Preprints.org

Article

Not peer-reviewed version

MV-S2CD: A Modality-Bridged Vision Foundation Model-Based Framework for Unsupervised Optical–SAR Change Detection

Yonggi Shi , Ruopeng Yang * , Changsheng Yin , Yiwei Lu , Bo Huang , Yonggi Wen , Yihao Zhong , Zhaoyang Gu

Posted Date: 31 January 2026

doi: 10.20944/preprints202601.2350.v1

Keywords: unsupervised change detection; optical–SAR; vision foundation model; modality-bridged representation

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

MV-S2CD: A Modality-Bridged Vision Foundation Model-Based Framework for Unsupervised Optical–SAR Change Detection

Yongqi Shi [1], Ruopeng Yang [2,*], Changsheng Yin [2], Yiwei Lu [2], BO Huang [1], Yongqi Wen [1], Yihao Zhong [1] and Zhaoyang Gu [1]

[1] National University of Defense Technology, Wuhan 430030, China

[2] Information Support Force Engineering University, Wuhan 430030, China

* Correspondence: yangruopeng17@nudt.edu.cn

Highlights

What are the main findings?

- MV-S2CD introduces a modality-bridged VFM based framework that reduces the optical and SAR feature discrepancy through modality specific adapters, LoRA, and a shared projector for unsupervised change detection.
- MV-S2CD designs a dual branch change modeling module that separately models semantic consistency and structure sensitive differences to better handle heterogeneous modality effects.

What are the implications of the main findings?

- The results suggest that explicitly bridging cross modal representations, together with parameter efficient adaptation of pretrained VFM features, can improve robustness in unsupervised optical and SAR change detection without relying on pixel level change annotations.
- The framework provides a flexible and parameter efficient recipe for transferring large scale pretrained visual representations to multimodal remote sensing change analysis, and can be instantiated with different pretrained backbones by updating only small adaptation modules.

Abstract

Unsupervised change detection (UCD) from heterogeneous bitemporal optical–SAR imagery is challenging due to modality discrepancy, speckle/illumination variations, and the absence of change annotations. We propose MV-S2CD, a vision foundation model (VFM)-based framework that learns a modality-bridged latent space and produces dense change maps in a fully unsupervised manner. To robustly adapt pretrained VFM priors to heterogeneous inputs with minimal task-specific parameters, MV-S2CD incorporates lightweight modality-specific adapters and parameter-efficient low-rank adaptation (LoRA) in high-level layers. A shared projector embeds the two observations into a common geometry, enabling consistent cross-modal comparison and reducing sensor-induced domain shift. Building on the bridged representation, we design a dual-branch change reasoning module that decouples structure-sensitive cues from semantic-consistency cues: a structure pathway preserves fine boundaries and local variations, while a semantic-consistency pathway employs reliability gating and multi-scale context aggregation to suppress pseudo-changes caused by modality-specific nuisances and residual misregistration. For label-free optimization, we develop a difference-centric self-supervision scheme with two perturbation views and reliability-guided pseudo partitioning, jointly enforcing pseudo-unchanged invariance, pseudo-changed/unchanged separability, and sparsity and edge-preserving regularization. Experiments on three heterogeneous optical–SAR benchmarks demonstrate that MV-S2CD consistently improves the precision–recall trade-off and achieves state-of-the-art performance among unsupervised baselines, while remaining backbone-flexible and efficient.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

Keywords: unsupervised change detection; optical–SAR; vision foundation model; modality-bridged representation

1. Introduction

Change detection (CD) using multitemporal remote sensing imagery is a fundamental technique for Earth observation applications such as urban expansion monitoring, disaster assessment, land cover mapping, and environmental surveillance [1,2]. With the growing availability of heterogeneous sensors, especially optical and synthetic aperture radar (SAR), CD systems can benefit from complementary observations. Optical imagery provides rich spectral and textural cues under favorable illumination, whereas SAR offers all weather and day-night sensing that is less affected by clouds and illumination. Effectively leveraging these complementary modalities is therefore important for robust change analysis across diverse conditions [3,4].

However, UCD on heterogeneous optical and SAR pairs remains difficult [5,6]. The intrinsic modality discrepancy in radiometry, noise, and structural responses can generate strong pseudo changes [7], and the lack of pixel level annotations further limits the applicability and scalability of supervised solutions [1,3]. While traditional UCD methods based on handcrafted features, statistical modeling, or low level transformations often struggle with complex semantic changes [1,8], many deep UCD approaches still depend on task specific architectures or implicit alignment at the loss level, which may be insufficient to reconcile cross modal structural inconsistencies [9–11].

To address these challenges, we propose MV-S2CD, a backbone agnostic framework that leverages VFM priors and modality-bridged representation learning for heterogeneous CD. MV-S2CD introduces lightweight modality specific adapters and parameter efficient LoRA [12] to stabilize and adapt the pretrained backbone, and employs a shared projector to build a common latent geometry for reliable cross observation comparison. On this bridged space, a dual branch change module decouples structure sensitive difference cues from semantic consistency cues using reliability gating and multi scale context aggregation. The model is trained without change labels via difference centric self-supervision with perturbation views, including pseudo unchanged invariance, pseudo changed separability, and sparsity and structure aware regularization.

The main contributions of this work can be summarized as follows:

1. A modality-bridged VFM based UCD framework with adapters, LoRA, and a shared projector for explicit cross modal alignment.
2. A dual branch change reasoning module that combines structure sensitivity with semantic consistency for robust heterogeneous CD.
3. Extensive experiments on three heterogeneous benchmarks demonstrate the effectiveness of MV-S2CD and its key components, and show competitive performance against representative state of the art unsupervised methods with a small number of trainable parameters.

2. Related Work

Recent years have seen rapid advances in remote sensing CD, spanning classical unsupervised methods, multimodal frameworks, and VFM-based approaches. Section 2.1 reviews representative unsupervised CD methods, Section 2.2 summarizes multimodal CD with a focus on optical–SAR scenarios, and Section 2.3 discusses VFM-based CD and generic remote sensing VFMs, positioning our MV-S2CD framework within this emerging line of work.

2.1. Unsupervised Change Detection

UCD aims to delineate land-cover changes from multitemporal remote sensing data without pixel-level labels. Classical clustering- and statistics-based methods have been extended along several directions. Pattern-analysis approaches driven by Floating References reduce the need for strict radiometric normalization and are robust on uncalibrated imagery [13]. Fuzzy-topology-based

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

majority voting fuses multiple difference images in a fuzzy topological space to resolve conflicting pixels and improve fusion reliability [14]. Local-texture descriptors and adaptive thresholding further enhance performance on very-high-resolution (VHR) data: an extended center-symmetric LBP with local histogram similarity and a progressive Otsu strategy reduces false alarms and improves UCD in 2–5 m multispectral images [15]. Training-free spatial-context modeling has also been explored; for example, SiROC models each pixel as a linear combination of distant neighbors to detect temporal deviations, achieving competitive results on Sentinel-2 and PlanetScope images while providing calibrated uncertainty estimates [16]. Bayesian nonparametric clustering with Dirichlet process mixtures has been applied to PolSAR imagery, where unsupervised classification maps from a product model are compared across times to detect changes [17]. These approaches are training-light and label-free but struggle to capture complex semantics and multimodal discrepancies.

Deep learning has substantially advanced UCD by enabling more expressive representations and task-adaptive priors. Reconstruction- and autoencoder-based models learn features or priors from unlabeled data and then derive changes from feature differences or reconstruction errors. Super-resolution convolutional autoencoders reconstruct spatial details so that learned features preserve fine geometrical information for CD [18]. A total-variation-regularized bipartite network jointly optimizes feature differences and a TV-regularized change probability map, producing smoother, more intrinsic change maps in both homogeneous and heterogeneous scenes [19]. Deep kernel PCA–based Siamese convolutional mapping networks extract high-level spatial–spectral features and map feature differences into a polar domain for thresholding or clustering, enabling label-free UCD in VHR imagery [20]. Image-reconstruction-loss–based methods can even be trained on single-temporal images and then use high reconstruction errors under temporal mismatch at inference to localize changes [21]. Other works explicitly address acquisition-induced style discrepancies: STFL-CD transforms multitemporal multispectral images to a common style via unmixing and reconstruction, then learns joint spatial–spectral features with attention, mitigating the "same object with different spectra" problem and approaching supervised or semi-supervised performance [22]. Content-invariant translation with adversarial and hybrid attention mechanisms learns one-sided cross-domain mappings to suppress radiometric differences while preserving content, thereby highlighting real changes [23]. Beyond end-to-end training on bitemporal pairs, progressive learning frameworks iteratively refine pseudo-labels by selecting reliable regions and expanding them with label-selection filters, improving optical aerial UCD without external annotations [24]. Overall, current UCD methods have greatly improved robustness to illumination, noise, and subtle radiometric variations, but they are still predominantly designed for single-modality settings and struggle with substantial cross-sensor gaps, which motivates our study of multimodal change detection (MCD).

2.2. Multimodal Change Detection

MCD aims to identify land-cover changes from bitemporal images acquired by heterogeneous sensors such as optical and SAR. Compared with unimodal UCD, MCD must distinguish true changes from strong cross-sensor discrepancies in radiometry, spatial resolution and imaging geometry, which are particularly severe for optical–SAR pairs. Early works extend Bayesian and Markov random field models or structural regression to heterogeneous inputs by designing modality-robust cues or cross-modal mappings. Pixel-pairwise MRFs and fractal projection approaches build quasi–modality-invariant observation fields or project one image into the modality of the other before Markovian segmentation [25,26]. To better exploit structural information, a series of methods perform structure-constrained regression between modalities: structured graph regression, iterative structure transformation with CRF, structural regression fusion and SDIR regress one modality into the other (or both directions) under global/local structure constraints, often within Markovian or CRF-based segmentation frameworks and, more recently, with explicit modeling of both similarity and dissimilarity relationships via k-nearest and k-farthest neighbor graphs [6,27–29]. Building on the observation that structural relationships are more stable than raw spectra, graph-based methods represent multimodal images as structured graphs or superpixels and measure

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

structural differences. Locality-preserving energy models directly optimize change maps from modality-invariant structural links [30]; FD-MCD, GSGM and AOSG leverage local/nonlocal structures or global structure graphs—sometimes in the graph Fourier domain—to derive and fuse change intensity maps [7,31,32]. SR-GCAE and SDC-GAE further employ graph convolutional autoencoders and structural difference compensation to reconstruct features across modalities, where reconstruction errors or compensation values indicate change intensity under reconstruction, sparsity and structural consistency losses [5,33].

Deep learning has substantially advanced multimodal optical–SAR CD by learning cross-modal mappings, shared latent spaces and self-supervised representations. Deep image translation with an affinity-based change prior down-weights changed pixels in unsupervised translation losses via affinity-derived priors combined with cycle consistency and adversarial training, obtaining pseudo-aligned images for change analysis [4]. Commonality-feature–based frameworks such as AEKAN and CFRL employ Siamese encoders and dual decoders to reconstruct original and cross-reconstructed pseudo-images, enforcing hierarchical commonality or feature-distance constraints so that latent commonality features in unchanged regions form a comparable shared space, while their discrepancies highlight changes [34,35]. Non-Siamese architectures like HFA-PANet perform hierarchical feature alignment using multiple-kernel MMD at each level and progressively aggregate difference features to enhance contrast between changed and unchanged areas [36]. PMGN explicitly designates a primary modality and uses primary-modality-guided feature exchange and adaptive decision fusion to handle large resolution and radiometric gaps [37], while HF-MCD introduces heterogeneous collaborative and adaptive fusion modules to reconcile spatial-resolution gaps and balance consistent versus complementary cues across modalities [38]. Transformer-based MCD for DSM–aerial pairs leverages cross-attention and multitask consistency constraints between semantic and height changes [39], and CMMAN couples a Swin-based CD branch with two GAN-based modality transformation branches, using change-masked alignment and weakly modality-correlated feature enhancement to decouple modality heterogeneity from true changes in optical–SAR settings [40]. In parallel, self-/unsupervised frameworks such as CDR-Net and MaCon unify mask reconstruction and contrastive learning: mask reconstruction focuses on common low-level representations and implicit cross-modal transformation, whereas contrastive learning emphasizes semantic discrepancies, with tailored sampling and silent attention to stabilize training [11,41]. Iterative optimization-enhanced contrastive learning further uses a common projection layer and iterative optimization module to unify features into the same space and enlarge the separation between changed and unchanged regions [9]. Other methods explicitly distill common and discrepant representations via dual self-supervised branches or combine multiscale adversarial domain adaptation with divergence-aware contrastive regularization to jointly achieve modality alignment and change separability [42,43]. Overall, these multimodal methods have significantly improved robustness to cross-sensor discrepancies—especially for optical–SAR pairs—by exploiting structural graphs, deep regression, feature alignment and self-supervised representation learning [44]. However, they still rely on task-specific encoders and handcrafted objectives, are mostly tailored to particular modality pairs or resolution configurations, and rarely leverage the rich semantic priors of generic VFMs. This gap motivates our modality-bridged VFM-based framework for unsupervised optical–SAR CD in the subsequent sections.

2.3. Vision Foundation Models for Change Detection

Recent advances in large-scale VFMs have inspired a shift in remote sensing CD from fully task-specific architectures toward reusing pretrained generic visual priors. A first line of methods directly adapts segmentation-oriented VFMs, especially SAM-like models, to bitemporal VHR images. VFM-ReSCD integrates side adapters into FastSAM and introduces a recurrent module to model temporal semantic correlations, enabling semantic change detection (SCD) with joint change localization and land-cover classification in VHR RSIs [45]. SAM-CD employs the visual encoder of FastSAM and adds a convolutional adaptor to aggregate task-oriented change cues together with a task-agnostic

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

semantic learning branch, effectively exploiting the inherent semantic representations of SAM for binary CD and achieving performance competitive with fully/semisupervised methods [46]. Similar ideas are explored for MobileSAM in SCD scenarios, where fine-tuning the encoder helps adapt to the specific imaging characteristics of high-resolution RS images [47]. SAM-FDN further couples SAM with frequency-domain analysis via a low-rank fine-tuning strategy, introducing a Fourier-domain enhancement module that emphasizes meaningful changes while suppressing high-frequency noise, thereby reducing false alarms and missed detections in complex scenes [48]. From a different perspective, the burden-free distillation framework (BFD) uses VFMs only during training and transfers their general knowledge into lightweight CD models through dual-temporal feature matching and patch contrastive distillation, avoiding any additional computational burden from large models at inference time [49]. Other VFM-constrained CD networks explicitly inject VFM features via cross-attention–based semantic feature transfer and feature-constrained decoders, combined with feature-level loss terms, to guide the CD backbone toward semantically consistent change maps [50]. Beyond binary CD, AdaptVFMs-RSCD jointly leverages SAM and CLIP: CLIP provides image–text alignment for recognizing land-cover categories, while a semantic information-based CD module fuses change and semantic cues, significantly improving semantic CD and even enabling the conversion of binary CD datasets into semantic ones [51]. In addition, Siamese InternImage has been proposed as a CNN-based CD "vision foundation model" built upon deformable convolution, which aims to capture long-range dependencies and global context while preserving precise local details, and achieves strong performance on common CD benchmarks [52].

In parallel, several works in the broader remote sensing community have developed generic VFMs that, although not dedicated to CD, provide powerful backbones or priors for downstream CD tasks [53]. CROMA learns rich radar–optical representations via a joint self-supervised pretraining scheme that combines cross-modal contrastive learning and masked sensor modeling, and further improves scalability to larger images through relative positional biases (2D-/X-ALiBi), offering optionally multimodal encodings suitable for diverse downstream tasks including multimodal interpretation [54]. Multimodal Earth observation models based on lightweight VQ-VAE families extend CORSA-like architectures to handle both Sentinel-1 and Sentinel-2 and evaluate them on land-cover classification and CD, illustrating the promise of unified multimodal VFMs for Earth observation applications [55]. RemoteCLIP introduces the first vision–language foundation model tailored to remote sensing by scaling heterogeneous annotations into a unified image–caption format and incorporating UAV imagery, producing robust visual–text representations that support zero-shot and few-shot recognition and retrieval [56]. Large plain-ViT-based RS foundation models with rotated varied-size window attention demonstrate strong transferability to detection, classification and segmentation tasks and thus provide competitive generic representations for CD backbones [57]. The MTP framework advances RS foundation models via multitask pretraining on SAM-annotated RS segmentation data, jointly training semantic segmentation, instance segmentation and rotated detection with a shared encoder and task-specific decoders, and shows that such models can be effectively fine-tuned for CD alongside other downstream tasks [58]. Beyond pure vision models, PromptMID combines diffusion models, VFMs and text prompts to construct modality-invariant descriptors for optical–SAR image matching, highlighting the potential of foundation model-based priors to bridge heterogeneous modalities in tasks closely related to change analysis [59]. A recent survey systematically reviews VFMs in remote sensing, emphasizing that large-scale pretraining, self-supervised objectives such as contrastive learning and masked autoencoding, and multimodal designs are key to building robust and transferable RS foundation models [60]. Overall, current VFM-based CD methods mainly adapt or distill VFMs trained on natural images, focusing on binary or semantic CD in optical RSIs, and only marginally touch multimodal scenarios such as optical–SAR. Designing modality-bridged, VFM-based frameworks that explicitly handle large radiometric and structural gaps in heterogeneous CD, while remaining label-free, is still largely unexplored and motivates the MV-S2CD framework proposed in this work.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

3. Methods

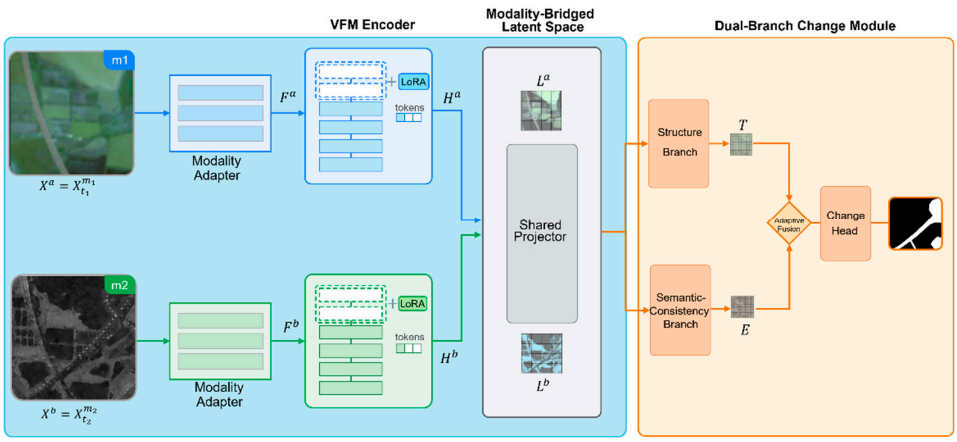
In this section, we first introduce MV-S2CD for unsupervised cross-modal bitemporal CD in Section 3.1. We then describe the modality-specific adapters and backbone-flexible VFM encoder in Section 3.2, followed by the modality-bridged latent space projection in Section 3.3 and the dual-branch change modeling module in Section 3.4. Finally, we present the difference-centric unsupervised learning objectives with perturbation views and regularization in Section 3.5.

3.1. Overall Architecture

The proposed MV-S2CD framework targets unsupervised cross-modal bitemporal CD from a single co-registered heterogeneous pair (X^a, X^b), by leveraging VFM priors and a modality-bridged representation learning scheme. As illustrated in Figure 1, MV-S2CD adopts a dual-stream encoder that processes two observations acquired at different times and different modalities, and produces a dense change map without requiring any change annotations. Importantly, we do not assume a fixed correspondence between time and modality: the earlier observation may be optical and the later SAR, or vice versa.

A pretrained VFM provides semantic priors for heterogeneous change reasoning. To preserve generality, MV-S2CD is backbone-flexible: the VFM module can be instantiated as (i) a single shared encoder applied to both modalities, or (ii) two modality-specific encoders (one per modality) that output dense token features on a compatible grid. In this work, we instantiate the VFM with CROMA [54], using its optical encoder for the optical observation and radar encoder for the SAR observation. We further augment the VFM module with (i) lightweight modality-specific adapters to reduce sensor-induced domain shift, (ii) LoRA modules injected into high-level layers for parameter-efficient adaptation, and (iii) a shared projector that constructs a modality-bridged latent space. On top of this latent space, MV-S2CD employs a dual-branch change reasoning module that decouples structure-sensitive differences from semantic-consistency cues: the structure branch emphasizes locality-preserving variations (e.g., boundaries and shape changes), while the semantic-consistency branch aggregates VFM-enhanced context and suppresses pseudo-changes caused by speckle, illumination/phenology variations, and residual misregistration. The two branches are fused adaptively to obtain pixel-wise change probabilities.

To enable learning from one heterogeneous pair, we design a fully unsupervised objective inspired by difference-centric self-supervision. Specifically, we generate two perturbation views of the same input pair and enforce: (i) cross-view consistency of learned change embeddings on pseudo-unchanged regions, (ii) explicit separability between pseudo-changed and pseudo-unchanged embeddings via a triplet-style objective, and (iii) grid sparsity and structure-aware smoothness to suppress insignificant changes and promote spatial coherence. These objectives jointly supervise the adapters, LoRA-augmented backbone, bridged latent space, and the dual-branch change module, resulting in discriminative change representations in a fully unsupervised manner.



© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

Figure 1. Overall architecture of MV-S2CD. Given one co-registered heterogeneous pair $(X^a, X^b) = (X_{t_1}^{m_1}, X_{t_2}^{m_2})$ with $t_1 \neq t_2$ and $m_1 \neq m_2$, each observation is first processed by a modality-specific adapter and then encoded by a VFM backbone. In this work, the backbone is instantiated with CROMA, using its unimodal optical encoder for optical inputs and unimodal radar encoder for SAR inputs; LoRA is applied to high-level transformer layers for parameter-efficient adaptation. A shared projector maps dense token features into a modality-bridged latent space, and a dual-branch change module produces a dense change map. The framework is backbone-flexible and can be instantiated with other transformer-style EO backbones that output dense token features.

3.2. Modality-Specific Adapters and VFM Backbone

3.2.1. Modality-Specific Adapters

Directly feeding raw optical and SAR images into a VFM may introduce severe domain shift, especially for SAR imagery whose radiometric statistics and speckle characteristics differ fundamentally from optical observations. To mitigate this issue while keeping the framework backbone-flexible, we introduce lightweight modality-specific adapters before the VFM encoders. The adapters transform modality-dependent inputs into a unified, backbone-compatible representation and format them to satisfy the input specification of the chosen VFM (e.g., channel configuration, normalization/value range, and tensor layout), facilitating stable optimization and downstream cross-observation interaction without constraining the method to any specific VFM. Although CROMA is pretrained on radar–optical data, its input conventions and the distribution of our target optical–SAR CD data may differ; the proposed adapters provide a lightweight, backbone-flexible interface that improves optimization stability and preserves portability across sensors and VFMs, such that switching to another VFM only requires updating this adapter/interface layer.

Let the input be a single co-registered heterogeneous pair:

$$X^a = X_{t_1}^{m_1} \in \mathbb{R}^{H\times W\times C_{m_1}}, X^b = X_{t_2}^{m_2} \in \mathbb{R}^{H\times W\times C_{m_2}} \quad (1)$$

where $m_1, m_2 \in \{\text{opt}, \text{sar}\}$ denote the two modalities (optical-like and SAR-like), and C_m is the channel number of modality m. We define two modality-specific adapters $A_{\text{opt}}(\cdot)$ and $A_{\text{sar}}(\cdot)$. The adapted feature maps are:

$$F^a = A_{m_1}(X^a), F^b = A_{m_2}(X^b) \quad (2)$$

where $F^a, F^b \in \mathbb{R}^{H\times W\times C_b}$, and C_b denotes the channel dimension expected by the backbone interface. Unless otherwise specified, we preserve the image-like spatial layout in the adapters and keep the spatial resolution unchanged.

Each adapter is implemented as a shallow and parameter-efficient network tailored to the statistics of its corresponding sensor modality. In this work, we adopt simple convolutional designs. For the optical adapter, we use 1×1 and 3×3 convolutions with batch normalization:

$$A_{\text{opt}}(X) = \text{Conv}_{C_b}^{3\times 3}\left(\sigma\left(\text{BN}\left(Conv_{C_{mid}}^{1\times 1}(X)\right)\right)\right) \quad (3)$$

where $Conv_C^{k\times k}$ denotes a convolution with kernel size $k\times k$ and C output channels, BN is batch normalization, $\sigma(\cdot)$ is a nonlinear activation (e.g., GELU), and C_{mid} is an intermediate channel width. For the SAR adapter, we employ instance normalization to better handle strong local contrast and multiplicative noise:

$$A_{sar}(X) = \text{Conv}_{C_b}^{3\times 3}\left(\sigma\left(\text{IN}\left(Conv_{C_{mid}}^{3\times 3}(X)\right)\right)\right) \quad (4)$$

where IN denotes instance normalization.

Overall, the modality-specific adapters (i) reduce modality-induced domain shift by aligning low-level statistics and channel dimensions, (ii) enable modality-aware preprocessing without modifying the backbone, and (iii) provide a unified interface for subsequent representation learning.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

3.2.2. VFM Encoder Adaptation with LoRA

Based on the modality-specific adapters, we employ a transformer-based VFM module to extract high-level semantic representations from the heterogeneous pair (F^a, F^b). MV-S2CD is backbone-flexible: the VFM module can be instantiated either as a single shared encoder or as two modality-specific encoders, as long as dense token features on a compatible spatial grid are provided for subsequent modality-bridged projection.

Instantiation with CROMA. In this work, we instantiate the VFM with CROMA and adopt its unimodal encoders: the optical encoder processes the optical observation, and the radar encoder processes the SAR observation. Let $B_{\text{opt}}(\cdot)$ and $B_{\text{sar}}(\cdot)$ denote the corresponding pretrained transformer encoders. Given F^a and F^b, we follow the backbone's tokenization interface and perform patch embedding, yielding token sequences:

$$Z^a, Z^b \in \mathbb{R}^{N\times D} \quad (5)$$

where N is the number of tokens and D is the embedding dimension. The tokens are then fed into the modality-matched LoRA-augmented encoders:

$$H^a = \tilde{B}_{m_1}(Z^a), H^b = \tilde{B}_{m_2}(Z^b) \quad (6)$$

where $m_1, m_2 \in \{\text{opt}, \text{sar}\}$ follow the modalities of (X^a, X^b), and $\tilde{B}_{\text{opt}}, \tilde{B}_{\text{sar}}$ denote CROMA's optical and radar encoders augmented with LoRA. Here $H^a, H^b \in \mathbb{R}^{N\times D}$ are high-level feature tokens. Positional encodings follow the adopted backbone implementation (e.g., attention biases in CROMA), and we do not impose additional positional embeddings, which preserves compatibility with other VFMs.

Fully fine-tuning a VFM is computationally expensive and may overfit, especially under unsupervised objectives. We therefore adopt LoRA to adapt the backbone in a parameter-efficient manner. Let $W \in \mathbb{R}^{d_{in}\times d_{out}}$ be a pretrained weight matrix of a selected linear layer. LoRA keeps W frozen and introduces a learnable low-rank update ΔW:

$$\Delta W = AB^\top, A \in \mathbb{R}^{d_{in}\times r}, B \in \mathbb{R}^{d_{out}\times r} \quad (7)$$

where $r \ll \min(d_{in}, d_{out})$. The effective weight becomes:

$$W' = W + \alpha\Delta W \quad (8)$$

with α being a scaling factor. Only A and B are updated during training.

In MV-S2CD, we apply LoRA to the top L_{LoRA} transformer blocks of each modality encoder and insert LoRA into the query and value projections of the MSA sub-layer. For a block input $X \in \mathbb{R}^{N\times D}$, the projections are:

$$Q = XW^Q + \alpha XA^Q(B^Q)^\top, K = XW^K, V = XW^V + \alpha XA^V(B^V)^\top \quad (9)$$

The attention output is computed as:

$$\text{MSA}(X) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V \quad (10)$$

where d_h denotes the head dimension. By freezing the original backbone and learning only low-rank updates in a few high-level layers, we substantially reduce task-specific parameters while preserving pretrained priors. The subsequent shared projector (Section 3.3) maps (H^a, H^b) into a common latent geometry, enabling robust cross-observation comparison for CD.

Generality remark. Replacing CROMA with another transformer-style VFM only requires substituting the encoder(s) $B_{\text{opt}}, B_{\text{sar}}$ (or a single shared B) while keeping the adapters, projector, change module, and unsupervised objectives unchanged.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

3.3. Modality-Bridged Latent Space Projection

A key component of MV-S2CD is a modality-bridged latent space built on top of the VFM features. It maps the two heterogeneous observations into a common embedding geometry so that cross-observation similarity and difference can be computed consistently for downstream change reasoning (Figure 2). Given token features H^a, H^b from Section 3.2.2, we reshape tokens back to spatial maps and apply a lightweight shared projector g_θ to obtain latent maps L^a, L^b.

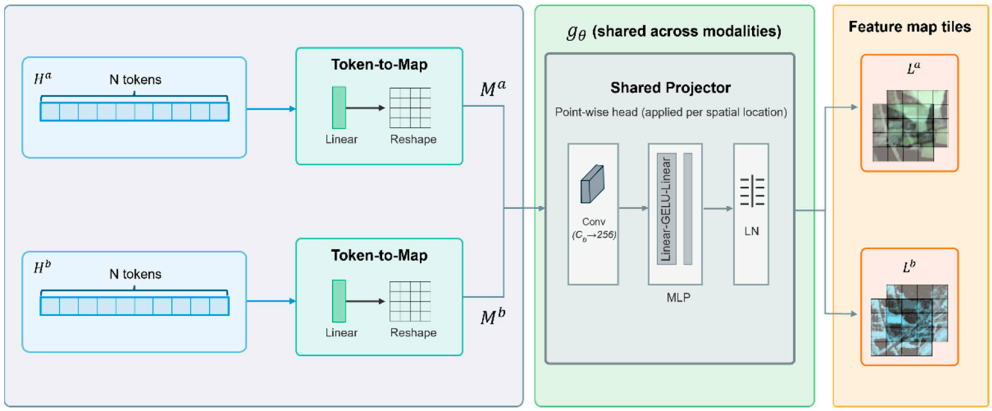


Figure 2. Modality-bridged latent space projection. Two co-registered observations are encoded by the VFM backbone to obtain token features H^a, H^b (in our instantiation, from CROMA optical and radar encoders, respectively). Tokens are reshaped to spatial maps M^a, M^b and projected by a shared point-wise head g_θ (1×1 Conv → 2-layer MLP (Linear–GELU–Linear) → LN) into latent maps $L^a, L^b \in \mathbb{R}^{H' \times W' \times D_z}$. Sharing g_θ embeds both representations into a common geometry for reliable cross-observation comparison.

3.3.1. Token-to-Map Reshaping

Let $H^{\cdot} \in \mathbb{R}^{N \times D}$ denote the output token features of the backbone encoder(s), where $N = H'W'$ with $H' = H/P$ and $W' = W/P$. We convert tokens back to spatial feature maps via a linear channel projection followed by reshaping:

$$M^a = \text{Reshape}_{H' \times W'}\left(H^a W_{\text{map}}\right), M^b = \text{Reshape}_{H' \times W'}\left(H^b W_{\text{map}}\right) \quad (11)$$

where $M^a, M^b \in \mathbb{R}^{H' \times W' \times C_b}$ and $W_{\text{map}} \in \mathbb{R}^{D \times C_b}$. If $D = C_b$, W_{map} can be treated as identity; otherwise it is a learnable projection.

3.3.2. Shared Projector and Latent Feature Maps

We introduce a shared projector $g_\theta(\cdot)$ whose parameters are shared across both observations:

$$L^a = g_\theta(M^a), L^b = g_\theta(M^b) \quad (12)$$

where $L^a, L^b \in \mathbb{R}^{H' \times W' \times D_z}$ and we set $D_z = 256$. For any $M \in \mathbb{R}^{H' \times W' \times C_b}$, the projector is:

$$g_\theta(M) = \text{LN}\left(\text{MLP}_2\left(\text{Conv}_{1\times1}(M)\right)\right) \quad (13)$$

where $\text{Conv}_{1\times1}: C_b \to 256$, MLP_2 is a point-wise 2-layer MLP (Linear–GELU–Linear: $256 \to 256 \to 256$) applied independently at each spatial location; and LN is performed over channels per location.

The latent feature maps

$$\{L^a, L^b\} \quad (14)$$

serve as the interface between the semantic encoder (adapters + VFM encoder(s)) and the subsequent change reasoning module. All downstream modules operate on these latent maps defined on the same spatial grid.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

3.4. Dual-Branch Change Modeling

As shown in Figure 3, we perform change reasoning on the modality-bridged latent maps $\{L^a, L^b\}$ obtained in Section 3.3 and derive a discriminative change representation C for dense prediction. Cross-modal CD requires both (i) locality-preserving modeling to capture fine structural variations and boundaries, and (ii) semantic-level reasoning to robustly identify land-cover transitions while suppressing pseudo-changes induced by speckle, illumination differences, and residual misregistration. To this end, we adopt a dual-branch design consisting of a structure branch and a semantic-consistency branch, and fuse their outputs adaptively.

Unless otherwise specified, all operators in this section are applied on the spatial grid of the latent maps $L^a, L^b \in \mathbb{R}^{H' \times W' \times D_z}$.

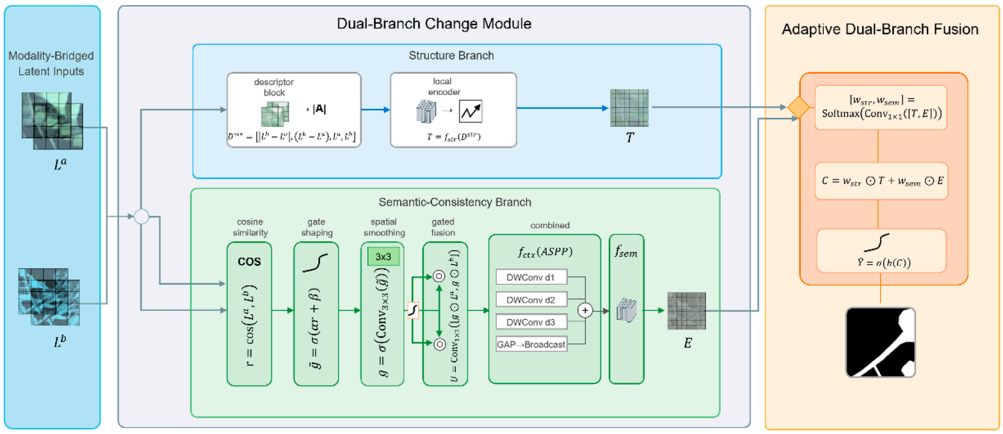


Figure 3. Dual-branch change module. The structure branch extracts locality-preserving cross-observation differences from (L^a, L^b). The semantic-consistency branch computes a reliability gate g from the cosine similarity between (L^a, L^b), aggregates multi-scale context, and outputs robust semantic change cues. The two branches are adaptively fused for the final prediction.

3.4.1. Structure Branch

The structure branch models cross-observation differences to preserve sharp boundaries and fine spatial details. We construct a bidirectional descriptor:

$$D^{str} = [|L^b - L^a|, (L^b - L^a), L^a, L^b] \quad (15)$$

where $[\cdot]$ denotes channel-wise concatenation. We obtain structure change features via a lightweight local encoder $f_{str}(\cdot)$:

$$T = f_{str}(D^{str}) \quad (16)$$

3.4.2. Semantic-Consistency Branch

This branch aims to suppress pseudo-changes by exploiting cross-observation agreement (enabled by the bridged latent space) and aggregating multi-scale context.

Reliability gating. We compute a cosine similarity between L^a and L^b at each location:

$$r(h,w) = \frac{\langle L^a(h,w), L^b(h,w) \rangle}{\|L^a(h,w)\|_2 \|L^b(h,w)\|_2 + \epsilon} \quad (17)$$

where $r(h,w)$ is a scalar and ϵ is a small constant.

We map the similarity to a soft gate:

$$\tilde{g} = \sigma(\alpha r + \beta) \quad (18)$$

$$g = \sigma\big(\mathrm{Conv}_{3\times3}(\tilde{g})\big) \quad (19)$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

where $\sigma(\cdot)$ is sigmoid and α, β are learnable scalars. Spatial smoothing reduces sensitivity to local outliers caused by speckle and residual misregistration.

Using g, we form a gated cross-modal representation:

$$U = \mathrm{Conv}_{1\times1}([g \odot L^a, g \odot L^b]) \quad (20)$$

where $\odot$ denotes element-wise multiplication (broadcast over channels).

Multi-scale context aggregation. We enrich semantic context via a lightweight module $f_{ctx}(\cdot)$:

$$\bar{U} = f_{ctx}(U) \quad (21)$$

We adopt an ASPP-style design:

$$f_{ctx}(U) = \mathrm{Conv}_{1\times1}\left(\left[\mathrm{DWConv}^{d_1}(U), \mathrm{DWConv}^{d_2}(U), \mathrm{DWConv}^{d_3}(U), \mathrm{Broadcast}\left(\mathrm{GAP}(U)\right)\right]\right) \quad (22)$$

where $\mathrm{DWConv}^{d}(\cdot)$ denotes depthwise atrous convolution with dilation rate d, $\mathrm{GAP}(\cdot)$ performs global average pooling, and $\mathrm{Broadcast}(\cdot)$ expands the pooled feature to match spatial size for concatenation.

Semantic change encoding. We obtain the semantic-consistency change feature via:

$$E = f_{sem}(\bar{U}) \quad (23)$$

where $f_{sem}(\cdot)$ is implemented with lightweight convolutions.

3.4.3. Adaptive Dual-Branch Fusion

The structure feature T is spatially precise, whereas the semantic-consistency feature E is more robust to modality-specific disturbances. We fuse them using a location-dependent gate:

$$[w_{str}, w_{sem}] = \mathrm{Softmax}\left(\mathrm{Conv}_{1\times1}([T, E])\right) \quad (24)$$

where $w_{str}, w_{sem} \in \mathbb{R}^{H'\times W'\times1}$. The final change representation is:

$$C = w_{str} \odot T + w_{sem} \odot E \quad (25)$$

The fused representation C is fed into a change head $h(\cdot)$ for dense prediction:

$$\hat{Y} = \sigma\left(h(C)\right) \in \mathbb{R}^{H'\times W'\times1} \quad (26)$$

During training, we compute predictions from two perturbation views (Section 3.5) and use $\hat{Y} = \frac{1}{2}(\hat{Y}_1 + \hat{Y}_2)$ in regularization terms unless otherwise specified.

3.5. Unsupervised Learning Objectives

We train MV-S2CD without pixel-level change annotations by optimizing a reliability-aware unsupervised objective consistent with the modality-bridged representation learning (Section 3.3) and the dual-branch change modeling with cross-observation gating (Section 3.4). Unlike settings that permit same-time cross-modal alignment or within-modality temporal differences, our objectives are designed to be computable from one heterogeneous pair and robust to heterogeneous nuisance factors.

Notation. Let Ω denote spatial locations on the $H' \times W'$ grid. For any tensor $A \in \mathbb{R}^{H'\times W'\times d}$, $A(p) \in \mathbb{R}^d$ denote the vector at $p \in \Omega$. Cosine similarity is:

$$\mathrm{sim}(a, b) = \frac{\langle a,b\rangle}{\|a\|_2\|b\|_2+\epsilon} \quad (27)$$

3.5.1. Perturbation Views and Contrastive Projector

We generate two stochastic views $v \in \{1,2\}$ of the same heterogeneous pair:

$$(X_v^a, X_v^b) = \mathcal{T}_v(X^a, X^b) \quad (28)$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

Spatial transforms (crop/flip/rotation) are applied identically to both observations to preserve pixel correspondence, while radiometric perturbations are modality-appropriate and do not alter geometry. Each view yields $\{L_v^a, L_v^b, g_v, C_v, \hat{Y}_v\}$.

To decouple dense prediction from contrastive self-supervision, we introduce a lightweight projection head $q(\cdot)$ (2-layer MLP with GELU and LN) and define:

$$Z_v(p) = q\big(C_v(p)\big) \in \mathbb{R}^{D_c} \quad (29)$$

All contrastive terms below are computed on Z_v, while the change head is trained on C_v.

3.5.2. Reliability-Guided Pseudo-Partitioning

We first compute a latent discrepancy score for each view:

$$s_v(p) = \|L_v^b(p) - L_v^a(p)\|_2 \quad (30)$$

and a reliability-guided score using the view-specific gate g_v:

$$\tilde{s}_v(p) = \big(1 - g_v(p)\big) \cdot s_v(p) \quad (31)$$

We select pseudo-unchanged and pseudo-changed sets by percentile thresholding:

$$\Omega_0^v = \{p \in \Omega | \tilde{s}_v(p) \le \tau_v\}, \Omega_1^v = \Omega \setminus \Omega_0^v \quad (32)$$

where $\rho \in (0,1)$ is the assumed change ratio and $\tau_v = \text{Quantile}_{1-\rho}(\{\tilde{s}_v(p) | p \in \Omega\})$ is the $(1-\rho)$-quantile of $\tilde{s}_v$, so that approximately a fraction ρ of pixels are assigned to Ω_1^v.

Cross-view agreement filtering. To improve pseudo-set purity without introducing an EMA teacher, we only keep locations whose pseudo status is consistent across both views:

$$\Omega_0 = \Omega_0^1 \cap \Omega_0^2, \Omega_1 = \Omega_1^1 \cap \Omega_1^2 \quad (33)$$

and ignore ambiguous pixels outside $\Omega_0 \cup \Omega_1$ when computing contrastive losses.

3.5.3. Difference-Centric Contrastive Learning

Pseudo-unchanged invariance. For pseudo-unchanged locations, the learned change embedding should be invariant across views:

$$\mathcal{L}_{inv} = \frac{1}{|\Omega_0|} \sum_{p \in \Omega_0} \Big(1 - \text{sim}\big(Z_1(p), Z_2(p)\big)\Big) \quad (34)$$

Pseudo-changed separability. Inspired by S2C's difference-centric learning [61], we explicitly encourage pseudo-changed embeddings to be consistent across views at the same location while separable from pseudo-unchanged embeddings. For each anchor $p \in \Omega_1$, we sample a negative location $n(p) \in \Omega_0$ and define:

$$\mathcal{L}_{tri} = \frac{1}{|\Omega_1|} \sum_{p \in \Omega_1} \Big[d\big(Z_1(p), Z_2(p)\big) - d\Big(Z_1(p), Z_2\big(n(p)\big)\Big) + m\Big]_+ \quad (35)$$

where $d(x,y) = 1 - \text{sim}(x,y)$, $m > 0$ is a margin, and $[\cdot]_+ = \max(0, \cdot)$.

Negative sampling. We adopt in-batch negatives: $n(p)$ is sampled uniformly from Ω_0 within the current mini-batch, providing diverse negatives at negligible overhead.

3.5.4. Grid Sparsity and Structure-Aware Regularization

Unsupervised heterogeneous CD is prone to pseudo-changes. We regularize $\hat{Y}$ using grid sparsity, reliability-guided sparsity, and structure-aware smoothness.

Grid sparsity. Partition Ω into K non-overlapping grids $\{\mathcal{G}_k\}$. Let

$$\bar{y}_k = \frac{1}{|\mathcal{G}_k|} \sum_{p \in \mathcal{G}_k} \hat{Y}(p) \quad (36)$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

$$\mathcal{L}_{gs} = \frac{1}{K}\sum_{k=1}^{K}|\bar{y}_k| \quad (37)$$

Reliability-guided sparsity. Let $\bar{g}(p) = \frac{1}{2}(g_1(p) + g_2(p))$. We impose:

$$\mathcal{L}_{spa} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\big(1 - \bar{g}(p)\big)\big|\hat{Y}(p)\big| \quad (38)$$

Structure-aware edge-preserving smoothness. Guided by the structure response T (computed from the unperturbed forward or averaged across views), we use:

$$\mathcal{L}_{tv} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\omega(p)\big(\big|\nabla_x\hat{Y}(p)\big| + \big|\nabla_y\hat{Y}(p)\big|\big), \omega(p) = \exp(-\kappa\|T(p)\|_2) \quad (39)$$

where ∇_x, ∇_y are finite differences and $\kappa > 0$.

3.5.5. Prediction Consistency and Gate Prior

Prediction consistency. We enforce view-consistent predictions:

$$\mathcal{L}_{pc} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\big\|\hat{Y}_1(p) - \hat{Y}_2(p)\big\|_2^2 \quad (40)$$

Gate prior. To prevent trivial solutions (e.g., $g_v \to 0$), we impose a weak prior:

$$\mathcal{L}_g = \sum_{v\in\{1,2\}}\Big(\frac{1}{|\Omega|}\sum_{p\in\Omega} g_v(p) - \pi\Big)^2 \quad (41)$$

where $\pi \in (0,1)$ is a weak prior on the expected fraction of reliable pixels.

3.5.6. Overall Objective and Warm-Up Strategy

The final unsupervised objective is

$$\mathcal{L} = \lambda_{inv}\mathcal{L}_{inv} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{gs}\mathcal{L}_{gs} + \lambda_{spa}\mathcal{L}_{spa} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_{pc}\mathcal{L}_{pc} + \lambda_g\mathcal{L}_g \quad (42)$$

where $\lambda_{\cdot}$ are weighting factors.

Warm-up. Since pseudo partitioning can be unreliable at the beginning, we adopt a short warm-up stage for the first E_w epochs, during which λ_{inv} and λ_{tri} are disabled and we optimize only:

$$\mathcal{L}_{warm} = \lambda_{gs}\mathcal{L}_{gs} + \lambda_{spa}\mathcal{L}_{spa} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_{pc}\mathcal{L}_{pc} + \lambda_g\mathcal{L}_g \quad (43)$$

After warm-up, we switch to the full objective in (42).

4. Experiments and Results

In this section, we present experimental results to evaluate the proposed MV-S2CD framework. Section 4.1 introduces the datasets and metrics, Section 4.2 presents implementation details, Section 4.3 reports comparison results on Gloucester, Shuguang, and California, and Section 4.4 provides ablation studies.

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

We evaluate MV-S2CD on three heterogeneous (optical–SAR) bitemporal change-detection datasets that cover different regions, sensor pairs, spatial resolutions, and change types. All image pairs are co-registered, and the corresponding binary change masks are available for quantitative

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

evaluation. The dataset statistics are summarized in Table 1, and qualitative examples are shown in Figure 4.

Gloucester (QB2/TSX) [26] is a very-high-resolution optical–SAR dataset acquired over Gloucester, UK (QuickBird-2 and TerraSAR-X, 0.65m). It captures flooding-related changes under pronounced cross-modal discrepancies and strong radiometric variations, posing challenges for direct pixel-wise comparison.

Shuguang [26] is a SAR–optical dataset collected over Shuguang, China (RADARSAT-2 and Google Earth imagery, 8m). The dominant changes correspond to urbanization, characterized by heterogeneous appearance shifts and fine structural modifications across modalities.

California [62] is a medium-resolution optical–SAR dataset acquired over California (Sentinel-1A and Landsat 8, ≈15m). It also targets flooding events, where coarser boundaries and mixed pixels increase ambiguity between subtle changes and background variability.

Table 1. Dataset Summary.

Name	Region	Date	Modality pair	Sensor	Event	Spatial Resolution
Gloucester (QB2/TSX)	Gloucester, UK	Jul 2006/Jul 2007	Optical–SAR	QuickBird2/TerraSAR-X	Flooding	0.65 m
Shuguang	Shuguang, China	Jun 2008/Sep 2012	SAR–Optical	Radarsat-2/Google Earth	Urbanization	8 m
California	California, USA	Jan 2017/Feb 2017	Optical–SAR	Sentinel-1A/Landsat 8	Flooding	≈15 m

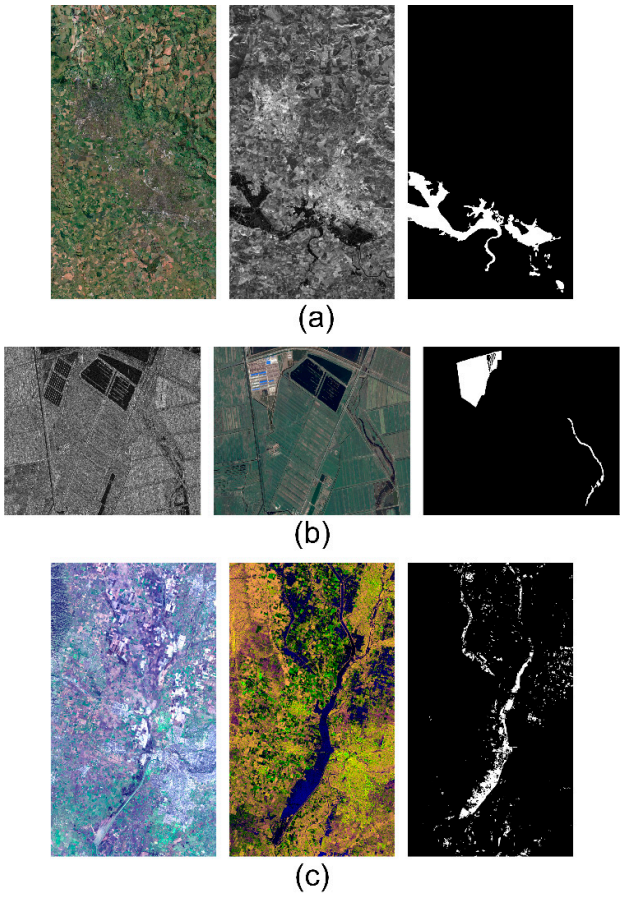


Figure 4. Datasets. (a) Gloucester. (b) Shuguang. (c) California.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

4.1.2. Evaluation Metrics

In MV-S2CD, the optical–SAR CD task is formulated as a binary pixel-wise classification problem, where each pixel is categorized as changed (positive class) or unchanged (negative class). Although the proposed framework is trained in an unsupervised manner, quantitative evaluation is conducted using the available ground-truth change masks for benchmarking. Following common practice in heterogeneous (optical–SAR) CD, we adopt four widely used metrics: Overall Accuracy (OA), Precision (P), Recall (R), and F1-score (F1).

Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively, computed by comparing the predicted binary change map with the ground truth. The metrics are defined as:

$$\mathrm{OA} = \frac{TP+TN}{TP+TN+FP+FN} \quad (44)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (45)$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \quad (46)$$

$$\mathrm{F1} = \frac{2TP}{2TP+FP+FN} \quad (47)$$

OA measures the overall proportion of correctly classified pixels. Precision indicates the reliability of detected change pixels by penalizing false alarms (FP), whereas Recall reflects the completeness of detected changes by penalizing missed detections (FN). The F1-score provides a balanced summary of Precision and Recall and is particularly informative for optical–SAR CD, where the changed class is often significantly smaller than the unchanged class.

To ensure fair comparison across methods and datasets, all metrics are computed on the final binary change maps using the same evaluation protocol.

4.2. Implementation Details

4.2.1. Environment and Pretraining

Training is conducted on two NVIDIA RTX 6000 Ada GPUs (48 GB each). We instantiate the backbone VFM with CROMA and initialize it with the released pretrained weights. During training, the original CROMA pretrained weights are kept frozen; we only optimize the modality-specific adapters, the shared projector, the dual-branch change module, the change head, the contrastive projection head, and the inserted LoRA parameters.

4.2.2. Hyperparameters

We set the modality-bridged latent dimension to $D_z = 256$ and the contrastive embedding dimension to $D_c = 256$.

LoRA is inserted into the query and value projections of the top $L_{\mathrm{LoRA}} = 4$ transformer blocks in each modality encoder, with rank $r = 8$ and scaling $\alpha = 16$.

For pseudo partitioning, we set the change-ratio hyperparameter ρ to 0.06; the triplet margin is set to $m = 0.2$. The gate prior uses $\pi = 0.7$.

For structure-aware smoothness, we set $\kappa = 1.0$. Grid sparsity is computed on non-overlapping 16×16 grids on the latent $H' \times W'$ map.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

Loss weights are fixed as: $\lambda_{inv} = 1.0$, $\lambda_{tri} = 1.0$, $\lambda_{gs} = 0.5$, $\lambda_{spa} = 1.0$, $\lambda_{tv} = 0.1$, $\lambda_{pc} = 1.0$, and $\lambda_g = 0.1$.

4.2.3. Training Details

Training is performed on randomly cropped paired patches sampled from the full-resolution images. The patch size is 512×512 for Gloucester and California, and 256×256 for Shuguang.
For each patch, we generate two stochastic perturbation views. Spatial transforms are applied identically to both modalities to preserve correspondence, while radiometric perturbations are applied independently per modality.
We use AdamW with learning rate 1×10^{-4} and weight decay 1×10^{-2}. The global batch size is 16 (over two GPUs). We run 10k iterations for Gloucester and California, and 3k iterations for Shuguang.

4.3. Comparison Experiments

To validate the effectiveness of MV-S2CD on heterogeneous (optical–SAR) binary CD, we compare it with representative unsupervised baselines spanning Bayesian statistical segmentation, structure-consistency graph mapping, projection/translation-based alignment, reconstruction-driven anomaly learning, and affinity-aware self-supervised translation. For each method, we first attempt to reproduce the reported results by using the official implementation (when available) and following the original experimental protocol. When the reproduced performance is consistent with that reported in the corresponding paper, we directly report the numbers from the original paper.
M3CD (Multimodal MRF change detector) [63]. M3CD formulates heterogeneous CD in an unsupervised Bayesian framework. It constructs a modality-robust observation field via pixel-pairwise modeling and estimates the binary change map through MRF-based spatial regularization and MAP inference. The likelihood parameters are iteratively estimated, and spatial coherence is enforced by the Markov prior.
Fractal (Fractal projection with Markovian segmentation) [26]. This line of work first performs fractal-based cross-modal projection, mapping one modality into the other by exploiting self-similarity structures that are comparatively stable across modalities. The projected pair is then compared in a common domain, and the resulting difference map is binarized using a Markovian segmentation scheme, typically with EM-based parameter estimation and ICM-style local refinement.
DSRM (Deep Sparse Residual Model) [64]. DSRM casts heterogeneous CD as unsupervised anomaly detection in a learned residual space. A stacked sparse autoencoder is trained to capture normal (unchanged) patterns, and pixel-/patch-wise reconstruction errors are used as change indicators. The final change map is obtained by clustering the residual errors, without requiring change annotations.
IRG-McS.dist / IRG-McS.sim (Iterative Robust Graph with Markovian co-segmentation)[65]. IRG-McS leverages modality-invariant structural consistency by constructing robust (often superpixel-based) kNN graphs in each modality. It computes forward and backward difference images through intra-domain graph mapping, and fuses them using an MRF co-segmentation model solved by co-graph cut. An iterative feedback mechanism updates graph construction by reducing the influence of detected change regions, improving robustness to “changed neighbors”.
X-Net / ACE-Net (Affinity-prior-guided deep image translation) [4]. These methods learn unsupervised cross-domain translation while mitigating the impact of change pixels during training. They derive an unsupervised change prior from domain-specific affinity matrices (relational pixel information) and use it to weight translation objectives, typically combined with cycle consistency and (for some variants) adversarial learning. The translated outputs enable constructing comparable representations and difference images for subsequent change extraction.
CAA (Code-Aligned Autoencoders) [10]. CAA improves translation-based UCD by explicitly enforcing latent code-space alignment across modalities. Using affinity-derived relational constraints, it encourages pixels with consistent affinity relations in the two domains to remain correlated in latent space, thereby reducing cross-modal discrepancy and suppressing the influence of changed pixels during representation learning.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

4.3.1. Results on Gloucester

Table 2 reports quantitative results on the Gloucester (QB2/TSX) dataset. MV-S2CD achieves the best overall performance, obtaining an OA of 0.978 and an F1-score of 0.817, which indicates a more favorable precision–recall trade-off under severe optical–SAR discrepancies. Compared with the strongest competing method in terms of F1 (Fractal [26], 0.790), MV-S2CD improves F1 by +2.7 points, while also delivering the highest precision (0.893), suggesting fewer false alarms. Although Fractal attains the highest recall (0.855), it does so at the cost of reduced precision, reflecting a tendency to over-detect changes in challenging cross-modal regions. By contrast, MV-S2CD better suppresses pseudo-changes while maintaining competitive recall (0.753), benefiting from the modality-bridged latent space and the complementary structure/semantic-consistency reasoning.

Figure 5 provides a qualitative comparison. Traditional unsupervised baselines exhibit either scattered false positives in homogeneous background regions or missed detections along thin/irregular flooded boundaries. In contrast, MV-S2CD produces cleaner predictions with sharper object boundaries and fewer spurious responses, aligning more closely with the ground truth. This visual evidence corroborates the quantitative gains in Table 2 and highlights MV-S2CD's robustness to strong radiometric differences and speckle-induced artifacts on Gloucester.

Table 2. Results on Gloucester dataset. The best result for each metric is highlighted in bold.

Method	OA	Precision	Recall	F1
M3CD [63]	0.955	0.638	0.693	0.665
Fractal [26]	0.971	0.733	**0.855**	0.790
DSRM [64]	0.964	0.703	0.680	0.691
IRG-McS.dist [65]	0.972	0.852	0.672	0.751
IRG-McS.sim [65]	0.972	0.853	0.676	0.755
MV-S2CD (ours)	**0.978**	**0.893**	0.753	**0.817**

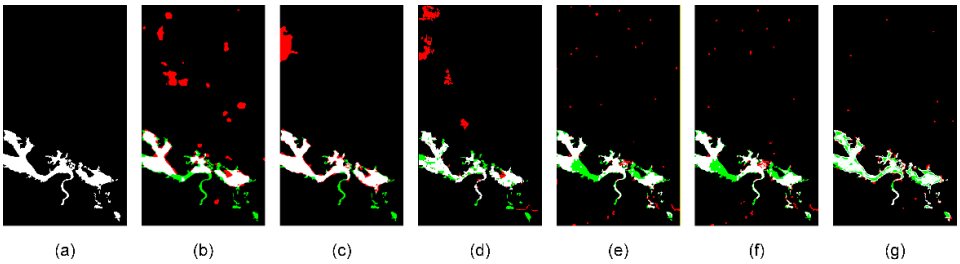


Figure 5. Qualitative comparison on the Gloucester dataset. From left to right: (a) Ground Truth, (b) M3CD, (c) Fractal, (d) DSRM, (e) IRG-McS.dist, (f) IRG-McS.sim, and (g) MV-S2CD (ours). Red indicates false positives, green indicates false negatives, and white indicates true positives.

4.3.2. Results on Shuguang

Table 3 summarizes the performance on the Shuguang (RADARSAT-2/Google Earth) dataset, which features pronounced SAR–optical appearance gaps and fine-grained urbanization changes. MV-S2CD achieves the best results across all metrics, reaching an OA of 0.989 and an F1-score of 0.821. Compared with the strongest baseline (IRG-McS.dist [65]), MV-S2CD improves F1 by +1.7 points (0.821 vs. 0.804) while also yielding higher precision (0.857 vs. 0.843) and recall (0.788 vs. 0.769). These gains indicate that MV-S2CD not only reduces false alarms but also recovers more true change regions, resulting in a consistently better precision–recall balance.

Figure 6 provides qualitative evidence. Methods based on statistical segmentation or reconstruction tend to produce fragmented detections or miss small structural modifications, whereas translation-based approaches may introduce pseudo-changes due to imperfect cross-domain alignment. In contrast, MV-S2CD generates more coherent change masks with fewer isolated

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

false positives and better completeness on thin or discontinuous urban change patterns, consistent with its improved quantitative performance.

Table 3. Results on Shuguang dataset. The best result for each metric is highlighted in bold.

Method	OA	Precision	Recall	F1
M3CD [63]	0.962	0.574	0.682	0.624
DSRM [64]	0.982	0.818	0.663	0.732
IRG-McS.dist [65]	0.983	0.843	0.769	0.804
IRG-McS.sim [65]	0.978	0.751	0.782	0.767
ACE-Net [4]	0.982	0.804	0.662	0.726
X-Net [4]	0.984	0.816	0.665	0.731
MV-S2CD (ours)	**0.989**	**0.857**	**0.788**	**0.821**

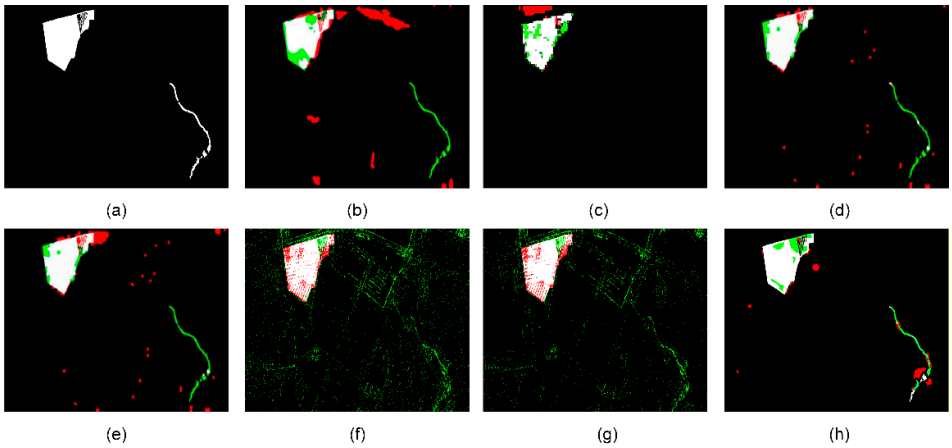


Figure 6. Qualitative comparison on the Shuguang dataset. From left to right: (a) Ground Truth, (b) M3CD, (c) DSRM, (d) IRG-McS.dist, (e) IRG-McS.sim, (f) ACE-Net, (g) X-Net, and (h) MV-S2CD (ours). Red indicates false positives, green indicates false negatives, and white indicates true positives.

4.3.3. Results on California

Table 4 reports results on the California (Sentinel-1A/Landsat 8) dataset, a medium-resolution setting where flood boundaries are coarser and mixed pixels are prevalent, making it difficult to distinguish subtle inundation from background variability. MV-S2CD achieves the best overall performance, with an OA of 0.967 and an F1-score of 0.685. Notably, MV-S2CD substantially outperforms the strongest baseline (CAA [10]) by +11.6 points in F1 (0.685 vs. 0.569), indicating more reliable change extraction under larger cross-sensor gaps and resolution-induced ambiguity.

From a precision–recall perspective, MV-S2CD attains the highest precision (0.719) while also providing the highest recall (0.654), demonstrating that it simultaneously reduces false alarms and recovers a larger portion of true flood pixels. In comparison, graph-based methods (IRG-McS.dist/sim [65]) achieve moderate F1 scores (~0.51) but tend to under-detect changes, whereas translation-based approaches (ACE-Net/X-Net [4]) favor high precision at the expense of low recall, missing substantial flooded regions. M3CD [63] performs poorly in this dataset, suggesting that its statistical assumptions are less suitable under this medium-resolution, highly imbalanced scenario.

Qualitative results (Figure 7) further confirm these observations: MV-S2CD produces more spatially coherent flood extents with fewer scattered false positives and fewer missed detections along elongated or fragmented inundation patterns, aligning better with the ground truth than competing methods.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

Table 4. Results on California dataset. The best result for each metric is highlighted in bold.

Method	OA	Precision	Recall	F1
M3CD [63]	0.592	0.043	0.390	0.077
Fractal [26]	0.952	0.487	0.389	0.433
IRG-McS.dist [65]	0.959	0.534	0.492	0.512
IRG-McS.sim [65]	0.958	0.526	0.491	0.508
ACE-Net [4]	0.915	0.714	0.338	0.459
X-Net [4]	0.911	0.684	0.332	0.447
CAA [10]	0.924	0.518	0.631	0.569
MV-S2CD (ours)	**0.967**	**0.719**	**0.654**	**0.685**

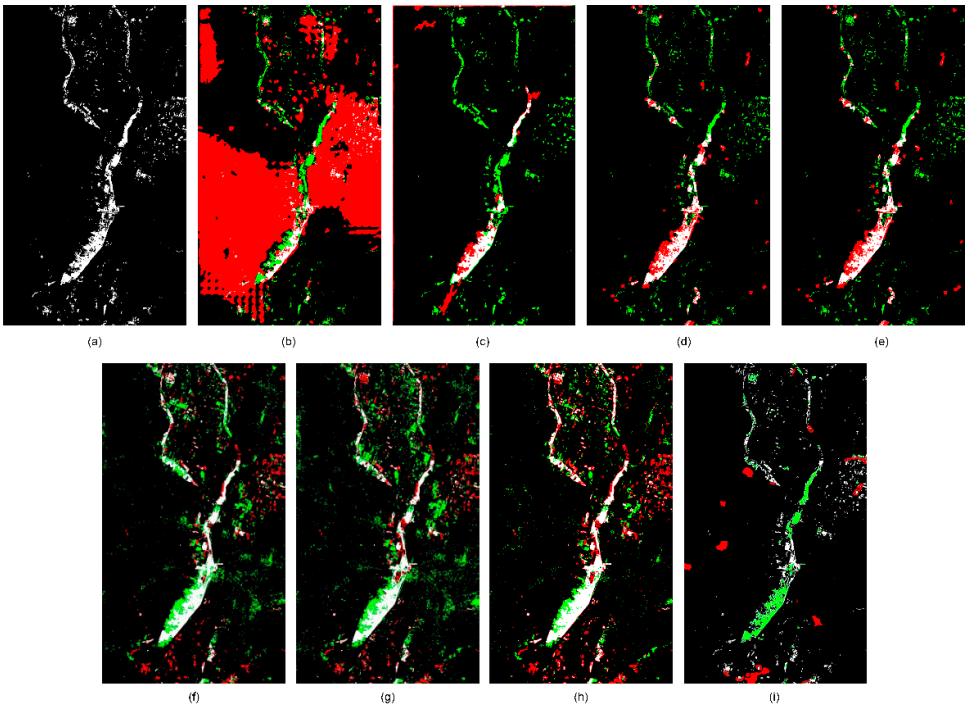


Figure 7. Qualitative comparison on the California dataset. From left to right: (a) Ground Truth, (b) M3CD, (c) Fractal, (d) IRG-McS.dist, (e) IRG-McS.sim, (f) ACE-Net, (g) X-Net, (h)CAA, and (i) MV-S2CD (ours). Red indicates false positives, green indicates false negatives, and white indicates true positives.

4.4. Ablation Studies

We conduct ablation studies on the Gloucester dataset to quantify the contribution of each key design choice. Unless otherwise specified, all ablated variants follow the same training protocol and hyper-parameters as the full model; we only remove (or disable) the component under study. We report OA, Precision, Recall, and F1-score for the change class.

We do not perform an ablation over different VFM backbones because only a limited number of publicly available VFMs currently support both optical and SAR inputs; most VFMs are trained on optical imagery only, which prevents a fair and consistent cross-modal comparison. As more optical–SAR-capable VFMs are actively being developed by the community, we will further analyze backbone choices and generalization once such models become available.

4.4.1. Architecture Ablation

We first evaluate the effect of the main architectural components of MV-S2CD, including the modality-specific input adapters, parameter-efficient backbone adaptation (LoRA), the modality-

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

bridged latent space projection, and the dual-branch change reasoning module (structure branch and semantic-consistency branch). The ablated variants are defined as follows:

w/o Adapter removes the modality-specific adapters while keeping all remaining components unchanged. This variant tests whether explicit low-level modality alignment is necessary for stable cross-modal optimization.

w/o LoRA disables LoRA and keeps the VFM backbone fully frozen, assessing the benefit of lightweight high-level adaptation beyond training only the task heads.

w/o Bridge (No-Bridge) removes the proposed modality-bridging constraint by discarding the shared projector. Concretely, we replace the shared projector with two independent projectors (one per modality) while keeping the same projector architecture and output dimensionality. All subsequent modules and training objectives operate on the projected latent features in the same way as the full model. This variant evaluates whether enforcing a shared latent geometry through a shared projector is critical for robust cross-modal comparison.

w/o semantic-consistency branch (Structure-only) removes the semantic-consistency branch (including reliability gating and context aggregation) and predicts changes solely from the structure branch. The fusion is accordingly bypassed so that the final change representation is produced only by the structure pathway. This isolates the contribution of semantic-consistency modeling in suppressing pseudo-changes.

w/o structure branch (Semantic-only) removes the structure branch and predicts changes solely from the semantic-consistency branch. The fusion is accordingly bypassed so that the final change representation is produced only by the semantic-consistency pathway. This variant evaluates whether boundary/detail recovery is mainly provided by the structure pathway.

Table 5 and Figure. 8 report the architectural ablation results on the Gloucester dataset. The full MV-S2CD achieves the best performance (OA = 0.978, Precision = 0.893, Recall = 0.753, F1 = 0.817), indicating that the proposed components contribute jointly to robust cross-modal CD. Removing the modality-bridged projection (w/o Bridge) causes the most severe degradation (F1 drops to 0.623 and OA to 0.937), with both Precision and Recall decreasing simultaneously, which confirms that enforcing a shared latent geometry is crucial for reliable cross-modal comparison.

Table 5. Architecture ablation on the Gloucester dataset. The best result for each metric is highlighted in bold.

Method	OA	Precision	Recall	F1
w/o Adapter	0.951	0.782	0.604	0.682
w/o LoRA	0.955	0.813	0.569	0.669
w/o Bridge	0.937	0.723	0.547	0.623
w/o Sem-Cons	0.943	0.637	0.728	0.679
w/o Str	0.932	0.847	0.476	0.603
MV-S2CD(Full)	**0.978**	**0.893**	**0.753**	**0.817**

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

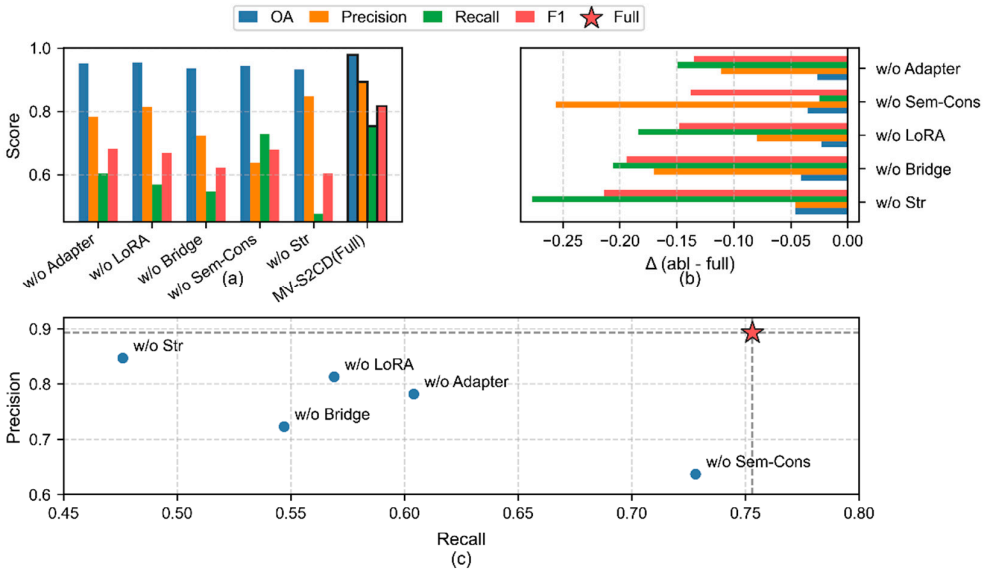


Figure 8. Architecture ablation results on the Gloucester dataset. (a) Absolute performance in terms of OA, Precision, Recall, and F1. (b) Performance drop relative to the full model (abl − full). (c) Precision–Recall trade-off, with the full model highlighted.

The adaptation mechanisms also play a non-negligible role. Without the modality-specific adapters or LoRA, the F1 score decreases to 0.682 and 0.669, respectively, and the Recall drops markedly (0.604 and 0.569 vs. 0.753 for the full model). This suggests that low-level modality alignment and lightweight high-level backbone adaptation are both important to mitigate modality discrepancies and to avoid overly conservative predictions that miss true changes.

Finally, the dual-branch design shows clear complementarity. Removing the semantic-consistency branch (Structure-only) yields a relatively high Recall (0.728) but a pronounced Precision drop (0.637), implying increased pseudo-changes when semantic constraints are absent. Conversely, removing the structure branch (Semantic-only) preserves high Precision (0.847) but leads to the lowest Recall (0.476), indicating that fine boundary/detail recovery and small-change sensitivity largely rely on the structure pathway. Overall, MV-S2CD provides a better Precision–Recall trade-off by integrating both branches, as evidenced by the largest F1 and the favorable operating point in Figure 8(c).

4.4.2. Objective Ablation

We further ablate the key learning strategy and objectives that implement our difference-centric unsupervised representation learning.

No-Warmup removes the warm-up schedule in Section 3.5.6, i.e., L_{inv} and L_{tri} are enabled from the first epoch and the full objective in Eq. (42) is optimized throughout training. This variant tests whether warm-up is necessary to avoid early-stage instability caused by unreliable pseudo partitioning under the one-pair training setting.

w/o L_{inv} removes the pseudo-unchanged invariance term in Eq. (34) while keeping all remaining losses unchanged. This evaluates whether explicitly enforcing cross-view invariance on pseudo-unchanged regions is critical to prevent over-responding to cross-modal appearance discrepancies.

w/o L_{tri} removes the triplet-style separability term in Eq. (35) while keeping the remaining losses unchanged. This tests the role of explicitly separating pseudo-changed and pseudo-unchanged embeddings in improving change/unchanged discriminability.

These ablations demonstrate the necessity of the proposed learning objectives. As reported in Table 6, the full MV-S2CD achieves the best overall performance (OA/F1 = 0.978/0.817). Removing the warm-up schedule leads to a clear degradation (F1 drops to 0.755), with a pronounced decrease

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

in recall (0.681 vs. 0.753), indicating that enabling all objectives from the beginning can amplify the impact of unreliable early pseudo partitioning and hinder the model from recovering true changes in the one-pair setting.

Moreover, disabling either objective term consistently harms the precision–recall trade-off. Without L_{inv}, precision decreases substantially (0.791 vs. 0.893) while recall remains relatively close to the full model (0.732 vs. 0.753), suggesting that invariance on pseudo-unchanged regions is crucial for suppressing false alarms induced by cross-modal appearance discrepancies. In contrast, removing L_{tri} causes the largest recall drop (0.623), even though precision remains comparatively high (0.871), implying that explicit separability is essential for reducing missed detections by enlarging the margin between pseudo-changed and pseudo-unchanged embeddings.

Table 6. Objective ablation on the Gloucester dataset. The best result for each metric is highlighted in bold.

Method	OA	Precision	Recall	F1
No-Warmup	0.966	0.847	0.681	0.755
w/o L_{inv}	0.963	0.791	0.732	0.760
w/o L_{tri}	0.959	0.871	0.623	0.726
MV-S2CD(Full)	**0.978**	**0.893**	**0.753**	**0.817**

5. Discussion

MV-S2CD shows that explicitly bridging the optical–SAR modality gap at the representation level is critical for unsupervised heterogeneous CD. The ablation results indicate that enforcing a shared latent geometry via the shared projector contributes most to performance, suggesting that loss-level alignment alone is often insufficient when cross-modal structural discrepancies are strong. In addition, the modality-specific adapters and LoRA-based high-level adaptation provide complementary benefits, improving optimization stability and change sensitivity with limited trainable parameters.

The dual-branch change module offers a practical way to balance boundary fidelity and robustness. The structure branch enhances fine-grained localization but is more prone to pseudo-changes, while the semantic-consistency branch suppresses modality-induced artifacts yet can miss small or thin changes. Their adaptive fusion yields a better precision–recall trade-off than either branch alone, consistent with the observed trends in the architectural ablations.

From the learning perspective, difference-centric self-supervision is effective under the one-pair setting, but it is sensitive to early-stage noisy pseudo partitioning. The warm-up strategy and the combination of pseudo-unchanged invariance and triplet-style separability help stabilize training and improve discriminability, as reflected in the objective ablations. Remaining challenges include dependence on co-registration quality and the hyper-parameterization in percentile-based pseudo partitioning, as well as extending the framework to broader sensor combinations or longer temporal sequences without sacrificing the fully unsupervised and parameter-efficient setting.

6. Conclusions

This paper presented MV-S2CD, a VFM-based framework for unsupervised heterogeneous bitemporal CD between optical and SAR imagery. The method introduces modality-specific adapters and LoRA-based parameter-efficient backbone adaptation, and constructs a modality-bridged latent space using a shared projector to enable consistent cross-modal comparison. A dual-branch change modeling module further decouples structure-sensitive differences from semantic-consistency cues, and an unsupervised difference-centric learning scheme with perturbation views, warm-up training, and sparsity and structure-aware regularization provides effective supervision without change labels.

Experiments on three heterogeneous benchmarks demonstrate that MV-S2CD yields competitive performance against representative unsupervised methods. Ablation studies verify the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

importance of the bridged latent space, the adaptation mechanisms, the dual-branch design, and the proposed unsupervised objectives. Future work will focus on improving robustness to larger misregistration, reducing sensitivity to hyper-parameters involved in pseudo partitioning, and extending the framework to multitemporal sequences and broader sensor configurations while maintaining parameter-efficient adaptation.

Author Contributions: Conceptualization, Y.S. and R.Y.; methodology, Y.S. and R.Y.; software, Y.S., R.Y., and C.Y.; validation, Y.S., R.Y., Y.Z., and Z.G.; formal analysis, Y.S. and R.Y.; investigation, R.Y., C.Y., and Y.L.; resources, B.H. and Y.W.; data curation, C.Y. and Y.L.; writing—original draft preparation, Y.S. and R.Y.; writing—review and editing, Y.S., R.Y., B.H., Y.W., Y.Z., and Z.G.; visualization, Y.S., R.Y., and Y.L.; supervision, B.H., Y.W., and Y.Z.; project administration, B.H. and Y.W.; funding acquisition, Y.S. and R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Planning Project of Military Construction during the 14th Five-Year Plan period (No. 145BWX053021000X).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to express our sincere gratitude to the editors and the anonymous reviewers for their valuable time, careful evaluation, and constructive comments, which have greatly improved the quality of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CD	Change detection
UCD	Unsupervised change detection
MCD	Multimodal change detection
SCD	Semantic change detection
VFM	Vision foundation model
SAR	Synthetic aperture radar
VHR	Very-high-resolution
BFD	Burden-free distillation framework
LoRA	Low-rank adaptation
OA	Overall Accuracy
P	Precision
R	Recall
F1	F1-score

References

1. Bai, T.; Wang, L.; Yin, D.; Sun, K.; Chen, Y.; Li, W.; Li, D. Deep Learning for Change Detection in Remote Sensing: A Review. *Geo-Spat. Inf. Sci.* **2023**, *26*, 262–288, doi:10.1080/10095020.2022.2085633.
2. Jiang, H.; Peng, M.; Zhong, Y.; Xie, H.; Hao, Z.; Lin, J.; Ma, X.; Hu, X. A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1552, doi:10.3390/rs14071552.
3. Saidi, S.; Idbraim, S.; Karmoude, Y.; Masse, A.; Arbelo, M. Deep-Learning for Change Detection Using Multi-Modal Fusion of Remote Sensing Images: A Review. *Remote Sens.* **2024**, *16*, 3852, doi:10.3390/rs16203852.
4. Luppino, L.T.; Kampffmeyer, M.; Bianchi, F.M.; Moser, G.; Serpico, S.B.; Jenssen, R.; Anfinsen, S.N. Deep Image Translation With an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–22, doi:10.1109/TGRS.2021.3056196.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

5. Chen, H.; Yokoya, N.; Wu, C.; Du, B. Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18, doi:10.1109/TGRS.2022.3229027.
6. Sun, Y.; Lei, L.; Liu, L.; Kuang, G. Structural Regression Fusion for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–18, doi:10.1109/TGRS.2023.3294884.
7. Chen, H.; Yokoya, N.; Chini, M. Fourier Domain Structural Relationship Analysis for Unsupervised Multimodal Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 99–114, doi:10.1016/j.isprsjprs.2023.03.004.
8. Khelifi, L.; Mignotte, M. Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. *IEEE Access* **2020**, *8*, 126385–126400, doi:10.1109/ACCESS.2020.3008036.
9. Tang, Y.; Yang, X.; Han, T.; Sun, K.; Guo, Y.; Hu, J. Iterative Optimization-Enhanced Contrastive Learning for Multimodal Change Detection. *Remote Sens.* **2024**, *16*, 3624, doi:10.3390/rs16193624.
10. Luppino, L.T.; Hansen, M.A.; Kampffmeyer, M.; Bianchi, F.M.; Moser, G.; Jenssen, R.; Anfinsen, S.N. Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 60–72, doi:10.1109/TNNLS.2022.3172183.
11. Wang, J.; Yan, L.; Xie, H.; Zhou, T.; Shi, W.; Atkinson, P.M. Unsupervised Multimodal Change Detection by Distilling Common and Discrepant Representations. In Proceedings of the 2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2024); IEEE: Athens, GREECE, 2024; pp. 7817–7820.
12. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models 2021.
13. Negri, R.G.; Frery, A.C. Unsupervised Change Detection Driven by Floating References: A Pattern Analysis Approach. *Pattern Anal. Appl.* **2021**, *24*, 933–949, doi:10.1007/s10044-020-00954-w.
14. Shao, P.; Shi, W.; Liu, Z.; Dong, T. Unsupervised Change Detection Using Fuzzy Topology-Based Majority Voting. *Remote Sens.* **2021**, *13*, 3171, doi:10.3390/rs13163171.
15. Shen, Y.; Wei, Y.; Zhang, H.; Rui, X.; Li, B.; Wang, J. Unsupervised Change Detection in HR Remote Sensing Imagery Based on Local Histogram Similarity and Progressive Otsu. *Remote Sens.* **2024**, *16*, 1357, doi:10.3390/rs16081357.
16. Kondmann, L.; Toker, A.; Saha, S.; Scholkopf, B.; Leal-Taixe, L.; Zhu, X.X. Spatial Context Awareness for Unsupervised Change Detection in Optical Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15, doi:10.1109/TGRS.2021.3130842.
17. Bdiri, W.; Bouhlel, N.; Meric, S.; Pottier, E.; Kallel, F. A Bayesian Nonparametric Model for Unsupervised Change Detection of Fully Polarimetric Sar Images. In Proceedings of the IGARSS 2024-2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, IGARSS 2024; IEEE: Athens, GREECE, 2024; pp. 2789–2794.
18. Bergamasco, L.; Martinatti, L.; Bovoloni, F.; Bruzzone, L. An Unsupervised Change Detection Technique Based on a Super-Resolution Convolutional Autoencoder. In Proceedings of the 2021 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM IGARSS; IEEE: Electr Network, 2021; pp. 3337–3340.
19. Hu, L.; Liu, J.; Xiao, L. A Total Variation Regularized Bipartite Network for Unsupervised Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18, doi:10.1109/TGRS.2022.3224293.
20. Wu, C.; Chen, H.; Du, B.; Zhang, L. Unsupervised Change Detection in Multitemporal VHR Images Based on Deep Kernel PCA Convolutional Mapping Network. *IEEE Trans. Cybern.* **2022**, *52*, 12084–12098, doi:10.1109/TCYB.2021.3086884.
21. Noh, H.; Ju, J.; Seo, M.; Park, J.; Choi, D.-G. Unsupervised Change Detection Based on Image Reconstruction Loss. In Proceedings of the 2022 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, CVPRW 2022; IEEE: New Orleans, LA, 2022; pp. 1351–1360.
22. Liu, G.; Yuan, Y.; Zhang, Y.; Dong, Y.; Li, X. Style Transformation-Based SpatialSpectral Feature Learning for Unsupervised Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15, doi:10.1109/TGRS.2020.3026099.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

23. Fang, B.; Chen, G.; Kou, R.; Paoletti, M.E.; Haut, J.M.; Plaza, A. CIT: Content-Invariant Translation with Hybrid Attention Mechanism for Unsupervised Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 321–339, doi:10.1016/j.isprsjprs.2023.09.012.
24. Zhou, Y.; Li, X.; Chen, K.; Kung, S.-Y. Progressive Learning for Unsupervised Change Detection on Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13, doi:10.1109/TGRS.2023.3235981.
25. Touati, R.; Mignotte, M.; Dahmane, M. Multimodal Change Detection in Remote Sensing Images Using an Unsupervised Pixel Pairwise-Based Markov Random Field Model. *IEEE Trans. Image Process.* **2020**, *29*, 757–767, doi:10.1109/TIP.2019.2933747.
26. Mignotte, M. A Fractal Projection and Markovian Segmentation-Based Approach for Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8046–8058, doi:10.1109/TGRS.2020.2986239.
27. Sun, Y.; Lei, L.; Guan, D.; Wu, J.; Kuang, G. Iterative Structure Transformation and Conditional Random Field Based Method for Unsupervised Multimodal Change Detection. *Pattern Recognit.* **2022**, *131*, 108845, doi:10.1016/j.patcog.2022.108845.
28. Sun, Y.; Lei, L.; Li, Z.; Kuang, G. Similarity and Dissimilarity Relationships Based Graphs for Multimodal Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 70–88, doi:10.1016/j.isprsjprs.2024.01.002.
29. Sun, Y.; Lei, L.; Tan, X.; Guan, D.; Wu, J.; Kuang, G. Structured Graph Based Image Regression for Unsupervised Multimodal Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *185*, 16–31, doi:10.1016/j.isprsjprs.2022.01.004.
30. Sun, Y.; Lei, L.; Guan, D.; Kuang, G.; Li, Z.; Liu, L. Locality Preservation for Unsupervised Multimodal Change Detection in Remote Sensing Imagery. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 6955–6969, doi:10.1109/TNNLS.2024.3401696.
31. Han, T.; Tang, Y.; Chen, Y.; Zou, B.; Feng, H. Global Structure Graph Mapping for Multimodal Change Detection. *Int. J. Digit. Earth* **2024**, *17*, doi:10.1080/17538947.2024.2347457.
32. Han, T.; Tang, Y.; Zou, B.; Feng, H. Unsupervised Multimodal Change Detection Based on Adaptive Optimization of Structured Graph. *Int. J. Appl. Earth Obs. Geoinformation* **2024**, *126*, 103630, doi:10.1016/j.jag.2023.103630.
33. Han, T.; Tang, Y.; Chen, Y.; Yang, X.; Guo, Y.; Jiang, S. SDC-GAE: Structural Difference Compensation Graph Autoencoder for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16, doi:10.1109/TGRS.2024.3396141.
34. Liu, T.; Xu, J.; Lei, T.; Wang, Y.; Du, X.; Zhang, W.; Lv, Z.; Gong, M. AEKAN: Exploring Superpixel-Based AutoEncoder Kolmogorov-Arnold Network for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14, doi:10.1109/TGRS.2024.3515258.
35. Liu, T.; Zhang, M.; Gong, M.; Zhang, Q.; Jiang, F.; Zheng, H.; Lu, D. Commonality Feature Representation Learning for Unsupervised Multimodal Change Detection. *IEEE Trans. Image Process.* **2025**, *34*, 1219–1233, doi:10.1109/TIP.2025.3539461.
36. Liu, T.; Pu, Y.; Lei, T.; Xu, J.; Gong, M.; He, L.; Nandi, A.K. Hierarchical Feature Alignment-Based Progressive Addition Network for Multimodal Change Detection. *Pattern Recognit.* **2025**, *162*, 111355, doi:10.1016/j.patcog.2025.111355.
37. Cai, L.; Xu, S.; Sun, H.; Sun, X.; Yang, L.; Gao, L. Primary Modality Guided Multimodal Change Detection. In Proceedings of the 2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2024); IEEE: Athens, GREECE, 2024; pp. 10323–10327.
38. Cai, L.; Sun, H.; Sun, X.; Yan, H.; Gao, L. HF-MCD: A Heterogeneous Fusion Framework for Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15, doi:10.1109/TGRS.2025.3606546.
39. Liu, B.; Chen, H.; Li, K.; Yang, M.Y. Transformer-Based Multimodal Change Detection with Multitask Consistency Constraints. *Inf. Fusion* **2024**, *108*, 102358, doi:10.1016/j.inffus.2024.102358.
40. Jiang, F.; Huang, B.; Wu, H.; Feng, D.; Zhou, Y.; Zhang, M.; Gong, M.; Zhao, W.; Guan, Z. Change Masked Modality Alignment Network for Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16, doi:10.1109/TGRS.2024.3516001.
41. Wang, J.; Yan, L.; Yang, J.; Xie, H.; Yuan, Q.; Wei, P.; Gao, Z.; Zhang, C.; Atkinson, P.M. MaCon: A Generic Self-Supervised Framework for Unsupervised Multimodal Change Detection. *IEEE Trans. Image Process.* **2025**, *34*, 1485–1500, doi:10.1109/TIP.2025.3542276.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

42. Pu, Y.; Gong, M.; Liu, T.; Zhang, M.; Gao, T.; Jiang, F.; Hu, X. Adversarial Feature Equilibrium Network for Multimodal Change Detection in Heterogeneous Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17, doi:10.1109/TGRS.2024.3480091.
43. Hou, X.; Bai, Y.; Xie, Y.; Zhang, Y.; Fu, L.; Li, Y.; Shang, C.; Shen, Q. Self-Supervised Multimodal Change Detection Based on Difference Contrast Learning for Remote Sensing Imagery. *Pattern Recognit.* **2025**, *159*, 111148, doi:10.1016/j.patcog.2024.111148.
44. Huang, B.; Lu, Y.; Yin, C.; Yang, R.; Tao, Y.; Shi, Y.; Wang, S.; Zhao, Q. DBASNet: A Double-Branch Adaptive Segmentation Network for Remote Sensing Image. *Pattern Recognit. Lett.* **2026**, *201*, 9–14, doi:10.1016/j.patrec.2025.11.043.
45. Zhang, J.; Ding, L.; Zhou, T.; Wang, J.; Atkinson, P.M.; Bruzzone, L. Recurrent Semantic Change Detection in VHR Remote Sensing Images Using Visual Foundation Models. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14, doi:10.1109/TGRS.2025.3546808.
46. Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Yang, K.; Bruzzone, L. Adapting Segment Anything Model for Change Detection in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11, doi:10.1109/TGRS.2024.3368168.
47. Cheng, Y.; Yang, W.; Li, Y.; Xu, W.; Chang, L.; Li, R. Remote Sensing Semantic Change Detection Based on the Visual Foundation Model. In Proceedings of the 2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2024); IEEE: Athens, GREECE, 2024; pp. 10237–10240.
48. Peng, S.; Li, J.; Zhang, T. SAM-FDN: A SAM Fine-Tuning Adaptation Remote Sensing Change Detection Method Based on Fourier Frequency Domain Analysis Difference Reinforcement. *Remote Sens.* **2025**, *17*, 3842, doi:10.3390/rs17233842.
49. Wang, S.; Lv, C.; Quan, D.; Huyan, N.; Cao, X.; Sun, J.; Jiao, L. Burden-Free Distillation From Foundation Model for Efficient Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–13, doi:10.1109/TGRS.2025.3584094.
50. Wu, Z.; Zan, L.; Chen, Z.; Cai, M.; Li, Y.; Wang, Z.; Xie, J.; Shi, X. A Remote Sensing Image Change Detection Network With Feature Constraints From a Visual Foundation Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 28939–28956, doi:10.1109/JSTARS.2025.3626022.
51. Jiang, W.; Sun, Y.; Lei, L.; Kuang, G.; Ji, K. AdaptVFMs-RSCD: Advancing Remote Sensing Change Detection from Binary to Semantic with SAM and CLIP. *ISPRS J. Photogramm. Remote Sens.* **2025**, *230*, 304–317, doi:10.1016/j.isprsjprs.2025.09.010.
52. Shen, J.; Huo, C.; Xiang, S. Siamese InternImage for Change Detection. *Remote Sens.* **2024**, *16*, 3642, doi:10.3390/rs16193642.
53. Shi, Y.; Yang, R.; Yin, C.; Lu, Y.; Huang, B.; Tao, Y.; Zhong, Y. Two-Stage Fine-Tuning of Large Vision-Language Models with Hierarchical Prompting for Few-Shot Object Detection in Remote Sensing Images. *Remote Sens.* **2026**, *18*, 266, doi:10.3390/rs18020266.
54. Fuller, A.; Millard, K.; Green, J.R. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders 2023.
55. Beusen, B.; Luyts, A.; Ivashkovych, X.; van Achteren, T. Lightweight and Efficient: A Family of Multimodal Earth Observation Foundation Models. In Proceedings of the IGARSS 2024-2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, IGARSS 2024; IEEE: Athens, GREECE, 2024; pp. 2841–2846.
56. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16, doi:10.1109/TGRS.2024.3390838.
57. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15, doi:10.1109/TGRS.2022.3222818.
58. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing Remote Sensing Foundation Model via Multitask Pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11632–11654, doi:10.1109/JSTARS.2024.3408154.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2026 doi:10.20944/preprints202601.2350.v1

59. Nie, H.; Luo, B.; Liu, J.; Fu, Z.; Zhou, H.; Zhang, S.; Liu, W. PromptMID: Modal Invariant Descriptors Based on Diffusion and Vision Foundation Models for Optical-SAR Image Matching. *ISPRS J. Photogramm. Remote Sens.* **2025**, *230*, 192–207, doi:10.1016/j.isprsjprs.2025.08.030.
60. Lu, S.; Guo, J.; Zimmer-Dauphinee, J.R.; Nieusma, J.M.; Wang, X.; Vanvalkenburgh, P.; Wernke, S.A.; Huo, Y. Vision Foundation Models in Remote Sensing: A Survey. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 190–215, doi:10.1109/MGRS.2025.3541952.
61. Ding, L.; Zuo, X.; Hong, D.; Guo, H.; Lu, J.; Gong, Z.; Bruzzone, L. S2C: Learning Noise-Resistant Differences for Unsupervised Change Detection in Multimodal Remote Sensing Images 2025.
62. Luppino, L.T.; Bianchi, F.M.; Moser, G.; Anfinsen, S.N. Unsupervised Image Regression for Heterogeneous Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9960–9975, doi:10.1109/TGRS.2019.2930348.
63. Touati, R.; Mignotte, M.; Dahmane, M. Multimodal Change Detection in Remote Sensing Images Using an Unsupervised Pixel Pairwise-Based Markov Random Field Model. *IEEE Trans. Image Process.* **2020**, *29*, 757–767, doi:10.1109/TIP.2019.2933747.
64. Touati, R.; Mignotte, M.; Dahmane, M. Anomaly Feature Learning for Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 588–600, doi:10.1109/JSTARS.2020.2964409.
65. Sun, Y.; Lei, L.; Guan, D.; Kuang, G. Iterative Robust Graph for Unsupervised Change Detection of Heterogeneous Remote Sensing Images. *IEEE Trans. Image Process.* **2021**, *30*, 6277–6291, doi:10.1109/TIP.2021.3093766.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.