A Novel Dual-Attention Network for Change Detection in High-resolution Remote Sensing Images

Afsaneh Talebizadeh Sardari [1]*, Saeid Niazmardi[1], Tayeb Alipour Fard[1]

[1] Department of Surveying Engineering, Faculty of Civil and Surveying Engineering, Graduate University of Advanced Technology, Kerman, Iran – (a.talebizadeh@student.kgut.ac.ir, (s.niazmardi, t.alipour)@kgut.ac.ir))

KEY WORDS: Change detection, high-resolution remote sensing, convolutional neural networks, dual attention network.

ABSTRACT:

Change detection in high-resolution remote sensing imagery is essential for a wide range of applications, including urban sprawl monitoring, conducting environmental assessments, and responding to disasters. Despite significant advancements in deep learning, challenges remain in capturing subtle changes, preserving spatial detail, and minimizing false detections. This work proposes a novel change detection framework, in which a high-resolution feature extraction backbone is integrated with dual attention mechanisms and multiscale contextual aggregation. In particular, HRNet serves as the backbone to obtain an informative representation of high-resolution images, while both channel-wise and spatial attention modules are incorporated to enhance the representation's discriminative capability. A residual change decoder jointly encodes absolute feature differences and semantic content, while a pyramid pooling module captures contextual dependencies across multiple scales. Finally, a lightweight refinement block is introduced to improve boundary sharpness and reduce noise. Extensive experiments on the LEVIR-CD dataset demonstrate that the proposed method achieves superior performance compared to state-of-the-art baselines, with improvements observed across major evaluation metrics. The obtained accuracy of the proposed model (F1-score of 89.39% and IoU of 80.82%) substantiates the robustness and effectiveness of the proposed architecture for reliable change detection in high-resolution remote sensing imagery.

1. INTRODUCTION

One of the primary objectives of remote sensing is to monitor land surface dynamics. It is possible to systematically track global environmental and landscape alterations through satellite and aerial imagery, encompassing densely inhabited regions and isolated, inaccessible locations. This capacity underscores why change detection (CD) has emerged as a critical area of research within remote sensing studies (Coppin et al. 2004). Change detection involves identifying areas within remote sensing imagery that have changed over specific observation intervals (Bovolo and Bruzzone 2015). The precise meaning of change often varies according to the particular context or application. Relevant changes of interest typically include alterations in man-made structures (e.g., buildings, roads), shifts in vegetation, and environmental transformations such as polar ice cap melting, deforestation, and damage resulting from natural disasters. Accordingly, change detection has found various applications across numerous fields, including fire detection (Giglio et al., 2016), environmental monitoring (Liu, Kuffer, and Persello 2019), disaster assessment (Voigt et al. 2007), urban dynamics analysis (Huang, Cao, and Li 2020), and land management (Jin et al. 2013). An effective CD model must be capable of accurately identifying these changes while minimizing the detection of irrelevant variations caused by seasonal cycles, building shadows, atmospheric disturbances, and fluctuations in illumination conditions (Bandara and Patel 2022).

Many traditional change detection approaches rely on manual feature extraction techniques, such as principal component analysis (PCA) (Celik, 2009), Gabor filters (Z Li et al. 2017), multivariate alteration detection (MAD) (Nielsen et al., 1998), and change vector analysis (CVA) (Bovolo 2006). While these methods can achieve a certain level of accuracy, they suffer from several significant limitations (Bandara & Patel, 2022). Features extracted by traditional methods are often sensitive to seasonal changes, lighting variations, and the characteristics of different satellite sensors, rendering them less robust for achieving high CD accuracy. Although these strategies attempt to mitigate false detections by integrating shape and texture features, these techniques typically involve intensive computational demands and tuning numerous hyperparameters, leading to reduced robustness and increased computational cost. Moreover, the reliance on manually engineered features necessitates substantial prior domain knowledge, ultimately constraining the generalization ability and broader applicability of these traditional CD methods.

In recent years, deep learning-based change detection methods for remote sensing imagery have gained widespread attention, due to substantial improvements in accuracy and robustness over traditional approaches (Zhai et al., 2021). Due to their strong discriminative capabilities, deep convolutional neural networks (CNNs) have been successfully applied to remote sensing image analyses and demonstrated impressive performance in CD tasks (Shi et al., 2020). Most contemporary supervised CD methods employ CNN-based architectures to extract high-level features from each bi-temporal image pair, effectively capturing the interest changes (J. Chen et al., 2020). The evolution of CNN-based deep learning models for CD began with the introduction of the U-Net (O Ronneberger et al. 2015), which established a benchmark model for change detection. Building upon this foundation, Siamese network architectures emerged as a dominant paradigm in the field (Daudt et al., 2018). These models process pre- and post-change images through twin branches with shared weights, allowing for direct comparison of their learned feature representations. For instance, in (Zhang et al., 2020), a Siamese CNN that extracts features from bi-temporal

* Corresponding author

image pairs was proposed, followed by a specialized decoder network that fuses these features to capture change information more effectively. Similarly, in (Zhang et al., 2018), a deep CNN backbone (ResNet101) was employed to extract image features, and at the same time, dilated convolutions were utilized to expand the model's receptive field, allowing for more effective capture of contextual information across different spatial scales.

In parallel with these CNN-based advancements, transformer-based architectures have recently been introduced into remote sensing change detection, bringing substantial modeling flexibility and increased accuracy. For example, ChangeFormer introduces a multi-scale vision transformer framework that fuses features across temporal dimensions using cross-attention, achieving superior performance on high-resolution datasets (W. Bandara et al., 2022). Similarly, BITNet (H. Chen et al., 2021) proposes a bi-temporal image transformer that captures long-range dependencies and contextual relationships across both time stamps and spatial locations, demonstrating the capacity of transformers to model subtle and large-scale changes. Other models, such as SNUNet-Transformer (Fang et al., 2021), integrate convolutional backbones with transformer layers to harness both local spatial detail and global semantics.

The advent of high-resolution sensors posed CD technology to new challenges. It has also heightened the demand for more sophisticated and intelligent methods capable of effectively analyzing these increasingly detailed changes (W. Liu et al., 2024). As noted in (Volpi et al., 2013), change detection in low-resolution and very high-resolution (VHR) imagery presents distinct challenges. In low-resolution images, individual pixels often encompass information from multiple objects within their spatial extent, meaning a single pixel may simultaneously represent changed and unchanged surfaces between image pairs. In contrast, VHR images are more vulnerable to issues such as parallax effects, significant reflectance variability among objects of the same class, and co-registration inaccuracies, all of which complicate the reliable detection of changes.

Numerous researchers have tried to address the challenges of high-resolution change detection. Given redundant spatial information in high-resolution (HR) remote sensing images, multiscale features are frequently leveraged to suppress spatial noise (Hou et al., 2021). The attention mechanism has also proven effective in enhancing the detection of semantic changes. For instance, in (Li et al., 2022), channel attention is applied to refine multiscale features, thereby improving the model's ability to focus on the most relevant information. Furthermore, context aggregation through methods such as pyramid pooling (Zhao et al., 2017) and atrous spatial pyramid pooling (ASPP) (L. Chen et al., 2017) has proven highly effective in capturing multi-scale contextual information, a critical capability for detecting changes that manifest at varying spatial resolutions and semantic levels. In remote sensing, where the size, shape, and texture of changed objects can differ dramatically (e.g., small vehicles versus large buildings or agricultural plots), integrating both global context and local detail is indispensable for robust change interpretation.

Despite these advancements, several challenges remain unresolved in change detection. One of the most significant issues in prior deep learning-based models is their limited ability to preserve high-resolution spatial features, which are critical for accurately localizing changes in VHR remote sensing images. The reliance on encoder-decoder structures frequently leads to a loss of spatial detail due to repeated downsampling operations, which is detrimental in high-resolution imagery where precise boundary localization is essential. While attention mechanisms and context modules like ASPP or pyramid pooling offer improvements, their integration is often limited or not sufficiently optimized for the unique requirements of high-resolution remote sensing imagery.

To address these limitations, a novel deep learning architecture for change detection is proposed, in which multi-scale contextual aggregation, dual attention mechanisms, and a high-resolution feature extraction backbone are integrated into a unified, end-to-end framework. The proposed model adopts high-resolution network (HRNet) as the backbone to preserve high-resolution representations throughout the network, mitigating the loss of spatial precision commonly associated with downsampling heavy architectures (Sun et al. 2019). To enhance the discriminative power of feature representations, both channel-wise and spatial attention mechanisms are incorporated, allowing informative features to be emphasized. At the same time, the network suppresses irrelevant or redundant signals. Furthermore, a residual change decoder is designed to explicitly model the absolute difference between bi-temporal features while simultaneously considering their semantic content. A pyramid pooling module is embedded within this decoder to capture rich contextual information at multiple spatial scales, thereby improving the model's ability to detect coarse and subtle changes. To refine the coarse change predictions, a lightweight RefineBlock is introduced to progressively enhance the final change maps' spatial consistency and edge clarity. One of the most important factors in the success of deep learning-based algorithms is finding configurations or conditions that fully align with the target problem. Therefore, in this study, based on various techniques and sub-techniques, we proposed a framework whose key features that distinguish this research from others are outlined as follows:

1. A creative solution change detection network is introduced, in which a high-resolution feature extraction backbone (HRNet) is combined with channel and spatial attention mechanisms (Woo et al. 2018), enabling fine spatial details to be maintained while informative regions are selectively emphasized.
2. A residual change decoder is designed to jointly leverage absolute feature differences and semantic information from both input images, allowing subtle and contextually rich change patterns to be effectively captured.
3. A pyramid pooling module is added in the decoder to improve multi-scale contextual awareness, allowing for the detection of changes across a broad range of spatial resolutions and object scales.
4. A lightweight refinement module is developed to progressively sharpen the change map, thereby improving boundary localization and reducing false positives.

This letter is organized as follows: Section 2 details the proposed change detection methodology, including the overall network architecture and key components. Section 3 presents a comprehensive evaluation through quantitative experiments and a comparative analysis with state-of-the-art methods. Section 4 discusses the results in depth, highlighting the strengths of the proposed approach and analysing the observed performance patterns. Finally, Section 5 concludes the letter by summarizing the main findings and outlining potential directions for future research.

2. METHODOLOGY

2.1 Overall Architecture

This section presents the architecture and components of the proposed deep learning framework for change detection in high-resolution remote sensing imagery. The model, shown in Figure 1, integrates multi-resolution feature extraction, dual attention mechanisms, contextual aggregation, and a residual refinement

strategy into an end-to-end trainable network. The framework is designed to detect both coarse and subtle changes between bi-temporal satellite images while preserving spatial detail and semantic coherence.

Let $I_1, I_2 \in \mathbb{R}^{C\times H\times W}$ represent the pre-change and post-change RGB image pair, respectively, where C is the number of channels, H is the height, and W is the width of the input image. The goal of the model is to learn a function $\mathcal{F}: (I_1, I_2) \mapsto \hat{Y}$, where $\hat{Y} \in \mathbb{R}^{1\times H\times W}$ that predicts a binary change map.

2.2 High-resolution Feature Extraction

The first component of the proposed method is a high-resolution feature extraction module, presented in Figure 2. To retain fine spatial details and semantic consistency, the HRNet-W48 architecture (Sun et al. 2019), pre-trained on ImageNet (Deng et al. 2009), is utilized as the backbone for feature extraction. Its multi-resolution parallel branches and cross-scale information fusion make it appropriate for remote sensing applications requiring accurate localization of subtle changes. Feature maps are obtained at five resolution levels from the images I_1 and I_2. These features are denoted as $\{F_0, F_1, F_2, F_3, F_4\}$, where F_0 corresponds to the highest spatial resolution (i.e., closest to the input) and F_4 the lowest. This clear hierarchy allows for effective differentiation between various levels of resolution.

To enable effective multi-scale fusion, all feature maps are bilinearly upsampled to match the spatial size of F_1 (the second-highest resolution), balancing spatial detail preservation with computational efficiency. Then, a lightweight convolutional module, denoted as FuseF0, is applied to the upsampled F_0 to enhance low-level semantics, which contain subtle details such as edges, textures, and structural boundaries that are crucial for precise change localization in high-resolution remote sensing imagery:

$$F_0' = \mathrm{SE}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}(F_0)\right)\right)\right) \quad (1)$$

where Conv = a 2D convolution layer
BN = batch normalization
ReLU = the rectified linear unit
SE = the Squeeze-and-Excitation module

The Squeeze-and-Excitation module (SE) (Hu et al. 2018) applies channel-wise attention to recalibrate the feature response. In SE, given an input tensor $X \in \mathbb{R}^{B\times C\times H\times W}$, where B is the batch size, the channel descriptor $z_c \in \mathbb{R}^C$ is computed using global average pooling:

$$z_c = \frac{1}{H\cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{c,i,j} \quad (2)$$

where H = the height of the input image
W = the width of the input image
z_c = the sigmoid function

This descriptor is passed through two fully connected (FC) layers with a ReLU and a sigmoid activation to produce channel attention weights $s \in \mathbb{R}^C$, which are used to recalibrate the input:

$$X' = X\cdot\sigma\left(W_2\delta(W_1 z_c)\right) \quad (3)$$

where W_1 and W_2 = the weights of the FC layers
$\delta(.)$ = the ReLU activation
$\sigma(.)$ = the sigmoid function
X' = the recalibrated feature map

Finally, all five feature maps are concatenated along the channel dimension to form a unified high-dimensional feature tensor:

$$F = \mathrm{Concat}(F_0', F_1, F_2, F_3, F_4) \quad (4)$$

where Concat = concatenation along the channel dimension
F_0' = the recalibrated highest-resolution feature map

2.3 Channel Reduction and Difference Encoding

The high-dimensional feature map F is projected into a lower-dimensional space via a 1×1 convolution:

$$F_{redu} = \mathrm{Conv}_{1\times1}(F) \quad (5)$$

where $\mathrm{Conv}_{1\times1}$ = 1×1 convolution
F = the high-dimensional feature map
F_{redu} = the feature map after channel reduction

Given the reduced features $F_{redu}^{(1)}$ and $F_{redu}^{(2)}$ for images I_1 and I_2, a change embedding tensor F_Δ is computed as:

$$F_\Delta = \mathrm{Concat}\left(\left|F_{redu}^{(1)} - F_{redu}^{(2)}\right|, F_{redu}^{(1)}, F_{redu}^{(2)}\right) \quad (6)$$

The element-wise absolute difference and channel-wise concatenation of these tensors highlight change areas by emphasizing discrepancies between the two temporal features.

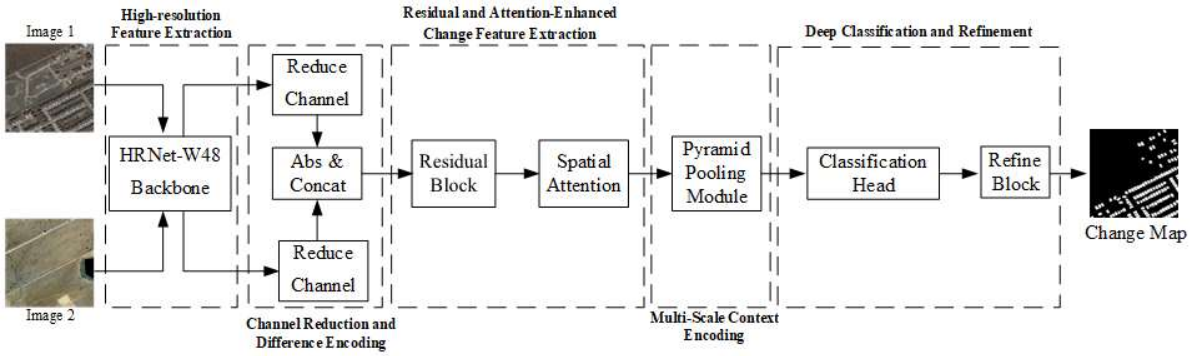


Figure 1. Overall architecture of the proposed model

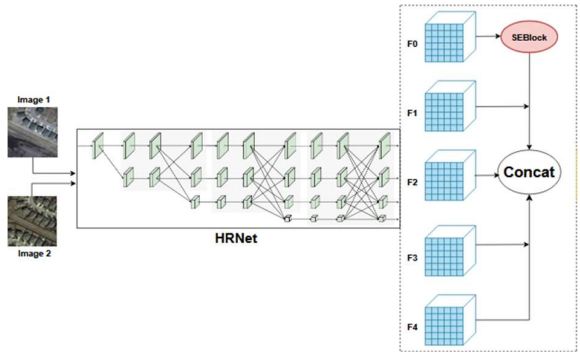


Figure 2. High-resolution feature extraction

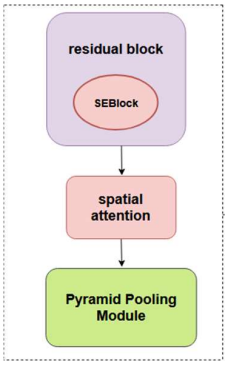


Figure 3. Feature enhancement

2.4 Residual and Attention-Enhanced Change Feature Extraction

To extract deep semantic differences, F_{Δ} is processed through a residual network composed of six SE-enhanced residual blocks:

$$F_{res} = ResCD(F_{\Delta}) \quad (7)$$

Each residual block R operates as:

$$(x) = ReLU\left(x + SE\left(B_2\left(Conv_2\left(ReLU(B_1(Conv_1(x)))\right)\right)\right)\right) \quad (8)$$

where $Conv_1$ and $Conv_2$ = two 3×3 convolutional layers
B_1 and B_2 = batch normalization layers

Spatial attention is subsequently applied to emphasize salient spatial patterns:

$$F_{att} = F_{res} \cdot \sigma\left(Conv_7(Concat[Avg(F_{res}), Max(F_{res})])\right) \quad (9)$$

where $Avg(F_{res})$ = Channel-wise average pooling of F_{res}
$Max(F_{res})$ = Channel-wise max pooling of F_{res}
$Conv_7$ = a 7×7 convolution

2.5 Multi-Scale Context Encoding

A Pyramid Pooling Module (Zhao et al. 2017) is employed to incorporate contextual information at multiple scales. For a set of pooling scales $S = \{1,2,3,6\}$, feature maps are pooled, convolved, and upsampled to match the input dimensions:

$$F_{ppm} = Concat\left(F_{att}, \left\{Upsample\left(Conv_1\left(Pool_s(F_{att})\right)\right) \mid s \in S\right\}\right) \quad (10)$$

where $Pool_s$ = an adaptive average pooling operation
$Upsample(.)$ = upsampling

The adaptive average pooling operation reduces the input feature map to a fixed spatial size of $s \times s$ for each $s \in S$, and each convolved feature map is upsampled using bilinear interpolation to match the original spatial dimensions. This module is presented in Figure 3.

2.6 Deep Classification and Refinement

The concatenated multi-scale features are processed through a classifier comprising five convolutional layers interleaved with batch normalization, ReLU activation, and dropout. This results in an intermediate change map (C_{raw}):

$$C_{raw} = ClassifierCD\left(F_{ppm}\right) \quad (11)$$

A dedicated RefineBlock, which consists of five iterative residual convolutional sub-blocks, is applied to further refine the spatial precision of change boundaries:

$$C_{refined}^{(i+1)} = ReLU\left(C_{refined}^{(i)} + Conv_3\left(C_{refined}^{(i)}\right)\right) \quad (12)$$

where $Conv_3$ = 3×3 convolution

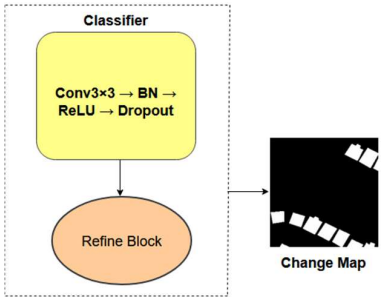


Figure 4. Prediction head

The final refined change map C_{final} is produced by:

$$C_{final} = Conv_{1\times1}\left(C_{r\,refined}^{(5)}\right) \quad (13)$$

The result is then upsampled to the original input size ($H \times W$) using bilinear interpolation. This module is presented in Figure 4.

$$\hat{C} = Upsample(C_{final}) \quad (14)$$

2.7 Loss Function

The proposed network is trained in a supervised manner using the binary cross-entropy loss with logits, which is ideal for binary pixel-wise classification tasks such as change detection. This loss function is defined as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log\left(\sigma(\hat{y}_i)\right) + (1-y_i) \cdot \log\left(1-\sigma(\hat{y}_i)\right)\right] \quad (15)$$

where $\hat{y}_i$ = the predicted logit for the i-th pixel
N = the total number of pixels in the batch
$y_i \in \{0,1\}$ = the corresponding ground truth label

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental Design

3.1.1 Datasets: The proposed model is evaluated using the LEVIR-CD dataset (Chen and Shi 2020), a publicly available large-scale dataset curated for building change detection. LEVIR-CD comprises 637 pairs of high-resolution (0.5-meter) remote sensing images, each with a spatial dimension of 1024 × 1024 pixels. The image pairs were collected from Google Earth over time intervals ranging from 5 to 14 years, enabling the study of long-term urban structural changes. The official dataset split is adopted, with 445 image pairs used for training, 64 image pairs for validation, and 129 image pairs for testing. All images are normalized using the standard ImageNet mean and standard deviation and converted to tensors for model input.

3.1.2 Implementation Details: The proposed model is implemented in the PyTorch deep learning framework and trained on a workstation equipped with a Quadro RTX 8000 GPU. Across all experiments, the batch size is fixed at 4. To optimize the model parameters, the AdamW optimizer is adopted, in which the adaptive learning rate strategy of Adam is integrated with decoupled weight decay for improved generalization (Loshchilov and Hutter 2017). The initial learning rate is set to 1×10^{-3}, with a weight decay of 1×10^{-4} to mitigate overfitting. To promote stable convergence, a cosine annealing learning rate scheduler is employed, gradually reducing the learning rate over training epochs to allow for finer updates and escape from local minima. To enhance training efficiency and reduce GPU memory usage without compromising numerical precision, the model is trained using automatic mixed precision (AMP), facilitated by PyTorch's GradScaler. Training on the LEVIR-CD dataset is performed for 25 epochs, with model performance monitored on a held-out validation set. An early stopping mechanism is applied with a patience threshold of 5 epochs, halting training if validation performance ceases to improve. The model checkpoint corresponding to the lowest validation loss is retained and saved for subsequent evaluation and inference.

3.1.3 Comparative Experiments: To evaluate the effectiveness of the proposed method, a comparative study was conducted against several state-of-the-art change detection approaches, including both convolutional-based and attention-enhanced architectures. Specifically, three purely convolutional methods: Fully Convolutional Early Fusion (FC-EF), Fully Convolutional Siamese-Difference (FC-Siam-Di), and Fully Convolutional Siamese-Concatenation (FC-Siam-Conc) (Daudt et al. 2018), as well as two attention-based methods: Dual-Task Constrained Deep Siamese Convolutional Network (DTCDSCN) (Liu et al. 2020) and SNUNet-CD (Fang et al. 2021).
All baseline models are implemented using their publicly available code repositories and trained with default hyperparameters provided by the respective authors to ensure a consistent and fair comparison.

3.1.4 Evaluation Metrics: In the proposed model experiments, the performance of the proposed model is evaluated using five standard evaluation metrics: precision (Pre), recall (Rec), F1-score (F1), overall accuracy (OA), and Intersection over Union (IoU). These metrics collectively provide a comprehensive evaluation of classification accuracy, spatial agreement, and consistency between predictions and ground truth (Powers 2020). These metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{19}$$

where FP = the number of false positives
TN = the number of true negatives
FN = the number of false negatives

3.2 Results

The experimental results, presented in Table 1, demonstrate the superior performance of the proposed method compared to several established baselines, including both convolution-based and attention-enhanced change detection models. The proposed model achieves the highest scores considering all evaluation metrics, indicating a balanced and robust detection capability.
The improvement in precision suggests that the proposed model is particularly effective at suppressing false positives, likely due to the integration of both channel-wise (SEBlock) and spatial attention mechanisms, which help the network focus on semantically meaningful regions and filter out irrelevant or noisy features. At the same time, the model achieves high recall, demonstrating its sensitivity to subtle and small-scale changes that are often overlooked by shallower or more coarsely fused architectures. This balance is reflected in the F1-score, where the proposed model surpasses all others by 1% compared to the second-best performer (SNUNet-CD).

Model	Evaluation metric				
	Pre.	**Rec.**	**F1.**	**IoU.**	**OA.**
FC-EF	86.91	80.17	83.40	71.53	98.39
FC-Siam-Di	89.53	83.31	86.31	75.92	98.67
FC-Siam-Conc	91.99	76.77	83.69	71.96	98.49
DTCDSCN	88.53	**86.83**	87.67	78.05	98.77
SNUNet-CD	89.18	87.17	88.16	78.83	98.82
Proposed Model	**92.46**	86.52	**89.39**	**80.82**	**99.14**

Table 1. The quantitative comparison results on the LEVIR-CD test set. All evaluation metrics are reported as percentages (%), and the highest score for each metric is highlighted in bold.

A notable improvement is observed in the IoU metric, which directly measures the spatial alignment between predicted and ground truth change regions. The proposed model achieves an IoU of 80.82%, substantially improving over the best-performing baseline, SNUNet-CD, with an IOU of 78.83%. This gain can be attributed to the high-resolution HRNet backbone, which preserves spatial detail throughout the network, and incorporates a pyramid pooling module, which enhances the model's contextual understanding at multiple scales. These components collectively allow for more precise boundary localization and better discrimination of change versus background.
The consistently high OA of 99.14% further confirms the model's robustness in detecting change and correctly identifying unchanged regions, which is a key challenge in imbalanced binary classification tasks like change detection.
In contrast, convolutional methods such as FC-EF, FC-Siam-Di, and FC-Siam-Conc rely primarily on early or shallow fusion

strategies and lack multi-scale context modeling and attention-guided refinement mechanisms. As a result, these models show relatively lower performance, particularly in IoU and F1-score, indicating their limitations in capturing the subtle spatial and semantic nuances required for high-resolution change detection. While attention-based methods such as DTCDSCN and SNUNet-CD offer improved performance, they still fall short in fully leveraging multi-resolution feature alignment and structured refinement, key innovations in the proposed design.

Figures 5–8 present the visual results of change detection models applied to representative samples from the LEVIR-CD dataset. These comparisons include two input images (pre- and post-change), the ground truth change mask, and the predictions generated by six methods: FC-EF, FC-Siam-Di, FC-Siam-Conc, DTCDSCN, SNUNet-CD, and the proposed model. The predictions are color-coded for diagnostic clarity, where true positives, true negatives, and false positives are shown in white, black, and green, respectively.

Across all cases, the proposed method produced visually cleaner and more accurate change maps, with notably lower rates of false positives and false negatives. In contrast, the baseline models, particularly FC-EF and FC-Siam variants, frequently misclassify unchanged areas as changed (red regions) or miss actual changes (green regions).

Figure 5 shows small and spatially isolated buildings. While most methods detect the main structures, FC-EF and FC-Siam-Di show substantial false positives around edges. SNUNet-CD detects the buildings but introduces many artifacts (red and green regions). The proposed method shows a strong alignment with the ground truth, achieving sharp object boundaries and minimum misclassification. This illustrates its potential for superior spatial precision, highlighting areas for further enhancement and application.

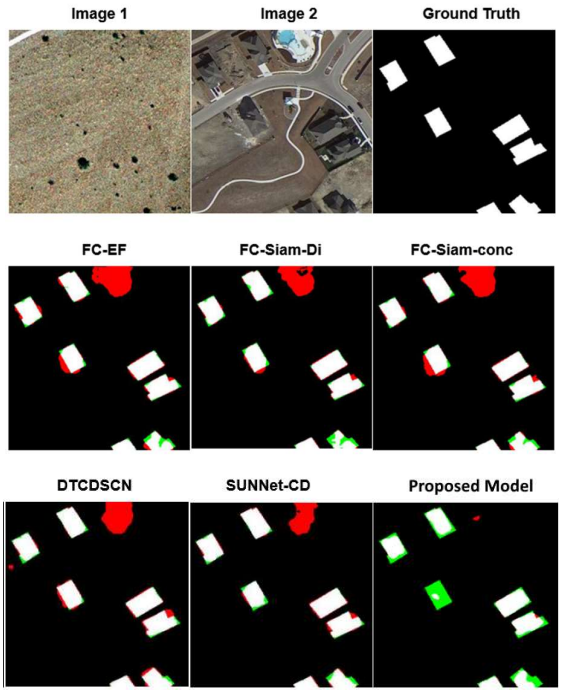


Figure 5. Visualization results of different methods on the LEVIR-CD

Figure 6 illustrates a significant change due to a newly constructed building. Although many methods initially identify the structure, they struggle with issues of over-segmentation (red) and under-segmentation (green). In contrast, the proposed model achieves clear and precise object boundaries, nearly perfectly overlapping with the ground truth. This success draws from its high-resolution features and effective context aggregation, which enhance its accuracy in capturing the scene's details.

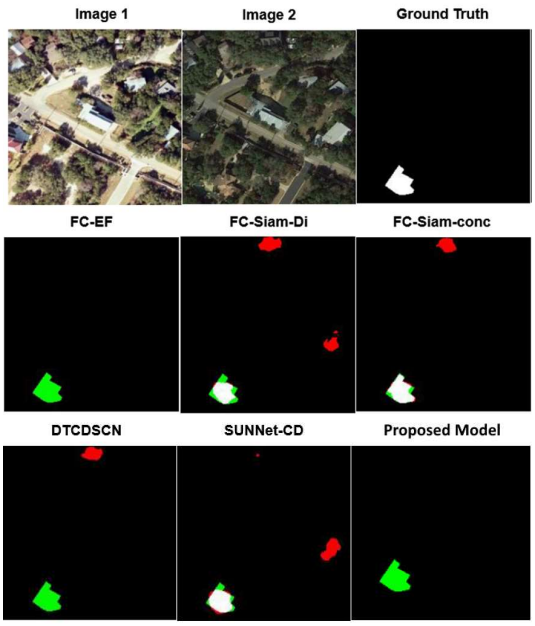
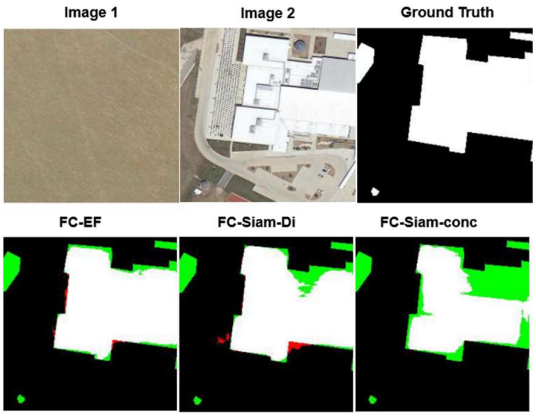


Figure 6. Visualization results of different methods on the LEVIR-CD

Figure 7 shows a challenging large-scale change involving a newly constructed building. Although most methods capture the structure, they suffer from severe over-segmentation (red) or under-segmentation (green). The proposed model shows clear object boundaries with near-perfect overlap with the ground truth, benefiting from the high-resolution features and context aggregation.

In Figure 8, a dense urban area with many small buildings, the baseline models exhibit extensive false positives and fragmented predictions due to their limited spatial resolution and weak multi-scale reasoning. In contrast, the proposed method produces well-structured predictions, with accurate separation between adjacent objects and fewer spurious detections.



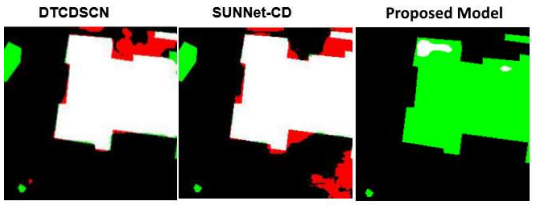


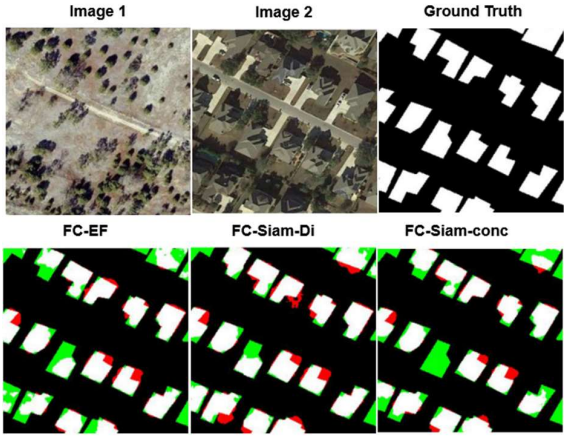
Figure 7. Visualization results of different methods on the LEVIR-CD

4. DISCUSSION

The experimental results, encompassing both quantitative evaluations and qualitative visual comparisons, clearly demonstrate the effectiveness and robustness of the proposed change detection framework. On the LEVIR-CD test set, the proposed method consistently outperforms a broad range of state-of-the-art baselines across all key metrics, including precision, recall, F1-score, IoU, and overall accuracy. These gains are consistent across diverse change scenarios ranging from small isolated structures to large-scale urban structures, indicating strong generalization ability.

The superior performance is rooted in several carefully designed architectural components. Adopting HRNet as the backbone allows for high-resolution feature preservation, critical for delineating fine object boundaries in high-resolution remote sensing imagery. The dual attention mechanisms, which include channel-wise squeeze-and-excitation blocks and spatial attention modules, provide an excellent way for the network to focus on meaningful features while minimizing noise. This capability is especially advantageous in complex scenes that involve shadows and clutter, as it enhances overall performance. Furthermore, the pyramid pooling module significantly strengthens the model's ability to capture contextual cues at different scales, enabling it to adeptly recognize both broad and localized changes. This combination of features empowers the model to operate with greater efficiency and accuracy in challenging environments.

In the qualitative results, the proposed method exhibits notably fewer false positives and false negatives compared to all baselines. This highlights its ability not only to detect where changes occur but also to maintain precise spatial boundaries and suppress spurious activations. Unlike many Siamese or encoder-decoder-based networks that suffer from spatial resolution loss or weak multi-scale fusion, the proposed framework delivers sharp, semantically coherent, and spatially accurate change maps. Collectively, these results affirm that the proposed method effectively addresses longstanding challenges in change detection for high-resolution remote sensing applications.



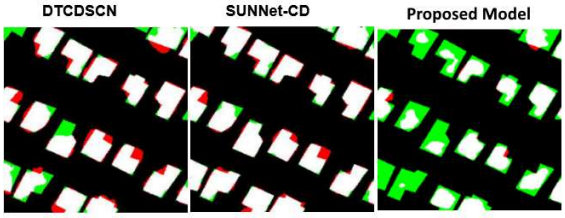


Figure 8. Visualization results of different methods on the LEVIR-CD

5. CONCLUSION

This work proposes a novel deep learning-based framework for change detection in high-resolution remote sensing imagery. The architecture integrates high-resolution feature extraction through HRNet, dual attention mechanisms for enhanced feature representation, pyramid pooling for multi-scale context aggregation, and a refinement module for spatial precision. This unified design enables the model to accurately detect changes of varying size and complexity while minimizing false alarms.

Extensive experiments on the LEVIR-CD dataset demonstrate that the proposed model significantly outperforms existing state-of-the-art methods across all major evaluation metrics. Quantitative and qualitative analyses confirm the model's robustness, accuracy, and capacity to generalize to a wide range of change scenarios. The proposed method effectively balances semantic understanding and spatial detail, offering a reliable solution for large-scale, real-world change detection tasks.

Future work will explore the extension of this framework to multi-temporal and multi-modal settings, as well as its application to other domains such as disaster management, agricultural monitoring, and urban planning.

REFERENCES

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

Bovolo, F., Bruzzone, L., 2006: A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(1), 218–236.

Bovolo, F., Bruzzone, L., 2015: The time variable in data fusion: A change detection perspective. *IEEE Geoscience and Remote Sensing Magazine*, *3*(3), 8–26.

Celik, T., 2009: Unsupervised change detection in satellite images using principal component analysis and $ k $-means clustering. *IEEE Geoscience and Remote Sensing Letters*, *6*(4), 772–776.

Chen, H., Qi, Z., Shi, Z., 2021: Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–14.

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.

Chen, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y., Li, H., 2020: DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied*

Earth Observations and Remote Sensing, *14*, 1194–1206.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., Lambin, E., 2004: Review ArticleDigital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, *25*(9), 1565–1596.

Daudt, R. C., Le Saux, B., Boulch, A., 2018: Fully convolutional siamese networks for change detection. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 4063–4067.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009: Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Fang, S., Li, K., Shao, J., Li, Z., 2021: SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Giglio, L., Schroeder, W., Justice, C. O., 2016: The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sensing of Environment*, *178*, 31–41.

Hou, X., Bai, Y., Li, Y., Shang, C., Shen, Q., 2021: High-resolution triplet network with dynamic multiscale feature for change detection on satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *177*, 103–115.

Hu, J., Shen, L., Sun, G., 2018: Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, X., Cao, Y., Li, J., 2020: An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sensing of Environment*, *244*, 111802.

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., Xian, G., 2013: A comprehensive change detection method for updating the National Land Cover Database to circa 2011. *Remote Sensing of Environment*, *132*, 159–175.

Li, Z., Shi, W., Zhang, H., Hao, M., 2017: Change detection based on Gabor wavelet features for very high resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 783–787.

Li, Z., Tang, C., Wang, L., Zomaya, A. Y., 2022: Remote sensing change detection via temporal feature interaction and guided refinement. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11.

Liu, R., Kuffer, M., Persello, C., 2019: The temporal dynamics of slums employing a CNN-based change detection approach. *Remote Sensing*, *11*(23), 2844.

Liu, W., Kang, Z., Liu, J., Lin, Y., Yu, Y., Li, J., 2024: A Multitask CNN-Transformer Network for Semantic Change Detection From Bi-temporal Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X., 2020: Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, *18*(5), 811–815.

Loshchilov, I., Hutter, F., 2017: Decoupled weight decay regularization. *ArXiv Preprint ArXiv:1711.05101*.

Nielsen, A. A., Conradsen, K., Simpson, J. J., 1998: Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, *64*(1), 1–19.

Powers, D. M. W., 2020a: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*.

Ronneberger, O., Fischer, P., Brox, T., 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241.

Shi, W., Zhang, M., Zhang, R., Chen, S., Zhan, Z., 2020: Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, *12*(10), 1688.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J., 2019: High-resolution representations for labeling pixels and regions. *ArXiv Preprint ArXiv:1904.04514*.

Voigt, S., Kemper, T., Riedlinger, T., Kiefl, R., Scholte, K., Mehl, H., 2007: Satellite image analysis for disaster and crisis-management support. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(6), 1520–1528.

Volpi, M., Tuia, D., Bovolo, F., Kanevski, M., Bruzzone, L., 2013: Supervised change detection in VHR images using contextual information and support vector machines. *International Journal of Applied Earth Observation and Geoinformation*, *20*, 77–85.

Zhai, H., Zhang, H., Li, P., Zhang, L., 2021: Hyperspectral image clustering: Current achievements and future lines. *IEEE Geoscience and Remote Sensing Magazine*, *9*(4), 35–67.

Zhang, M., Shi, W., 2020: A feature difference convolutional neural network-based change detection method. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(10), 7232–7246.

Zhang, M., Xu, G., Chen, K., Yan, M., Sun, X., 2018: Triplet-based semantic relation learning for aerial remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, *16*(2), 266–270.

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.