

# Cross-Dataset Generalization of Compact Building Change Detectors: From LEVIR-CD to SpaceNet 7

**Bowen Zheng\***

*College of Science and Technology, Ningbo University, Ningbo, China*

*\*Corresponding author: Bowen Zheng.*

---

## Abstract

Within-dataset scores do not establish performance after sensor, geography, and label definitions change. We compare Change Vector Analysis (CVA) with three compact learned controls under a common LEVIR-CD training regime, then freeze checkpoints, source RGB moments, and validation-selected thresholds before external evaluation. The external set contains 13 SpaceNet 7 areas of interest (AOIs), each with fixed 10- and 22-month pairs, for 26 pairs in total. On LEVIR-CD, the learned controls achieved F1 scores of 0.6541-0.7425, compared with 0.1398 for CVA. The ranking reversed on SpaceNet 7: CVA reached 0.1740 macro-F1, whereas the learned controls reached 0.0334-0.0650. Their pooled AUPRC and target-oracle F1 also remained low, so threshold recalibration alone could not recover the source advantage. An eightfold scale-degradation proxy reduced learned LEVIR-CD F1 by 0.6229-0.7197. A label-free target-statistics radiometric control improved all three learned controls, but it did not belong to the strict source-only result. These findings define a protocol-specific generalization boundary for the evaluated compact controls and support external validation before building-change maps are used as auxiliary evidence for urban-stock monitoring.

## Keywords

building change detection, cross-dataset generalization, LEVIR-CD, SpaceNet 7, source-only transfer

---

## 1. Introduction

Paired overhead images allow building change detectors to identify footprint additions, removals, and shape changes [1–3]. When resolution, radiometry, viewing geometry, prevalence, and annotation rules shift together, within-dataset accuracy becomes an incomplete test.

LEVIR-CD provides high-resolution bitemporal pairs with binary building-change labels [2]; SpaceNet 7 provides monthly, approximately 4-m imagery and tracked footprints across many geographies [4]. Its official task is SpaceNet Change and Object Tracking, not generic two-date pixel classification. Side-looking data and registration studies show that geometry and residual displacement alter measured change [5, 6]. Transfer evidence must therefore state both frozen source decisions and the derived target protocol.

We conduct an empirical cross-dataset diagnostic study of whether compact learned detectors retain their source advantage without target-label adaptation. The controls are CVA, FC-EF-Compact, FC-Siam-Diff-Compact, and Compact Token-Fusion. They are intentionally small, from-scratch implementations rather than

official reproductions. The goal is to test whether source-domain ranking survives a documented domain shift under one shared training budget, not to benchmark state-of-the-art architectures.

The paper's main contribution is a source-only, AOI-level LEVIR-CD-to-SpaceNet 7 protocol in which checkpoints, source RGB moments, and decision thresholds are fixed before external evaluation. Macro-F1, pooled AUPRC, and target-oracle F1 then separate operating-point failure from weak probability ranking. Eightfold scale degradation and label-free target-statistics radiometric normalization probe two contributors to the observed gap. All conclusions remain limited to the evaluated compact controls and the registered 13-AOI protocol.

## 2. Related Work

Early fusion and Siamese encoding are common bitemporal detector designs [1]. Later variants use dense skips, attention, or contrastive objectives [7-9]. BIT introduces token interaction, while ChangeFormer uses hierarchical transformers [3, 10]. Our reduced implementations borrow these design families but are labeled as compact controls to avoid implying reproduction of the official systems.

LEVIR-CD combines high-resolution bitemporal pairs with building-change masks [2]. S2Looking emphasizes off-nadir and illumination variation [5], while SpaceNet 7 provides monthly imagery, persistent building identifiers, and SCOT evaluation [4]. Our pairwise footprint-raster XOR task differs from SCOT and is used only to examine cross-dataset pixel classification. Misregistration is a known source of false change near object boundaries [6].

Target-assisted methods such as SDACD use target-domain supervision or adaptation [11]. They address a different question because no SpaceNet 7 label changes our fitted models, source moments, or thresholds. Controlled studies also warn that architectural complexity can obscure the value of simple baselines [12]. We therefore ask whether a source ranking survives transfer, rather than whether one compact architecture wins within LEVIR-CD.

Cross-dataset evaluation addresses a different question from within-dataset model comparison. A detector can preserve high source accuracy while its score ranking, calibration, or spatial response changes after sensor and annotation conditions change. Source-only evaluation isolates this problem by prohibiting target-label tuning, whereas supervised domain adaptation asks how much target information is needed to recover accuracy [11]. These settings should remain separate because even a small target-labeled validation set changes the meaning of the external score. We therefore treat SpaceNet 7 as a frozen external test and identify every target-dependent analysis as a diagnostic rather than a deployment result.

A second distinction concerns the statistical unit. Pixel pooling lets large or change-rich scenes dominate and does not reflect the dependence of nearby pixels. Pair-level averaging also overstates sample size when two intervals share the same geography and first-date image. We instead pool both temporal pairs within each AOI and average across AOIs. This choice does not remove spatial dependence, but it aligns the uncertainty analysis with geographic units and exposes whether an aggregate change is broadly distributed or driven by a few areas.

Prior syntheses distinguish pixel-difference detection from object-aware analysis and identify sensor geometry, radiometry, registration, and label definition as persistent comparison risks [13], [14]. These works motivate separating a frozen transfer protocol from a claim of general deployment performance.

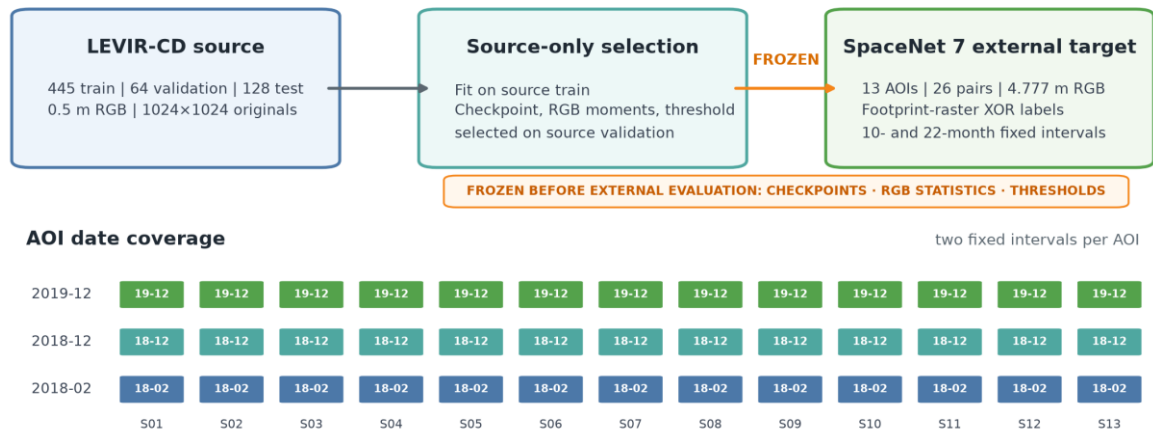
## 3. Comparative Study Design

### 3.1 Source Data and Compact Controls

LEVIR-CD follows its public 445/64/128 train, validation, and test partition [2]. Each complete 1,024 x 1,024 image and mask was resized to 256 x 256 using bilinear and nearest-neighbor interpolation, respectively. RGB channel moments were computed from the training split only. For seeds 0, 1, and 2, each learned control used AdamW with a learning rate of 1e-3, weight decay of 1e-4, batch size 16, and at most 15 epochs. Training stopped after five validation epochs without improvement. Horizontal and vertical flips were applied jointly to each image pair and mask. The loss was positive-weighted binary cross-entropy plus Dice loss.

CVA thresholds the Euclidean change magnitude between the two RGB images. FC-EF-Compact concatenates the dates before a four-stage width-16 encoder and skip-connected decoder (483,169 trainable parameters). FC-Siam-Diff-Compact applies one shared encoder and decodes absolute multiscale feature differences (482,737 parameters). Compact Token-Fusion adds four-head self-attention over a 16 x 16 pooled token map at the deepest shared feature level before decoding the difference pyramid (1,024,817 parameters). For each method and seed, the checkpoint and probability threshold were selected by validation F1 over thresholds from 0.05 to 0.95 in steps of 0.05.

Figure 1: Source-only evaluation protocol. LEVIR-CD supplies 445/64/128 train, validation, and test pairs. Checkpoints, RGB moments, and thresholds are frozen before 13 SpaceNet 7 AOIs are evaluated over fixed 10- and 22-month intervals



### 3.2 Source-Only SpaceNet 7 Protocol

The external set contains the first 13 lexicographically ordered AOIs in the retained public SpaceNet 7 collection. Each AOI contributes the fixed pairs February 2018 to December 2018 and February 2018 to December 2019, giving 26 intervals of 10 and 22 months. The RGB mosaics retain their native 1,023/1,024-pixel extent and approximately 4.777-m spacing. For each date pair, the target mask is the logical XOR of the two building-footprint rasters. Pixels are valid only when both RGB dates contain at least one nonzero channel.

Figure 1 summarizes the fixed acquisition dates. S01-S13 follow the manifest order and map to the full AOI identifiers in the released protocol file. That manifest records the image and label paths for all 26 pairs, while the result files retain the pair dates, valid-area fraction, change prevalence, thresholds, and confusion counts. Across the 26 pairs, change prevalence ranged from 0.00034 to 0.01209 of valid pixels, which confirms severe and geographically variable class imbalance.

All target AOIs remain external to fitting and selection. Learned checkpoints, one set of source training moments, and the per-seed validation thresholds were fixed before SpaceNet 7 inference. Target labels enter only final metrics and the separately identified target-oracle F1 diagnostic. The XOR task is neither the official SpaceNet 7 SCOT task nor an object-tracking evaluation.

Pair selection was fixed before model scoring. The same two calendar intervals were used for every AOI, avoiding manual selection of visibly favorable change episodes. This gives two temporal separations per geography while retaining AOI as the independent statistical unit. The complete manifest and file checksums preserve the selected data independently of the model outputs.

The XOR labels require a narrow interpretation. Minor footprint edits can create positive boundary pixels without a substantive building event, while mosaicking or residual registration error can move edges without changing building stock. Conversely, roof replacement or internal renovation may be absent when the mapped footprint is unchanged. The derived masks therefore indicate possible footprint change and do not identify construction causes, permits, occupancy, or regeneration outcomes.

### 3.3 Metrics and Domain Diagnostics

For each seed, confusion counts from the two pairs were pooled within each AOI and converted to precision, recall, F1, and intersection over union. Macro-F1 is the mean of the 13 AOI-level F1 scores; its 95% interval uses 5,000 AOI bootstrap resamples. AUPRC and target-oracle F1 are calculated after pooling valid target pixels and are then averaged across learned-model seeds. Target-oracle F1 chooses a threshold on target labels only to measure threshold recoverability. It is not part of the source-only performance claim.

Two controls probe components of the domain gap without retraining. The eightfold scale-degradation proxy downsamples LEVIR-CD images by a factor of eight and restores them bilinearly before inference, while labels and source decisions remain fixed. The factor approximates the order of the 0.5-m to 4.777-m ground-sampling-distance difference but is not a sensor simulation. The radiometric control matches each target image's valid-pixel channel mean and standard deviation to the source moments without using target labels.

The metrics answer separate questions. Macro-F1 evaluates each frozen operating point while giving every AOI equal weight. Pooled AUPRC tests whether changed target pixels tend to receive larger scores over all thresholds. Target-oracle F1 estimates how much performance is recoverable by changing only the target threshold. A low oracle value therefore indicates that calibration is not the only failure, although it does not identify which representation component failed.

Native and radiometrically normalized predictions are paired within the same AOI after averaging learned-model seeds. This comparison measures whether the control changes performance consistently across geography without treating seeds as independent AOIs. Because target image statistics are used at test time, these radiometric values are reported only in Table 2 as a label-free diagnostic and are excluded from the strict result in Table 1.

## 4. Results

### 4.1 Controlled Performance Does Not Transfer

Table 1 reports the controlled source benchmark and strict external evaluation. On LEVIR-CD, FC-EF-Compact reached  $0.7425 \pm 0.0150$  F1 and Compact Token-Fusion reached  $0.7399 \pm 0.0138$ . FC-Siam-Diff-Compact was lower at  $0.6541 \pm 0.0161$ , but all three learned controls exceeded CVA at 0.1398.

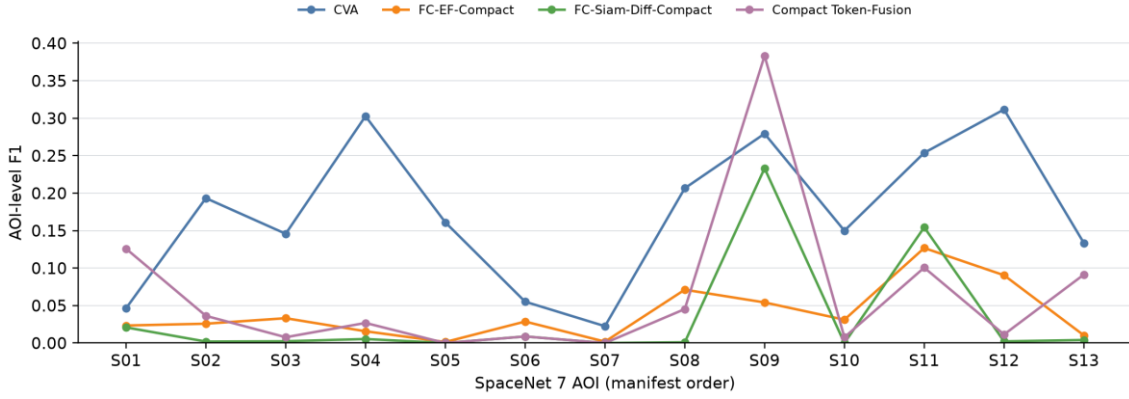
*Table 1: LEVIR-CD and strict source-only SpaceNet 7 results. Learned values are mean  $\pm$  SD over three seeds. The macro-F1 interval uses 5,000 AOI bootstrap resamples; AUPRC and target-oracle F1 use pooled valid target pixels*

| Method               | Params    | LEVIR-CD F1         | SN7 macro-F1 (95% CI)                | SN7 AUPRC           | Target-oracle F1    |
|----------------------|-----------|---------------------|--------------------------------------|---------------------|---------------------|
| CVA                  | 0         | 0.1398              | 0.1740 [0.1234, 0.2227]              | 0.1043              | 0.1996              |
| FC-EF-Compact        | 483,169   | $0.7425 \pm 0.0150$ | $0.0395 \pm 0.0133$ [0.0220, 0.0597] | $0.0144 \pm 0.0085$ | $0.0394 \pm 0.0125$ |
| FC-Siam-Diff-Compact | 482,737   | $0.6541 \pm 0.0161$ | $0.0334 \pm 0.0211$ [0.0029, 0.0766] | $0.0529 \pm 0.0112$ | $0.1370 \pm 0.0089$ |
| Compact Token-Fusion | 1,024,817 | $0.7399 \pm 0.0138$ | $0.0650 \pm 0.0445$ [0.0218, 0.1277] | $0.0579 \pm 0.0238$ | $0.1386 \pm 0.0349$ |

The external ranking reversed. CVA reached 0.1740 macro-F1 (95% CI, 0.1234-0.2227), while the learned controls ranged from 0.0334 to 0.0650. CVA also had the highest pooled AUPRC at 0.1043, compared with 0.0144-0.0579 for the learned controls. Target-oracle F1 was 0.1996 for CVA and 0.0394-0.1386 for the learned controls. Since the learned oracle values remained low, changing only the decision threshold did not restore their source-domain advantage.

The reversal was geographically broad rather than driven by one AOI. After averaging the three learned seeds, CVA had the highest AOI-level F1 in 11 of 13 areas, and Compact Token-Fusion led in the remaining two. Absolute scores still varied substantially across AOIs, as shown in Figure 2. This spread supports AOI-level reporting because a single pooled pixel score would obscure how strongly a few geographies influence the aggregate.

Figure 2: Source-frozen AOI-level F1 on SpaceNet 7. Two fixed pairs are pooled within each AOI; learned values average three seeds. S01-S13 follow the manifest order



## 4.2 Scale and Radiometry Explain Part of the Gap

Table 2 separates two domain controls from the strict main result. Eightfold degradation reduced learned LEVIR-CD F1 by 0.6229-0.7197, whereas CVA decreased by 0.0338. The change was largest for Compact Token-Fusion, whose F1 fell from 0.7399 to 0.0202. These results show strong scale sensitivity in the evaluated learned controls, but the proxy does not reproduce SpaceNet 7 optics, mosaicking, or annotation rules.

Table 2: Scale and label-free radiometric diagnostics. Eightfold degradation is a scale proxy. Target-statistics normalization uses no target labels and is excluded from the strict source-only scores

| Method               | LEVIR native<br>F1 | 8x scale<br>F1 | Scale<br>delta | SN7 strict<br>F1 | Target-stat.<br>F1 | Target-stat.<br>delta |
|----------------------|--------------------|----------------|----------------|------------------|--------------------|-----------------------|
| CVA                  | 0.1398             | 0.1059         | -0.0338        | 0.1740           | 0.1659             | -0.0081               |
| FC-EF-Compact        | 0.7425             | 0.0634         | -0.6791        | 0.0395           | 0.1622             | +0.1227               |
| FC-Siam-Diff-Compact | 0.6541             | 0.0312         | -0.6229        | 0.0334           | 0.1262             | +0.0927               |
| Compact Token-Fusion | 0.7399             | 0.0202         | -0.7197        | 0.0650           | 0.1309             | +0.0659               |

Target-statistics radiometric normalization increased FC-EF-Compact macro-F1 from 0.0395 to 0.1622, FC-Siam-Diff-Compact from 0.0334 to 0.1262, and Compact Token-Fusion from 0.0650 to 0.1309. The corresponding gains were 0.1227, 0.0927, and 0.0659. CVA changed from 0.1740 to 0.1659. Thus, the simple control recovered part of the learned-model deficit but did not replace the strict source-only comparison.

The direction of the radiometric effect was also visible across AOIs. FC-EF-Compact improved in all 13 areas, FC-Siam-Diff-Compact in 12, and Compact Token-Fusion in 10. These paired counts make radiometric sensitivity plausible, but they do not establish radiometry as the sole cause. The control uses target-image statistics at test time, so its values should be read as a diagnostic upper comparison rather than deployable source-only performance.

## 5. Discussion

The source-domain ranking did not survive the registered external shift. Low frozen-threshold F1 could have been a calibration problem, but the learned controls also had low pooled AUPRC and target-oracle F1. Their external failure therefore involved weak target-domain score ranking as well as operating-point mismatch. This conclusion is protocol-specific: it applies to the evaluated compact, from-scratch controls and does not show that all modern building-change detectors fail to generalize.

Scale and radiometric sensitivity explain different parts of the gap. The eightfold control removed most of the learned LEVIR-CD F1, which is consistent with a loss of small-building and edge information after coarse resampling. Target-statistics normalization recovered sizeable external F1 for every learned control, which is consistent with a radiometric mismatch in the frozen source moments. Neither result isolates causality because resolution, radiometry, mosaicking, geometry, prevalence, and label construction change together between datasets.

The external error pattern also cautions against relying on one aggregate score. CVA was more stable across most AOIs, yet Compact Token-Fusion exceeded it in two areas and varied strongly between others. Such variation may arise from local roof scale, background texture, registration quality, or the prevalence of footprint edits. The present experiment records AOI-level scores and pair statistics, but it cannot attribute each error to one of these factors without additional controlled labels.

Several limitations bound the inference. The compact controls are not official or state-of-the-art reproductions, and from-scratch training may understate the transfer behavior of pretrained systems. Thirteen AOIs provide a useful conference-scale external test but do not represent all SpaceNet 7 cities. Footprint XOR combines additions, removals, boundary edits, and annotation differences, while the scale experiment is only an order-of-magnitude proxy. No field survey or planning record validates the causes of mapped changes.

For urban-stock monitoring, the implication is procedural. A high score on a single-city benchmark should not be treated as evidence that the same model will screen another sensor or geography reliably. A deployment study should first freeze its source decisions, report target AOIs separately, and inspect calibration and scale sensitivity. Even then, the output is remote-sensing auxiliary evidence for possible footprint change, not proof of demolition, extension, renovation, permits, occupancy, or policy effectiveness.

A stronger follow-up can retain the same leakage controls while adding an official or near-official detector, more AOIs, pretrained weights, and explicit registration perturbations. Object-aware evaluation could separate missed buildings from boundary-only pixel errors, while event labels could distinguish construction and demolition from annotation revision. These extensions would test whether the observed boundary persists beyond the present small-budget regime rather than assuming that it does.

## 6. Conclusion

This study provides an AOI-level, source-only protocol for testing whether building-change rankings survive transfer from LEVIR-CD to derived SpaceNet 7 pairs. Under the evaluated compact, from-scratch regime, learned controls reached 0.6541-0.7425 F1 on LEVIR-CD but only 0.0334-0.0650 macro-F1 externally, below CVA at 0.1740. Low AUPRC and target-oracle F1 showed that threshold recalibration alone was insufficient, while scale degradation and label-free radiometric normalization exposed two contributors to the gap. The result establishes a protocol-specific generalization boundary and supports external validation before footprint-change maps are used as auxiliary evidence for urban-stock monitoring.

## References

- [1] Daudt, R. C., Saux, B. L., and Boulch, A., *Fully convolutional siamese networks for change detection*, 2018 *IEEE international conference on image processing (ICIP)*, 4063–4067, 2018, <https://doi.org/10.1109/ICIP.2018.8451652>.
- [2] Chen, H., and Shi, Z., “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sensing*, 12, (10), 1662, 2020, <https://doi.org/10.3390/rs12101662>.
- [3] Chen, H., Qi, Z., and Shi, Z., “Remote sensing image change detection with transformers,” *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14, 2022, <https://doi.org/10.1109/TGRS.2021.3095166>.
- [4] Etten, A. V., Hogan, D., Martinez-Manso, J., Shermeyer, J., Weir, N., and Lewis, R., *The multi-temporal urban development SpaceNet dataset*, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6398–6407, 2021, <https://doi.org/10.1109/CVPR46437.2021.00633>.
- [5] Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., and Jiang, B., *S2Looking: A satellite side-looking dataset for building change detection*, 2021, <https://arxiv.org/abs/2107.09244v3>.
- [6] Townshend, J. R. G., Justice, C. O., Gurney, C., and McManus, J., “The impact of misregistration on change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, 30, (5), 1054–1060, 1992, <https://doi.org/10.1109/36.175340>.

- [7] Fang, S., Li, K., Shao, J., and Li, Z., “SNUNet-CD: A densely connected siamese network for change detection of VHR images,” *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5, 2022, <https://doi.org/10.1109/LGRS.2021.3056416>.
- [8] Chen, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y., and Li, H., “DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, <https://doi.org/10.1109/JSTARS.2020.3037893>.
- [9] Wang, Z., Peng, C., Zhang, Y., Wang, N., and Luo, L., “Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss,” *Neurocomputing*, 2021, <https://doi.org/10.1016/j.neucom.2021.06.059>.
- [10] Bandara, W. G. C., and Patel, V. M., *A transformer-based siamese network for change detection*, 2022 *IEEE international geoscience and remote sensing symposium*, 207–210, 2022, <https://doi.org/10.1109/IGARSS46834.2022.9883686>.
- [11] Liu, J., Xuan, W., Gan, Y., Zhan, Y., Liu, J., and Du, B., “An end-to-end supervised domain adaptation framework for cross-domain change detection,” *Pattern Recognition*, 132, 108960, 2022, <https://doi.org/10.1016/j.patcog.2022.108960>.
- [12] Corley, I., Robinson, C., and Ortiz, A., *A change detection reality check*, *ICLR workshop on machine learning for remote sensing*, 2024, <https://arxiv.org/abs/2402.06994>.
- [13] Radke, R. J., Andra, S., Al-Kofahi, O., and Roysam, B., “Image change detection algorithms: A systematic survey,” *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005, doi: 10.1109/TIP.2004.838698.
- [14] Tewkesbury, A. P., Comber, A. J., Tate, N. J., Lamb, A., and Fisher, P. F., “A critical synthesis of remotely sensed optical image change detection techniques,” *Remote Sensing of Environment*, vol. 160, pp. 1–14, 2015, doi: 10.1016/j.rse.2015.01.006.

## Funding

This research received no external funding.

## Conflicts of Interest

The authors declare no conflict of interest.

## Acknowledgment

This paper is an output of the science project.

## Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

