UniChange: Unifying Change Detection with Multimodal Large Language Model

Supplementary Material

Overview

This supplementary material provides additional details to support the main manuscript. The document is structured as follows: Sec. A breaks down the evaluation metrics for BCD and SCD tasks, while Sec. B offers qualitative visual comparisons against state-of-the-art methods. Section C describes the experimental datasets, including their technical specifications. Finally, Sec. D and Sec. E demonstrate UniChange's generalization capability and provide comprehensive performance comparisons against prior methods, respectively.

A. Evaluation Metrics

A.1. BCD Metrics

To evaluate the performance of UniChange on the binary change detection (BCD) task, we employ four standard pixel-level metrics: Precision (P), Recall (R), F1 Score (F1), and Intersection over Union (IoU). Precision measures the proportion of correctly predicted change pixels among all pixels classified as change. Recall indicates the proportion of true positive pixels among all truly positive pixels in the ground truth. The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances these two measures. Finally, IoU measures the overlap between predicted and ground-truth positive regions, serving as a robust measure of segmentation quality.

The metrics are individually defined as follows, where TP, TN, FP, and FN represent the number of true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively:

$$\begin{aligned} \mathrm{P} &= \mathrm{TP} \,/\, (\mathrm{TP} + \mathrm{FP}), \\ \mathrm{R} &= \mathrm{TP} \,/\, (\mathrm{TP} + \mathrm{FN}), \\ \mathrm{F1} &= 2 \times \mathrm{P} \times \mathrm{R} / (\mathrm{P} + \mathrm{R}), \\ \mathrm{IoU} &= \mathrm{TP} \,/\, (\mathrm{TP} + \mathrm{FP} + \mathrm{FN}). \end{aligned} \tag{1}$$

A.2. SCD Metrics

The assessment of semantic change detection (SCD) model performance is executed using a collection of specialised metrics, all derived from the confusion matrix $Q = \{q_{ij}\}$, where q_{ij} records the count of pixels classified as class i with a ground truth label of j.

The mean Intersection over Union for SCD (mIoU) is utilised to evaluate overall segmentation quality, established as the arithmetic mean of the Intersection over Union for the regions without change (IoU_{nc}) and all changing regions (IoU_{c}):

$$\mathrm{mIoU} = (\mathrm{IoU}_{nc} + \mathrm{IoU}_{c})/2. \tag{2}$$

IoU_{nc} measures the overlap between the predicted unchanged regions and the ground-truth unchanged regions:

$$\mathrm{IoU}_{nc} = q_{00} / (\sum_{i=0}^{N} q_{i0} + \sum_{j=0}^{N} q_{0j} - q_{00}). \tag{3}$$

Conversely, IoU_{c} measures the overall segmentation quality of all change regions, treating all distinct semantic change categories as a single change class:

$$\mathrm{IoU}_{c} = \sum_{i=1}^{N}\sum_{j=1}^{N} q_{ij} / (\sum_{i=0}^{N}\sum_{j=0}^{N} q_{ij} - q_{00}). \tag{4}$$

The Separation kappa coefficient (SeK) provides a valuable measure of semantic discrimination amidst class imbalance, particularly designed to diminish the influence of the prevalent unchanged class. SeK is computed from the confusion matrix $\hat{Q} = \{\hat{q_{ij}}\}$, where $\hat{q}_{ij} = q_{ij}$, but $\hat{q}_{00} = 0$, and is calculated as

$$\begin{aligned} \rho &= \sum_{i=0}^{N} \hat{q}_{ii} \Big/ \sum_{i=0}^{N}\sum_{j=0}^{N} \hat{q}_{ij}, \\ \eta &= \sum_{i=0}^{N} \left(\sum_{j=0}^{N} \hat{q}_{ij} \times \sum_{j=0}^{N} \hat{q}_{ji} \right) \Big/ \left(\sum_{i=0}^{N}\sum_{j=0}^{N} \hat{q}_{ij} \right)^{2}, \\ SeK &= e^{IoU_c - 1} \cdot (\rho - \eta)/(1 - \eta). \end{aligned} \tag{5}$$

The F1-Score for SCD (F_{scd}) offers a focused quantification of semantic transition accuracy within change regions. This metric is derived from the precision (P_{scd}) and recall (R_{scd}) over the pixels determined to have changed.

The SCD precision P_{scd} and recall R_{scd} are defined as

$$\begin{aligned} P_{scd} &= \sum_{i=1}^{N} q_{ii} \Big/ \sum_{i=1}^{N}\sum_{j=0}^{N} q_{ij}, \\ R_{scd} &= \sum_{i=1}^{N} q_{ii} \Big/ \sum_{i=0}^{N}\sum_{j=1}^{N} q_{ij}. \end{aligned} \tag{6}$$

The final F_{scd} is the harmonic mean of these two components:

$$F_{scd} = 2 \times P_{scd} \times R_{scd} / (P_{scd} + R_{scd}). \tag{7}$$

Table 1. Detailed information about change detection datasets used for experiments.

Dataset	Resolution	Image Size	Image Number	Evaluation Task	Class
WHU-CD	0.075m	32507 × 15354	1	BCD	Building
S2Looking	0.5-0.8m	1024 × 1024	5000	BCD	Building
LEVIR-CD+	0.5m	1024 × 1024	985	BCD	Building
SECOND	0.5-3m	512 × 512	4662	SCD	Building, Low Vegetation, Tree, Water, Playground, Bare Ground

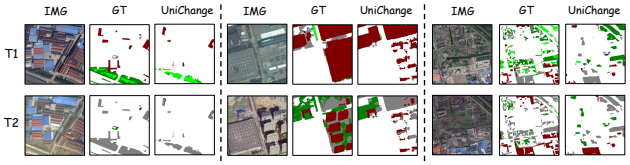

Figure 1. Failure cases.

The performance of UniChange on the SCD task is comprehensively evaluated using BCD metrics: IoU and F1 (F_{bcd}) and SCD metrics: mIoU, F_{scd} and SeK.

B. Visual Comparisons

To qualitatively evaluate the performance of our proposed UniChange, we provide comprehensive visual comparisons against other state-of-the-art (SOTA) methods on both binary and semantic change detection tasks.

Fig. 1 illustrates the inherent limitations of the current UniChange framework. Specifically, the model exhibits a performance bottleneck when encountering infinitesimal change regions or targets characterized by highly irregular geometric topologies.

Fig. 2 illustrates the visual results for binary change detection (BCD). The comparisons on the WHU-CD, S2Looking, and LEVIR-CD+ datasets all show that UniChange performs better than competing methods. Its generated change masks are more complete and have fewer false positives (FP) and false negatives (FN).

Furthermore, Fig. 3 presents the qualitative results for semantic change detection (SCD). As shown in the figure, UniChange not only accurately locates the changed regions but also exhibits a strong capability in correctly identifying the specific semantic categories of the changes. These visual results underscore the superior performance and generalisation ability of our unified model.

C. Dataset Details

To comprehensively evaluate our model, we utilise four diverse, publicly available remote sensing datasets, covering both binary change detection (BCD) and semantic change detection (SCD) tasks. The detailed specifications of these datasets are summarised in Tab. 1.

For the BCD task, we utilise three datasets, all of which focus on identifying building changes. The WHU-CD dataset provides a single, massive-scale image pair (32507 × 15354 pixels), captured at an ultra-high 0.075m resolution. It is particularly notable for its coverage, capturing a 20.5 km^2 area in Christchurch, New Zealand. The dual-temporal images, captured in 2012 and 2016, document the region's reconstruction following a major earthquake. This provides an excellent case for studying large-scale urban recovery. In addition, we use S2Looking, a large-scale dataset consisting of 5000 dual-temporal image pairs, which are 1024 × 1024 pixels in size and have a resolution of 0.5-0.8m. Its defining characteristic is the use of satellite side-looking images captured at various off-nadir angles, a sharp contrast to typical near-nadir (or top-down) imagery. This dataset focusses on globally distributed rural areas. It challenges models with large illumination variances and complex rural scenes. It also features geometric distortions from the oblique angles. Finally, we utilise LEVIR-CD+, an expanded version of the LEVIR-CD dataset. It contains 985 image pairs of 0.5m resolution Google Earth imagery. Each image pair is 1024 × 1024 pixels. In contrast to S2Looking, LEVIR-CD+ features near-nadir images and focusses on building changes in urban and suburban environments, primarily covering 20 different regions in Texas, USA.

For the more complex SCD task, we use the SECOND dataset. This is a large-scale benchmark containing 4,662 pairs of aerial images, each 512×512 pixels in size. The images come from several platforms and sensors. They cover major cities in China, including Hangzhou, Chengdu, and Shanghai. Unlike the BCD datasets, SECOND requires the model to identify semantic transitions among six distinct land-cover classes: building, low vegetation, tree, water, playground, and bare ground. A defining feature of this dataset is its annotation method; it provides land-cover map pairs and nonchange masks. This structure is specifically designed to challenge models. It tests their ability to detect changes that occur between the same land-cover class. For example, a model must find where an old playground is removed and a new one is built in its place. This is a critical capability that many other datasets cannot evaluate.

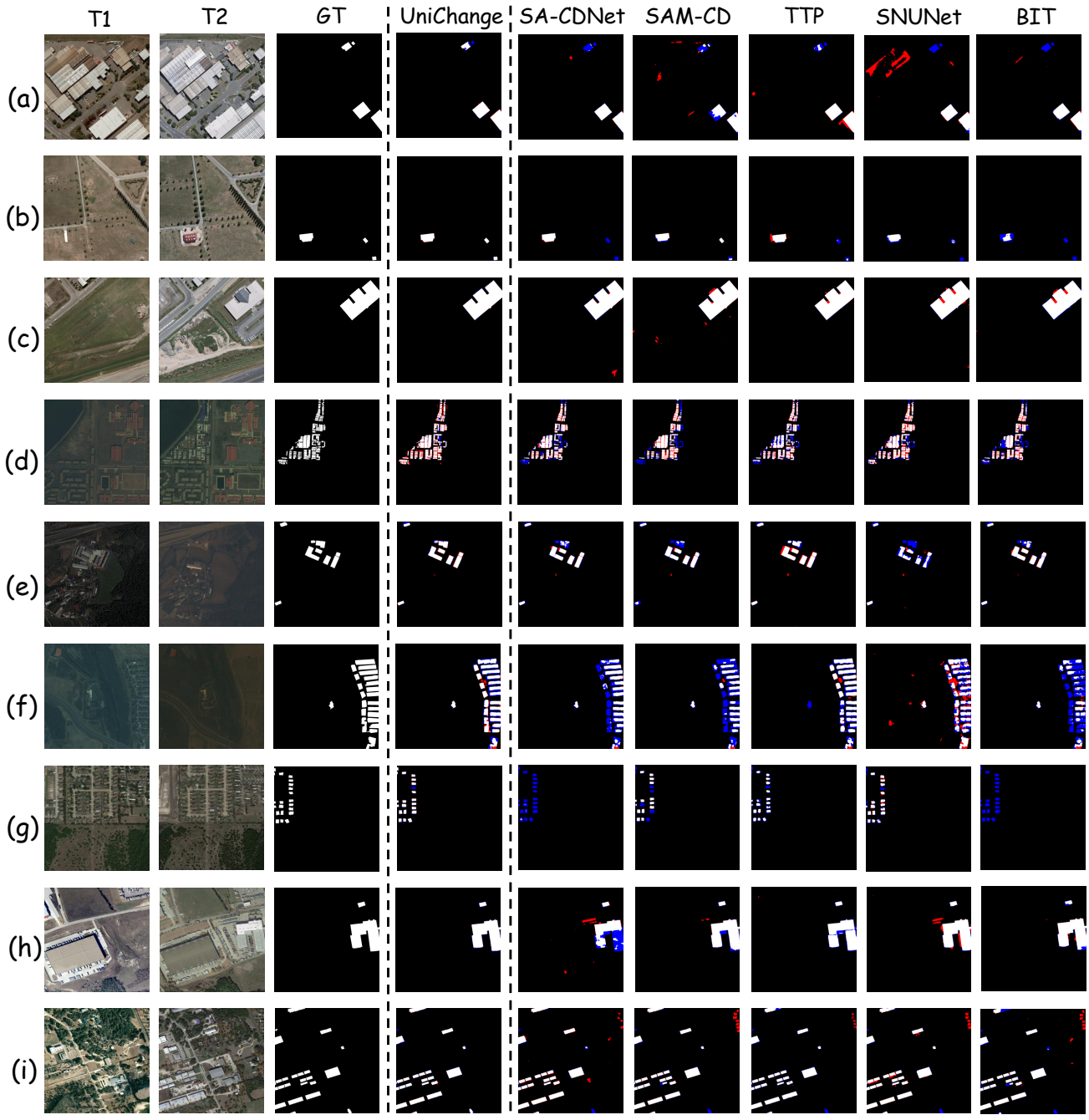

Figure 2. Visual comparisons of the UniChange with other state-of-the-art methods for binary change detection. **Red** means false positives (FP), while **Blue** denotes false negatives (FN). Samples (a) (b) (c) are from the WHU-CD dataset, (d) (e) (f) are from the S2Looking dataset, and (g) (h) (i) are from the LEVIR-CD+ dataset.

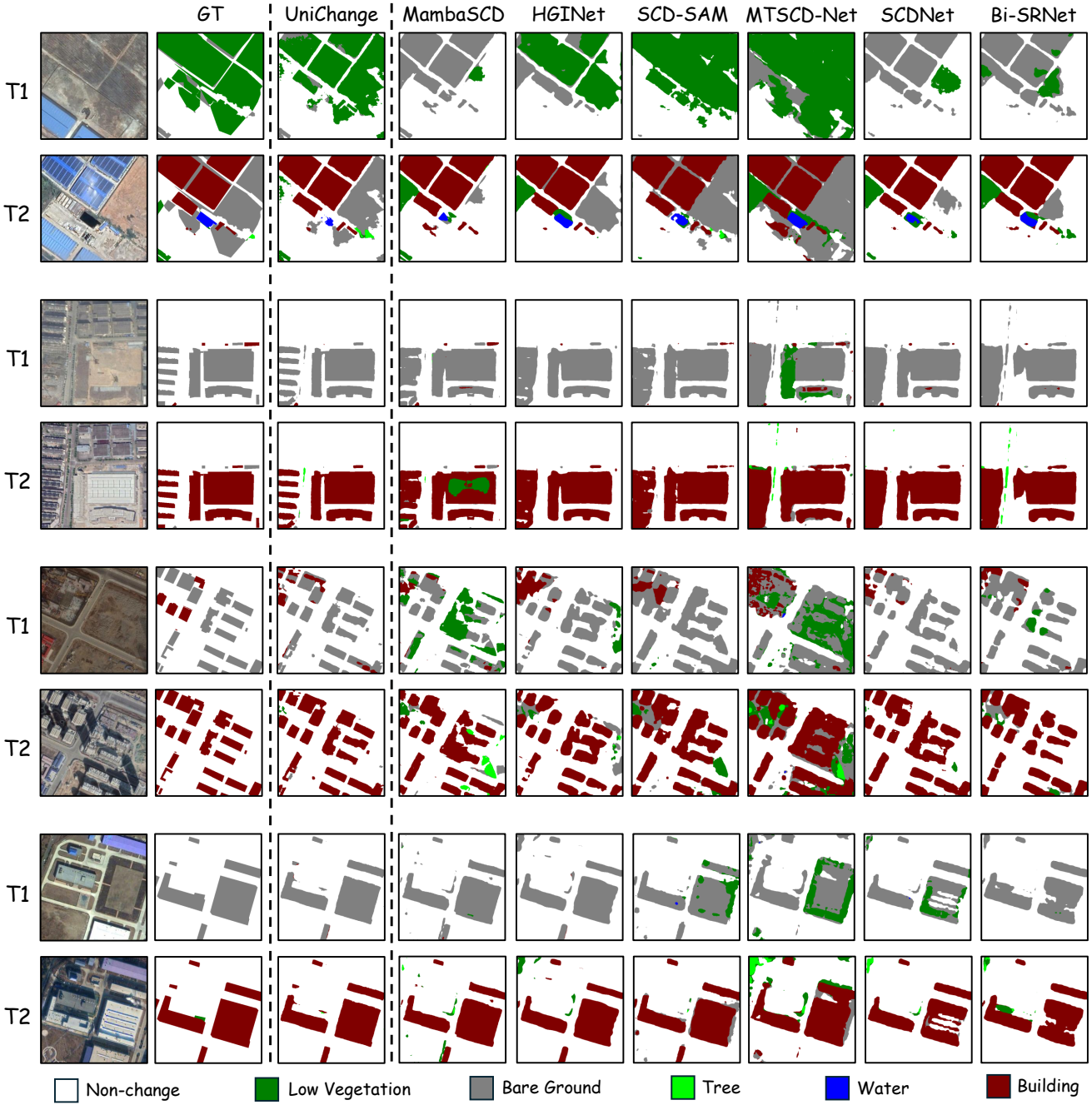

Figure 3. Visual comparisons of the UniChange with other state-of-the-art methods for semantic change detection. The colour legend is as follows: **White** represents Non-change, **dark green** represents Low Vegetation, **grey** represents Bare Ground, **bright green** represents Tree, **blue** represents Water, and **dark red** represents Building.

Table 2. Experimental Results on Model Generalization.

Method	EGY-BCD		HRCUS-CD	
	IoU	F1	IoU	F1
Changer(Joint)	8.10	14.99	8.67	15.96
DynamicEarth-C1	33.15	49.97	19.46	32.58
DynamicEarth-C2	39.16	56.28	20.58	34.14
UniChange	**61.56**	**76.21**	**24.76**	**39.69**

Table 3. Comparison of our separate and joint training performance against previous SOTA results.

Method	WHU	S2L	LEV+	SEC	
	IoU	IoU	IoU	mIoU	SeK
Previous SOTA	90.08	50.96	76.12	72.73	21.31
UniChange(Separate)	89.68	52.47	78.36	72.74	22.54
UniChange(Joint)	**90.41**	**53.04**	**78.87**	**72.85**	**23.02**

D. Generalization

To evaluate the generalization capability of UniChange, we conduct zero-shot experiments on the unseen EGY-BCD and HRCUS-CD datasets. We compare our model against Changer (training jointly on UniChange's datasets.) and the state-of-the-art zero-shot change detection framework, DynamicEarth, using two distinct configurations: C1 (SAM-DINOv2-SegEarth-OV) and C2 (APE-/-DINO). As reported in Tab. 2, UniChange consistently outperforms these competitive baselines, demonstrating superior cross-domain robustness and adaptability to diverse remote sensing scenarios.

E. Comprehensive Comparison

In this section, we present a thorough evaluation of UniChange under various training paradigms across four benchmark datasets: **WHU-CD (WHU)**, **S2Looking (S2L)**, **LEVIR-CD+ (LEV+)**, and **SECOND (SEC)**. We categorize the training strategies into **"Separate"** (training on each dataset individually) and **"Joint"** (raining jointly on UniChange's datasets.) to examine the model's versatility.

As summarized in Tab. 3, UniChange's performance in the separate-training setting already surpasses the **Previous SOTA** results, which represent the highest performance achieved by any prior specialized method for each specific metric. This validates the fundamental architectural strength of our approach. Furthermore, Tab. 4 demonstrates that UniChange in the joint training configuration consistently outperforms other unified or joint-training frameworks (e.g., Falcon and RSBuilding). For a rigorous and comprehensive comparison, we also trained Changer (a representative method originally designed for separate training) under the joint setting to serve as a unified comparative baseline. The results highlight that UniChange not only resolves semantic conflicts across diverse datasets but also benefits from joint supervision to achieve optimal performance.

Table 4. Comparison of our joint training performance against other joint training methods.

Method	WHU		S2L		LEV+	
	IoU	mIoU	IoU	mIoU	IoU	mIoU
Falcon	-	53.10	-	57.00	-	57.00
RSUniVLM	54.19	-	-	-	-	-
Changer(Joint)	79.19	89.27	41.36	70.24	68.93	83.71
RSBuilding	88.84	94.25	51.30	75.28	76.37	87.65
UniChange(Joint)	**90.41**	**95.13**	**53.04**	**76.17**	**78.87**	**88.93**