OpenDPR: Open-Vocabulary Change Detection via Vision-Centric Diffusion-Guided Prototype Retrieval for Remote Sensing Imagery

Supplementary Material

1. Additional Dataset Details

WHU-CD [2]. The WHU-CD dataset is designed for building change detection in Christchurch, New Zealand. It consists of a pair of high-resolution aerial images acquired in 2012 and 2016, with an original size of 32,507 × 15,354 pixels and a spatial resolution of 0.3m. The dataset is cropped into 256 × 256 patches, yielding 7434 training pairs, 743 validation pairs, and 744 testing pairs.

LEVIR-CD [1]. The LEVIR-CD dataset targets building change detection in Texas, USA, covering various building types such as villas, high-rise apartments, garages, and warehouses. It consists of 637 pairs of high-resolution aerial images from Google Earth, each with a size of 1024 × 1024 pixels and a spatial resolution of 0.5m. The temporal span ranges from 5 to 14 years, with significant land-use changes, particularly due to construction activities. The dataset is divided into 445 training pairs, 64 validation pairs, and 128 testing pairs. To support the weakly supervised pre-training of S2C, the images are further cropped into non-overlapping 256 × 256 patches, resulting in 7120 training pairs, 1024 validation pairs, and 2048 testing pairs. During inference, the predictions are stitched back to the original 1024 × 1024 size for final change localization.

Hi-UCD mini [5]. The Hi-UCD mini dataset is designed for semantic change detection, covering approximately 30 square kilometers in Tallinn, Estonia. It consists of 745 pairs of remote sensing images acquired in 2017, 2018, and 2019, each sized at 1024 × 1024 pixels with a spatial resolution of 0.1 m. The dataset includes images with partially annotated regions, with labeled areas covering one "no-change" class and nine "change" land cover types: water, grass, building, greenhouse, road, bridge, bareland, woodland, and others. The images are cropped into 512 × 512 patches, resulting in 1200 training pairs, 236 validation pairs, and 1544 testing pairs. Since the "others" category is not a clearly defined semantic category, only the remaining eight classes are considered for evaluation. Similar to LEVIR-CD, the images are further cropped into 256 × 256 patches for the weakly supervised pre-training of S2C, yielding 4800 training pairs, 944 validation pairs, and 6176 testing pairs. During inference, the predictions are stitched back to the 512 × 512 size to generate the final change localization results of S2C.

SECOND [6]. The SECOND dataset is also intended for semantic change detection, covering multiple urban areas in China, including Hangzhou, Chengdu, and Shanghai. Each image pair has a size of 512 × 512 pixels, with a spatial resolution ranging from 0.5 to 3m. The dataset includes one "no-change" class and six "change" land cover types: water, ground, low vegetation, tree, building, and playground. It is divided into 4069 training pairs and 593 testing pairs. As each image pair in the manually curated SECOND dataset inherently contains semantic changes, only the training-free OpenDPR framework is evaluated on this dataset.

2. Additional Implementation Details

2.1. Computation Cost and Scalability of OpenDPR

We record the time cost of the prototype construction process on the SECOND dataset. On a single RTX 3090, constructing 10 prototypes for each of the 6 categories (6×10 in total) takes only 52.8 minutes, with the breakdown of each step summarized in Table 1. As the number of target categories grows, the overall computation cost increases linearly, therefore constructing 10 prototypes for an additional category requires only about 9 minutes. Since this procedure is entirely offline and the resulting prototypes can be reused across downstream scenarios, OpenDPR demonstrates strong practical efficiency and scalability.

Table 1. Time Cost of Prototype Construction on SECOND.

Step	Time (min)
LLM Prompting	1.0
Image Generation	28.3
Target Localization	23.1
Feature Extraction	0.3
Feature Clustering	0.1
Total	52.8

For the inference stage, since both our methods and the baselines rely on similarity-based matching, the overall computation cost remains largely comparable. Moreover, our methods not only achieve better performance but also significantly reduce inference time through efficient simultaneous multi-class inference, as they avoid the need for complex post-processing steps to merge per-class predictions. This highlights the strong practicality and scalability of our methods for OVCD applications.

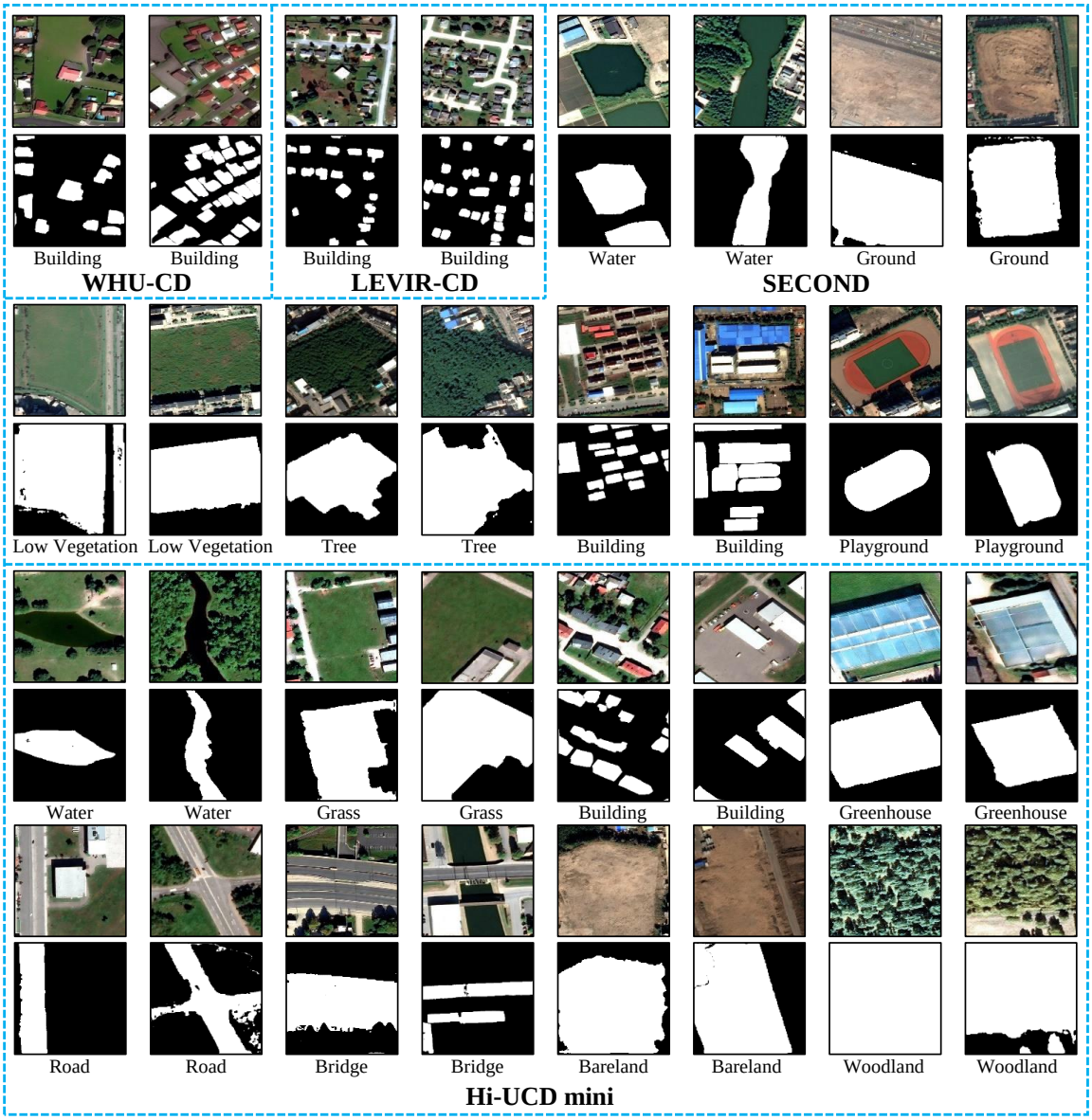

Figure 1. Examples of synthesized support images and their corresponding land-cover masks generated from LLM-guided prompts on the WHU-CD, LEVIR-CD, SECOND, and Hi-UCD mini datasets.

2.2. Prompts of the Extended M-C-I Variants

To enable a comprehensive comparison with the current state-of-the-art M-C-I method, and since that M-C-I was originally designed only for per-class inference, we extend it to support a more practical multi-class inference setting. Specifically, M-C-I (SP) and M-C-I (MP) denote the variants that use a single and multiple textual prompts per category within the open-vocabulary identifier, respectively. The complete prompt sets used for the SECOND and Hi-UCD mini datasets are provided in Table 2.

3. Additional Prototype Analysis

3.1. Analysis of the Diffusion-guided Support Set

Figure 1 shows the support images synthesized by DiffusionSat [3] from LLM-generated prompts, together with the corresponding class-specific land-cover masks extracted by APE [4]. Although these synthesized images still differ in style and distribution from real remote sensing data, they generally capture the semantic characteristics of various land-cover categories.

However, due to the inherent complexity of remote sens-

Table 2. Prompt sets used for the extended M-C-I variants.

Dataset	Single Prompt	Multi Prompts
SECOND	Water	Water, River, Pond, Sea
	Ground	Ground, Bareland
	Low Vegetation	Low Vegetation, Grass
	Tree	Tree, Forest, Woodland
	Building	Building, Roof, House
	Playground	Playground, Sports Field
Hi-UCD mini	Water	Water, River, Pond, Sea
	Grass	Grass, Low Vegetation
	Building	Building, Roof, House
	Greenhouse	Greenhouse, Glasshouse
	Road	Road, Street, Highway
	Bridge	Bridge, Overpass
	Bareland	Bareland, Barren Land
	Woodland	Woodland, Forest, Tree

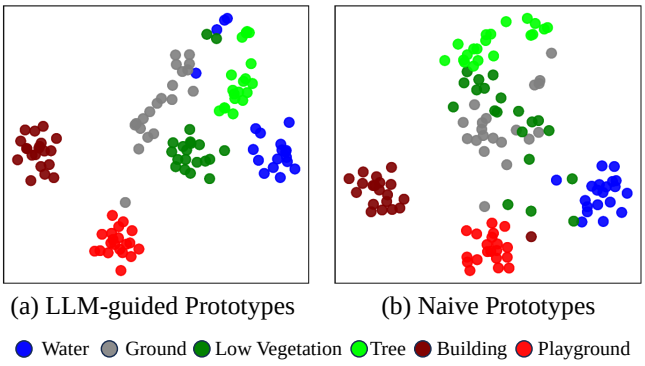

Figure 2. t-SNE visualization of category prototypes constructed from LLM-guided prompts and naive prompts on SECOND.

ing scenes, particularly for rare and complex categories such as pedestrian bridge in the Hi-UCD mini, semantic biases may arise during image generation and target localization. Such biased samples can degrade the quality of the constructed category prototypes, thereby reducing the reliability of retrieval between prototypes and change proposals and the overall OVCD performance. In future work, we plan to fine-tune foundation models for more robust support-set construction.

3.2. Analysis of Prototype Distribution

As shown in Figure 2, we perform t-SNE visualization of the prototype features constructed from LLM-guided and naive prompts on the SECOND dataset. The LLM-derived prototypes demonstrate more compact intra-class distributions and enhanced inter-class separability in the feature space, benefiting from the richer semantic information introduced during support image synthesis. Nevertheless, due to the inherent noise in the support set, partial confusion remains for semantically similar land-cover categories such as ground and low vegetation. Based on the above analysis, we answer the following questions:

(1) *Why does the global-max strategy outperform the category-mean strategy?* This is because the global-max strategy emphasizes the most discriminative prototype feature, while the category-mean strategy may lead to performance degradation due to averaging ambiguous prototypes.

(2) *Why does multi-class inference outperform per-class inference?* This is because multi-class inference builds separate prototypes for each category, yielding cleaner and more discriminative representations. In contrast, per-class inference mixes all non-target categories into a single background class, leading to biased prototypes.

4. Additional Qualitative Results

Figures 3, 4, 5, and 6 present additional qualitative comparisons of the proposed methods across four change detection datasets, offering a more comprehensive view of their effectiveness in diverse scenarios. In addition, Figure 7 shows results on complex in-the-wild scenes, further highlighting the potential of our methods to advance change detection beyond the traditional closed-set paradigm toward more general open-world understanding.

References

[1] 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. 1

[2] 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. 1

[3] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[4] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203, 2024. 2

[5] Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*, 2020. 1

[6] 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. 1

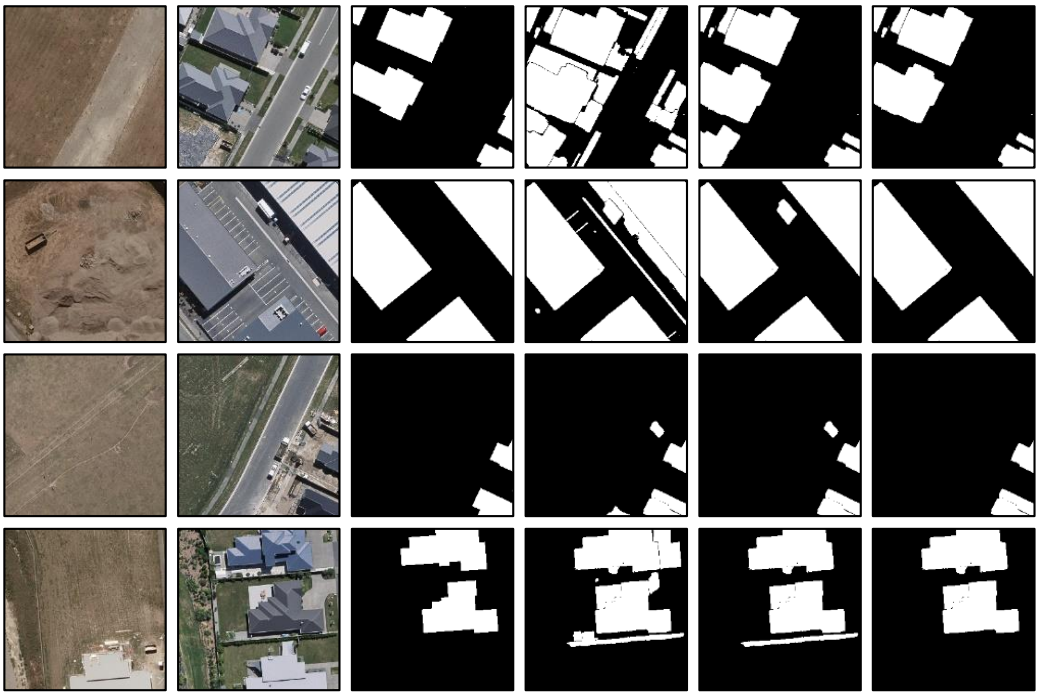

Figure 3. Qualitative comparison of different OVCD methods on the building change detection dataset WHU-CD.

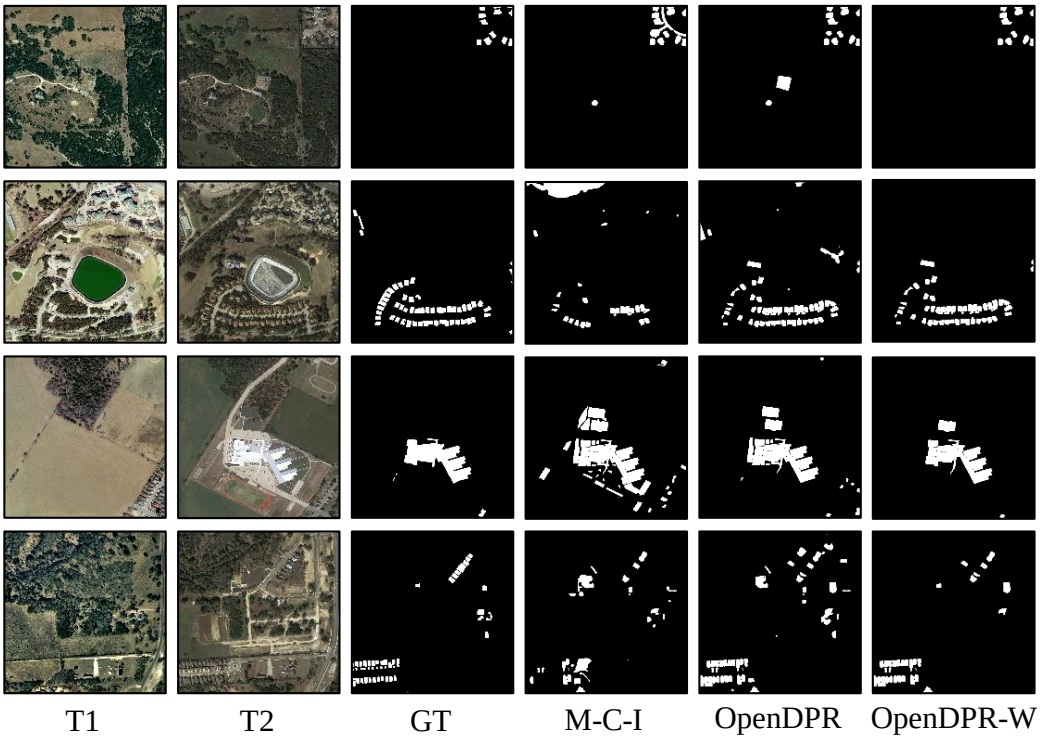

Figure 4. Qualitative comparison of different OVCD methods on the building change detection dataset LEVIR-CD.

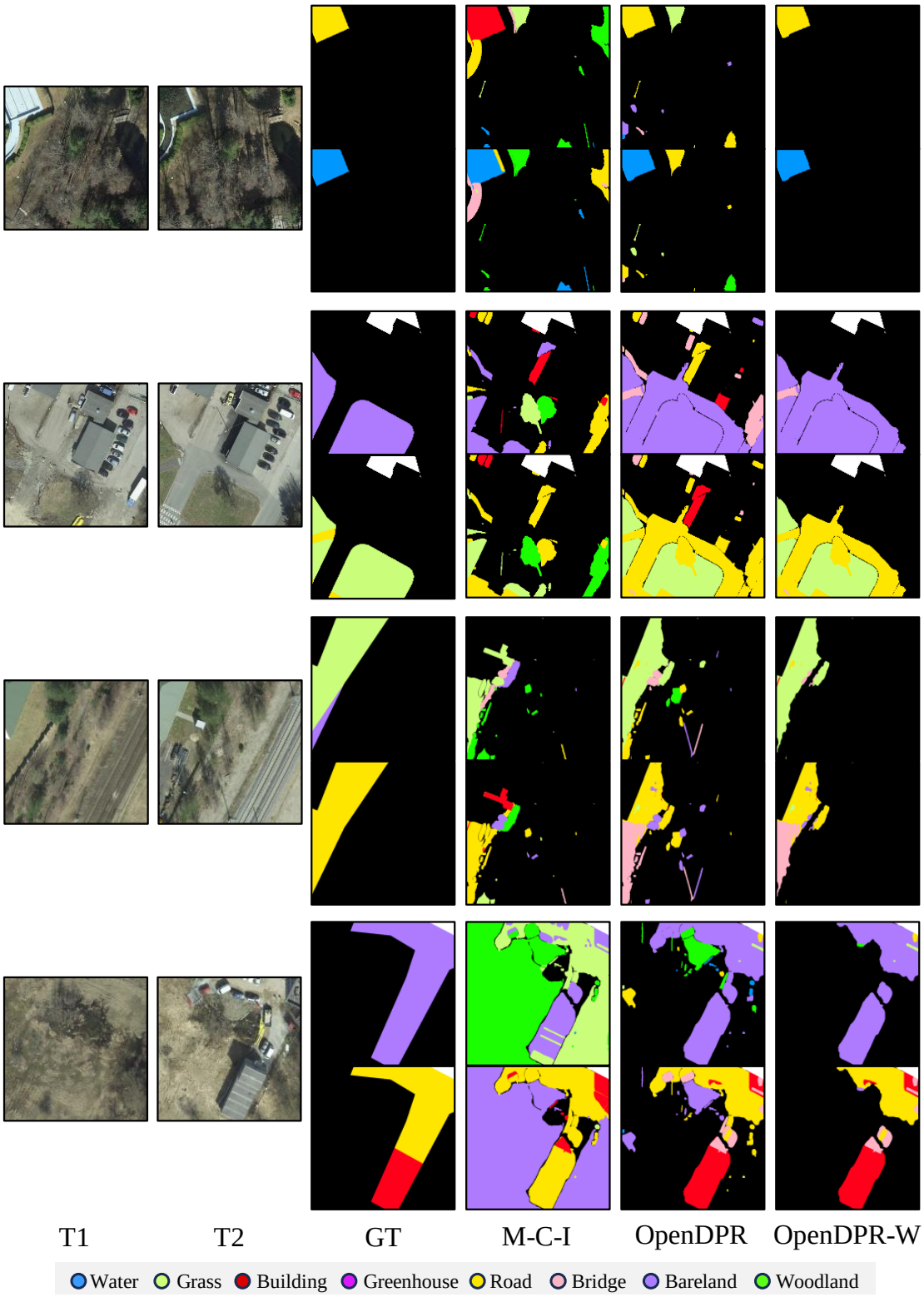

Figure 5. Qualitative comparison of different OVCD methods on the semantic change detection dataset Hi-UCD mini.

Figure 6. Qualitative comparison of different OVCD methods on the semantic change detection dataset SECOND.

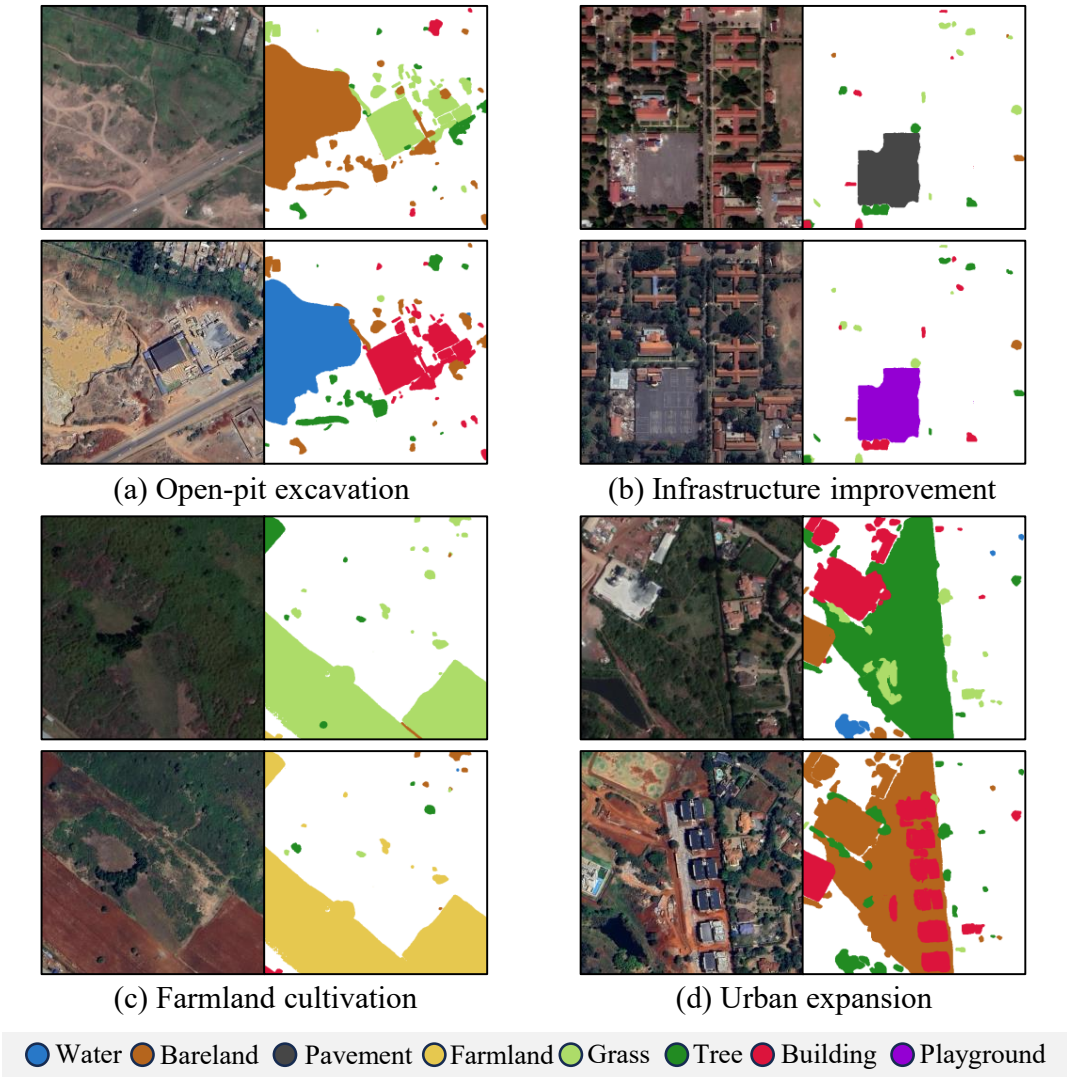

Figure 7. Real-world Case Study in Nairobi, Kenya.