International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhanced Building Change Detection in Aerial Imagery Using Siamese U-Net Architectures

Dr. G. Malini Devi [1], Wajiha Kulsum[2], Ch. Radhika [3]

[1]*Department of CSE, G. Narayanamma Institute of Technology and Science*
[2] *Department of CSE, G. Narayanamma Institute of Technology and Science*
[3] *Department of CSE, G. Narayanamma Institute of Technology and Science*

Abstract - Monitoring urban development and infrastructure changes is essential for sustainable city planning and disaster management. This paper investigates building change detection using two enhanced Siamese U-Net architectures: one utilizing feature concatenation and the other leveraging element-wise difference between bi-temporal aerial images. The LEVIR-CD dataset, comprising high-resolution paired images of urban landscapes, is employed to train and evaluate the models. Preprocessing includes slicing large images into manageable patches, applying augmentation techniques, and normalizing inputs. Experimental results demonstrate that the Siamese U-Net with difference outperforms its concatenation counterpart, achieving superior accuracy, F1-score, and intersection-over-union metrics. The findings highlight the effectiveness of dual-branch networks, particularly difference-based approaches, in capturing subtle structural transformations in complex urban environments. These results support the adoption of advanced Siamese architectures as robust tools for automated building change detection in remote sensing applications.

Key Words: Building change detection, Siamese U-Net, Aerial imagery, Urban monitoring, Remote sensing, Bi-temporal analysis.

1. INTRODUCTION

Urban areas undergo continuous transformation due to construction, renovation, and demolition of buildings, driven by socio-economic, environmental, and political factors. Detecting and monitoring these building-level changes is vital for informed urban planning, infrastructure management, and post-disaster recovery [2], [3]. Remote sensing has emerged as a powerful means to observe urban dynamics, providing wide-area, repeatable, and objective data about the Earth's surface [4]. Among remote sensing modalities, high-resolution aerial imagery is particularly well-suited for building change detection, as it captures fine-grained structural details necessary for analyzing individual buildings [7].

Historically, building change detection relied on handcrafted features and thresholding techniques combined with classical machine learning models, such as Support Vector Machines and Random Forests [5]. Although successful in a controlled environment, the methods failed in urban settings given noise, occlusions and texture heterogeneity. The introduction of deep learning specifically convolutional neural networks (CNNs) has provided a new edge to the development of end-to-end extracted features and semantic segmentation [1], [6]. To the best of our knowledge, the U-Net architecture is one of the most popular CNN-based architecture used for its encoder–decoder structure and skip connections, which help preserve spatial information during segmentation tasks [1]. Despite the success of U-Net, it does not explicitly model bi-temporal relationships between pre- and post-change images. To address this, Siamese network architectures have been developed, in which two parallel branches process image pairs and extract comparative features [8]. Such networks can highlight subtle structural differences by either concatenating the extracted features or computing their element-wise difference [6].

This study investigates and compares two enhanced Siamese U-Net variants: Siamese U-Net Concatenation and Siamese U-Net Difference, for detecting building changes in urban scenes using high-resolution aerial imagery. These models were trained and evaluated on the publicly available LEVIR-CD dataset [7], which provides a diverse benchmark of urban landscapes for change detection. By analyzing their relative performance, this work aims to assess the suitability of these architectures for accurate and automated building change detection in remote sensing applications.

2. LITERATURE REVIEW

Change detection generation in buildings has evolved so drastically over the past couple of years through the conventional techniques down to the more complex deep learning models. Early methods, instead, were focused on handmade features, thresholding and, one the conventional machine learning techniques, like the Support Vector Machines and Random Forests. Though such methods have been demonstrated to perform successfully in limited environments, they have not been proven effective in the presence of tough urban textures, obscures, and corrupted data [5]. A significant breakthrough in the field was conditioned by the invention of deep learning (more precisely in the convolutional neural networks (CNNs) format). The U-Net architecture, presented by Ronneberger et al. [1] gave a model framework to the later segmentation methods because of its encoder-decoder model and its skip connections, the spatial information is retained. U-Net and its versions have achieved performance satisfaction in developing segmentation and change detection task based on high-resolution remote sensing imagines [7].

In order to compensate for the lack of explicitness from U-Net with respect to bi-temporal relationships, Siamese network architectures were proposed. These architectures use parallel branches for pre- and post-change images, and thus directly compare feature representations. Peng et al. [8], authors introduced a Siamese U-Net with spatial attention for better change localization, focusing more on the salient regions. Zhang et al. [6] proposed SMD-Net, a Siamese multi-scale difference-enhancement network focusing on subtle structural

differences by enhancing multi-scale features. Recent studies have incorporated attention mechanisms and transformer-based components to further improve performance. Feng et al. [3] introduced SGNet, a semantic-guided transformer-based network that achieved high accuracy in complex urban environments by integrating semantic guidance and multi-stage fusion. Chen et al. [9] proposed SMADNet, which combined multiscale decoding with attention modules in a Siamese framework, demonstrating superior results on the LEVIR-CD dataset.

Additional advancements have focused on improving data efficiency and robustness. Benchabana et al. [2] employed enhanced super-pixel segmentation alongside deep learning models to better preserve object boundaries in high-resolution imagery. Sirko [4] explored the use of medium-resolution Sentinel-2 data for scalable building detection, offering a more accessible alternative to high-resolution datasets. Corley et al. [5] conducted a critical evaluation of recent change detection methods, emphasizing the importance of reproducibility and realistic benchmarking. Cui et al. [10] introduced a Siamese Swin-Unet, combining Swin Transformers with the Siamese architecture to improve accuracy and contextual understanding in change detection tasks.

Overall, these studies demonstrate the evolving trends from traditional methods to modern deep learning frameworks, in which the efficacy of Siamese networks and attention mechanisms for capturing temporal dynamics is emphasized. However, the choice of feature fusion strategy in Siamese U-Nets (e.g., concatenation vs. element-wise difference) is still an open question. This paper fills this gap by conducting a comprehensive comparison of the two approaches of building change detection over high-resolution aerial imagery.

3. METHODOLOGY

The methodology describes the dataset, and preprocessed activity, model architectures, training, and evaluation metrics used to compare two Siamese U-Net models on an application for building change detection.

In this paper, there is a comparative study of the performance of Siamese U-Net Concatenation and Siamese U-Net Difference architectures based on high-resolution aerial imagery.

3.1. Dataset

The LEVIR-CD dataset [7] was chosen as reference for this work. It consists of 637 pairs of high-resolution aerial images (also known as orthophotos) with a spatial resolution of 0.5 meters and the size of 1024×1024 pixels.

The two datasets have different urban or suburban imagery and are paired with binary change masks of building level changes between two times. To adapt the data for training, the images were divided into non-overlapping patches of size 256×256 pixels, ensuring manageable input dimensions while preserving spatial context.

3.2. Preprocessing

To enhance model generalization and address data imbalance, several augmentation techniques were applied during training. These included random horizontal and vertical flips to introduce variability, random cropping at multiple scales to simulate different spatial resolutions, and normalization of pixel intensities to a standardized range. The dataset was partitioned into training, validation, and test sets in proportions of 70%, 10%, and 20% respectively, ensuring unbiased evaluation.

3.3. Model Architectures

Two Siamese U-Net variants [6], [8] were implemented to perform building change detection by comparing bi-temporal aerial image pairs. Both architectures employ dual-branch encoders and a shared decoder. Each encoder processes one image from the input pair, extracting hierarchical feature representations independently. The outputs of the two encoders are then combined and passed to the decoder, which generates a pixel-wise binary change mask delineating areas of structural transformation. In Siamese U-Net Concatenation, both the encoder feature maps are concatenated in terms of the channels followed by feeding the concatenated feature maps into the decoder. This methodology conserves feature diversity because all the data in both the images remains so that the decoder gets to learn intricate interactions between the unmodified portions and the altered. Siamese U-Net Difference, contrastingly, calculates absolute difference between feature maps of the two encoders prior to decoding. These highlights areas of non-agreement thus becoming more responsive to small structural variation and masks overrepresented data residing in areas lacking structural change.

The two architectures use ResNet-50 backbone encoder since it has already shown to extract rich and multi-scale features in high-resolution imagery. The decoder is a conventional U-Net structure [1], and the transposed convolutional layers and skip links are sequentially used to restore the spatial information, and feature information of the earlier encoder stages are added. Through such an arrangement they can define boundaries of change accurately and less spatial information is lost in down-sampling. Both fusion strategies concatenation and difference are introduced to be compared as both of them are radically different ways of modeling change in terms of carrying all the information about features or focusing on differences. The relative comparison in terms of their effectiveness provides details in regard to feature fusion and how it impacts the accuracy of building change detection [6], [8].

3.4. Training and Evaluation

The corresponding models were trained with the framework of PyTorch, and the objective was cross-entropy loss. The model was optimized using stochastic gradient descent (learning rate = 0.01) and the model parameters were iteratively set and optimized until the validation performance stabilized. The performance metrics were accuracy, F1 score, intersection-over-union (IoU), precision, and recall that were computed on the held-out test set to evaluate the detection comprehensively [5].

The inclusion of IoU provides insight into the spatial overlap between predicted and actual change regions, complementing other metrics.

4. RESULTS AND DISCUSSION

The performance of the Siamese U-Net Concatenation and Siamese U-Net Difference models was evaluated using five standard metrics: accuracy, precision, recall, F1-score, and intersection over union (IoU). These metrics are mathematically defined as follows:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Intersection over Union (IoU)

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Together, these metrics provide a comprehensive view of model performance, capturing overall correctness (Eq. (1)), ability to avoid false positives (Eq. (2)), sensitivity to true changes (Eq. (3)), harmonic balance between precision and recall (Eq. (4)), and spatial overlap with the ground truth (Eq. (5)).

4.1. Overall Performance

Table 1 presents the overall test-set performance of the two Siamese U-Net variants. The Siamese U-Net Difference outperformed the concatenation model across all evaluation metrics defined in Eqs. (1)–(5). Specifically, the difference model achieved an accuracy of 94.25%, an F1-score of 95.23%, and an IoU of 0.9124, compared to the concatenation model's accuracy of 88.21%, F1-score of 91.12%, and IoU of 0.8457. The difference model also exhibited higher precision and recall, indicating better capability in identifying changed regions while minimizing false alarms.

Table 1- Overall performance metrics of Siamese U-Net variants.

Model Name	Accuracy (%)	F1-Score (%)	IoU (%)	Precision (%)	Recall (%)
Siamese U-Net Concatenation	88.21%	91.12%	84.57%	97.95%	99.30%
Siamese U-Net Difference	94.25%	95.23%	91.24%	99.10%	99.45%

The superior performance of the difference model can be attributed to its explicit modeling of change through element-wise difference of feature maps, which enhances discrimination between changed and unchanged areas. This aligns with findings from prior studies [6], [8], where difference-based fusion was shown to emphasize subtle structural variations more effectively than concatenation. Furthermore, the difference model achieved an IoU (Eq. (5)) exceeding 0.91, suggesting strong spatial alignment of predicted change regions with the ground truth masks, a critical requirement in practical urban analysis applications.

The concatenation model, while outperforming baseline U-Net variants (not shown here for brevity), was less sensitive to fine-grained changes. This may be due to its strategy of preserving full feature information from both images without explicitly emphasizing differences, which can lead to retention of redundant information from unchanged areas.

4.2. Sample-wise Analysis: Siamese U-Net Concatenation

To better understand the models' behavior on individual test samples, results for three representative image pairs are presented in Table 2. The concatenation model demonstrated consistent performance across the samples, achieving accuracy (Eq. (1)) values between 97.8% and 99.0%, with F1-scores (Eq. (4)) above 89.6% and IoU (Eq. (5)) above 80%.

Visual inspection of the results revealed that while most major changes were correctly detected, the model tended to over-segment regions in cluttered areas, leading to minor false positives. This tendency can be linked to its fusion strategy, which retains all feature information without explicitly suppressing irrelevant patterns. Nevertheless, the high precision (Eq. (2)) observed in all three samples suggests that when it predicts a change, it is generally correct.

In Sample 1, which contains moderately dense urban development with clear change boundaries, the model achieved an accuracy of 98.1% and an IoU of 82.1% (Eq. (5)). The predicted change mask captured the major transformations but included some over segmentation at building edges. This suggests that while concatenation retains rich feature representations, it may struggle to suppress fine-grained irrelevant textures.

In Sample 2, representing a suburban area with smaller and more dispersed structures, the model recorded the lowest F1-score (89.6%) and IoU (81.1%) among the three samples. Visual inspection of the predictions indicates that the model failed to completely capture several small-scale changes, likely due to its tendency to preserve all feature information, including redundant or misleading patterns from unchanged areas.

Table 2- Sample-wise results of Siamese U-Net Concatenation.

Original image	Changed image	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)	Detection Result
		98.1	92.3	88.1	90.2	82.1	
		97.8	97.5	82.8	89.6	81.1	
		99.0	98.4	83.2	90.2	80.1	

In contrast, Sample 3**,** which featured a relatively open area with a few isolated structures, yielded the highest accuracy (99.0%) and a comparable IoU of 80.1%. In this case, the model effectively identified the changed regions with minimal false positives, reflecting its strength in less cluttered scenes where preserved feature diversity is advantageous.

Across all three samples, the model consistently achieved high precision (Eq. (2)), exceeding 92%, which indicates a low rate of false positives. However, the comparatively lower recall (Eq. (3)) in Samples 2 and 3 suggests that some true changes were missed, a common limitation of concatenation-based fusion. This behavior is consistent with findings in the literature [8], which reported that concatenation retains more complete information but lacks the explicit focus on differences that aids in detecting subtle changes. Furthermore, the relatively stable performance in various conditions proves the high degree of uniformity in the use of concatenation strategy as the measure of maintaining the general proportion of accuracy high, and overidentifications of the unchanged areas and under detections of minor changes prove the types of improvement that are to be made. These findings are confirmed by the fact that Siamese U-Net Concatenation provides a satisfactory level of feature diversity once more they can be reliably used to perform acceptably with open or simply complex city views though their performance is less impressive when they are dealing with dense or fine-grained scenes. This appears to indicate the trade-off in the capability of retaining features richness and outlining the structural variations in change detection tasks.

4.3. Sample-wise Analysis: Siamese U-Net Difference

And to be able to extend such comparison of the effectiveness of the Siamese U-Net Difference model, Table 3 provides us with the results that the same three examples of the test samples take. This review enables us to know the nature of the model in different urban cases as well as the whereabouts and how it is superior to the concatenation one.

Table 3- Sample-wise results of Siamese U-Net Difference.

Original image	Changed image	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)	Detection Result
		99.3	93.5	92.1	92.8	85.2	
		98.6	96.1	88.4	92.1	84.1	
		99.4	97.2	90.7	93.8	84.9	

In Sample 1, corresponding to a highly built-up urban region with strict boundaries between change, the Difference model provided an accuracy of 99.3% as well as an IoU of 85.2% (Eq. (5)). The model was able to identify most change areas with less false positive outcomes than the concatenation model. The improved IoU recommends that the absolute difference computation of encoder features will effectively discard irrelevant information resulting in the ability to provide more accurate change localization.

In Sample 2, a residential part with scattered, low-level modifications, the model showed that its performances were robust at an accuracy of 98.6% and an IoU of 84.1%. Interestingly, F1-score (Eq. (4)) went up to 92.1% compared to that of concatenation model, which showed that it was more balanced between precision (Eq. (2)) and recall (Eq. (3)). Visual evaluation showed that the difference model identified some slight variations that could not be identified by the concatenation model and that the former was more sensitive to smaller structural modifications.

Among the three samples, the highest scores were recorded on the sample 3 that had an open area and less density of buildings 99.4%, F1-score 93.8%, and IoU 84.9% according to the difference model. The mask of detection was similar to the ground truth with well-marked contours and had limited numbers of false positives. It means that the difference model does more than excelling in cluttered conditions as it also exhibits a high level of robustness in rich scenes. The difference

model yielded improved performance in all three of the samples in comparison to their concatenation model analog further regarding the value of the IoU (Eq. (5)), i.e., direct overlapping measure of predictions against the ground truth. The enhanced recollection (Eq. (3)) in every instance demonstrates the enhanced potential to identify the actual alterations and does not disregard the essential regions which is one of the most vital limitations which have been pointed to the concatenation method. These values stand by the hypothesis that features level of emphasis guarantees that the model focuses more on the region of change boundaries and more briefly interrupted by constant regions.

This is in agreement with the literature in the past findings that the difference-based fusion strategies have been reported as performing excellently in the complex urban settings [6], [9]. The gradual enhancements of the chaotic, scattered and free urban landscapes youth the soundness of difference model and corroborate the proper decision to employ it to the practice of building change detection in buildings. Though it has some portion of excessive segmentation as well in very intricate areas, the results are significantly better than those of fusion using concatenation.

4.4. Comparative Analysis of Models

The contrast analysis of the two versions of the Siamese U-Net shows that difference model responded mostly efficiently on all the measures of evaluation provided in Eqs. (1)-(5) and within any situations tested. According to Table 1, the difference model output had an overall accuracy of 94.25% and IoU of 91.24%, which was better than that of concatenation model with accuracy value of 88.21% and IoU of 84.57%.

This tendency was also observed in the sample-wise analysis given at Tables 2 and 3 whereby the difference model was the superior performer on various urban scenes, such as densely populated ones, suburban areas, and open territories. The enhanced performance can be explained by the fact that the difference model attaches an explicit emphasis on structural inconsistency in the feature space that allows one to eliminate redundant information in the stable areas and improve change detection of weak differences.

These results ensure that difference-based feature fusion is a more certain and accurate method of constructing change detection in high-resolution aerial images, and therefore it is a more effective alternative in real-life use of the situation in actual cities surveillance.

5. CONCLUSION

This paper did a thorough benchmarking experiment of the two Siamese U-Net architectures, Siamese U-Net Concatenation and Siamese U-Net Difference, to generate building change detection in high resolution aerial imagery. The benchmark was the publicly available LEVIR-CD dataset that depicts a multi-various urban and suburban environment. The fair and transparent evaluation was carried out using standard preprocessing methods as well as stringent evaluation metrics as stated in Eqs. (1)–(5), were employed to ensure fair and transparent assessment.

Both models proved to be very efficient regarding their possibility to discover building-level changes, which proves that Siamese architectures are appropriate to detect changes. In all the metrics, the difference model had surpassed the concatenation model in all forms of accuracy, F1-score, and IoU at both the overall and sample-wise levels. In particular, the difference model achieved an IoU of 0.9124, representing accurate spatial correspondence of the inferred and truth change area and pointing out strong results under many urban settings such as dense, cluttered, scattered landscapes.

These results are consistent with the fact that predicting a difference in features between bi-temporal images via direct modeling of the difference facilitates the network to concentrate on the areas of real change and ignore the insignificant data on areas that have not changed. The analysis further revealed that the concatenation model retained more feature diversity, which was beneficial in simpler scenes but led to over segmentation and false positives in more complex environments. In contrast, the difference model maintained high precision and recall across all samples, demonstrating its robustness and reliability. The results align with observations from prior studies that have emphasized the advantages of difference-based fusion in accurately detecting subtle and localized structural changes.

Although results of this study are encouraging a number of opportunities on further improvement exist. The two models had problems of being ineffective in highly congested urban environments, especially in environments where tiny small-scale changes occurred and were not captured. Future study of the insertion of attention mechanisms [9], [10], that would allow the network to pay special attention to the eye-catching areas and enhance sensitivity in detecting the changes could be discussed to overcome these challenges. Also, inclusion of transformer-based architecture, that are feature known to identify long-range dependencies and contextual relations, could also help in improving the model capacity to work with complex spatial patterns [3]. The other area of improvement is to use self-supervised learning that may minimize the need to use concern labels and improve generalization to unseen settings.

Lastly, this area can be further expanded to multi-classes change detection and real-time inference that can expand its applicability to real-time-monitoring and humanitarian-based urban-monitoring situations. Altogether, the results of the study indicate the high efficiency of Siamese U-Net Difference architectures in the development of change detection and support several promising avenues of further development of deep learning-based change detection methods in remote sensing.

ACKNOWLEDGEMENT

The authors are grateful to their institute for providing the facilities and support that enabled this research. Special thanks are extended to the department leadership and research mentors for their guidance, encouragement, and continued support throughout the project.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[2] A. Benchabana, M.-K. Kholladi, R. Bensaci, and B. Khaldi, "Building detection in high-resolution remote sensing images by enhancing superpixel segmentation and classification using deep learning approaches," *Buildings*, vol. 13, no. 7, p. 1649, 2023.

[3] J. Feng, X. Yang, and Z. Gu, "SGNet: A Transformer-Based Semantic-Guided Network for Building Change Detection," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 9922–9935, 2024.

[4] W. Sirko, "High-Resolution Building and Road Detection from Sentinel-2," *arXiv preprint* arXiv:2310.11622, 2023.

[5] I. Corley, C. Robinson, and A. Ortiz, "A Change Detection Reality Check," *arXiv preprint* arXiv:2402.06994, 2024.

[6] X. Zhang, L. He, K. Qin, Q. Dang, H. Si, X. Tang, and L. Jiao, "SMD-Net: Siamese Multi-Scale Difference-Enhancement Network for Change Detection in Remote Sensing," *Remote Sensing*, vol. 14, no. 7, 1580, 2022. https://doi.org/10.3390/rs14071580

[7] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020. https://doi.org/10.3390/rs12101662

[8] D. Peng, Q. Zhang, and H. Guan, "High-Resolution Building Change Detection via Siamese Network with Spatial Attention," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 110–123, 2021. https://doi.org/10.1016/j.isprsjprs.2021.05.014

[9] Y. Chen, J. Zhang, Z. Shao, X. Huang, Q. Ding, X. Li, and Y. Huang, "SMADNet: A Siamese Multiscale Attention Decoding Network for Building Change Detection on High-Resolution Remote Sensing Images," *Remote Sensing*, vol. 15, no. 21, p. 5127, 2023. https://doi.org/10.3390/rs15215127

[10] Z. Cui, L. Wang, S. Cheng, and Y. Li, "Siamese Swin-Unet for Image Change Detection," *Scientific Reports*, vol. 14, p. 54096, 2024. https://doi.org/10.1038/s41598-024-54096-8