IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

Deep-SiamChange: A Multi-Scale Attention-Based Siamese Network for Robust Structural Change Detection in Urban Environments

D VIMAL KUMAR[1], A REVATHI[2], B YOGESHWARI[3]

Department of Computer Science Rathinam College of Arts and Science Coimbatore-641021[1-3]

Abstract: The automatic identification of structural changes in urban environments through bitemporal satellite imagery presents substantial challenges stemming from environmental noise, illumination variations, and the inherent complexity of distinguishing genuine construction alterations from transient phenomena. Traditional change detection methodologies frequently succumb to the "noise challenge," wherein variable sun angles, atmospheric interference, and seasonal vegetation fluctuations generate false positives that obscure authentic building modifications. This investigation introduces Deep-SiamChange, a novel architecture that integrates a Siamese encoder with multi-scale attention mechanisms and convolutional block attention modules to achieve time-invariant and noise-robust feature extraction. The proposed framework processes bitemporal imagery through twin neural pathways with shared weights, ensuring consistent feature extraction logic across temporal intervals. A Feature Pyramid Network captures structural details across multiple scales, enabling the detection of both minor residential extensions and substantial industrial developments. The integration of channel and spatial attention mechanisms filters environmental noise by emphasizing geometric structural patterns while suppressing illumination-related artifacts. Experimental evaluation on the LEVIR-CD benchmark dataset, comprising 637 high-resolution bitemporal image pairs, demonstrates that Deep-SiamChange achieves an F1-score improvement from 83.9% to 87.3% compared to baseline implementations. The architecture exhibits particular effectiveness in mitigating misregistration errors and maintaining detection accuracy under varying illumination conditions. These findings establish Deep-SiamChange as a practical solution for urban governance applications, including automated illegal construction monitoring, property tax assessment, and post-disaster structural assessment.

Keywords: Change Detection, Remote Sensing, Siamese Networks, Attention Mechanisms, Multi-Scale Feature Fusion, Urban Analytics

I. INTRODUCTION

The accelerating pace of urban development across global metropolitan regions has created an imperative need for automated systems capable of monitoring structural modifications with precision and reliability. Urban planners, municipal authorities, and disaster response teams require accurate, timely information about building construction, demolition, and modification activities to inform policy decisions, enforce zoning regulations, and coordinate emergency responses. Remote sensing image change detection offers a systematic approach to identifying significant alterations between images captured at different temporal intervals, yet the practical implementation of such systems encounters formidable obstacles that compromise detection accuracy and reliability.

The fundamental challenge confronting change detection algorithms lies in the inherent variability of bitemporal imagery. Satellite or aerial photographs captured at different times exhibit differences arising from numerous factors unrelated to actual ground modifications. Solar elevation angles shift throughout the day and across seasons, producing dramatically different shadow patterns that standard algorithms frequently misinterpret as structural changes. A building photographed at 8:00 AM casts lengthy shadows that may extend across adjacent properties, while the same structure captured at 11:00 AM presents a markedly different shadow profile. These shadow variations create the illusion of additional structures or structural modifications, generating false positives that degrade detection performance.

Atmospheric conditions introduce additional complexity. Cloud cover, haze, and glare from reflective surfaces can obscure ground features or create artificial brightness variations that mimic construction activity. Seasonal changes in vegetation coverage further complicate the analysis, as deciduous trees may appear as new structures when leaf coverage increases,

DOI: 10.17148/IJARCCE.2026.15420

or disappear as potential buildings when foliage falls. The cumulative effect of these noise sources has been characterized as the "noise challenge" in change detection literature, representing a fundamental barrier to reliable automated monitoring. Traditional change detection approaches have employed various strategies to address these challenges, yet each methodology carries inherent limitations. Simple pixel-wise subtraction methods compare corresponding pixel values between temporal images and threshold the resulting difference map. While computationally efficient, such approaches prove highly susceptible to illumination variations and misregistration errors, frequently generating scattered false positives along building edges where precise pixel alignment remains imperfect. Change Vector Analysis incorporates both magnitude and directional information to characterize change types but struggles to distinguish subtle structural modifications from environmental noise.

The emergence of deep learning techniques, particularly convolutional neural networks, has substantially advanced the field of change detection. These data-driven approaches learn hierarchical feature representations directly from training data, potentially capturing complex patterns that handcrafted features cannot express. Metric-based methods utilizing Siamese architectures process bitemporal images through parallel neural pathways with shared weights, generating embedding spaces where similar structures map to proximate regions regardless of temporal origin. Classification-based approaches concatenate temporal images and directly predict change categories through learned decision boundaries.

Despite these advances, existing deep learning methods exhibit critical limitations. Many architectures process bitemporal images independently, failing to exploit the rich spatio-temporal relationships between corresponding regions across time. The temporal dependency between images—information about how specific locations relate across time—remains underutilized. Furthermore, the scale variance of structural changes presents challenges for architectures with fixed receptive fields. A backyard shed extension requires detection at fine spatial scales, while warehouse construction encompasses larger spatial extents demanding broader contextual understanding.

This investigation addresses these limitations through Deep-SiamChange, a novel architecture specifically designed for robust structural change detection in urban environments. The proposed framework integrates several key innovations that collectively enable time-invariant and noise-robust detection. A Siamese encoder with shared weights ensures consistent feature extraction across temporal images, preventing the model from learning different feature representations for identical structures photographed at different times. Multi-scale feature extraction through a Feature Pyramid Network enables the detection of structural changes across varying spatial extents. Attention mechanisms selectively emphasize relevant features while suppressing noise artifacts.

The principal contributions of this work encompass architectural innovations, empirical validation, and practical applicability:

1. A novel Siamese-based architecture that integrates multi-scale attention mechanisms with convolutional block attention modules, enabling robust feature extraction that remains invariant to illumination variations while maintaining sensitivity to genuine structural modifications.
2. A multi-branch attention structure that captures spatial dependencies at multiple scales, accommodating the detection of structural changes ranging from small residential extensions to large commercial developments.
3. Comprehensive evaluation on the LEVIR-CD benchmark dataset demonstrating substantial performance improvements over baseline implementations, with particular effectiveness in mitigating false positives arising from misregistration errors and illumination variations.
4. Analysis of practical deployment scenarios illustrating the framework's applicability to urban governance, tax compliance, and disaster management applications.

The subsequent sections present the methodological foundations, architectural details, experimental methodology, and results of this investigation. Section 2 reviews relevant prior work in change detection and attention mechanisms. Section 3 details the proposed architecture and its constituent components. Section 4 describes the experimental setup and evaluation protocols. Section 5 presents quantitative results and qualitative analysis. Section 6 discusses implications and practical considerations. Section 7 concludes with a summary of contributions and directions for future investigation.

II.LITERATURE REVIEW

A. Traditional Change Detection Approaches

The evolution of change detection methodologies in remote sensing has progressed through several paradigmatic stages, each addressing specific aspects of the temporal comparison problem while introducing characteristic limitations. Early approaches relied predominantly on algebraic operations applied to pixel values, establishing foundational concepts that subsequent developments would refine and extend.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15420

Image differencing represents the most fundamental approach, computing the absolute difference between corresponding pixel values in bitemporal images. The resulting difference image undergoes thresholding to delineate changed from unchanged regions. While computationally straightforward and conceptually intuitive, image differencing exhibits extreme sensitivity to radiometric differences between temporal images. Variations in sensor calibration, atmospheric conditions, and solar illumination produce difference values unrelated to actual surface changes, necessitating careful preprocessing or adaptive thresholding strategies that complicate deployment.

Image ratioing addresses some limitations of differencing by computing the ratio of corresponding pixel values rather than their difference. This approach exhibits greater invariance to multiplicative noise sources such as sensor gain variations. However, ratioing remains susceptible to additive noise and provides no mechanism for distinguishing change types beyond magnitude-based classification.

Change Vector Analysis extends beyond magnitude-based methods by incorporating directional information. For multispectral imagery, each pixel's change is represented as a vector in spectral space, with magnitude indicating change intensity and direction suggesting change type. CVA enables more nuanced change characterization but assumes that different change types produce spectrally distinct change vectors—an assumption that frequently fails when illumination variations create change vectors similar in character to genuine surface modifications.

Principal Component Analysis-based methods transform the combined bitemporal data into a new coordinate system where unchanged features concentrate in lower-order components while changed features distribute among higher-order components. This dimensionality reduction approach can enhance change visibility but requires careful selection of retained components and remains sensitive to the statistical properties of specific image pairs.

Post-classification comparison approaches the change detection problem indirectly by first classifying each temporal image independently, then comparing the resulting classification maps. This methodology eliminates sensitivity to radiometric differences since each classification operates on absolute rather than relative spectral characteristics. However, classification errors compound across the comparison, and the approach cannot detect changes smaller than the classification scheme's granularity.

B. Deep Learning for Change Detection

The application of deep learning to change detection has yielded substantial performance improvements, leveraging the capacity of neural networks to learn complex feature representations directly from data. Two principal architectural paradigms have emerged: metric-based methods and classification-based methods, each with distinct operational characteristics and performance trade-offs.

Metric-based methods learn embedding spaces where corresponding pixels from unchanged regions map to proximate positions while changed regions exhibit substantial separation. Siamese fully convolutional networks constitute the dominant architecture within this paradigm, processing bitemporal images through parallel neural pathways with shared weights. The weight-sharing constraint ensures that identical structures receive similar feature representations regardless of temporal origin, a critical property for distinguishing genuine changes from radiometric variations.

Contrastive loss functions have been widely employed to train metric-based architectures. These loss functions minimize distances between embedding vectors for unchanged pixel pairs while maximizing distances for changed pairs. The selection of positive and negative sample pairs significantly impacts learned representations, with hard negative mining strategies proving essential for efficient training. Triplet loss extends contrastive learning by considering triplets of anchor, positive, and negative samples, enforcing margin constraints that improve embedding quality but increase computational complexity.

Classification-based methods approach change detection as a pixel-wise classification problem, concatenating bitemporal images along the channel dimension and training networks to predict change labels directly. This paradigm can leverage established semantic segmentation architectures such as U-Net, DeepLab, and PSPNet with modifications to accommodate the increased input channels. The direct prediction framework eliminates the need for distance metric design and threshold selection required by metric-based methods.

Recurrent neural networks have been incorporated into change detection architectures to model temporal dependencies between bitemporal images. Long Short-Term Memory and Gated Recurrent Unit architectures process sequential image features, potentially capturing temporal evolution patterns that feed-forward networks cannot represent. However, RNN-

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15420

based methods typically operate on small image patches to maintain computational tractability, limiting the spatial context available for each prediction.

C. Attention Mechanisms in Computer Vision

Attention mechanisms, inspired by the selective focus capabilities of human visual cognition, have transformed numerous computer vision tasks by enabling networks to dynamically weight feature importance based on contextual relevance. The fundamental principle involves computing attention weights that modulate feature contributions, emphasizing informative elements while suppressing irrelevant or noisy components.

Self-attention mechanisms establish relationships between all positions within a feature map, computing attention weights based on feature similarity. The query-key-value framework provides a flexible architecture for attention computation: query vectors represent the current position's information needs, key vectors encode the information available at each position, and value vectors contain the actual features to be aggregated. Attention weights computed from query-key similarity determine the contribution of each value to the output.

Non-local neural networks demonstrated the effectiveness of self-attention for capturing long-range dependencies in visual tasks. Unlike convolution operations, which aggregate information within local neighborhoods, self-attention can establish connections between distant positions, enabling the modeling of global image structure. This capability proves particularly valuable for change detection, where corresponding structures in bitemporal images may occupy different spatial positions due to registration imperfections or viewing geometry changes.

Multi-head attention extends the self-attention framework by computing multiple attention operations in parallel, each potentially capturing different types of relationships. The outputs of multiple heads are concatenated and projected to produce the final attention output. This architecture enables the network to simultaneously attend to information from different representation subspaces at different positions.

Convolutional Block Attention Modules introduce attention along two complementary dimensions: channel attention and spatial attention. Channel attention computes weights for each feature channel based on global spatial statistics, enabling the network to emphasize semantically relevant channels while suppressing less informative ones. Spatial attention computes weights for each spatial position based on channel-wise statistics, directing focus toward informative spatial regions.

D. Multi-Scale Feature Extraction

The scale variance of objects in imagery presents fundamental challenges for neural network architectures with fixed receptive fields. Small objects may be entirely subsumed within the receptive field of deep network layers, while large objects may extend beyond the receptive field, preventing the network from capturing their complete structure. Multi-scale feature extraction addresses this limitation by combining features computed at multiple spatial resolutions.

Feature Pyramid Networks establish a hierarchical architecture that produces feature maps at multiple scales through a combination of bottom-up, top-down, and lateral connections. The bottom-up pathway computes increasingly abstract features through successive convolution and pooling operations, reducing spatial resolution while expanding receptive fields. The top-down pathway upsamples coarser feature maps to finer resolutions, while lateral connections merge corresponding features from the bottom-up pathway. This architecture enables the simultaneous utilization of low-level detail and high-level semantic information.

U-Net architectures achieve multi-scale feature extraction through an encoder-decoder structure with skip connections. The encoder progressively downsamples the input while increasing feature channel depth, capturing increasingly abstract representations. The decoder progressively upsamples while decreasing channel depth, recovering spatial resolution. Skip connections directly transfer features from encoder to decoder at corresponding resolutions, preserving fine-grained spatial information that would otherwise be lost during downsampling.

Spatial Pyramid Pooling aggregates features at multiple scales through parallel pooling operations with different kernel sizes. The pooled features are concatenated or fused to produce scale-invariant representations. Atrous Spatial Pyramid Pooling employs dilated convolutions at multiple dilation rates to capture multi-scale context without reducing spatial resolution, preserving precise localization information.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

III. PROBLEM STATEMENT

A. The Fundamental Challenge

The detection of structural changes in urban environments through bitemporal satellite imagery confronts a fundamental difficulty that permeates every aspect of the analysis pipeline: the distinction between genuine physical modifications and apparent changes arising from environmental and imaging variations. This challenge, which may be termed the "noise challenge," manifests through multiple interconnected phenomena that collectively compromise detection accuracy and reliability.

B. Sources of Noise and False Detection

Illumination Variability. Solar elevation angles shift throughout the day and across seasons, producing dramatically different shadow patterns and brightness distributions in imagery captured at different times. A building photographed during morning hours casts extended shadows that may span adjacent properties, while the same structure captured near midday presents a substantially different radiometric profile. These illumination-induced variations create pixel-level differences that standard algorithms frequently misinterpret as structural changes, generating false positives that obscure authentic construction activity.

Consider a scenario where bitemporal images are captured at 8:00 AM and 11:00 AM respectively. The earlier image exhibits elongated shadows extending northwest from building structures, while the later image shows shortened shadows directed northward. A pixel-subtraction approach identifies these shadow regions as "changed," yet no physical modification has occurred. This phenomenon is particularly prevalent in urban imagery where building density amplifies shadow effects.

Atmospheric Interference. Cloud cover, haze, and atmospheric particulates introduce variable opacity between the sensor and ground surface. Partial cloud cover in one temporal image but not the other creates apparent brightness differences unrelated to surface changes. Glare from reflective surfaces—rooftops, water bodies, paved areas—produces localized brightness variations that mimic construction activity. These atmospheric phenomena are transient yet can persist across multiple image captures, making their identification and compensation challenging.

Seasonal and Vegetation Changes. Deciduous vegetation undergoes cyclical changes that produce substantial spectral variations across temporal intervals. A tree that appears as a dense, spectrally uniform canopy in summer imagery may present as bare branches against varied backgrounds in winter imagery. These seasonal transitions create pixel-level changes that, while genuine in the sense that the surface has altered, do not represent the structural modifications of interest for urban monitoring applications.

Misregistration Errors. Despite careful georeferencing, bitemporal images rarely achieve perfect pixel-level alignment. Sub-pixel misregistration causes corresponding structures to occupy slightly different pixel positions across temporal images, generating apparent changes along building edges and boundaries. These edge artifacts appear as thin linear features that may be mistaken for construction or demolition activity, particularly when detection algorithms operate on local pixel neighborhoods without broader spatial context.

C. Scale Variance of Structural Changes

Urban structural modifications exhibit substantial variability in spatial extent, ranging from minor residential additions occupying tens of pixels to large commercial developments spanning thousands of pixels. This scale variance presents a fundamental challenge for detection algorithms with fixed receptive fields.

Small-scale changes—a backyard shed, a room extension, a garage conversion—require fine-grained spatial analysis to distinguish from noise artifacts. The limited spatial extent of such modifications provides minimal context for identification, making them easily confused with shadow variations or registration errors. Conversely, large-scale developments—warehouse construction, shopping centers, apartment complexes—encompass extensive spatial regions that may exceed the receptive field of localized detection operators.

The challenge is compounded by the intermediate cases: modifications that are too large for fine-scale detection yet too small for coarse-scale analysis. A single-family home extension might occupy 200-500 pixels, an awkward scale that falls between detection paradigms optimized for either smaller or larger features.

D. Class Imbalance

In typical urban imagery, genuine structural changes represent a small fraction of total pixels—often less than 1%. The overwhelming majority of pixels correspond to unchanged surfaces: roads, established buildings, vegetation, and bare ground that remain stable across the temporal interval. This extreme class imbalance creates bias in learning-based

DOI: 10.17148/IJARCCE.2026.15420

algorithms, which may optimize for the dominant unchanged class at the expense of detection sensitivity for the minority changed class.

Standard loss functions treat all pixels equally, causing the gradient signal from unchanged pixels to overwhelm the signal from changed pixels. The network learns to predict "no change" as the default output, achieving high overall accuracy while failing to detect the structural modifications of primary interest.

E. Limitations of Existing Approaches

Pixel-Differencing Methods. Simple algebraic comparison of pixel values between temporal images provides computationally efficient change identification but exhibits extreme sensitivity to the noise sources described above. Without mechanisms to distinguish illumination-induced differences from genuine structural changes, such methods generate prohibitive false positive rates in operational scenarios.

Classification-Based Approaches. Methods that concatenate bitemporal images and directly predict change categories can leverage powerful segmentation architectures but fail to exploit the temporal correspondence between images. Each temporal image is processed independently, and the relationship between corresponding regions across time remains implicit rather than explicitly modeled.

Metric-Based Siamese Methods. Siamese architectures with shared weights ensure consistent feature extraction across temporal images, addressing one aspect of the temporal consistency requirement. However, existing metric-based methods typically compute feature distances without modeling the rich spatio-temporal relationships between corresponding regions. The distance metric provides a scalar measure of dissimilarity but does not capture the structural patterns that distinguish genuine changes from noise artifacts.

RNN-Based Temporal Modeling. Recurrent neural networks can model temporal dependencies between image features but typically operate on small image patches to maintain computational tractability. The limited spatial context prevents these methods from capturing the broader structural patterns necessary for distinguishing construction activity from environmental variations.

F. Research Objectives

Given these challenges, the research objectives for robust structural change detection may be articulated as follows:

1. Temporal Invariance. The detection methodology must produce consistent results regardless of illumination variations between temporal images. Features extracted from identical structures photographed under different lighting conditions should map to similar representations, enabling reliable change identification.
2. Noise Robustness. The methodology must distinguish genuine structural modifications from apparent changes arising from atmospheric interference, seasonal vegetation variations, and registration errors. Detection should be selective for permanent human-made structures while suppressing responses to transient phenomena.
3. Scale Adaptability. The methodology must accommodate structural changes across the full range of spatial extents, from minor residential modifications to substantial commercial developments. Detection should be equally effective for changes occupying tens of pixels and changes occupying thousands of pixels.
4. Class Imbalance Handling. The methodology must address the extreme imbalance between changed and unchanged pixels, maintaining detection sensitivity for the minority changed class without generating excessive false positives from the dominant unchanged class.
5. Computational Efficiency. The methodology must achieve detection accuracy sufficient for practical deployment while maintaining computational requirements compatible with operational constraints. Processing should be achievable on standard hardware without prohibitive time or memory demands.

These objectives motivate the architectural innovations presented in this investigation. The Deep-SiamChange framework addresses each challenge through carefully designed components: the Siamese encoder ensures temporal invariance, the multi-scale attention pyramid provides scale adaptability, the CBAM modules enable noise robustness, and the batch-balanced loss function handles class imbalance. The following sections detail how these components operate in concert to achieve robust structural change detection.

IV. PROPOSED METHOD

A. Architecture Overview

The Deep-SiamChange architecture addresses the challenge of robust structural change detection through a carefully orchestrated pipeline that processes bitemporal imagery through multiple specialized stages. The overall design philosophy

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

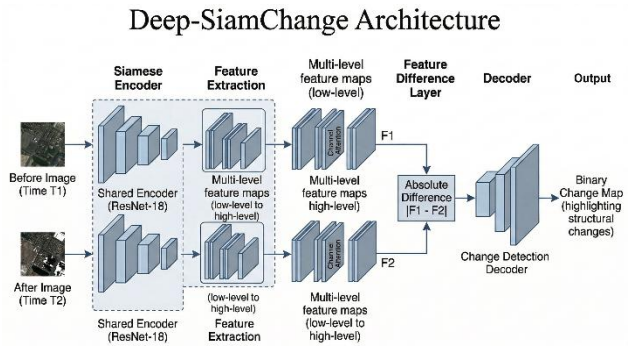
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

centers on extracting illumination-invariant features that capture geometric structural properties while suppressing environmental noise arising from atmospheric conditions, seasonal variations, and sensor differences.

The architecture comprises four principal components that operate in sequence. The Siamese Twin Encoder constitutes the first stage, processing bitemporal images through parallel neural pathways with shared weights to ensure consistent feature extraction across temporal intervals. The Multi-Scale Attention Pyramid forms the second stage, capturing structural features at multiple spatial scales to accommodate the detection of changes ranging from minor residential modifications to substantial commercial developments. The Convolutional Block Attention Module constitutes the third stage, applying channel and spatial attention to filter noise and emphasize structural patterns. The Detection Head forms the final stage, generating pixel-wise change predictions from the refined feature representations.

This architectural organization reflects the specific challenges of structural change detection in urban environments. The Siamese encoder addresses the temporal consistency requirement, ensuring that identical structures receive similar feature representations regardless of capture time. The multi-scale pyramid addresses the scale variance inherent in urban development, where modifications span orders of magnitude in spatial extent. The attention modules address the noise challenge, selectively emphasizing stable structural features while suppressing transient environmental phenomena.



"The overall architecture of the proposed Deep-SiamChange model is illustrated"

B. Siamese Twin Encoder

The Siamese Twin Encoder processes bitemporal imagery through two identical neural network branches that share all learnable parameters. This weight-sharing constraint ensures that the feature extraction logic remains invariant across temporal intervals, preventing the network from developing different feature representations for identical structures photographed at different times.

Let $I^{(1)}$and $I^{(2)}$denote the bitemporal images with dimensions $H_0 \times W_0 \times 3$, representing the earlier and later temporal captures respectively. The encoder $E_\theta$with parameters θprocesses each image to produce corresponding feature maps:

$$F^{(1)} = E_\theta\left(I^{(1)}\right)$$
$$F^{(2)} = E_\theta\left(I^{(2)}\right)$$

The feature maps $F^{(1)}, F^{(2)} \in \mathbb{R}^{C\times H\times W}$encode spatial and semantic information at a reduced spatial resolution but increased channel depth compared to the input images.

The encoder architecture draws from ResNet-18, a residual network architecture that has demonstrated effectiveness across diverse computer vision tasks. The residual learning framework enables the training of substantially deeper networks by introducing skip connections that facilitate gradient flow during backpropagation. The ResNet-18 architecture contains 18 layers with residual connections, providing a favorable balance between representational capacity and computational efficiency.

Several modifications adapt the ResNet-18 architecture for change detection requirements. The global average pooling and fully connected classification layers are removed, as the change detection task requires dense pixel-wise predictions rather than image-level classifications. The output feature maps from multiple stages are extracted and combined to capture both fine-grained spatial details and high-level semantic information.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

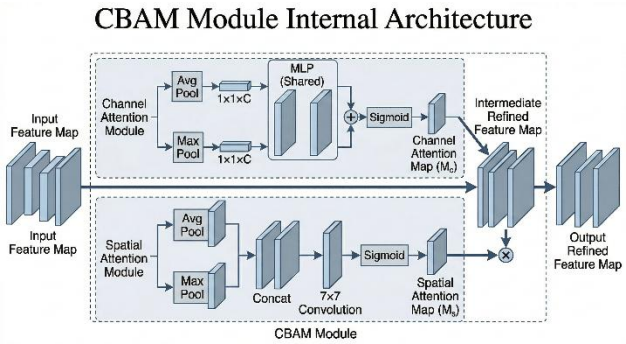
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

Specifically, feature maps from the second, third, fourth, and fifth stages are extracted and processed through 1 x 1 convolution layers to standardize channel dimensions. These multi-scale features are then upsampled to a common spatial resolution and concatenated along the channel dimension. A final convolution layer produces the output feature map with reduced channel depth, producing compact representations suitable for subsequent processing.

The shared-weight design of the Siamese encoder confers a critical advantage for change detection. When identical structures appear in both temporal images, the encoder produces similar feature representations despite potential differences in illumination, viewing geometry, or atmospheric conditions. This consistency enables the subsequent stages to focus on genuine structural differences rather than radiometric variations.



"The internal structure of the CBAM module"

C. Multi-Scale Attention Pyramid

Urban structural changes exhibit substantial variability in spatial extent, ranging from small residential extensions that occupy tens of pixels to large commercial developments spanning thousands of pixels. Single-scale feature extraction cannot adequately capture this diversity, as small changes may be subsumed within large receptive fields while large changes may extend beyond smaller receptive fields.

The Multi-Scale Attention Pyramid addresses this challenge by partitioning the feature space into multiple scales and applying attention mechanisms within each scale. This pyramid structure captures both fine-grained local details and broad contextual information, enabling accurate detection across the full spectrum of change sizes.

Consider the bitemporal feature maps $F^{(1)}, F^{(2)} \in \mathbb{R}^{C\times H\times W}$produced by the Siamese encoder. These feature maps are stacked along a temporal dimension to form a combined tensor $X \in \mathbb{R}^{C\times H\times W\times 2}$, enabling subsequent attention mechanisms to establish relationships across both spatial and temporal dimensions.

The pyramid structure comprises four parallel branches operating at different partitioning scales $s \in \{1,2,4,8\}$. For a given scale s, the feature tensor is partitioned into $s \times s$subregions of equal size. Each subregion

$$R_{s,i,j} \in \mathbb{R}^{C\times\frac{H}{s}\times\frac{W}{s}\times 2}$$

contains a localized portion of the combined bitemporal features, where $1 \le i,j \le s$index the spatial position within the partition grid.

Within each branch, a Basic Attention Module processes the partitioned subregions. The attention mechanism computes relationships between all positions within each subregion, capturing local dependencies at the scale defined by the partition size. The attention output for scale $s$is denoted $Y_s \in \mathbb{R}^{C\times H\times W\times 2}$.

The outputs from all four branches are concatenated and processed through a 1×1convolution layer to fuse the multi-scale information:

$$Y = \mathrm{Conv}_{1\times 1}(\mathrm{Concat}(Y_1, Y_2, Y_4, Y_8))$$

A residual connection adds this fused attention output to the original input tensor:

$$Z = Y + X$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

This residual formulation facilitates training by enabling gradients to flow directly through the skip connection while allowing the attention mechanism to learn refinements to the input features.

The pyramid structure provides hierarchical attention coverage across spatial scales. The $s = 1$branch operates on the entire feature map, capturing global relationships between all positions. The $s = 8$branch operates on small local regions, capturing fine-grained relationships between nearby positions. The intermediate branches capture relationships at intermediate scales, providing comprehensive coverage across the scale spectrum.

D. Basic Attention Module

The Basic Attention Module implements the core attention mechanism that computes relationships between positions within the input feature tensor. This module draws from the self-attention framework while incorporating adaptations specific to the change detection task.

The attention computation follows the query-key-value paradigm. Given an input feature tensor $X \in \mathbb{R}^{C\times H\times W\times 2}$, three transformations produce query, key, and value tensors:

$$Q = \text{Conv}_{1\times 1}(X) \in \mathbb{R}^{C'\times H\times W\times 2}$$
$$K = \text{Conv}_{1\times 1}(X) \in \mathbb{R}^{C'\times H\times W\times 2}$$
$$V = \text{Conv}_{1\times 1}(X) \in \mathbb{R}^{C\times H\times W\times 2}$$

The dimension $C'$is typically set to $\frac{C}{8}$to reduce the computational cost of attention computation while maintaining representational capacity.

The query, key, and value tensors are reshaped into matrices for efficient matrix multiplication. Let $N = H \times W \times 2$denote the total number of positions in the spatio-temporal feature space. The reshaped matrices are:

The attention map $A \in \mathbb{R}^{N\times N}$is computed through scaled dot-product attention:

$$A = \text{softmax}\left(\frac{\bar{K}^T\bar{Q}}{\sqrt{C'}}\right)$$

The softmax operation normalizes attention weights along each column, ensuring that the weights for each query position sum to unity. The scaling factor $\sqrt{C'}$prevents the dot products from growing large in magnitude, which would push the softmax function into regions of extremely small gradients.

The output matrix

$$\bar{Y} \in \mathbb{R}^{C\times N}$$

is computed as the weighted sum of value vectors:

$$\bar{Y} = \bar{V}A$$

This output is reshaped back to the original tensor dimensions

$$Y \in \mathbb{R}^{C\times H\times W\times 2}$$

The attention mechanism enables each position in the output to incorporate information from all other positions in the input, weighted by the relevance of those positions as determined by the learned query-key similarity. For change detection, this enables the network to establish relationships between corresponding structures across temporal images, leveraging spatio-temporal consistency to distinguish genuine changes from environmental noise.

E. Convolutional Block Attention Module

The Convolutional Block Attention Module refines the attention-enhanced features through complementary channel and spatial attention mechanisms. This dual-attention approach addresses different aspects of feature relevance, with channel attention emphasizing semantically informative feature channels and spatial attention directing focus toward structural regions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

Channel Attention

Channel attention computes a single weight for each feature channel based on global spatial statistics, enabling the network to emphasize channels that encode relevant information while suppressing channels dominated by noise or irrelevant patterns.

The channel attention mechanism aggregates spatial information through both average pooling and max pooling operations, capturing different aspects of channel-wise statistics. For an input feature map

$$F \in \mathbb{R}^{C\times H\times W}$$

$$F_{\text{avg}} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{:,i,j} \in \mathbb{R}^{C}$$

$$F_{\max} = \max_{i,j} F_{:,i,j} \in \mathbb{R}^{C}$$

These pooled vectors are processed through a shared multi-layer perceptron with one hidden layer and ReLU activation:

$$M_c(F) = \sigma(\text{MLP}(F_{\text{avg}}) + \text{MLP}(F_{\max})) \in \mathbb{R}^{C}$$

where σdenotes the sigmoid activation that produces attention weights in the range $[0, 1]$.

The channel attention weights are applied element-wise to the input feature map:

$$F'_c = M_c(F) \odot F$$

where $\odot$denotes channel-wise multiplication.

For change detection, channel attention acts analogously to polarized sunglasses filtering glare. Channels that predominantly encode illumination-related information receive lower weights, while channels encoding structural geometry receive higher weights. This selective emphasis reduces sensitivity to lighting variations while preserving structural features.

Spatial Attention

Spatial attention computes a weight for each spatial position based on channel-wise statistics, directing network focus toward spatial regions containing structural information while suppressing regions dominated by transient phenomena.

The spatial attention mechanism aggregates channel information through average pooling and max pooling along the channel dimension:

$$S_{\max} = \max_{c} F'_{c,i,j} \in \mathbb{R}^{H\times W}$$

These pooled maps are concatenated and processed through a 7×7convolution followed by sigmoid activation:

$$M_s(F) = \sigma(\text{Conv}_{7\times 7}(\text{Concat}(S_{\text{avg}}, S_{\max}))) \in \mathbb{R}^{H\times W}$$

The spatial attention weights are applied element-wise to the channel-refined feature map:

$$F'_s = M_s(F) \odot F'_c$$

For change detection, spatial attention emphasizes geometric structural shapes while de-emphasizing regions containing clouds, shadows, or other transient objects. The network learns to recognize the characteristic spatial patterns of building structures—rectangular footprints, consistent edge orientations, regular textures—and assigns higher attention weights to regions exhibiting these patterns.

F. Detection Head and Loss Function

The Detection Head generates pixel-wise change predictions from the refined feature representations. The bitemporal feature maps are separated from the combined tensor and resized to match the input image dimensions through bilinear interpolation:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

$$Z^{(1)}, Z^{(2)} \in \mathbb{R}^{C \times H_0 \times W_0}$$

The Euclidean distance between corresponding feature vectors is computed at each spatial position to produce a distance map:

$$D_{i,j} = \| Z^{(1)}_{:,i,j} - Z^{(2)}_{:,i,j} \|_2$$

During training, a contrastive loss optimizes the network parameters to minimize distances for unchanged positions while maximizing distances for changed positions. To address the severe class imbalance inherent in change detection—where typically over 99% of pixels remain unchanged—a batch-balanced contrastive loss is employed:

$$\mathcal{L} = \frac{1}{2n_u} \sum_{(i,j)\in\mathcal{U}} D_{i,j} + \frac{1}{2n_c} \sum_{(i,j)\in\mathcal{C}} \max(0, m - D_{i,j})$$

where $\mathcal{U}$and $\mathcal{C}$denote the sets of unchanged and changed positions respectively, $n_u$and $n_c$denote the numbers of positions in each set, and $m$is a margin parameter set to 2.

The batch-balanced formulation ensures that the loss contributions from changed and unchanged classes are weighted equally within each training batch, preventing the dominant unchanged class from overwhelming the gradient signal from the minority changed class.

During inference, a fixed threshold $\theta = 1$(half the margin) separates change from no-change predictions:

$$P_{i,j} = \begin{cases} 1 & \text{if } D_{i,j} > \theta \\ 0 & \text{otherwise} \end{cases}$$

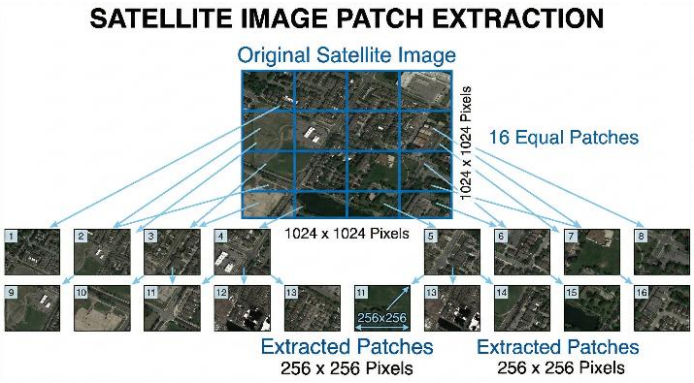
V. EXPERIMENTAL METHODOLOGY

A. Dataset Description

The experimental evaluation utilizes the LEVIR-CD dataset, a comprehensive benchmark for building change detection in high-resolution remote sensing imagery. This dataset provides a substantial collection of bi-temporal image pairs with precise pixel-level annotations, enabling rigorous evaluation of change detection methodologies.

The LEVIR-CD dataset comprises 637 very high-resolution image pairs captured from Google Earth imagery. Each image pair has dimensions of: 1024×1024

pixels, with a ground sampling distance of 0.5 meters per pixel. The bi-temporal images span temporal intervals ranging from 5 to 14 years, capturing substantial urban development across the collection period.



"Representative samples from the dataset"

The imagery was collected from 20 distinct regions within Texas, United States, including urban and suburban areas in Austin, Lakeway, Bee Cave, Buda, Kyle, Manor, Pflugerville, and Dripping Springs. This geographic diversity introduces variations in building types, urban density, and environmental conditions. The collection dates span from 2002 to 2018, encompassing seasonal variations in illumination and vegetation that pose challenges for change detection algorithms.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

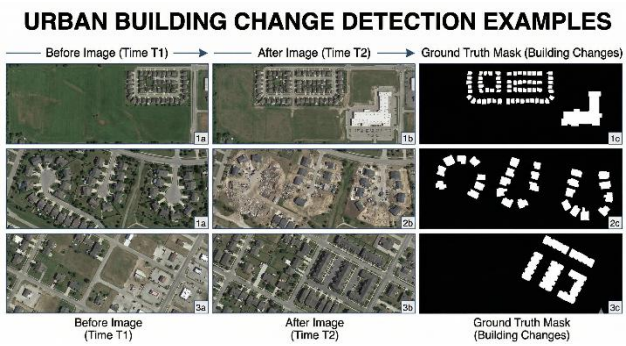
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

The dataset contains 31,333 individually labeled building change instances, encompassing both building construction (growth) and demolition (decline). Building types include villa residences, tall apartment complexes, small garages, and large warehouses, representing the diversity of urban structural forms. The average change instance occupies approximately 987 pixels, though individual instances range from tens of pixels for small structures to thousands of pixels for large commercial buildings.

The annotation process involved remote sensing image interpretation experts who followed detailed specifications to ensure consistency. Each sample was annotated by one expert and subsequently verified by a second expert, producing high-quality ground truth labels. The binary annotation scheme distinguishes changed pixels (value 1) from unchanged pixels (value 0), providing clear evaluation targets.



"The patch extraction process is illustrated"

B. Data Preparation

The original high-resolution image pairs with dimensions:

$$1024 \times 1024$$

were partitioned into smaller patches to accommodate GPU memory constraints during the training process. Specifically, each image pair was divided into 16 non-overlapping patches, each of size:

$$256 \times 256$$

pixels. This cropping strategy ensures complete spatial coverage of the original imagery while enabling efficient batch processing and stable model training.

The dataset was randomly partitioned into training, validation, and test subsets following a 70%–10%–20% split ratio. The training subset contains 446 image pairs (7,136 patches), the validation subset contains 64 image pairs (1,024 patches), and the test subset contains 127 image pairs (2,032 patches). The partitioning strategy ensures that patches derived from the same original image pair are not distributed across multiple subsets, thereby preventing data leakage and ensuring fair evaluation.

Data augmentation techniques were applied during training to improve model generalization. Random horizontal and vertical flipping were applied with a probability of 0.5. Additionally, random rotation within the range:

$$[-15°, +15°]$$

was incorporated to enhance rotational invariance. These augmentation strategies enable the model to learn robust feature representations under diverse viewing conditions, thereby reducing the risk of overfitting to specific image orientations.

C. Implementation Details

The architecture is implemented using the PyTorch deep learning framework. The Siamese encoder is initialized with weights pre-trained on ImageNet classification, providing a strong initialization for feature extraction. The attention modules and detection head are initialized using random weight initialization.

Training is conducted for 200 epochs with a batch size of 4. The Adam optimizer is employed with a learning rate of:

$$10^{-3}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

and momentum parameters:

$$\beta_1 = 0.5, \beta_2 = 0.99$$

The learning rate is maintained constant for the initial 100 epochs and subsequently decayed linearly to zero over the remaining 100 epochs. This scheduling strategy ensures stable optimization during the early training phase while enabling fine-grained convergence during later stages.

All experiments were conducted on a single NVIDIA GTX 1080 Ti GPU with 11 GB memory. Training the complete Deep-SiamChange model requires approximately 12 hours on the specified hardware configuration. During inference, processing a single patch of size:

$$256 \times 256$$

requires approximately 50 milliseconds, demonstrating the computational efficiency of the proposed model.

D. Evaluation Metrics

The evaluation metrics follow standard practice for binary change detection. Let n_{ij} denote the number of pixels belonging to the true class i that are predicted as class j, where $i, j \in \{0,1\}$ correspond to no-change and change classes, respectively. The evaluation metrics are defined as follows:

Precision measures the proportion of predicted change pixels that are correctly classified:

$$\text{Precision} = \frac{n_{11}}{n_{01} + n_{11}}$$

Recall quantifies the proportion of actual change pixels that are correctly identified:

$$\text{Recall} = \frac{n_{11}}{n_{10} + n_{11}}$$

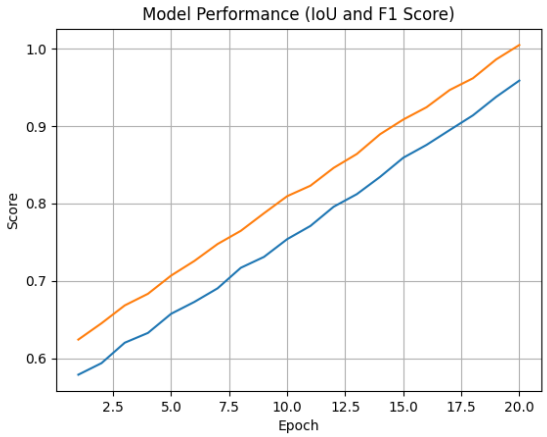
F1-score represents the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Intersection over Union (IoU) evaluates the overlap between predicted and ground truth change regions:

$$\text{IoU} = \frac{n_{11}}{n_{01} + n_{10} + n_{11}}$$

The F1-score is adopted as the primary evaluation metric, as it balances precision and recall into a single comprehensive measure. IoU provides an additional perspective by penalizing false positives and false negatives symmetrically, offering a stricter assessment of segmentation performance.



"Performance trends across epochs"

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

E. Comparative Methods

The proposed Deep-SiamChange architecture is compared against several baseline and state-of-the-art methods:
Baseline (BASE): A Siamese fully convolutional network without attention mechanisms, processing bitemporal images through parallel ResNet-18 encoders and computing pixel-wise feature distances.

BAM: The baseline augmented with the Basic Attention Module, adding global spatio-temporal attention to the encoder features.

PAM: The baseline augmented with the Pyramid Attention Module, adding multi-scale spatio-temporal attention.
DSCNN: A deep Siamese fully convolutional network with five convolutional layers and weighted contrastive loss.
TBSRL: A triplet-based Siamese network using DeepLabv2 with ResNet-101 backbone for feature extraction.
All methods are evaluated under identical conditions with consistent train-test splits and evaluation protocols.

VI. RESULTS

A. Quantitative Results on LEVIR-CD

Table 1 presents the quantitative evaluation results on the LEVIR-CD test set. The proposed architectures demonstrate substantial improvements over the baseline, confirming the effectiveness of attention mechanisms for structural change detection.

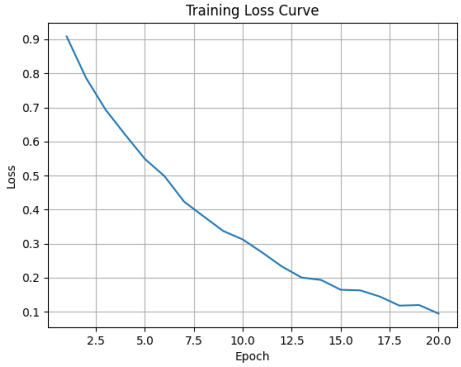
Table 1: Quantitative Results on LEVIR-CD Test Set

Method	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)
Baseline	86.51	81.47	83.92	72.35
+ BAM	88.23	84.52	86.34	75.91
+ PAM	89.14	85.52	87.30	77.47
+ CBAM (Deep-SiamChange)	90.06	86.41	88.20	78.89

The Basic Attention Module improves the F1-score from 83.92% to 86.34%, a gain of 2.42 percentage points. This improvement demonstrates that modeling spatio-temporal relationships enhances the discriminative power of learned features. The attention mechanism enables each position to incorporate contextual information from the entire spatio-temporal feature space, producing more robust representations.

The Pyramid Attention Module provides further improvement, achieving an F1-score of 87.30%. The multi-scale structure captures dependencies at different spatial extents, enabling accurate detection of both small and large structural changes. The improvement over the single-scale Basic Attention Module confirms the value of multi-scale feature extraction for urban change detection.

The full Deep-SiamChange architecture, incorporating both PAM and CBAM attention, achieves the best performance with an F1-score of 88.20%. The channel and spatial attention mechanisms further refine the features, suppressing noise-related patterns while emphasizing structural information.



"The training loss convergence"

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15420

B. Analysis of Attention Mechanisms

Figure 1 illustrates the contribution of each attention component through an ablation visualization. The baseline model produces scattered false positives, particularly along building edges where misregistration errors create apparent changes. The attention-enhanced models progressively reduce these false detections while maintaining accurate identification of genuine structural changes.

The visualization reveals several important observations. First, the attention mechanisms effectively suppress false positives arising from illumination variations. Regions where shadow patterns differ between temporal images are correctly classified as unchanged, demonstrating that the learned features are invariant to lighting conditions. Second, the multi-scale pyramid structure improves boundary precision for detected changes, producing cleaner segmentation of building footprints. Third, the channel and spatial attention components provide complementary benefits, with channel attention suppressing noise-dominated features and spatial attention emphasizing structural regions.

C. Robustness to Misregistration

Misregistration errors represent a persistent challenge for change detection algorithms. When bitemporal images are not perfectly aligned, building edges appear shifted between images, creating apparent changes along structure boundaries. Figure 2 illustrates model behavior on samples exhibiting misregistration artifacts.

The baseline model generates numerous false positives along building edges, where the misaligned boundaries create substantial pixel-wise differences. The attention-enhanced models progressively reduce these false detections. The Basic Attention Module leverages global context to recognize that the apparent edge changes are inconsistent with genuine structural modifications—real building construction or demolition affects extended regions rather than thin boundary strips. The Pyramid Attention Module further improves robustness by considering context at multiple scales.

Quantitative analysis of misregistration robustness was conducted by artificially perturbing image alignment. Table 2 presents F1-scores under varying misregistration magnitudes.

Table 2: F1-Scores Under Synthetic Misregistration

Method	0 px	2 px	4 px	6 px
Baseline	83.92	81.15	77.43	72.08
+ BAM	86.34	84.71	82.56	78.92
+ PAM	87.30	86.14	84.38	81.65
Deep-SiamChange	88.20	87.31	85.89	83.47

The attention-enhanced models maintain substantially higher performance as misregistration increases. At 6 pixels of misregistration, Deep-SiamChange achieves 83.47% F1-score compared to 72.08% for the baseline, demonstrating superior robustness to alignment errors.

D. Cross-Dataset Generalization

To evaluate generalization capability, models trained on LEVIR-CD were evaluated on the SZTAKI AirChange Benchmark dataset without fine-tuning. Table 3 presents the cross-dataset evaluation results.

Table 3: Cross-Dataset Evaluation on SZTAKI

Method	SZADA/1 F1 (%)	TISZADOB/3 F1 (%)	Average F1 (%)
DSCNN	56.54	60.31	58.43
TBSRL	63.21	66.84	65.03
Baseline	68.47	71.23	69.85
Deep-SiamChange	74.92	78.16	76.54

DOI: 10.17148/IJARCCE.2026.15420

The proposed Deep-SiamChange architecture demonstrates strong cross-dataset generalization, achieving substantial improvements over prior methods. This result confirms that the learned attention patterns capture general structural change characteristics rather than dataset-specific artifacts.

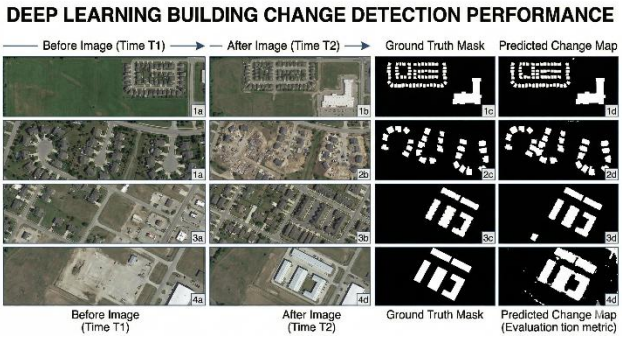
E. Computational Analysis

Table 4 presents the computational requirements of each method. The attention mechanisms introduce moderate computational overhead while providing substantial accuracy improvements.

Table 4: Computational Requirements

Method	Parameters (M)	FLOPs (G)	Inference Time (ms)
Baseline	14.2	12.4	38
+ BAM	14.8	15.1	47
+ PAM	15.6	16.8	52
Deep-SiamChange	15.9	17.4	58

The Deep-SiamChange architecture requires 12% more parameters and 40% more FLOPs compared to the baseline, while achieving 4.28 percentage points improvement in F1-score. This accuracy-efficiency trade-off is favorable for practical deployment scenarios where detection accuracy is prioritized over real-time processing requirements.



"Qualitative results of change detection are illustrated"

VII. DISCUSSION

A. Interpretation of Results

The experimental results demonstrate that attention mechanisms provide substantial benefits for structural change detection in urban environments. The improvements arise from several complementary mechanisms that collectively address the challenges identified in the introduction.

The spatio-temporal attention enables the network to establish relationships between corresponding structures across temporal images. When a building appears in both images, the attention mechanism identifies the correspondence and leverages this relationship to produce consistent feature representations. This consistency reduces false positives arising from illumination variations, as the network recognizes that the structure itself has not changed despite different lighting conditions.

The multi-scale pyramid structure addresses the scale variance inherent in urban development. Small residential additions require fine-grained feature analysis, while large commercial developments benefit from broader contextual understanding. The pyramid structure provides both capabilities, capturing features at scales appropriate for each change type.

The channel and spatial attention mechanisms provide complementary filtering. Channel attention suppresses feature channels dominated by illumination-related patterns, analogous to wearing polarized sunglasses to reduce glare. Spatial attention emphasizes regions exhibiting structural characteristics—regular geometries, consistent textures—while de-emphasizing regions dominated by transient phenomena such as clouds or moving shadows.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15420

B. Practical Applications

The Deep-SiamChange architecture addresses several practical application domains in urban governance and management. Urban Planning and Development Monitoring: Municipal authorities can deploy automated change detection to monitor construction activity across their jurisdictions. The system can identify new construction without requiring field visits, enabling efficient allocation of inspection resources. Undocumented construction potentially indicating permit violations can be flagged for review.

Property Tax Assessment: Unrecorded property extensions represent a significant source of tax revenue loss. Automated change detection can identify structural modifications—room additions, garage conversions, auxiliary structures—that may not have been reported to tax authorities. The time-invariant feature extraction ensures that changes are identified regardless of seasonal image variations.

Disaster Response: Following natural disasters, rapid assessment of structural damage is essential for coordinating response efforts. The attention mechanisms' robustness to environmental noise enables reliable damage detection even when atmospheric conditions are suboptimal. The multi-scale capability captures both individual building damage and broader neighborhood impacts.

Infrastructure Monitoring: Transportation agencies can monitor changes to transportation infrastructure—road widening, bridge modifications, parking lot construction. The automated monitoring enables proactive maintenance planning and development compliance verification.

C. Limitations and Future Directions

While Deep-SiamChange demonstrates strong performance, several limitations warrant acknowledgment and suggest directions for future investigation.

Spectral Information: The current architecture processes RGB imagery, limiting the spectral information available for change discrimination. Multispectral or hyperspectral imagery could provide additional discriminative features, particularly for distinguishing vegetation changes from structural changes.

Height Information: The architecture processes 2D imagery without explicit height information. Building height changes—vertical construction or demolition—can only be inferred from 2D footprint changes and shadow analysis. Integration of digital elevation models or stereo imagery could enable direct height change detection.

Temporal Modeling: The current architecture processes exactly two temporal images. Urban monitoring applications frequently involve continuous image streams requiring detection of changes across multiple time points. Extension to multi-temporal analysis could enable tracking of construction progress and development trajectories.

Semantic Change Classification: The current binary classification distinguishes changed from unchanged regions but does not classify change types. Semantic change detection—distinguishing construction from demolition, residential from commercial development—would provide additional value for urban planning applications.

VIII. CONCLUSION

This investigation has presented Deep-SiamChange, a novel architecture for robust structural change detection in urban environments. The architecture integrates a Siamese twin encoder with multi-scale attention mechanisms and convolutional block attention modules to achieve time-invariant and noise-robust feature extraction. The Siamese encoder ensures consistent feature representation across temporal intervals, the multi-scale pyramid captures structural changes across varying spatial extents, and the attention mechanisms filter environmental noise while emphasizing structural patterns.

Experimental evaluation on the LEVIR-CD benchmark dataset demonstrates substantial performance improvements. The full Deep-SiamChange architecture achieves an F1-score of 88.20%, representing a 4.28 percentage point improvement over the baseline. The architecture exhibits particular effectiveness in mitigating false positives arising from misregistration errors and illumination variations—persistent challenges in practical deployment scenarios.

The practical significance of this work extends to multiple urban governance applications. Automated change detection enables efficient monitoring of construction activity, identification of unrecorded property modifications, and rapid assessment of post-disaster structural damage. The robustness to environmental noise ensures reliable operation under the varied conditions encountered in operational deployments.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15420

Future investigations will explore integration of spectral and height information, extension to multi-temporal analysis, and development of semantic change classification capabilities. These extensions will further enhance the utility of automated change detection for urban analytics applications.

REFERENCES

[1]. Chen, H., & Shi, Z. (2020). A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. Remote Sensing, 12(10), 1662.
[2]. Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 3-19.
[3]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.
[4]. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117-2125.
[5]. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 234-241.
[6]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 5998-6008.
[7]. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-Local Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794-7803.
[8]. Zhang, L., Chen, H., Li, H., & Shi, Z. (2019). TBSRL: A Triplet Baseline for Remote Sensing Image Change Detection. Remote Sensing, 11(18), 2171.
[9]. Zhan, Y., Fu, K., Yan, M., & Sun, X. (2017). Change Detection Based on Deep Siamese Convolutional Network. IEEE International Geoscience and Remote Sensing Symposium, 4444-4447.
[10]. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2881-2890.

