IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1295-1305
Journal homepage: http://iieta.org/journals/ts

HieraChange: A Hierarchical Vision Transformer Model for Change Detection in Satellite Images

Rajalakshmi V[1*], Kala A[2], Sharon Femi P[2], Khanaghavalle G R[1]

[1] Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur 602117, India
[2] Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur 602117, India

Corresponding Author Email: vraji@svce.ac.in

https://doi.org/10.18280/ts.430317

Received: 18 March 2026
Revised: 30 May 2026
Accepted: 12 June 2026
Available online: 30 June 2026

Keywords:
hierarchical vision transformer, building change detection, satellite imagery, LEVIR-change detection dataset, Siamese network, Swin Transformer, remote sensing, multi-scale feature extraction

ABSTRACT

Satellite-based building change detection (CD) plays a vital role in urban planning, disaster assessment, and environmental monitoring by enabling accurate identification of infrastructure changes over time. This paper proposes HieraChange, a hierarchical vision transformer framework for building CD using bi-temporal satellite imagery. The proposed model employs a shared-weight Siamese encoder based on the Swin Transformer to extract multi-scale hierarchical features from pre-change and post-change images. Feature differencing and multi-scale aggregation are performed to capture temporal variations, while a lightweight multilayer perceptron (MLP) based decoder reconstructs the final change map. To improve segmentation performance under class imbalance, the model is optimized using a hybrid Binary Cross-Entropy (BCE) and Dice loss function. The proposed framework is evaluated on the LEVIR-CD dataset using Precision, Recall, F1-score, and Intersection over Union (IoU). Experimental results demonstrate that HieraChange achieves a precision of 97.56%, recall of 96.12%, F1-score of 96.23%, and IoU of 92.70%, indicating competitive performance for satellite-based building CD. The proposed framework provides an effective approach for accurate infrastructure CD in high-resolution remote sensing (RS) imagery.

1. INTRODUCTION

The rapid expansion of urban areas, increasing frequency of natural disasters, and growing environmental concerns necessitate efficient monitoring of infrastructure changes. Satellite imagery has emerged as a crucial tool for tracking dynamic transformations in cities, ecosystems, and disaster regions. However, accurately detecting and analyzing these changes remains a complex challenge due to data heterogeneity, scale variations, and computational constraints. Traditional methods often struggle to provide high-precision results while maintaining efficiency across diverse scenarios. Recent advancements in deep learning (DL), particularly in transformer-based models and Siamese architectures, provide strong and effective approaches for tackling these challenges.

The HieraChange model presented in this work has a hierarchical design that plays a pivotal role in analyzing complex satellite data, where spatial details vary significantly across scales. By employing a hierarchical transformer encoder, HieraChange progressively extracts multi-scale features, ensuring the model captures both fine-grained and large-scale patterns. This design improves the ability of the model to detect even slight changes that may otherwise go unnoticed. The hierarchical structure also enables progressive refinement of feature maps, enhancing the model's ability to generalize across diverse environments such as urban areas, forests, and coastal regions. The primary objective of this study is to develop and evaluate a hierarchical vision transformer-based framework, named HieraChange, for binary building change detection (CD) using bi-temporal high-resolution satellite imagery. The proposed framework employs a shared-weight Siamese Swin Transformer encoder, hierarchical feature differencing, multi-scale feature aggregation, and a multilayer perceptron (MLP)-based decoder to accurately identify building changes. The experimental evaluation is conducted exclusively on the LEVIR-CD dataset using standard CD metrics, including precision, recall, F1-score, and Intersection over Union (IoU). Although building CD is a fundamental component of broader remote sensing (RS) applications such as urban planning, disaster assessment, environmental monitoring, and land-use analysis, these applications are beyond the scope of the present study and are considered potential directions for future research.

Existing transformer-based CD models, including DHT-CD, GAS-Net, and EfficientCD, have demonstrated the effectiveness of hierarchical feature extraction, Siamese learning, and feature fusion for RS CD. DHT-CD primarily employs hierarchical transformer encoders with a convolutional (U-Net-based) decoder for multi-scale feature reconstruction, whereas GAS-Net focuses on enhancing global contextual representation through a global-aware Siamese architecture. EfficientCD emphasizes computational efficiency by exchanging bi-temporal features while maintaining competitive detection performance. In contrast, the proposed HieraChange framework integrates these

strengths within a unified architecture by performing explicit stage-wise hierarchical difference computation at every transformer encoder level before feature aggregation, thereby preserving temporal variations across multiple spatial scales. The resulting multi-level difference features are reconstructed using a lightweight MLP decoder instead of a computationally intensive convolutional decoder, reducing computational complexity while maintaining high detection accuracy. Furthermore, the integration of Binary Cross-Entropy (BCE) and Dice loss provides balanced optimization, improving the localization of small infrastructure changes and mitigating class imbalance. These architectural enhancements enable HieraChange to achieve improved accuracy and efficiency, making it a scalable solution for large-scale satellite image CD.

The main contributions of this work are as follows:

1. A hierarchical vision transformer framework with stage-wise multi-scale difference computation for enhanced temporal feature representation.
2. A lightweight MLP decoder that reconstructs change maps with lower computational complexity than conventional convolutional decoders.
3. A hybrid BCE-Dice loss optimization strategy that improves localization of small changes while addressing severe class imbalance.
4. Comprehensive evaluation on the LEVIR-CD dataset, demonstrating superior Precision (97.56%), Recall (96.12%), F1-score (96.23%), and IoU (92.70%) compared with recent state-of-the-art methods.

2. RELATED WORK

Song et al. [1] introduced an Object-Based Image Analysis (OBIA) approach that proves more effective than traditional pixel-level methods for identifying changes in very high-resolution (VHR) satellite imagery. Their work presents a DL-driven OBIA framework that incorporates object-level uncertainty estimation to perform CD without relying on ground truth labels. The method begins by generating change objects using unsupervised techniques, which are then refined through a deep network employing 3D convolutions and ConvLSTMs. Object labels are iteratively updated based on uncertainty measures until all regions are confidently classified. Experimental results demonstrate that this approach outperforms conventional methods, offering strong noise resistance and precise object-level change identification.

Siddique and Ahmed [2] introduced a DL-based coherent change detection model (CCD-Conv1D) using Sentinel-1 SAR images to monitor floods and predict affected areas. The model combines image segmentation, log-ratio analysis, and DL to generate accurate flood maps. Experiments in Ayodhya and Basti showed improved CD accuracy and significant inundation over vegetation areas. The DL-based forecasts using Conv1D and Naïve Forecast models proved more reliable than traditional methods, supporting future development of effective flood monitoring systems.

Zhu et al. [3] emphasized that high spatial resolution (HSR) RS images are crucial for examining land-use transitions, yet many traditional methods fail to integrate semantic class information. To overcome this limitation, a new Siamese Global Learning (Siam-GL) framework is introduced, integrating a shared-parameter Siamese network with a global hierarchical sampling strategy to address sample imbalance and improve semantic CD. Experimental results across several HSR datasets demonstrate that Siam-GL delivers higher accuracy and stronger generalization capabilities than existing techniques.

Xiao et al. [4] addressed multi-spatial-resolution CD, which predicts the differences between image pairs of varying resolutions by aligning them and using unsupervised deep feature learning. The proposed framework uses co-registration, denoising autoencoders, and a mapping network to bridge resolution gaps and generate accurate change maps, showing superior results across four real datasets.

Farooq and Manocha [5] discussed that satellite-based CD is crucial for monitoring urban growth, land use, disasters, and environmental changes using multi-temporal RS data. DL methods have shown superior performance over traditional techniques, though some challenges remain. The review also highlights AI-based approaches and outlines future directions for advancing urban CD research.

Patil et al. [6] presented EffCDNet, which integrates a Siamese EfficientNet encoder with an Attention-UNet decoder and UDWT-based post-processing to improve feature extraction and produce noise-resistant change maps. Results on benchmark datasets demonstrate that EffCDNet surpasses existing approaches in both accuracy (IoU) and inference speed.

Gite and Gupta [7] proposed a novel approach that integrates Fuzzy Neural Network (FNN) classification with segmentation performed using a Taylor Shuffled Shepherd Optimization (TSSO)-based GAN. The TSSO method combines the Taylor series with Shuffled Shepherd Optimization for more effective feature learning. Experimental results confirm that the method performs highly with 0.932 accuracy, 0.0704 error rate, and 0.911 kappa coefficient. Compared to KPCA-MNet, MSVM, PPCNET, and IFN, the approach improves accuracy by 2.28%, 4.78%, 0.33%, and 13.26%, respectively, proving its superiority.

Dong et al. [8] introduced EfficientCD, a DL framework that employs EfficientNet for feature extraction along with a novel ChangeFPN module to enhance bi-temporal feature interaction. The model also incorporates a progressive upsampling module that uses Euclidean distance to improve feature fusion and reconstruction. Experiments on the LEVIR-CD, SYSU-CD, CLCD, and WHU-CD datasets indicate that EfficientCD achieves state-of-the-art performance in CD.

Chughtai et al. [9] compared traditional and modern approaches, highlighting the effectiveness of post-classification techniques. Among them, the Maximum Likelihood Classifier (MLC) consistently delivers the highest accuracy across diverse regions and remains the most reliable method for LULC CD.

Lim et al. [10] proposed a CNN-based CD method for high-resolution satellite imagery that removes the need for preprocessing steps such as ortho-rectification or classification. One of the major challenges in multi-temporal image analysis is separating genuine changes from variations in viewpoint or color, particularly in urban areas where registration errors from tall buildings often undermine traditional techniques. To overcome this, the authors designed three encoder-decoder CNNs capable of generating change maps directly from RGB image pairs. They also built a large, fully annotated dataset from Google Earth images spanning different years and seasons to enable supervised training. Experimental results demonstrate that the CNN models can reliably identify true changes despite imperfect alignment, and

their ensemble outperforms each individual model.

Afaq and Manocha [11] provided a comprehensive review of digital CD techniques, noting that change vector analysis delivers the highest accuracy among algebra-based methods, while discrete wavelet transformation outperforms other transformation-based approaches. The study also highlights that neural networks and fuzzy-based techniques generally surpass traditional methods, and although DL shows strong potential, continued advancements are still needed within the RS field.

Satellite imagery received at ground stations is frequently degraded by noise such as impulse, speckle, and Gaussian noise. Conventional denoising techniques can reduce noise but often blur edges, resulting in a loss of important details. To overcome this issue, Asokan and Anitha [12] developed an optimized bilateral filter enhanced with nature-inspired optimization algorithms. The filter's parameters are fine-tuned using Particle Swarm Optimization (PSO), Cuckoo Search (CS), and Adaptive Cuckoo Search (ACS). Experimental results indicate that the ACS-optimized bilateral filter outperforms traditional methods, including Median and RGB spatial filters, across metrics such as PSNR, MSE, FSIM, entropy, and processing time.

Feizizadeh et al. [13] proposed an OBIA method for landslide mapping and CD using multi-temporal satellite imagery. The approach integrates spatial and spectral features, including shape, texture (via GLCM), and fuzzy logic after initial image segmentation. Multi-resolution segmentation was applied to IRS-1D, SPOT-5, and ALOS datasets, incorporating DEM-derived slope and flow direction information. Classification was performed using fuzzy membership functions and operators, with Fuzzy Synthetic Evaluation (FSE) used for validation. The AND operator produced the highest accuracy (93.87% in 2005 and 94.74% in 2011), demonstrating the effectiveness of object-based CD in monitoring landslide evolution in northern Iran.

Zhang et al. [14] proposed GAS-Net, a global-aware Siamese network developed to model comprehensive contextual relationships between scenes and foreground regions. The framework integrates a Global-Attention Module (GAM) and a Foreground-Awareness Module (FAM), enhancing contextual reasoning and joint feature learning. Experiments conducted on the LEVIR-CD and Lebedev-CD datasets demonstrate strong robustness, yielding F1-scores of 91.21% and 95.84%, respectively.

Liu et al. [15] proposed a hierarchical transformer architecture that leverages self-attention to efficiently capture long-range dependencies in multi-temporal imagery. Local windowed self-attention is used to lower computational cost while improving feature representation. The extracted features are fused and processed through a nested U-Net, with an additional model fusion strategy further enhancing accuracy. Evaluation on the LEVIR-CD and SYSU-CD datasets demonstrates superior performance compared to existing methods.

Priyadarshi and Singh [16] presented a fully convolutional Siamese auto-encoder framework that produces change maps from bi-temporal, co-registered satellite imagery. The method is applied to detect land-cover changes such as encroachment, building expansion, and unauthorized construction. Due to the absence of a public multi-label CD dataset, a new hand-annotated dataset was created. The proposed approach achieves a 3.9% improvement in F1-score over state-of-the-art methods and significantly reduces training requirements, needing only 50 hours for any number of labels, compared to 30 hours per label in previous CNN-based techniques.

Wang et al. [17] introduced a Spatiotemporal Enhanced CD (STECD) algorithm for monitoring flood changes using multi-source heterogeneous earth observation image time series (EOITS). The method combines spatial clustering, where a sparse Markov random field (SMRF) optimizes flood boundaries using contextual spatial features, with temporal enhancement, which leverages historical flood information to reduce errors caused by terrain, cloud, and topographic shadows. By jointly exploiting spatial and temporal dependencies, the STECD algorithm improves the accuracy of flood area detection. Experimental results on real EOITS datasets confirm that STECD outperforms widely used CD methods. Overall, it provides a more robust solution for flood hazard monitoring.

Cheng et al. [18] proposed a Multi-scale Cross-modal Fusion Network (MCFNet) that leverages style transfer to integrate content and style features across multiple scales for improved fusion. Experimental evaluations on three datasets demonstrate that MCFNet surpasses existing approaches, achieving at least a 4% improvement in Kappa accuracy.

Recent research shows that translating multimodal RS images into a common domain can reduce incompatibility issues in CD. However, most methods ignore global context, causing semantic confusion and reduced accuracy. To overcome this, a pseudo-siamese GAN (PS-GAN) is proposed by Zhu et al. [19], which extracts all global and local features collectively via a two-branch encoder and fuses them using an adaptive multi-scale module, ensuring structurally intact translations. A U2Net+-based CD model with gated feature modulation and enhanced squeeze-excitation further improves semantic feature representation. Experiments on four datasets show PS-GAN outperforms 11 state-of-the-art multimodal CD methods with higher accuracy.

Heterogeneous RS with SAR and optical data enables reliable anytime, all-season CD. To address feature loss and noise in existing two-stage methods, a domain adaptation-based multi-source CD network (DA-MSCDNet) is proposed by Zhang et al. [20], aligning deep feature spaces across heterogeneous data for improved detection. Experiments on multiple datasets, including Sentinel, Landsat, QuickBird, and COSMO-SkyMed, show DA-MSCDNet outperforms six existing methods with higher accuracy (F1-score 82.58%) and 10× faster training efficiency.

Zhang et al. [21] introduced a bipartite unsupervised residual network (ResNet) for CD using heterogeneous radar and optical images captured at different times. Unlike traditional shallow pixel-level methods, the proposed approach operates at the feature level to better capture differences. The bipartite deep ResNet fully leverages joint multi-temporal information, generating representative features of image changes. A novel loss function is designed to optimize the ResNet training process. Experiments on both heterogeneous and homogeneous datasets show that the method outperforms existing CD approaches.

CD in heterogeneous RS imagery is crucial for accurate disaster damage evaluation, yet current homogeneous transformation techniques often degrade semantic information because they rely solely on low-level feature representations. To address this, a Deep Homogeneous Feature Fusion (DHFF) model based on image style transfer (IST) is proposed by Jiang et al. [22], which separates semantic content and style features for accurate transformation. An iterative IST strategy further

enhances feature homogeneity across new subspaces, improving detection performance. Experiments on SAR and optical satellite images show that DHFF significantly improves accuracy and Kappa index compared to existing methods.

CD using RS images is essential for monitoring surface changes, but existing Siamese networks often lack sufficient contextual information, leading to false or missed detections and poor edge refinement. To overcome this, SSANet was proposed by Jiang et al. [23]. It is a spatial-temporal attention network that is self-weighted with a combined learning framework of different subnetworks that include fusion, difference, and decoding. The fusion subnetwork contains the MCA module that aggregates multiscale semantic context, while the difference subnetwork, which contains the FDR module, extracts and reconstructs temporal change features. An AW module in the decoder balances these features, enabling SSANet to deliver superior change maps and achieve state-of-the-art performance.

Wang et al. [24] introduced an unsupervised CD method tailored for optical sensors, leveraging a fusion framework that is robust. By augmenting a pre-trained optical image fusion network with a parallel network of identical architecture and integrating both into an adversarial learning framework, the method facilitates effective CD. Experimental comparisons show that the proposed method is both versatile and superior to existing state-of-the-art techniques.

Li et al. [25] proposed a spatially self-paced convolutional network (SSPCN) for unsupervised CD. Pseudo labels are first generated, and self-paced learning (SPL) dynamically selects reliable samples, starting with easier ones and gradually incorporating complex ones. The model assigns adaptive weights to samples, updating them during training to improve robustness and convergence. By integrating spatial information, SSPCN enhances stability against noisy data, and experiments on four heterogeneous datasets demonstrate its superior effectiveness.

Tong et al. [26] presented a novel method for detecting multiple types of changes in bi-temporal hyperspectral images using only a small set of source-image samples. The workflow consists of four key stages: performing unsupervised CD through uncertain area analysis, classifying the source image via active learning, classifying the target image using an enhanced transfer learning strategy, and finally generating a multiclass change map. Experiments on both synthetic and real datasets demonstrate substantial improvements in binary and multiclass CD performance, surpassing state-of-the-art methods.

Sun et al. [27] proposed an unsupervised CD framework built on the patch similarity graph matrix (PSGM), which assumes that the graph structures of homogeneous or heterogeneous image pairs remain stable in unchanged regions. The PSGM for the first image is constructed using the self-expressive property, where each patch is represented as a graph vertex. Change intensities are then derived by assessing how closely the second image adheres to this learned structure. A sparse prior is applied to further refine the generated change map, and experiments on multiple datasets confirm the approach's robustness and strong effectiveness.

Sun et al. [28] proposed a structure consistency-based CD method that compares image structures instead of pixel values, offering robustness to noise, illumination, and modality differences. The approach uses an advanced nonlocal patch-based graph (NLPG) for structural comparison while avoiding data leakage. It is validated across homogeneous and heterogeneous CD, including the rarely explored multichannel SAR data. Experiments on six scenarios with 12 datasets demonstrate its robustness and accuracy as a unified CD framework.

Sun et al. [29] proposed a robust graph-mapping-based framework for unsupervised heterogeneous CD in RS images. Because heterogeneous image pairs still share modality-invariant structural information of ground objects, the method builds K-nearest neighbor graphs for each image and compares them via graph mapping to produce forward and backward difference images. These different images are then integrated using a Markovian co-segmentation model with co-graph cut to accurately identify changed regions. An iterative feedback mechanism further refines the process by updating graph construction to minimize the influence of changed neighbors. Experimental results across multiple datasets confirm the effectiveness and stability of the proposed approach.

Unsupervised CD is often limited to optical images from sensors with the same spectral and spatial resolution in RS, but this is not always feasible in real-world scenarios like emergencies or defense. Existing methods typically handle only specific cases of resolution disparity, often with resampling-based preprocessing that can discard useful information. Ferraris et al. [30] proposed a robust fusion-based framework that models both observed images as degraded versions of latent high-resolution images, differing only in sparse regions where changes occur. The approach formulates CD as an inverse problem solved via alternating minimization, demonstrating improved accuracy and versatility over state-of-the-art methods on real multi-band optical datasets.

3. DATASET DESCRIPTION

In this work, the LEVIR-CD dataset serves as a widely used benchmark for evaluating CD methods in RS and satellite imagery. It was introduced to facilitate the development and evaluation of DL models for identifying infrastructure changes over time. The dataset consists of 637 pairs of high-resolution (1024 × 1024) RS images that span a variety of urban environments, capturing real-world instances of infrastructure growth, expansion, and transformation. The dataset covers various change types such as building construction, demolition, road extensions, and other urban developments, in which the urban development dataset is used in this work. The image pairs are captured at different time periods, providing realistic temporal variations for both training and evaluation. The dataset is commonly split into training, validation, and test sets to enable systematic assessment of model performance.

4. PROPOSED HIERACHANGE: A HIERARCHICAL VISION TRANSFORMER MODEL

HieraChange is an advanced model designed for detecting infrastructure changes in satellite imagery. By combining a hierarchical transformer encoder with a lightweight MLP decoder, it effectively captures both fine-grained and large-scale patterns. The Siamese architecture in the model enables precise identification of meaningful differences in the images. The detailed architectural description is as follows:

4.1 Overall architecture

A Hierarchical Transformer Encoder is a specialized architecture designed to extract multi-scale features from complex data, particularly in vision tasks like satellite image analysis. It employs multiple transformer layers that progressively refine feature maps, capturing both local details and global patterns. By processing data at varying resolutions, it efficiently handles scale variations and enhances contextual understanding. This design enhances the model's capacity to identify subtle changes and complex patterns across varied environments. Its hierarchical structure effectively balances computational efficiency with strong feature extraction performance. In this framework, the encoder processes both input images through transformer blocks at multiple levels, generating rich multi-scale feature representations for accurate CD.

Figure 1 represents a hierarchical vision transformer CD network designed to compare the two remote-sensing images that are captured at various periods of time. Each image is passed through an identical encoder that uses downsampling layers and Transformer blocks to progressively extract multi-scale feature representations. Because the two encoder branches share weights, they produce comparable feature maps from both time periods, ensuring consistent feature extraction.

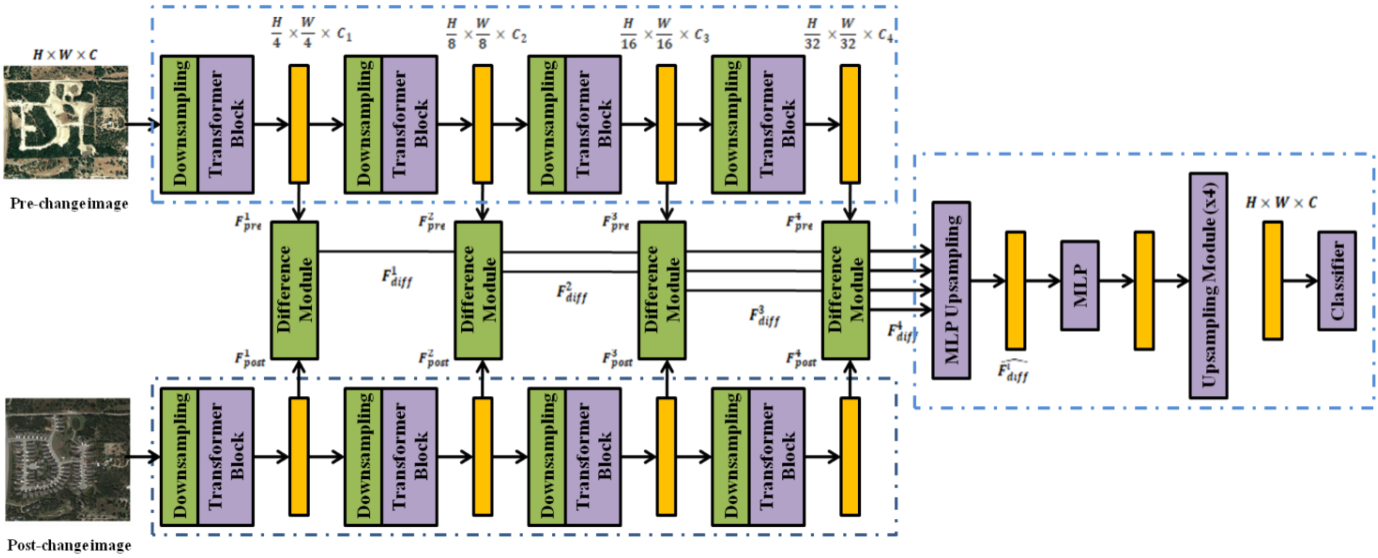

Figure 1. Architecture of the proposed HieraChange hierarchical vision transformer model

At every stage of the encoder, the respective feature maps from both images are sent to a difference module. These modules compute differences at multiple resolutions, capturing both fine-grained and large-scale changes. This multi-level differencing helps the model identify changes in complex scenes, even when the spatial resolution varies or the changes are subtle.

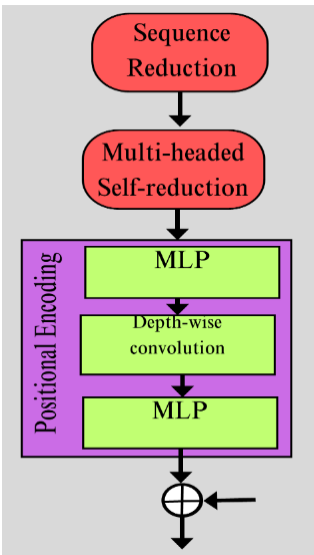

Figure 2. Transformer block

All the different features are then processed by a decoder that uses MLP transformations and multi-scale upsampling to gradually restore full spatial resolution. Finally, a classifier produces an H × W change map, identifying which pixels have changed between the two time steps. This combination of multi-level transformer features, difference modules, and upsampling enables accurate and detailed CD.

The transformer block, as shown in Figure 2, begins with a Sequence Reduction step, which compresses the input token sequence to a shorter one. This reduces computational cost by lowering the number of tokens processed in attention. After that, the compact sequence is passed through a Multi-Headed Self-Reduction module, a variation of multi-head self-attention that operates on the reduced sequence. This step allows the model to capture global contextual relationships among tokens while maintaining efficiency.

The output of the reduction modules enters a block that integrates positional encoding, ensuring the model retains information about spatial order. Inside this block, the features pass through an MLP layer, followed by a depth-wise convolution. The depth-wise convolution enhances local spatial information-something traditional Transformers lack-making the representation more suitable for vision tasks. Another MLP further refines the representation, allowing the model to combine features that are local as well as global effectively.

The MLP upsampling module, as shown in Figure 3, takes a multi-scale difference feature F^i_{diff} of size $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$

and first passes it through an MLP to unify or project the channel dimension to a fixed size CCC. This transformed feature is then fed into an upsampling layer, which increases its spatial resolution from $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$ to the target resolution $\frac{H}{4} \times \frac{W}{4}$. The final output, $\widehat{F_{diff}^{i}}$, is therefore a spatially enlarged and channel-aligned version of the original difference feature, allowing multi-level features from different scales to be fused consistently in the decoder.

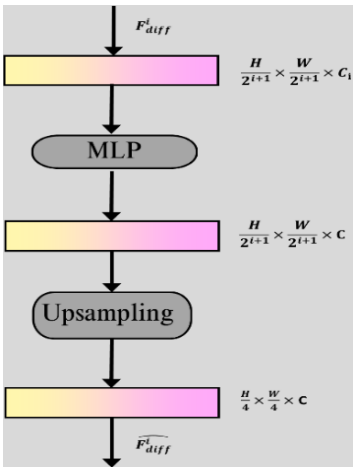

Figure 3. Multilayer perceptron (MLP) upsampling process

4.1.1 Feature extraction

Feature extraction refers to the process of converting raw data into informative representations that emphasize key patterns and characteristics. In DL models, this is typically achieved through convolutional layers, transformers, or specialized encoders that capture spatial, structural, or contextual information. Effective feature extraction reduces data complexity while retaining essential details for accurate predictions. It plays a crucial role in tasks like image recognition, CD, and object tracking. Well-extracted features enhance model performance by improving generalization and reducing noise. At each stage *'l'*, the encoder produces feature maps as shown in Eqs. (1) and (2):

$$F_{pre}^{l} = Transformer\ Block^{l}(I_{pre}) \tag{1}$$

$$F_{post}^{l} = Transformer\ Block^{l}(I_{post}) \tag{2}$$

4.1.2 Self-attention mechanism

The self-attention mechanism is a fundamental component of transformer models that allows each input element to evaluate and weight its relevance to every other element in the sequence. It calculates attention scores using query, key, and value matrices, allowing the transformer to focus on the most relevant regions of the input. This mechanism captures long-range relationships and complex patterns, improving the model's contextual understanding. By continuously adjusting attention weights, it enhances performance across tasks such as natural language processing, image interpretation, and CD. Its flexibility makes self-attention well-suited for handling variable-length inputs and diverse data types. In each transformer block, the self-attention mechanism computation is carried out in Eq. (3):

$$Attention(Q,K,V) = softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{3}$$

where

Q = $Wt.X$ is the query;
K = $Wt.X$ is the key;
V = $Wt.X$ is the value;
d is the dimension of the key vector.

4.1.3 Fusion of pre- and post-change features

The fusion of pre- and post-change features combines information from both time-stamped data to enhance CD accuracy. Techniques such as concatenation, element-wise addition, or attention mechanisms are used to merge features, highlighting meaningful differences. This fusion process strengthens the ability of the model to distinguish the changes from noise and irrelevant variations. This captures structural differences between the two images at each hierarchical level as stated in Eq. (4):

$$F_{diff}^{l} = F_{pre}^{l} - F_{post}^{l} \tag{4}$$

4.2 Feature aggregation

Feature aggregation is the process of combining multiple feature maps to create a richer and more informative representation. Methods like summation, concatenation, or weighted averaging are commonly used to combine the features from various layers or scales. This enhances the model's ability to capture diverse patterns and improves overall performance. The extracted features are aggregated across multiple scales as shown in Eq. (5):

$$F_{agg} = Concat(F_{diff}^{1}, F_{diff}^{2}, \ldots\ldots, F_{diff}^{L}) \tag{5}$$

where, L is the total number of hierarchical stages and *Concat(.)* is the channel-wise concatenation.

4.3 Lightweight multilayer perceptron decoder

A lightweight MLP decoder is a simplified yet effective module designed to refine extracted features and generate precise output predictions. It typically consists of a few fully connected layers with minimal parameters, ensuring fast computation and reduced memory usage. Despite its compact design, it efficiently maps complex feature representations to output spaces like segmentation masks or change maps. The lightweight structure makes it ideal for real-time applications and deployment on resource-constrained devices. Its balance between efficiency and accuracy enhances model performance in tasks like CD and image analysis. The decoder takes the aggregated features and processes them through several fully connected layers to predict change maps as shown in Eq. (6).

$$Y = \sigma(W_2.ReLU(W_1 F_{agg} + b_1) + b_2) \tag{6}$$

where, Y is the predicted output, W_1 and W_2 are the weights, b_1 and b_2 are the biases, σ is the sigmoid activation function for binary CD. Finally, the aggregated features are reconstructed as:

$$F_{dec} = MLP(F_{agg})$$

where, *MLP(.)* denotes the lightweight 'multilayer perceptron' decoder.

4.4 Loss function

A loss function is a mathematical tool used to measure how far a model's predictions deviate from the true target values. It plays a crucial role in optimization by supplying feedback that helps adjust the model's parameters. In this work, a combination of loss functions is employed during training to achieve precise and reliable CD.

4.4.1 Binary Cross-Entropy loss

BCE loss is widely used for binary classification problems. It measures how much the predicted probabilities differ from the actual labels. The loss is computed using the following formula in Eq. (7):

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i)] \quad (7)$$

This loss effectively penalizes incorrect predictions, encouraging the model to produce confident and accurate outputs.

4.4.2 Dice loss (for handling class imbalance)

Dice loss is a popular loss function in image segmentation, used to evaluate how well the predicted mask matches the ground-truth mask. It is based on the Dice Similarity Coefficient (DSC) and is expressed in Eq. (8) as:

$$L_{Dice} = 1 - \frac{2\sum y_i \hat{y}_i + \epsilon}{\sum y_i \hat{y}_i + \epsilon} \quad (8)$$

4.4.3 Total loss

The total loss is often a weighted combination, which is given in Eq. (9):

$$L_{Total} = \alpha L_{BCE} + \beta L_{Dice} \quad (9)$$

where, α and β are hyperparameters balancing the two losses.

4.4.4 Final prediction

The binary change mask is obtained as stated in Eq. (10):

$$M(x) = \begin{cases} 1, & P(x) \geq 0.5 \\ 0, & otherwise \end{cases} \quad (10)$$

The encoder used in the proposed HieraChange framework is based on the standard Swin Transformer architecture with a shared-weight Siamese configuration for processing pre-change and post-change satellite images. The encoder comprises four hierarchical stages, each consisting of alternating Window-based Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) blocks, followed by Patch Merging layers for progressive spatial downsampling and hierarchical feature learning. The model accepts an input image of size 256 × 256 × 3, which is partitioned into 4 × 4 non-overlapping patches before feature extraction. Multi-scale feature maps are generated at 1/4, 1/8, 1/16, and 1/32 of the original image resolution with progressively increasing channel dimensions, enabling the extraction of both local and global contextual information. The shared Siamese encoder processes the bi-temporal images using identical network weights to ensure consistent feature representation. The extracted multi-scale features are subsequently passed to the proposed hierarchical differencing module, where temporal differences are computed and aggregated across all encoder stages before being reconstructed using a lightweight MLP decoder.

5. RESULTS AND DISCUSSION

5.1 Input data

The LEVIR-CD dataset is a high-resolution benchmark tailored for detecting building changes in RS imagery. It consists of paired satellite images of the same geographic area taken at different time points, capturing changes in infrastructure and land use. Each image pair includes a "before" and "after" snapshot, accompanied by a ground truth change map that marks regions where changes have occurred. The dataset offers detailed imagery of buildings, roads, and other structures, enabling accurate identification of changes such as new constructions, demolitions, or structural modifications with a spatial resolution of 0.5 meters. Covering a range of urban and suburban environments, it presents diverse conditions ideal for evaluating the performance of CD algorithms. The dataset provides binary change maps as annotations, highlighting regions where changes have occurred, which supports supervised learning for CD models.

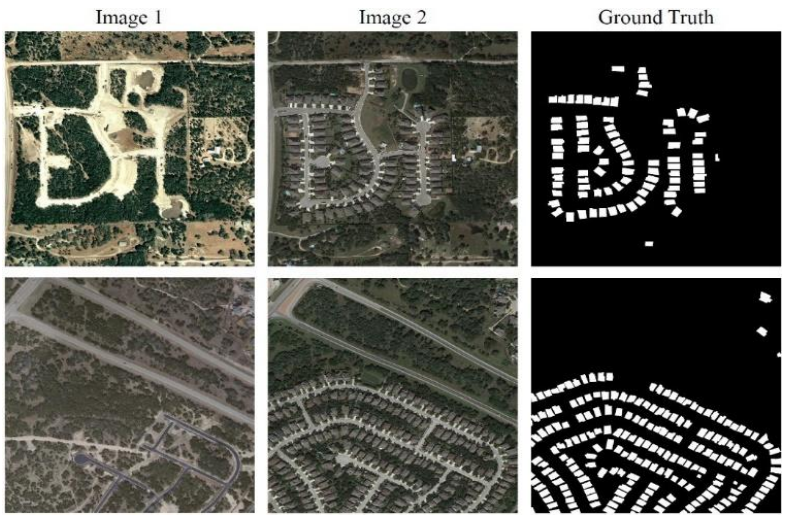

Figure 4. Sample data for training

Figure 4 illustrates the sample data used for training the model, which includes two sets, each containing three images. In each set, the first two sections display satellite images of the particular location taken at various periods of time, one before and one after a specific period. These images reveal various features such as buildings, roads, and open spaces. The visual differences between them indicate changes over time, including new constructions, demolitions, or alterations to existing structures. The third section presents the corresponding ground truth output, a binary change map essential for model training and evaluation. In this map, areas where significant changes have occurred are highlighted in white, while regions that remain unchanged are shown in black.

5.2 Change prediction model training and prediction

The hierarchical vision transformer model receives two satellite images as input: a pre-change I_{pre} and a post-change image I_{post}. Both images have dimensions $H \times W \times C$, corresponding to the height, width, and number of channels, respectively. The model processes a pair of satellite images, one captured before and the other after a change of the particular geographic location taken at various periods of time. It analyzes these images to identify alterations in infrastructure and land use. The result is a binary change map, resembling the ground truth, which visually highlights the regions where changes have taken place.

After training and evaluating the model, the focus shifts to interpreting the insights derived from the output binary change map. This map identifies the regions where changes have occurred in the original satellite images. However, while the change map highlights these areas, it can be difficult for a typical viewer to interpret the results clearly. To enhance clarity, the model is tuned to perform an additional step of shading the detected change regions in green color. This green overlay is then superimposed onto the original satellite image, effectively highlighting the changed areas as depicted on the right side of Figure 5. This visual enhancement makes it easier for users to understand and interpret the changes at a glance.

To gain a deeper understanding of the landscape, it's important to quantify the extent of change that has occurred. This is achieved by calculating the percentage of white and black pixels in the binary change map. Since the map highlights changed areas in white, determining the proportion of white pixels provides the percentage of land that has undergone development or modification. By subtracting this from the total number of pixels, we can determine the number of black pixels, which represent unchanged or undeveloped regions. These undeveloped areas can potentially be targeted for future infrastructure development.

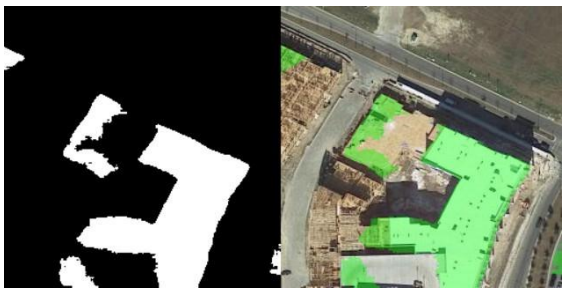

Figure 5. Visualization of predicted change regions with green overlay

Figure 6 displays a change map that visually depicts the detected changes in a specific area using black and white pixels. The white pixels highlight regions where significant modifications have occurred, such as new constructions or alterations to existing structures. In contrast, black pixels indicate areas where no changes have been detected. The accompanying text provides the calculated proportions of white and black pixels within the map. In this example, white pixels make up 12.93% of the total image, signifying that this portion of the area has undergone change. The remaining 87.07%, shown in black, corresponds to regions that have remained unchanged.

Figure 6. Bitmap image of change prediction

5.3 Experimental setup

The proposed HieraChange model was trained and evaluated on the LEVIR-CD dataset, which contains high-resolution bi-temporal satellite images and corresponding binary change masks. The original images have a resolution of 1024 × 1024 pixels and were preprocessed into 256 × 256 patches for efficient model training. Each pair includes a pre-change and post-change RGB image of size 256 × 256 pixels, along with a corresponding binary change mask. A total of 637 image pairs are utilized, divided into a 70/30 split for training and testing. The model is based on the Swin Transformer architecture, which employs two parallel transformer encoders to extract multi-scale features from the before-and-after images. These features, generated at multiple resolutions (such as 1/4, 1/8, and 1/16 of the original size), are integrated using a feature fusion module that may apply cross-attention or feature differencing, supported by skip connections to retain spatial information. A decoder then progressively upsamples and merges these features to create a binary change map, using a sigmoid activation at the output to determine changed versus unchanged regions.

The original LEVIR-CD dataset consists of bi-temporal satellite image pairs with a spatial resolution of 1024 × 1024 pixels. To facilitate efficient training and inference, each image pair was partitioned into 256 × 256 non-overlapping image patches, which served as the input to the proposed HieraChange model. This preprocessing step preserves the spatial information while reducing computational complexity. A total of 637 image pairs were used in this study and divided into 70% for training (446 image pairs) and 30% for testing (191 image pairs). A validation set comprising 10% of the training data was used for model selection. Data augmentation techniques, including random horizontal and vertical flipping, rotation, and random cropping, were applied during training to improve the model's generalization capability. A fixed random seed of 42 was used to ensure the reproducibility of the experimental results.

During training, the model is optimized using the AdamW optimizer with a batch size of 8, 100 training epochs, 1×10^{-5} weight decay, dropout of 0.1, an initial learning rate of 1×10^{-4}, and cosine annealing for learning-rate scheduling. No structural modifications are introduced to the Swin Transformer backbone; instead, the novelty of HieraChange lies in the hierarchical feature differencing strategy, multi-scale feature aggregation, lightweight MLP decoder, and hybrid BCE-Dice loss optimization, which together improve change localization while maintaining computational efficiency. Model performance is evaluated using precision, recall, F1-score, IoU, and overall accuracy. After prediction, the detected change regions are highlighted in green and overlaid on the original satellite images for easier visual interpretation. The percentage of land-use change is also computed by calculating the proportion of white pixels in the generated binary change map.

The proposed HieraChange model was implemented in PyTorch and trained on a workstation equipped with an NVIDIA RTX 3060 GPU (12 GB memory), Intel Core i7 processor, and 32 GB RAM. The implementation employed the AdamW optimizer and mixed-precision training to improve computational performance.

5.4 Performance analysis

Table 1 summarizes the proportion of pixels predicted as change and no-change for representative test images. The predicted change percentage varies from 12.93% to 25.26%, reflecting differences in the extent of infrastructure changes present in the selected scenes. Correspondingly, the predicted no-change percentage ranges from 74.74% to 87.07%, indicating that most image regions remain unchanged. These values describe the spatial distribution of the model's predictions for individual test images and illustrate the variability of detected changes across different urban environments. The quantitative performance of the proposed model is evaluated separately using Precision, Recall, F1-score, and IoU, as presented in Table 2.

Table 2 compares the proposed HieraChange model with state-of-the-art methods on the LEVIR-CD dataset using Precision, Recall, F1-score, and IoU. Early CNN-based approaches, such as Ensemble CNN and EffCDNet, achieved F1-scores of 86.60% and 90.80%, respectively, while transformer-based methods, including DHT-CD (Deep Hierarchical Transformer) and EfficientCD, further improved performance through enhanced feature extraction and temporal modeling. Among the existing approaches, GAS-Net (Global-Aware Siamese network) reported the best results with a Precision of 96.08%, Recall of 95.68%, F1-score of 95.88%, and IoU of 92.09%.

Table 1. Predicted percentage of change and no-change pixels in selected test images

Image Name	Predicted Change (%)	Predicted No-Change (%)
test_55_0256_0000.png	12.93%	87.07%
test_77_0512_0256.png	19.33%	80.67%
test_102_0512_0000.png	20.63%	79.37%
test_121_0768_0256.png	16.56%	83.44%
test_2_0000_0000.png	25.26%	74.74%
test_7_0256_0512.png	13.98%	86.02%
test_2_0000_0512.png	18.26%	81.74%

Table 2. Comparison with the existing models

Method	Year	Precision	Recall	F1-Score	IoU
Ensemble CNN [10]	2018	88.60	84.70	86.60	76.40
EffCDNet [6]	2021	91.80	89.90	90.80	83.10
GAS-Net [14]	2023	96.08	95.68	95.88	92.09
DHT-CD [15]	2023	93.70	92.50	93.10	87.00
EfficientCD [8]	2024	94.60	93.40	94.00	88.70
Proposed HieraChange	2026	97.56	96.12	96.23	92.7

The proposed HieraChange model achieved a Precision of 97.56%, Recall of 96.12%, F1-score of 96.23%, and an IoU of 92.70% on the LEVIR-CD dataset. The proposed HieraChange demonstrates competitive performance compared with recent state-of-the-art methods, including EffCDNet, DHT-CD, EfficientCD, and GAS-Net. Since these results are collected from different studies that may employ different experimental settings, data splits, and implementation details, the comparison is intended to provide a qualitative performance overview rather than a direct benchmark under identical evaluation conditions. IoU is a particularly important metric for CD because it directly measures the overlap between predicted change regions and ground-truth annotations, thereby providing a reliable evaluation of segmentation quality. The high IoU achieved by HieraChange indicates excellent localization of changed regions with minimal missed detections and false alarms. The hierarchical architecture captures both fine-grained local information and global contextual dependencies, enabling accurate and computationally efficient CD. The proposed HieraChange framework achieves competitive performance by integrating hierarchical feature extraction, Siamese feature differencing, multi-scale feature aggregation, and an MLP-based decoder. However, since an ablation study was not conducted in the present work, the individual contribution of each module has not been quantified. Future work will investigate the relative impact of these components through comprehensive ablation experiments. Overall, the experimental results demonstrate that the proposed HieraChange framework produces accurate and computationally efficient CD. Although the results indicate promising performance for satellite-based infrastructure monitoring, further evaluation on additional datasets and real-world scenarios is necessary to comprehensively assess its robustness and generalization capability.

6. CONCLUSION

The proposed HieraChange framework demonstrates effective performance for binary CD in new buildings using high-resolution bi-temporal satellite imagery. By integrating a shared-weight Siamese Swin Transformer encoder, hierarchical feature differencing, multi-scale feature aggregation, an MLP-based decoder, and a hybrid BCE–Dice loss function, the proposed HieraChange framework effectively detects the building changes in the LEVIR-CD dataset. Experimental results achieved a Precision of 97.56%, Recall of 96.12%, F1-score of 96.23%, and IoU of 92.70%, indicating competitive performance for building CD.

The present study is limited to evaluation on the LEVIR-CD benchmark dataset, and therefore the generalization of the

proposed framework to other datasets and RS scenarios has not been investigated. Future work will focus on cross-dataset validation, comprehensive computational efficiency analysis, ablation studies, and the extension of the proposed framework to broader RS applications, including urban monitoring and disaster assessment.

ACKNOWLEDGMENT

We would like to express our sincere appreciation to our colleagues at our institution for their valuable insights, thoughtful discussions, and constructive feedback, which significantly contributed to the successful completion of this research.

REFERENCES

[1] Song, A., Kim, Y., Han, Y. (2020). Uncertainty analysis for object-based change detection in very high-resolution satellite images using deep learning network. Remote Sensing, 12(15): 2345. https://doi.org/10.3390/rs12152345

[2] Siddique, M., Ahmed, T. (2025). CCD-Conv1D: A deep learning based coherent change detection technique to monitor and forecast floods using Sentinel-1 images. Remote Sensing Applications: Society and Environment, 37: 101440. https://doi.org/10.1016/j.rsase.2024.101440

[3] Zhu, Q.Q., Guo, X., Deng, W.H., et al. (2022). Land-Use/Land-Cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 184: 63-78. https://doi.org/10.1016/j.isprsjprs.2021.12.005

[4] Xiao, P.F., Zhang, X.L., Wang, D.G., Yuan, M., Feng, X.Z., Kelly, M. (2016). Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. ISPRS Journal of Photogrammetry and Remote Sensing, 119: 402-414. https://doi.org/10.1016/j.isprsjprs.2016.07.003

[5] Farooq, B., Manocha, A. (2024). Satellite-based change detection in multi-objective scenarios: A comprehensive review. Remote Sensing Applications: Society and Environment, 34: 101168. https://doi.org/10.1016/j.rsase.2024.101168

[6] Patil, P.S., Holambe, R.S., Waghmare, L.M. (2021). EffCDNet: Transfer learning with deep attention network for change detection in high spatial resolution satellite images. Digital Signal Processing, 118: 103250. https://doi.org/10.1016/j.dsp.2021.103250

[7] Gite, K.R., Gupta, P. (2023). GAN-FuzzyNN: Optimization based generative adversarial network and fuzzy neural network classification for change detection in satellite images. Sensing and Imaging, 24(1): 1. https://doi.org/10.1007/s11220-022-00404-3

[8] Dong, S.J., Zhu, Y.W., Chen, G., Meng, X.L. (2024). EfficientCD: A new strategy for change detection based with Bi-temporal layers exchanged. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-13. https://doi.org/10.1109/TGRS.2024.3433014

[9] Chughtai, A.H., Abbasi, H., Karas, I.R. (2021). A review on change detection method and accuracy assessment for land use land cover. Remote Sensing Applications: Society and Environment, 22: 100482. https://doi.org/10.1016/j.rsase.2021.100482

[10] Lim, K., Jin, D., Kim, C.S. (2018). Change detection in high resolution satellite images using an ensemble of convolutional neural networks. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, pp. 509-515. https://doi.org/10.23919/APSIPA.2018.8659603

[11] Afaq, Y., Manocha, A. (2021). Analysis on change detection techniques for remote sensing applications: A review. Ecological Informatics, 63: 101310. https://doi.org/10.1016/j.ecoinf.2021.101310

[12] Asokan, A., Anitha, J. (2020). Adaptive Cuckoo search based optimal bilateral filtering for denoising of satellite images. ISA Transactions, 100: 308-321. https://doi.org/10.1016/j.isatra.2019.11.008

[13] Feizizadeh, B., Blaschke, T., Tiede, D., Moghaddam, M.H.R. (2017). Evaluating fuzzy operators of an object-based image analysis for detecting landslides and their changes. Geomorphology, 293: 240-254. https://doi.org/10.1016/j.geomorph.2017.06.002

[14] Zhang, R.Q., Zhang, H.C., Ning, X.G., Huang, X., Wang, J., Cui, W. (2023). Global-aware siamese network for change detection on remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 199: 61-72. https://doi.org/10.1016/j.isprsjprs.2023.04.001

[15] Liu, B., Yu, A.Z., Zuo, X.B., Wang, R.R., Qiu, C.P., Yu, X.C. (2023). Deep hierarchical transformer for change detection in high-resolution remote sensing images. European Journal of Remote Sensing, 56(1): 2196641. https://doi.org/10.1080/22797254.2023.2196641

[16] Priyadarshi, A., Singh, P.K. (2021). Change detection in satellite imagery: A multi-label approach using convolutional neural network. In Hybrid Intelligent Systems, pp. 242-252. https://doi.org/10.1007/978-3-030-73050-5_24

[17] Wang, Z.H., Wang, X.Q., Li, G. (2022). Change detection of flood hazard areas in multi-source heterogeneous earth observation image time series based on spatiotemporal enhancement strategy. In Artificial Intelligence, pp. 453-465. https://doi.org/10.1007/978-3-031-20497-5_37

[18] Cheng, W., Feng, Y.N., Sun, Y.C., Wang, X.H. (2025). Heterogeneous remote sensing image change detection network based on multi-scale feature modal transformation. Applied Soft Computing, 170: 112725. https://doi.org/10.1016/j.asoc.2025.112725

[19] Zhu, Z.F., Yuan, X.P., Gan, S., Li, R.B., Bi, R., Luo, W.D.. (2025). PS-GAN: A novel pseudo-siamese generative adversarial network for multimodal remote sensing image change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 16609-16632. https://doi.org/10.1109/JSTARS.2025.3582803

[20] Zhang, C.X., Feng, Y.K., Hu, L., et al. (2022). A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images. International Journal of Applied Earth Observation and Geoinformation, 109: 102769. https://doi.org/10.1016/j.jag.2022.102769

[21] Zhang, H.C., Liu, J., Xiao, L. (2020). Bipartite residual network for change detection in heterogeneous optical and radar images. In IGARSS 2020-2020 IEEE

International Geoscience and Remote Sensing Symposium. Waikoloa, HI, USA, pp. 332-335. https://doi.org/10.1109/IGARSS39084.2020.9323786
[22] Jiang, X., Li, G., Liu, Y., Zhang, X.P., He, Y. (2020). Change detection in heterogeneous optical and SAR remote sensing images via deep homogeneous feature fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 1551-1566. https://doi.org/10.1109/JSTARS.2020.2983993
[23] Jiang, K.X., Zhang, W.H., Liu, J., Liu, F., Xiao, L. (2022). Joint variation learning of fusion and difference features for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-18. https://doi.org/10.1109/TGRS.2022.3226778
[24] Wang, J.J., Dobigeon, N., Chabert, M., Wang, D.C., Huang, T.Z., Huang, J. (2024). CD-GAN: A robust fusion-based generative adversarial network for unsupervised remote sensing change detection with heterogeneous sensors. Information Fusion, 107: 102313. https://doi.org/10.1016/j.inffus.2024.102313
[25] Li, H., Gong, M.G., Zhang, M.Y., Wu, Y. (2021). Spatially self-paced convolutional networks for change detection in heterogeneous images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 4966-4979. https://doi.org/10.1109/JSTARS.2021.3078437
[26] Tong, X.H., Pan, H.Y., Liu, S.C., Li, B.B., Luo, X., Xie, H. (2020). A novel approach for hyperspectral change detection based on uncertain area analysis and improved transfer learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 2056-2069. https://doi.org/10.1109/JSTARS.2020.2990481
[27] Sun, Y.L., Lei, L., Li, X., Tan, X., Kuang, G.Y. (2021). Patch similarity graph matrix-based unsupervised remote sensing change detection with homogeneous and heterogeneous sensors. IEEE Transactions on Geoscience and Remote Sensing, 59(6): 4841-4861. https://doi.org/10.1109/TGRS.2020.3013673
[28] Sun, Y.L., Lei, L., Li, X., Tan, X., Kuang, G.Y. (2022). Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-21. https://doi.org/10.1109/TGRS.2021.3053571
[29] Sun, Y.L., Lei, L., Guan, D.D., Kuang, G.Y. (2021). Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images. IEEE Transactions on Image Processing, 30: 6277-6291. https://doi.org/10.1109/TIP.2021.3093766
[30] Ferraris, V., Dobigeon, N., Chabert, M. (2020). Robust fusion algorithms for unsupervised change detection between multi-band optical images-A comprehensive case study. Information Fusion, 64: 293-317. https://doi.org/10.1016/j.inffus.2020.08.008

NOMENCLATURE

CD	Change Detection
RS	Remote Sensing
CNN	Convolutional Neural Nework
ViT	Vision Transformer
MLP	Multilayer Perceptron
BCE	Binary Cross-Entropy
IoU	Intersection over Union
F1	F1-score
SAR	Synthetic Aperture Radar
GAN	Generative Adversarial Network
LEVIR-CD	LEVIR Building Change Detection Dataset
LULC	Land Use and Land Cover
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative