BSSRS: Bridging Semantic and Structural Manifolds for Zero-Shot Single-Temporal Remote Sensing Anomaly Detection

Shih-Chih Lin [1] **Jia-Xian Jian** [2] **YunTung Chu** [3] **Wei-Chieh Sun** [3] **Fang-Yi Lin** [2]

[1]*National Tsing Hua University, Taiwan* [2]*National Cheng Kung University, Taiwan* [3]*University of Washington, Seattle, USA*. Correspondence to: Shih-Chih Lin leolin65@gapp.nthu.edu.tw.

Abstract

Traditional remote sensing change detection relies on bi-temporal image pairs, which are often unavailable or unreliable in time-critical scenarios such as post-disaster assessment. We propose **BSSRS**, a zero-shot single-temporal framework that reformulates building change detection as structural anomaly detection from a single post-event image. BSSRS preserves CLIP as an invariant semantic reference and injects frozen DINOv3 structural evidence through a constrained residual blending strategy in the dense pixel pathway, while image-level anomaly scoring is kept fully decoupled in an isolated vision–language branch. This design explicitly mitigates structural over-specialization under cross-domain transfer. Under zero-shot cross-dataset evaluation, training on LEVIR-CD post-event images only and testing directly on WHU, BSSRS achieves **95.35%** Pixel AUC, **59.58%** Pixel F1, and **98.60%** Image AUC without target-domain fine-tuning.

1. Introduction

Remote sensing change detection is typically formulated using registered bi-temporal image pairs. In practice, however, pre-event imagery may be unavailable, misaligned, or too costly to acquire. This limitation is especially problematic in emergency response and rapid urban monitoring, where only a single post-event image may be available. We therefore study zero-shot *single-temporal* anomaly detection, treating newly constructed buildings as structural deviations from surrounding natural backgrounds.

A key challenge under cross-domain deployment is *structural over-specialization*: trainable modules can overfit to source-domain geometric statistics and fail on unseen target distributions. We address this with **BSSRS**, a dual-vision framework that bridges semantic and structural manifolds while keeping their roles decoupled. DINOv3 contributes local geometric cues for dense localization, whereas CLIP remains the dominant semantic reference and exclusively governs global image-level scoring.

2. Experiments

We evaluate BSSRS under zero-shot cross-dataset transfer. Training uses *only* post-event imagery from LEVIR-CD, while testing is performed directly on WHU without fine-tuning. This setup removes any reliance on pre-event images at inference time and provides a strict test of cross-domain robustness.

Table 1: Zero-shot cross-dataset performance of BSSRS.

Source → Target	P-AUC	P-F1	P-IoU	I-AUC
LEVIR-CD → WHU	**95.35**	**59.58**	**42.43**	**98.60**
WHU → LEVIR-CD	90.41	29.31	17.18	92.74

3. Method

Let f_v^{clip} and f_t denote the frozen CLIP vision and text encoders, and let f_v^{dino} denote a frozen DINOv3 encoder. Given a single remote sensing image x, CLIP extracts semantic patch tokens $Z_c^{(l)}$ and DINOv3 extracts structural patch tokens $Z_d^{(l)}$ from selected layers $\mathcal{L} = \{6, 12, 18, 24\}$. We construct adapted normal and abnormal text anchors

$$t_n = \psi_t(f_t(p_n)), \qquad t_a = \psi_t(f_t(p_a)), \tag{1}$$

where $\psi_t(\cdot)$ is a lightweight text adaptation module.

To preserve CLIP as the dominant manifold, DINOv3 tokens are projected, spatially aligned to the CLIP grid, and injected via constrained residual blending:

$$Z_{\text{fuse}}^{(l)} = \text{Norm}\big(\hat{Z}_c^{(l)} + \alpha \tilde{Z}_d^{(l)}\big), \qquad \alpha = 0.05. \tag{2}$$

The fused tokens are matched with $[t_n, t_a]$ to produce dense anomaly maps, which are averaged across layers for pixel-level localization.

For image-level scoring, we deliberately avoid using the blended tokens. Instead, a standalone CLIP detection token yields the global anomaly score:

$$\ell = z_{\text{det}}^{\top}[t_n, t_a], \qquad s_{\text{img}} = \text{Softmax}(\ell)_a. \tag{3}$$

This decoupled design prevents local structural noise from corrupting global semantic decisions.

4. Results and Discussion

BSSRS achieves strong zero-shot transfer from LEVIR-CD to WHU, reaching 95.35% Pixel AUC and 59.58% Pixel F1, while maintaining highly stable image-level performance with 98.60% Image AUC. The reverse transfer is noticeably harder, reflecting morphological differences between the two datasets and validating the importance of structural robustness. Qualitatively, BSSRS produces compact heatmaps aligned with building footprints while avoiding diffuse false positives. These results support the central claim that constrained structural injection improves dense localization, whereas isolated CLIP scoring preserves global semantic stability under domain shift.

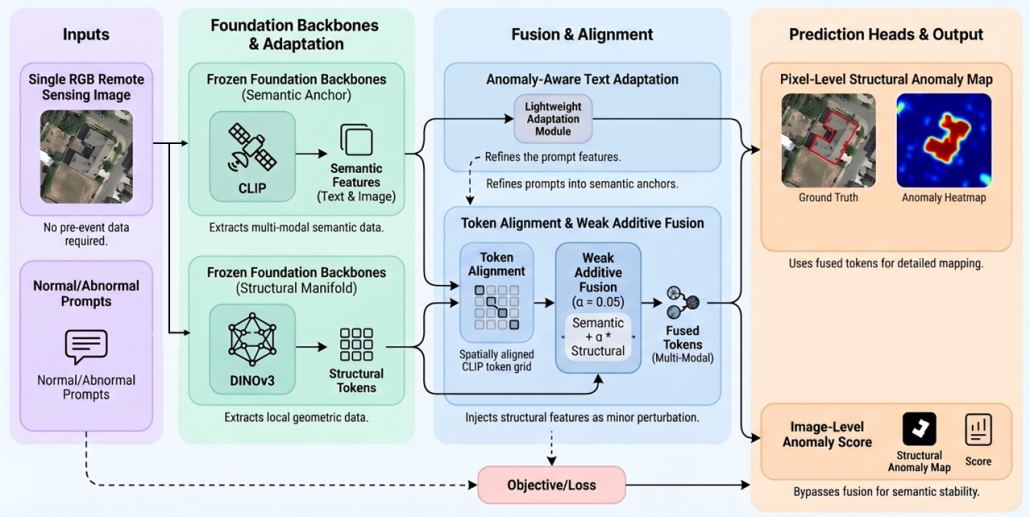

Fig. 1: Overview of BSSRS. A frozen CLIP branch provides invariant semantic context, while frozen DINOv3 features are spatially aligned and injected into the pixel-level pathway through constrained residual blending. Image-level scoring is fully decoupled and derived from an isolated CLIP vision–language branch.

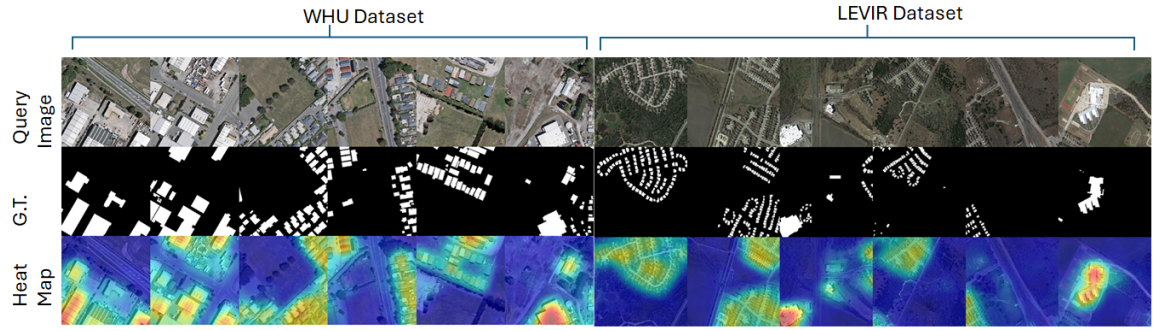

Fig. 2: **Qualitative visualization of BSSRS.** From top to bottom: input image, ground-truth mask, and predicted anomaly heatmap. Results on LEVIR-CD and WHU under zero-shot cross-dataset transfer show that BSSRS produces spatially compact responses aligned with building footprints. By bridging semantic and structural manifolds through constrained residual blending, the model improves boundary sensitivity while preserving global semantic stability.

5. Conclusion

BSSRS provides a practical route to single-temporal remote sensing anomaly detection when pre-event imagery is unavailable. Its main principle is controlled structural injection for dense localization together with strict decoupling of image-level reasoning from fused spatial features. This makes the method suitable for real-world post-event and cross-domain deployment settings where data pairing and adaptation budgets are limited.

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.

[2] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multi-source building extraction from an open aerial and satellite imagery dataset. *IEEE TGRS*, 57(1):574–586, 2018.

[3] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision-language foundation model for remote sensing. *IEEE TGRS*, 62:1–16, 2024.

[4] 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 *ICML*, 2021.

[5] Oriane Simeoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, et al. DINOv3. arXiv preprint arXiv:2508.10104, 2025.

[6] Yu Zhang and Zhi Gao. RSAD-CLIP: Zero-shot remote sensing anomaly detection of the earth's surface based on pre-trained vision-language model. In *IEEE*, 2025.