International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n2.08
Journal homepage: https://wepub.org/index.php/IJCSIT/index

The Improved Changeformer for Remote Sensing Change Detection

Shanshan Zhu *, Yuxi Wu

College of Computer Science, Nanjing Audit University, Nanjing, Jiangsu, China
* Corresponding Author: Shanshan Zhu

ABSTRACT

With advancements in remote sensing satellite technology, change detection (CD) has become a crucial technique in remote sensing image processing. Traditional CD methods, including algebra-based, transformation-based, and classification-based approaches, have contributed significantly to change analysis but face challenges such as misclassification and limited adaptability. Recent developments in deep learning, particularly convolutional neural networks (CNNs) and fully convolutional networks (FCNs), have improved CD accuracy by enabling pixel-level predictions. However, CNN-based methods struggle with multi-scale feature extraction and distinguishing between change and static information. To address these limitations, attention mechanisms and Transformer-based architectures, such as ChangeFormer, have been introduced to enhance long-range dependency modeling and spatial feature representation. Additionally, the Atrous Spatial Pyramid Pooling (ASPP) module further improves multi-scale feature extraction by expanding the network's receptive field without reducing resolution. This study proposes integrating ASPP into deep learning models to enhance the efficiency, accuracy, and robustness of change detection in complex remote sensing applications.

KEYWORDS

Change Detection; Atrous Spatial Pyramid Pooling; Transformer.

1. INTRODUCTION

In light of advances in remote sensing satellite image technology, which are marked by significant improvements in temporal, spatial, and spectral resolutions, the change detection (CD) technique using satellite data has become central to remote sensing image processing. Initially, high-resolution images of a designated area are acquired by remote sensing satellites at different time points. Subsequently, the CD method automatically evaluates alterations between these images. Analyzed parameters include, but are not limited to, increases or decreases, displacement, and morphological changes of surface features, all of which are crucial for comprehensively understanding surface dynamics. In summary, change detection methods accurately delineate areas of temporal alteration. Given the pivotal role of satellite remote sensing data in natural resource development and conservation management, CD technology shows substantial promise for applications in environmental monitoring and resource management, early warning and assessment of natural disasters, urban planning and construction, agricultural surveillance, and military operations.[1]

Since the late 1970s, change detection technology has undergone significant advancements. Researchers worldwide have continuously explored and proposed various traditional methodologies. These methods can be broadly categorized into algebra-based, transformation-based, and classification-based approaches, each employing different technical strategies. Algebra-based

methods detect changes by computing algebraic relationships between remote sensing images from different time phases. Common algorithms include image quantization[2], image differencing[3], and image regression[4]. Image subtraction[5] is the most fundamental approach, identifying changes by calculating differences in grayscale values between two images. [5]While straightforward and intuitive, it struggles to determine the nature and type of changes. Transformation-based methods extract features from remote sensing images to distinguish changed and unchanged pixels, thereby improving detection accuracy. Common techniques include the Tasseled Cap Transform (TCT)[6], Principal Component Analysis (PCA)[7], and Change Vector Analysis (CVA)[8]. The TCT, introduced by Kauth and Thomas in 1976, converts multispectral images into orthogonal components, where brightness, greenness, and wetness serve as indicators of land cover changes. PCA transforms original data into a new coordinate system through orthogonal transformation to extract key change information. The CVA algorithm, proposed by Malila in 1980, analyzes shifts in ground cover by examining vectorial changes in pixels across different time phases, improving classification accuracy. These transformation techniques reduce data dimensionality and noise, enhancing detection precision. However, they struggle with the complexity of ground cover types and diverse change patterns, leading to potential misclassification. Classification-based methods assign semantic labels to each pixel using classifiers, then generate change maps by comparing pre-change and post-change classifications. Common algorithms include maximum likelihood estimation[9], support vector machines (SVMs)[10], random forests[11], and artificial neural networks (ANNs)[12]. SVM, proposed by Vapnik in 1995, identifies the optimal classification boundary by maximizing the margin between different classes, demonstrating strong performance in remote sensing change detection. ANNs, inspired by human neural networks, offer high adaptability and robustness. However, classification-based methods require extensive training data and prior knowledge, limiting their adaptability to diverse and complex ground cover types.

In recent years, deep learning methodologies, particularly convolutional neural networks (CNNs), have been widely applied in remote sensing tasks such as hyperspectral image classification and aerial image segmentation, leading to significant advancements. As a result, CNNs have been adopted for change detection by directly comparing corresponding image blocks from dual-temporal high-resolution satellite imagery to derive detection results. [13]However, traditional CNNs rely on fully connected layers, limiting their ability to produce dense pixel-level predictions. To address this, fully convolutional networks (FCNs), composed exclusively of convolutional operations, were developed. Since their introduction, FCNs and their variants, such as U-shaped networks, have become the dominant approach for pixel-level change detection. For example, Daudt et al. proposed a U-network-based framework that processes dual-temporal high-resolution satellite imagery to generate pixel-level detection results.[14] To enhance boundary sharpness, Gu Lian et al. integrated a refinement module into the U-network decoder, enabling each layer to independently predict detection results. [15]While U-networks preserve spatial detail through cross-layer connections, they struggle to extract multiscale object features. To address this limitation, Peng et al. introduced dense cross-layer connections within a U-shaped network, facilitating multiscale feature extraction from dual-temporal images. [16]Building on this, Yu et al. utilized the difference map of dual-temporal images as input to refine detection results. [17]These approaches, termed "early fusion," concatenate dual-temporal images or their difference maps as input to CNNs for change detection. Another strategy, known as "late fusion," employs twin networks to independently extract features from two temporal phases before fusing them for detection. Daudt et al. proposed two late fusion methods based on U-networks, both of which duplicate the U-network encoder into two encoders with shared parameters to extract bi-temporal features. [14]The difference between these methods lies in their decoding process—one concatenates the encoder outputs with the decoded features, while the other concatenates the absolute difference of the two encoder outputs with the decoded features. Similarly, Zhang et al. applied a difference discrimination operation within the decoder to enhance detection accuracy. [18]These methods primarily focus on optimizing connection and fusion strategies to extract contextual information within the spatial neighborhood. However, they do not effectively distinguish between

changing and static information, and the inclusion of static information can degrade detection performance. Attention mechanisms, widely used in computer vision, help networks focus on regions of interest. Consequently, they have been incorporated into change detection models to enhance change features while suppressing static information. Researchers have introduced convolutional attention modules to emphasize changing objects in both spatial and channel domains. [19]Shi et al. proposed a deep supervision module that provides supervisory information to the network's middle layers, generating more robust features. [20]Cheng et al. combined channel attention, spatial attention, and distance maximization modules to enhance both change and static features.[21] Fang et al. integrated a channel attention module into a U-network to refine features across different semantic levels.[22] Additionally, researchers have employed self-attention mechanisms to capture relationships between any two pixel points. However, the computational cost of applying self-attention to all pixels is high. To mitigate this, Chen et al. proposed an improved self-attention mechanism that operates only on a limited number of high-level semantic tokens rather than all pixels.[23] Overall, attention-based change detection methods primarily rely on convolutional networks (ConvNets) for feature extraction and use attention mechanisms to emphasize key features. While these methods can capture global details, they struggle with long-range spatiotemporal dependencies due to their reliance on reweighting ConvNet-derived features. To overcome this limitation, Bandara et al. introduced ChangeFormer, a dual-network architecture based on the Transformer framework.[24] ChangeFormer leverages the non-local self-attention mechanism of Transformers, offering a larger effective receptive field (ERF) and superior context modeling compared to ConvNets. By incorporating Transformer architecture, ChangeFormer significantly improves change detection performance, enhancing the ability to capture long-range contextual information while reducing dependence on ConvNets. This results in notable advancements in remote sensing image change detection.

The ASPP (Atrous Spatial Pyramid Pooling) module offers distinct advantages in overcoming the limitations of the previously mentioned methods, which disrupt spatial order and degrade the spatial structure of images during segmentation, leading to deficiencies in spatial feature extraction. [25]Designed to expand the network's receptive field without reducing resolution, the ASPP module enhances the ability to capture multi-scale contextual information. It achieves this by employing multiple parallel dilated convolutional layers with different dilation rates, where features extracted at each rate are processed separately and then fused to generate the final output. Building on this, this paper proposes integrating the ASPP module into a deep learning model to develop a more efficient, accurate, and robust system for image segmentation and change detection. This approach aims to better address complex and evolving real-world applications.

2. METHOD

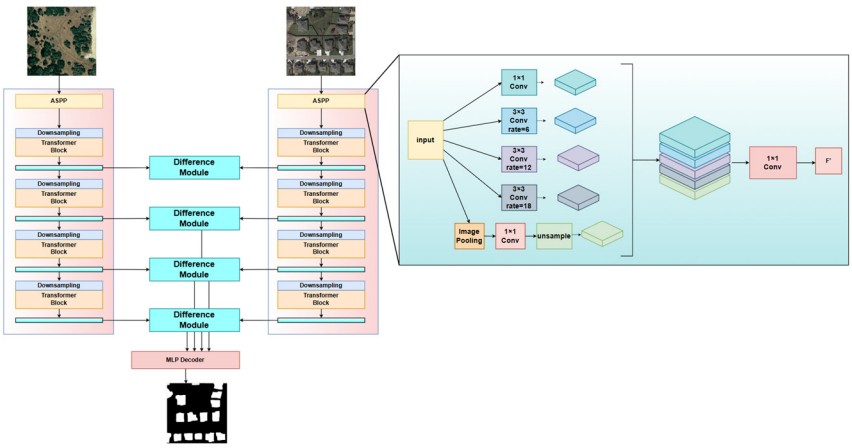

Figure 1. The proposed ChangeFormer+ network for CD

The proposed change detection model consists of two core components: the ChangeFormer module and the ASPP (Atrous Spatial Pyramid Pooling) module. Built upon the ChangeFormer framework, the model utilizes a hierarchical Transformer encoder to capture long-range contextual features from images. Specifically, bi-temporal remote sensing images are first processed by the ASPP module, which extracts multi-scale spatial features. These features are then fed into ChangeFormer to enhance long-range contextual representation. Finally, the processed features pass through an MLP decoder to generate the final change detection results.

2.1. Changformer

Traditional convolutional neural networks (CNNs) struggle to capture long-range contextual features in remote sensing images, especially when detecting large-scale changes in high-resolution imagery. This limitation arises from the local receptive fields of convolution operations, which restrict the effective extraction of global information. To address this, we introduce the Transformer-based ChangeFormer module, which leverages a self-attention mechanism to model global contextual relationships, thereby improving the accuracy and robustness of change detection.

The core of the ChangeFormer module is composed of multiple hierarchical Transformer Blocks. The Transformer Block effectively extracts global information from image features via the self-attention mechanism, generating multi-level features akin to those in Convolutional Networks, such as high-resolution coarse features and low-resolution fine-grained features. For the input image features F′, each Transformer Block calculates the correlation between each pixel and other pixels in the image using three matrices, subsequently adjusting the pixel representations. Specifically, the self-attention calculation is performed as follows:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

Here, (Q,K,V) represent the query, key, and value matrices, respectively, and d denotes the dimensionality of the matrices. Through this mechanism, the Transformer Block is able to efficiently capture long-range contextual information. Subsequently, the Downsampling Block is employed to downsample the feature maps produced by the Transformer Block. With each Transformer Block, the spatial resolution of the feature maps is reduced by half, extracting increasingly hierarchical features. To capture the change regions between pre- and post-change images, the Difference Module computes the differences in feature maps at each layer, generating multi-scale change information. Specifically, the Difference Module is computed as follows:

$$F_{diff} = BN\left(ReLu\left(C\left(Cat\left(F_{pre}, F_{post}\right)\right)\right)\right) \tag{2}$$

Here, F_{pre}, F_{post} represent the feature maps of the pre-change and post-change images at each layer, Cat denotes tensor concatenation, and C represents a $3 \times 3 \times 3$ convolution operation. These multi-scale change features are then passed to the MLP Decoder, which fuses the extracted features and generates the final change detection output.

2.2. Atrous Spatial Pyramid Pooling

In change detection tasks, differences in object scale across various scenes present a common challenge. Objects may undergo significant changes in size and shape at different time points, necessitating the model's ability to extract spatial features at multiple scales. If these scale differences are not properly addressed, it could result in inaccurate detection outcomes. In the ChangeFormer module, however, cropping the images prior to input disrupts the original spatial structure, hindering the model's ability to learn contextual information and extract multi-scale features. To address this issue, we introduce the ASPP (Atrous Spatial Pyramid Pooling) module, which is designed to enhance

the model's ability to extract multi-scale spatial features, thereby improving the accuracy of change detection.

Given input images $X_1 \in \mathbb{R}^{H\times W\times 3}$ and $X_2 \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the images, respectively, the ASPP consists of five branches, each performing a distinct operation. The first branch applies a 1×1 convolution to produce a feature map, while the second, third, and fourth branches employ dilated convolutions with increasing dilation rates of 6, 12, and 18, respectively, to capture multi-scale spatial information. The fifth branch performs adaptive average pooling, followed by a 1×1 convolution and upsampling to restore the image size, producing another feature map. These five branches generate feature maps with identical spatial dimensions and 256 channels, which are then concatenated along the feature dimension to form a combined feature map containing multi-scale spatial features. A final 1×1 convolution is applied to compress these multi-scale features, resulting in the final feature map F', which serves as the input to the ChangeFormer module.

3. EXPERIMENTS

3.1. Datasets

This study employs two widely used change detection datasets: DSIFN-CD[26] and LEVIR-CD[27]. The DSIFN-CD dataset captures diverse land cover changes, including deforestation, urban expansion, and agricultural transitions, with each remote sensing image having a resolution of 512 × 512 pixels. To facilitate analysis, we cropped the images into 256 × 256 pixel patches and divided them into a training set (14,400 samples), validation set (1,360 samples), and test set (192 samples).

In contrast, the LEVIR-CD dataset specializes in building change detection within urban areas, capturing the emergence, disappearance, and renovation of structures. It consists of 637 pairs of high-resolution images with a spatial resolution of 0.5 meters, each measuring 1024 × 1024 pixels. Using the same patching strategy, we divided the images into 256 × 256 pixel patches, distributing them across the training set (7,120 samples), validation set (1,024 samples), and test set (2,048 samples).

3.2. Implementation Details

In the experimental setup, we utilized the PyTorch deep learning framework to construct and implement the proposed model, leveraging an NVIDIA Quadro RTX 8000 GPU for high-performance computation, ensuring efficient and stable model training. Network parameters were initialized using a random initialization strategy. For data preprocessing, various augmentation techniques were applied to enhance data diversity and improve generalization. These included random image flipping, random scaling (0.8 to 1.2), random cropping, Gaussian blur, and random color adjustments, simulating real-world variations to enhance model adaptability. During training, we employed the Cross-Entropy loss function to quantify the discrepancy between predictions and ground truth labels. The AdamW optimizer was used for weight updates, with a weight decay of 0.01 to mitigate overfitting. The beta parameters were set to (0.9, 0.999), and the initial learning rate of 0.0001 was linearly decayed to zero over 200 epochs to balance rapid convergence and fine-tuned adjustments. A batch size of 16 was chosen to optimize efficiency and memory usage.

To comprehensively evaluate model performance and analyze the impact of the ASPP module, we selected key metrics, including overall accuracy (OA), Intersection over Union (IoU), F1 score , and precision and recall for the change classes.

3.3. Results And Discussion

Table 1. Evaluation of ChangeFormer+ on LEVIR-CD[27] and DSIFN-CD[26] Datasets

Method	LEVIR-CD[27]					DSIFN-CD[26]				
	Precision	Recall	F1	IoU	OA	Precision	Recall	F1	IoU	OA
BIT[28]	89.58	82.30	85.79	75.11	98.61	84.99	73.79	78.99	65.28	93.33
DTCDSCN[29]	90.22	82.78	86.34	75.96	98.67	86.42	80.76	83.49	71.66	94.57
ChangeFormer[24]	92.13	88.79	90.43	82.53	99.04	87.10	84.89	85.98	75.41	95.30
ChangeFormer+(ours)	92.38	89.68	91.01	83.50	99.10	87.42	86.48	86.95	76.91	95.59

Table 1 presents the performance of all comparison methods on the LEVIR-CD[27] and DSIFN-CD[26] test sets. ChangeFormer[24] outperforms BIT[28] and DTCDSCN[29] by leveraging the advanced SegFormer backbone, which enhances feature extraction and contextual modeling. Compared with ChangeFormer, our method further improves the detection performance. This is because our method integrates ASPP into ChangeFormer, addressing the limitation of Transformers in effectively extracting spatial features. Specifically, on LEVIR-CD, ChangeFormer+ improves F1 by 0.58%, IoU by 0.97%, and OA by 0.06%, while on DSIFN-CD, F1 increases by 0.97%, IoU by 1.5%, and OA by 0.29%.

4. SUMMARY

In this paper, we address the challenge of detecting large-scale changes in high-resolution satellite imagery, where traditional convolutional networks (CNNs) struggle to capture long-range contextual information and multi-scale features. To overcome these limitations, we propose a novel model, ChangeFormer+, which integrates the Transformer-based architecture with the ASPP module. This combination enhances the model's ability to capture both global contextual relationships and multi-scale spatial features, leading to improved detection accuracy. Our experimental results on LEVIR-CD and DSIFN-CD datasets show that ChangeFormer+ outperforms state-of-the-art methods, achieving significant improvements in key metrics, including F1 score, IoU, and overall accuracy (OA).

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by the National College Students' Innovation and Entrepreneurship Training Program under Grant 202411287030Z.

REFERENCES

[1] Sun, J., Zhao, M., & Hao, X. (2024). A review of remote sensing image change detection methods. Computer Engineering and Applications, 20, 30-48.

[2] Dai, X., & Khorram, S. (1997). Quantification of the impact of misregistration on the accuracy of remotely sensed change detection. In IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing—A Scientific Vision for Sustainable Development (Vol. 4, pp. 1763–1765). IEEE. https://doi.org/10.1109/IGARSS.1997.609062.

[3] Peng, X., Zhong, R., Li, Z., & Li, Q. (2021). Optical remote sensing image change detection based on attention mechanism and image difference. IEEE Transactions on Geoscience and Remote Sensing, 59(9), 7296–7307. https://doi.org/10.1109/TGRS.2020.3033009.

[4] Luppino, L. T., Bianchi, F. M., Moser, G., & Anfinsen, S. N. (2018). Remote sensing image regression for heterogeneous change detection. In 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1–6). IEEE. https://doi.org/10.1109/MLSP.2018.8517033.

[5] Alard, C., & Lupton, R. H. (1998). A method for optimal image subtraction. The Astrophysical Journal, 503(1), 325.

[6] Kauth, R. J., & Thomas, G. S. (1976). The tasseled cap—A graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT. LARS Symposia, 159.

[7] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

[8] Johnson, R. D., & Kasischke, E. S. (1998). Change vector analysis: A technique for the multispectral monitoring of land cover and condition. International Journal of Remote Sensing, 19(3), 411–426. https://doi.org/ 10.1080/ 014311698216062.

[9] Lombardo, P., & Pellizzeri, T. M. (2002). Maximum likelihood signal processing techniques for detecting step changes in multitemporal SAR images. IEEE Transactions on Geoscience and Remote Sensing, 40(4), 853–870. https://doi.org/10.1109/TGRS.2002.1006363.

[10] Cao, G., Li, Y., Liu, Y., & Shang, Y. (2014). Automatic change detection in high-resolution remote-sensing images by means of level set evolution and support vector machine classification. International Journal of Remote Sensing, 35(16), 6255–6270. https://doi.org/10.1080/01431161.2014.951740.

[11] Feng, W., Sui, H., Tu, J., Huang, W., & Sun, K. (2018). A novel change detection approach based on visual saliency and random forest from multi-temporal high-resolution remote-sensing images. International Journal of Remote Sensing, 39(22), 7998-8021.

[12] Dai, X. L., & Khorram, S. (1999). Remotely sensed change detection based on artificial neural networks. Photogrammetric Engineering and Remote Sensing, 65, 1187-1194.

[13] Stent, S., Gherardi, R., Stenger, B., & Cipolla, R. (2015, September). Detecting change for multi-view, long-term surface inspection. In BMVC (pp. 127-1).

[14] Daudt, R. C., Le Saux, B., & Boulch, A. (2018, October). Fully convolutional siamese networks for change detection. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 4063-4067). IEEE.

[15] Gu, L., Xu, S., & Zhu, L. (2020). Building change detection in remote sensing images based on FlowS-Unet. Acta Automatica Sinica, 46(6), 1291-1300.

[16] Peng, D., Zhang, Y., & Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing, 11(11), 1382.

[17] Yu, X., Fan, J., Chen, J., Zhang, P., Zhou, Y., & Han, L. (2021). NestNet: A multiscale convolutional neural network for remote sensing image change detection. International Journal of Remote Sensing, 42(13), 4898-4921.

[18] Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., & Liu, G. (2020). A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 166, 183-200.

[19] Wang, D., Chen, X., Jiang, M., Du, S., Xu, B., & Wang, J. (2021). ADS-Net: An attention-based deeply supervised network for remote sensing image change detection. International Journal of Applied Earth Observation and Geoinformation, 101, 102348.

[20] Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., & Zhang, L. (2021). A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-16.

[21] Cheng, G., Wang, G., & Han, J. (2022). ISNet: Towards improving separability for remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-11.

[22] 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.

[23] Chen, H., Qi, Z., & Shi, Z. (2021). Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

[24] Bandara, W. G. C., & Patel, V. M. (2022). Transformer-based Siamese change detection network. In IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium (pp. 207-210). IEEE. https://doi. org/ 10. 1109/ IGARSS46834.2022.9883686.

[25] Lian, X., Pang, Y., Han, J., & Pan, J. (2021). Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation. Pattern Recognition, 110, 107622.

[26] Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., & Liu, G. (2020). A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 166, 183-200.

[27] 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.

[28] Chen, H., Qi, Z., & Shi, Z. (2021). Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

[29] 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.