scientific reports

Check for updates

OPEN

# Dual-stream hybrid architecture with adaptive multi-scale boundary-aware mechanisms for robust urban change detection in smart cities

Israr Ahmad[1], Fengjun Shang[1,5], Muhammad Salman Pathan[2], Ahsan Wajahat[3] & Yun-Su Kim[4]✉

**Urban environments undergo continuous changes due to natural processes and human activities, which necessitates robust methods for monitoring changes in land cover and infrastructure for sustainable developments. Change detection in remote sensing plays a pivotal role in analyzing these temporal variations and supports various applications, including environmental monitoring. Many deep learning-based methods have been widely investigated for change detection in the literature. Most of them are typically regarded as per-pixel labeling and show their dominance, but they still struggle in complex scenarios with multi-scale features, imprecise & blurring boundaries, and domain shifts between temporal shifts. To address these challenges, we propose a novel Dual-Stream Hybrid Architecture (DSHA) that combines the strengths of ResNet34 and Modified Pyramid Vision Transformer (PVT-v2) for robust change detection for smart cities. The decoder integrates a boundary-aware module, along with multiscale attention for accurate object boundary detection. For the experiments, we incorporated the LEVIR-MCI dataset, and the results demonstrate the superior performance of our approach by achieving an mIoU of 92.28% and an F1 score of 92.50%. Ablation studies highlight the contribution of each component by showing significant improvements in the evaluation metrics. In comparison with existing methods, DSHA outperformed the existing state-of-the-art methods on the benchmark dataset. These advancements demonstrate our proposed approach's potential for accurate and reliable urban change detection, making it highly suitable for smart city monitoring applications focused on sustainable urban development.**

**Keywords** Change Detection, Remote Sensing, Dual-Stream Encoder, Smart Cities Monitoring

The dynamic properties of urban environments are continuously influenced by both natural processes and human activities, resulting in the constant transformation of the Earth's surface. Understanding and monitoring these changes has become increasingly critical for urban planning, environmental management, and sustainable development. Change detection (CD), as a fundamental technology in earth observation, provides an essential tool for analyzing and quantifying these temporal variations in land cover and urban structures. The basic objective of CD is to detect and identify significant changes between bi-temporal remote sensing images of the same geographical region, enabling comprehensive interpretation of surface modifications over time. This capability has proven invaluable across numerous applications, including urban expansion monitoring, natural disaster assessment, land-use change analysis, and environmental protection.

In the context of smart cities, CD systems are essential for supporting automated urban planning[1], real-time infrastructure monitoring[2], and evidence-based policy decisions[3]. The integration of robust change detection systems into smart city frameworks facilitates continuous assessment of urban development patterns and serves

[1]Department of Computer Science & Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. [2]School of Computing, Dublin City University, Dublin, Ireland. [3]School of Software, Northwestern Polytechnical University, Xi'an, China. [4]Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. [5]Fengjun Shang contributed equally to this work. ✉email: yunsukim@gist.ac.kr

as a critical tool for urban planners and municipal decision-makers. By providing comprehensive monitoring capabilities, CD systems enhance informed governance, infrastructure management, and regulatory compliance. These systems enable data-driven governance through real-time monitoring of urban transformations and facilitate urban planning and resource allocation that directly support the achievement of Sustainable Development Goal 11: making cities and human settlements inclusive, safe, resilient, and sustainable[4,5].

Traditional change detection approaches, primarily based on manual feature extraction and pixel-level comparison, have shown significant limitations in handling complex scenarios and large-scale datasets[6]. These methods often struggle with the inherent variability in remote sensing imagery, including illumination changes, seasonal variations, and atmospheric effects. Moreover, their reliance on hand-crafted features limits their ability to capture subtle changes and adapt to diverse urban landscapes[7]. Over the past decade, deep learning technology has revolutionized the field of change detection through Convolutional Neural Networks (CNNs), which demonstrate remarkable success due to their superior learning ability and automatic feature extraction capabilities[8,9]. Early CNN-based approaches focused on adapting semantic segmentation architectures and treating the change detection as a binary classification problem at the pixel level[10]. These methods typically process bi-temporal images either through a single-stream architecture with early fusion[11] or siamese networks with separate feature extraction paths[12].

The advancement of these deep learning methods employs model fusion that combine different satellite imagery, like multispectral[13], multi-layer attention mechanisms for precise feature extraction[8], generative models like GANs for improved change mapping in noisy environments[14,15], 3D-CNNs for multi-temporal analysis[16], and temporal feature refinement for continuous urban monitoring[17]. Furthermore, the introduction of Vision Transformers (ViTs), marks a significant milestone in deep learning advancement for object detection. A Vision Transformer in principle divides an image into patches, processes them through layers to capture features, and outputs bounding box coordinates and class predictions for detected objects[6,18]. Several studies have explored ViTs in various configurations, and have showed significant improvements in detecting changes in urban features such as buildings and roads[19].

While these deep learning-based approaches have shown promising results, but using them as a single solution often faces problems. Specially, the CNN-based models often struggle to effectively capture the complex relationships between temporal features[16] and maintain spatial consistency in the change detection results. Similarly, the transformer-based models alone face challenges in effectively integrating local and global features while maintaining temporal coherence throughout the network, particularly when dealing with small objects and complex urban structures[20]. The primary challenge lies in the network's inability to simultaneously preserve fine-grained spatial details and capture broader contextual information[21]. This limitation often leads to either over-segmentation or missed changes, particularly in complex urban environments where transformations occur at various spatial scales[22]. The situation becomes more complex due to suboptimal handling of multi-scale temporal relationships and domain shifts between different temporal states, which impacts the system's overall robustness and generalization capability[23]. When we look at the mathematical formulation of these challenges, it can be expressed as minimizing the empirical loss $L(F\theta(X1, X2), Y)$, where X1 and X2 represent bi-temporal images and Y shows the ground truth change map. What makes this different from regular semantic segmentation is the need for state-of-the-art mechanisms that can analyze corresponding features at different time points while dealing with real-world complications like climate changes, preprocessing variations, and different imaging conditions[24]. Although the hybrid models can be a solution to these challenges[25], but simple hybrid approaches still face challenges in achieving robust change detection performance. Issues such as false positives, missed detections, and boundary blurring persist due to factors like lighting changes and scale differences[15,26].

To address these challenges, we propose a sophisticated hybrid architecture, a novel dual-stream hybrid architecture that combines the strengths of both CNN and transformer architectures. We introduced a modified U-Net with a state-of-the-art dual-stream integration mechanism as the backbone, incorporating a customized PVT-v2 and ResNet34 that work in parallel, where PVT-v2 handles global context through its transformer-based structure while ResNet34 preserves fine-grained spatial details. This dual-stream approach is enhanced by cross-attention fusion mechanisms and adaptive boundary-aware modules that enable effective feature interaction while maintaining the distinctive characteristics of each temporal state, allowing for more precise detection of changes at various scales. The core innovation lies in our hierarchical feature processing pipeline, which integrates our carefully designed cross-stream connections and boundary-aware modules within an improved decoder framework. Our architecture maintains distinct processing streams while enabling controlled interaction through cross-attention fusion mechanisms, allowing for effective modeling of temporal dependencies. It is further strengthened by a multi-scale attention mechanism that dynamically processes features across different spatial scales, enhancing change detection accuracy in urban environments. The main contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first work that synergistically combines a customized Pyramid Vision Transformer (PVT-v2) and ResNet34 in a dual-stream architecture as the backbone for a U-Net framework. The customization includes modified 6-channel inputs for bi-temporal images, enhanced spatial-channel attention mechanisms in PVT-v2 for global context, and attention gates in ResNet34 for adaptive feature fusion. This design enables simultaneous capture of both global context and fine-grained details.
- We propose a specialized boundary-aware module (BAM) that integrates Sobel operators with a multi-scale attention mechanism. Unlike existing fixed edge-detection methods, our approach dynamically adjusts to different scales of urban changes, ensuring improved boundary detection. We also incorporated a hierarchical cross-temporal attention (CTA) mechanism that intelligently fuses features across temporal states.
- A combined loss function incorporating four losses addresses class imbalance and helps the proposed model to adjust its weights to improve the segmentation overlap.

Through experiments, we demonstrate that these innovations work together to achieve superior performance in challenging urban scenarios, particularly in cases where subtle changes and complex structures are involved.

## Related works

### CNN-based remote sensing change detection methods

CNN-Based Remote Sensing Change Detection Methods In the recent literature, CNN-based remote sensing change detection methods are widely employed. In these studies, most approaches adopted an encoder and decoder architecture along with the Siamese encoder, which extracts deep representation from the bi-temporal images, followed by a decoder that generates a pixel-wise change map. For example, Zhang et al.[27] proposed a two-channel CNN with shared weights to generate multi-scale feature difference maps in remote sensing change detection. Similarly, Mou et al.[28] introduced the CNN combined with a recurrent neural network to learn the temporal dependencies in remote sensing images. On top of this, to capture high contextual information, the researchers focused on incorporating multiscale feature learning strategies such as deep supervision[29] and dense connection[29]. For the rich contextual information, the method MFINet[30] incorporating cross-scale interactions to improve the change-aware representations, was also used, and other studies[31,32] also focused on the cross-scale feature interaction and feature fusion methods to obtain the change-aware representations.

However, the traditional CNN-based change detection methods often struggled with disintegrated object boundaries and noise due to pixel-wise classification. To address this problem, object-based methods gained importance. For instance, Wang et al.[33] used the ensembled learning on multiple features to preserve object boundaries in urban environments. Zhang et al.[34] proposed the squeeze-and-excitation W-Net to fuse the multi-source data to reduce the noise in prediction changes. Recent advances in CNN architectures have introduced sophisticated boundary-aware mechanisms to address these limitations. Yang et al.[35] proposed an enhanced hybrid CNN and transformer network (EHCTNet) that integrates dual-branch feature extraction with boundary refinement modules for robust change detection. Additionally, Li et al.[36] developed a change detection network based on transformer and transfer learning, focusing on boundary completeness and internal compactness in change regions.

Furthermore, attention mechanisms such as multiscale features were developed to capture the object information at multiple scales from remote sensing images[37,38]. The evolution of CNN-based approaches has led to increasingly sophisticated architectures that address the challenges of arbitrary-oriented object detection in aerial imagery. Notably, Huang et al.[39] introduced task-wise sampling convolutions (TS-Conv) for arbitrary-oriented object detection in aerial images, demonstrating adaptive feature sampling from task-specific sensitive regions. This approach addresses the inconsistent features between localization and classification tasks that often constrain detection performance in complex aerial scenes. The existing studies predict the change detection in remote sensing images using the pixel-wise assignment of changed or unchanged. Although this approach tries to capture various changes in the bi-temporal images, it often generates the detected objects in fragmented boundaries and isolated noises due to the convolutional operations.

### Transformer-based remote sensing change detection methods

Recent transformers with their self-attention mechanisms have been widely adopted to address the shortcomings of the CNN-based methods in remote sensing change detection. These models can model the long-range dependencies in the remote sensing images. Chen et al.[40] proposed a transformer-based framework to enhance the performance in capturing global semantic information for remote sensing change detection. Bandra et al.[41] presented a hierarchical transformer encoder for better feature learning. Similarly, Liu et al.[42] proposed AMTNet, a multi-scale transformer network, to combine its strength with CNNs. Likewise, Zheng et al.[43] introduced a hybrid architecture, L-Former, to get good results on the remote sensing benchmark datasets. In another study[44] transformer was used to tokenize the global contextual features obtained into patch-wise features.

The advancement of transformer architectures has led to more sophisticated change detection frameworks that specifically address the challenges boundaries and diverse shapes in change areas. Advanced transformer-based methods have also incorporated domain-specific attention mechanisms to improve change detection accuracy. Yin et al.[45] proposed CTCANet, a CNN-transformer network combining convolutional block attention module (CBAM) for change detection in high-resolution remote sensing images, demonstrating superior performance in capturing both local details and global context. Although transformers have gained significant success in context modeling in remote sensing change detection, they often struggle to extract the local and fine-grained information of objects in long-range modeling dependencies.

### Attention mechanism

The attention mechanism is widely adapted in the many deep learning-based models, including encoder-decoder-based architectures, to focus on the spatial and channel features. The attention mechanism also plays a pivotal role. In remote sensing change detection tasks, as these images contain objects at different scales and color ranges. Peng et al.[46] proposed attention in image difference to overcome false positives to improve the accuracy of change detection in remote sensing images. Similarly, Eftekhari et al.[47] presented a parallel spatial attention block for the change detection task to reduce the false alarms caused by occlusions. Feng et al.[48] proposed ICIF-Net using multiscale attention to capture the local and global contextual information. Furthermore, Jiang et al.[49] introduced a multi-scale difference and attention network (MDANet) for high-resolution city change detection. Building upon this work, Zhan et al.[50] proposed AMFNet, an attention-guided multi-scale fusion network for bi-temporal change detection that integrates innovative feature fusion techniques for enhanced performance.

Li et al.[51] introduced a multi-scale fusion Siamese network based on a three-branch attention mechanism for high-resolution remote sensing image change detection, addressing challenges in edge detection and small target detection. Likewise, Farooque et al.[52] proposed a dual attention-driven multi-scale multi-level feature

fusion approach for hyperspectral image classification. Recent studies highlighted the importance of attention mechanisms in improving the detection performance in remote sensing change detection[53]. Furthermore, Guo et al.[54] developed MSFNet, a multi-scale spatial-frequency feature fusion network that replaces traditional CNN operations with shift window self-attention (SWSA) for direct processing of remote sensing images. However, attention-based models often need careful design to balance between computational efficiency and accuracy because the unbalanced and excessive attention mechanism in any model can lead to reduced computational efficiency.

### Feature fusion

In many deep learning tasks, feature fusion is considered an important process, including segmentation[55], multimodal tasks[56], and classification[57]. Liu et al.[58] introduced SoftFormer, a SAR-optical fusion transformer for urban land use and land cover classification, demonstrating the effectiveness of transformer-based fusion strategies. Their approach employs an interior self-attention mechanism that enables shallow transformer layers to extract local features similar to CNNs while maintaining the global modeling capabilities of transformers. The fusion in deep learning models happens at multiple scales, with heterogeneous features and multiple levels. The feature fusion in some cases is just a simple operation based on the problem level; for some cases, simple concatenation can be enough[59]. The feature fusion can be more reliable and flexible when using multiple attention-based techniques[60–62]. In remote sensing, feature fusion mechanisms are employed using attention mechanisms, temporal, and spatial-temporal attentions. Furthermore, using the flow field and deformable convolution[63] to focus on the alignment-based fusion to align features of different levels in the spatial dimensions[64,65].

## Materials and methods

The proposed methodology introduces a novel hybrid semantic segmentation architecture for change detection in remote sensing imagery for smart cities, monitoring, and development. At its core, the system employs a modified U-Net framework in which the traditional single encoder is replaced by an enhanced dual-stream encoder as the backbone, along with multiple specialized modules. The Channel Attention module utilizes average and maximum pooling with shared MLP networks for channel-wise attention, while the Spatial Attention module employs convolution for spatial relationships. A key innovation in the architecture is the Boundary Aware Module using Sobel operators for object boundary detection, the Multi-Scale Attention module processing at different scales for multi-scale context, and the Improved Decoder module combining boundary awareness with multi-scale attention. The architecture is enhanced by the Cross Attention Fusion module for transformer-like feature exchange and dynamic feature combination through channel attention.

### Dual stream encoder

The modified UNet encompasses a dual-encoder stream that integrates the strengths of both transformer-based and convolutional neural network-based architectures for advanced semantic segmentation for change detection in smart cities. As the image pairs in CD are captured across time, changes occur at different scales and vary at different scales, like small changes (e.g., new small buildings), medium changes (e.g., road construction in the t2 image), and large-scale changes like urban development. To extract the hierarchical changes, the Pyramid Vision Transformer (PVT-v2) is employed to process the input through multiple stages. This architecture gives a bigger picture, like a bird's-eye view. In parallel, we incorporated the ResNet34 encoder to capture complementary features, which provides strong local feature extraction capabilities and focuses on fine details, much like having a magnifying glass. To adapt the ResNet34 encoder for 6-channel input (3 channels $\times$ 2 temporal images), we initialize the first convolutional layer by averaging the pretrained weights across the RGB channels and replicating them for the temporal pairs. Let $W_{\text{pretrained}} \in \mathbb{R}^{64\times3\times7\times7}$ represent the pretrained weights from ImageNet. We compute channel-averaged weights $\mathrm{W}_{\text{avg}}$ as follows:

$$W_{\text{avg}} = \frac{1}{3}\sum_{c=1}^{3} W_{\text{pretrained}}[:, c, :, :]$$

These averaged weights are then replicated across the temporal pairs to initialize the 6-channel input:

$$W_{\text{init}} = \text{Repeat}(W_{\text{avg}}, \text{dim} = 1, \text{repeats} = 2)$$

This initialization preserves the spectral characteristics of the pretrained model while enabling effective processing of bi-temporal imagery. The dual-stream encoder architecture processes image pairs *t*1, and *t*2, each image is represented as a three-dimensional tensor $I_{t1}, I_{t2} \in \mathbb{R}^{3\times H\times W}$ $X = [I_{t1}||I_{t2}] \in \mathbb{R}^{6\times H\times W}$, where 3 represents the RGB color channels. So, each image is represented as a tensor of size 3×256×256. These two images are concatenated to form a single input . The || symbol represents the concatenation of 3 channels from the first image with the 3 channels from the second image, resulting in a 6-channel input and creating a tensor of size 6×256×256.

*Pyramid vision transformer stream processing*

The modified UNet encompasses a dual-encoder stream that integrates the strengths of both transformer-based and convolutional neural network-based architectures for advanced semantic segmentation for change detection in smart cities. As the image pairs in CD are captured across time, changes occur at different scales and vary at different scales, like small changes (e.g., new small buildings), medium changes (e.g., road construction in the t2 image), and large-scale changes like urban development. To extract the hierarchical changes, the Pyramid Vision

Transformer (PVT-v2) is employed to process the input through multiple stages. This architecture gives a bigger picture, like a bird's-eye view. In parallel, we incorporated the ResNet34 encoder to capture complementary features, which provides strong local feature extraction capabilities and focuses on fine details, much like having a magnifying glass. To adapt the ResNet34 encoder for 6-channel input (3 channels × 2 temporal images), we initialize the first convolutional layer by averaging the pretrained weights across the RGB channels and replicating them for the temporal pairs. Let $W_{\text{pretrained}} \in \mathbb{R}^{64\times3\times7\times7}$ represent the pretrained weights from ImageNet. We compute channel-averaged weights $\mathrm{W}_{\text{avg}}$ as follows:

$$W_{\text{avg}} = \frac{1}{3}\sum_{c=1}^{3} W_{\text{pretrained}}[:, c, :, :]$$

These averaged weights are then replicated across the temporal pairs to initialize the 6-channel input:

$$W_{\text{init}} = \text{Repeat}(W_{\text{avg}}, \text{dim} = 1, \text{repeats} = 2)$$

This initialization preserves the spectral characteristics of the pretrained model while enabling effective processing of bi-temporal imagery. The dual-stream encoder architecture processes image pairs *t*1, and *t*2, each image is represented as a three-dimensional tensor $I_{t1}, I_{t2} \in \mathbb{R}^{3\times H\times W}$ $X = [I_{t1}||I_{t2}] \in \mathbb{R}^{6\times H\times W}$, where 3 represents the RGB color channels. So, each image is represented as a tensor of size 3×256×256. These two images are concatenated to form a single input . The || symbol represents the concatenation of 3 channels from the first image with the 3 channels from the second image, resulting in a 6-channel input and creating a tensor of size 6×256×256.

*Pyramid vision transformer stream processing*

The PVT-v2 encoder serves as the primary feature extraction stream, with a hierarchical transformer structure to process visual information at multiple scales. The PVT-v2 encoder stream starts with an overlapping patch embedding layer, which projects the input image into a sequence of tokens:

$$P_{emb}(x) = Conv2D(x, k_{patch}, s_{stride}) + PE$$

Where $P_{emb}(x)$ represents the patch embedding output, *PE* are learnable position embeddings to encode spatial information, $k_{patch}$ is the patch size, and $s_{stride}$ is the stride. This encoder then processes these tokens through four progressive stages, each maintaining different feature dimensions:

$$F_{PVT} = \{f_1^p \in \mathbb{R}^{64}, f_2^p \in \mathbb{R}^{128}, f_3^p \in \mathbb{R}^{320}, f_4^p \in \mathbb{R}^{512}\}$$

Here, $f_1^P, f_2^P, f_3^P, f_4^P$ represent multi-scale feature maps extracted from the 1st, 2nd, 3rd, and 4th encoder blocks of PVT-v2 backbone respectively, with channel dimensions 64, 128, 320, and 512.

Each stage captures increasingly complex temporal changes by implementing a modified transformer block that includes multi-head self-attention (MHSA) with spatial reduction. In which, *Q*, *K*, *V* represent query, key, and value, feature maps respectively, $W^O$ is the output projection matrix, and $W_i^Q, W_i^K, W_i^V$ are per-head projection matrices for $i = 1, ..., h$ where *h* is total the number of attention heads.

$$MHSA(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

where each head computes:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

And a feed-forward network (FFN) with depth-wise convolution processes through three sequential operations:

$$FFN(x) = MLP(DWConv(MLP(x))$$

The structural diagram of the modified PVT-V2 is presented in Fig. 1.

*Residual network stream processing*

The ResNet34 stream processes in parallel, providing complementary analysis through its initial modified convolution operation:

$$x_{conv1} = Conv2d(x_{concat}, channels_{out} = 64, kernel = 7, stride = 2)$$

Where $x_{concat}$ represents the bi-temporal image pair input. The initialization process includes ReLU activation and batch normalization. The ResNet stream processes the input through modified convolution layers:

$$X_{init} = ReLU(BN(Conv_{7\times7}(X)))$$

Where, weights specifically adapted for 6-channel input through channel-wise averaging:

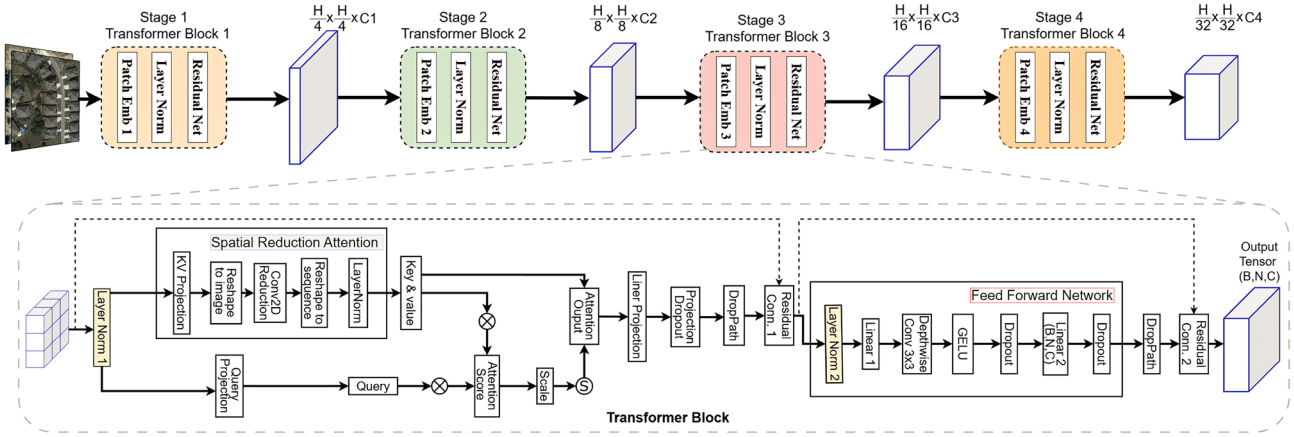


**Fig. 1**. Structural diagram of the modified Pyramid Vision Transformer (PVT-v2) encoder. The architecture processes the input through overlapping patch embedding, hierarchical transformer blocks, and MHSA mechanisms. The figure highlights the four progressive stages (f p 1 to f p 4) with varying feature dimensions, which enables the model to extract hierarchical representations for precise change detection. The satellite image on the left side is from the Levir-MCI dataset[66].

$$W_{6ch} = \frac{1}{3}\sum_{i=1}^{3} W_{3ch} \cdot 1_{1\times 6\times 7\times 7}$$

This stream maintains its hierarchical feature extraction through residual blocks $R_i = ResBlock_i\left(X_{i-1}\right),\ i \in 1, 2, 3, 4$, producing a feature hierarchy:

$$F_{ResNet} = \{f_1^r \in {}^{\mathbb{R}64}, f_2^r \in \mathbb{R}^{128}, f_3^r \in \mathbb{R}^{256\to 320}, f_4^r \in \mathbb{R}^{512}\}$$

With dimensions $C_i = \{64, 128, 256, 512\}$, where each level serves specific purposes: the first level $(f_1^r)$ captures fine-grained temporal differences in textures and edges, the second level $(f_2^r)$ identifies changes in local patterns and shapes, the third level $(f_3^r)$ adapts its features to match the PVT-v2's dimensional space for better feature fusion, and the fourth level $(f_4^r)$ comprehends complex local transformation patterns. The structural diagram for the ResNet34 encoder is presented in the Fig. 2.

## Channel and spatial attention mechanisms

The proposed methodology introduces a dual-attention mechanism specifically designed for detecting changes in remote sensing imagery from the Levir-MCI dataset, which contains bi-temporal high-resolution satellite images focusing on building and road changes. The architecture implements two complementary attention modules: Channel Attention (CA) and Spatial Attention (SA), working in concert to enhance feature representation for precise change detection.

*Channel attention mechanism*

The channel attention (CA) module advances the feature representation by dynamically computing the channel-wise importance weights for change detection. First, the module computes global average and max pooling to capture channel-wise statistics:

$$F_{avg} = \frac{1}{H \times W}\sum_{ij} X(i,\ j), \quad F_{\max} = \max_{ij} X(i,j)$$

Here, $F_{avg}$, $F_{\max}$ are average-pooled and max-pooled feature maps, $H$, $W$ are the height and width (spatial dimensions) of input feature map, $X_{(i,j)}$ input feature at spatial position. Through average and max pooling, the mean activation and most prominent activation per channel are captured.

$$A_r = Conv1d(F_{avg},\ W_r)$$

In this operation, a shared bottleneck is achieved through dimension reduction. The $A_r$ represents the reduced feature map, $Conv1d$ represents the convolution operation, and $W_r$ represents the learnable weights for dimension reduction ratio.

$$A_e = \mathrm{Re}LU(Conv1d(A_r,\ W_e))$$

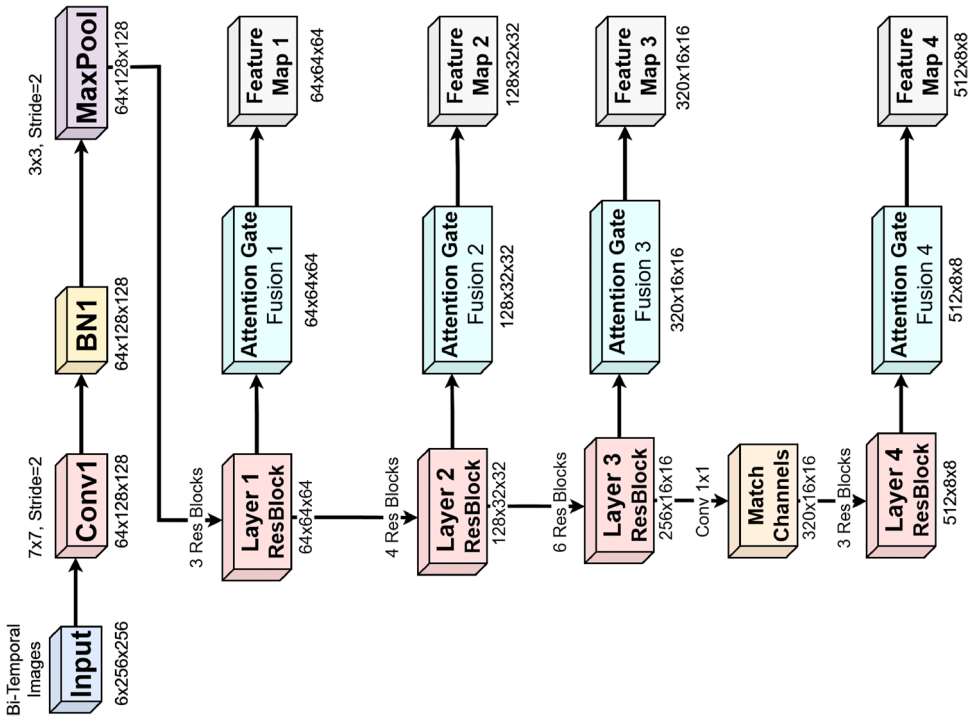


**Fig. 2**. Structural diagram of the modified ResNet34 encoder. The input is modified to accept 6 channels (3 channels x 2 temporal images), Conv1 weights are initialized by averaging pretrainted RGB weights. The match channel layer converts the 256 to 320 channels to align with PVT-v2, and attention gates fuses features from ResNet34 and PVT-v2 streams.

Through this Equation, $\mathrm{Re}LU$ adds the non-linearity, and $W_e$ represents to learnable weights for channel expansion. This operation generates channel-wise importance weights.

$$F_c = \sigma(A_e) \odot X$$

Finally, the refined features are derived; $\sigma(A_e)$ representing the normalization of the attention weights for channel importance and $\odot$ representing the channel-wise multiplication. The final operation emphasizes on critical channels for detecting changes from the remote sensing images. The CA mechanism is like an adaptive lens system that functions like an intelligent filter that provides both a bird's eye view and a magnifying glass. Just as an aerial photographer uses specialized color filters to highlight specific features in the urban environment.

*Spatial attention*

The spatial attention (SA) module highlights the spatial locations, which plays a critical role in detecting the localized changes; in our case, these changes could be the development of the road or building or the demolishing of these. SA first calculates the channel-wise average and max feature maps and concatenates both maps and processes through the convolution layer to generate a spatial attention map.

$$S = \sigma(Conv2d(Concat(F_{avg},\ F_{\max});\ W_s))$$

Here, $W_s$ represents the learnable weights for spatial attention. The refined feature map $F_c$ is obtained by multiplying the channel-wise attention feature map with the spatial attention feature map *S*. $F' = S \odot F_c$ The combination of CA and SA modules ensures the model focuses on spatial regions where changes are most likely to occur, improving detection accuracy for small-scale urban changes. The SA mechanism acts like a dynamic spotlight system that enhances both bird's eye view and magnifying glass capabilities. The SA module creates intelligent attention maps that guide both global and local processing streams to focus computational resources on spatial regions where changes are most likely to occur.

## Boundary-aware feature enhancement

The proposed methodology introduces an advanced boundary-aware module (BAM) specifically designed to emphasize on the boundaries of the objects for detecting changes in remote sensing imagery from the Levir-MCI dataset. The BAM incorporates an advanced dual-directional gradient computation approach based on the Sobel operators ($3 \times 3$ matrices) to extract boundary information from feature maps *X*, representing the number of channels, height, and width. The horizontal and vertical gradients are computed using two-dimensional convolution operations with Sobel kernels:

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

Here, $K_x$ represents the horizontal Sobel Kernel, and $K_y$ represents the vertical Sobel Kernel. The horizontal and vertical gradients ($E_x,\ E_y$) are computed using the two-dimensional convolution operations with Sobel Kernels:

$$E_x = Conv2d(X,\ K_x), \qquad E_y = Conv2d(X,\ K_y)$$

After computing the horizontal and vertical gradients, these gradients are combined through the Pythagorean theorem to compute the edge magnitude of each pixel. For each pixel position, it takes the square root of the sum of squared gradients, giving us the total edge strength.

$$E = \sqrt{E_X^2 + E_y^2}$$

The final-boundary enhanced feature map $F'$ is obtained by element-wise multiplication with sigmoid-activated $\sigma$ edge map.

$$F' = F \odot \sigma(E)$$

The integration of BAM significantly improves feature representation by emphasizing boundary information, leading to more accurate change detection in urban environments from remote sensing images.

## Multi-scale attention mechanism

The multi-scale attention (MSA) mechanism is incorporated for change detection in remote sensing from smart cities. The changes in urban environments (like single buildings, stacks of buildings, and roads)happen at various scales, and these changes look different from different perspectives. The MSA framework decomposes the input into a feature map. We implemented the three scales decomposition strategy using bilinear interpolation, which can be represented as:

$$F_s = \psi\left(F, H_s, W_s\right), \quad S \in \ \{1.0,\ 0.5,\ 0.25\}$$

Here, $H_s,\ W_s$ is the target height and width for the scaled feature map, *F* which represents the input feature map, and $\psi$ represents the bilinear interpolation function, which resizes the feature map to desired dimensions. Each scaled feature map $F_s$ is then processed using scale-specific convolutional layers to capture multi-scale contextual information.

$$F'_s = Conv2d(F_s,\ W_s)$$

The $W_s$ in this context represents learnable weights for each scale. The bilinear interpolation is again used to upsample the enhanced feature for all scales to the original resolution.

$$F_s^{up} = \psi(F',\ \ H, W)$$

Here, $F_s^{up}$ represents the upsamples feature map at scale *s*. All three upsampled feature maps are then concatenated to form the final multi-scale attention feature map, the mathematical representation is as follows:

$$F'_{msa} = Concat(F_{s1}^{up}, F_{s2}^{up}, F_{s3}^{up})$$

This concatenated feature map is then processed through a convolutional layer to refine the multi-scale features.

$$F'_{msa} = Conv2d(F_{msa},\ \ W_{msa})$$

Here, $W_{msa}$ represents the learnable weights for the convolutional layer. The refined $F'_{msa}$ enhances the model's capability to capture the changes at multiple scales and improves detection accuracy for both small and large-scale buildings and roads from remote sensing images. The MSA mechanism extends the dual perception system by implementing a telescopic zoom array that operates simultaneously at multiple magnification levels. The primary bird's-eye view provides the standard aerial perspective and the magnifying glass focuses on fine details. The MSA creates additional viewing scales, like deploying multiple surveillance drones at different altitudes (100%, 50%, and 25% zoom levels) over the same urban area.

## Cross-temporal attention mechanism

The cross-temporal attention (CTA) module is incorporated to enable the model to correlate the features from bi-temporal remote sensing images. The CTA enhances the ability to detect structural changes (such as buildings and roads) over time in evolving smart city environments. The CTA mechanism first projects feature maps from the two temporal states into query (*Q*), key (*K*), and value (*V*) spaces using the learnable metrics, mathematically represented as follows:

$$Q = W_q.F_1, \qquad K = W_k.F_2, \qquad V = W_v.F_2$$

Here, $F_1$, $F_2$ represents feature maps from the first and second temporal states. This formulation transforms the features into spaces like $Q$, $K$, and $V$ which are helpful in the attention computation. To compute the attention scores scaled dot-product attention is used.

$$A = SoftMax(\frac{Q.K^T}{\sqrt{d_k}})$$

Here, $d_k$ represents the dimension of the key vectors. This equation calculates the similarity between $Q$ and $K$ vectors, and enables the model to identify corresponding regions in both temporal states. From this a context-aware map is generated

$$C = A\,.\,V$$

This operation captures the relevant information from the second temporal state based on the computed attention scores. Next a residual connection is applied.

$$F_{cta} = F_1 + C$$

Finally, the module combines the original feature map from the first temporal state $F_1$ with a context-aware map $C$. The attained fused feature map integrates temporal dependencies and enhances the model's capability to detect changes from bi-temporal remote sensing images while skipping the irrelevant variations. The CTA operates like a time-lapse analyst who not only captures changes over time but also intelligently correlates specific regions between temporal states. It is a sophisticated photo alignment system that works through both our bird's-eye and magnifying glass perspectives.

### Decoder module

To process the feature maps from the encoder stages, we implemented a three-stream feature processing decoder as depicted in the Fig. 4 (block diagram for the proposed model). It refines BAM features and multi-scale features to generate the final change detection output. The decoder integrates features from the encoder and enhances them through up sampling and feature fusion. For each stage, the module first applies BAM to emphasize the boundaries of objects and fine details.

$$F_{bam} = BAM(F_{in})$$

Here, the represents to the boundary-aware module, and $F_{in}$ are the input features maps from the encoder stages. The BAM uses the Sobel operators (as explained in section 3.3) and MSA to refine the boundaries. Where the MSA is applied to capture the contextual information at different resolution or scales:

$$F_{msa} = MSA(F_{bam})$$

This operation combines the features from the multiple scales to pay attention to the fine-grained and contextual information. Then these feature maps are up sampled to the next resolution using the bilinear interpolation.

$$F_{up} = \psi(F_{MSA}, \quad scale = 2)$$

The scale factor, such as s=2, doubles the resolution for fine-grained analysis. The upsampled features are combined with residual connections from previous decoder steps.

$$F_{out} = F_{prev} + F_{up}$$

This final output is passed through as a convolutional layer to generate the predicted change map.

$$F_{final} = Conv2d(F_{out}, \quad W_{final})$$

The final feature is obtained through convolution. This step refines the features and predicts the changes from bi-temporal remote sensing images for smart cities monitoring and other applications. The detailed block diagram of our proposed DSHA architecture is presented in Fig. 3, which illustrates the integration of the different modules, and the visualisation of the DSHA architecture's processing pipeline, from input to images to the final change map, is shown in Fig. 4.

## Experimental results

### Datasets

We conducted experiments using the two change detection datasets.

**LEVIR-MCI dataset**: The LEVIR-MCI dataset[66] contains the 10077 bi-temporal aerial image pairs. The dataset contains a pre-image (A), an image taken before, and a post-image (B), which was taken after a time gap to indicate whether the area is changed or not, along with the grayscale & RGB segmentation masks. We have used the RGB masks in this study. The colormap of masks consists of three labels: (0, 0, 0) for the background

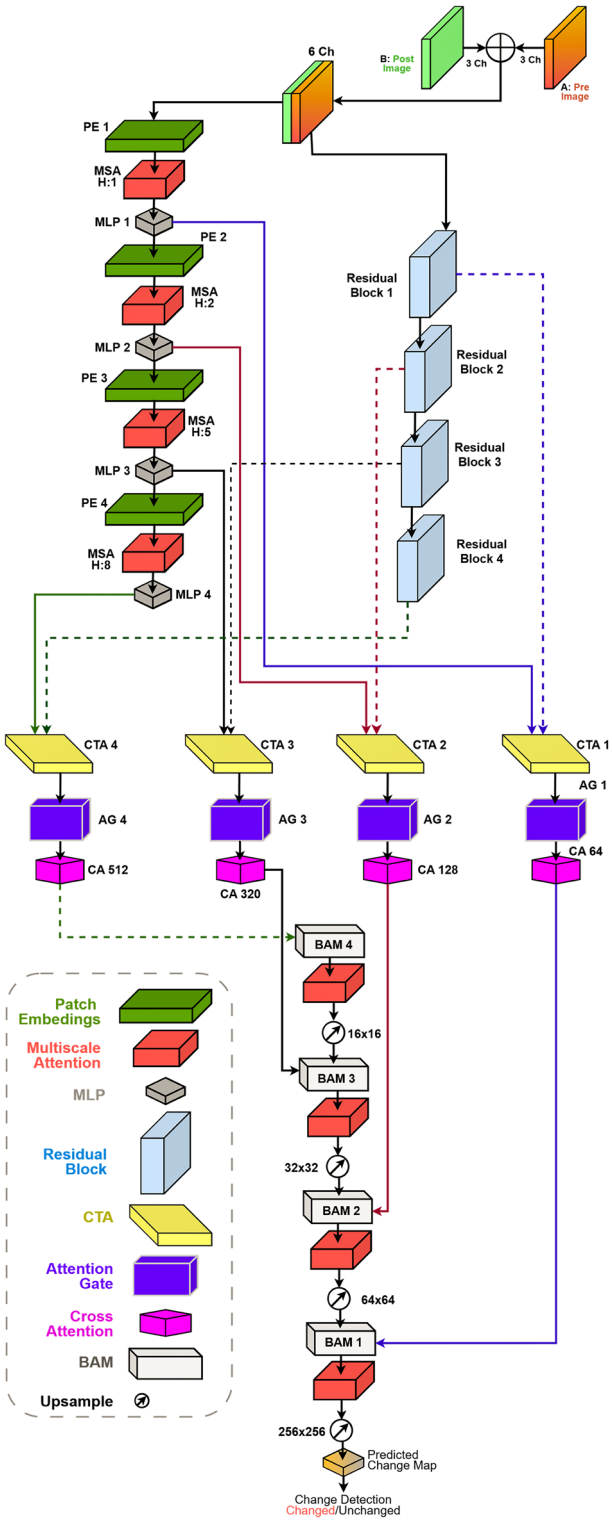


**Fig. 3**. Block diagram of the proposed dual-stream hybrid architecture. The architecture integrates a dual-stream encoder (PVT-v2 on the left and ResNet34 on the right), cross-attention fusion modules, boundary-aware modules, and a multi-scale attention mechanism. The diagram illustrates the flow of features through the network, highlighting the hierarchical processing, feature fusion, and refinement stages to generate the final change map.

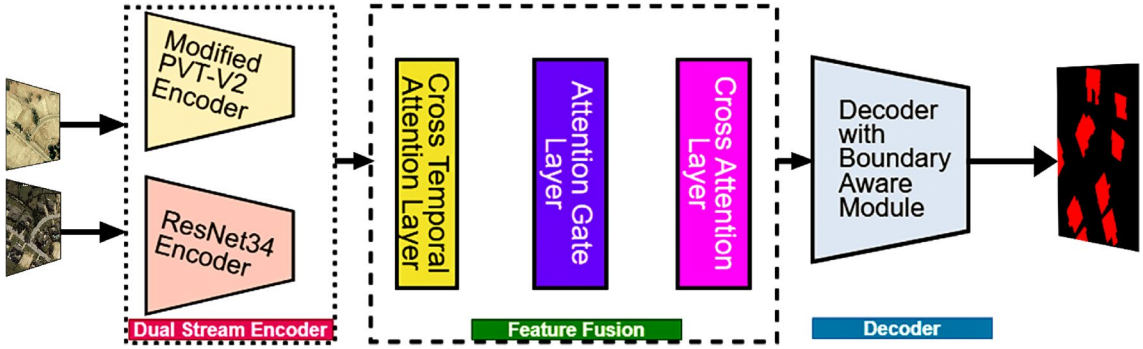


**Fig. 4**. Visualization of the proposed DSHA architecture's processing pipeline. (Left) Input bi-temporal satellite images form LEVIR-MCI[66]. (Center) Feature extraction via the dual-stream encoder, followed by cross-temporal attention and feature fusion. (Right) Decoder with boundary-aware modules, producing the final change detection output with refined boundaries.

pixel represented with 0; (255, 255, 0), which represents the roads with hot encoded with 1; and (255, 0, 0) for buildings, represented with 2. The distribution of the dataset is as follows: 6815 training image pairs, 1333 validation image pairs, and 1929 testing image pairs.

Before feeding the bi-temporal images pairs into the dual-stream encoder, our preprocessing pipeline applies the minimal transformations by resizing the images to 256x256 pixels and converting them to tensors with pixel values in the [0,1] range to perserve the spectral integrity of remote sensing imagery. We do not apply the ImageNet normalization paramters (also known as ImageNet stats) because our preliminary experiments revealed severe degradation loss of important spectral information by making images too dark when these stats were applied to LEVIR-MCI images; resultantly, it was compromising the change detection performance.

**Change Detection Dataset**: The Change Detection Dataset (CDD) consists of season-varying remote sensing images of the same region, obtained from Google Earth (DigitalGlobe)[67]. The CDD dataset contains cropped images with a size of 256 × 256 pixels and includes multiple augmentations[67]. The spatial resolution of CDD ranges from 3 to 100 cm/px. The dataset comprises 10,000 training image pairs, 3,000 validation image pairs, and 3,000 test image pairs. For preprocessing, we used the same settings as LEVIR-MCI. The images from the LEVIR-MCI dataset and CDD datasets are shown in Fig. 5.

## Loss function and evaluation metrics

To handle the class imbalance and improve the segmentation overlap, we incorporated the combined loss for the deep supervision. The combined loss comprises the four multiple losses: CrossEntropy (CE) loss, Dice loss, Focal loss, and Boundary Aware (BA) loss. The mathematical representation of CE is as follows:

$$L_{CE}^{(i)} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=0}^{C-1} y_{b,c}^{(i)} \log(p_{b,c}^{(i)})$$

Here, $B$ is the batch size of the images, $C$ is the number of classes (3 in our case), and $y_{b,c}^{(i)}$ is the ground truth label for class $c$ in the batch sample. $b$, $p_{b,c}^{(i)}$ is the predicted probability for class $c$ in output $i$. It guides the model to adjust weights so that the prediction matches with the ground truth. The Dice loss can be represented as below:

$$L_{Dice}^{(i)} = 1 - \frac{2.TP^{(i)} + \varepsilon}{2.TP^{(i)} + FP^{(i)} + FN^{(i)} + \varepsilon}$$

In Dice loss, the $\varepsilon$ is $1e-5$ as a soothing factor, $TP^{(i)}$ is true positives, $FP^{(i)}$ is false positives, and $FN^{(i)}$ is false negatives. The mathematical formulation of focal loss is as follows:

$$L_{Focal}^{(i)} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=0}^{C-1} a_c.(1 - p_{b,c}^{(i)})^{\gamma} \log(p_{b,c}^{(i)})$$

$\gamma$ is $\gamma = 2$, which is the focusing parameter, and $a_c$ are class weights (computed for dataset imbalance in our case). The formulation for BA loss is as follows:

$$L_{Boundary}^{(i)} = MSE(\partial_{pred}^{(i)},\ \partial_{t\arg et}^{(i)})$$

Here $\partial$ represents the gradient computation using the Sobel filters. The final output is the weighted sum of all these losses. It means each loss function has a different weight (importance) in the final calculation. The

**Fig. 5**. Sample bi-temporal satellite image pairs from LEVIR-MCI and CDD datasets showing various urban changes including building construction and road development. The upper row shows LEVIR-MCI dataset images[66] and bottom row shows CDD dataset images[67].

weighting scheme for deep supervision is as follows: for CE, 0.7; for Dice loss, 0.7; for Focal loss, 0.3; and for BA loss, 0.5. The total loss calculation formula is as follows:

$$L_{total} = \frac{1}{N}\sum_{i=1}^{N}(w_i.(0.3.L_{CE}^{(i)} + 0.7L_{DICE}^{(i)} + 0.3L_{Focal}^{(i)} + 0.5L_{BA}^{(i)}))$$

In the combined loss function, we set the CE weight to 0.3, which provided fundamental classification guidance for changed and unchanged pixels. Since our study focused on the CD task where spatial coherence matters more than individual pixel accuracy, we set the Dice loss weight to 0.7 to detect coherent changed regions (buildings and roads) rather than scattered pixels. Further we included the Focal loss was to address the imbalance between changed and unchanged pixels and to focus on learning hard-to-detect changes, like smaller buildings and narrow roads. To prevent over-emphasis on hard examples that could lead to noise detection, we set the Focal loss weight to a moderate value of 0.3. Finally, we incorporated BA loss to address blurry boundary issues, with a balance weight of 0.5 to precisely balance boundary detection and overall performance. This combination of the loss functions facilitates our model in dealing with class imbalance, smooth boundary prediction, and improving segmentation overlap, resulting in better performance.

For the evaluation, five commonly used metrics are employed: Accuracy (A), F1-Score (F1), precision (P), recall (R), and mean Intersection over Union (mIoU). Each evaluation metric serves an important purpose. The overaccuracy characterises the percentage of correctly detected pixels among all the samples. The F1 represents the harmonic mean of the precision and recall. The P represents the ratio of correctly detected changed pixels to all the pixels that are identified as changed in the map. The recall represents the percentage of correctly detected changed pixels to the number of all pixels that should be detected as changed pixels. While the mIoU, an important evaluation metric in the change detection task, represents the overlap between prediction and ground truth, it reveals the precise coverage of changed and unchanged pixels in the detected change map compared with ground truth. The mathematical representations of these metrics are as follows:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, *TP* is the correctly detected changed pixels, *TN* is the number of accurately detected unchanged pixels, *FP* is the number of incorrectly detected pixels, and *FN* is the number of missed detected changed pixels.

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = 2.\frac{P\,.\,R}{P + R}$$

$$mIoU = \frac{\Sigma_{c=0}^{C=1}(\frac{TP_c}{TP_c + FP_c + FN_c}) \times weight_c}{\Sigma_{c=0}^{C=1} weight_c}$$

Here, $weight_c$ is the number of ground truth pixels in class *c*.

## Implementation details

The proposed DSHA model is implemented based on the open-source deep learning framework PyTorch and initialized with two pretrained models, PVTv2 and ResNet34, for the backbone of the U-Net architecture. For model training, the AdamW optimizer was adopted with an initial learning rate of 5e-4, a weight decay of 0.0005, and parameters $\beta_1$ and $\beta_2$ as 0.9 and 0.999, respectively. The batch size was set to 8 for each GPU. In addition, we employ a OneCycleLR scheduler with a maximum learning rate of 1e-3, a cosine annealing strategy, and gradual reduction to ensure better model convergence. For the loss function, a combined loss is used, which incorporates CE loss, Dice loss, Focal loss, and BA loss with weights [0.3, 0.7, 0.3, 0.5], along with class weighting and deep supervision weights [1.0, 0.8, 0.6, 0.4]. All the experiments were conducted on the Nvidia RTX 4060 Ti with 16 GB RAM with the Windows 11 Pro operating system.

## Model complexity

The DSHA architecture comprises 25.69M parameters, which are strategically distributed across components. The base architecture contains 25.63M parameters, and our final dual-stream architecture with all other configurations introduces a modest 60K additional parameters. This represents a parameter efficiency of 0.23% overhead for the achieved performance improvements.

## Ablation studies

To validate the effectiveness of our proposed method, we conducted a series of ablation experiments. We removed the following components from the U-Net architecture: the attention mechanism, multiscale processing, the boundary-aware module, the encoder & dual encoder, and integrated them one by one. Their results are presented in numeric form in Table 1.

As can be observed from the results, the baseline UNet + Attention Mechanism, although achieving a modest performance, struggled with high validation loss (0.4303), low validation accuracy (63.80%), lowest P (62.34%), lowest R (61.40%), lowest F1 or F1-Score (61.01%), and low validation mIoU (61.81%), which limits its ability to process all features within context. The addition of multiscale processing in the U-Net + attention mechanism significantly improved the performance. Specifically, the validation loss was reduced to 0.3886, representing a 9.69% decrease compared to the baseline configuration. Besides the decrease in validation loss, the validation accuracy increased to 71.69%, precision improved to 69.91%, recall increased to 67.92%, the F1-score reached 68.89%, and the mIoU was elevated to 69.45%. The achieved scores of these metrics collectively indicate that the integration of multiscale processing not only reduced prediction errors but also enhanced the model's overall robustness capabilities by allowing the model to look at the feature at different scales. The integration of the PvT-V2 encoder in the U-Net + attention mechanism + multiprocessing further strengthened the model's performance.

The validation loss decreased by an additional 10.11%, reaching 0.3493, while the validation accuracy increased to 80.55%, precision improved to 77.20%, recall amplified to 81.85%, the F1 score improved to 78.15%, and the mIoU reached 78.03%. This enhancement underscores the importance of leveraging advanced transformer-based architectures in our change detection task, which made it easier for the model to strongly capture the global context and long-range dependencies. Introducing the dual-stream encoder in place of the single-stream PvT-V2 encoder in the U-Net + attention mechanism + multiprocessing marked another significant milestone in the ablation study. The dual-stream architecture allowed the model to process bitemporal images more effectively by separately extracting local and global features from each temporal input before fusing them. As a result, the validation loss was further reduced by 9.91%, dropping to 0.3147, while the validation accuracy rose to 90.51%, precision improved to 88.97%, recall increased to 86.26%, the F1 score reached 87.50%, and the mIoU climbed to 87.67%.

The final configuration, which incorporated the boundary-aware module alongside all previously mentioned components, achieved the best overall performance. The boundary-aware module played a crucial role in refining edges and fine details, which are critical for accurately detecting changes in remote sensing images. With this addition, the validation loss was decreased to its lowest value of 0.2997, while the validation accuracy peaked

| Metrics | | U-Net+AM | U-Net+ AM+ MP | U-Net+ AM+ MP+PE | U-Net+ AM+ MP+DSEnc | U-Net+AM+ MP+DS |
|---|---|---|---|---|---|---|
| A (%) | Train | 65.09 | 73.13 | 82.17 | 92.33 | 97.19 |
| | Val | 63.80 | 71.69 | 80.55 | 90.51 | 95.27 |
| Loss | Train | 0.3709 | 0.3341 | 0.3010 | 0.2712 | 0.2583 |
| | Val | 0.4303 | 0.3886 | 0.3493 | 0.3147 | 0.2997 |
| P (%) | Train | 63.80 | 72.88 | 81.21 | 92.38 | 96.54 |
| | Val | 62.34 | 69.91 | 77.20 | 88.97 | 93.86 |
| R (%) | Train | 61.40 | 69.14 | 83.39 | 87.99 | 92.37 |
| | Val | 59.83 | 67.92 | 81.85 | 86.26 | 91.28 |
| F1 (%) | Train | 62.24 | 70.12 | 78.97 | 89.24 | 93.05 |
| | Val | 61.01 | 68.89 | 78.15 | 87.50 | 92.51 |
| mIoU (%) | Train | 63.10 | 70.90 | 79.66 | 89.50 | 94.21 |
| | Val | 61.81 | 69.45 | 78.03 | 87.67 | 92.28 |

**Table 1**. Comprehensive Analysis of Training and Validation Metrics Across Different Model Configurations in the Ablation Study. The metrics include Loss, A (%), P (%), R (%), F1 (%), and mIoU (%) for both training and validation phases, demonstrating progressive improvement. Abbreviations: AT - Attention Mechanism; MP - Multiscale Processing; PvT-V2 Enc - Pyramid Vision Transformer V2 Encoder; DSEnc - Dual Stream Encoder; BAM - Boundary Aware Module

at 95.27%, precision reached 93.86%, recall improved to 91.28%, the F1 score increased to 92.51%, and the mIoU achieved an impressive 92.28%. These results demonstrate the effectiveness of the boundary-aware module in addressing challenges related to edge detection and segmentation accuracy, which are common limitations in baseline models.

On comparing the final configuration (our proposed) with the baseline model configuration, it showed remarkable improvement. Our proposed method achieved a 49.29% improvement in mIoU, a 50.41% increase in accuracy, a 51.29% increase in precision, a 52.62% improvement in recall, and a 51.63% improvement in F1-score.

These enhancements in the results suggest the effectiveness of our proposed model over the simple baseline model in better handling the global and local context in segmentations. Therefore, our suggested model has significant potential in smart cities for accurate and reliable change detection from bitemporal remote sensing images, which can be extremely helpful for urban planning policymakers to monitor land use changes and track infrastructure development and environmental impacts over time. Figure 7 presents the ablation study results across validation metrics results in graphical form. The graphical representation of the results of our proposed model also shows the significance and effectiveness in the change detection task by achieving the peak mIoU and overall accuracy, along with other metrics, and the loss significantly decreased to its minimalist level. The graphical results representation is divided into two sub-images Fig. 6 and Fig. 7 for the clarity of analysis.

After training and validation of the proposed DSHA model, we exported the model and tested it on the test dataset. We performed the testing with two configurations: one with a single stream PVT-V2 encoder, and the final configuration of U-Net with a dual stream module as a backbone coupled with a boundary aware module. Their qualitative results are displayed in Fig. 8.

By looking closely at the results in Fig. 9, the customized single stream encoder detected the changes from the given bitemporal images but lacks the fine details and failed to identify the exact boundaries of the buildings and roads. Therefore, it produced blurry, smooth-out segmentation masks, and in some areas, it overly predicted. In contrast, our proposed model with the boundary-aware module detected the objects' edges with fine details, resulting in detecting almost the same boundaries as in the ground truth. Our proposed model performed equally well in detecting small building objects as well as a series of building objects; also, road structures are perfectly detected. These results further validate that our proposed hybrid model is the optimal solution for detecting changes at multiscale, including the small and complex building structures and roads from the remote sensing bi-temporal images for smart cities.

## Comperative experiment

To further validate the robustness of the DSHA, we conducted the comparative experiments on CDD. For the experiments we used the baseline UNet + AM and final configuration of our proposed DSHA for testing on CDD with the same training parameters settings as before. The results of these experiments are presented in Table 2, which further contributes to the analysis of the performance metrics of the different models.

The experimental results demonstrate significant performance improvements across all evaluation metrics when comparing the baseline UNet+AM with our proposed DSHA architecture. Specifically, the DSHA model achieved an accuracy of 92.17%, representing a substantial improvement of 23.14 percentage points over the baseline UNet+AM (69.03%). Similarly, the precision metric also showed remarkable improvement, with DSHA achieving 91.78% compared to 68.72% for the baseline, indicating a 23.06 percentage point improvement. The recall performance displayed the higher substantial improvement, with DSHA achieving 87.83% compared 62.11% for UNet+AM, representing a significant improvement of 25.72 percentage points. Furthermore, the F1-score, which provides a balanced measure of precision and recall, also demonstrated substantial improvement

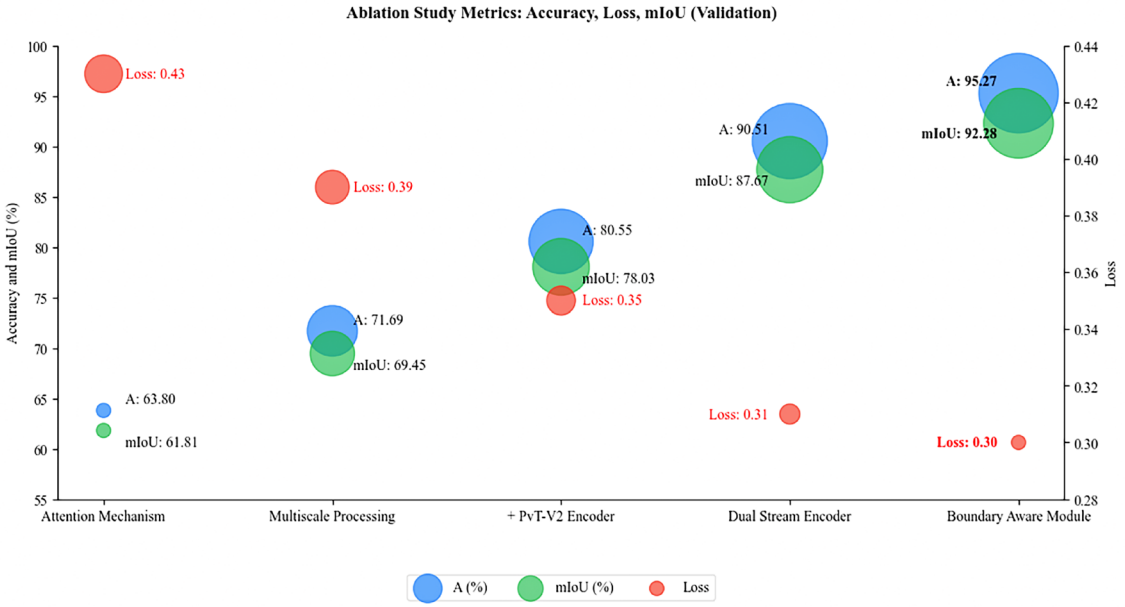


**Fig. 6**. Visual comparisons of the validation performance metrics of different model configurations in the ablation study. The left y-axisrepresents A and mIoU in percentage, while the right y-axis shows the loss values.

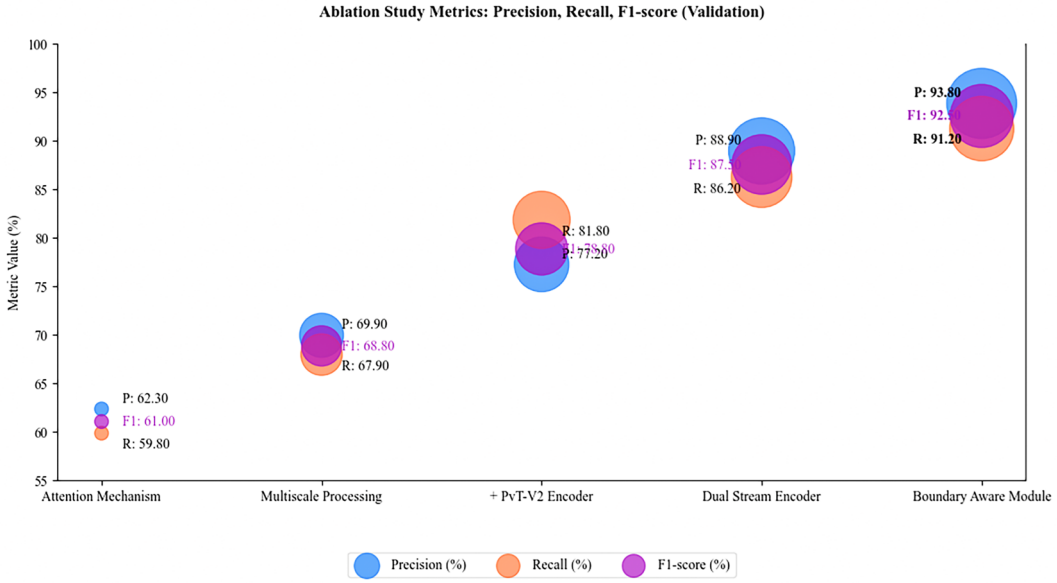


**Fig. 7**. Visual comparisons of the validation performance metrics of different model configurations in the ablation study. The validationresults for the three-evaluation metrics (P, R, and F1) are shown.

from 65.31% (UNet+AM) to 89.81% (DSHA), representing a performance enhancement of 24.50 percentage points. Likewise, the mIoU, which is considered the primary evaluation metric, demonstrated exceptional improvement from 65.24% to 89.97%, achieving a significant improvement of 24.73 percentage points. These comprehensive comparative results on the CDD dataset further validate the effectiveness and robustness of our proposed DSHA architecture, demonstrating its superior capability in handling diverse urban change detection scenarios across different datasets. The consistent performance improvements across all evaluation metrics highlight the architecture's ability to generalize effectively to various remote sensing imagery characteristics and change detection challenges, so, confirming its potential for practical deployment in smart city monitoring applications.

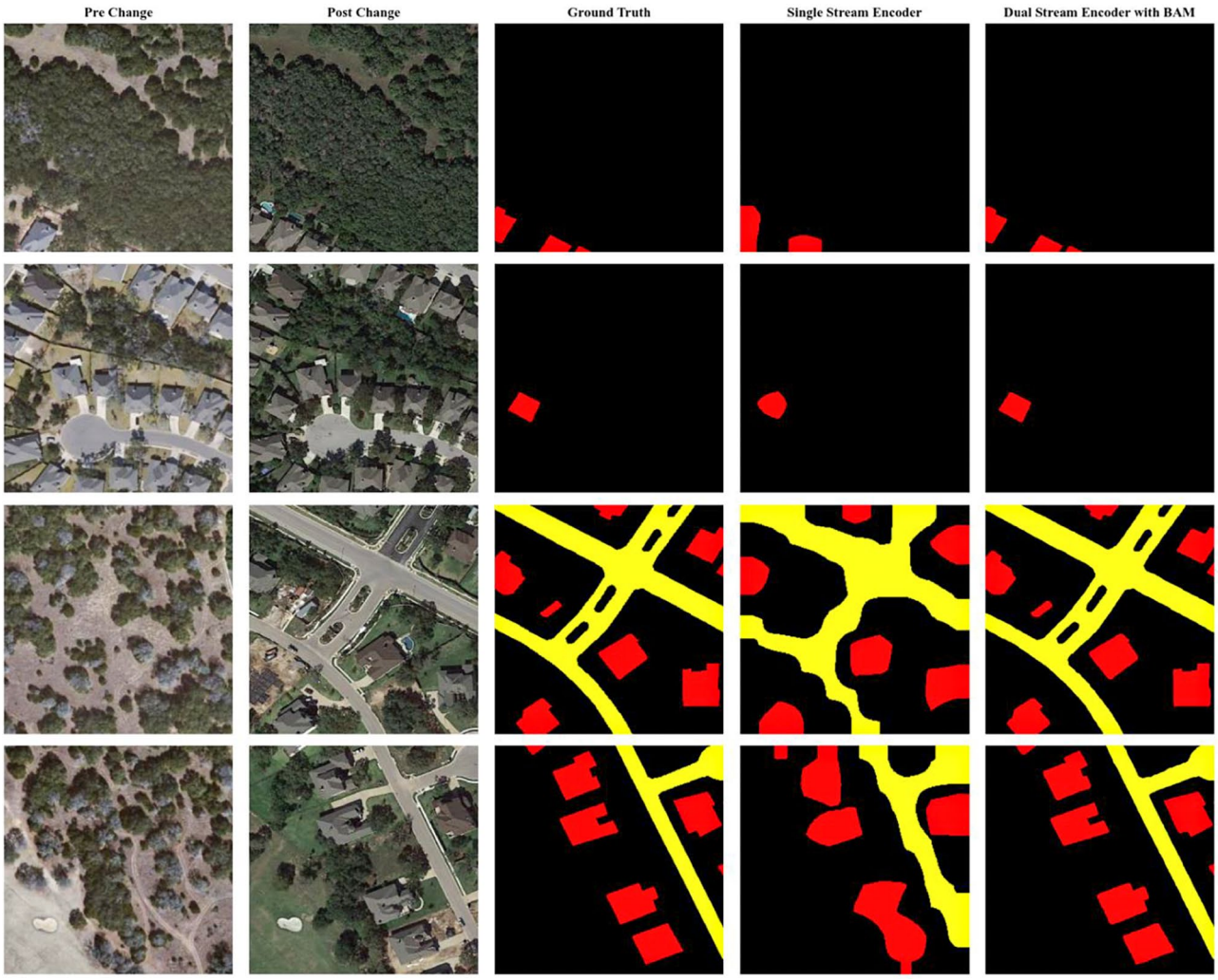


**Fig. 8**. Visual comparison of change detection results on the LEVIR-MCI[66]: Left side (column a) Pre-change images, (column b) Post-change images, (column c) Ground truth masks, (column d) Predictions using single-stream modified PVT-V2 encoder backbone, and right side (column e) Predictions using proposed DSHA model with dual-stream encoder. The first and second rows (columns d and e) represent building change predictions in red color masks, while rows 3 and 4 (columns d and e) show road changes in yellow and building change predictions.

**Fig. 9**. Closeup results for the comparison are presented. (left side mask) ground truth, (center mask) PvT-v2's detection, (right side mask) our proposed model's detection.

| Models | A(%) | P(%) | R(%) | F1-Score(%) | mIoU(%) |
|---|---|---|---|---|---|
| UNet+AM | 69.03 | 68.72 | 62.11 | 65.31 | 65.24 |
| DSHA | 92.17 | 91.78 | 87.83 | 89.81 | 89.97 |

**Table 2**. Performance comparison of baseline UNet+AM and the Proposed DSHA Model on CDD Dataset. The evlaution metrics include A(%), P(%), R(%), F1-Score(%) and mIoU(%).

| Method | Dataset | IoU(%) | mIoU(%) | F1-Score (%) |
|---|---|---|---|---|
| MaskChanger | LEVIR CD | 85.12 | | 91.96 |
| Mask Classification (MaskCD) | LEVIR CD | | 91.13 | 90.84 |
| ChangerEx | LEVIR CD | x | x | 91.77 |
| BIT | LEVIR CD | | 73.54 | 0.6683 |
| SRC-Net | LEVIR CD | 85.60 | | 92.24 |
| DMINet | LEVIR CD | | 83.57 | 80.57 |
| SiamixFormer5 | LEVIR CD | 85.38 | | 91.58 |
| SNUNet | LEVIR CD | | 87.07 | 86.08 |
| LightCDNet | LEVIR CD | 84.21 | | 91.43 |
| CrossCDNet | LEVIR CD | 84.65 | | 91.69 |
| RS-Mamba | LEVIR CD | 83.66 | | 91.1 |
| USSFC-Net | LEVIR CD | 83.55 | | 91.04 |
| Chage Former | LEVIR CD | 82.48 | | 90.40 |
| **Ours** | LEVIR CD | | **92.28** | **92.50** |

**Table 3**. Comparison of different methods on LEVIR CD Dataset. The table shows IoU(%), mIoU(%), and F1-Score(%) metrics. Best results are highlighted in bold.

## Comparison with SOTA

The proposed DSHA is compared with several recent state-of the-art deep learning-based remote sensing change detection methods (2023-2025); a Siamese network integrated MaskChanger[68], a cross-level change representation perceiver based MaskCD[69], Residual networks based ChangerEx[70], a bitemporal image transformer (BIT)[69], a patch-mode joint feature fusion model SRC-Net[71], a dual branch multi-level inter-temporal network (DMINet)[69], a two-transformer-based SiamixFormer[72], an integrated Siamese network and nested U-Net (SNUNet)[69], an early fusion and deep supervision based LightCDNet[73], a global and lightweight decoder based CrossCDNet[74], an omnidirectional selective scan based RS-Mamba[75], a U-Net based USSFC-Net[76], and transformer-based ChangerFormer[77].

In Table 3, we compared the performance of our proposed DSHA model with several state-of-the-art methods of deep learning-based remote sensing change detection methods on the LEVIR-MCI dataset. Maskchanger adopted the segmentation-specialized Mask2Former architecture by incorporating Siamese networks to extract features separately from bi-temporal images, while retaining the original mask transformer decode. Maskchanger used IoU evaluation to calculate the degree of overlap between the detected masks and the ground truth, and it achieved an IoU score of 85.12% and an F1 score of 91.96% on the LEVIR-CD dataset. Mask (MaskCD) reformulated change detection as a mask classification problem by using a Cross-Level Change Representation Perceiver (CLCRP) with deformable attention to generate change-aware representations, followed by a Masked Attention-based Detection Transformer (MA-DETR) decoder that predicts object masks and their corresponding change/no-change classifications instead of performing pixel-wise labeling. It achieved an mIoU score of 91.13% and an F1-score of 90.84%, showcasing its effectiveness in handling complex tasks. ChangerEx used parameter-free feature exchange operations (spatial exchange in early stages and channel exchange in later stages) between bi-temporal features during feature extraction, combined with Flow Dual-Alignment Fusion (FDAF) for interactive alignment and fusion to achieve effective change detection. ChangerEx demonstrated the high scores with 91.77% of F1-score, a recall of 90.61%, and precision of 92.97%, however, the IoU or mIoU score was not mentioned in the results. BIT uses a transformer-based approach that incorporates self-attention mechanisms to model long range dependencies with deep features but achieved relatively lower scores, with a mIoU of 73.54% and F1-score of 66.83%.

SRC-Net employed a Perception and Interaction Module with cross-branch perception mechanisms and a Patch-Mode joint Feature Fusion Module to leverage bi-temporal spatial relationships between features at the same location at different times for enhanced change detection and used the IoU for the evaluation. It achieved a relatively high IoU score of 85.60 and an F1-score of 92.24% compared to BIT on the LEVIR-CD dataset. A dual-branch multi-level inter-temporal network DMINet achieved an mIoU of 83.57% and F1-score of 80.57% on the LEVIR-CD dataset, which demonstrates a reasonably good performance in change detection. A two SegFormers used two parallel SegFormer encoders to extract hierarchical features from bi-temporal images, applies temporal transformers at each stage for cross-attention fusion (query from T1, key-value from T2), and

employs a lightweight MLP decoder to generate building detection or change detection maps, and achieved an IoU score of 85.38 and F1-score of 91.57% on LEVIR-CD dataset. SiamixFormer employed two SegFormers to independently extract features from bi-temporal images and achieved an IoU score of 85.38% on the LEVIR-CD dataset, particularly excelling in detecting building changes. SNUNet, an integrated Siamese network combined with a nested U-Net architecture, achieved an mIoU score of 87.07% and an F1-score of 86.08%. The score of SNUNet shows its ability to capture the features for change detection from bi-temporal remote sensing images. LightCDNet, used an early fusion backbone network with a Deep Supervised Fusion Module (DSFM) to guide the fusion of primary features from bi-temporal images, combined with a pyramid decoder for end-to-end lightweight change detection while preserving input information; achieved an IoU score of 84.21% and an F1-score of 91.43% on the LEVIR-CD dataset.

CrossCDNet employed a Siamese neural network with an Instance Normalization and Batch Normalization Module (IBNM) as the encoder backbone to extract and fuse bi-temporal feature maps, followed by a simple MLP decoder for cross-domain change detection with enhanced generalization capability. It used the IoU for measuring the degree of overlap between detection and ground truth. CrossCDNet achieved an IoU score of 84.65% and an F1-score of 91.96% on the LEVIR-CD dataset. The RS-Mamba incorporated an omnidirectional selective scan module to capture global context in multiple spatial directions with linear complexity, enabling efficient dense prediction on large VHR remote sensing images without the quadratic computational overhead of transformers. RS-Mamba achieved an IoU score of 83.66% and F1-score of 91.1% on the LEVIR-CD dataset. USSFC-Net used multi-scale decoupled convolution (MSDConv) for efficient multi-scale feature extraction and a spatial-spectral feature cooperation strategy (SSFC) that generates 3D attention weights without additional parameters to model spatial-spectral feature interactions for ultra-lightweight change detection, and achieved an IoU score of 83.55% and F1-score of 91.04% for remote sensing change detection task. Finally, ChangeFormer, used a hierarchical transformer encoder in a Siamese network to extract multi-scale features from bi-temporal images, computes feature differences through difference modules at multiple scales, and employs a lightweight MLP decoder to fuse these multi-level feature differences for change detection. It achieved an IoU score of 82.48% and an F1-score of 90.40% on the LEVIR-CD dataset.

Our proposed DSHA model incorporating a dual stream encoder with multi-scale attention and boundary aware module achieved a better mIoU score of 92.28% and a better F1-score of 92.50% and outperformed the currently existing state-of-the-art models on the LEVIR-CD dataset for change detection. This comparison shows the superior capability of DSHA in detecting changes at multiple scales, including the small and complex building structures as well as road structures, which makes it an optimal solution for remote sensing bi-temporal images' image analysis in smart cities. The superior performance demonstrates DSHA's practical viability for deployment in smart city monitoring networks, where accurate detection of small-scale urban changes is crucial for municipal asset management, building permit compliance verification, and infrastructure development oversight. This enhanced precision in multi-scale change detection fills a critical gap in operational urban monitoring systems, providing city administrators with reliable automated analysis tools necessary for maintaining comprehensive urban development records and supporting regulatory enforcement processes. This evidence-based support for policymakers in sustainable urban planning and controlling urban expansion and infrastructural changes directly contributes to the Sustainable Cities and Communities goal of the Sustainable Development Goals[4].

## Conclusion and disscussion

In this article, we have proposed the DSHA with adaptive multi-scale boundary-aware mechanisms for robust urban change detection in smart cities. Our approach synergistically combines the strengths of both CNN and transformer by employing ResNet34 and customized PVT-v2 in a dual-stream encoder as the backbone for a U-Net framework. This integration enables the simultaneous capture of both the global context and fine-grained details. The feature fusion layers fuse the global and fine-grained features. The integration of a boundary-aware module, along with multi-scale attention in the decoder, significantly enhances the model's ability to detect object boundaries and capture changes at various scales accurately. Furthermore, the combined loss function, which is a combination of four losses, also helps the model to adjust its weights for better detections. The experimental results of our proposed DSHA, ablation studies, and comparison with SOTA demonstrate a substantial improvement over the existing state-of-the-art methods on the change detection benchmark dataset. The proposed DSHA model achieved an mIoU score of 92.28, which is a primary evaluation metric in segmentation tasks that calculates the ratio of overlap between ground truth and detections. Moreover, the DSHA also showed its superior performance in other evaluation metrics, such as achieving the F1-score of 92.50, precision of 93.86, recall of 91.28, and accuracy of 95.27. The qualitative results also show the better detection capability of our proposed DSHA, including the small and complex building and road structures. These advancements in the results clearly demonstrate a significant advancement in urban change detection from bi-temporal remote sensing images for smart cities. Future research could explore extending to land use changes and other remote sensing applications and scenarios to enhance its robustness.

Beyond the technical achievements demonstrated, this research contributes significantly to the smart cities paradigm by providing urban planners and policymakers with a robust technological tool for evidence-based decision-making in sustainable urban development. The DSHA model's superior performance in detecting dense urban changes aligns with Sustainable Development Goal 11's targets for making cities inclusive, safe, resilient, and sustainable, while also supporting SDG 15 through accurate monitoring of land use changes and environmental impacts. The real-time monitoring capabilities enabled by this approach represent a crucial component of smart city infrastructure, facilitating data-driven governance and supporting the achievement of multiple Sustainable Development Goals through comprehensive urban change analysis. This work demonstrates how advanced remote sensing technologies can bridge the gap between technical innovation and practical urban

planning applications, providing the monitoring foundation necessary for sustainable smart city development and contributing to global efforts toward achieving the 2030 Sustainable Development Agenda.

## Data availability

The datasets used in this study are publicly available at: Agent, L. C. LEVIRMCl Dataset. https://huggingface.co datasetslcybuaalevir-mcitreemain/ (2025). Accessed:20250308, and https://drive.google.comfiled/1GX656JqqOyBi_Ef0w65kDGVtonHrNs9edit. Accessed: 20250613.

Received: 2 March 2025; Accepted: 13 August 2025
Published online: 21 August 2025

## References

1. Hadiyana, T. & Ji-hoon, S. Ai-driven urban planning: Enhancing efficiency and sustainability in smart cities. *ITEJ (Information Technology Engineering Journals)* **9**, 23–35 (2024).
2. Adreani, L., Bellini, P., Fanfani, M., Nesi, P. & Pantaleo, G. Smart city digital twin framework for real-time multi-data integration and wide public distribution. *IEEE Access* (2024).
3. Onoja, J. P. & Ajala, O. A. Smart city governance and digital platforms: A framework for inclusive community engagement and real-time decision-making. *GSC Adv. Res. Rev.* (2023).
4. Sharifi, A., Allam, Z., Bibri, S. E. & Khavarian-Garmsir, A. R. Smart cities and sustainable development goals (sdgs): A systematic literature review of co-benefits and trade-offs. *Cities* **146**, 104659 (2024).
5. Kellison, T. An overview of sustainable development goal 11. *The Routledge handbook of sport and sustainable development* 261–275 (2022).
6. Huang, Y., Wei, M., Ge, B., Zhang, Y. & Ji, Z. Change detection in dual-temporal remote sensing data based on a lightweight siamese network with effective preprocessing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, 8397–8400 (IEEE, 2024).
7. Javed, A., Kim, T., Lee, C. & Han, Y. Deep learning framework for semantic change detection in urban green spaces along with overall urban areas. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, 10039–10043 (IEEE, 2024).
8. Adiga, S., Parthasarthy, S., Vivek, R., Sarvade, A. & Natarajan, S. Mlafe-net-multi-layer attention and feature extraction for change detection. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 1–6 (IEEE, 2024).
9. Yu, W., Zhang, X., Das, S., Zhu, X. X. & Ghamisi, P. Maskcd: A remote sensing change detection network based on mask classification. *IEEE Trans. Geosci. Remote Sens.* (2024).
10. Jayarajan, K., Alzubaidi, L. H., Vasanthakumar, G., Lande, J. & Deiwakumari, K. Domain adaptive based convolutional neural network for change detection using time series satellite imagery. In *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)*, 1–4 (IEEE, 2024).
11. Gao, Y. et al. Relating cnn-transformer fusion network for remote sensing change detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6 (IEEE, 2024).
12. Wang, X., Guo, Z. & Feng, R. A cnn-and transformer-based dual-branch network for change detection with cross-layer feature fusion and edge constraints. *Remote Sens.* **16** (2024).
13. Saidi, S., Idbraim, S., Karmoude, Y., Masse, A. & Arbelo, M. Deep-learning for change detection using multi-modal fusion of remote sensing images: A review. *Remote Sens.* **16**, 3852 (2024).
14. Kiruluta, A., Lundy, E. & Lemos, A. Novel change detection framework in remote sensing imagery using diffusion models and structural similarity index (ssim). arXiv preprint arXiv:2408.10619 (2024).
15. Sghaier, M. O., Hadzagic, M., Yu, J. Y., Shton, S. & Shahbazian, E. Leveraging generative deep learning models for enhanced change detection in heterogeneous remote sensing data. In *2024 27th International Conference on Information Fusion (FUSION)*, 1–8 (IEEE, 2024).
16. Yu, S., Tao, C., Zhang, G., Xuan, Y. & Wang, X. Remote sensing image change detection based on deep learning: Multi-level feature cross-fusion with 3d-convolutional neural networks. *Appl. Sci.* **14**, (2076-3417) (2024).
17. Hafner, S., Fang, H., Azizpour, H. & Ban, Y. Continuous urban change detection from satellite image time series with temporal feature refinement and multi-task integration. arXiv preprint arXiv:2406.17458 (2024).
18. He, Q. et al. Ast: Adaptive self-supervised transformer for optical remote sensing representation. *ISPRS J. Photogramm. Remote Sens.* **200**, 41–54 (2023).
19. Deng, K. et al. Cross-modal change detection using historical land use maps and current remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **218**, 114–132 (2024).
20. Jiao, W., Persello, C. & Vosselman, G. Polyr-cnn: R-cnn for end-to-end polygonal building outline extraction. *ISPRS J. Photogramm. Remote Sens.* **218**, 33–43 (2024).
21. Wang, Y. et al. Msgfnet: Multi-scale gated fusion network for remote sensing image change detection. *Remote Sens.* **16**, 572 (2024).
22. Yu, B. et al. Multi-scale differential network for landslide extraction from remote sensing images with different scenarios. *Int. J. Digit. Earth* **17**, 2441920 (2024).
23. Wang, L., Zhang, M., Gao, X. & Shi, W. Advances and challenges in deep learning-based change detection for remote sensing images: A review through various learning paradigms. *Remote Sens.* **16**, 804 (2024).
24. Vincent, E., Ponce, J. & Aubry, M. Satellite image time series semantic change detection: Novel architecture and analysis of domain shift. arXiv preprint arXiv:2407.07616 (2024).
25. Zou, C. & Wang, Z. A semi-parallel cnn-transformer fusion network for semantic change detection. *Image Vis. Comput.* **149**, 105157 (2024).
26. Chen, M., Jiang, W. & Zhou, Y. Dtt-cginet: A dual temporal transformer network with multi-scale contour-guided graph interaction for change detection. *Remote Sens.* **16**, 844 (2024).
27. Zhang, M. & Shi, W. A feature difference convolutional neural network-based change detection method. *IEEE Trans. Geosci. Remote Sens.* **58**, 7232–7246 (2020).
28. Mou, L., Bruzzone, L. & Zhu, X. X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **57**, 924–935 (2018).
29. Fang, S., Li, K., Shao, J. & Li, Z. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters* **19**, 1–5 (2021).
30. Ren, W., Wang, Z., Xia, M. & Lin, H. Mfinet: Multi-scale feature interaction network for change detection of high-resolution remote sensing images. *Remote Sens.* **16**, 1269 (2024).
31. Feng, Y., Jiang, J., Xu, H. & Zheng, J. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15 (2023).
32. Fang, S., Li, K. & Li, Z. Changer: Feature interaction is what you need for change detection. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–11 (2023).

33. Wang, X. et al. Object-based change detection in urban areas from high spatial resolution images based on multiple features and ensemble learning. *Remote Sens.* **10**, 276 (2018).
34. Zhang, H. et al. A novel squeeze-and-excitation w-net for 2d and 3d building change detection with multi-source and multi-feature remote sensing data. *Remote Sens.* **13**, 440 (2021).
35. Yang, J., Wan, H. & Shang, Z. Enhanced hybrid cnn and transformer network for remote sensing image change detection. *Sci. Rep.* **15**, 10161 (2025).
36. Li, H. et al. Change detection network based on transformer and transfer learning. *IEEE Access* (2025).
37. Han, C., Wu, C., Guo, H., Hu, M. & Chen, H. Hanet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **16**, 3867–3878 (2023).
38. Li, Z. et al. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–12 (2023).
39. Huang, Z., Li, W., Xia, X.-G., Wang, H. & Tao, R. Task-wise sampling convolutions for arbitrary-oriented object detection in aerial images. *IEEE Trans. Neural Netw. Learn. Syst.* (2024).
40. Chen, H., Qi, Z. & Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2021).
41. Bandara, W. G. C. & Patel, V. M. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 207–210 (IEEE, 2022).
42. Liu, W., Lin, Y., Liu, W., Yu, Y. & Li, J. An attention-based multiscale transformer network for remote sensing image change detection. *ISPRS J. Photogramm. Remote Sens.* **202**, 599–609 (2023).
43. Ramsey, N. Some dynamics in real quadratic fields with applications to inhomogeneous minima. arXiv preprint arXiv:2206.12345 (2022).
44. Kirillov, A., Girshick, R., He, K. & Dollár, P. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6399–6408 (2019).
45. Yin, M., Chen, Z. & Zhang, C. A cnn-transformer network combining cbam for change detection in high-resolution remote sensing images. *Remote Sens.* **15**, 2406 (2023).
46. Peng, X., Zhong, R., Li, Z. & Li, Q. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **59**, 7296–7307 (2020).
47. Eftekhari, A., Samadzadegan, F. & Javan, F. D. Building change detection using the parallel spatial-channel attention block and edge-guided deep network. *Int. J. Appl. Earth Obs. Geoinf.* **117**, 103180 (2023).
48. Feng, Y., Xu, H., Jiang, J., Liu, H. & Zheng, J. Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–13 (2022).
49. Jiang, S. et al. Mdanet: A high-resolution city change detection network based on difference and attention mechanisms under multi-scale feature fusion. *Remote Sens.* **16**, 1387 (2024).
50. Zhan, Z. et al. Amfnet: Attention-guided multi-scale fusion network for bi-temporal change detection in remote sensing images. *Remote Sens.* **16**, 1765 (2024).
51. Li, Y., Weng, L., Xia, M., Hu, K. & Lin, H. Multi-scale fusion siamese network based on three-branch attention mechanism for high-resolution remote sensing image change detection. *Remote Sens.* **16**, 1665 (2024).
52. Farooque, G., Xiao, L., Sargano, A. B., Abid, F. & Hadi, F. A dual attention driven multiscale-multilevel feature fusion approach for hyperspectral image classification. *Int. J. Remote Sens.* **44**, 1151–1178 (2023).
53. Sun, L., Wang, X., Zheng, Y., Wu, Z. & Fu, L. Multiscale 3-d-2-d mixed cnn and lightweight attention-free transformer for hyperspectral and lidar classification. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–16 (2024).
54. Guo, Z., Chen, H. & He, F. Msfnet: Multi-scale spatial-frequency feature fusion network for remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* (2024).
55. Wang, A., Cai, J., Lu, J. & Cham, T.-J. Modality and component aware feature fusion for rgb-d scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5995–6004 (2016).
56. Sun, K., Xiao, B., Liu, D. & Wang, J. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703 (2019).
57. Lin, T.-Y., RoyChowdhury, A. & Maji, S. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, 1449–1457 (2015).
58. Liu, R., Ling, J. & Zhang, H. Softformer: Sar-optical fusion transformer for urban land use and land cover classification. *ISPRS J. Photogramm. Remote Sens.* **218**, 277–293 (2024).
59. Bandara, W. G. C. & Patel, V. M. Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1767–1777 (2022).
60. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y. & Barnard, K. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3560–3569 (2021).
61. Li, X. et al. Gated fully fusion for semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* **34**, 11418–11425 (2020).
62. Zhou, H. et al. Canet: Co-attention network for rgb-d semantic segmentation. *Pattern Recognit.* **124**, 108468 (2022).
63. Dai, J. et al. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773 (2017).
64. Huang, Z. et al. Alignseg: Feature-aligned segmentation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 550–557 (2021).
65. Li, X. et al. Semantic flow for fast and accurate scene parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 775–793 (Springer, 2020).
66. Liu, C. et al. Change-agent: Toward interactive comprehensive remote sensing change interpretation and analysis. *IEEE Trans. Geosci. Remote Sens.* 1–1, https://doi.org/10.1109/TGRS.2024.3425815 (2024).
67. Lebedev, M., Vizilter, Y. V., Vygolov, O., Knyaz, V. A. & Rubis, A. Y. Change detection in remote sensing images using conditional adversarial networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. S* **42**, 565–571 (2018).
68. Ebrahimzadeh, M. & Manzuri, M. T. Maskchanger: A transformer-based model tailoring change detection with mask classification. In *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*, 1–6 (IEEE, 2024).
69. Yu, W., Zhang, X., Das, S., Zhu, X. X. & Ghamisi, P. Maskcd: A remote sensing change detection network based on mask classification. *IEEE Trans. Geosci. Remote Sens.* (2024).
70. Fang, S., Li, K. & Li, Z. Changer: Feature interaction is what you need for change detection. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–11 (2023).
71. Chen, H., Xu, X. & Pu, F. Src-net: Bi-temporal spatial relationship concerned network for change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* (2024).
72. Mohammadian, A. & Ghaderi, F. Siamixformer: A fully-transformer siamese network with temporal fusion for accurate building detection and change detection in bi-temporal remote sensing images. *Int. J. Remote Sens.* **44**, 3660–3678 (2023).
73. Xing, Y. et al. Lightcdnet: Lightweight change detection network based on vhr images. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023).
74. Song, Y. et al. A cross-domain change detection network based on instance normalization. *Remote Sens.* **15**, 5785 (2023).
75. Zhao, S. et al. Rs-mamba for large remote sensing image dense prediction. *IEEE Trans. Geosci. Remote Sens.* (2024).
76. Lei, T. et al. Ultralightweight spatial-spectral feature cooperation network for change detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–14 (2023).

77. Bandara, W. G. C. & Patel, V. M. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 207–210 (IEEE, 2022).

## Acknowledgements
This work was supported by the Gwangju Institute of Science and Technology (GIST) research fund (Future-leading Specialized Research Project, 2025).

## Author contributions
I.A. conceived and designed the study, established the experiments of the proposed model, and wrote the initial manuscript draft. F.S. supervised the experiments, formally analyzed the results, and guided the preparation of the figures and tables. A.W. proofread the manuscript and assisted with the conversion to LaTeX format. M.S.P also supervised the manuscript, reviewed it, provided suggestions repeatedly until the final version. Y.S.K (CA) Directed and supervised the additional experiments specifically requested by the reviewers, providing essential resources and technical expertise crucial for addressing the review comments. Secured funding for the research. And Participated equally in the revision of the final manuscript. All authors participated in the revision of the final manuscript.

## Declarations

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to Y.-S.K.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.