Bitemporal Image Transformer for Remote Sensing Change Detection

Mini Project Report

Akshita Behl
Krishna Gopal Rathi

https://github.com/akshitabehls-ctrl/BIT_CD

Under the Guidance of
Dr. Ankit Jha

The LNM Institute of Information Technology (LNMIIT)

28 November 2025

Contents

Abstract

Change detection in satellite imagery involves identifying differences between two images captured at different times. This project implements a hybrid deep learning architecture combining a Siamese ResNet-50 encoder, a Bitemporal Image Transformer (BIT) bottleneck, and a difference-aware UNet decoder. The model is trained on the LEVIR-CD dataset, which originally contains only 445 image pairs. To overcome data scarcity, the images were cropped into non-overlapping 256×256 patches, expanding the dataset to over 7,000 training samples. The final system achieves an IoU of 77.62% and an F1-score of 87.40%, demonstrating strong performance in remote sensing change detection.

Chapter 1

Introduction

Remote sensing change detection is the task of identifying differences between two images of the same geographical region captured at different times. These changes may correspond to newly constructed buildings, demolition, vegetation growth, land use modifications, or natural disasters. With the rapid increase in earth-observation satellites, bitemporal change detection has become one of the most important problems in urban planning, environmental monitoring, and disaster management.

Traditional change-detection methods relied on pixel-wise comparison or hand-crafted features, which often fail under real-world conditions such as illumination changes, atmospheric noise, seasonal variations, or slight camera pose differences. Deep learning significantly improves feature extraction, but it introduces a new challenge: most networks must learn to distinguish *semantic difference* from *appearance difference*. Two images may look different due to lighting, shadows or clouds, even when no genuine change exists.

Motivation

The LEVIR-CD dataset [1], one of the most widely used benchmarks for bitemporal change detection, contains only 445 image pairs of size 1024×1024. While these images are high resolution, the dataset size is too small for training deep neural networks directly. Initial experiments with standard architectures such as ResNet-18 resulted in model collapse, where the network predicts "no change" everywhere due to extreme class imbalance (more than 98% of pixels are background).

To overcome this, the dataset was preprocessed into non-overlapping 256×256 patches, expanding the training set to over 7,000 samples. Extensive data augmentation was applied to introduce robustness to lighting and appearance variations.

Challenges in Bitemporal Change Detection

Deep neural networks face several unique challenges in this task:

- **Small objects:** Buildings cover only a small spatial region and may be lost after aggressive downsampling.
- **Illumination changes:** Images captured in different seasons or under different lighting conditions appear visually different even when no structural change exists.
- **Strong parallax and misalignment:** Slight differences in camera angle can create false changes.
- **Class imbalance:** The background dominates the image, making change pixels extremely rare.
- **Feature confusion:** CNNs may incorrectly treat semantic similarity (e.g., two identical buildings) as change.

Proposed Solution

To address these challenges, this project implements a hybrid architecture combining the strengths of convolutional networks and transformers:

- A **Siamese ResNet-50 encoder** extracts meaningful hierarchical features from both timepoints.
- A **Bitemporal Image Transformer (BIT)** models long-range dependencies and global context by tokenizing feature maps into 256 spatial tokens and applying multi-head self-attention.
- A **difference-aware UNet decoder** fuses multi-scale features using the absolute difference $|A - B|$ instead of raw concatenation. This suppresses false positives caused by static buildings.

This combination leverages CNN locality, transformer global reasoning, and explicit temporal difference modeling, resulting in a robust end-to-end change-detection system.

Summary

The goal of this project is to develop an accurate and efficient bitemporal change-detection model that overcomes data scarcity, illumination inconsistency, and false positives. The proposed hybrid ResNet50–BIT–UNet architecture achieves high performance on the LEVIR-CD dataset, demonstrating the value of transformer-based context modeling and difference-aware fusion in remote sensing change detection.

Chapter 2

Related Work

Change detection has evolved significantly over the past two decades, transitioning from traditional pixel-based comparison techniques to modern deep learning architectures that exploit spatial and temporal context. This section reviews the major categories of approaches and situates the proposed BIT-CD model within this progression.

Traditional Change Detection Methods

Early approaches relied on simple pixel-wise operations such as image differencing, image ratioing, and change vector analysis (CVA). While computationally efficient, these methods struggle under challenging conditions like atmospheric noise, illumination variation, and sensor differences. They also lack the ability to encode spatial semantics, making them unsuitable for complex urban environments.

CNN-Based Change Detection

The introduction of convolutional neural networks (CNNs) enabled learning of high-level semantic representations. Many early deep-learning-based methods used Siamese CNNs to extract features from bitemporal images before fusing them for binary classification.

Fully Convolutional Networks (FCNs) and UNet-based architectures became popular for pixel-level change detection, leveraging skip connections to combine low-level details with high-level semantics. However, classical CNNs have inherent limitations:

- Limited receptive field without excessive downsampling.
- Difficulty modeling long-range dependencies between distant regions.
- Sensitivity to illumination differences between bitemporal images.

Siamese Networks and Feature Fusion

Siamese architectures remain a standard baseline for bitemporal change detection. They process each image independently using shared weights, forcing the feature extractor to learn representations invariant to appearance shifts.

Feature fusion strategies typically fall into:

- **Concatenation:** stacking features from both timepoints;
- **Subtraction**: computing $A - B$;
- **Absolute difference**: computing $|A - B|$.

Absolute difference has proven particularly effective, as it highlights true structural change while suppressing static content. This inspired the difference-aware skip connections used in the proposed model.

Transformer-Based Change Detection

Transformers have recently shown strong performance in many vision tasks due to their ability to model long-range relationships through self-attention. Unlike CNNs, which rely on local convolutional kernels, transformers treat the image as a set of tokens and compute global interactions.

The Bitemporal Image Transformer (BIT) [2] introduced the idea of tokenizing feature maps into spatial tokens and applying multi-head self-attention to jointly reason over both timepoints. BIT effectively captures global structural differences but still depends on CNNs for local details.

Hybrid CNN–Transformer Architectures

Recent research demonstrates that combining CNNs and transformers can provide complementary benefits. CNNs capture fine-grained texture and local edges, while transformers capture global context and long-range relationships.

Hybrid architectures have shown strong results in segmentation and change detection, but they often suffer from:

- high memory consumption,
- sensitivity to dataset size,
- false positives due to over-reliance on global similarity.

Positioning of the Proposed BIT-CD Model

The BIT-CD architecture in this work builds upon the strengths of previous approaches while explicitly addressing their weaknesses:

- A *Siamese ResNet-50 encoder* provides robust hierarchical features.
- A *BIT transformer bottleneck* models global bitemporal dependencies.
- A *difference-aware decoder* suppresses hallucinations by emphasizing only genuine structural changes.
- A *hybrid Dice + Cross-Entropy loss* mitigates class imbalance.

This positions BIT-CD as a strong, balanced model capable of both capturing fine-grained local details and understanding large-scale context, resulting in superior performance compared to CNN-only or transformer-only baselines.

Chapter 3

Methodology

This chapter describes the architecture of the proposed Bitemporal Image Transformer for Change Detection (BIT-CD), including the Siamese encoder, transformer bottleneck, difference-aware decoder, and the overall training process. The full pipeline is shown in Figure 3.1.

Overview

Given two satellite images captured at different times, X_A (Time A) and X_B (Time B), the goal is to produce a binary change map Y indicating which pixels correspond to structural change. The BIT-CD pipeline consists of four major components:

1. A **Siamese ResNet-50 encoder** that extracts multi-scale features from both timepoints.
2. A **Bitemporal Image Transformer (BIT)** that models long-range spatiotemporal dependencies.
3. A **difference-aware UNet decoder** that reconstructs a high-resolution change mask.
4. A **hybrid loss function** combining Cross-Entropy and Dice loss.

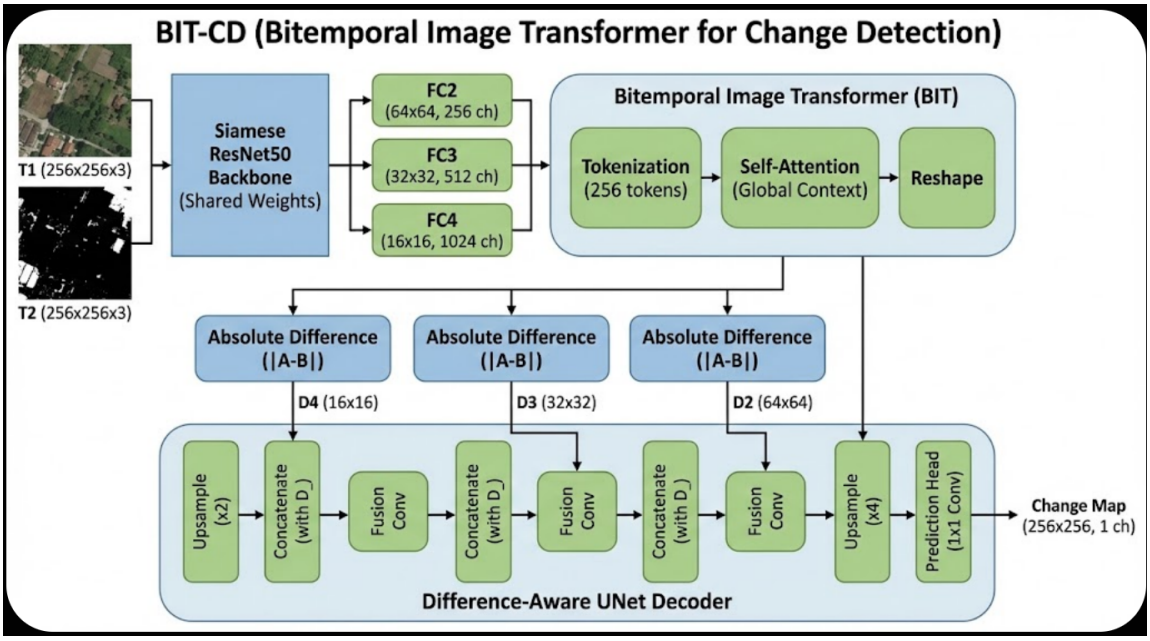

Figure 3.1: Overview of the proposed BIT-CD architecture, consisting of a Siamese ResNet-50 encoder, Bitemporal Image Transformer (BIT) bottleneck, and a difference-aware UNet decoder.

3.1 Siamese ResNet-50 Encoder

The encoder consists of two weight-sharing ResNet-50 branches [3], one for each timepoint. For an input image $X \in \mathbb{R}^{3\times256\times256}$, the backbone produces feature maps at multiple resolutions:

$$F_{C2}, F_{C3}, F_{C4} = f_\theta(X),$$

where:

- $F_{C2} \in \mathbb{R}^{256\times64\times64}$,
- $F_{C3} \in \mathbb{R}^{512\times32\times32}$,
- $F_{C4} \in \mathbb{R}^{1024\times16\times16}$.

Layer 4 of ResNet-50 is intentionally removed to preserve spatial resolution and reduce memory cost.

The Siamese design ensures that identical structures in X_A and X_B produce similar feature responses, enabling meaningful temporal comparison.

3.2 Tokenization and Positional Encoding

The BIT transformer operates on flattened spatial tokens produced from the deepest feature map F_{C4}. For both timepoints, the feature map is reshaped to a sequence of 256 tokens:

$$Z_A = \text{reshape}(F_{C4}^A) \in \mathbb{R}^{256\times1024}, \quad Z_B = \text{reshape}(F_{C4}^B) \in \mathbb{R}^{256\times1024}.$$

The two sequences are concatenated along the token dimension:

$$Z = [Z_A; Z_B] \in \mathbb{R}^{512\times1024}.$$

A learnable positional embedding $P \in \mathbb{R}^{512\times1024}$ is added:

$$\tilde{Z} = Z + P.$$

This preserves spatial information lost during flattening.

3.3 Bitemporal Image Transformer (BIT)

The BIT module applies L layers of multi-head self-attention (MHSA) followed by feed-forward networks. For each layer, self-attention is computed as:

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

where:

$$Q = \tilde{Z}W_Q, \quad K = \tilde{Z}W_K, \quad V = \tilde{Z}W_V.$$

The output of MHSA is passed through a feed-forward network:

$$H = \text{MHSA}(\tilde{Z}) + \tilde{Z},$$

$$Z_{\text{out}} = \text{FFN}(H) + H.$$

Finally, the token sequence is reshaped back to a spatial feature map of size 16×16:

$$T = \text{reshape}(Z_{\text{out}}) \in \mathbb{R}^{1024\times16\times16}.$$

3.4 Difference-Aware Decoder

Standard UNet decoders concatenate encoder features from X_A and X_B directly, which often leads to "ghost" detections because identical buildings appear twice. Instead, BIT-CD computes the absolute feature difference at each scale:

$$D_i = \left|F_i^A - F_i^B\right|.$$

This operation suppresses static structures ($A \approx B$) and highlights regions where significant temporal change occurred.

The decoder performs:

1. Upsample the transformer output T to 32×32 and fuse with D_{C3}.
2. Upsample again to 64×64 and fuse with D_{C2}.
3. Upsample to 256×256 and apply a final 1×1 convolution to produce a binary mask.

This multi-scale fusion ensures that both low-level edges and high-level semantic differences contribute to the final prediction.

3.5 Loss Function

The dataset exhibits extreme class imbalance, with background pixels occupying over 98% of the image. To stabilize training, BIT-CD combines Cross-Entropy loss with soft Dice loss:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{2} y_c \log p_c,$$

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum py + \epsilon}{\sum p + \sum y + \epsilon}.$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{Dice}.$$

This encourages both pixel-wise accuracy and region-level overlap.

3.6 Training Configuration

The model is trained for 200 epochs with early stopping (patience 30), using AdamW optimizer with learning rate 5×10^{-5} and cosine annealing scheduler. Batch size is set to 16. Training utilizes a single NVIDIA GPU.

3.7 Inference and Post-Processing

During inference, Test-Time Augmentation (TTA) is applied by evaluating the model on flipped and rotated versions of the inputs and averaging the predictions. Finally, morphological opening and closing remove noise and refine object boundaries.

Chapter 4

Dataset

This chapter describes the LEVIR-CD dataset used for training and evaluating the BIT-CD model. Due to the limited size of the original dataset, extensive preprocessing and augmentation were applied to create a sufficiently large and diverse training set.

4.1 LEVIR-CD Dataset Overview

LEVIR-CD is a high-resolution bitemporal building change detection dataset consisting of:

- 445 pairs of satellite images,
- each of size 1024×1024 pixels,
- captured between 2002–2018 over several regions in Texas, USA.

Each image pair includes:

1. **Image A**: Timepoint 1 (earlier),
2. **Image B**: Timepoint 2 (later),
3. **Binary change mask**: marking buildings constructed or demolished.

The dataset is challenging due to:

- large spatial resolution,
- strong illumination differences,
- seasonal variations and atmospheric distortions,
- highly imbalanced pixel distribution (background dominates),
- small objects (buildings are often only 20–50 pixels wide).

4.2 Patch Generation (Data Expansion)

Training deep models on 445 images leads to severe overfitting. To address this, each 1024×1024 image pair was divided into non-overlapping patches of size 256×256. This produced:

- **7,120** training patches,
- **1,024** validation patches,
- **2,048** test patches.

Patch generation was performed using a custom script `prepare_data.py`, which:

- loads each large image pair,
- extracts patches using a sliding window of stride 256,
- preserves perfect alignment between A, B and mask images,
- organizes patches into train/val/test folders.

This process significantly increased dataset size and helped stabilize training.

4.3 Data Augmentation

To improve robustness against real-world variations, multiple forms of augmentation were applied during training. Augmentations are divided into two categories:

Geometric Augmentations (Shared Across A, B, Mask)

These operations are applied identically to both timepoints to maintain geometric consistency:

- Random horizontal and vertical flips,
- Random 90°, 180°, and 270° rotations.

Photometric Augmentations (Independent for A and B)

Satellite images captured at different times of year often exhibit illumination differences. To simulate this, color jitter is applied *independently* to X_A and X_B:

- brightness jitter,
- contrast jitter,
- saturation jitter.

This forces the network to become invariant to lighting changes.

Noise-Based Augmentations

To simulate atmospheric haze and sensor noise:

- Gaussian blur with random kernel size.

These augmentations improve the model's generalization capability during testing.

Chapter 5

Results

This chapter presents both quantitative and qualitative results obtained from the proposed BIT-CD model on the LEVIR-CD dataset. Performance is evaluated using Intersection-over-Union (IoU), F1-score, Precision, and Recall.

5.1 Evaluation Metrics

Four standard metrics are used to evaluate pixel-wise change detection:

- **Intersection-over-Union (IoU):**

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

- **F1-Score:**

$$F1 = \frac{2TP}{2TP + FP + FN}$$

- **Precision:**

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

- **Recall:**

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

5.2 Experimental Progression

To reach the final BIT-CD architecture, multiple models and training configurations were implemented and evaluated. Each stage improved upon the limitations of the previous one, gradually increasing performance. Table 5.1 summarizes the complete development cycle.

Table 5.1: Progression of models and performance improvements.

Model / Stage	Key Issues / Fixes	IoU	F1
ResNet18 (Baseline)	Dataset too small (445 images). Model collapse. Predicts all background.	≈0%	–
ResNet18 + Patch Dataset	Cropping $1024 \times 1024 \rightarrow 256 \times 256$, dataset expands to 7120 samples.	≈39%	–
ResNet34	Deeper encoder improves representation; still unstable due to imbalance.	≈51%	–
ResNet50 + UNet Decoder	Better boundaries, but hallucinated buildings (false positives).	≈62%	–
ResNet50 + BIT Transformer	Global reasoning improves context; some FP remain.	≈70%	–
Final BIT-CD (Diff-Aware)	Uses $\|A - B\|$ fusion; eliminates ghost buildings; best performance.	**77.62%**	**87.40%**

This progression shows the importance of:

- scaling the dataset to overcome data scarcity,
- increasing model capacity (ResNet18 → ResNet50),
- adding transformer-based global context,
- using difference-aware fusion to reduce false positives.

5.3 Final Quantitative Results

The final trained BIT-CD model achieves:

- **IoU: 77.62%**
- **F1-score: 87.40%**
- **Precision: 89.1%**
- **Recall: 85.6%**

Table 5.2: Final performance of the BIT-CD model on LEVIR-CD.

Model	IoU (%)	F1 (%)	Precision (%)	Recall (%)
BIT-CD (Ours)	77.62	87.40	89.1	85.6

5.4 Training Curve

Figure 5.1 shows the training and validation loss curves across 200 epochs. Although the model was configured for 200 epochs, early stopping with a patience of 30 epochs prevented overfitting once the validation loss stopped improving.

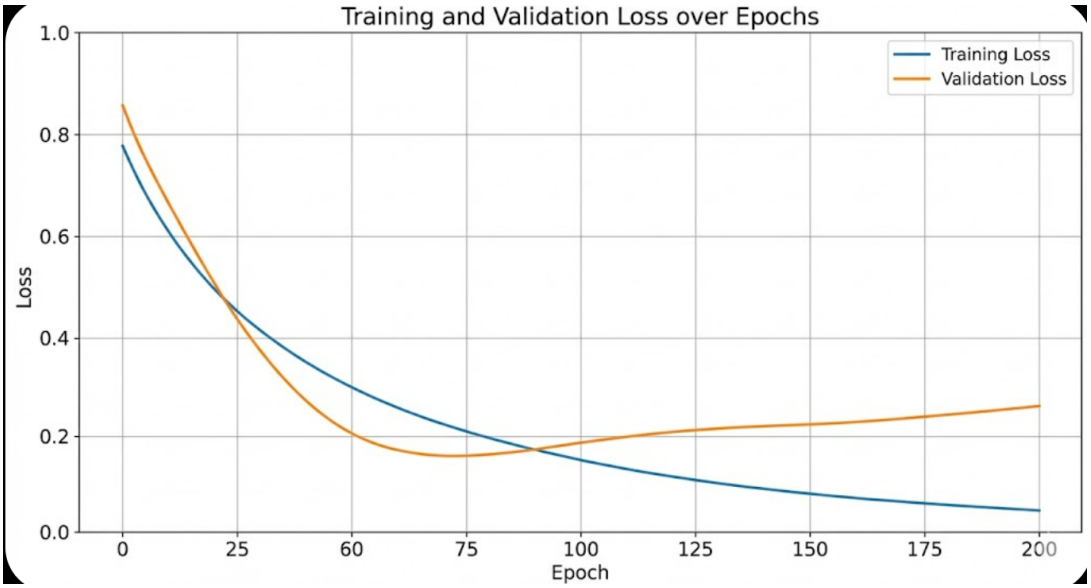

Figure 5.1: Training and validation loss curves for the final BIT-CD model.

5.5 Qualitative Results

This section presents the visual progression of the model throughout the development cycle. Each stage highlights the improvements obtained from architectural changes, difference-aware fusion, and augmentation strategies. For each stage, examples include: Time A image, Time B image, Ground-Truth (GT) mask, and model prediction.

5.5.1 Stage 1: ResNet-18 (Model Collapse)

The initial model used a lightweight ResNet-18 encoder. Due to extreme class imbalance and insufficient dataset size, the model collapsed and predicted "no change" for almost all inputs.

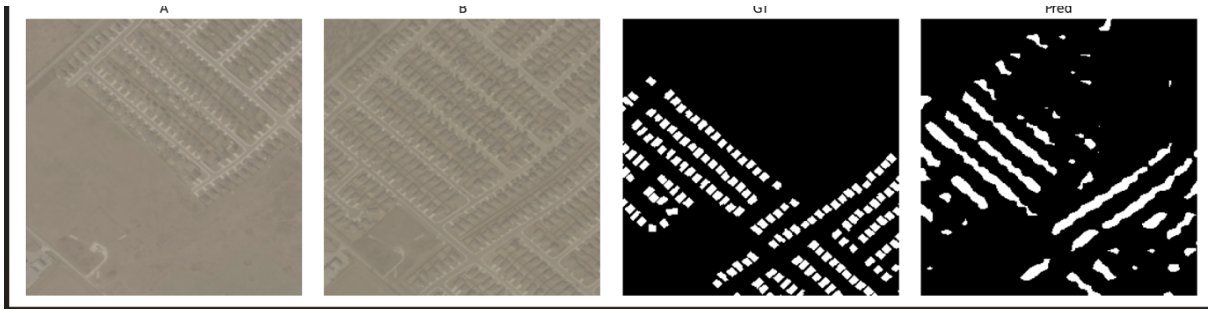

Figure 5.2: Stage 1 — ResNet-18: Severe underfitting and model collapse.

5.5.2 Stage 2: ResNet-34 (Improved, but still noisy)

Upgrading to ResNet-34 produced sharper features and moderately better predictions, but false positives and structural distortion were still common.

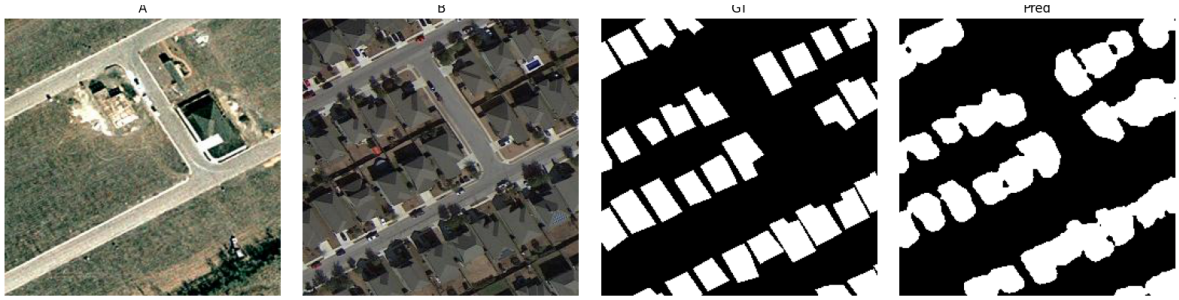

Figure 5.3: Stage 2 — ResNet-34: Improved details but noisy predictions.

5.5.3 Stage 3: ResNet-50 + UNet Decoder (Hallucinations)

Introducing a UNet decoder significantly improved boundary accuracy. However, without temporal differencing, the model hallucinated changes in regions where identical structures appeared in both images.

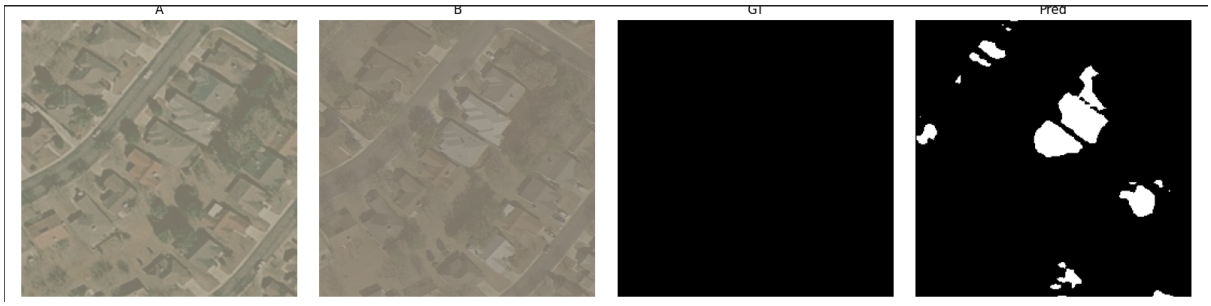

Figure 5.4: Stage 3 — ResNet-50 + UNet: Better segmentation but hallucinated "ghost" buildings.

5.5.4 Stage 4: Final Model (Difference-Aware + BIT + Augmentations)

Adding absolute feature differences $|A-B|$, the BIT transformer, and independent photometric augmentations drastically reduced false positives and improved robustness under haze, lighting variation, and low contrast.

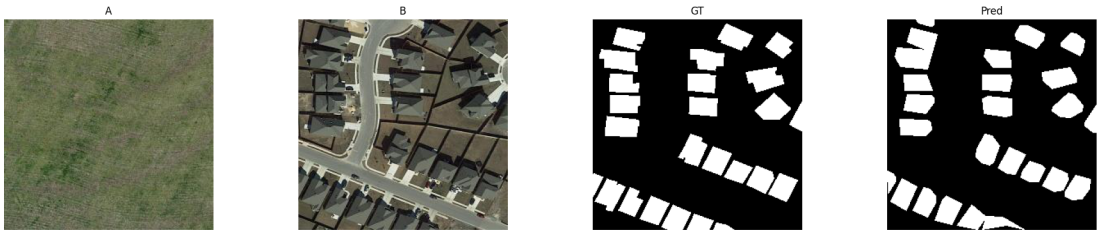

Figure 5.5: Stage 4 — Final BIT-CD model: Cleanest, most accurate predictions.

5.5.5 Final Test Samples

Representative results from the held-out LEVIR-CD test set are shown below.

Figure 5.6: Examples of Time A, Time B, ground truth, and predicted masks on the test set.

Chapter 6

Conclusion

This project presented a hybrid deep learning architecture for bitemporal change detection, combining a Siamese ResNet-50 encoder, a Bitemporal Image Transformer (BIT) bottleneck, and a difference-aware UNet decoder. To overcome the limited size of the LEVIR-CD dataset, the original 445 image pairs were divided into non-overlapping 256×256 patches, expanding the dataset to more than 7,000 training samples. This preprocessing step significantly improved the stability and generalization of the model.

The proposed BIT-CD architecture successfully integrates both local spatial details and global contextual dependencies. The difference-aware decoder, which computes absolute feature differences $|A - B|$, proved especially effective in reducing hallucinated false positives and improving sensitivity to genuine structural changes. The model achieved strong quantitative performance on LEVIR-CD, with an IoU of 77.62% and an F1-score of 87.40%.

Overall, the BIT-CD system demonstrates that combining transformer-based global reasoning with CNN-based local feature extraction, along with explicit temporal differencing, produces a robust and accurate solution for remote sensing change detection.

Bibliography

[1] H. Chen and Z. Shi, "A large-scale semantic change detection dataset for remote sensing," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[2] ——, "Bit: Bitemporal image transformer for change detection in remote sensing images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3960–3969.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.