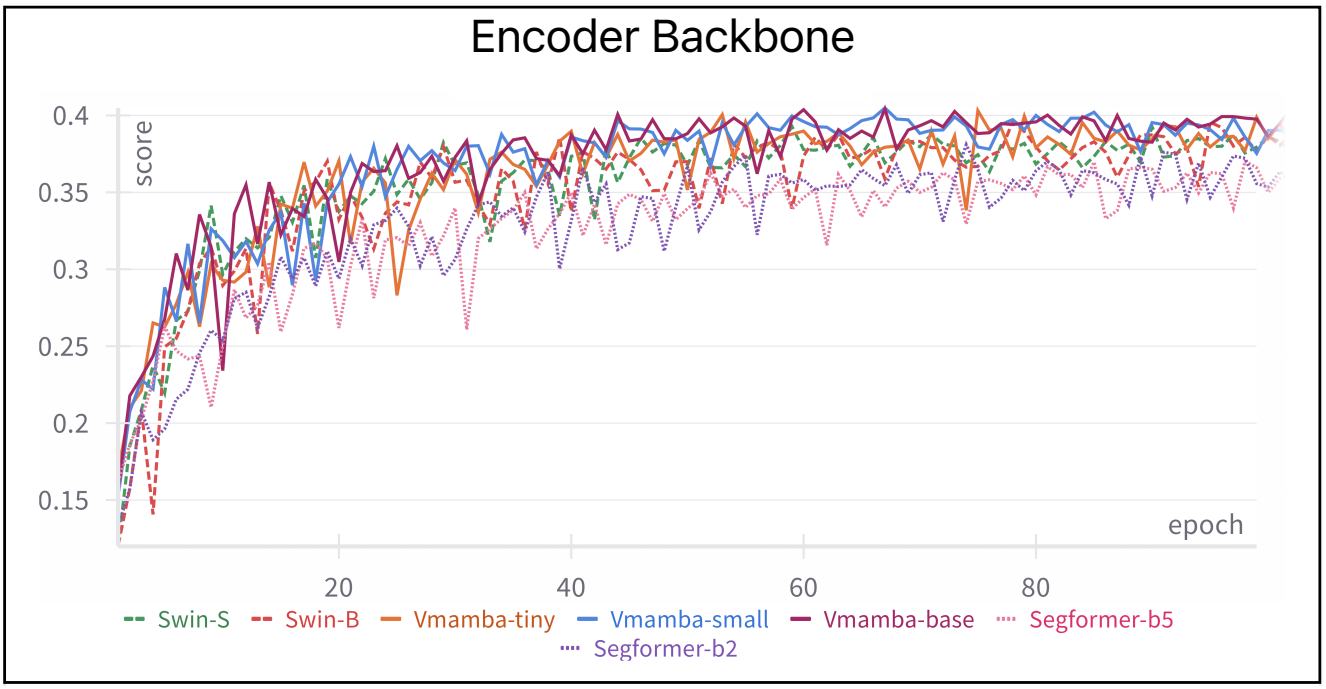

Figure 12. Ablation results comparing various encoder backbones.

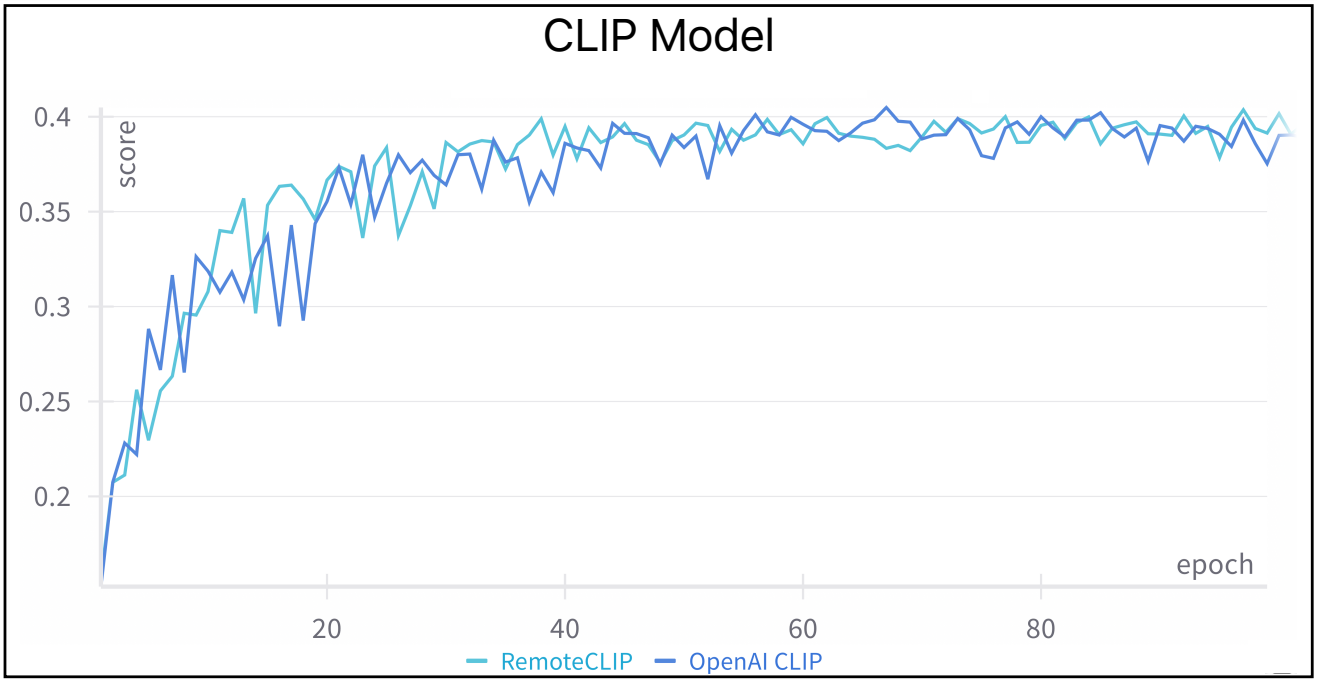

Figure 13. Ablation results comparing different CLIP models.

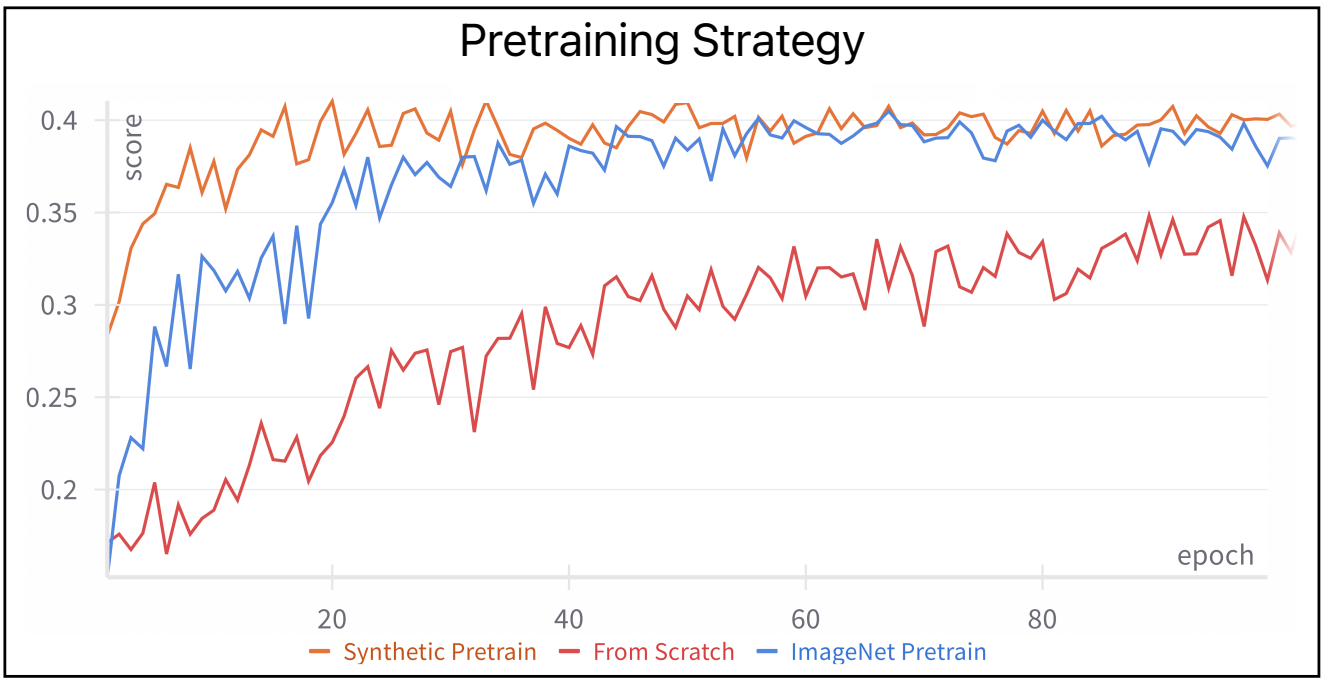

Figure 14. Ablation results comparing various pretraining strategies.

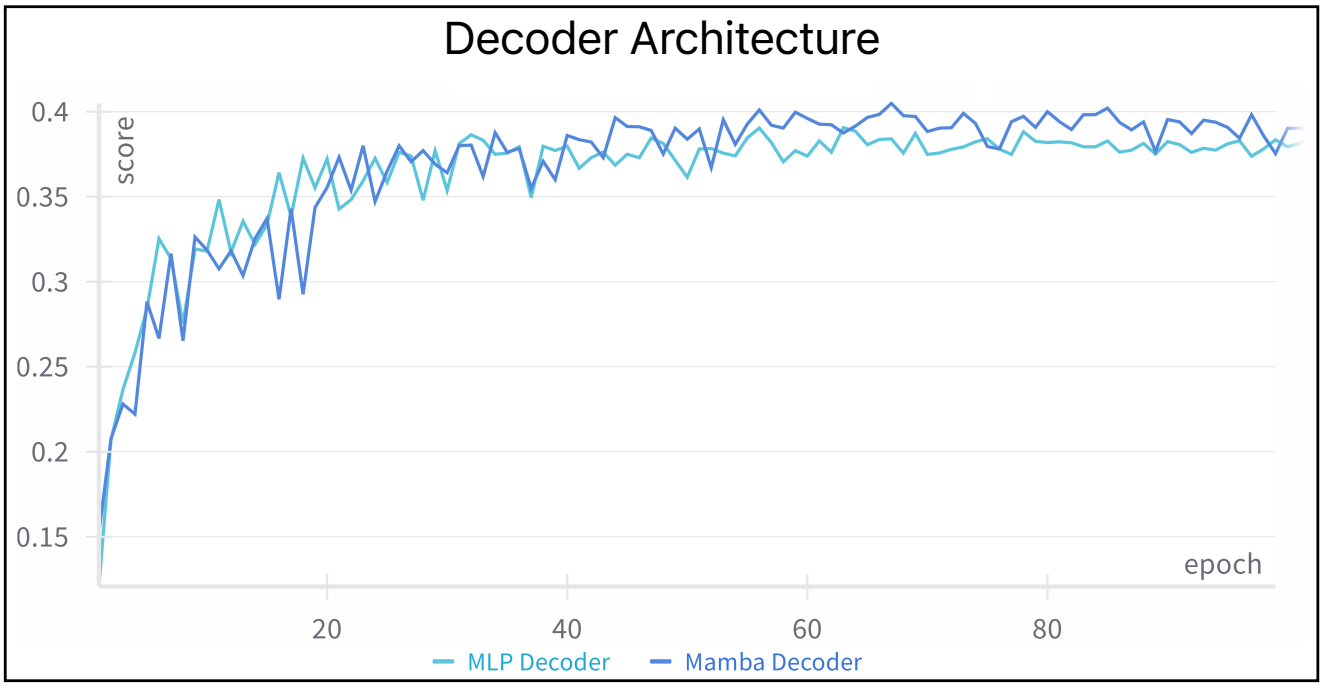

Figure 15. Ablation results comparing different decoder backbones.

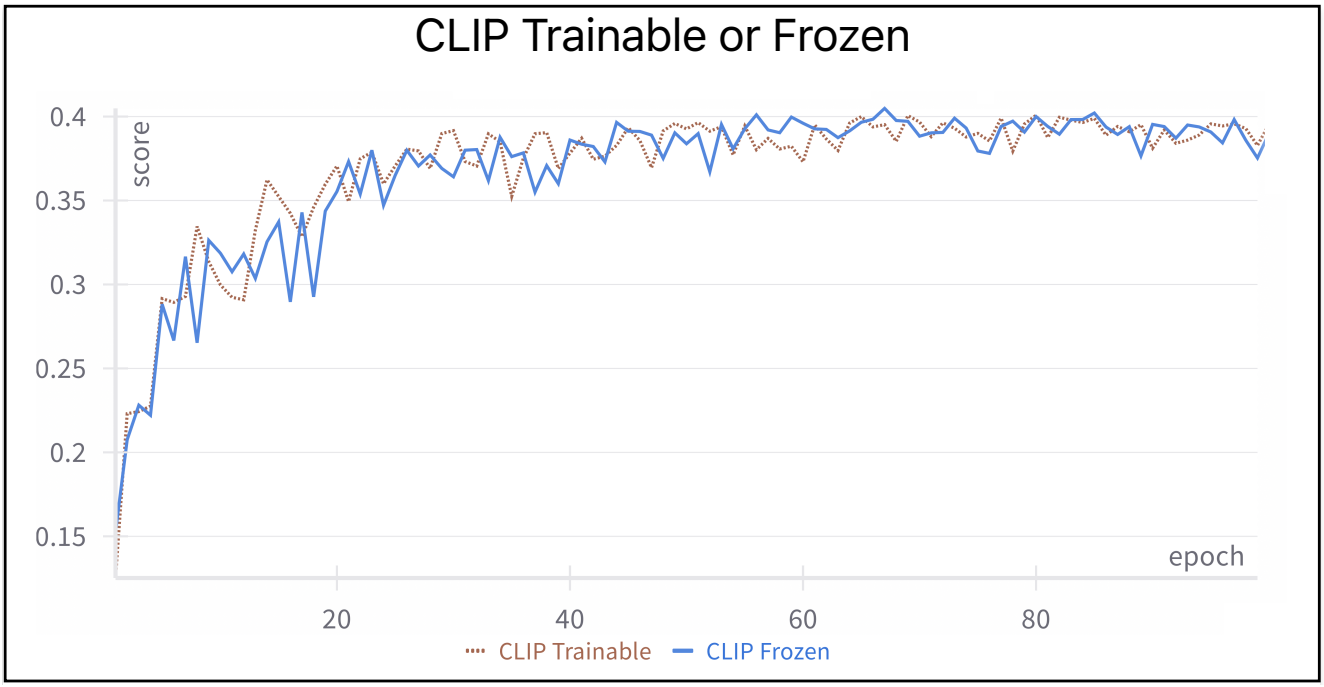

Figure 16. Ablation results for freezing or fine-tuning CLIP weights using LoRA [27] with ImageNet pretraining.

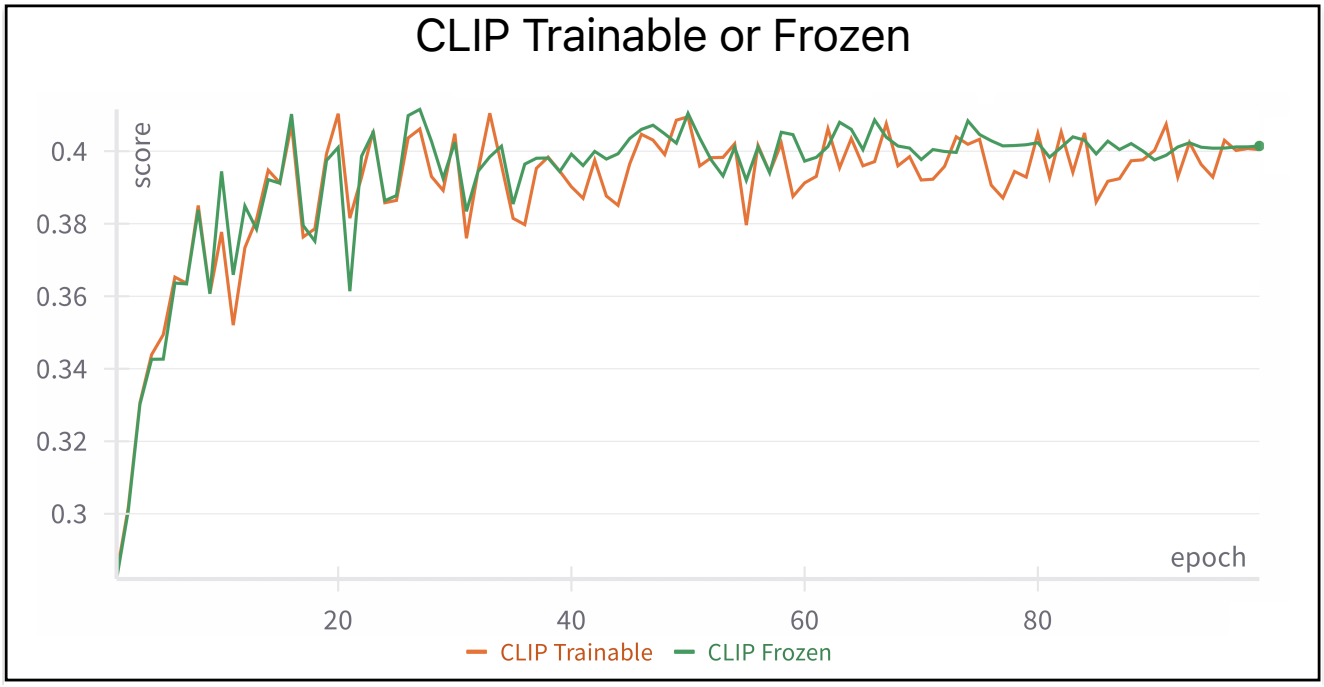

Figure 17. Ablation results for freezing or fine-tuning CLIP weights using LoRA [27] with synthetic pretraining.