University of Tehran

Earth Observation and Geomatics Engineering

Online ISSN: 2588-4360

Homepage: https://eoge.ut.ac.ir/

ChangeCoTNet: Building Change Detection by Contextual Transformer Deep Network

Akram Eftekhari [1]✉, **Farhad Samadzadegan** [2], and **Farzaneh Dadrass Javan** [3]

1. Corresponding author, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran. E-mail: a.eftekhari@ut.ac.ir@ut.ac.ir
2. School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran. E-mail: samadz@ut.ac.ir @ut.ac.ir
3. Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 NB Enschede, the Netherlands. E-mail: f.dadrassjavan@utwente.nl

Article Info

Article type:
Research Article

Article history:
Received 2024-06-23
Received in revised form 2025-01-24
Accepted 2025-02-19
Available online 13 May 2025

Keywords:
Remote Sensing,
Building Change Detection,
Contextual Transformer,
Dual Branch Network,
Self-Attention Mechanisms.

ABSTRACT

Building change detection (BCD) is a critical task in remote sensing, with applications in urban management and disaster assessment. However, achieving high accuracy in high-resolution BCD remains challenging due to the complexity of urban scenes. In this study, we propose ChangeCoTNet, a novel dual-branch deep learning model that integrates Contextual Transformer (CoT) blocks in the encoder and a Convolutional Neural Network (CNN) in the decoder. The CoT blocks enable the extraction of both static and dynamic contextual representations, while the Channel Attention Block (CAB) enhances discriminative feature extraction. The proposed model was implemented and evaluated on the LEVIR-CD and 2DCD datasets using a PyTorch backend. Experimental results demonstrate that ChangeCoTNet outperforms state-of-the-art methods, achieving F1-score improvements of 1.1% and 1.9% for the respective datasets. These results validate the effectiveness and efficiency of the proposed model in detecting changes with high precision and recall, making it a valuable tool for real-world applications.

Cite this article: Eftekhari, A., Samadzadegan, F., & Dadrass Javan, F. (2025). ChangeCoTNet: Building Change Detection by Contextual Transformer Deep Network. *Earth Observation and Geomatics Engineering*, Volume 8 (Issue 1), Pages 1-12. http//doi.org/10.22059/eoge.2025.378306.1153

Publisher: University of Tehran.
DOI: http//doi.org/10.22059/eoge.2025.378306.1153

1. Introduction

Building change detection (BCD) is one of the important and significant research topics in remote sensing. It entails recognizing and measuring spatial changes in building using two or more co-registered satellite or aerial images (R. Qin et al., 2016). BCD is a critical component of urban planning and development (Stilla & Xu, 2023). It involves using remote sensing technology to identify and track changes in the built environment over time. This information is essential for a range of applications, including disaster response (Zheng et al., 2021), land use planning (Chughtai et al., 2021), and environmental monitoring (Padró et al., 2019). BCD can help identify areas that are experiencing rapid growth or decline, as well as areas that are at risk of natural disasters or other hazards (Mazzanti et al., 2022). Additionally, BCD is a valuable tool for monitoring the impacts of human activity on the environment, such as deforestation or urbanization (Alzu'bi & Alsmadi, 2022). In recent years, a new breed of very high-resolution (VHR) satellites has been put into orbit, equipped with the ability to capture images with a resolution of 1 meter or greater (Wen et al., 2021). VHR images can provide more detailed information on building footprints, roof shapes, and other physical characteristics of structures (Y. Qin et al., 2019). This information is useful for detecting and quantifying changes in the built environment, such as new construction or demolition of buildings (Yan et al., 2022). The correct selection of data, the type of change detection algorithm, and the extraction and selection of features that is suitable for the size, shape, texture, and spectral signature of the building are among the most important effective factors in the accurate identification of building changes (Bai et al., 2022).

In remote sensing, changes are primarily identified through pixel-based and object-based methods (Wen et al., 2021). Pixel-based methods compare individual pixels in the images and identify changes based on differences in their spectral values (Hussain et al., 2013). Despite its simplicity, this approach does not consider the spatial context information which leads to the presence of a considerable amount of salt and pepper noise in the resulting images, especially when using high-resolution (HR) and VHR images. Thus, it is more suitable to apply this approach to identify changes in images of moderate resolution (Tewkesbury et al., 2015). Object-based methods, on the other hand, consider groups of pixels that form objects or features and compare them to identify changes (Zhang et al., 2018). Considering the level of detail in HR and VHR images, the use of object-based methods is more suitable for detecting changes due to the extraction of rich spectral, texture, structural, and geometric features. Features in object-based methods can be extracted manually and automatically. While manual methods are simpler, they are very time-consuming and do not result in acceptable outcomes. However, automatic methods based on machine learning algorithms can extract accurate and suitable features (Khelifi & Mignotte, 2020). Deep learning (DL) is one of the machine learning methods have become popular for feature extraction from images due to their ability to automatically learn and extract features from raw data (Ball et al., 2017). CNNs are a common type of deep learning model used for image feature extraction, as they can learn hierarchical representations of features that capture both low-level and high-level details in an image (Z. Li et al., 2022). These learned features can then be used for a variety of tasks, including object recognition (Groener et al., 2019), image classification (Zhao et al., 2022), and change detection (Khelifi & Mignotte, 2020; D. Peng et al., 2019).

In recent years, remote sensing image change detection (CD) methods have got extraordinary advancement based on CNNs. Supervised CD methods that have been developed primarily depend on CNN-based architectures to extract high-level semantic features that determine the relevant changes between each temporal image (Khusni et al., 2020). While convolution kernels are excellent at extracting local features from an image by processing a small neighborhood of pixels at a time, they struggle to capture long-range dependencies between image features (Vaswani et al., 2017). Convolutional operations are a key component in the feature-based CD and the image-level fusion- and segmentation-based CD methods, which are currently the dominant approaches for detecting changes in remote sensing images. As a result, these methods excel in capturing local image content relationships, but they fall short in representing long-range global interactions (M. H. Guo et al., 2022). Utilizing the global contextual relationships within an image to compute individual pixel values leads to the generation of more resilient feature maps, capturing extensive, long-range global interactions (Chen et al., 2021). However, models neglecting these interactions might not deliver optimal performance in situations where understanding global contextual information is pivotal for accurate localization, especially in cases involving extensive land-use changes. To address this constraint, it is imperative to incorporate a nonlocal self-attention mechanism, enabling the model to grasp long-range global relationship details effectively. This technique will enable the model to better capture the dependencies between image features, and thus improve its ability to recognize complex patterns and structures in the image data (G. Wang et al., 2022).

A transformer is a unique form of self-attention mechanism, which calculates global contextual relationships to automatically identify significant information locations and perform adaptive weighting of inputs (Lan et al., 2023). The transformer architecture has achieved remarkable success in natural language processing tasks (Gillioz et al., 2020) and speech recognition (Y. Wang et al., 2020). Given the importance of global information in vision tasks, adapting the transformer architecture could potentially address a limitation of CNNs, which typically increase their receptive field by adding more layers. One of the most widely used transformers in computer vision and image processing is the vision transformer (ViT). ViT was introduced by (Dosovitskiy et al., 2021). In recent years, ViT have been applied to various tasks in the field of remote sensing as well, including land cover classification, object detection and hyperspectral image processing (M. Li et al., 2023; Y. Li et al., 2022; H. Wang et al., 2022). Also, less research has been done in the field of using transformers for CD. SiamixFormer (Mohammadian & Ghaderi, 2023) is a Siamese based network that utilizes hierarchical transformer architecture with two encoders and temporal transformers for feature fusion, which helps maintain large receptive fields. However, the transformer-based mechanism employed by this model is complicated and has a high computational cost. Another effective approach is to combine CNNs with transformers. (Chen et al., 2021) using a combination of CNNs and transformer-based architecture to model the spatial and temporal relationships between the image features and identify regions that have changed between the two images. Bit-CD employs a transformer-based decoder network to enrich the contextual information derived from Conv-Net features. This enables the model to capitalize on the efficient training capabilities of convolutional networks, while simultaneously harnessing the benefits of capturing extensive dependencies within the input data (Chen et al., 2021). Generally, models that combine CNNs with transformers have a lower computational cost than models that exclusively rely on transformers in both the encoder and decoder. Nonetheless, these models still require significant computational time. In this article, we have used a structure based on hybrid Transformer-Conv-Net model, which is a light network, in order to detect changes.

The conventional self-attention blocks used in transformer models primarily rely on the isolated interaction between query-key pairs to calculate the attention matrix, disregarding the abundant contextual information shared among adjacent keys. In the work by (Y. Li et al., 2023), they introduced Contextual Transformer (CoT) blocks, which adeptly utilize contextual information among input keys. This approach guides the learning process of a dynamic attention matrix, notably improving the model's capability to represent visual information. This design integrates contextual analysis among keys and self-attention learning across a two-dimensional feature map within a unified structure, eliminating the need for an additional branch dedicated to context mining. Based on this idea, we introduced ChangeCoTNet that is a dual-branch CD network in which CoT is used in down-sampling instead of 3×3 convolution in order to extract global contextual information.

The primary contributions of this study can be outlined as follows:

1. Introducing ChangeCoTNet: This novel approach integrates Contextual Transformers and CNNs for high-resolution remote sensing CD.
2. Developing Contextual Representations: A method is developed to create both static and dynamic contextual representations using a CoT block. This enhances the detection process by modeling dense pixel relations.
3. Enhancing Feature Extraction: The model employs a multi-head attention matrix and a channel attention block to extract discriminative features, thereby improving accuracy in identifying changes in complex scenes.

2. Methodology

Given the need to extract global information for better determine building changes, we have proposed ChangeCoTNet architecture. It is a network based on the self-attention mechanism with additional exploitation of contextual information among input keys. The visual representation of our proposed approach is depicted in Figure 1. In the following subsection, we will discuss the details of each component of the proposed architecture.

2.1. The overview of ChangeCoTNet

The proposed ChangeCoTNet technique involves a dual-branch deep learning network that is specifically tailored to extract profound characteristics from two multi-temporal images captured under distinct exposure factors. The dual-branch network incorporates inputs on both sides, connected by a single expansive path in the middle. In the presented network, CoT blocks are employed in place of conventional 3×3 convolutions. Therefore, every contraction side consists of four encoding levels, encompassing CoT, batch normalization, and dropout layers. At the end of the encoding stage, the Euclidean function calculates the dissimilarity between features on both sides of the network. Subsequently, a Channel Attention Block (CAB) is utilized for extracting distinct discriminator characteristics, enabling a more effective differentiation between modified and unmodified regions. In the expansion path, the deconvolution process is applied, wherein low-level features extracted by CoT are symmetrically copied from both sides of the network and then merged with high-level information. The deconvolution step ensures the independence of network weights on both sides. During backpropagation, the generated loss values are concurrently propagated to both sides of the network, leading to simultaneous updates of network weights. This approach enables a nonlinear simulation of various data sources and diverse conditions concurrently.

2.2. Contextual Transformer Block

Self-attention is a mechanism used in the Transformer model architecture to capture the dependencies between different positions (or tokens) within a sequence (L. Wang et al., 2022). The self-attention mechanism requires three kinds of representations for each token: key, query, and value. These representations are obtained by linearly transforming the input embedding using learned weight matrices. The resulting weighted sum is the output representation for the token. It captures the contextual information from the entire sequence, as it combines information from all tokens in a weighted manner. This enables the model to consider the relationships and dependencies between tokens when generating representations, facilitating better understanding and capturing of long-range dependencies in the data (Anonymous, 2021).

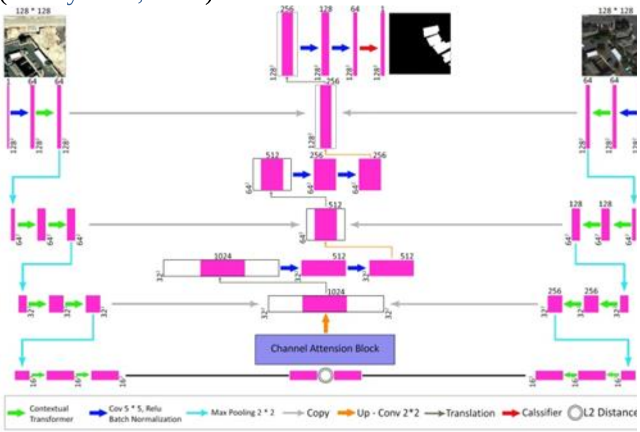

Figure 1. ChangeCOTNet - A Dual Branch Network for BCD based on CoT and CAB.

However, with all the advantages of self-attention, there are some limitations for it: it learns pairwise relationships independently for each query-key pair without considering the rich contextual information in between. As a result, its ability to learn self-attention across 2D feature maps for visual representation learning is significantly constrained. In order to alleviate these problems, CoT block has been introduced by (Y. Li et al., 2023). In this block, whose architecture is also given in Figure 2, contextual information mining is given along with self-attention in an integrated new architecture. This approach is centered on maximizing the utilization of contextual information between adjacent keys, aiming to enhance the efficiency of self-attention learning and augment the representational capacity of the resulting combined feature map.

To illustrate, let's consider a 2D feature map I with dimensions h × w × c (h: height, w: width, c: channel) as an input. For this case, the keys, queries, and values are defined as follows: K = I, Q = I, and V = IW$_v$, respectively. Rather than using a 1×1 convolution, which is commonly employed in standard self-attention, the CoT block takes a different approach. It initiates with a group convolution of size k×k to process neighboring keys within the same grid, allowing for contextualization of individual key representations in a spatially aware manner. The learned contextualized keys, denoted as K^1 with dimensions R$^{h \times w \times c}$, inherently capture the unchanging contextual details within adjacent keys. K^1 was considered as the fixed contextual interpretation of the input X. Subsequently, by combining contextualized keys (K^1) with queries (Q) and applying consecutive 1×1 convolutions, the attention matrix is calculated. The first convolution W_θ incorporates a ReLU activation function, while the second convolution W_δ does not have an activation function.

$$A = [K^1, Q]W_\theta W_\delta \tag{1}$$

Certainly, within every attention head, the system calculates the attention matrix for precise spatial positions in matrix A is calculated by analyzing both the query feature and the contextualized key feature. This approach improves self-attention learning by leveraging the contextual information provided by the mined static context K^1, rather than relying solely on isolated query-key pairs. Subsequently, utilizing the contextualized attention matrix A, we generate the attended feature map K^2 by combining all the values V, employing a procedure akin to conventional self-attention mechanisms.

$$K^2 = V * A \tag{2}$$

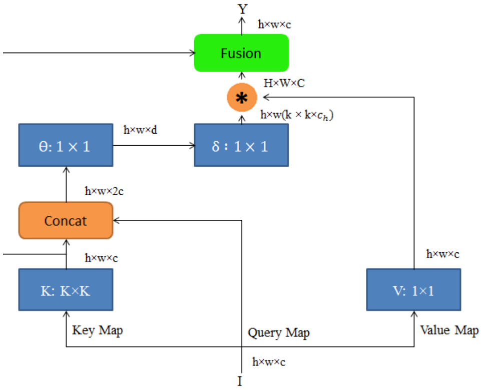

Figure 2. A view of CoT block

2.3. Channel Attention Block

CAB refers to a specific architectural component that incorporates the channel attention mechanism into a deep learning model (M. H. Guo et al., 2022). Within deep networks, distinct channels in feature maps usually represent different objects. Channel attention modifies the importance of each channel, serving as a method to decide where to concentrate when identifying an object (Eftekhari et al., 2023). In our proposed network, we have applied CAB to the feature distance function of both sides of the network. As a result, channels that capture significant changes are emphasized, leading to improved accuracy in identifying and discerning the changes. Figure 3 illustrates CAB, which avoids using convolution for generating new features. Instead, the input feature D is reshaped from C×H×W to C×N size, where N represents the product of H and W. This reshaped D is then multiplied by its transpose to construct the channel-wise attention mechanism by size N×N. To apply attention, a Softmax operation is employed using Formula (3),

$$\boldsymbol{Dx_{ij} = softmax(\frac{D_i^T D_j}{\sqrt{C}})} \tag{3}$$

Here Dx_{ji} measures the influence of the jth channel on the ith channel, with higher values signifying a more robust connection between them. The reshaped D is subsequently multiplied by C×N with Dx_{ji}, yielding the resultant output. To this outcome, a coefficient δ is added, as demonstrated in Equation (4), to achieve the final output.

$$Dca_j = \delta \sum_{i=1}^{C} (Dc_{ji} Di) + D_j \tag{4}$$

The coefficient δ is initially set to 0 and is determined during training. The ultimate feature of each channel is a weighted combination of all channels and the initial feature, as outlined in the previously mentioned equation.

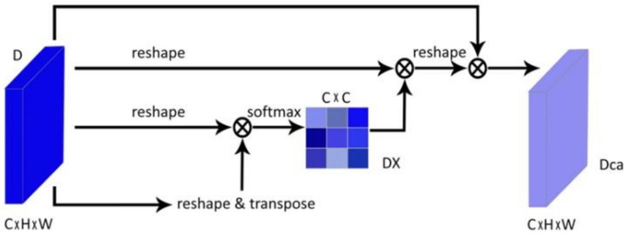

Figure 3. The general structure of CAB is outlined as follows: D represents the input feature; Dx corresponds to the channel attention module, while Dca signifies the outcome of channel attention operations implemented on input feature D.

2.4. Accuracy Assessment

Precision-recall serves as a valuable metric for predicting outcomes in scenarios where class imbalances are prominent (Fang et al., 2021). In the context of CD problems, where the number of altered points is typically significantly lower than the unchanged points, and an inherent imbalance exists between these two types of data, employing precision-recall metrics is fitting. Hence, the evaluation of 2D results entails utilizing precision (Pr), recall (Re), and F1-score (F1). The mathematical expressions for these measures can be found in the subsequent equations (Fang et al., 2021).

$$Pr = \frac{TP}{TP + FP} \tag{5}$$

$$Re = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2PrRe}{Pr + Re} \tag{7}$$

In the CD domain, TP (true positive) signifies the accurate identification of altered pixels, FP (false positive) refers to erroneously identified changed pixels, and FN (false negative) represents the changed pixels mistakenly classified as unchanged. In the context of CD, a lower rate of false detections enhances precision, while a reduced number of missed predictable outcomes increases recall. The F1 measure serves as a comprehensive criterion for evaluating results, with higher values indicating more desirable and appropriate fitting outcomes.

3. Experiments and Results

3.1. Dataset

A) LEVIR-CD Dataset: LEVIR-CD, a freely available dataset for BCD, was introduced by (Chen & Shi, 2020). It comprises 637 VHR images obtained from Google Earth, with a pixel resolution of 0.5 meters. These images, captured between 2002 and 2018, were sourced from various cities in Texas, USA. Notably, the dataset encompasses images captured during different seasons, clearly exhibiting changes in brightness between the image pairs. This comprehensive database effectively models real-world changes, minimizing the influence of spurious changes caused by factors like seasonal variations. The dataset includes positive building changes, encompassing alterations in land cover like the transition from soil, vegetation, or structures under construction to newly erected buildings, as well as instances of building removals. In total, the labeled LEVIR-CD dataset comprises 31,333 instances of distinct building changes, averaging approximately 50 changed buildings per 1024 × 1024 image. Most of the changes correspond to the construction of new buildings, occupying roughly 987 pixels per image.

B) 3DCD Dataset: The 3DCD dataset, introduced by Valerio Marsoccia et al., represents a pioneering resource in the domain of two-dimensional (2D) and three-dimensional (3D) change detection (Marsocci et al., 2023). This dataset comprises a total of 472 image pairs, meticulously extracted from optical orthophotos acquired during distinct aerial surveys conducted in the years 2010 and 2017. Each image pair is characterized by optical imagery complemented by 2D CD maps. Of particular significance to our study, we focus on the utilization of the 2D data component. Furthermore, accompanying these elements are corresponding 3D CD maps that encapsulate elevation alterations. For illustrative purposes, a selection of 2D sample images from the dataset is thoughtfully presented in Figure 4.

The dataset encompasses the historical and downtown areas of Valladolid in Spain, along with nearby commercial districts, excluding agricultural regions due to minimal elevation variations. The 2D CD maps are binary, categorizing areas into two classes: no change (ΔH= 0) and changes resulting from construction (ΔH > 0) or demolition (ΔH < 0) of human-made structures such as buildings, roads, and bridges. Only significant elevation changes affecting artificial structures were included in the 2D CD maps, while elevation changes below one meter were set to zero as negligible. It should be mentioned that the 2D images have 400 × 400 pixels with a 0.5 meter GSD. In this article, due to the use of 2D data, we call it 2DCD from now on.

3.2. Technical Implementation

The implementation of the proposed ChangeCoTNet model involves a systematic process that integrates CoT blocks and a CAB to enhance BCD. The following steps outline the detailed implementation process:

1. Input Data Processing
    - Bi-temporal Images: The input consists of two co-registered high-resolution images from different time periods. These images are normalized and resized to a fixed size of 128×128 pixels for computational efficiency.
    - Data Augmentation: To address limited training data and improve model generalization, augmentation techniques such as random horizontal flipping, rotation, and scaling are applied.
2. Encoder: Contextual Transformer Blocks
    - The encoder processes the input images through four hierarchical CoT blocks. Unlike traditional convolutions, CoT blocks are designed to model both static and dynamic contextual representations by capturing the relationships among adjacent keys.
    - Each CoT block includes:
        - Static Contextual Representation: A group convolution applied to neighboring keys to extract spatially aware features.
        - Dynamic Contextual Representation: A multi-head attention matrix that dynamically combines queries and contextualized keys to enhance feature representation.
    - These blocks enable the model to capture both local and long-range dependencies, which are critical for identifying subtle and complex changes.
3. Feature Comparison and Discriminative Enhancement
    - At the end of the encoding stage, features from the two temporal images are compared using the Euclidean distance function to generate a difference map.
    - The CAB:
        - This module analyzes the generated difference map to emphasize channels that capture significant changes.
        - The CAB uses a channel-wise attention mechanism, which assigns importance weights to each channel, ensuring the model focuses on the most relevant features.
4. Decoder and Output Generation
    - The decoder reconstructs the feature map into a binary change map, indicating changed and unchanged regions.
    - Low-level features from the encoder are symmetrically copied and merged with high-level features in the decoder to enhance the accuracy of localization.
    - The output is a high-resolution change map, where white pixels represent changed areas and black pixels represent unchanged regions.
5. Training Details
    - Optimizer: The Adam optimizer is employed, with a learning rate initialized at 1e-1 and gradually decreased to 1e-4.
    - Loss Function: A weighted binary cross-entropy loss function is used to address the class imbalance between changed and unchanged pixels.
    - Dataset Split:
        - LEVIR-CD dataset: 70% training, 20% testing, 10% validation.
        - 2DCD dataset: 68% training, 23% testing, 9% validation.
    - Batch Size and Epochs: A batch size of 10 is used, and the model is trained for 50 epochs to ensure convergence.

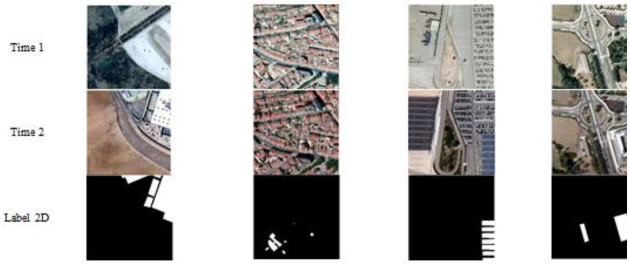

Figure 4. The samples of 2DCD dataset (2DCD)

3.3. Experimental Results

We devised several experiments to demonstrate the impact of the suggested CoT-based approach. In the dual-branch architecture depicted in Figure 1, we initially employed a CNN for the encoding phase. The outcomes of this identical network were obtained by incorporating channel attention during the decoding stage. Subsequently, we present the outcomes of the dual-branch network, contrasting the use of CoT for encoding instead of CNN, both with and without channel attention.

3.3.1. Implementation Results of LEVIR-CD

The LEVIR-CD dataset was utilized to implement the proposed approach in a sequential manner. The outcomes are presented in Table 1. In order to more clearly highlight the efficacy of the proposed method, a systematic approach was adopted. Initially, a basic dual model grounded in CNN architecture was employed. Subsequently, a CAB was incorporated into the decoding segment. Progressing further, the conventional use of CNNs in the dual network was replaced by CoT. Finally, our proposed method (ChangeCoTNet) involved implementing a decoding component founded on both CoT and a CAB.

As depicted in **Error! Reference source not found.**, the integration of CAB led to a 1.7% enhancement in F1 within the context of the simple dual-branch network, and a 1.2%

increase within the CoT network. Meanwhile, employing the CoT network itself yielded a notable 3% increment in the F1. Ultimately, the innovative network architecture that combines CoT and CAB demonstrated substantial advancements, resulting in a remarkable 4.7% enhancement in precision, a notable 3.8% elevation in recall, and a significant 4.2% improvement in the F1 metric.

Figure 5 also presents the visual outcomes of progressively applying the proposed model to the LEVIR-CD dataset. In lines 4 to 7 of the Figure 5, we present the outcomes stemming from a stepwise implementation approach, particularly in regions exhibiting changes of varying sizes. The findings underscore the limitations of employing a dual-branch network modeled CNN as it has not comprehensively captured the changes. However, when employing the CAB model, the incorporation of distinctive feature extraction mechanisms notably enhances the fidelity of CD. Furthermore, the application of the CoT model, as depicted in line 6, and its fusion with CAB in line 7, demonstrates that integrating contextual feature extraction within the network architecture significantly improves the capacity to identify even subtle and minute changes with a high degree of accuracy.

Table 1. An investigation of the proposed method through an ablation study on the LEVIR-CD validation set

Method	Pr (%)	Re (%)	F1 (%)
dual-branch network by CNN 3*3	88.42	86.49	87.44
dual-branch by CNN 3*3 +CAB	90.27	87.62	88.93
dual-branch by CoT	91.30	88.86	90.06
dual-branch by CoT +CAB (ChangeCoTNet)	**92.53**	**89.79**	**91.14**

3.3.2. Implementation Results of 2DCD dataset

The results of applying our proposed methodology to the 2DCD dataset are presented in detail in Table 2. This table shows that our proposed method, ChangeCoTNet, achieved highly promising performance. This method exhibits superlative performance, achieving the highest levels of Precision, Recall, and F1-score. It is worth noting that the incorporation of CAB consistently contributes to the augmentation of model performance. This augmentation is palpably demonstrated in the enhanced metrics discernible in both the straightforward dual-branch network and the 'ChangeCoTNet' methods.

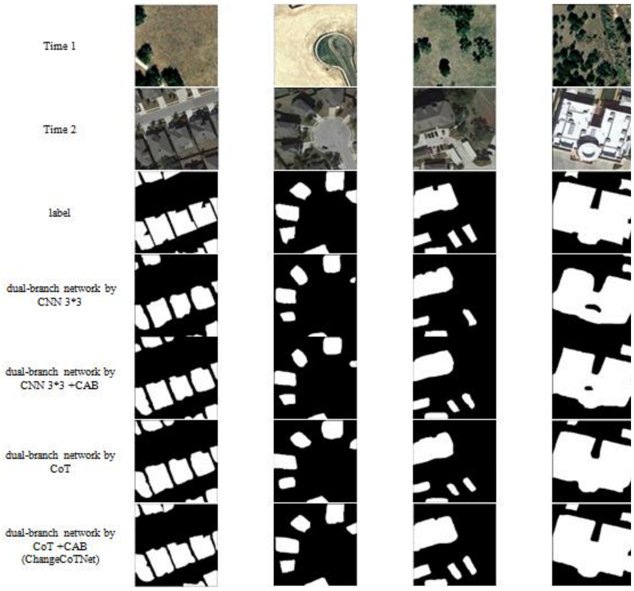
Time 1
Time 2
label
dual-branch network by CNN 3*3
dual-branch network by CNN 3*3 +CAB
dual-branch network by CoT
dual-branch network by CoT +CAB (ChangeCoTNet)

Figure 5. The visual outcomes of progressively applying the proposed model to the LEVIR-CD dataset. White signifies the changed region, while black signifies the unchanged region

Simultaneously, when considering two-dimensional data, the employment of the CoT network in isolation yields a noteworthy 8.2% escalation in the F1 score. However, the introduction of our pioneering network architecture that amalgamates CoT with CAB reveals substantial strides forward. This architectural innovation culminates in a remarkable 11.9% enhancement in the F1 metric within the two-dimensional data subset of the 3DCD dataset. It is imperative to note that the results obtained on the 2D data segment of the 3DCD dataset exhibit lower performance compared to the LEVIR-CD dataset. This discrepancy can primarily be attributed to the scarcity of training data within the 2D dataset, which impedes the model's capacity to generalize effectively onto the evaluation data.

Figure 6 also depicts the visual outcomes achieved through the progressive application of the proposed model to the aforementioned dataset. Within the Figure 6, specifically in lines 4 to 7, we illustrate the results obtained through a stepwise implementation strategy, focusing on regions that show building changes of different sizes: small (first column), medium (second column), and large (third and fourth column. These observations highlight the inherent limitations of employing a dual-branch network based on CNNs, as it fails to comprehensively capture the changes present in the dataset.

However, when utilizing the CAB model, a framework designed with distinctive feature extraction mechanisms, a notable enhancement in CD fidelity is observed. Additionally, the integration of the CoT model, as illustrated in line 6, and its fusion with CAB in line 7, exemplifies the substantial improvements achieved by incorporating contextual feature extraction within the network architecture. This enhancement significantly enhances the model's ability to identify even subtle and minute changes with a high degree of accuracy.

Table 2. An investigation of the proposed method through an ablation study on the 2DCD validation set

Method	Pr (%)	Re (%)	F1 (%)
dual-branch network by CNN 3*3	80.16	64.57	71.53
dual-branch by CNN 3*3 +CAB	84.51	66.62	74.51
dual-branch by CoT	87.04	69.94	77.56
dual-branch by CoT +CAB (ChangeCoTNet)	**88.47**	**73.09**	**80.05**

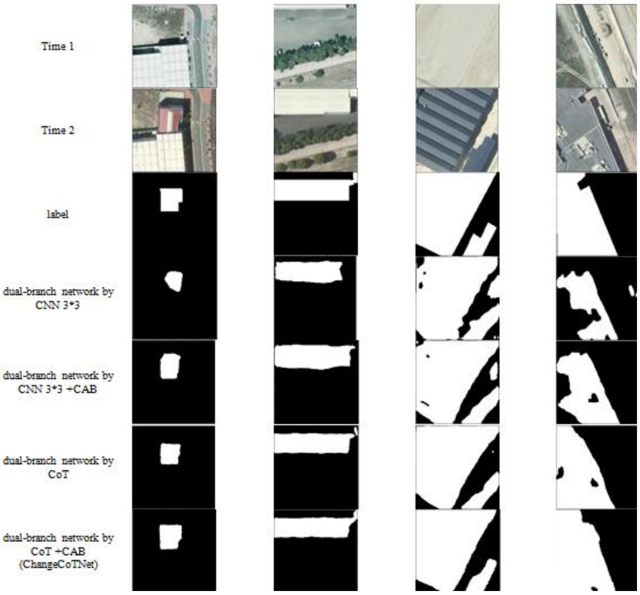
Time 1
Time 2
label
dual-branch network by CNN 3*3
dual-branch network by CNN 3*3 +CAB
dual-branch network by CoT
dual-branch network by CoT +CAB (ChangeCoTNet)

Figure 6. The visual outcomes of progressively applying the proposed model to the 2DCD dataset. White signifies the changed region, while black signifies the unchanged region

4. Discussion

In this section, we will discuss the results by comparing our proposed network with state-of-the-art (SOTA) methods. The evaluation of our method's performance incorporates various contemporary techniques. For instance, The STANet, introduced by the developers of the LEVIR-CD dataset in 2020, utilizes the self-attention mechanism to capture spatial-temporal dependencies (Chen & Shi, 2020). DDCNN, as introduced by Peng et al. in 2020, utilizes dense attention and several up-sample attention components for dual-temporal image processing. Additionally, it incorporates a DE unit to enhance network efficiency (X. Peng et al., 2020). Additionally, AGCDetNet (Song & Jiang, 2021) integrates spatial attention alongside a module featuring channel-wise attention-guided interference filtering. This approach enhances features at multiple levels and context across various scales, enhancing attention-guided CD capabilities.

In transformer-based methods, SiamixFormer (Mohammadian & Ghaderi, 2023) is a novel Siamese model designed for CD. It operates by employing a dual-encoder setup within hierarchical transformer structure in the encoding stage. The inclusion of temporal transformers in feature fusion offers the added benefit of effectively maintaining the expansive receptive fields established by transformer encoders. Temporal transformer output subsequently passes through a simple MLP decoder at each stage. BIT (Chen et al., 2022) designed to effectively model spatial-temporal contexts. By representing bitemporal images as semantic tokens, the approach employs a transformer encoder to capture context in a condensed token-based space-time representation. (Q. Guo et al., 2022) presented an innovative module named iterative difference-enhanced transformers (IDET). This module stems from a novel perspective in CD, emphasizing the enhancement of feature differences to accentuate changes and diminish unchanged regions. (Bandara & Patel, 2022) introduced the ChangeFormer network, which is a hierarchical transformer encoder with a Multi-Layer Perception (MLP) decoder in a Siamese network. This integration enables the efficient capture of multi-scale long-range details necessary for accurate Change Detection (CD).

To ensure a fair comparison, all techniques were instantiated using the hyper parameters outlined in **Error! Reference source not found.**. The input data dimensions were confined to 128 × 128 due to graphic card constraints. These methodologies were executed on the LEVIR-CD and 2DCD datasets.

Table 3. Parameters employed for executing the SOTA techniques

Method	Train crop size	Batch size	Optimization Algorithm	No. of epoch
STANet	128 × 128	10	Adam	50
DDCNN	128 × 128	10	Adam	50
AGCDetNet	128 × 128	10	Adam	50
SiamixFormer	128 × 128	10	Adam	50

The comparison results of the implementation of the proposed model on the SOTA methods mentioned above on the LEVIR and 2DCD data sets are given in **Error! Reference source not found.**. ChangeCoTNet consistently outperforms other methods on both datasets in terms of Precision, Recall, and F1-Score. It especially shows significant improvement over other methods on the 2DCD dataset. Hence, there has been an increase in the F1 score by 1.1% in the LEVIR-CD dataset and 1.9% in the 2DCD dataset. Particularly noteworthy is the exceptional precision observed in the 2DCD dataset, showing a remarkable improvement of 2.42% compared to the performance achieved by previous optimal methods. A high Pr signifies the extraction of localized information through dense connections among multiscale features. Furthermore, the Re parameter can compete with SOTA methods and even in 2DCD data, it has increased by about 1.2 %. This rise underscores the effective utilization of contextual information within neighboring keys, enabling the model to learn to focus on itself. Additionally, it highlights the strategic incorporation of channel attention mechanisms, facilitating the extraction of distinctive and comprehensive long-range features.

To effectively showcase the performance of the proposed methodology, visual outcomes for both LEVIR-CD and 2DCD datasets are juxtaposed with those of SOTA methods in Figure 7. In these visual representations, correctly identified changes (TP) are depicted in white, while accurately detected stability (TN) is represented in black. Pixels erroneously identified as changed (FP) are denoted in red, while pixels inaccurately identified as unchanged (FN) are highlighted in blue. In the illustrated examples, certain methods displayed limitations in capturing all changes comprehensively. Specifically, STANet and DDCNN failed to identify subtle alterations in the second column of Figure 7. In contrast, our proposed method, leveraging contextual information and a dynamically learned attention matrix, demonstrated superior efficacy in detecting both minor and significant changes within an image scene. Also, in some methods such as SiamixFormer, the color changes of the road or vehicles have been extracted as structural changes, and an FP error has been created, which cannot be seen in the proposed method.

The notable aspect regarding the 2DCD dataset revolves around labeling inaccuracies present within the data, potentially leading to a reduction in overall accuracy across all methods. As depicted in the fourth column of Figure 7, the alterations manifest as the construction of two distinct buildings with a considerable gap between them, and one of the structures is still under construction towards the right side of the image. However, the dataset labels this scenario as a single isolated building undergoing change. Contrarily, our results demonstrate that all methods have correctly identified these as two separate structures, aligning with the actual scenario. Additionally, the southern section of the building on the left is in the process of being built, a detail accurately captured by the proposed method but missed by previous approaches

The noteworthy aspect concerning the 2DCD dataset revolves around labeling errors inherent in the data, potentially leading to an overall decrease in accuracy across all methods. As depicted in the fourth column of Figure 7, the changes appear in the form of constructing two distinct buildings with a considerable gap between them, and one of the structures is still under construction towards the left side of the image. However, the dataset labels this scenario as a single isolated building undergoing change. Conversely, our results demonstrate that all methods have correctly identified these as two separate structures, aligning with the actual scenario. Additionally, the southern section of the building on the left is under construction, a detail accurately captured by the proposed method but missed by previous approaches.

Table 4. Comparative Assessment of proposed BCD Algorithms: LEVIR-CD vs 2DCD

Methods	**LEVIR-CD**			**2DCD**		
	Pre (%)	**Re (%)**	**F1 (%)**	**Pr (%)**	**Re (%)**	**F1 (%)**
STANet	84.59	**91.00**	87.68	79.37	50.77	61.93
DDCNN	91.88	88.29	90.05	81.03	60.41	69.22
AGCDetNet	91.07	88.93	89.98	86.38	69.65	77.12
SiamixFormer	89.51	88.93	89.22	84.43	69.07	75.98
BIT	90.16	89.11	89.63	83.41	68.92	75.46
IDET	91.30	87.05	89.12	85.71	70.49	77.36
ChangeFormer	92.16	88.17	90.12	86.04	72.25	78.54
ChangeCoTNet (ours)	**92.53**	89.79	**91.14**	**88.47**	**73.09**	**80.05**

5. Conclusion

In this study, we proposed ChangeCoTNet, a novel hybrid network that integrates Contextual Transformer (CoT) blocks in the encoder and CNN in the decoder for high-resolution building change detection. The model effectively addresses challenges related to complex urban environments by leveraging both static and dynamic contextual representations and channel attention mechanisms.

The key contributions of this work include:

1. The introduction of a dual-branch architecture that combines CoT blocks and CNN to enhance feature extraction and improve the model’s ability to capture global and local contextual information.
2. The use of a channel attention block to emphasize discriminative features, leading to more accurate detection of subtle and complex changes.
3. Extensive evaluation on two benchmark datasets, LEVIR-CD and 2DCD, demonstrating significant improvements in F1-scores compared to state-of-the-art methods.

Our results highlight the practical advantages of the proposed method, including a 1.1% improvement in F1-score for the LEVIR-CD dataset and a remarkable 1.9% improvement for the 2DCD dataset. These improvements underline the effectiveness of incorporating CoT and CAB in achieving superior performance in change detection tasks.

Moreover, the efficiency of ChangeCoTNet makes it well-suited for real-world applications, such as urban planning, disaster response, and environmental monitoring, where accurate and timely detection of changes is critical.

Future work will explore the application of the proposed method to larger datasets and more diverse scenarios to further validate its generalizability and scalability.

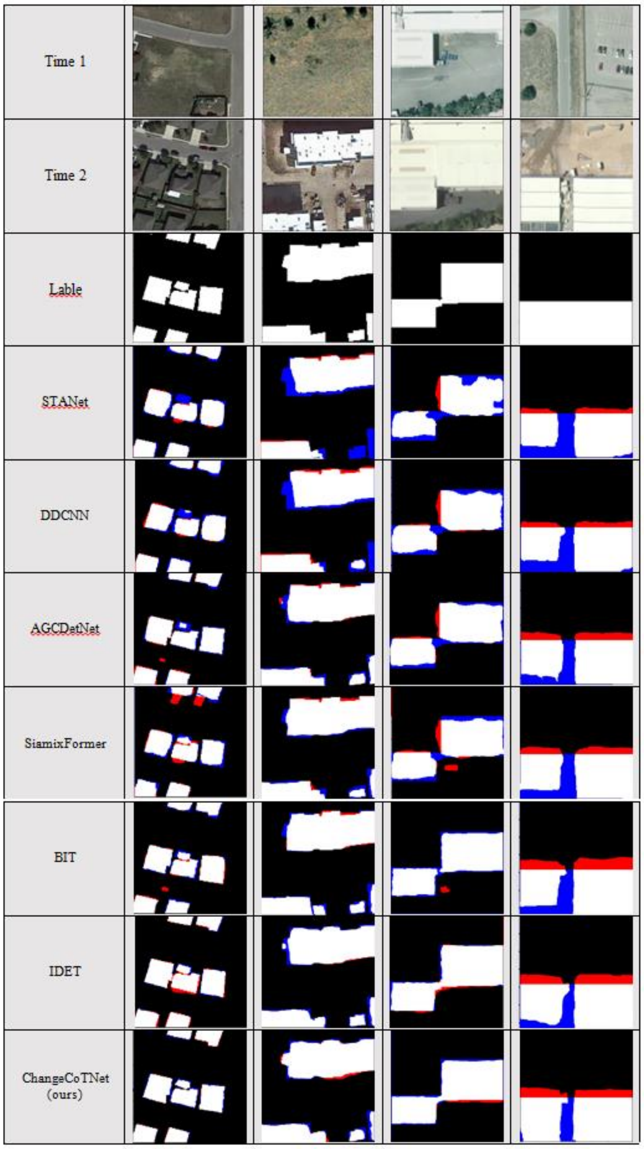

Figure 7. Visual evaluation contrasting SOTA methodologies with the proposed ChangeCoTNet approach on both the LEVIR-CD and 2DCD datasets

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest: No conflict of interest exists in the submission of this manuscript, and the manuscript is approved by all authors for publication.

References

Alzu'bi, A., & Alsmadi, L. (2022). Monitoring deforestation in Jordan using deep semantic segmentation with satellite imagery. Ecological Informatics, 70, 101745. https://doi.org/10.1016/j.ecoinf.2022.101745

Anonymous. (2021). Global Self-Attention Networks. ICLR Submissions, Mil. https://doi.org/10.48550/arXiv.2010.03019

Bai, T., Wang, L., Yin, D., Sun, K., Chen, Y., Li, W., & Li, D. (2022). Deep learning for change detection in remote sensing: a review. In Geo-Spatial Information Science. https://doi.org/10.1080/10095020.2022.2085633

Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. Journal of Applied Remote Sensing. https://doi.org/10.1117/1.jrs.11.042609

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. https://doi.org/10.48550/arXiv.2201.01293

Chen, H., Qi, Z., & Shi, Z. (2021). Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14. https:// doi.org/10.1109/TGRS.2021.3095166

Chen, H., Qi, Z., & Shi, Z. (2022). Remote Sensing Image Change Detection with Transformers. IEEE Transactions on Geoscience and Remote Sensing, 60. https://doi.org/10.1109/TGRS.2021.3095166

Chen, H., & Shi, Z. (2020). A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing. https://doi.org/10.3390/rs12101662

Chughtai, A. H., Abbasi, H., & Karas, I. R. (2021). A review on change detection method and accuracy assessment for land use land cover. Remote Sensing Applications: Society and Environment, 22. https://doi.org/10.1016/j.rsase.2021.100482

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. ICLR 2021 - 9th International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929

Eftekhari, A., Samadzadegan, F., & Dadrass Javan, F. (2023). Building change detection using the parallel spatial-channel attention block and edge-guided deep network. International Journal of Applied Earth Observation and Geoinformation, 117. https://doi.org/10.1016/j.jag.2023.103180

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. https://doi.org/10.1109/LGRS.2021.3056416

Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). Overview of the Transformer-based Models for NLP Tasks. Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020. https://doi.org/10.15439/2020F20

Groener, A., Chern, G., & Pritt, M. (2019). A Comparison of Deep Learning Object Detection Models for Satellite Imagery. Proceedings - Applied Imagery Pattern Recognition Workshop, 2019-October. https://doi.org/10.1109/AIPR47015.2019.9174593

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., & Hu, S. M. (2022). Attention mechanisms in computer vision:

A survey. In Computational Visual Media. https://doi.org/10.1007/s41095-022-0271-y

Guo, Q., Wang, R., Huang, R., Sun, S., & Zhang, Y. (2022). IDET: Iterative difference-enhanced transformers for high-quality change detection. ArXiv Preprint ArXiv:2207.09240. https://doi.org/10.48550/arXiv.2207.09240

Hussain, M., Chen, D., Cheng, A., Wei, H., & Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. In ISPRS Journal of Photogrammetry and Remote Sensing. https://doi.org/10.1016/j.isprsjprs.2013.03.006

Khelifi, L., & Mignotte, M. (2020). Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3008036

Khusni, U., Dewangkoro, H. I., & Arymurthy, A. M. (2020). Urban Area Change Detection with Combining CNN and RNN from Sentinel-2 Multispectral Remote Sensing Data. 2020 3rd International Conference on Computer and Informatics Engineering, IC2IE 2020. https://doi.org/10.1109/IC2IE50715.2020.9274617

Lan, J., Zhang, C., Lu, W., & Gu, N. (2023). Spatial-Transformer and Cross-Scale Fusion Network (STCS-Net) for Small Object Detection in Remote Sensing Images. Journal of the Indian Society of Remote Sensing, 51(7). https://doi.org/10.1007/s12524-023-01709-w

Li, M., Fu, Y., & Zhang, Y. (2023). Spatial-Spectral Transformer for Hyperspectral Image Denoising. Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023, 37. https://doi.org/10.1609/aaai.v37i1.25221

Li, Y., Mao, H., Girshick, R., & He, K. (2022). Exploring Plain Vision Transformer Backbones for Object Detection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13669 LNCS. https://doi.org/10.1007/978-3-031-20077-9_17

Li, Y., Yao, T., Pan, Y., & Mei, T. (2023). Contextual Transformer Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2). https://doi.org/10.1109/TPAMI.2022.3164083

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12). https://doi.org/10.1109/TNNLS.2021.3084827

Marsocci, V., Coletta, V., Ravanelli, R., Scardapane, S., & Crespi, M. (2023). Inferring 3D change detection from bitemporal optical images. ISPRS Journal of Photogrammetry and Remote Sensing, 196. https://doi.org/10.1016/j.isprsjprs.2022.12.009

Mazzanti, P., Scancella, S., Virelli, M., Frittelli, S., Nocente, V., & Lombardo, F. (2022). Assessing the Performance of Multi-Resolution Satellite SAR Images for Post-Earthquake Damage Detection and Mapping Aimed at Emergency Response Management. Remote Sensing, 14(9). https://doi.org/10.3390/rs14092210

Mohammadian, A., & Ghaderi, F. (2023). SiamixFormer: a fully-transformer Siamese network with temporal Fusion for accurate building detection and change detection in bi-temporal remote sensing images. International Journal of Remote Sensing, 44(12). https://doi.org/10.1080/01431161.2023.2225228

Padró, J. C., Muñoz, F. J., Planas, J., & Pons, X. (2019). Comparison of four UAV georeferencing methods for environmental monitoring purposes focusing on the combined use with airborne and satellite remote sensing platforms. International Journal of Applied Earth Observation and Geoinformation, 75. https://doi.org/10.1016/j.jag.2018.10.018

Peng, D., Zhang, Y., & Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing, 11(11). https://doi.org/10.3390/rs11111382

Peng, X., Zhong, R., Li, Z., & Li, Q. (2020). Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference. IEEE Transactions on Geoscience and Remote Sensing, 59(9). https://doi.org/10.1109/tgrs.2020.3033009

Qin, R., Tian, J., & Reinartz, P. (2016). 3D change detection – Approaches and applications. In ISPRS Journal of Photogrammetry and Remote Sensing (Vol. 122). https://doi.org/10.1016/j.isprsjprs.2016.09.013

Qin, Y., Wu, Y., Li, B., Gao, S., Liu, M., & Zhan, Y. (2019). Semantic segmentation of building roof in dense urban environment with deep convolutional neural network: A case study using GF2 VHR imagery in China. Sensors, 19(5), 1164. https://doi.org/10.3390/s19051164

Song, K., & Jiang, J. (2021). AGCDetNet:An Attention-Guided Network for Building Change Detection in High-Resolution Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14. https://doi.org/10.1109/JSTARS.2021.3077545

Stilla, U., & Xu, Y. (2023). Change detection of urban objects using 3D point clouds: A review. In ISPRS Journal of Photogrammetry and Remote Sensing (Vol. 197). https://doi.org/10.1016/j.isprsjprs.2023.01.010

Tewkesbury, A. P., Comber, A. J., Tate, N. J., Lamb, A., & Fisher, P. F. (2015). A critical synthesis of remotely sensed optical image change detection techniques. In Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2015.01.006

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem. https://doi.org/10.48550/arXiv.1706.03762

Wang, G., Li, B., Zhang, T., & Zhang, S. (2022). A Network Combining a Transformer and a Convolutional Neural Network for Remote Sensing Image Change Detection. Remote Sensing, 14(9). https://doi.org/10.3390/rs14092228

Wang, H., Xing, C., Yin, J., & Yang, J. (2022). Land Cover Classification for Polarimetric SAR Images Based on Vision Transformer. Remote Sensing, 14(18). https://doi.org/10.3390/rs14184656

Wang, L., Fang, S., Meng, X., & Li, R. (2022). Building Extraction With Vision Transformer. IEEE Transactions on Geoscience and Remote Sensing, 60. https://doi.org/10.1109/TGRS.2022.3186634

Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., Zhang, F., Fuegen, C., Zweig, G., & Seltzer, M. L. (2020). Transformer-Based Acoustic Modeling for Hybrid Speech Recognition. ICASSP, IEEE

International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May. https://doi.org/10.1109/ICASSP40776.2020.9054345

Wen, D., Huang, X., Bovolo, F., Li, J., Ke, X., Zhang, A., & Benediktsson, J. A. (2021). Change Detection from Very-High-Spatial-Resolution Optical Remote Sensing Images: Methods, applications, and future directions. IEEE Geoscience and Remote Sensing Magazine, 9(4). https://doi.org/10.1109/MGRS.2021.3063465

Yan, L., Yang, J., & Zhang, Y. (2022). Building Instance Change Detection from High Spatial Resolution Remote Sensing Images Using Improved Instance Segmentation Architecture. Journal of the Indian Society of Remote Sensing, 50(12). https://doi.org/10.1007/s12524-022-01601-z

Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., & Qiu, X. (2017). Change Detection Based on Deep Siamese Convolutional Network for Optical Aerial Images. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2017.2738149

Zhang, Y., Peng, D., & Huang, X. (2018). Object-based change detection for VHR images based on multiscale uncertainty analysis. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2017.2763182

Zhao, J., Wang, X., Dou, X., Zhao, Y., Fu, Z., Guo, M., & Zhang, R. (2022). A high-precision image classification network model based on a voting mechanism. International Journal of Digital Earth, 15(1). https://doi.org/10.1080/17538947.2022.2142306

Zheng, Z., Zhong, Y., Wang, J., Ma, A., & Zhang, L. (2021). Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. Remote Sensing of Environment, 265. https://doi.org/10.1016/j.rse.2021.112636