Bridging Semantic and Structural Manifolds: Zero-Shot Single-Temporal Change Detection in Remote Sensing

Shih-Chih Lin
National Tsing Hua University
Hsinchu, Taiwan
leolin65@gapp.nthu.edu.tw

Jia-Xian Jian
National Cheng Kung University
Tainan, Taiwan
n28101173@gs.ncku.edu.tw

YunTung Chu
University of Washington
Seattle, USA
yuntungc@uw.edu

Wei-Chieh Sun
University of Washington
Seattle, USA
wsun12@uw.edu

Fang-Yi Lin
National Cheng Kung University
Tainan, Taiwan
n28111518@gs.ncku.edu.tw

Abstract

Traditional remote sensing change detection paradigms typically rely on bi-temporal image pairs to identify surface variations. However, in time-critical scenarios such as post-disaster assessment, pre-event imagery may be unavailable or subject to severe registration errors, necessitating zero-shot single-temporal inference. A critical bottleneck in such cross-domain transfer is structural over-specialization, where models overfit to source-domain geometric statistics, leading to degraded generalization on unseen target distributions. To address this limitation, we propose BSSRS, a zero-shot framework that reformulates building change detection as a single-temporal structural change localization problem. By integrating the complementary inductive biases of two frozen foundation encoders—CLIP and DINOv3—BSSRS explicitly counteracts this domain shift. While DINOv3 emphasizes local geometric consistency for dense localization, trainable modules can easily over-specialize to these source-domain features. To prevent domain collapse, the frozen CLIP backbone acts as an invariant contextual reference. Instead of complex token-mixing, patch tokens from multiple selected transformer layers are integrated via a constrained residual blending strategy. Specifically, spatially aligned DINOv3 tokens are injected into the primary CLIP embedding space as a soft geometric stimulus, enabling robust dense change localization without full backbone adaptation. Furthermore, to avoid disturbing the original global semantics, image-level change scores are derived directly from an isolated vision-language pathway, deliberately omitting the blended patch tokens. Extensive experiments reveal a clear training divergence that validates our decoupled design: while pixel-level metrics may eventually degrade due to structural overfitting, the image-level semantic stability preserved by the standalone CLIP classifier remains stable. Evaluated in a zero-shot setting—trained on LEVIR-CD using only post-event imagery and tested on the WHU Building Dataset—BSSRS achieves 95.38% Pixel AUC and a 63.99% F1-score. These results indicate that our dual-manifold synthesis effectively mitigates structural over-specialization, providing a highly robust solution for single-temporal remote sensing analysis. The code is available at https://github.com/leolin65/BSSRS.git.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Neural networks**.

Keywords

remote sensing, zero-shot change detection, vision-language models, CLIP, DINOv3

1 Introduction

The unprecedented pace of global urban expansion has intensified the demand for efficient Earth surface monitoring. In rapidly developing regions, timely identification of building changes is critical for sustainable urban planning and proactive resource allocation. Remote sensing (RS), characterized by large-scale coverage and high revisit frequency, has become a cornerstone technology for urban change analysis [9].

Despite substantial progress, conventional building change detection (CD) methods remain heavily dependent on precisely registered bi-temporal image pairs and dense pixel-level annotations. In practice, acquiring well-aligned pre-event imagery is often costly or infeasible, particularly in time-critical scenarios such as disaster response or rapid urban expansion [15]. Moreover, supervised CD models frequently suffer severe performance degradation under cross-domain deployment due to variations in sensor characteristics, illumination conditions, and architectural styles [6]. These limitations hinder the scalability and real-world applicability of traditional pipelines.

To overcome these challenges, we reformulate building change detection as a zero-shot, single-temporal structural change localization problem. Rather than explicitly modeling temporal differences, we identify visually prominent built structures as single-temporal change candidates from a single post-event image. This formulation eliminates the reliance on paired temporal inputs during inference and enables deployment in scenarios where only post-event imagery is available.

Recent vision-language foundation models such as CLIP [13] offer powerful open-vocabulary representations. However, transferring these models to overhead remote sensing imagery presents a major challenge. While CLIP captures broad conceptual alignment, it is dominated by high-level contextual cues and lacks sensitivity to fine-grained geometric variations critical for structural reasoning in aerial scenes. Conversely, structure-sensitive self-supervised models (e.g., DINOv3 [16]) excel at capturing local spatial coherence, but adapting them to unseen domains inevitably leads to a critical bottleneck: structural over-specialization. This phenomenon

occurs when trainable modules overfit to source-domain geometric statistics, leading to degraded performance under distribution shifts. Bridging this domain gap requires complementary modeling beyond prompt engineering or lightweight semantic calibration alone.

In this work, we propose BSSRS (Bridging Semantic and Structural manifolds in Remote Sensing), a dual-vision framework designed to explicitly counteract this bottleneck. To preserve the respective inductive biases of both foundation models, BSSRS maintains frozen CLIP and DINOv3 backbones. Rather than employing complex token-mixing modules that risk severe overfitting, we integrate multi-depth patch representations through a constrained residual blending strategy. Specifically, spatially aligned DINOv3 tokens are injected into the primary CLIP embedding space as a soft geometric stimulus. This configuration guarantees precise dense change localization while explicitly preventing the architecture from over-specializing to source-domain geometric features.

Furthermore, to prevent the injection of spatial details from disturbing image-level semantic reasoning, global change scoring is executed entirely by an isolated vision-language pathway, deliberately omitting the blended patch features. Extensive experiments reveal a clear training divergence that corroborates our decoupled topology: although dense spatial metrics may gradually decay due to geometric overfitting, the global stability maintained by the standalone CLIP classifier remains stable.

The main contributions of this work are summarized as follows:

- We reformulate remote sensing construction monitoring as a zero-shot, single-temporal structural change detection task, completely eliminating the reliance on precisely registered bi-temporal inputs during inference.
- We introduce BSSRS, a dual-vision framework that harmonizes semantic and structural priors via a constrained residual blending strategy. By spatially aligning and injecting DINOv3 tokens as a soft geometric stimulus into multi-level CLIP layers, we provide a robust solution to mitigate structural over-specialization.
- We propose a decoupled scoring topology that extracts dense predictions from the integrated spatial features while reserving an isolated CLIP pathway for global change classification, thereby ensuring stable semantic behavior across heterogeneous remote sensing datasets.

2 Related Work

Modern vision–language foundation models, such as CLIP [13] and its remote sensing variants [12], aim to capture rich visual semantics without relying on exhaustive pixel-level annotations. This paradigm is particularly relevant to remote sensing (RS) applications such as change detection and rapid disaster assessment [1, 15, 19], where constructing large-scale, densely annotated datasets is costly and labor-intensive [9]. By leveraging large-scale visual–semantic pretraining, such models demonstrate strong potential for identifying structural change cues—e.g., newly constructed buildings—across diverse geographic regions without task-specific supervision.

Our work is primarily related to four research directions: remote sensing change detection and domain generalization, vision–language modeling for zero-shot change detection in RS, foundation visual encoders for structure-aware modeling, and lightweight adapter-based adaptation.

Remote Sensing Change Detection and Domain Generalization. Building change detection (CD) has traditionally relied on supervised bi-temporal frameworks, including early Siamese architectures [7] and more recent transformer-based models such as ChangeFormer [2]. These approaches explicitly model pixel-wise differences between pre- and post-event images and therefore assume accurate co-registration and access to paired observations. As a result, they are highly sensitive to misalignment, seasonal variations, and domain shifts across geographic regions and sensors. To alleviate these limitations, prior studies have explored domain adaptation and domain generalization strategies [11, 18]. However, most existing methods still operate within supervised segmentation paradigms or require access to target-domain samples during training. In contrast, BSSRS reformulates building change detection as a single-temporal structural change detection problem. Beyond eliminating the reliance on bi-temporal inputs, BSSRS explicitly addresses the bottleneck of structural over-specialization by bridging dual-vision manifolds via a decoupled architecture, providing a zero-shot solution that is inherently more robust for real-world cross-domain deployment.

Vision–Language Models and Zero-Shot Change Detection in RS. The success of CLIP [13] has inspired growing research on vision–language modeling for remote sensing, including RemoteCLIP [12], GeoCLIP [4], and open-vocabulary RS segmentation frameworks [3]. While these models demonstrate strong semantic alignment, many require RS-specific pretraining or supervised fine-tuning. More recently, change-aware vision–language frameworks such as RSAD-CLIP [21] have reframed building change detection as a zero-shot change localization problem by introducing unchanged and changed textual prototypes. Despite their effectiveness, these methods primarily rely on semantic similarity between visual features and textual anchors, and may lack sufficient sensitivity to fine-grained geometric deviations in high-resolution RS imagery. To counteract this limitation without sacrificing semantic stability, BSSRS incorporates a structure-sensitive encoder to capture local geometric consistency. Crucially, to prevent domain collapse, we utilize the frozen CLIP backbone as an invariant contextual reference and retain an isolated vision-language pathway for image-level scoring, ensuring stable image-level semantic reasoning.

Foundation Visual Encoders for Structure-Aware Modeling. Recent studies reveal that large-scale self-supervised vision transformers exhibit strong geometry-aware representations even without language supervision. In the remote sensing domain, foundation models such as RingMo [17] leverage masked image modeling to extract robust structural features. Furthermore, general vision models such as Segment Anything (SAM) [10] and its remote sensing variants [20] have demonstrated powerful zero-shot structural parsing capabilities. However, while SAM relies heavily on explicit prompts to delineate discrete boundaries, models such as DINOv3 [16] learn continuous structural consistency and local feature

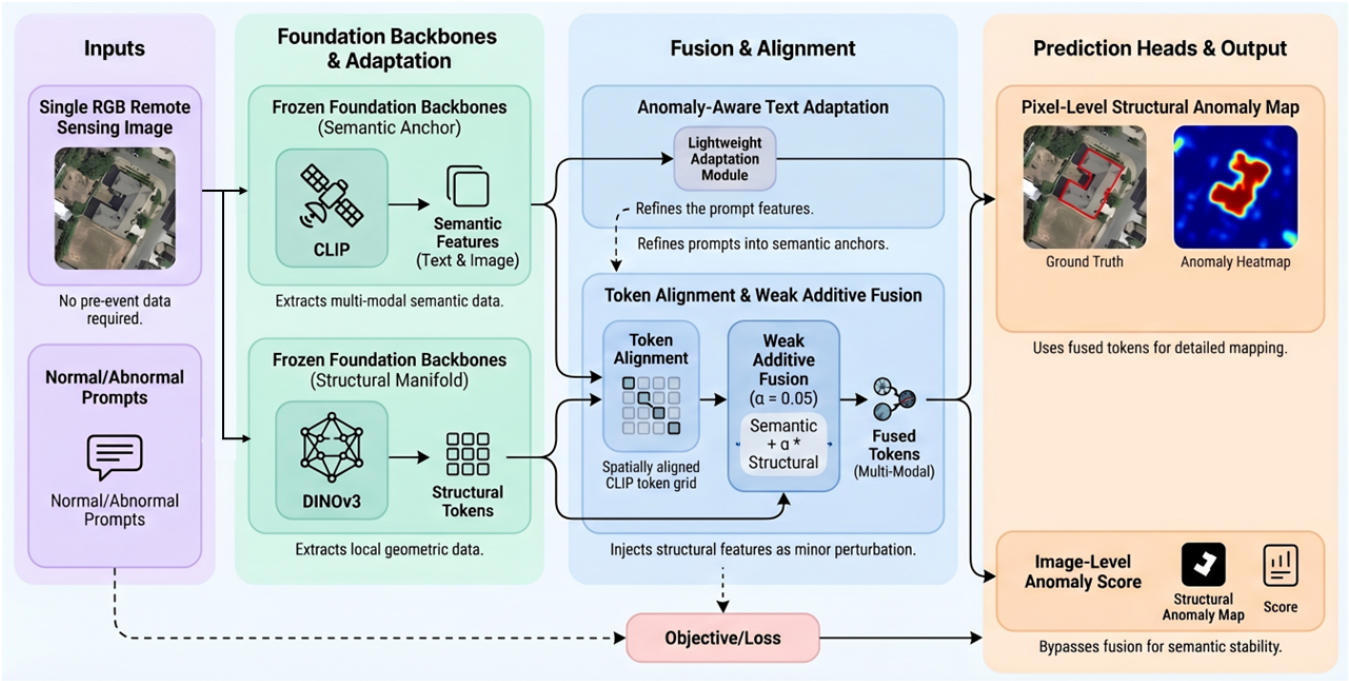

Figure 1: Architecture of the proposed BSSRS framework. Given a single post-event remote sensing image, BSSRS explicitly bridges dual manifolds by integrating two frozen foundation encoders: an invariant contextual reference (CLIP ViT-L/14@336) and a structure-sensitive encoder (DINOv3 ViT-L) [16]. Instead of complex token-mixing, patch-level tokens from multiple selected transformer layers are integrated via a constrained residual blending strategy. Specifically, spatially aligned DINOv3 features are injected as a soft geometric stimulus into the primary CLIP embedding space to produce a precise dense change map. Crucially, to avoid disturbing the original global semantics, robust image-level scores are derived directly from an isolated vision-language pathway, deliberately omitting the blended patch tokens.

coherence through self-distillation. This makes them inherently sensitive to geometric irregularities and boundary discontinuities without manual prompting. The need for such robust structure-aware modeling is further underscored by challenging remote sensing scenarios involving severe side-looking angles and off-nadir distortions, as characterized by datasets such as S2Looking [14]. These findings suggest that semantic alignment and structural sensitivity arise from distinct inductive biases and can provide complementary change evidence.

Unlike purely semantic CLIP-based approaches, BSSRS explicitly bridges these dual manifolds. Instead of employing complex token-mixing modules, we integrate patch-level representations via a constrained residual blending strategy across multiple selected layers. By spatially aligning and injecting DINOv3 tokens as a soft geometric stimulus into the primary CLIP embedding space, the framework captures both anthropogenic semantic cues and geometry-driven structural deviations. This dual-vision synergy improves robustness while explicitly mitigating the risk of structural over-specialization under cross-domain deployment.

Lightweight Adapter-Based Adaptation. Adapting large vision–language models to downstream RS tasks via full fine-tuning is computationally expensive and frequently compromises zero-shot generalization. Adapter-based learning introduces lightweight trainable modules while preserving the pretrained backbone, striking a balance between flexibility and robustness. Such strategies have been widely adopted in multimodal learning to stabilize downstream adaptation by relying on frozen models as structural or semantic anchors. In contrast to heavy fine-tuning or domain-specific pretraining, BSSRS maintains frozen foundation encoders and introduces only lightweight projection and adaptation modules alongside the residual blending. For dense localization, structural tokens are aligned and injected into the pixel-level pathway. Meanwhile, image-level change decisions deliberately omit the blended patch tokens and rely instead on the standalone CLIP classifier. This decoupled adaptation strategy preserves the complementary inductive biases of both manifolds while enabling efficient and stable cross-domain change localization.

3 Proposed Method

Figure 1 illustrates the overall pipeline of our proposed BSSRS framework. Given a single RGB remote sensing image $x \in \mathbb{R}^{3\times H\times W}$, the model predicts (i) a dense change map $A \in \mathbb{R}^{H\times W}$ for spatial localization and (ii) an image-level change score s_{img} for global decision-making. The framework is built upon a frozen CLIP visual–language backbone equipped with lightweight text and image adapters. To improve sensitivity to fine-grained structural irregularities, we further incorporate a frozen DINOv3 branch into the pixel-level pathway

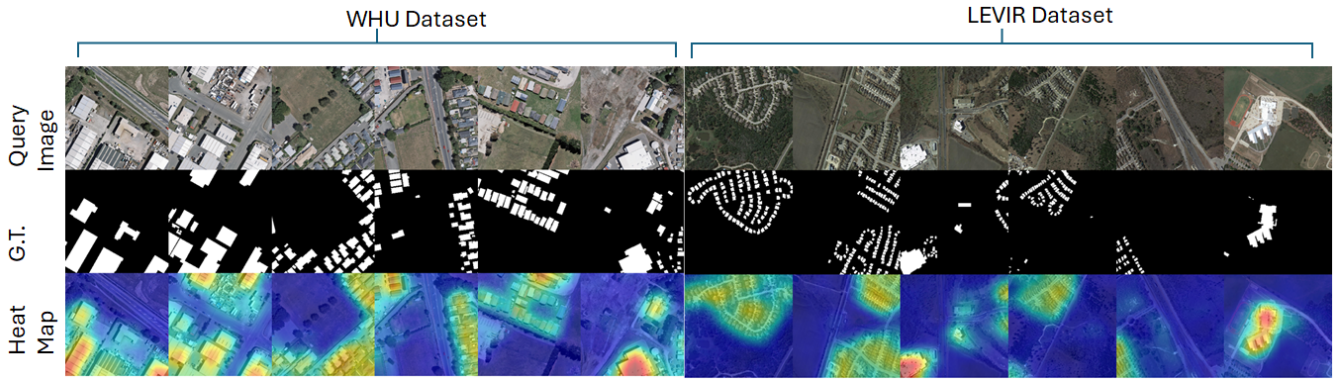

Figure 2: Qualitative visualization of BSSRS. From top to bottom: input image, ground-truth mask, and predicted change heatmap. Results are shown on LEVIR-CD and WHU under zero-shot cross-dataset transfer. By explicitly bridging dual manifolds via a constrained residual blending strategy, BSSRS generates spatially compact change heatmaps that closely align with ground-truth building footprints. The invariant contextual reference effectively prevents diffuse responses and semantic drift, while injecting geometry-aware structural representations as a localized soft geometric stimulus preserves fine-grained boundaries. This synergistic design demonstrates robust dense localization that effectively mitigates structural over-specialization without compromising global semantic integrity.

through a conservative training-time constrained residual blending strategy. Importantly, while DINOv3 provides complementary geometric cues for dense localization, the image-level prediction branch remains governed by an isolated vision-language pathway to preserve semantic stability.

3.1 Problem Setup

Let f_v^{clip} denote the CLIP vision encoder and f_t the CLIP text encoder. For a selected transformer layer set $\mathcal{L}$, the CLIP visual branch extracts semantic patch tokens:

$$Z_c^{(l)} \in \mathbb{R}^{N_c^{(l)} \times C_c}, \quad l \in \mathcal{L}, \tag{1}$$

while a frozen DINOv3 encoder extracts structure-aware patch tokens:

$$Z_d^{(l)} \in \mathbb{R}^{N_d^{(l)} \times C_d}. \tag{2}$$

Because the two backbones adopt different tokenization schemes and patch geometries, their spatial resolutions $N_c^{(l)}$ and $N_d^{(l)}$ generally differ. To capture both shallow structural patterns and deep semantic representations, we explicitly select a multi-level feature set for constrained residual blending:

$$\mathcal{L} = \{6, 12, 18, 24\}, \tag{3}$$

which allows the model to integrate multi-depth geometric stimuli into the primary CLIP embedding space.

3.2 Change-Aware Text Adaptation

To improve the separation between unchanged and changed semantics in the visual–textual space, the text branch refines change-aware prompt embeddings instead of directly relying on raw textual features. Let p_n and p_a denote the unchanged and changed prompts, respectively. Their adapted embeddings are defined as:

$$t_n = \psi_t(f_t(p_n)), \quad t_a = \psi_t(f_t(p_a)), \tag{4}$$

where $\psi_t(\cdot)$ is a lightweight text adaptation module. These adapted embeddings serve as invariant contextual references for both dense patch-level similarity matching and image-level classification.

3.3 Visual Alignment and Constrained Residual Blending

For each selected layer $l \in \mathcal{L}$, the CLIP and DINOv3 patch tokens are first projected into a shared embedding space through lightweight projection modules:

$$\hat{Z}_c^{(l)} = \phi_c^{(l)}(Z_c^{(l)}), \quad \hat{Z}_d^{(l)} = \phi_d^{(l)}(Z_d^{(l)}). \tag{5}$$

Since the token grids of the two encoders are not naturally aligned, the projected DINOv3 tokens are spatially aligned to the CLIP token layout before integration. Denoting this alignment operator by $\mathcal{A}(\cdot)$, we obtain:

$$\tilde{Z}_d^{(l)} = \mathcal{A}(\hat{Z}_d^{(l)}), \tag{6}$$

where $\tilde{Z}_d^{(l)}$ matches the spatial arrangement of $\hat{Z}_c^{(l)}$. This design keeps CLIP as the primary semantic reference manifold while introducing DINOv3 as an auxiliary structural cue.

Instead of using aggressive token interaction modules such as cross-attention, we adopt a constrained residual blending rule:

$$Z_{\text{fuse}}^{(l)} = \text{Norm}\left(\hat{Z}_c^{(l)} + \alpha \tilde{Z}_d^{(l)}\right), \tag{7}$$

where $\text{Norm}(\cdot)$ denotes feature normalization and α is a small scalar fusion coefficient. In our implementation, we set $\alpha = 0.10$. This formulation preserves CLIP as the dominant semantic manifold while allowing DINOv3 to contribute only a soft geometric stimulus, thereby reducing semantic distortion under severe modality shifts.

3.4 Dense Pixel-Level Change Localization

Dense change evidence is obtained by matching the blended visual tokens against the adapted text anchors $[t_n, t_a]$. For each selected

layer, the similarity map is computed as:

$$S^{(l)} = \mathrm{Sim}\left(Z_{\mathrm{fuse}}^{(l)}, [t_n, t_a]\right), \quad l \in \mathcal{L}, \tag{8}$$

where $\mathrm{Sim}(\cdot,\cdot)$ denotes cosine-style similarity in the aligned visual–textual space. The changed channel is then spatially reshaped to obtain the layer-wise change map $A^{(l)} \in \mathbb{R}^{H\times W}$.

The final dense change map is obtained by aggregating the predictions from all selected layers:

$$A = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} A^{(l)}. \tag{9}$$

This multi-level formulation allows the framework to integrate structural evidence across different representation depths while maintaining the semantic stability of the CLIP backbone.

3.5 Image-Level Scoring via Isolated Vision-Language Pathway

A key design choice of our framework is that the image-level branch is not derived from the blended visual tokens. To avoid structural noise interfering with global semantic reasoning, we preserve a standalone CLIP detection token, denoted as z_{det}, for image-level prediction.

The image-level logits are computed against the semantic text anchors as:

$$\ell = z_{\mathrm{det}}^{\top}[t_n, t_a], \tag{10}$$

followed by softmax normalization to obtain the final image-level change score:

$$s_{\mathrm{img}} = \mathrm{Softmax}(\ell). \tag{11}$$

Therefore, DINOv3 only refines the dense spatial localization pathway, while the global classification branch deliberately omits the blended patch tokens and remains entirely grounded in the isolated vision-language pathway.

3.6 Training Objective

The framework follows a lightweight CLIP-adaptation paradigm. The overall objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{text}} + \lambda\mathcal{L}_{\mathrm{image}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{text}}$ refines change-aware textual semantics and $\mathcal{L}_{\mathrm{image}}$ supervises the visually adapted patch-level pathway. During training, both CLIP and DINOv3 remain frozen, and only the lightweight adaptation and projection modules are optimized.

4 Experiments

Datasets and Benchmarks. We evaluate BSSRS on remote sensing benchmarks to assess single-temporal structural change detection and cross-domain generalization. LEVIR-CD [5] is repurposed by using only post-event images (T_2) and interpreting change masks as structural change labels, enabling training without pre-event imagery. WHU Building Dataset [8] is used for zero-shot cross-dataset evaluation, where building footprints are treated as change masks without any fine-tuning on the target domain.

Implementation Details. All experiments are implemented in PyTorch using a single NVIDIA RTX 4090 GPU with a batch size of 32. We adopt frozen CLIP ViT-L/14@336 and DINOv3 ViT-L as the dual foundation encoders. Images are resized to 518×518, yielding a 37×37 (1369) patch grid.

To explicitly align with our decoupled architecture, both foundation backbones remain strictly frozen. Only the lightweight projection modules and text/visual adaptation heads are trainable. Models are optimized using AdamW (lr = 1×10^{-4}, weight decay 1×10^{-4}) for 20 epochs. For the constrained residual blending strategy, we specifically select a multi-level feature set ($\mathcal{L} = \{6, 12, 18, 24\}$) and apply a conservative blending coefficient of $\alpha = 0.10$ across all selected layers, intentionally treating the spatially aligned DINOv3 features as a localized soft geometric stimulus. Furthermore, unlike previous pixel-grounded methods, image-level change scores are derived directly from the isolated vision-language pathway without any patch-level aggregation. Heatmap smoothing is performed via Gaussian blur with $\sigma = 1.0$.

Evaluation Metrics. We report Pixel AUC for localization, and Image AUC for image-level detection. For remote sensing benchmarks, we additionally report Average Precision (AP) and F1-score and IoU computed at the optimal threshold that maximizes the F1-score, which is essential under severe class imbalance.

4.1 Quantitative Performance and Generalization Analysis

The quantitative evaluation of BSSRS highlights the intrinsic challenges of zero-shot cross-dataset reasoning in remote sensing. Unlike conventional supervised learning, the objective is not to maximize source-domain accuracy, but to explicitly mitigate structural over-specialization via a decoupled architecture to learn representations that generalize across unseen domains.

Remote sensing datasets often exhibit substantial distribution shifts caused by sensor differences, acquisition conditions, and geographic variability. Therefore, cross-dataset transfer serves as a stringent test of whether a model successfully bridges dual manifolds to capture domain-invariant structural patterns, rather than merely memorizing dataset-specific appearance and geometric statistics.

4.1.1 Transferability and Structural Robustness. As summarized in Table 1, BSSRS shows reasonable cross-dataset transferability when trained on LEVIR-CD and directly evaluated on WHU without fine-tuning. In this setting, the model achieves a Pixel AUC of 95.38%, together with a Pixel F1-score of 63.99% and a Pixel IoU of 47.05%. Although the AUC is relatively high, the F1-score and IoU remain moderate, indicating that fine-grained localization precision and boundary delineation under zero-shot transfer still have room for improvement.

Table 1 also shows a clear performance asymmetry when the source and target datasets are swapped. When trained on WHU and evaluated on LEVIR-CD, the Pixel F1-score drops to 29.31% and the Pixel IoU decreases to 17.18%. A plausible explanation is that the two datasets exhibit different morphological characteristics. LEVIR-CD contains more widely distributed suburban structures and more heterogeneous natural backgrounds, whereas WHU more often contains denser and larger building instances with comparatively simpler surrounding contexts. As a result, training only on WHU

may provide less exposure to diverse background-to-structure transitions, which can reduce generalization when the model is tested on the more varied scenes in LEVIR-CD.

Despite this asymmetry, the results suggest that injecting structure-sensitive representations from DINOv3 as a localized soft geometric stimulus into the primary CLIP embedding space is highly beneficial for cross-domain change localization. The DINOv3 tokens contribute local geometric sensitivity through the constrained residual blending, while the frozen CLIP branch provides an invariant contextual reference under severe domain shift.

At the image level, the model remains stable across the two transfer directions. As reported in Table 1, the Image-level AP reaches 99.54% for LEVIR-CD → WHU and 99.31% for WHU → LEVIR-CD. This result supports our decoupled scoring architecture: by deriving image-level decisions directly from an isolated vision-language pathway rather than aggregating blended visual tokens, the model successfully preserves stable global semantic behavior, even when pixel-level boundary quality varies across heterogeneous datasets.

4.2 Qualitative Visualization

To further validate the effectiveness of explicitly bridging dual manifolds, we present qualitative results of BSSRS under zero-shot cross-dataset settings. As illustrated in Fig. 2, the model consistently highlights building-change candidates across diverse urban scenes without any target-domain fine-tuning.

Structural Boundary Precision. A key advantage of BSSRS lies in its ability to explicitly mitigate structural over-specialization while preserving fine-grained structural boundaries. The predicted change heatmaps are spatially compact and closely aligned with ground-truth building footprints, even under cluttered backgrounds and complex urban layouts.

This behavior reflects the successful bridging of complementary foundation priors. While the frozen CLIP backbone acts as an invariant contextual reference to maintain global consistency and prevent domain collapse, DINOv3 is injected as a localized soft geometric stimulus via residual blending, enhancing sensitivity to fine-grained geometric deviations without dominating the semantic manifold. This synergy results in sharp and coherent building contours rather than the diffuse responses typically seen in purely semantic models. These qualitative observations demonstrate that our constrained blending strategy effectively improves dense structural change localization without over-specializing to source geometries.

Failure Case Analysis and Structural Sensitivity. Despite strong overall performance, BSSRS encounters challenges in scenes with extremely small building footprints or severe shadow occlusion. In such cases, change responses remain spatially coherent but may exhibit reduced confidence, leading to partial detections. Importantly, the model rarely produces large false-positive regions.

This behavior highlights a crucial property of our decoupled design: maintaining structural sensitivity while utilizing the semantic anchor to prevent excessive semantic drift. Rather than grounding image-level decisions in blended dense pixel evidence, BSSRS reserves an isolated vision-language pathway for global scoring. Simultaneously, in the pixel-level pathway, structural features act merely as a soft stimulus to the primary semantic space. This separation ensures that localized activations are strictly regularized by the contextual reference, preventing change responses from spuriously spreading across visually similar yet structurally irrelevant areas under severe domain shift.

Table 1: Zero-shot cross-dataset performance of BSSRS. Source and Target denote training and testing datasets. By explicitly bridging dual manifolds via a decoupled architecture, BSSRS effectively mitigates structural over-specialization. It maintains strong pixel-level ranking performance through constrained residual blending (P-metrics) and stable image-level semantic behavior derived from the isolated vision-language pathway (I-metrics) across diverse urban morphologies. All metrics are reported in percentage (%).

Source	Target	P-AUC	P-AP	P-F1	P-IoU	I-AUC	I-AP
LEVIR-CD	WHU	**95.38**	71.00	**63.99**	**47.05**	**98.74**	**99.54**
WHU	LEVIR-CD	90.41	31.74	29.31	17.18	92.74	99.31

Limitations of Single-Temporal Reasoning. A limitation arises from the single-temporal formulation itself. On datasets such as LEVIR-CD, which provide bi-temporal annotations, the model may occasionally highlight pre-existing buildings as change candidates. Since BSSRS operates strictly without pre-event (T_1) imagery at inference, any prominent anthropogenic structure may be assigned high structural-change likelihood because no temporal reference is available.

This ambiguity is important for interpreting the quantitative results. For example, the relatively high scores obtained on the WHU dataset may be influenced in part by its dataset characteristics. WHU contains many standalone buildings set against relatively uniform natural backgrounds; under our formulation, such visually prominent structures are more likely to be flagged as change candidates. In contrast, for datasets with more complex urban textures and denser pre-existing structures (e.g., LEVIR-CD), the model may also highlight pre-existing buildings, which can be penalized by traditional change detection metrics. As a result, the current framework is better interpreted as measuring structural distinctiveness in a single image, rather than strictly capturing chronological change. However, this specific property makes the framework exceptionally well-suited for real-world monitoring of illegal land abuse or unauthorized building extensions, where any prominent anthropogenic structure appearing in a target restricted zone constitutes a potential violation.

This limitation reflects the inherent ambiguity of zero-shot structural reasoning without explicit temporal references. Incorporating temporal cues or consistency constraints into the bridged manifolds is a meaningful direction for future work.

5 Ablation Study

To quantitatively assess the contribution of each component in the proposed BSSRS framework, we conduct ablation experiments on the WHU Building Dataset under a zero-shot cross-dataset setting. All models are trained on LEVIR-CD and directly evaluated on WHU without any target-domain fine-tuning. Results are summarized in Table 2.

Table 2: Ablation study of BSSRS under zero-shot transfer (LEVIR-CD → WHU). T (Text Adp.) denotes change-aware text adaptation, and V (Vis. Adp.) denotes pixel-level visual adaptation via constrained residual blending with DINOv3. The decoupled design improves dense localization without affecting image-level scoring.

Method	T	V	P-AUC	P-AP	P-F1	P-IoU	I-AUC	I-AP
Vanilla CLIP	×	×	33.31	7.78	15.65	8.48	17.85	56.48
+ Vis. Adp. (wo Text Adp.)	×	✓	95.23	70.05	37.50	23.08	98.29	99.30
Full ($\alpha = 0.05$)	✓	✓	95.35	**71.05**	59.58	42.43	98.60	99.48
Full ($\alpha = 0.10$)	✓	✓	**95.38**	71.00	**63.99**	**47.05**	**98.74**	**99.54**

CLIP Baseline and Semantic Limitation. The vanilla CLIP model serves as a zero-shot baseline, struggling significantly under this cross-domain setting by achieving a Pixel F1-score of only 15.65% and a Pixel IoU of 8.48%. Because CLIP is primarily optimized for global image–text alignment rather than fine-grained structural reasoning, its change responses remain spatially diffuse and inaccurate. In remote sensing scenarios, change detection requires sensitivity to subtle geometric deviations rather than coarse semantic similarity. Therefore, while the frozen CLIP backbone acts as a strong global prior, it lacks the structural discrimination required for precise dense localization.

Effect of Visual Adaptation (wo Text Adp.). We first evaluate the isolated impact of the Visual Adaptation pathway, which integrates structure-sensitive cues from the DINOv3 branch without change-aware text prompts. As shown in Table 2, explicitly bridging the dual manifolds provides complementary structural information, boosting Pixel AUC from 33.31% to 95.23% and Pixel F1 from 15.65% to 37.50%. Unlike CLIP, which aligns visual tokens with language semantics, DINOv3 is trained via self-supervised learning to capture intrinsic geometric regularities and boundary continuity. By injecting DINOv3 tokens as a localized soft geometric stimulus into the primary CLIP embedding space, the model significantly enhances sensitivity to structural deviations. However, without adapted text anchors, the overall Pixel F1 (37.50%) and Pixel IoU (23.08%) remain constrained by semantic ambiguity.

Effect of Text Adaptation (Full Architecture). Introducing the Change-Aware Text Adaptation alongside Visual Adaptation completes our full decoupled architecture. This indicates that explicitly disentangling normal and changed prototypes within the semantic manifold is crucial for change-aware reasoning. By refining language prompts into an invariant contextual reference, the model effectively reduces semantic ambiguity and stabilizes cross-domain transfer against semantic drift. Consequently, adding Text Adaptation leads to massive gains in fine-grained localization: compared to the Visual Adaptation-only baseline, the Full model ($\alpha = 0.10$) jumps from 37.50% to 63.99% in Pixel F1, and from 23.08% to 47.05% in Pixel IoU.

Complementarity and Impact of Fusion Coefficient (α). To further investigate the impact of the constrained residual blending, we compare two fusion coefficients in the full architecture: $\alpha = 0.05$ and $\alpha = 0.10$. As shown in Table 2, both settings significantly outperform the partial adaptations. The model with $\alpha = 0.05$ achieves a highly robust 95.35% Pixel AUC and 71.05% Pixel AP. Slightly increasing the geometric stimulus to $\alpha = 0.10$ yields the best overall dense localization, boosting Pixel F1 to 63.99% and Pixel IoU to 47.05%, while simultaneously maintaining peak image-level stability (99.54% Image AP). These results confirm that semantic alignment and structural representation learning are complementary. The carefully tuned geometric stimulus refines spatial precision without overwhelming the semantic space, while the isolated vision-language pathway independently maintains stable global semantic behavior under severe domain shifts [1, 2].

6 Conclusion

We presented BSSRS, a dual-vision framework that reformulates remote sensing change detection as single-temporal structural change detection, completely eliminating the need for bi-temporal image pairs at inference. By explicitly bridging a frozen semantic encoder (CLIP) and a structure-sensitive encoder (DINOv3) via a decoupled architecture, BSSRS provides a highly robust solution for zero-shot cross-domain analysis.

Through constrained residual blending, DINOv3 features act as a localized soft geometric stimulus to refine dense structural boundaries without overwhelming the semantic space. Simultaneously, an isolated vision-language pathway deliberately bypasses this token fusion to guarantee stable global semantic behavior under severe domain shifts. This synergistic design effectively mitigates structural over-specialization, enabling BSSRS to achieve 95.38% Pixel AUC and 63.99% Pixel F1-score without target-domain fine-tuning.

Future Work. Future work includes incorporating temporal cues to better distinguish pre-existing anthropogenic structures from genuine changes, and evaluating BSSRS on broader geospatial benchmarks with varying sensor characteristics.

Acknowledgments

The authors would like to express their sincere gratitude to Prof. Jenq-Neng Hwang (University of Washington, Seattle, USA) and Prof. Shang-Hong Lai (Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan) for their insightful discussions and invaluable guidance. This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants NSTC 114-2221-E-007-030-MY2 and NSTC 113-2221-E-007-104-MY3. Furthermore, this research was made possible by the advanced computational infrastructure and academic resources generously provided by the National Center for High-Performance Computing (NCHC), National Institutes of Applied Research (NIAR), Taiwan.

References

[1] K. Amini, Y. Liu, J. E. Padgett, et al. 2025. Debris segmentation using post-hurricane aerial imagery. *Computer-Aided Civil and Infrastructure Engineering* 40, 25 (2025), 4116–4131.

[2] Wele Gedara Chaminda Bandara and Vishal M. Patel. 2022. A Transformer-Based Siamese Network for Change Detection. *CoRR* abs/2201.01293 (2022). doi:10.48550/arXiv.2201.01293

[3] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. 2024. Open-Vocabulary Remote Sensing Image Semantic Segmentation. *CoRR* abs/2409.07683 (2024). doi:10.48550/arXiv.2409.07683

[4] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. *CoRR* abs/2309.16020 (2023). doi:10.48550/arXiv.2309.16020

[5] Hao Chen and Zhenwei Shi. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing* 12, 10 (2020), 1662.

[6] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, H. Zhao, Q. Zhao, and S. Xiang. 2024. Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review. *Remote Sensing* 16, 13 (2024), 2355.

[7] R. C. Daudt, B. L. Saux, and A. Boulch. 2018. Fully Convolutional Siamese Networks for Change Detection. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 4063–4067.

[8] Shunping Ji, Shiqing Wei, and Meng Lu. 2018. Fully Convolutional Networks for Multi-Source Building Extraction From an Open Aerial and Satellite Imagery Dataset. *IEEE Transactions on Geoscience and Remote Sensing* 57, 1 (2018), 574–586. doi:10.1109/TGRS.2018.2858817

[9] H. Jiang, M. Peng, H. Zhong, Y.and Xie, Z. Hao, J. Lin, X. Ma, and X. Hu. 2022. A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images. *Remote Sensing* 14, 7 (2022), 1552.

[10] Alexander Kirillov. 2023. Segment Anything. In *ICCV*.

[11] Chenbin Liang, Weibin Li, Yunyun Dong, and Wenlin Fu. 2024. Single Domain Generalization Method for Remote Sensing Image Segmentation via Category Consistency on Domain Randomization. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16. doi:10.1109/TGRS.2024.3379669

[12] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16.

[13] A. Radford, J. W. Kim, C. Hallacy, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning* 139 (2021), 8748–8763.

[14] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. 2021. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing* 13, 24 (2021), 5094.

[15] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan. 2020. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sensing* 12, 10 (2020), 1688.

[16] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. 2025. Dinov3. *arXiv preprint arXiv:2508.10104* (2025).

[17] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, et al. 2023. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–22.

[18] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. 2016. Recent Advances in Domain Adaptation for the Classification of Remote Sensing Data. *IEEE Geoscience and Remote Sensing Magazine* 4, 2 (2016), 41–57. doi:10.1109/MGRS.2016.2548504

[19] G. Wang, S. Y. Shin, and G. Jo. 2024. An improved post-hurricane building damaged detection method based on transfer learning. *Indonesian Journal of Electrical Engeening and Computer Science* 33, 3 (2024), 1546–1556.

[20] Lei Zhang. 2024. SAM for remote sensing. *Remote Sensing Letters* (2024).

[21] Yu Zhang and Zhi Gao. 2025. RSAD-CLIP: Zero-Shot Remote Sensing Anomaly Detection of the Earth's Surface Based on Pre-Trained Vision-Language Model. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1–5.