Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

FIFA-RS: Fine-grained Image-Feature Alignment for Structural Anomaly Reasoning in Remote Sensing

SHIH-CHIH LIN[1] Jia-Xian Jian[2] YunTung Chu[3] Wei-Chieh Sun[3]
[1]National Tsing Hua University
[2]National Cheng Kung University
[3]University of Washington

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 images may be unavailable or subject to severe registration errors. To address this limitation, we propose **FIFA-RS**, a zero-shot framework that formulates change detection as a **single-temporal structural anomaly reasoning** problem.

FIFA-RS enhances the ability of vision–language models to characterize anthropogenic structures without relying on temporal references. Built upon a frozen CLIP backbone, the proposed framework adopts a lightweight two-stage adaptation strategy that combines token-level high-pass adaptation with an image-only 2D spatial high-pass enhancement branch. The former suppresses token-level common bias and emphasizes relative feature differences, while the latter sharpens local geometric structures such as building contours and boundaries. These structurally enhanced features are further aggregated through learnable multi-scale fusion for dense pixel-level anomaly localization.

Extensive experiments indicate that FIFA-RS exhibits strong cross-dataset generalization across diverse remote sensing scenarios. When trained on LEVIR-CD using only post-event images and evaluated on the WHU Building Dataset in a zero-shot setting, the proposed method achieves a **94.91% Pixel AUC** and a **61.60% F1-score**. These results suggest that lightweight structural adaptation provides an effective and efficient solution for single-temporal remote sensing analysis. The code is available at https://github.com/leolin65/FIFA-RS.git.

Keywords: Single-temporal Change Detection, Structural Anomaly Detection, Zero-Shot Learning, Remote Sensing, Vision-Language Adaptation

1. Introduction

The rapid pace of global urban expansion has intensified the demand for efficient Earth surface monitoring. In rapidly developing regions, timely detection of newly constructed buildings is crucial for sustainable urban planning and proactive resource allocation. Remote sensing (RS), with its large-scale coverage and high revisit frequency, has become a key tool for urban change analysis [1].

Despite substantial progress, conventional building change detection (CD) methods still rely heavily on precisely registered bi-temporal image pairs and dense pixel-level annotations. In practice, acquiring well-aligned pre-event (T1) imagery is often costly or infeasible, especially in time-critical scenarios such as disaster response or rapid urban expansion [2]. Moreover, supervised CD models often suffer severe performance degradation under cross-domain deployment due to variations in sensor characteristics, illumination, and architectural styles [3].

To address these limitations, we reformulate the detection of newly constructed buildings as a **zero-shot, single-temporal structural anomaly reasoning** problem. By leveraging the open-vocabulary representation of Contrastive Language–Image Pre-training (CLIP) [4], our approach removes the reliance on paired temporal inputs during inference

leolin65@gapp.nthu.edu.tw, n28101173@gs.ncku.edu.tw, yuntungc@uw.edu, wsun12@uw.edu

and avoids costly fine-tuning of domain-specific remote sensing foundation models [5]. Although bi-temporal datasets are used during training to expose structural change patterns, **our framework operates strictly in a single-temporal manner at inference**, requiring only a post-event image.

However, directly transferring CLIP from natural-image domains to overhead remote sensing imagery remains challenging. Bridging the domain gap between web-scale pretraining data and aerial perspectives requires more than prompt engineering alone. While anomaly-aware frameworks such as AA-CLIP [6] adopt two-stage adaptation to stabilize semantic representations, they mainly operate in the spatial domain and may struggle with fine-grained structural variations such as building boundaries and geometric discontinuities.

To overcome these issues, we propose a lightweight adaptation framework that enhances CLIP with complementary structural inductive biases. The key idea of FIFA-RS is that lightweight structural adaptation, when aligned with a frozen CLIP prior, is sufficient to support robust single-temporal anomaly reasoning under severe cross-domain shifts. Concretely, instead of explicit frequency-domain decomposition, our design combines token-level high-pass adaptation with an image-only 2D spatial high-pass enhancement.

This design is related to prior anomaly detection approaches that improve structural sensitivity through learned representations or reconstruction constraints [7, 8], but differs in that it operates directly within a frozen vision–language backbone for zero-shot remote sensing reasoning. The former suppresses common token-level bias and highlights relative feature differences, while the latter sharpens local geometric structures on reconstructed patch grids. Together with learnable multi-scale fusion, these components address complementary failure modes, namely token bias, weak boundary sensitivity, and scale inconsistency. As a result, FIFA-RS enables more effective geometry-aware anomaly reasoning without heavy architectural changes.

The main contributions of this work are threefold:

(1) We formulate remote sensing construction monitoring as a zero-shot, single-temporal structural anomaly reasoning task, eliminating the need for bi-temporal inputs at inference.
(2) We propose a lightweight adaptation framework that combines token-level high-pass adaptation with an image-only 2D spatial high-pass enhancement branch to improve sensitivity to structural boundaries and geometric changes.
(3) We demonstrate strong cross-dataset generalization by training on LEVIR-CD and evaluating on the WHU Building Dataset, highlighting the effectiveness of the proposed structural inductive biases.

2. Related Work

Modern vision–language foundation models 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 [2, 9, 10], where constructing large-scale, densely annotated datasets is costly and labor-intensive [1]. By leveraging large-scale visual–semantic pretraining, such models exhibit strong potential for identifying structural anomalies—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 anomaly detection in RS, adapter-based model adaptation, and structural adaptation.

Remote Sensing Change Detection and Domain Generalization. Building change detection (CD) has traditionally relied on supervised bi-temporal frameworks, including

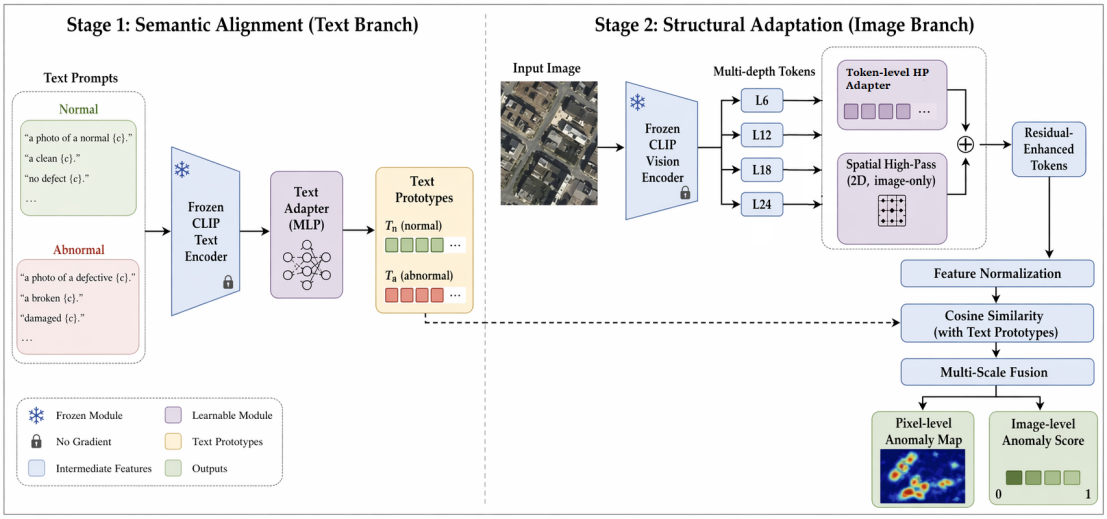

Figure 1. Architecture of the proposed FIFA-RS framework. Stage 1 performs semantic alignment by encoding normal and abnormal prompts with a frozen CLIP text encoder and learning task-adaptive text prototypes. Stage 2 performs structural adaptation by extracting multi-depth visual tokens from a frozen CLIP vision encoder and enhancing them using token-level high-pass adaptation and image-only 2D spatial high-pass filtering. The resulting features are matched with text prototypes through cosine similarity, and learnable multi-scale fusion produces the final pixel-level anomaly map and image-level anomaly score.

early Siamese architectures [11] and more recent transformer-based models such as ChangeFormer [12]. 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 mitigate these limitations, prior studies have explored domain adaptation and domain generalization strategies [13, 14]. However, most existing methods still operate within fully supervised segmentation paradigms or require access to target-domain samples during training. In contrast, our framework reformulates building change detection as a **single-temporal structural anomaly reasoning** problem, removing the dependency on paired temporal inputs during inference and improving applicability in real-world deployment.

Vision–Language Models and Zero-Shot Anomaly Detection in RS. The success of CLIP [4] has inspired a growing body of research on vision–language modeling for remote sensing, including RemoteCLIP [15], GeoCLIP [16], and open-vocabulary RS segmentation frameworks [17]. While these models demonstrate strong semantic alignment, many require extensive RS-specific pretraining or supervised fine-tuning.

More recently, anomaly-aware vision–language frameworks such as AA-CLIP [6] and RSAD-CLIP [18] have reformulated building change detection as a zero-shot anomaly localization problem by introducing normal and abnormal textual prototypes. In parallel, a complementary line of research has explored representation learning and reconstruction-based approaches for anomaly detection, including quantization-based models and CLIP-guided feature reconstruction methods [7, 8, 19, 20]. However, these methods are primarily developed for industrial or general anomaly detection settings and do not explicitly address structural reasoning in remote sensing imagery.

Adapter-Based Adaptation and Structural Adaptation. Adapting large vision–language models to downstream RS tasks via full fine-tuning is computationally expensive

and may compromise their generalization ability. Adapter-based learning offers a practical alternative by introducing lightweight trainable modules while preserving the pretrained backbone.

In parallel, enhancing structural sensitivity has been shown to be critical for dense prediction tasks. Prior studies [21–23] suggest that emphasizing high-frequency or structure-aware cues can improve robustness to appearance variations. Similar observations have also been reported in anomaly detection literature, where quantization- and reconstruction-based representations benefit from structure-sensitive feature modeling [19, 20]. However, many existing approaches rely on global spectral transformations or explicit Fourier decomposition, which introduce additional computational complexity and are not always necessary for capturing local geometric structures. In contrast, our approach adopts a simpler yet effective strategy by combining token- level high-pass adaptation with image-only 2D spatial high-pass enhancement applied to patch features. This design directly targets local geometric cues such as building edges and boundaries, while remaining lightweight and fully compatible with frozen vision–language backbones. Rather than introducing heavier spectral modules or target-domain adaptation, FIFA-RS focuses on lightweight structural adaptation that preserves the semantic prior of the frozen backbone. As a result, it provides a practical solution for zero-shot structural anomaly localization in remote sensing imagery.

3. Methodology

Problem Formulation. Given a single RGB remote sensing image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, our goal is to predict a pixel-level structural anomaly mask $\mathbf{m} \in \{0,1\}^{H\times W}$ and an image-level anomaly label $y \in \{0,1\}$. Unlike conventional bi-temporal change detection approaches, we consider a **single-temporal, zero-shot** setting, where only post-event imagery is available at inference time. This formulation is motivated by practical deployment scenarios in which pre-event references are unavailable or unreliable.

For each semantic category c, the scene is modeled using a binary state space (normal vs. anomalous), represented by a class-specific text prototype matrix

$$\mathbf{T}_c = [\mathbf{t}_c^n, \mathbf{t}_c^a] \in \mathbb{R}^{d\times 2},$$

where d is the embedding dimension, and $\mathbf{t}_c^n$ and $\mathbf{t}_c^a$ denote the normal and anomalous semantic prototypes, respectively.

Overall Architecture. Our framework is built upon a frozen CLIP ViT-L/14 backbone with embedding dimension $d = 768$. To capture structural anomalies at multiple resolutions, visual tokens are extracted from intermediate transformer layers

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

This yields multi-scale patch features $\mathbf{F}^\ell \in \mathbb{R}^{B\times N\times d}$, where N denotes the number of spatial patches, together with a global image-level feature $\mathbf{q} \in \mathbb{R}^{B\times d}$ for anomaly classification.

Lightweight adapters are inserted into both the visual and textual branches, while all CLIP backbone parameters remain frozen. This design preserves the open-vocabulary semantic prior of CLIP while introducing only a small number of trainable parameters for domain adaptation.

Token-Level High-Pass Adaptation and Image-Only 2D Spatial High-Pass Enhancement. Our latest implementation does not employ an FFT-based frequency decomposition module. Instead, it uses a simpler and more stable design that combines: (i) token-level high-pass adaptation implemented in `SimpleAdapter`, and (ii) an additional image-only 2D spatial high-pass enhancement branch applied to patch tokens.

Specifically, the existing adapter first projects the input token features and then applies a token-level high-pass residual:

$$\mathbf{Y}_{1D} = \mathrm{Adapter}_{1D}(\mathbf{X}),$$

which suppresses token-level common bias and enhances relative feature differences. This branch is shared by both the text and image pathways.

To further strengthen structural cues in the image branch, we introduce a lightweight 2D spatial high-pass enhancement branch that operates only on patch tokens. Given patch tokens $\mathbf{F}^{\ell} \in \mathbb{R}^{B\times N\times C}$, we first reshape them into a 2D spatial grid:

$$\mathbf{U}^{\ell} \in \mathbb{R}^{B\times C\times H\times W},$$

where $N = H \times W$. A fixed depthwise Laplacian filter is then applied channel-wise:

$$\mathbf{H}^{\ell}_{2D} = \mathrm{HP}_{2D}(\mathbf{U}^{\ell}),$$

where HP_{2D} denotes a lightweight 2D high-pass convolution. The resulting feature map is fused back through a learnable residual coefficient:

$$\widetilde{\mathbf{U}}^{\ell} = \mathbf{U}^{\ell} + \alpha_{2D}\mathbf{H}^{\ell}_{2D},$$

where α_{2D} is a learnable scalar initialized with a small value. The enhanced spatial features are then reshaped back into patch-token form for subsequent projection and anomaly reasoning.

This design provides two complementary inductive biases:

- **1D high-pass adaptation** suppresses token-level bias and emphasizes relative differences in the feature sequence.
- **Image-only 2D spatial high-pass enhancement** sharpens building contours, boundaries, and local geometric structures in the image branch.

Importantly, the 2D branch is applied only to image patch tokens and does not affect the text branch, ensuring that spatial priors are introduced only where geometric structure is meaningful.

More importantly, the key modules in FIFA-RS play complementary roles. Token-level high-pass adaptation suppresses common token bias, the image-only 2D branch enhances local geometric boundaries, and multi-scale fusion aggregates anomaly evidence across transformer depths. Their combination is therefore not a simple accumulation of modules, but a coordinated structural adaptation strategy for robust single-temporal reasoning.

Per-Layer Gating and Multi-Scale Fusion. To avoid applying the same adaptation strength to all transformer layers, the framework introduces learnable per-layer gates for both the text and image branches. These gates modulate the contribution of each adapter residual at different depths, allowing the model to automatically adjust the adaptation strength across layers.

For dense anomaly localization, scale-specific anomaly maps are first generated from the multi-scale patch features. Instead of simply averaging them, the framework employs a learnable multi-scale fusion module to combine anomaly predictions from different transformer depths:

$$\mathbf{M}_{\mathrm{fuse}} = \Phi_{\mathrm{ms}}(\mathbf{S}^{6}, \mathbf{S}^{12}, \mathbf{S}^{18}, \mathbf{S}^{24}),$$

where $\mathbf{S}^{\ell}$ denotes the class-aware similarity map at layer ℓ, and Φ_{ms} is a learnable fusion function. This improves robustness by allowing the model to adaptively aggregate complementary evidence from shallow and deep representations. In particular, shallower features preserve finer boundary detail, whereas deeper features provide stronger semantic stability. The fusion module balances both for dense anomaly localization.

Text Prototype Learning and Dense Prediction. To obtain stable semantic anchors, we adopt prompt ensembling for each category c. A set of normal prompts $\mathcal{P}^{n}_{c}$ and

anomalous prompts $\mathcal{P}_c^a$ are encoded by the adapted text branch and averaged to form:

$$\mathbf{t}_c^s = \mathrm{Norm}\left(\frac{1}{|\mathcal{P}_c^s|}\sum_{p\in\mathcal{P}_c^s}\mathrm{Norm}(E_{\text{text}}(p))\right), \quad s\in\{n,a\}.$$

To explicitly separate normal and anomalous semantics, we apply an orthogonality regularizer:

$$\mathcal{L}_{\text{orth}} = \left((\mathbf{t}_c^n)^\top \mathbf{t}_c^a\right)^2.$$

Given multi-scale patch features $\mathbf{F}^\ell$ and the prototype matrix $\mathbf{T}_c$, class-aware similarity maps are computed as

$$\mathbf{S}^\ell = 100\cdot\mathbf{F}^\ell\mathbf{T}_c \in \mathbb{R}^{B\times N\times 2},$$

and then reshaped into dense anomaly predictions. The final pixel-level anomaly map is obtained by the learnable multi-scale fusion module, while the global feature $\mathbf{q}$ is used for image-level anomaly prediction.

Two-Stage Adaptation and Inference. Training follows a two-stage strategy. In Stage-1 (*semantic alignment*), only the text adapters are optimized so that robust class-specific normal/anomalous prototypes can be established while the image branch remains frozen. In Stage-2 (*structural adaptation*), the text prototypes are fixed and the image adapters, image-only 2D spatial high-pass enhancement branch, per-layer image gates, and learnable multi-scale fusion module are optimized for structural anomaly localization.

At inference time, a single post-event image is forwarded through the adapted visual branch. Scale-specific anomaly maps are computed from visual-text similarity, fused by the learnable multi-scale fusion module, and used to produce the final pixel-level anomaly map. The global visual feature is simultaneously matched against the text prototypes to obtain the image-level anomaly score. Notably, the entire inference process is performed in a **single-temporal, zero-shot** manner without requiring any pre-event reference image.

4. Experiments

4.1. Datasets and Benchmarks

We evaluate our framework on remote sensing benchmarks to assess single-temporal structural anomaly reasoning and cross-domain generalization. LEVIR-CD [24] is repurposed by using only post-event images (T_2) during training and interpreting change masks as structural anomaly labels. This enables single-temporal learning without requiring pre-event imagery at inference. WHU Building Dataset [25] is used for zero-shot cross-dataset evaluation, where building footprints are treated as anomaly masks without any fine-tuning on the target domain.

4.2. Implementation Details

All experiments are implemented in PyTorch. We use a frozen CLIP ViT-L/14-336 backbone, and all images are resized to 518×518, yielding a 37×37 patch grid. Training is conducted in two stages. In Stage-1, the text adapters are optimized for semantic alignment using Adam with a learning rate of 1×10^{-5}, a batch size of 16, and 5 epochs. In Stage-2, the text prototypes are fixed, while the image adapters, the image-only 2D spatial high-pass enhancement branch, per-layer image gates, and the learnable multi-scale fusion module are optimized using Adam with a learning rate of 5×10^{-4}, a batch size of 2, and 20 epochs.

Unlike the earlier frequency-domain design, the current model does not employ FFT-based amplitude–phase decomposition. Instead, structural adaptation is achieved through the combination of (i) token-level high-pass adaptation and (ii) an image-only 2D spatial high-pass enhancement branch applied to reshaped patch grids. This design remains

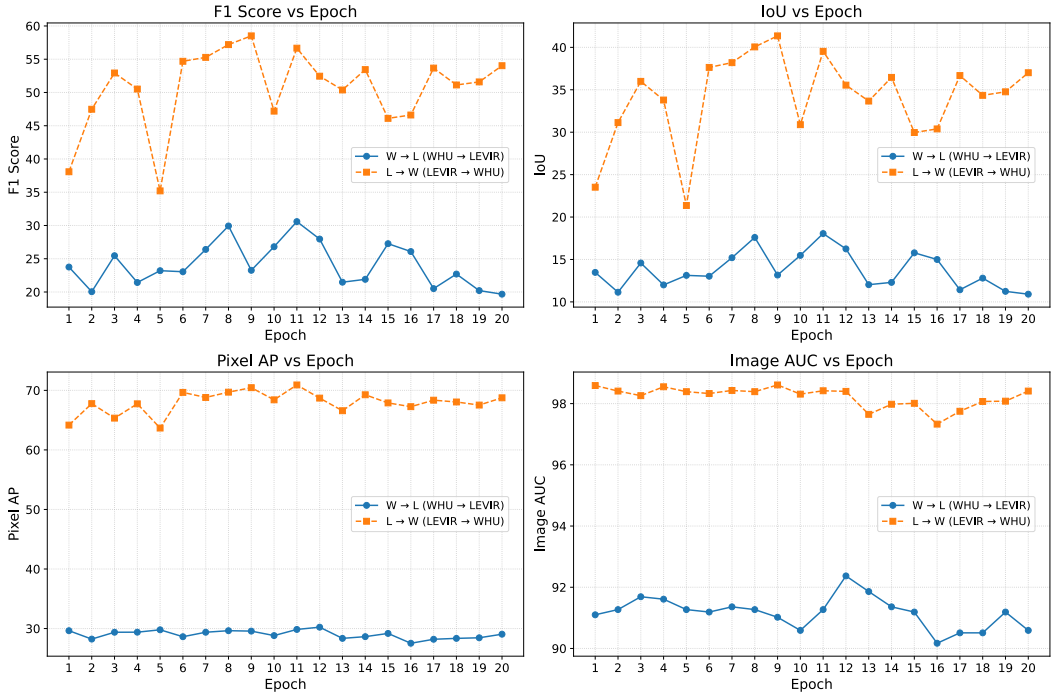

Figure 2. Cross-dataset training dynamics under domain shift (W: WHU, L: LEVIR). Performance over 20 epochs for bidirectional transfer (W→L and L→W) in terms of F1, IoU, Pixel AP, and Image AUC. Zero-shot generalization is non-monotonic, with LEVIR→WHU peaking near Epoch 9 before mild pixel-level degradation, indicating source-domain over-specialization. Image-level AUC remains comparatively stable, reflecting preserved semantic alignment in the frozen CLIP backbone.

lightweight while providing stronger sensitivity to building boundaries and local geometric variations.

4.3. **Evaluation Metrics**

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

4.4. **Quantitative Performance Evaluation and Generalization Analysis**

The quantitative evaluation highlights the intrinsic challenge of zero-shot cross-dataset reasoning in remote sensing. The central question is whether a lightweight structurally adapted model can remain effective on unseen target domains without temporal references or target-domain fine-tuning. Accordingly, the goal is not merely to optimize source-domain fitting, but to learn representations that remain effective under severe cross-domain shifts. This setting is substantially more difficult than conventional supervised segmentation or in-domain evaluation, since distribution shifts in sensor characteristics, illumination, geographic layouts, and building morphology may cause significant performance degradation [3, 13, 14].

4.4.1. **Transferability and Source-Domain Robustness**

As summarized in Table 1, the proposed framework demonstrates strong transferability when trained on the **source domain (LEVIR-CD)** and directly evaluated on the **target domain (WHU Building Dataset)** without any fine-tuning. The model achieves a **94.91% Pixel AUC** and a **61.60% F1-score**, indicating accurate localization of structural anomalies across previously unseen geographic regions.

These results suggest that the structural inductive biases introduced by the proposed adapters—namely, token-level high-pass adaptation and image-only 2D spatial high-pass enhancement—provide a robust prior for building anomaly reasoning. Rather than relying on dataset-specific textures, the model focuses on local geometric cues such as boundaries, contours, and structural irregularities. This improves transferability across domains.

At the image level, the framework also exhibits stable anomaly detection behavior under severe domain shifts. As reported in Table 1, the model consistently attains an **Image AP of 99.43%**, suggesting that the global reasoning token derived from the frozen CLIP backbone preserves strong semantic consistency across domains. This indicates that the model does not merely memorize source-domain appearance statistics, but instead learns transferable global representations that remain robust to environmental and structural variation.

4.4.2. **The Overfitting–Generalization Trade-off in Zero-shot Learning**

A critical observation revealed by the training dynamics, as illustrated in Fig. 2, is that zero-shot performance does not monotonically improve with additional training epochs. In the source-to-target transfer setting, where LEVIR-CD is the source and WHU is the target, the model reaches its optimal structural reasoning capability. Beyond this point, pixel-level metrics such as F1-score and IoU exhibit fluctuations or mild degradation, despite continued optimization on the source dataset.

This behavior reflects the classic **overfitting–generalization trade-off** in zero-shot adaptation. In early epochs, the lightweight adapters capture structural invariants shared across domains. However, excessive optimization encourages the model to absorb source-specific appearance statistics, such as sensor-dependent noise patterns, illumination conditions, or background textures, thereby reducing its flexibility on the target domain.

The adopted two-stage training strategy is designed to alleviate this issue. By freezing the CLIP backbone and restricting adaptation to lightweight text and image modules, the framework preserves CLIP's open-vocabulary semantic prior while allowing limited task-specific refinement. This design helps the model benefit from source-domain supervision without sacrificing cross-domain robustness.

4.4.3. **Discussion on Performance Disparity**

The asymmetric performance observed in the reverse transfer setting, where WHU serves as the source domain and LEVIR-CD as the target domain, further supports the above analysis. As reported in Table 1, this setting yields a significantly lower **Pixel F1-score of 19.26%**.

This performance gap can be attributed to the relatively homogeneous building styles and limited structural diversity in the WHU dataset. Consequently, models trained on WHU are more susceptible to distribution shifts when exposed to the more diverse urban morphologies and complex spatial layouts of LEVIR-CD. This observation highlights an important conclusion: **source-domain diversity strongly affects the upper bound of zero-shot generalization**. Datasets with richer structural variability enable the model to learn more universal geometric representations, whereas homogeneous source domains restrict transferability. In this regard, strategic early stopping together with structurally

diverse source-domain training data remains crucial for robust zero-shot remote sensing interpretation.

Source	Target	P-AUC	P-AP	P-F1	P-IoU	I-AUC	I-AP
LEVIR-CD	WHU	**94.91**	**70.36**	**61.60**	**44.51**	**98.44**	**99.43**
WHU	LEVIR-CD	88.63	28.06	19.26	10.65	91.44	99.14

Table 1. Zero-shot cross-dataset performance of FIFA-RS. The **Source** and **Target** columns denote the training and testing datasets, respectively. All metrics are reported in percentages (%).

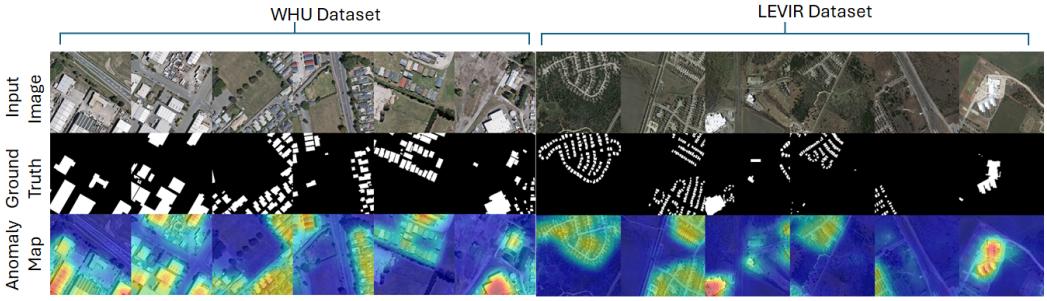

Figure 3. Qualitative results of FIFA-RS under cross-dataset zero-shot settings. Each column shows (top to bottom) the input image, ground-truth mask, and predicted anomaly map. The model accurately localizes structural anomalies across datasets, demonstrating strong generalization without target-domain training.

Method	Text Adp.	Image Adp.	P-AUC	P-F1	P-IoU	I-AUC	I-AP
Vanilla CLIP (Baseline)	✗	✗	35.97	5.38	2.76	43.1	70.32
w/ Text Adapter	✓	✗	66.8	26.60	15.34	87.2	95.05
Ours (Full Model)	✓	✓	**94.91**	**61.60**	**44.51**	**98.44**	**99.43**

Table 2. Ablation study of FIFA-RS on the WHU Building Dataset. **Vanilla CLIP** denotes direct evaluation using pretrained CLIP weights without any task-specific training. "Text Adp." and "Image Adp." denote the textual and visual adapters, respectively. All metrics are reported in percentages (%).

4.5. **Ablation Study**

The ablation results reveal three key observations. First, the **Vanilla CLIP** baseline performs poorly, indicating that direct transfer from natural-image pretraining is insufficient for structural anomaly reasoning in remote sensing.

Second, enabling only the **Text Adapter** significantly improves performance (from 35.97% to 66.8% Pixel AUC), demonstrating the importance of semantic alignment in establishing reliable normal/anomalous prototypes.

Finally, the **Full Model** further boosts performance to 94.91% Pixel AUC and 61.60% F1-score, showing that image-side structural adaptation is critical for precise localization. These results confirm that semantic alignment and structural adaptation are complementary.

4.6. **Qualitative Visualization**

To further validate the proposed structural adaptation strategy, we present qualitative results of **FIFA-RS** under zero-shot cross-dataset settings. As shown in Fig. 3, the model consistently highlights newly constructed buildings across diverse urban scenes, demonstrating robust generalization without target-domain fine-tuning.

Qualitative Strengths. A key advantage of FIFA-RS is its ability to preserve fine-grained structural boundaries. As shown by the anomaly heatmaps in Fig. 3, the predicted regions are spatially compact and closely aligned with ground-truth building footprints, even under cluttered backgrounds and complex urban layouts. This indicates that the proposed structural adaptation strategy improves sensitivity to local geometric discontinuities.

In particular, token-level high-pass adaptation together with image-only 2D spatial high-pass enhancement emphasizes both relative feature differences and spatial boundary cues, producing clearer building contours rather than diffuse or fragmented responses. These observations support our hypothesis that structural adaptation improves precise anomaly localization in high-resolution remote sensing imagery.

Failure Cases and Structural Sensitivity. Despite its strong overall performance, FIFA-RS remains challenged by very small building footprints and heavy shadow occlusion. In such cases, anomaly responses may become weaker or fragmented. Importantly, the model rarely produces large or chaotic false-positive regions.

This behavior reflects a desirable property of the proposed design: **structural sensitivity**. By emphasizing local geometric cues through token-level residual enhancement and spatial high-pass refinement, FIFA-RS reduces over-reliance on low-level texture or background appearance. As a result, erroneous activations tend to remain localized and structurally coherent.

Limitations of Single-Temporal Reasoning. A limitation of the single-temporal anomaly reasoning paradigm is that, in datasets such as **LEVIR-CD**, the model may occasionally highlight pre-existing buildings as anomalies. Since FIFA-RS does not access pre-event (T_1) imagery at inference time, prominent man-made structures may be interpreted as anomalous relative to the surrounding natural background.

Such false positives should therefore be viewed as a limitation of the single-temporal formulation rather than a contradiction of the proposed method. At the same time, this behavior highlights the effectiveness of token-level high-pass adaptation and image-only 2D spatial high-pass enhancement for delineating complex building boundaries. Incorporating an explicit T1 reference or temporal consistency constraints remains a promising direction for future work.

Ablation Insights. The qualitative observations in Fig. 3 are consistent with the quantitative ablation in Table 2. In particular, the large gap between **Vanilla CLIP** and the Text Adapter setting confirms the importance of semantic alignment, whereas the further improvement from the **Text Adapter** setting to the **Full Model** demonstrates that image-side structural adaptation is critical for precise anomaly localization.

5. **Conclusion**

In this paper, we presented **FIFA-RS**, a zero-shot framework for **single-temporal structural anomaly reasoning** in remote sensing. By reformulating building change detection as a structural anomaly localization task, we eliminated the dependency on co-registered bi-temporal image pairs at inference time. Instead of relying on explicit frequency-domain decomposition, our approach adopts a lightweight structural adaptation framework that combines token-level high-pass adaptation, image-only 2D spatial high-pass enhancement, and learnable multi-scale fusion. The resulting framework effectively bridges the gap

between natural-image pretraining and remote sensing structural reasoning without requiring localized fine-tuning or historical references.

Our results, particularly the **94.91% Pixel AUC** and **61.60% F1-score** achieved in zero-shot cross-dataset transfer, demonstrate that lightweight structural adaptation improves both boundary-aware localization and cross-domain robustness. More broadly, our results indicate that, under severe domain shift and without temporal references, structural adaptation is a more effective inductive bias than heavy task-specific fine-tuning. This property makes FIFA-RS a practical solution for rapid remote sensing monitoring in real-world deployment scenarios.

Future Work. To address the inherent ambiguity of single-temporal reasoning in static urban environments, we aim to extend FIFA-RS into a bi-temporal (A-B) comparison framework, using historical baselines to further suppress false positives from pre-existing structures. In addition, integrating structural adaptation with representation learning or reconstruction-based anomaly modeling [19, 20] may further improve robustness in complex remote sensing environments. Additionally, we plan to conduct broader cross-dataset evaluations on benchmarks like S2Looking [26] to further examine the robustness of the proposed structural adaptation modules across varying sensor resolutions and environmental conditions.

References

[1] H. Jiang, M. Peng, H. Zhong Y.and Xie, Z. Hao, J. Lin, X. Ma, and X. Hu. "A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images". In: *Remote Sensing* 14.7 (2022), p. 1552.

[2] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan. "Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges". In: *Remote Sensing* 12.10 (2020), p. 1688.

[3] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, H. Zhao, Q. Zhao, and S. Xiang. "Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review". In: *Remote Sensing* 16.13 (2024), p. 2355.

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

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

[6] W. Ma, X. Zhang, Y. Li, Q. Yao, F. Tang, C. Wu, R. Yan, Z. Jiang, and S. K. Zhou. "AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP". In: *arXiv preprint arXiv:2503.06661* (2025).

[7] S.-C. Lin and S.-H. Lai. "Clip-fsqae: Clip-guided finite scalar quantized autoencoder for few-shot anomaly detection". In: *2025 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2025, pp. 313–318.

[8] S. C. Lin, H.-W. Lee, Y.-S. Hsieh, C. Y. Ho, and S.-H. Lai. "Masked Attention ConvNeXt Unet with Multi-Synthesis Dynamic Weighting for Anomaly Detection and Localization." In: *BMVC*. 2023, p. 911.

[9] K. Amini, Y. Liu, J. E. Padgett, et al. "Debris segmentation using post-hurricane aerial imagery". In: *Computer-Aided Civil and Infrastructure Engineering* 40.25 (2025), pp. 4116–4131.

[10] G. Wang, S. Y. Shin, and G. Jo. "An improved post-hurricane building damaged detection method based on transfer learning". In: *Indonesian Journal of Electrical Engineering and Computer Science* 33.3 (2024), pp. 1546–1556.

[11] R. C. Daudt, B. L. Saux, and A. Boulch. "Fully Convolutional Siamese Networks for Change Detection". In: IEEE, 2018, pp. 4063–4067.

[12] W. G. C. Bandara and V. M. Patel. "A Transformer-Based Siamese Network for Change Detection". In: *CoRR* abs/2201.01293 (2022). DOI: 10.48550/arXiv.2201.01293.

[13] D. Tuia, C. Persello, and L. Bruzzone. "Recent Advances in Domain Adaptation for the Classification of Remote Sensing Data". In: *IEEE Geoscience and Remote Sensing Magazine* 4.2 (2016), pp. 41–57. DOI: 10.1109/MGRS.2016.2548504.

[14] C. Liang, W. Li, Y. Dong, and W. Fu. "Single Domain Generalization Method for Remote Sensing Image Segmentation via Category Consistency on Domain Randomization". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–16. DOI: 10.1109/TGRS.2024.3379669.

[15] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou. "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–16. DOI: 10.1109/TGRS.2024.3390838.

[16] V. V. Cepeda, G. K. Nayak, and M. Shah. "GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization". In: *CoRR* abs/2309.16020 (2023). DOI: 10.48550/arXiv.2309.16020.

[17] Q. Cao, Y. Chen, C. Ma, and X. Yang. "Open-Vocabulary Remote Sensing Image Semantic Segmentation". In: *CoRR* abs/2409.07683 (2024). DOI: 10.48550/arXiv.2409.07683.

[18] Y. Zhang and Z. Gao. "RSAD-CLIP: Zero-Shot Remote Sensing Anomaly Detection of the Earth's Surface Based on Pre-Trained Vision-Language Model". In: IEEE, 2025, pp. 1–5.

[19] S.-C. Lin and S.-H. Lai. "SQUAD: Scalar Quantized Representation Learning for Unsupervised Anomaly Detection and Localization". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2024.

[20] S.-C. Lin and S.-H. Lai. "LFQUIAD: Lookup-Free Quantized autoencoder for few-shot Unsupervised Industrial Anomaly Detection via Synthetic Diffusion Inpainting". In: *Synthetic Data for Computer Vision Workshop@ CVPR 2025*. 2025.

[21] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. C. Courville. "On the Spectral Bias of Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 5301–5310.

[22] Y. Yang and S. Soatto. "FDA: Fourier Domain Adaptation for Semantic Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 3996–4005.

[23] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan. "Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 5886–5895.

[24] H. Chen and Z. Shi. "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection". In: *Remote Sensing* 12.10 (2020), p. 1662.

[25] S. Ji, S. Wei, and M. Lu. "Fully Convolutional Networks for Multi-Source Building Extraction From an Open Aerial and Satellite Imagery Dataset". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1 (2018), pp. 574–586. DOI: 10.1109/TGRS.2018.2858817.

[26] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang. "S2Looking: A satellite side-looking dataset for building change detection". In: *Remote Sensing* 13.24 (2021), p. 5094.