Modular Segmentation-Agnostic Framework for Object-Based Change Detection in Multi-Temporal VHR Imagery

Abdul-Rashid Zakaria[1], **Teresa DiMeola**[1], **Charles Walter**[1], **Pasi Lautala**[2], **Thomas Oommen**[3], **Hong Xiao**[1]

[1]University of Mississippi, Department of Computer and Information Science, Oxford, USA
[2]Michigan Technological University, Department of Civil, Environmental, and Geospatial Engineering, Houghton, USA
[3]University of Mississippi, Department of Geology & Geological Engineering, Oxford, USA
{azakaria, tjdimeol}@go.olemiss.edu, {cwwalter, toommen, hxiao1}@olemiss.edu, ptlautal@mtu.edu,

Abstract

Detecting spatiotemporal anomalies that threaten critical transportation infrastructure, such as railway corridors, is essential for predictive maintenance and public safety. The growing availability of very high resolution (VHR) imagery from satellites and uncrewed aerial vehicles (UAVs) provides rich observational data, yet effectively utilizing these unlabeled images remains challenging due to illumination variability, noise, and the need for lightweight, real-time analysis on limited hardware. This paper presents a modular and segmentation-agnostic anomaly detection framework for multi-temporal VHR imagery that operates in a fully unsupervised and self-supervised setting. The framework integrates spatial and temporal reasoning through a hybrid architecture that combines segmentation, feature extraction, and change vector analysis to identify anomalous regions. A region correlation matrix (RCM) is introduced as an interpretable and label-free evaluation metric that quantifies segmentation consistency and anomaly structure, providing a foundation for explainable AI in unsupervised settings. Extensive experiments on UAV and satellite datasets of railway corridors and flood-affected regions demonstrate the framework's scalability, robustness, and cross-domain feasibility. Its modular design allows seamless integration of foundation and transformer-based vision-language models, enabling continual adaptation to evolving data distributions. The proposed approach thus contributes a flexible and robust solution for spatiotemporal anomaly detection in real-world infrastructure monitoring and other remote-sensing domains.

1 Introduction

We introduce a modular, segmentation-agnostic framework for unsupervised anomaly detection in multi-temporal very-high-resolution (VHR) imagery. The pipeline decouples segmentation, region-level feature extraction, and thresholding, enabling plug-and-play backbones and joint spatial–temporal reasoning without retraining. We further propose a self-supervised Region Correlation Matrix (RCM) with overlap/fragmentation/composite indices for label-free, interpretable assessment of segmentation consistency and anomaly structure (Sikka and Deserno 2010).

Detecting subtle, safety-critical changes in rail corridors and disaster zones is essential for predictive maintenance and response, yet labels are scarce and deployments often run on limited hardware. Classical pixel- and object-based approaches can be noise-sensitive or brittle across scenes, while recent deep models (e.g., transformer-based change detectors) often require dense supervision and substantial compute (Bandara and Patel 2022; Chen, Qi, and Shi 2022; Miao et al. 2024). Our design targets label-free, resource-aware detection robust to illumination, sensor, and domain shifts from UAV to satellite imagery.

Our method sits between pixel-based and object-based change detection: we stabilize geometry with object segmentation using classical methods such as (Felzenszwalb and Huttenlocher 2004; Comaniciu and Meer 2002; Vedaldi and Soatto 2008) and transformer-based architectures (Kirillov et al. 2023; Wu and Osco 2023), compute region-wise features and change vectors over time, and fuse them into a coherent change map.

The key contributions of this work are as follows:

- We introduce a modular and segmentation-agnostic anomaly detection framework for multi-temporal very high-resolution imagery. The framework unifies spatial and temporal analysis in an unsupervised manner and supports plug-and-play integration of new segmentation, feature extraction, and thresholding modules without retraining, enabling continual adaptation to evolving data.
- We propose a region correlation matrix (RCM) as a self-supervised evaluation metric that quantifies segmentation consistency and anomaly structure without ground truth, offering an interpretable and explainable basis for spatiotemporal anomaly reasoning.
- We demonstrate the scalability and cross-domain adaptability of the framework on UAV and satellite datasets, showing robust performance in resource-limited settings and across heterogeneous spatial domains, with applications in predictive maintenance and environmental risk monitoring.

2 Related Work

Change detection traditionally comprises methods that use each pixel in an image as the basic unit of analysis. These methods are commonly referred to as pixel-based change detection (PBCD) methods. With improved computation power and very high spatial resolution images, efforts have

evolved to aggregate pixels to form objects and extract features from these objects for change detection. These methods are known as object-based image analysis (OBIA). Object-based change detection (OBCD) is an extension of the OBIA, and involves extracting objects from images through segmentation, which assigns every pixel in an image to a label (Felzenszwalb and Huttenlocher 2004; Linares et al. 2017).

2.1 Pixel-Based Change Detection (PBCD)

PBCD methods can be implemented with machine learning techniques, both supervised and unsupervised. Recent PBCD methods utilize deep learning to generate feature maps for each image and compare the resulting feature maps across two or more temporal images of the same area of interest. A basic premise of most deep learning based PBCD methods is that image pixels can be classified into two categories: changed or unchanged, with methods proposed to create binary and semantic change detection using very high-resolution (VHR) Earth observations (Daudt et al. 2019).

Other methods include the deep slow feature analysis in which two deep neural networks are used to extract and represent the features of temporal images (Du et al. 2019). Deep learning-based methods generally require large datasets for training and high computational resources. In addition, supervised learning methods usually require high-quality ground truth labels, which is often tedious for computer vision tasks such as segmentation (Bansal, Vaid, and Gupta 2022) due to the dense maps required. For VHR imagery, PBCD methods are susceptible to noisy change pixels due to different illuminations and large reflectance that can be caused by different acquisition characteristics in temporal images, as well as image registration issues. This defect tends to result in a 'salt and pepper' change map, which could yield inaccurate detection results (Hussain et al. 2013; Niemeyer, Marpu, and Nussbaum 2008).

2.2 Object-Based Change Detection (OBCD)

OBCD methods, on the other hand, involve segmenting temporal images independently and comparing the resulting image objects. In addition to the comparison of objects' geometries, other properties such as compactness, texture, and spectral features can also be compared. The quality of the change detection depends on the segmentation quality, which could be impacted by common variations in sensor deployment conditions, illumination, and other atmospheric factors. For example, a common method of comparing the segmented features is the spatial comparison of related objects; however, for smaller objects generated from inconsistent segmentation, object 'slivers' are created, which can complicate the comparison, potentially reducing the usefulness of OBCD. A solution to this problem typically involves removing smaller objects using a threshold or knowledge of the actual ground condition, which can result in possible information loss (Boldt, Thiele, and Schulz 2012). OBCD methods, however, use contextual and feature information, which help in overcoming the 'salt and pepper' issue inherent to most PBCD methods.

From the above discussions, we can see that PBCD and OBCD methods are best suitable for different change detection tasks, depending on a plethora of conditions such as complexity of the algorithms, and availability of computational resources, user's needs, and the availability of training and testing datasets. Ultimately, the choice between the two approaches should also consider the model's scalability, reliability, and robustness in production environments.

2.3 Segmentation Algorithms Used in OBCD

We examine five total segmentation algorithms, three classical algorithms (Felzenszwalb, Mean Shift, and Quick Shift) and two advanced transformer-based models (SAMGeo and RESDA). These algorithms are discussed below.

Felzenszwalb was created in 1999 to separate images into regions via a greedy algorithm (Felzenszwalb and Huttenlocher 2004). This algorithm takes a graph-based approach to segmentation with the pixels as the set of elements to be segmented as vertices and edges connecting neighboring vertices. Each edge has a weight that measures the dissimilarity between two pixels (vertices), and the computational complexity is low, being linear to the number of graph edges.

Mean Shift Segmentation is based on the mean shift filtering procedure which is a mode seeking algorithm that finds the peaks of the density distribution of the image pixels. It is an adaptive gradient ascent method where regions of low-density values are negligible in the feature space and hence in such sparse areas mean shift steps are large (Comaniciu and Meer 2002). The mean shift segmentation involves an iterative search within the image for peaks. This algorithm requires an initial bandwidth or window size to be selected, and it is relatively computationally inexpensive and simple to implement.

Quick Shift produces superpixels by clustering pixels with similar color and spatial proximity. These superpixels, provide a dense overcomplete representation ofhte image that captures local details (Ibrahim and El-kenawy 2020). (Vedaldi and Soatto 2008) introduced quick shift as a competitive segmentation algorithm that can balance under- and over-fragmentation of clusters by the choice of a real parameter, which serves as a model selection threshold. Quick shift is relatively simple and faster than the mean shift algorithm.

Resource Efficient Domain Specialization: A Dual Zero-shot, Fine-tune Strategy (RESDA) (Anonymous 2025) introduces a resource-efficient approach to zero-shot segmentation in specialized domains such as aerial or medical imagery, where access to large foundation models or labeled datasets is limited. It utilizes a general-purpose vision-language model architecture, OFA (Wang et al. 2022), which is fine-tuned on small, domain-specific datasets comprising as few as 200 annotated images. During inference, RESDA can identify over 100 contextually relevant classes using a lightweight semantic token expansion strategy. Unlike traditional models, RESDA is optimized for inference on a single GPU and supports dynamic segmentation across datasets through zero-shot learning (ZSL). Its low computational demands and domain adaptability make it a compelling option for time-sensitive applications like disaster monitoring or rapid remote sensing analysis.

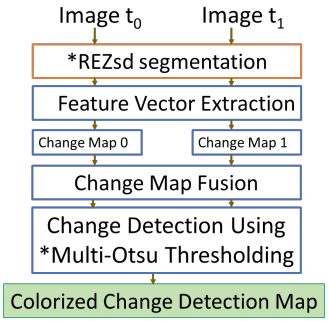

Figure 1: Proposed modular change-detection workflow (segmentation and thresholding methods are interchangeable).

Segment Anything Geospatial (SAMGeo) Segment Anything Model (SAM) is a transformer-based model developed by Meta (Kirillov et al. 2023), trained on more than one billion masks from eleven million images, and is capable of performing zero-shot inferences on new images from different domains. Wu and Osco created SAMGeo to leverage SAM for geospatial data analysis, and designed the system to require minimal coding effort, if so desired.

3 Proposed Framework

Our objective is to develop a framework capable of accurately detecting changes in multi-temporal VHR heterogeneous data. To achieve this, we propose a hybrid object-based change-detection methodology. The workflow consists of three main components:

1. **Image Segmentation:** partitioning VHR imagery into coherent regions representing distinct landscape or infrastructure elements or specific artifacts of interest depending on the application domain.
2. **Feature Extraction and Change Vector Analysis:** extracting informative attributes from each region and analyzing temporal differences to detect meaningful changes.
3. **Fusion and Change-Map Generation:** integrating temporal changes into a single coherent map highlighting significant differences across temporal images.

3.1 Segmentation Algorithms

Segmentation converts VHR input images into structured sets of meaningful regions or objects such as railway tracks, vegetation patches, rock formations, or damaged infrastructure areas. Each segmented object groups together pixels with similar visual characteristics such as color and texture, delineating areas where environmental changes occur. High-quality segmentation ensures precise spatial delineation and reduces erroneous or fragmented results. Because ground-truth segment labels are rarely available in remote-sensing tasks, segmentation quality is assessed *indirectly* through comparative evaluation of different algorithms.

Region Correlation Matrix (RCM) and Derived Indices
Segmentation and region sets Let $\Omega \subset \mathbb{Z}^2$ denote the image domain, and at time $t \in \{t_0, t_1\}$ let the segmentation be a finite partition $S_t = \{R_i^{(t)}\}_{i=1}^{n_t}$ of nonempty, pairwise-disjoint regions that cover the valid subset $\Omega_v^{(t)} = \{x \in \Omega : v^{(t)}(x) = 1\}$. Here $v^{(t)} : \Omega \to \{0,1\}$ is a binary validity mask indicating the reliability of a pixel with $v^{(t)}$ being 1 indicating a valid pixel, and 0 for an invalid or missing pixel. Each segmentation can equivalently be represented by a label map $\ell^{(t)} : \Omega \to \{1, \dots, n_t, \perp\}$ with region weights $w_i^{(t)} = |R_i^{(t)}|/N^{(t)}$, and all computations are restricted to the common valid domain $\Omega_c = \Omega_v^{(t_0)} \cap \Omega_v^{(t_1)}$ for consistent normalization.

Given two segmentation maps S_1 and S_2 with disjoint region sets $\{R_i^{(1)}\}_{i=1}^{n_1}$ and $\{R_j^{(2)}\}_{j=1}^{n_2}$, the Region Correlation Matrix (RCM) quantifies spatial overlap (Sikka and Deserno 2010) as

$$M_{ij} = |R_i^{(1)} \cap R_j^{(2)}|,$$

where $|\cdot|$ denotes the number of pixels. The normalized matrix is $\tilde{M}_{ij} = M_{ij}/N$ with $N = \sum_{i,j} M_{ij}$ the total pixel count. Row and column sums satisfy $\sum_j \tilde{M}_{ij} = |R_i^{(1)}|/N$ and $\sum_i \tilde{M}_{ij} = |R_j^{(2)}|/N$.

Overlap index.

$$E(S_1, S_2) = \sum_{i=1}^{n_1} \max_j \tilde{M}_{ij}, \quad 0 \le E \le 1.$$

Higher E implies stronger spatial alignment between the two maps.

Symmetric Overlap Index. The overlap index $E(S_1, S_2)$ measures how well regions in S_1 align with those in S_2 and is therefore *directional*. To obtain a symmetric measure of mutual alignment, we define

$$E_{\text{sym}}(S_1, S_2) = \tfrac{1}{2}\big(E(S_1, S_2) + E(S_2, S_1)\big), \quad 0 \le E_{\text{sym}} \le 1.$$

High values of E_{sym} indicate that each segmentation can be well explained by the other, reflecting strong bidirectional spatial correspondence.

Area-weighted fragmentation index. Let $w_i = |R_i^{(1)}|/N$ denote the relative area of region $R_i^{(1)}$, and let p_i be the number of nonzero entries in row i of $\tilde{M}$, corresponding to the number of regions in S_2 that overlap $R_i^{(1)}$. Define the per-row concentration ratio

$$m_i = \begin{cases} \dfrac{\max_j \tilde{M}_{ij}}{w_i}, & \text{if } w_i > 0, \\ 0, & \text{otherwise,} \end{cases} \quad 0 \le m_i \le 1.$$

The area-weighted fragmentation index is then

$$F_{\text{aw}}(S_1, S_2) = \begin{cases} \dfrac{\sum_i w_i (p_i - 1)[1 - m_i]}{\sum_i w_i (p_i - 1)}, & \text{if } \sum_i w_i (p_i - 1) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$0 \le F_{\text{aw}} \le 1. \tag{1}$$

This formulation weights each region by its area fraction w_i, penalizing fragmentation of large regions more heavily while normalizing by each region's own mass through m_i. Lower F_{aw} indicates more coherent and less fragmented correspondence between maps.

Area-weighted composite dissimilarity. To jointly capture overlap mismatch and fragmentation, we define

$$G_{\mathrm{aw}}(S_1, S_2) = \frac{\alpha\, F_{\mathrm{aw}}(S_1, S_2) + \beta\,[1 - E(S_1, S_2)]}{\alpha + \beta},$$
$$\alpha = \beta = 0.5,\ 0 \le G_{\mathrm{aw}} \le 1.$$

This index balances alignment and fragmentation into a single normalized measure, providing a compact, label-independent summary of segmentation consistency for unsupervised model selection.

Symmetric composite form. For direction-agnostic comparison, both E and F_{aw} can be averaged across directions:

$$F_{\mathrm{aw,sym}} = \tfrac{1}{2}\big(F_{\mathrm{aw}}(S_1, S_2) + F_{\mathrm{aw}}(S_2, S_1)\big),$$
$$G_{\mathrm{aw,sym}} = \frac{\alpha\, F_{\mathrm{aw,sym}} + \beta\,[1 - E_{\mathrm{sym}}]}{\alpha + \beta}.$$

These symmetric variants ensure that both segmentation maps contribute equally to the overall similarity assessment.

3.2 Feature Extraction and Change Vector Analysis

Region pairing (overlay). Let S_1 and S_2 be the segmentations at times t_0 and t_1 with region sets $\{R_i^{(1)}\}_{i=1}^{n_1}$ and $\{R_j^{(2)}\}_{j=1}^{n_2}$. We form the overlay partition

$$\mathcal{O} = S_1 \otimes S_2 = \big\{\, O_{ij} := R_i^{(1)} \cap R_j^{(2)} \;\big|\; |O_{ij}| > 0 \,\big\},$$

which partitions the image domain into atomic cells. Each O_{ij} provides a one-to-one locus for change computation and inherits its parent regions' attributes.

Features. For each O_{ij} we compute a K-dimensional feature vector $\mathbf{f}_{ij}^{(t)} \in \mathbf{R}^K$ at time $t \in \{t_0, t_1\}$. In this work K includes texture (GLCM contrast, homogeneity, dissimilarity at a fixed offset), gradient statistics (mean and variance of Sobel magnitude), shape/geometry (e.g., area, compactness of O_{ij}), and band/spectral indices when available. All features are standardized with a single z-score per feature using parameters (μ, σ) estimated from the pooled set $\{\mathbf{f}_{ij}^{(t_0)}\} \cup \{\mathbf{f}_{ij}^{(t_1)}\}$, and then applied to both times to preserve comparability.

Change Vector Analysis (CVA). The per-cell change magnitude is

$$c_{ij} = \big\|\mathbf{f}_{ij}^{(t_1)} - \mathbf{f}_{ij}^{(t_0)}\big\|_2,$$
$$\text{(optionally)}\quad c_{ij} = \big\|\mathbf{W}\big(\mathbf{f}_{ij}^{(t_1)} - \mathbf{f}_{ij}^{(t_0)}\big)\big\|_2 \text{ with diagonal } \mathbf{W}.$$

We form a single dense change map by assigning

$$C(x) = c_{ij} \qquad \text{for all } x \in O_{ij}.$$

3.3 Fusion and Change Map

In the absence of missing data, we simply set $C^* = C$. If feature availability differs across times (e.g., due to clouds or field-of-view (FOV) differences), let $v^{(t)}(x) \in \{0, 1\}$ indicate validity at time t and define

$$C^*(x) = \begin{cases} C(x), & v^{(t_0)}(x) = v^{(t_1)}(x) = 1, \\ \gamma\, C(x), & v^{(t_0)}(x) \oplus v^{(t_1)}(x) = 1, \\ 0, & \text{otherwise}, \end{cases}$$

with a small $\gamma \in (0, 1]$ if single-time evidence is to be downweighted.

Handling Missing or Invalid Data. In real-world imagery, some pixels may be unavailable or unreliable in one of the temporal acquisitions due to occlusion, sensor artifacts, or differing fields of view. To ensure that the resulting change map remains consistent and spatially coherent, we define a validity-aware fused change magnitude $C^*(x)$, which denotes the raw change magnitude at pixel x, $v^{(t_0)}(x)$ and $v^{(t_1)}(x)$ are binary validity masks for times t_0 and t_1, respectively, and $\oplus$ represents the exclusive OR (XOR) operator. The scalar $\gamma \in (0, 1]$ is a small downweighting factor that penalizes change values computed from only one valid temporal observation.

Interpretation. This formulation assigns different confidence levels to pixels depending on data availability:

- If both temporal images are valid ($v^{(t_0)} = v^{(t_1)} = 1$), the change value is fully trusted and retained as $C^*(x) = C(x)$.
- If only one time step is valid ($v^{(t_0)} \oplus v^{(t_1)} = 1$), the corresponding change magnitude is downweighted to $\gamma C(x)$ to reflect partial uncertainty.
- If neither image provides valid information, the pixel is excluded from analysis ($C^*(x) = 0$).

By integrating validity masks into the fusion process, this step prevents spurious changes caused by missing or unreliable pixels and ensures that non-overlapping image regions do not propagate false detections. In subsequent normalization and thresholding, only pixels with valid evidence contribute to the final change map, resulting in improved robustness under heterogeneous data quality or incomplete temporal coverage.

We then normalize to $[0, 255]$ using a robust percentile stretch:

$$\hat{C}(x) = 255\, \frac{\mathrm{clip}\big(C^*(x),\, P_1,\, P_{99}\big) - P_1}{P_{99} - P_1 + \varepsilon},$$

where P_1 and P_{99} denote the 1st and 99th percentiles of C^* over the image and $\varepsilon > 0$ avoids division by zero. A global Otsu (two classes) or Multi-Otsu (with classes merged to a binary label) threshold converts $\hat{C}$ to a binary change mask.

4 Experimental Setup

To showcase the general applicability of our framework, we selected orthophotos acquired from two distinct UAV data collection campaigns, conducted approximately one year

apart. The areas chosen for analysis include railway corridors in Montana and Michigan, providing a realistic representation of the geomorphic, vegetation, and infrastructure conditions. Each image covers more than 78,000 square meters (sqM) of highly detailed terrain.

Ten representative image pairs were deliberately selected from this broader dataset (shown in Figure 2), chosen specifically to include a wide variety of observable and subtle changes. These pairs capture instances such as rockfalls, vegetation growth, ballast displacement, as well as scenes with minimal change.

The images were carefully cropped to focus on key infrastructure elements, such as railroad tracks, embankments, adjacent slopes and areas susceptible to environmental disturbances. These smaller image subsets facilitate rapid model development and robust comparative evaluations while preserving the full-resolution detail required for accurate segmentation and change detection analyses.

To standardize comparisons across the image pairs, all images were resampled to a uniform spatial resolution of 0.010 m/pixel. However, subsequent segmentation and feature extraction experiments were performed using the original full-resolution data to ensure preservation of fine-scale details crucial for precise change detection. A summary of the properties of each selected image pair is provided in Table 1.

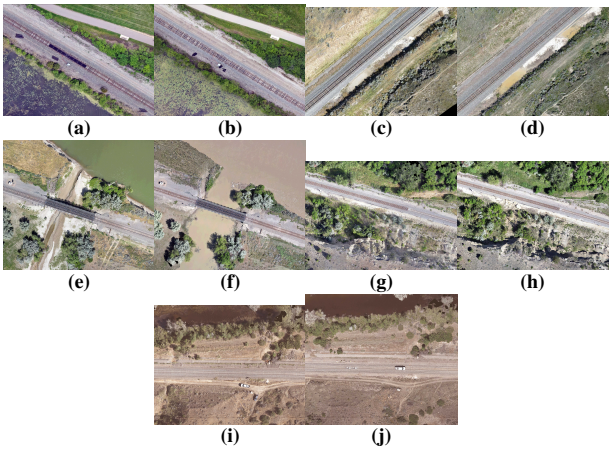

Figure 2: Multi-temporal images used in developing the algorithm **(a)** Ann Arbor 2022 **(b)** Ann Arbor 2023 **(c)** Marsh 2022 **(d)** Marsh 2023 **(e)** Hathaway South 2022 **(f)** Hathaway South 2023 **(g)** Rosebud 2022 **(h)** Rosebud 2023 **(i)** Hathaway North 2022 **(j)** Hathaway North 2023

5 Results

Evaluation Without Ground Truth. Since no annotated ground-truth change maps currently exist for the UAV or satellite datasets, the framework is evaluated in a fully unsupervised setting. We therefore focus on (i) label-independent quantitative metrics that measure segmentation stability and internal consistency, and (ii) qualitative validation based on physically interpretable changes observed in the imagery. This approach follows established unsupervised evaluation practices in object-based remote sensing (Hussain et al. 2013; Niemeyer, Marpu, and Nussbaum 2008), ensuring reproducibility without requiring pixel-wise reference data.

Image Pair	Resolution (m)	Approx. Area (sqM)
Ann Arbor	0.015	2,180
Hathaway	0.0077	10,350
Hathaway North	0.0076	21,278
Marsh	0.0078	24,360
Rosebud	0.015	2,641

Table 1: Summary of UAV image pairs used in the change detection experiments. Each pair covers key railroad corridor sites with different spatial resolutions and ground coverage.

5.1 Segmentation Results

This section compares the segmentation results obtained from the five segmentation methods. Rosebud segmentation results are discussed below; similar observations can be observed for other image pairs. Supplementary Material Section C contains the segmented images for all other areas of interest.

Visual Comparison. Visually comparing the results of each segmentation (Figure 3), SAMGeo results are visually clearer and tends to produce accurate masks of each object's instance. However, it tends to not segment ballasts along the right of way close to the rail tracks. The second-best result is from RESDA, which inclines to create fewer regions but captures, accurately, the key features such as the rockfall in 2023. The third-best result is from the Felzenszwalb algorithm, although its segmentation results tend to provide inaccurate instances when visually compared to the original image. In addition, most instances of the same object are segmented differently in terms of the shape of the masks, indicating the algorithm's inability to clearly differentiate objects such as trees, shrubs, ties, and ballast. Quick Shift and Mean Shift results are comparable and highly inaccurate. Quick shift tends to create a high number of super pixels, which is expected for a low-level segmentation method.

Similarity/Dissimilarity Evaluation Similarity/Dissimilarity Evaluation. Using the indices (overlap (E), fragmentation (F), and composite dissimilarity (G)), a pairwise comparison of the Rosebud image segmentations is presented in Table 2. These symmetric distances are computed by averaging both directions of comparison ("reference to comparison" and "comparison to reference") for each index, for example $E(S_1, S_2)$ and $E(S_2, S_1)$, where S_1 is the reference segmentation and S_2 is the comparison segmentation. A **lower** G value indicates stronger similarity between the segmentations, while higher values imply structural divergence.

The fragmentation index F measures the degree to which regions in one segmentation are split across multiple labels in another. A lower F denotes greater consistency (that is, each region in one map corresponds largely to a single region in the other), whereas a higher F signifies over segmentation or inconsistent labeling. From Table 2, the pair **RESDA and SAMGeo** exhibits the lowest fragmentation

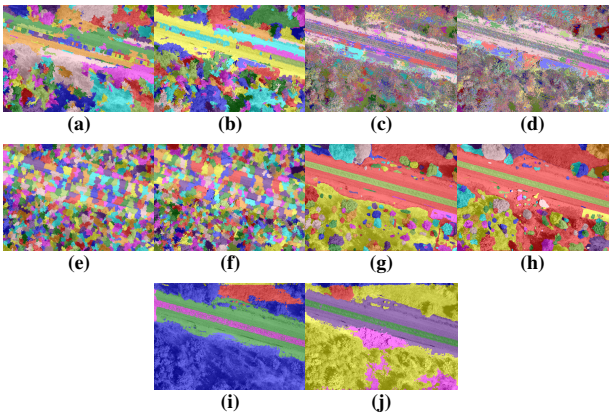

Figure 3: Rosebud segmented images using different algorithms: **(a)** Felzenszwalb 2022, **(b)** Felzenszwalb 2023, **(c)** Mean Shift 2022, **(d)** Mean Shift 2023, **(e)** Quick Shift 2022, **(f)** Quick Shift 2023, **(g)** SAMGeo 2022, **(h)** SAMGeo 2023, **(i)** RESDA 2022, **(j)** RESDA 2023.

indices across both temporal instances ($F_{\text{sym}}^{t_0} = \mathbf{0.3377}$ and $F_{\text{sym}}^{t_1} = \mathbf{0.5090}$), indicating that their segmentations maintain high internal coherence with minimal splitting. Conversely, the pair **Quick Shift and Mean Shift** yields the highest fragmentation values ($F_{\text{sym}}^{t_0} = 0.8173$, $F_{\text{sym}}^{t_1} = 0.8302$), reflecting severe over segmentation of Mean Shift regions by Quick Shift. This is consistent with Quick Shift's tendency to generate dense, fine grained superpixels that subdivide the broader homogeneous regions produced by Mean Shift.

A	B	$E_{\text{sym}}^{t_0}$	$F_{\text{sym}}^{t_0}$	$G_{\text{sym}}^{t_0}$	$E_{\text{sym}}^{t_1}$	$F_{\text{sym}}^{t_1}$	$G_{\text{sym}}^{t_1}$
Fel	MS	0.2370	0.7831	0.7731	0.2380	0.7934	0.7777
Fel	QS	0.5038	0.5901	0.5432	0.5256	0.5796	0.5270
Fel	RE	0.5321	0.5237	0.4958	0.5176	0.5260	0.5042
Fel	SA	0.5475	0.5376	0.4951	0.5641	0.5404	0.4881
MS	QS	0.1994	0.8173	0.8089	0.1962	0.8302	0.8170
MS	RE	0.4876	0.5294	0.5209	0.4779	0.5428	0.5324
MS	SA	0.3565	0.6735	0.6585	0.3699	0.6591	0.6446
QS	RE	0.4716	0.5729	0.5506	0.4733	0.5589	0.5428
QS	SA	0.4909	0.5931	0.5511	0.5209	0.5720	0.5256
RE	SA	**0.7521**	**0.3377**	**0.2928**	**0.6288**	**0.5090**	**0.4401**

Table 2: Pairwise segmentation comparison across algorithms using symmetric region correlation metrics at times t_0 and t_1. Green cells denote higher similarity (larger E_{sym}, smaller F_{sym}, G_{sym}), while red cells indicate weaker correspondence. The best-performing pair is **RESDA–SAMGeo**, consistent across both temporal datasets.

The overlap index E quantifies the spatial alignment between two segmentation maps. Higher values indicate stronger regional correspondence, whereas lower values imply that objects are fragmented or displaced across boundaries. As shown in Table 2, the pair **RESDA and SAMGeo** achieves the highest overlap scores at both times ($E_{\text{sym}}^{t_0} = \mathbf{0.7521}$, $E_{\text{sym}}^{t_1} = \mathbf{0.6288}$), demonstrating robust mutual alignment and confirming that both models capture similar region boundaries and object structures. In contrast, the pair **Quick Shift and Mean Shift** again shows the weakest alignment ($E_{\text{sym}}^{t_0} = 0.1994$, $E_{\text{sym}}^{t_1} = 0.1962$), underscoring their incompatibility in region delineation.

When interpreted jointly, the indices E, F, and G provide complementary insights into segmentation similarity. E reflects the global spatial alignment, F captures structural coherence, and G summarizes both effects into a unified dissimilarity score. Among all algorithm pairs, **RESDA and SAMGeo** consistently achieves the lowest composite distances ($G_{\text{sym}}^{t_0} = \mathbf{0.2928}$, $G_{\text{sym}}^{t_1} = \mathbf{0.4401}$), indicating a high level of mutual consistency and stability across time. This strong agreement may be attributed to their transformer based architectures, which produce semantically rich and spatially coherent regions. Conversely, **Mean Shift** appears in several high G pairs, reinforcing its reputation as the least consistent or desirable choice under fine grained, high resolution settings.

Overall, the results demonstrate that pairs involving deep or zero shot segmentation models (RESDA and SAMGeo) yield the most stable and interpretable results, while purely classical methods (Mean Shift, Quick Shift, and Felzenszwalb) show higher variability and weaker correspondence. These findings confirm that the proposed RCM based indices serve as reliable, label independent measures for assessing segmentation quality, enabling objective model selection in the absence of ground truth annotations.

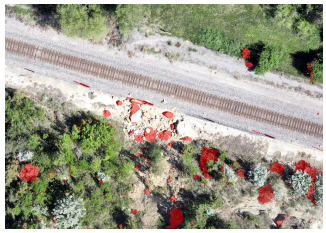

Figure 4: Detected change regions (in red) overlaid on the 2023 Rosebud UAV orthophoto. The SAMGeo segmentation and change vector analysis successfully delineate rockfall debris and hillside displacement adjacent to the railway corridor, while smaller vegetation variations are also highlighted. The results demonstrate the framework's ability to capture both large structural and subtle surface changes without labeled supervision.

5.2 Change Detection

After fusing the two change-magnitude maps, a binary change map (Figure 5) is generated to represent changed and unchanged pixels. In Figure 5 the top row shows the pre-event (*T0*) and post-event (*T1*) satellite images, the stretched CVA magnitude, and the resulting overlay using Multi-Otsu segmentation. Subsequent rows illustrate binary change masks produced by different thresholding algorithms, including global (Otsu, Triangle, Yen, Isodata, Mean, Median, Minimum), local (Niblack, Sauvola), adaptive, and manual thresholds. Global methods such as Otsu, Triangle, and Multi-Otsu capture the main flooded extent with ∼15–17% detected change, while adaptive and local methods (Sauvola, Niblack) tend to over-segment urban areas

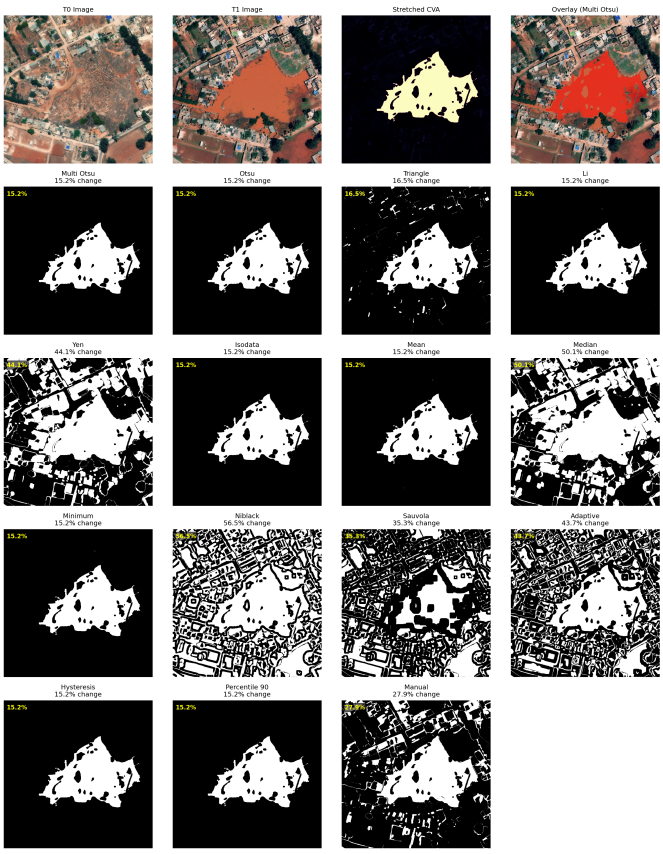

Figure 5: Comparison of global, local, and adaptive thresholding techniques applied to the 2023 Libya flood event using the stretched Change Vector Analysis (CVA) map.

with noisy boundaries. These results demonstrate that Multi-Otsu provides a balanced delineation of the inundation region with low false positives, validating its use as the default fusion threshold in the proposed change-detection framework. We benchmarked several histogram and region-based thresholding approaches and finally adopted the Multi-Otsu algorithm as the default, owing to its favorable accuracy. The multi-Otsu algorithm, an extension of Otsu thresholding, is a histogram-based algorithm that requires global image statistics, making it faster than adaptive local methods without a significant reduction in accuracy (Otsu et al. 1975). Although more sophiscated techniques exist (Ng 2006; Liao et al. 2001; Gurung and Tamang 2019; Huang and Wang 2009) they typically demand extensive hyperparameter tuning and have high computational costs. Multi-Otsu mitigates these burdens by automatically selecting thresholds that minimize intra-class variance or equivalently maximize inter-class variance in the change magnitude distribution.

The qualitative results in Figure 4 illustrate the framework's effectiveness. Using SAMGeo for segmentation, the algorithm delineates the 2023 Rosebud rockslide clearly. However, it also captures small and insignificant changes in the ballast, which may not be important depending on actual ground conditions. It uniquely detects the unstable hillside rock zones, which are confirmed by field reports, indicating the framework's sensitivity to subtle yet safety-critical changes. Comprehensive results of all sites, as well as further discussions of the RCM indices are provided in Supplementary Material, Sections C, D, and B, respectively.

5.3 Validation with Satellite Imagery

In designing this algorithm for change detection, we aimed to make it applicable to heterogeneous imagery of different resolutions, both pre- and post-event. We therefore tested our framework using VHR satellite imagery, including images from Maxar of a major flooding in Libya in 2023, which caused more than 6,000 deaths and the destruction of two dams, resulting in the widescale destruction of the city of Derna. For most satellite imagery, we observed that Multi-Otsu thresholding provided a better estimate of flooding, see Figure 5.

6 Conclusion

We introduced a fully unsupervised, modular object-based change detection framework that integrates multiple segmentation backends, region-level feature extraction, and a hybrid CVA plus fusion stage for multi-temporal VHR imagery. Instead of relying on labeled data, the framework evaluates segmentation stability with a Region Correlation Matrix (RCM) and three label independent indices: overlap (E), fragmentation (F), and composite dissimilarity (G).

Coupled with Multi-Otsu thresholding, the framework isolates changes such as rockfall debris and hillside erosion in UAV imagery and transfers to satellite scenes with minimal tuning.

For operations, **SAMGeo** and **Felzenszwalb** provide a good balance between detail and stability as region backbones for CVA, with **Quick Shift** serving as a high granularity reference and **RESDA** as a coarse baseline.

This design offers three principal advantages:

1. **Domain adaptability:** Because the framework is segmentation agnostic, users can select the algorithm that best captures object boundaries in their specific imagery, whether the target scene is a railway corridor, an urban setting, or a forest canopy.
2. **Scalable precision:** The modular fusion and thresholding block accommodates lightweight global thresholds for quick screening, as well as more sophisticated local or learning based methods when maximal accuracy is required.
3. **Future proofing:** As new segmentation models, feature descriptors, or thresholding techniques emerge, they can be dropped into the existing architecture with minimal integration overhead, keeping the workflow state-of-the-art.

Beyond unifying existing modules, the framework embodies conceptual novelty through its *segmentation agnostic* and *dynamically extensible* architecture. By decoupling the segmentation, feature extraction, and fusion components, it enables plug-and-play integration of advanced segmentation networks such as SNUNet3+ (Miao et al. 2024) and emerging transformer based descriptors without retraining.

References

Anonymous. 2025. Details withheld to preserve blind review.

Bandara, W. G. C.; and Patel, V. M. 2022. A Transformer-Based Siamese Network for Change Detection. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 207–210.

Bansal, P.; Vaid, M.; and Gupta, S. 2022. OBCD-HH: an object-based change detection approach using multi-feature non-seed-based region growing segmentation. *Multimedia Tools and Applications*, 81(6): 8059–8091.

Boldt, M.; Thiele, A.; and Schulz, K. 2012. Object-based urban change detection analyzing high resolution optical satellite images. In *Earth Resources and Environmental Remote Sensing/GIS Applications III*, volume 8538, 108–116. SPIE.

Chen, H.; Qi, Z.; and Shi, Z. 2022. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Comaniciu, D.; and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5): 603–619.

Daudt, R. C.; Le Saux, B.; Boulch, A.; and Gousseau, Y. 2019. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187: 102783.

Du, B.; Ru, L.; Wu, C.; and Zhang, L. 2019. Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12): 9976–9992.

Felzenszwalb, P. F.; and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International journal of computer vision*, 59: 167–181.

Gurung, A.; and Tamang, S. L. 2019. Image segmentation using multi-threshold technique by histogram sampling. *arXiv preprint arXiv:1909.05084*.

Huang, D.-Y.; and Wang, C.-H. 2009. Optimal multi-level thresholding using a two-stage Otsu optimization approach. *Pattern Recognition Letters*, 30(3): 275–284.

Hussain, M.; Chen, D.; Cheng, A.; Wei, H.; and Stanley, D. 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of photogrammetry and remote sensing*, 80: 91–106.

Ibrahim, A.; and El-kenawy, E.-S. M. 2020. Image segmentation methods based on superpixel techniques: A survey. *Journal of Computer Science and Information Systems*, 15(3): 1–11.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Liao, P.-S.; Chen, T.-S.; Chung, P.-C.; et al. 2001. A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.*, 17(5): 713–727.

Linares, O. A.; Botelho, G. M.; Rodrigues, F. A.; and Neto, J. B. 2017. Segmentation of large images based on superpixels and community detection in graphs. *IET Image Processing*, 11(12): 1219–1228.

Miao, L.; Li, X.; Zhou, X.; Yao, L.; Deng, Y.; Hang, T.; Zhou, Y.; and Yang, H. 2024. SNUNet3+: A Full-Scale Connected Siamese Network and a Dataset for Cultivated Land Change Detection in High-Resolution Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–18.

Ng, H.-F. 2006. Automatic thresholding for defect detection. *Pattern recognition letters*, 27(14): 1644–1649.

Niemeyer, I.; Marpu, P. R.; and Nussbaum, S. 2008. Change detection using object features. *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*, 185–201.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27.

Sikka, K.; and Deserno, T. M. 2010. Comparison of algorithms for ultrasound image segmentation without ground truth. In *Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment*, volume 7627, 423–431. SPIE.

Vedaldi, A.; and Soatto, S. 2008. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, 705–718. Springer.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. arXiv:2202.03052.

Wu, Q.; and Osco, L. P. 2023. samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM). *Journal of Open Source Software*, 8(89): 5663.