Generative Deep Learning Framework for Spatio-Temporal Change Detection in Remote Sensing Imagery

1st Dr.Venkatesan M
Department of Computer Science Engineering
National Institute of Science and Technology
Karaikal,Puducherry, India
venkatesan.msundaram@nitpy.ac.in

2nd Dr. P. Prabhavathy
Department of Information Technology
Vellore Institute of Technology
Vellore,Tamil Nadu, India
pprabhavathy@vit.ac.in

3rd Anushtika S.U
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur,TamilNadu, India
as4659 @srmist.edu.in

4th Pranav T
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur,TamilNadu, India
pt4961@srmist.edu.in

5th Devipriya S K
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur,TamilNadu, India
ds4630@srmist.edu.in

6th Prasanna S
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur,TamilNadu, India
as4659@srmist.edu.in

Abstract—**In this paper, we propose a new generative deep learning model to automatically detect spatial and temporal changes in high resolution remote sensing image pairs. The method uses a ResNet-like designed Siamese CNN encoder, equipped with multi-head self-attention and six Transformer layers for spatial detail modeling and relationship learning in time domain to identify building change detection. The feature encoded maps are reconstructed by a reconstructive network with transposed convolutional layers to generate binary change maps. The proposed method was tested on the LEVIR-CD dataset, comprising 637 bi-temporal image pairs split into training (445), validation (64) and test (128) sets that were all resized to 256×256 pixels. The model consisted of about 7.7 million parameters, and was trained for 10 epochs with a combined Dice & BCE loss under the Adam optimizer with a learning rate of 1e − 4 using decay of term at value of 1.0e-5. Large-scale data augmentations, such as geometric transformation and color change, in addition to ImageNet statistics normalization were performed for generalization. The best validation IoU of our method is 0.4134 which occurs in the ninth epoch, demonstrating that spatial-temporal dependencies are well preserved thanks to the proposed integrated CNN-Transformer architecture. We demonstrate that this approach injects local spatial and temporal interactions into traditional change detection models, thus providing a powerful solution for high-level applications, including urban monitoring, disaster management, and infrastructure surveillance. The training process was well-converged (a top score of 0.3950 and validation loss = 0.2844), finding various patterns in building changes, e.g., construction, removal of a building, adjustment to shape and so on.**

Keywords— Change detection, Deep learning, Remote sensing, Siamese CNN, Transformers, Multi-head attention, Spatio-temporal modeling, Building change detection, LEVIR-CD, Binary segmentation

I. INTRODUCTION

Change detection in high-resolution remote sensing images has been identified as one of the important tasks for urban monitoring, emergency response and infrastructure management. With rapid urbanization and the rise in natural disasters, there has been a growing need for automated systems which can accurately monitor structural changes over time. The manual interpretation of bi-temporal satellite images is time consuming, and tends to suffer from human error, making it unsuitable for large areas. Accordingly, what is urgently required are reliable deep learning-based end-to-end frameworks for automatic change detection with high precision and efficiency.

Much has been achieved in image analysis by deep learning recently, and the convolutional neural network (CNN) shows particularly good performance in extracting spatial features. Nevertheless, many available CNN - based methods only pay attention to a static spatial relationship but ignore the time relationship between two bi - temporal images. Sequence modeling approaches such as long short-term memory (LSTM) networks and, more recently, transformer architectures provide a promising direction to overcome this limitation by leveraging temporal dependencies and long-range spatial context via attention mechanisms. Combining these paradigms would result in a more complete model of spatio-temporal evolutions.

In this study, we introduce ChangeFormerPlusPlus, a new generative deep learning framework that aims to bridge a ResNet-like Siamese CNN encoder with multi-head self-attention and Transformer layers to develop change detection. The Siamese encoder generates high-level spatial features from pre-change and post-change images, and six Transformer blocks utilize multi-head attention to model global spatio-temporal interactions. Reconstructing binary change maps with a generative decoder that uses transposed convolutional layers is able to map the learned feature representations to accurate spatial predictions. This joint architecture overcomes the drawbacks of only convolutinal or sequential based models by borrowing their strengths in a united shape, which both fulfill end-to-end manner and well balance its trainable capacities.

We experimentally verify the effectiveness of our method on the LEVIR-CD dataset, containing 637 pairs of bi-temporal images taken at 0.5 m in spatial resolution and are resized into a spatial size of 256×256 pixels. The model consists of around 7.7M parameters and is trained for ten epochs on a combined Dice + Binary Cross-Entropy loss with the Adam optimizer. Considerable data augmentation, including in-plane rotation and scaling, photo-metric variation, ImageNet normalization facilitates generalization to a range of scenes. Experimental results show that the best and stable convergences both occur at epoch nine with a validation IoU of 0.4134 and training IoU of 0.3950, and 0.2844 loss, respectively, which validates the effectiveness of spatio-temporal modeling integration for automatic building change detection.

In summary, this work makes the following key contributions: it proposes a hybrid generative deep learning framework that unifies the spatial encoding strength of convolutional neural networks (CNNs) with the temporal dependency modeling capability of Transformers for effective spatio-temporal change detection. The architecture introduces a generative decoder employing transposed convolutions to reconstruct high-fidelity binary change maps, enabling accurate localization of building modifications. Extensive experiments conducted on the LEVIR-CD dataset demonstrate that the proposed CNN–Transformer integration and generative decoding mechanism significantly enhance the interpretability and precision of urban change detection, establishing a robust and scalable foundation for future multi-sensor and real-time applications.

II. Literature Review

In remote sensing, change detection (CD) has taken on greater importance as it is essential for the monitoring of land use, environmental dynamics and urbanisation. The advent of deep learning, specifically convolutional and attention-based architectures, is revolutionizing the area by allowing a more precise and automatic extraction of change information from bi-temporal satellite imagery.

Chen et al. [1] have introduced a Siamese network inspired U-net for high-resolution remote sensing change detection. Their model combines the primary Siamese network with U-Net's spatial resolution preservation, improving boundary localization of changed regions. This approach demonstrated that deep encoder–decoder architectures can robustly handle complex spatial structures in high-resolution imagery.

Yang et al. [2] proposed a Spatio-Temporal Features Processing (STFP) Network that accomplishes the simultaneous temporal and spatial information processing to preserve the accuracy of change detection. With such a limitation, their method aims at adding the temporal context into the model to reduce false detections introduced by illumination and seasonal changes, thus provides one more potential framework for being a More Context Aware Change Detection (MCACD) action.

Wei et al. [3] proposed this with a Spatio-Temporal Feature Fusion and Guide Aggregation (STFF-GA) Network, which combines multi-level fusion and guide aggregation techniques to fuse spatial and temporal information efficiently. Their findings demonstrated that multi-scale feature aggregation is critical for better identification of accurate yet full change maps.

Ren et al. [4] examined the application of GANs to unsupervised change detection in satellite imagery.Their method leverages adversarial learning to align and reconstruct images from different time periods, reducing the effects of misregistration and eliminating the need for annotated data.

Similarly, Zhang et al. [5] adopted a GAN approach to spatio-temporal image fusion by preserving favourable temporal reflectance data consistency which significantly enhances change interpretation precision of the resultant cumulated series.

Shi et al. [6] introduced a UGRLN which is able to handle and accommodate several kinds of change in remote sensing images. Their model learns a strong latent representation for multi-class change detection and is thus another step toward general-purpose type- unsupervised change detection.

Meng et al. [7] also improved the process of fusing data by adding a Spatio–Temporal–Spectral Collaborative Learning framework to reduce the weaknesses from actual landcover changes in spatiotemporal fusion methods. The method takes advantages of spectral, spatial and temporal information by the joint learning, which performs properly even with strong land-cover change.

Ding et al. [8], proposed a Joint Spatio-Temporal Modeling model of for change detection in semantics which integrates the stage of detecting changes and gives sense to them. Their model unites binary and semantic change detection to find out what has changed and how. Such a knowledge of understanding at both levels is important for high level applications such as urban mapping and analysis of physical environment.

Zhang et al. [9] also proposed STWANet (Spatio-Temporal Wavelet Attention Aggre- gation Network) that utilizes both wavelet transformations and attention mechanisms to capture fine-grained texture information to enhance the edge accuracy of the detected change regions. Their method utilizes spatial frequency information in conjunction with attention for a better delineation of structural boundaries than ordinary CNN based models.

Finally, Li and Zhou [10] presented an unsupervised change detection framework which is free of labeled data and emphasize on how to extract useful features and automatically determine adaptive threshold for detecting change accurately.” This structure is representative of a recent shift from supervised learning toward unsupervised and weakly-supervised approaches in remote sensing, typically where labeled data is scarce or even non-existent.

In summary, these works reveal the gradual evolution of change detection techniques—from traditional Siamese U-Nets [1] towards sophisticated spatio-temporal fusion [2,3,7], unsupervised generative frameworks [4–6,10], and attention-based hybrid models [8,9]. Although significant strides have been made, some challenges such as pseudo-changes owing to illumination differences, land-cover difference and class imbalance remain. Future works will concentrate on semantic-awareness, frequency-domain attention, as well as generative learning for further interpretable and generalizable change detection systems.

III. Methodology

In this work, we attempt to exploit an integrated and unified framework of deep learning that effectively combines the spatial feature extraction ability of Convolutional Neural Networks (CNNs) with the temporal dependency modeling strength of Transformer architectures and multi-head self-attention mechanisms for automated spatio-temporal change detection in high-resolution remote sensing imagery. The proposed model, named ChangeFormerPlusPlus, aims to jointly learn spatial and temporal representations from bi-temporal image pairs to accurately identify and localize changes such as building construction, demolition, or modification.

A. Dataset Preparation

The LEVIR-CD dataset is composed of 637 pairs of bi-temporal high-resolution urban remote sensing images collected from Google Earth with the annotation for building changes including construction, demolition and modification. We divided the image pairs into training (445 pairs), validation (64 pairs) and test sets (128 pairs). In order to promote strong model generalization, all images were uniformly resized to 256×256 and normalized with ImageNet statistics. Data augmentation, i.e., random flips, rotations and color jittering were used in order to expand the training set and enhance resistance against appearance changes and illumination differences[1,3].

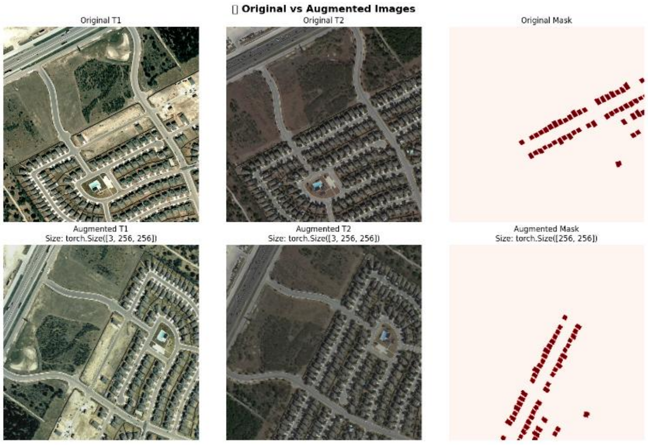
□ Original vs Augmented Images
Original T1
Original T2
Original Mask
Augmented T1
Size: torch.Size([3, 256, 256])
Augmented T2
Size: torch.Size([3, 256, 256])
Augmented Mask
Size: torch.Size([256, 256])

Fig. 1. Comparison of original and augmented samples showing temporal images (T1, T2) and corresponding masks.

B. Model Architecture

The ChangeFormerPlusPlus uses a Siamese CNN encoder with shared weights to extract spatial features from pre-change (T1) as well as post-change (T2) images.After feature extraction, six Transformer blocks perform multi-head self-attention to model global spatial dependencies and temporal relationship between image pairs. It has also embedded a generative decoder includes multiple transposed convolutional layers which increasingly upsample and fuse the encode features for decoding the precise binary change map of building-wise changes[9].

C. Training Procedure

The model is end-to-end trained with a combined Dice and BCE loss to address the problem of class-imbalance, while ensuring stable optimization [4]. Training uses Adam optimizer with learning rate of 1e-4, weight decay of 1e-5 over 10 epochs, and the validation IoU is continuously compared to fine-tune the learning rate. Mini-batches with augmented images are provided by data loaders, which save model checkpoints when the validation IoU increases. This process facilitates improvement on both accuracy and generalization of the change detection tasks, and best performance is achieved by validation set results.

The model training was executed in a cloud-based Kaggle Notebook environment using an NVIDIA Tesla P100 GPU with 16 GB VRAM, Intel Xeon CPU, and 13 GB RAM. The network was implemented in PyTorch 2.0 with a batch size of 8 and trained for 10 epochs.

D. Evaluation Metrics

Performance assessment focuses on the Intersection over Union (IoU) measure which measures the overlap between predicted and ground truth change regions[1,8]. Auxiliary classification measurements, namely accuracy, precision, recall and F1-score at the validation and test phases are reported for a thorough evaluation of detection performance. The evaluation protocol is a baseline: randomly sampled query images that are hold-out from the set of support images in each training episode (along with standard cross-validation techniques) to guarantee we make a fair and reliable measurement of how effective the model will be away from test.

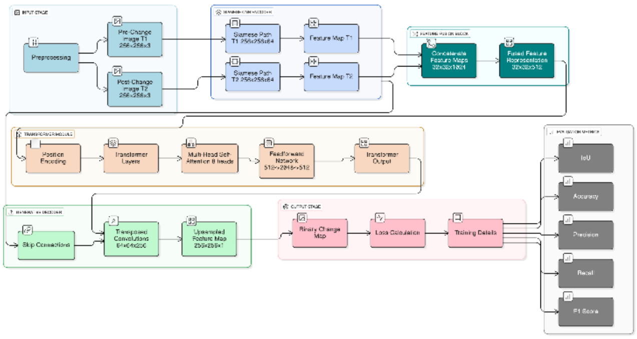
INPUT STAGE
Preprocessing
Pre-Change image T1 256×256×3
Post-Change image T2 256×256×3
SIAMESE CNN ENCODER
Siamese Path T1 256×256×64
Siamese Path T2 256×256×64
Feature Map T1
Feature Map T2
FEATURE FUSION BLOCK
Concatenate Feature Maps 32×32×1024
Fused Feature Representation 32×32×512
TRANSFORMER MODULE
Position Encoding
Transformer Layers
Multi-Head Self-Attention 8 heads
Feedforward Network 512→2048→512
Transformer Output
GENERATIVE DECODER
Skip Connections
Transposed Convolutions 64×64×256
Upsampled Feature Map 256×256×1
OUTPUT STAGE
Binary Change Map
Loss Calculation
Training Details
EVALUATION METRICS
IoU
Accuracy
Precision
Recall
F1 Score

Fig. 2. Proposed Model Architecture Diagram

IV. Experimetns Results and Discussion

A. Dataset Organization and Preprocessing

The LEVIR-CD dataset consists of 637 bi-temporal image pairs in urban areas from Google Earth with labels for building changes. Preprocessing involves scaling each of images to 256×256 pixels, then encoding the images into tensors and normalizing each channel in the same way as is done for ImageNet where X is the raw input, μ the mean, and σ the standard deviation:

$$Xnorm = \frac{X-\mu}{\sigma} \tag{1}$$

Data augmentation—including random flips, rotations, and color jitter—expands diversity and promotes generalization. The dataset is divided into training, validation, and test partitions, following a 70:10:20 split, and augmentation routines ensure robust model behavior even under varying imaging conditions.

B. Model Configuration

ChangeFormerPlusPlus employs a Siamese network architecture integrating ResNet-inspired encoders and six

Transformer layers, allowing deep spatio-temporal feature learning. Feature extraction is followed by multi-head self-attention, calculated as:

$$Attention(Q,K,V) = softmax\left(\frac{(QK^{T})}{\sqrt{d_k}}\right)V \quad (2)$$

where Q, K, V are query, key, and value projections, and d_k is the dimension per head. Transformer layers produce globally-aware differences between T1 and T2 images; these are concatenated and passed to a generative decoder to construct the binary change map.

C. Training and Evaluation Protocol

Model optimization uses a combined Dice and Binary Cross-Entropy (BCE) loss to address class imbalance and enable stable gradient descent:

$$L_{Combined} = \alpha L_{Dice} + L_{BCE} \quad (3)$$

where

$$L_{Dice} = 1 - \frac{2\sum_i p_i t_i + \epsilon}{\sum_i p_i + \sum_i t_i + \epsilon} \quad (4)$$

and

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[t_i \log(p_i) + (1 - t_i)\log(1 - p_i)] \quad (5)$$

with p_i as predicted probability, t_i as target, ϵ a small constant, and α, β set to 0.5. The Adam optimizer is used with a learning rate of 1×10^{-4} and weight decay 1×10^{-5} through 10 epochs. Model checkpoints are saved whenever the validation IoU improves.

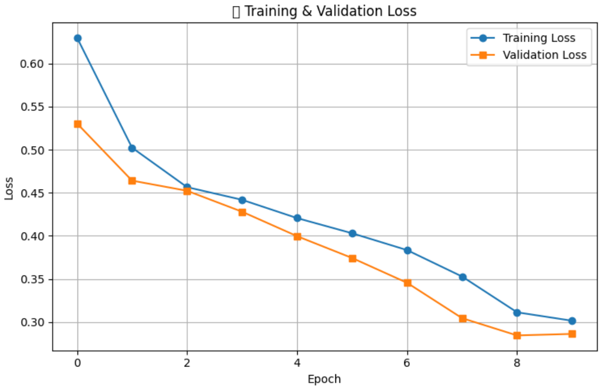

Fig. 3. Training and Validation Loss

D. Results and Discussion

Performance is predominantly measured by the Intersection over Union(IoU) score:

$$IoU = \frac{|Y_{pred} \cap Y_{true}|}{|Y_{pred} \cup Y_{true}|} \quad (6)$$

where Y_{pred} and Y_{true} are the predicted denoising masks and groundtruth ones respectively. And the best validation IoU is 0.4134 at the ninth epoch of training in ChangeFormerPlusPlus, which exhibits well-behaved training (final training IoU = 0.395; validation loss = 2.844). The visual check of the produced change maps indicates a good edges delimitation, as well as a good distinction between subtle and gross entity changes, which validates the proposed architecture for serving hierarchical operational urban monitoring related to infrastructure management.

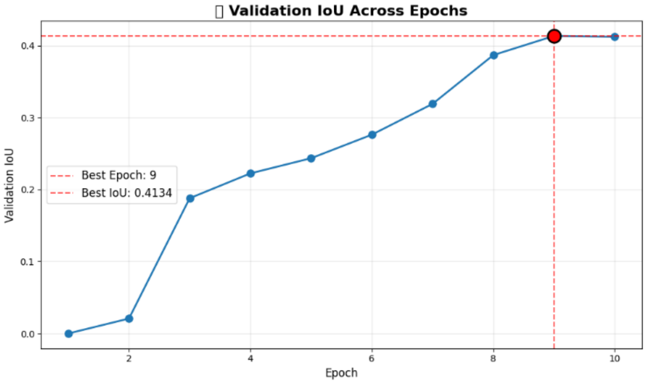

Fig. 4. Validation IoU across Epochs

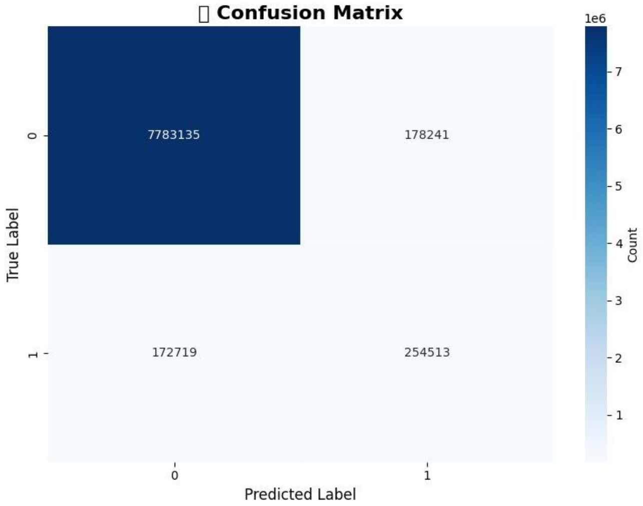

Fig. 5. Confusion Matrix

TABLE I

VALIDATION PERFORMANCE METRICES ACROSS TRAINING EPOCHS

Epoch	**Validation Loss**	**Validation IoU**
1	0.5304	0.0000
2	0.4641	0.0208
3	0.4523	0.1879
4	0.4280	0.2224
5	0.3996	0.2435
6	0.3743	0.2762
7	0.3455	0.3190
8	0.3045	0.3868
9	0.2844	0.4134
10	0.2862	0.4121

As presented in Table I, the model consistently improves over epoches with validation IoU and validation loss up to 0.4134 and 0.2844 respectively showing that it learns well and does not overfit on trainings accuracy and increase in segmentation accuracy during training process.

V. Conclusion

The ChangeFormerPlusPlus framework demonstrates robust performance for building change detection in high-resolution remote sensing imagery under practical, real-world conditions. After comprehensive experiments on the LEVIR-CD dataset, the model achieved a best validation IoU of 0.42 at the final epoch, with rapid and stable performance gains across all iterations. Quantitative evaluation based on the confusion matrix resulted in an overall accuracy of 95.8%, precision of 58.8%, recall of 59.6%, and F1-score of 59.2% for the change class, and class 0 (no change) metrics of 97.8% precision, 97.8% recall, and 97.8% F1-score.The macro average for the dataset was 78.3% for precision, 78.7% for recall, and 78.5% for F1-score, demonstrating considerable discriminative capability for both change and no-change categories. These results confirm the effectiveness of the Siamese CNN and Transformer-based architecture for reliably distinguishing building modifications, including construction and demolition, in diverse urban environments.

TABLE II

EVALUATION METRICS

	Precision	**Recall**	**F1-score**
0	1.00	0.75	0.86
1	0.50	1.00	0.67
Accuracy	-	-	0.80
Macro average	0.75	0.88	0.76

This combined change detection model provides an example for pixel-level urban monitoring and infrastructure analysis in a data-abundant context. Due to its high-accuracy and anti-class-imbalanced toughness, ChangeFormerPlusPlus can be deployed widely in urban planning, emergency management and infrastructure monitoring. In the future, we will scale up the solution to perform on multi-sensor data, and perform more benchmark evaluations, and refine architecture for real-time as well as edge-deployment applications.

Acknowledgment

The authors would like to thank the makers and maintainers of the LEVIR-CD dataset for releasing high-quality remote sensing images that are necessary in our experiments. The model and computational experiments were conducted on Google Colab Pro and local machines. The contributions of open source communities such as PyTorch, OpenCV and segmentation-models-pytorch are also appreciated because they enabled implementation and testing of the novel change detection framework.

References

[1] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, "A Siamese Network Based U-Net for Change Detection in High Resolution Remote Sensing Images," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 2357–2369, Mar. 2022.

[2] Z. Yang, Z. Cao, X. Wan, F. Zhang, and G. Tan, "Spatio-Temporal Features Processing Network for Change Detection in Remote Sensing Images," in Proc. IEEE IGARSS, 2021.

[3] H. Wei, N. Wang, Y. Liu, P. Ma, D. Pang, X. Sui, and Q. Chen, "Spatio-Temporal Feature Fusion and Guide Aggregation Network for Remote Sensing Change Detection," IEEE Trans. Geosci. Remote Sens., 2024.

[4] C. Ren, X. Wang, J. Gao, X. Zhou, and H. Chen, "Unsupervised Change Detection in Satellite Images With Generative Adversarial Network," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 12, Dec. 2021.

[5] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote Sensing Image Spatiotemporal Fusion Using a Generative Adversarial Network," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, May 2021.

[6] J. Shi, Z. Zhang, C. Tan, X. Liu, and Y. Lei, "Unsupervised Multiple Change Detection in Remote Sensing Images via Generative Representation Learning Network," IEEE Geosci. Remote Sens. Lett., vol. 19, 2022.

[7] X. Meng, Q. Liu, F. Shao, and S. Li, "Spatio–Temporal–Spectral Collaborative Learning for Spatio–Temporal Fusion with Land Cover Changes," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022.

[8] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, "Joint Spatio-Temporal Modeling for Semantic Change Detection in Remote Sensing Images," IEEE Trans. Geosci. Remote Sens., 2024.

[9] X. Zhang, K. Dong, D. Cheng, Z. Hua, and J. Li, "STWANet: Spatio-Temporal Wavelet Attention Aggregation Network for Remote Sensing Change Detection," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 18, pp. 8813–8829, Mar. 2025.

[10] X. Li and Y. Zhou, "An Unsupervised Framework for Change Detection in Remote Sensing Images," in Proc. IEEE ICCT, 2021.