buildings

MDPI

Article

Building Change Detection in Aerial Imagery Using End-to-End Deep Learning Semantic Segmentation Techniques

Tee-Ann Teo * and Pei-Cheng Chen

Department of Civil Engineering, National Yang Ming Chiao Tung University, Hsinchu City 300093, Taiwan; a4623217.en07@nycu.edu.tw

* Correspondence: tateo@nycu.edu.tw

Abstract: Automatic building change detection is essential for updating geospatial data, urban planning, and land use management. The objective of this study is to propose a transformer-based UNet-like framework for end-to-end building change detection, integrating multi-temporal and multi-source data to improve efficiency and accuracy. Unlike conventional methods that focus on either spectral imagery or digital surface models (DSMs), the proposed method combines RGB color imagery, DSMs, and building vector maps in a three-branch Siamese architecture to enhance spatial, spectral, and elevation-based feature extraction. We chose Hsinchu, Taiwan as the experimental site and used 1:1000 digital topographic maps and airborne imagery from 2017, 2020, and 2023. The experimental results demonstrated that the data fusion model significantly outperforms other data combinations, achieving higher accuracy and robustness in detecting building changes. The RGB images provide spectral and texture details, DSMs offer structural and elevation context, and the building vector map enhances semantic consistency. This research advances building change detection by introducing a fully transformer-based model for end-to-end change detection, incorporating diverse geospatial data sources, and improving accuracy over traditional CNN-based methods. The proposed framework offers a scalable and automated solution for modern mapping workflows, contributing to more efficient geospatial data updating and urban monitoring.

Keywords: buildings; change detection; deep learning; map updating

Academic Editor: Antonio Caggiano

Received: 18 January 2025
Revised: 17 February 2025
Accepted: 19 February 2025
Published: 23 February 2025

Citation: Teo, T.-A.; Chen, P.-C. Building Change Detection in Aerial Imagery Using End-to-End Deep Learning Semantic Segmentation Techniques. *Buildings* **2025**, *15*, 695. https://doi.org/10.3390/buildings15050695

1. Introduction

1.1. Motivation

Rapid urbanization has significantly increased the demand for frequent map updates, particularly in urban areas where construction and demolition occur frequently. Accurate and up-to-date building maps are essential for effective urban management. However, manually updating these maps is a labor-intensive and time-consuming process [1,2]. A more efficient approach for map updating could be focusing on changed areas. Advancements in geospatial mapping technologies have introduced automated and semi-automated processes, which show significant potential to improve map production processes. For instance, high-resolution remote sensing images from satellite or aerial photography can provide extensive coverage. Therefore, change detection techniques can be applied to identify hotspots that require updates, thereby enhancing the efficiency of urban maintenance and management efforts. The study [3] proposed a change detection-assisted mapping workflow and compared it with manual mapping efforts. The study demonstrated that incorporating change detection techniques significantly reduced mapping time while also

leading to slight improvements in the accuracy and quality of map updates. These findings show the potential of automated methods to enhance traditional mapping workflows, making them more efficient and scalable for large-scale urban environments.

1.2. Previous Studies

Modern change detection techniques have increasingly embraced machine learning approaches. The research [4] undertook a comprehensive review, tracing its evolution from early perceptron to the adoption of deep learning (DL) recently for remote sensing tasks. There has been a significant surge in publications related to DL in remote sensing since 2015 [5]. This growth can be mainly attributed to rapid advancements in convolutional neural network (CNN) architectures, which have dramatically improved the ability to analyze complex large-area remote sensing data. Among the most notable achievements has been performing very high-resolution (VHR) imagery semantic segmentation with remarkable precision.

In [6], Si Salah et al. highlighted the capacity and applications of change detection techniques in remote sensing. The authors stated that change detection techniques cannot effectively address complex remote sensing tasks without a well-designed workflow. Several critical dimensions, including input data, analysis units, targets, change categories, and temporal resolution, must be considered (Table 1).

Table 1. Dimensions in change detection [6].

Dimensions	Instance	Characteristic
Input data	Vector	Vector to raster conversion
	Raster	Color mode, spectral resolution, spatial resolution
Analysis unit	Pixel	Compare image pixels
	Object	Compare groups of contiguous pixels
Target	Buildings	Instance segmentation, semantic segmentation, panoptic segmentation
	Vegetation	
	Everything	
Change categories	Binary	Changed/unchanged
	Multi-class	Unchanged/new/destroyed/partially changed/modification
Temporal resolution	Bi-temporal	Two time epochs
	Multi-temporal	Series of time epochs
Operator	Statistical analysis	Calculating indices/statistical test
	Computer vision	Color/texture/shape
	Machine learning	SVM/deep learning

Semantic segmentation techniques have been widely used in image recognition. Applying such techniques for building change detection using aerial imagery enhances automation and accuracy. For instance, Ref. [4] used CNNs to detect building changes from high-resolution aerial imagery. Their results demonstrated that DL techniques can accurately identify change areas in images acquired at different time points. Furthermore,

Ref. [7] employed semantic segmentation to monitor urban development; the authors highlighted the superior performance of DL techniques in extracting building boundaries and classifying objects. These studies underscore the advantages of semantic segmentation in handling large-scale and complex imagery, particularly when images contain multiple object types.

The essential process of change detection algorithms involves extracting high-quality features and comparing their differences to identify areas of change. In [8], Jiang et al. systematically reviewed change detection methods for aerial imagery. Their analysis of 116 studies revealed the primary network architectures used for change detection, which included CNNs, generative adversarial networks (GANs), autoencoders (AEs), and recurrent neural networks (RNNs). Among these, CNNs are the most commonly used architecture, accounting for 62% of the studies, underscoring their widespread application in image change detection tasks. The popularity of CNNs can be attributed to their strengths in image processing and feature extraction. Moreover, 17% of the studies have used RNNs and GANs, indicating their specific applications in change detection. RNNs are typically employed for temporal data analysis, while GANs are used to generate and synthesize images to facilitate detection. AEs have been the least utilized; only 4% of studies have used this architecture, possibly due to their primary use in data compression and dimensionality reduction, which limits their broader application in change detection tasks.

Integrating different data sources, such as digital surface models (DSMs), with other optical imagery is crucial for improving change detection methodologies. DSMs provide elevation information that helps distinguish between buildings and ground features. For instance, Ref. [9] demonstrated the effectiveness of combining optical imagery with DSM data for building change detection; the authors revealed that including elevation information significantly enhanced the accuracy of identifying newly constructed buildings. Integrating multi-source data enables models to better recognize building changes over time, thus improving the overall reliability and precision of change detection tasks.

In this context, Siamese architecture offers a powerful approach to data fusion in change detection. In [10], Daudt et al. conducted experiments using a fully convolutional network to integrate RGB and synthetic aperture radar (SAR) imagery. They proposed three architectures: fully convolutional early fusion, fully convolutional Siamese concatenation, and fully convolutional Siamese difference. Among these, Siamese architectures stand out due to their ability to process two inputs through identical network branches with shared weights. This design ensures consistent feature extraction across both inputs, reducing bias and ensuring that the focus is on meaningful changes during comparison. The explicit comparison step inherent in Siamese architectures is particularly advantageous for multi-source data fusion, as it highlights differences between inputs and improves the sensitivity to subtle variations.

However, Ref. [11] highlighted the limitations of traditional two-dimensional (2D) change detection approaches, which mainly focus on detecting planimetric changes. The 2D change detection proved insufficient for applications requiring volumetric or topographic analysis, such as urban development monitoring, forest biomass estimation, or landform alteration [12,13]. The authors solved the issue by incorporating multi-source remote sensing data, such as optical, DSM, and SAR imagery into semantic segmentation models. The fusion of these data sources, combined with the robust feature extraction and comparison capabilities of Siamese architectures, represents a significant step forward in improving the precision and applicability of change detection in complex urban and environmental contexts.

Although CNNs have been widely and successfully applied in DL for image analysis, a notable limitation is that they treat all regions equally during convolution, focusing only on

global feature extraction. To overcome this issue, the vision transformer (ViT), introduced by [14], employs a self-attention mechanism that simultaneously considers both global and local features. A major breakthrough in transformer-based image segmentation was TransUNet [15], which demonstrated that the transformer-based model beat traditional CNN-based architectures in medical image segmentation. The success of TransUNet laid the foundation for transformer-driven image semantic segmentation models. According to a review by [16], numerous models derived from the ViT architecture have been developed and applied to various data sources, such as SAR images, hyperspectral images, and VHR images. These models can perform tasks such as image classification, object detection, and change detection. The ViT's superior ability to extract fine-grained features allows it to detect subtle changes in imagery.

When it comes to the applications of transformers in change detection, Ref. [17] proposed the bi-temporal image transformer (BIT) model, which replaced the transformer architecture as the traditional UNet bottleneck feature analyzer. In the BIT model, ResNet18 is employed for feature extraction, while the transformer structure focuses on analyzing features and contextual relationships between bi-temporal images. This hybrid approach effectively combines the strengths of both methods, with CNN excelling in capturing fine-grained spatial features and transformers, thus yielding superior contextual understanding. ChangeFormer [18] is a bi-temporal change detection framework that uses a Siamese transformer as the encoder and a multi-layer perceptron (MLP) as the decoder. The hierarchical multi-layer features serve as the optimal distance metrics between the features extracted from bi-temporal images to effectively form representations. The MLP decoder then performs upsampling and classification tasks. The authors compared the proposed method with the BIT model and demonstrated that their approach outperformed BIT. The results indicated that transformers deliver superior performance as feature encoders rather than decoders in hybrid models. This finding suggests that transformers excel at capturing and processing complex spatial and temporal relationships in input data. At the same time, simpler decoders such as MLPs can efficiently handle subsequent upsampling and classification stages. Pure transformer-based models, such as SwinSUNet [19], are built upon the Swin transformer [20], incorporating a windows-based multi-head self-attention (W-MSA) mechanism. This design allows the model to efficiently process input images at multiple scales, making it adaptable to varying image resolutions. The flexibility of the Swin transformer makes it an ideal choice for tasks involving diverse input sizes, such as training on small images (e.g., 512 $\times$ 512) and predicting on larger images (e.g., 1024 $\times$ 1024). For our research, this scalability proved particularly advantageous because, in VHR images, buildings often span multiple small image patches. The Swin transformer's hierarchical design effectively addresses this challenge by enabling the model to process local details and broader contextual information seamlessly.

The research [21] further advanced transformer-based architectures by introducing UNetFormer and FT-UNetFormer for efficient semantic segmentation on multiple datasets. The key distinction between these models lies in their encoder structure. UNetFormer uses a hybrid CNN-Transformer encoder, while FT-UNetFormer is fully transformer-based. Their study compared SwinUNet and TransUNet with their proposed methods, highlighting that both SwinUNet (41.1 M parameters) and TransUNet (90.7 M parameters) require significantly larger model sizes, whereas UNetFormer achieves better mean intersection over union (mIoU) scores with only 11.7 M parameters. Although the exact model scale of FT-UNetFormer was not specified (14.2 M parameters in our case), the study considered it as a potentially more powerful fully transformer-based model. The Swin transformer further improves the architecture by incorporating multi-level feature extraction to accommodate multi-scale variations. This enables the model to extract features at different scales, which is

crucial for detecting objects of varying sizes and shapes, such as buildings in aerial imagery. In addition to the enhanced encoder, the decoder is also designed based on the ViT structure, enabling the analysis of both global and local features. The decoder employs upsampling and feature reconstruction, facilitating the connection of high-dimensional features and allowing the effective fusion of semantic contextual information from surrounding areas. This enhanced feature integration improves the accuracy of change detection. Given these advantages, FT-UNetFormer was chosen for this study due to its efficiency, scalability, and improved feature extraction capabilities. It provides a balance between computational cost and model performance, making it an ideal choice for large-scale urban monitoring and geospatial data analysis.

1.3. Objectives and Contributions

This study aimed to develop an end-to-end deep learning framework for building change detection to facilitate map updating. The proposed method uses DL techniques to extract deep features from multispectral images, elevation models, and existing building polygons for accurate and efficient building change detection. The architecture is based on FT-UNetFormer, which integrates the strengths of the UNet structure for precise localization and the transformer mechanism for global context modeling. This data fusion approach enables the model to capture both fine-grained spatial details and long-range dependencies, enhancing the accuracy of detecting newly constructed, modified, and demolished buildings. This study's contribution is streamlining the process of building change detection by experimenting with different input data combinations and evaluating model performance. By enhancing prediction accuracy and reliability, this study addresses the challenges of timely and accurate map updating, thus providing valuable support for urban planning and management. The aforementioned details have been summarized in Table 2, which demonstrates the design of the research.

Table 2. Dimensions of the proposed change detection scheme.

Dimensions	Instance	Characteristic
Input data	Vector	1/1000 scale building map polygons
	Raster	Color mode: RGB Elevation mode: DSM
		Spatial resolution: 10 cm
		Spatial resolution: 8 bit, 16 bit
Analysis unit	Pixel	Accuracy, precision, recall, F1 score
	Object	Precision, recall, F1 score, Object area distribution analysis
Target	Buildings	Semantic segmentation
Change categories	Multi-class	Deconstruction, construction, no change, non-building
Temporal resolution	Bi-temporal	2017–2020 (train/val.) 2020–2023 (test)
Operator	Machine learning	FT-UNetFormer

The major contribution of this study is the extension of FT-UNetFormer from building detection to building change detection, enabling an end-to-end architecture that directly identifies changed areas. Additionally, this study evaluates the performance of building change detection using various data fusion approaches. This research addresses key gaps

in the field, as most existing methods primarily rely on a single data type, such as spectral imagery, while fusion-based approaches integrating spectral imagery with DSM remain limited. To enhance efficiency and accuracy, this study introduces a novel approach by integrating three diverse data types—RGB imagery, DSM, and building maps—into a unified one-stage change detection model.

2. Materials and Methods

2.1. Materials and Study Area

The experimental area for this study is located in the East District of Hsinchu City, Taiwan, an urban region characterized by rapid infrastructure development and frequent changes in the built environment. This dynamic setting provided a suitable context to evaluate the effectiveness of the proposed building change detection method.

To support the analysis, multi-temporal aerial imagery was acquired in 2017, 2020, and 2023. These 10 cm high-resolution aerial images were supplemented with corresponding 1:1000 digital topographic maps, including building polygon layers. The availability of multi-year datasets enabled us to comprehensively assess the model's ability to detect newly constructed, modified, and demolished buildings over time. Table 3 provides a detailed summary of the aerial imagery and the associated topographic maps, including information on the acquisition dates, spatial resolution, and relevant metadata. Figure 1 presents examples of the aerial images and corresponding building map polygons from 2017, 2020, and 2023. Two areas have been marked to emphasize the specific building change events: changes between 2017 and 2020 have been highlighted in red and those from 2020 to 2023 have been marked in orange.

Table 3. A summary of the parameters used in building change detection dataset.

	2017	**2020**	**2023**
Number of images	491	293	201
Sensor	DMC II	DMC II	DMC III
Avg. flight height	1280 m	1540 m	2240 m
Avg. spatial resolution	7.83 cm/pixel	9.39 cm/pixel	9.48 cm/pixel
Number of buildings	99,167 polygons	100,125 polygons	101,184 polygons

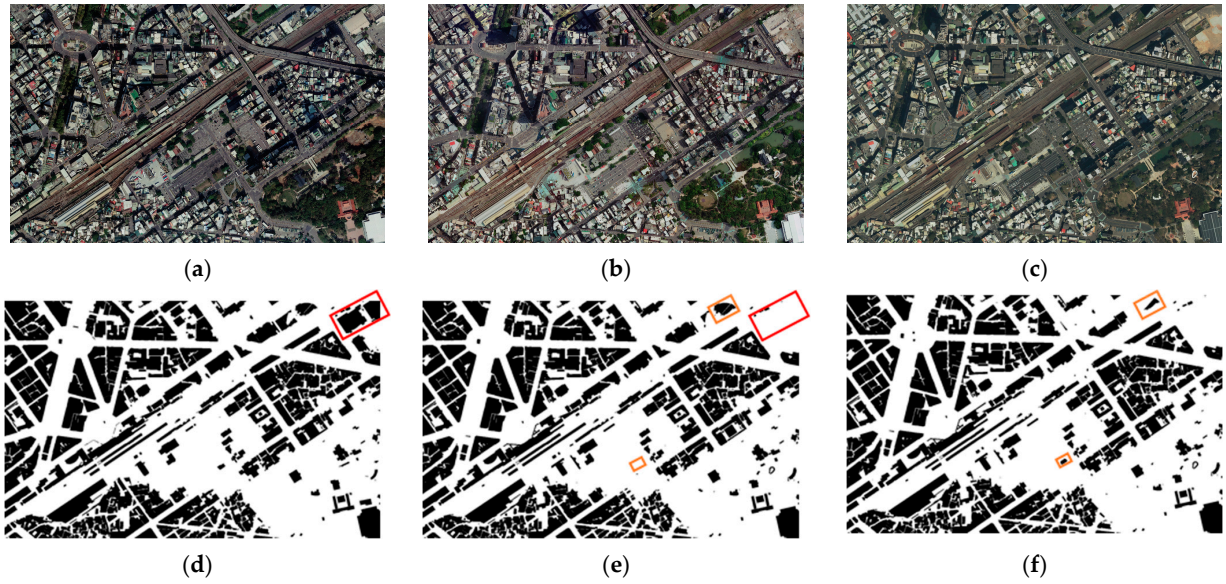

Figure 1. Aerial images and building polygons for a 1:1000 topographic map: (**a**–**c**) aerial images in 2017, 2020, and 2023 and (**d**–**f**) building polygons in 2017, 2020, and 2023. The red and orange boxes indicate some building change events.

The total coverage area spans 3552 hectares, distributed across 74 sheets of 1:1000 topographic maps. Each map sheet corresponds to a specific geographic section of the test area; this enabled the precise alignment of image data and vector building maps for training and validation purposes. Integrating aerial imagery with topographic maps enhanced the model's spatial analysis and change detection capacity. Figure 2 presents the spatial distribution of the 74 topographic map sheets, offering a visual overview of the study area and its partitioning into smaller mapping units.

Figure 2. Distribution of 74 frames of 1:1000 digital topographic maps.

2.2. Data Preprocessing

Data preprocessing is a crucial step in preparing input data for the deep learning-based building change detection model. This process includes generating DSMs and true orthophotos from aerial images, converting building vector polygons to raster data, and creating training datasets for the DL model. These preprocesses involve the following steps:

1. Generation of DSMs and true orthophotos. High-resolution DSMs and true orthophotos are generated through dense image matching to eliminate distortions and ensure accurate alignment. Traditional orthophotos (Figure 3a) are corrected using a digital terrain model (DTM), which only accounts for ground elevation. However, this approach does not consider the relief displacement caused by elevated structures such as buildings, leading to positional distortions in building outlines. This study adopts the image dense matching technique to generate a DSM that accurately represents surface elevation, including buildings, trees, and other structures. The DSM is then used to produce a true orthophoto (Figure 3b) [22,23], which provides a full nadir view, thereby eliminating the relief displacement of buildings. This process involves robust aerial triangulation, aligning images across different time periods using a consistent set of ground control points (GCPs) to ensure precise alignment of multi-temporal datasets. The aerial images used are cloud-free, ensuring clear and accurate data. Additionally, a mosaic image is generated by combining multiple source images captured from different viewpoints. This mosaic reconstruction effectively eliminates occluded regions, enhancing the completeness and reliability of the dataset. Consequently, the building edges are more precisely aligned, which improves the spatial accuracy of change detection. Figure 3c illustrates the differences between traditional and

true orthophotos by overlaying their respective color bands. This visual comparison highlights the positional shifts caused by relief displacement, which can be effectively resolved in true orthophotos. The DSM (Figure 3d) is created using the image dense matching method, which identifies corresponding points between multi-view aerial images to calculate elevation values. Unlike a DTM, which only represents bare earth, a DSM considers all objects' surface heights, including buildings and vegetation. The DSM plays a vital role in data preprocessing by providing critical elevation information for building change detection. It identifies changes in building height and helps differentiate newly constructed buildings from existing structures.

2. By refining building edges and ensuring precise spatial alignment, this methodology significantly improves the accuracy of change detection, minimizing errors caused by inconsistencies in aerial image acquisitions over time. Figure 3c illustrates the differences between traditional and true orthophotos by overlaying their respective color bands. This visual comparison highlights the positional shifts caused by relief displacement, which can be effectively resolved in true orthophotos. The DSM (Figure 3d) is created using the image dense matching method, which identifies corresponding points between multi-view aerial images to calculate elevation values. Unlike a DTM, which only represents bare earth, a DSM considers all objects' surface heights, including buildings and vegetation. The DSM plays a vital role in data preprocessing by providing critical elevation information for building change detection. It identifies changes in building height and helps differentiate newly constructed buildings from existing structures.
3. Vector-to-raster conversion. Digital building map (DBM) vector data are converted into raster format to meet the input requirements of the DL model, ensuring consistency in data representation. DBMs contain vectorized building footprints from previous periods and serve as reference layers to highlight regions where changes have occurred. To ensure accurate feature extraction, DBMs are first aligned to the corresponding map frame and then converted to raster format with the same spatial resolution (i.e., 10 cm). This process preserves high-resolution details while maintaining computational efficiency, providing a balance between processing speed and sufficient spatial context for precise change detection.
4. Training dataset preparation. Multi-temporal datasets, including RGB imagery, DSMs, and building maps, are carefully organized and labeled to create training and validation datasets suitable for the model's learning process. The dataset is divided into two pairs, 2017–2020 and 2020–2023, to evaluate the performance across different time intervals. The 2017–2020 dataset is used for training and validation, allowing the model to learn and optimize itself to detect changes. Meanwhile, the 2020–2023 dataset is reserved as an independent test set, serving as a new temporal pair to assess the model's generalization capability and robustness in identifying building changes over a different temporal interval.

Data augmentation techniques, such as random rotations, flips, color, and brightness adjustments, were applied to enhance model robustness and generalization. This data preparation approach ensured that the model was exposed to diverse input conditions, allowing it to effectively identify building changes in wide-ranging environmental and temporal scenarios. Figure 4 presents an example patch of the RGB, DSMs, and DBMs for 2017, 2020, and 2023, illustrating the elevation and spectral differences that the model learned to identify. This specific patch showcases significant changes across the dataset time periods, which are marked with red and black blocks to highlight areas of transformation.

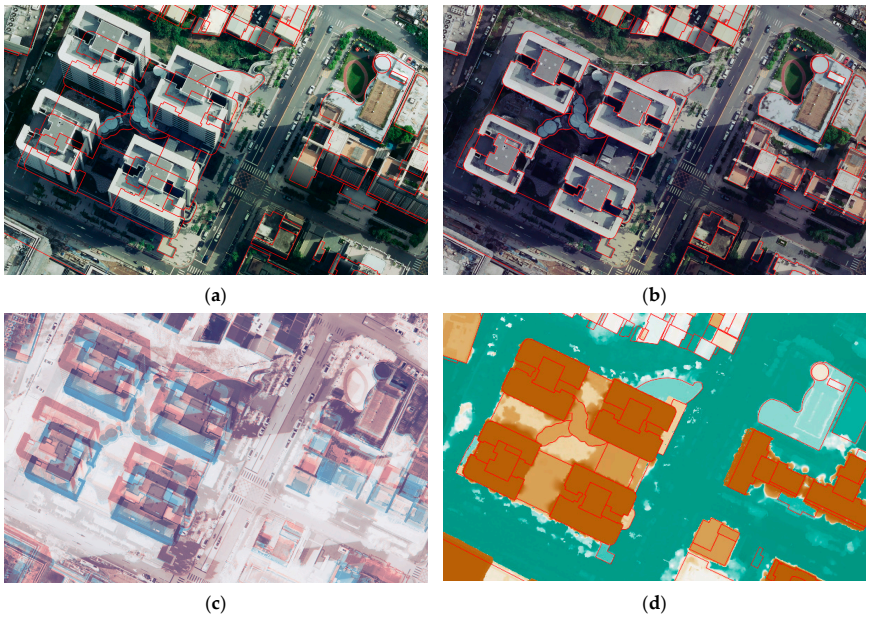

Figure 3. A comparison between traditional and true orthophotos. The red line boxes represent the building boundary: (**a**) a traditional orthophoto with significant building displacement; (**b**) a true orthophoto without building displacement; (**c**) superimposed traditional (red) and true (blue) orthophotos; and (**d**) a DSM and building boundary.

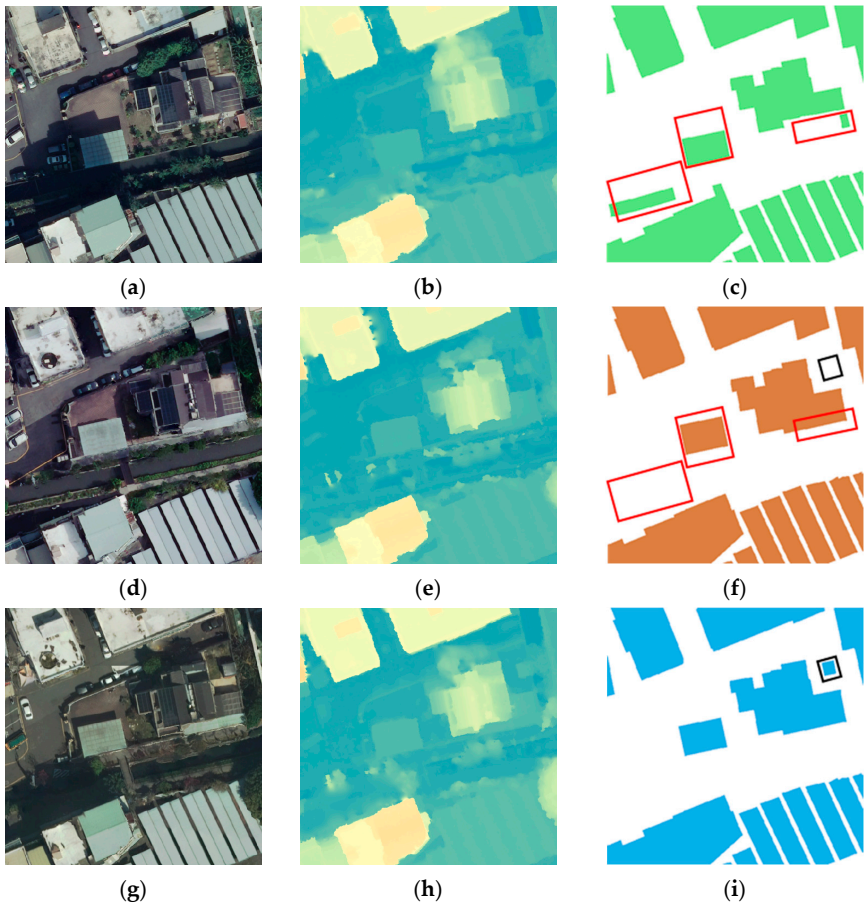

Figure 4. An example of the training dataset: (**a**,**d**,**g**) aerial images in 2017, 2020, and 2023; (**b**,**e**,**h**) DSMs in 2017, 2020, and 2023; and (**c**,**f**,**i**) building polygons in 2017, 2020, and 2023. The red and black boxes indicate building change events.

2.3. Methodology

This study employed an end-to-end deep learning semantic segmentation technique to detect areas of building change. The end-to-end approach is a one-step process that directly identifies building changes from multi-temporal datasets. In contrast, the non-end-to-end approach is a two-step process that first detects buildings separately for two periods and then compares the results to determine building changes. Since the end-to-end approach optimizes features specifically for change detection, it is expected to deliver better results compared with the non-end-to-end approach. In this study, orthorectified imagery from two time periods, DSMs from both periods, and building polygons from the earlier period were employed in the change detection process. Features were extracted and encoded using the transformer, and the changes were decoded and classified into three categories: unchanged, newly constructed, and demolished building areas. The overall workflow of the study is illustrated in Figure 5.

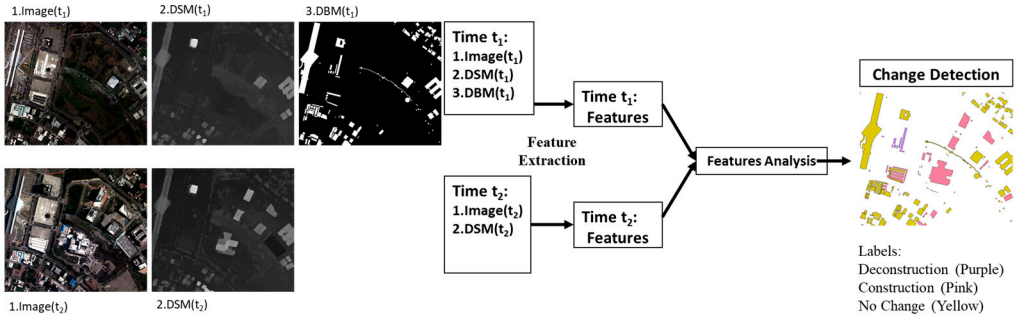

Figure 5. An end-to-end building change detection workflow.

The FT-UNetFormer, as depicted in Figure 6a, is a UNet-like architecture that is entirely made up of transformer-based components, making it a novel and highly effective model for various segmentation tasks. Unlike conventional UNet models that rely on the CNN structure for feature extraction, FT-UNetFormer leverages the Swin transformer as its encoder, thus yielding a more robust mechanism for balancing global and local feature extraction. The decoder adopted by this study was a global–local transformer block (GLTB) structure, which introduces a self-attention mechanism to effectively capture both local and global contexts. Three input datasets from two periods were used to fuse features from different datasets. Since some modifications had to be made to the model, we transformed the architecture into a three-branch Siamese structure to accommodate our dataset, as illustrated in Figure 6b. This modification enabled the model to process RGB, DSMs, and building polygons simultaneously, thereby maintaining the essential temporal comparison for the change detection task. We standardized the input shape and band composition to ensure compatibility and consistency across all the input data combinations, as detailed in Table 4. This approach allowed the model to effectively fuse multi-source and bi-temporal information, enhancing its ability to detect subtle changes with greater accuracy and robustness.

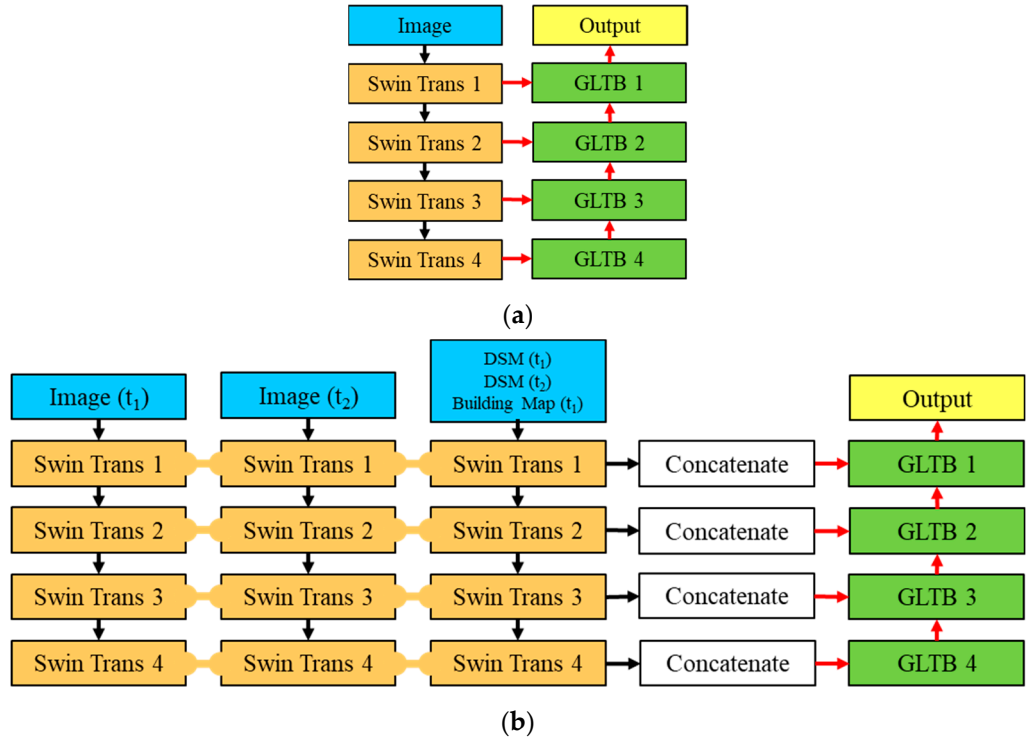

Figure 6. (**a**) FT-UNerFormer architecture and (**b**) the proposed modified architecture for RGB_DSM_Map input data combination.

Table 4. Input band composition.

	Brand 1	**Brand 2**	**Brand 3**	
	Image	**Image**	**DSM**	**Building Map**
Case 1. RGB_DSM	RGB (t_1) (3 bands)	RGB (t_2) (3 bands)	DSM (t_1), DSM (t_2), DiffDSM (3 bands)	-
Case 2. RGB_Map	RGB (t_1) (3 bands)	RGB (t_2) (3 bands)	-	Building map (t_1) (1 band × 3)
Case 3. RGB_DSM_Map	RGB (t_1) (3 bands)	RGB (t_2) (3 bands)	DSM (t_1), DSM (t_2) (2 bands)	Building map (t_1) (1 band)

The training process was run on an RTX Titan GPU (24 GB VRAM) to handle the computational demands of complex model structures and data combinations.

Table 5 highlights the hyperparameters used for training, which include essential configurations such as learning rate, batch size, and optimization strategy, ensuring the model's convergence during training. These parameters were initially derived from the original FT-UNetFormer training script for the ISPRS Vaihingen dataset and were subsequently fine-tuned through empirical experimentation to optimize performance for this study.

Data augmentation was applied dynamically during the batch loading phase to enhance the model's generalization robustness and prevent overfitting. The augmentation strategies, outlined in Table 6, included operations such as rotation, flipping, color, and brightness adjustment. These transformations enriched the training dataset by introducing variability in the input data, allowing the model to learn robust representations and become resilient to variant in orientation, scale, and lighting conditions.

Table 5. Training hyperparameters.

Parameters	Value
Input size	512 × 512 pixel
Batch size	8
Training epochs	50
Learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Backbone learning rate	1×10^{-4}
Backbone weight decay	3×10^{-5}
Best model monitor	Validation mIoU
Loss function	SoftCrossEntropy + DiceLoss

Table 6. Data augmentation during training.

Transform	RGB	DSM	DBM
Color shift	hue −0.1~0.1	-	-
Brightness shift	−0.1~0.1	−0.5~0.5	-
Rotation	−30~30°	−30~30°	−30~30°
Flip	Vertical and horizontal	Vertical and horizontal	Vertical and horizontal

3. Results and Evaluations

The performance of the three models (i.e., RGB_DSM, RGB_Map, and RGB_DSM_Map) on the independent test dataset (i.e., 2020–2023), using two types of analysis units (pixel-based and object-based evaluations), was evaluated, the results of which will be provided in this section. Pixel-based analysis assessed the accuracy of each individual pixel, whereas object-based analysis evaluated the accuracy of each changed unit. The former focused on metrics such as accuracy, precision, recall, and F1 score to assess the effectiveness of different data fusion strategies in capturing building changes; on the other hand, the latter analyzed groups of pixels to calculate the precision, recall, and F1 score, including the distribution of object areas, offering more profound insights into the potential for building change detection.

Table 7 presents the results of pixel-based analysis of the overall macro statistics evaluation of the training and validation datasets (i.e., 2017–2020) after the three training processes. The accuracy values across all three models indicate a highly reliable detection capability, with the RGB_DSM_Map model achieving the highest score of 99.74%. The precision and recall values suggest that each model effectively balanced the identification of true positive (TP) and false positive (FP) instances. As a result, we considered three models that converged to a stable status.

Table 7. Evaluation of the three models using validation datasets (2017–2020).

Statistic Metric	RGB_DSM	RGB_Map	RGB_DSM_Map
Accuracy	98.58%	99.46%	99.74%
Precision	89.28%	74.54%	86.19%
Recall	87.38%	87.55%	92.98%
F1 score	88.30%	79.19%	89.14%

3.1. Evaluation of the Test Dataset

Table 8 presents the overall macro statistics evaluation of the individual test datasets (2020–2023) after the three training processes. The assessment reveals that RGB_DSM_Map was the most balanced and robust model; it excelled in precision, recall, and F1 score due to the fusion of RGB, DSM, and DBM inputs. The RGB_DSM performed well in precision

but struggled with recall, while RGB_Map achieved the highest accuracy but fell short in recall and precision for complex scenarios. Overall, integrating diverse data sources in RGB_DSM_Map proved to be the most effective approach for change detection.

Table 8. Evaluation of the three models using individual test datasets (2020–2023).

Statistic Metric	RGB_DSM	RGB_Map	RGB_DSM_Map
Accuracy	96.75%	99.25%	99.56%
Precision	90.79%	72.21%	84.40%
Recall	63.94%	67.99%	82.17%
F1 score	69.95%	67.42%	82.94%

Figures 7–9 demonstrates a representative region exhibiting the three types of changes, allowing a comparison of the distinct contributions of the DSMs and building maps. As seen in Figures 7a, 8a and 9a, the deconstruction areas appeared unclear, but the inclusion of elevation information of the DSMs distinctly highlighted the construction areas. In contrast, the opposite characteristic can be observed in Figure 7b, Figure 8b, and Figure 9b, where the building map effectively captured the deconstruction areas that the DSMs struggled to identify. This observation highlights the complementary information provided by DSMs and the building polygons. Figure 7c, Figure 8c, and Figure 9c reinforce this synergy, as the fusion of DSMs and the building polygons produced a superior result, where the strengths of both data sources were combined for more comprehensive building change detection.

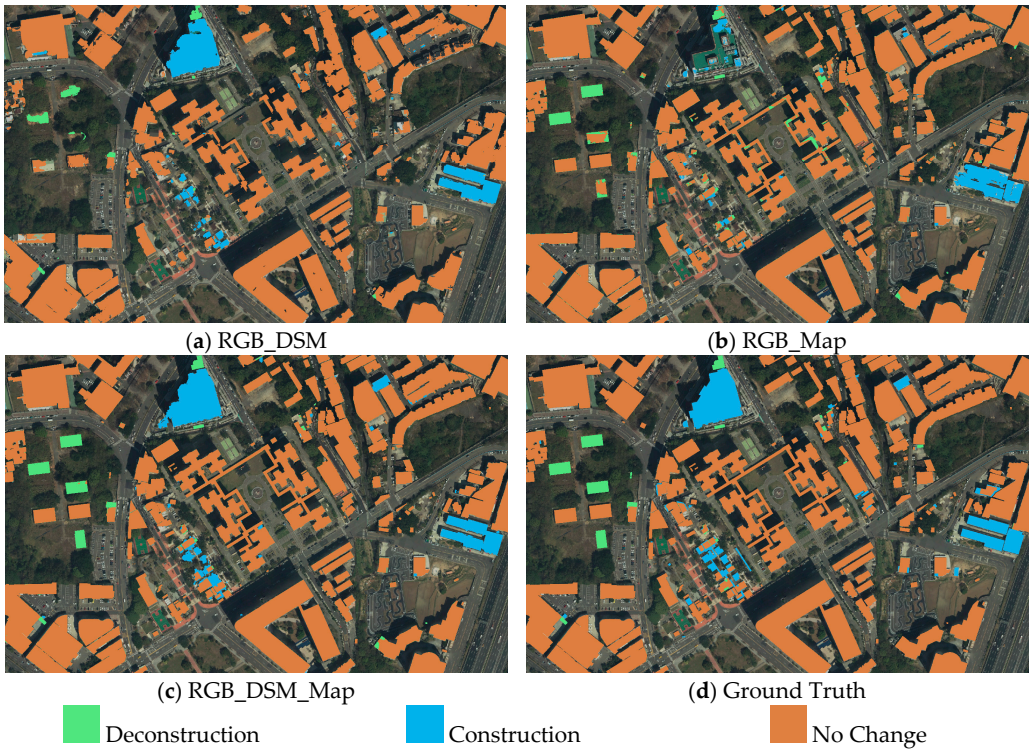

Figure 7. Comparison of change detection results for the three input combination models: (**a**) results of RGB_DSM; (**b**) results of RGB_Map; (**c**) results of RGB_DSM_Map; and (**d**) ground truth.

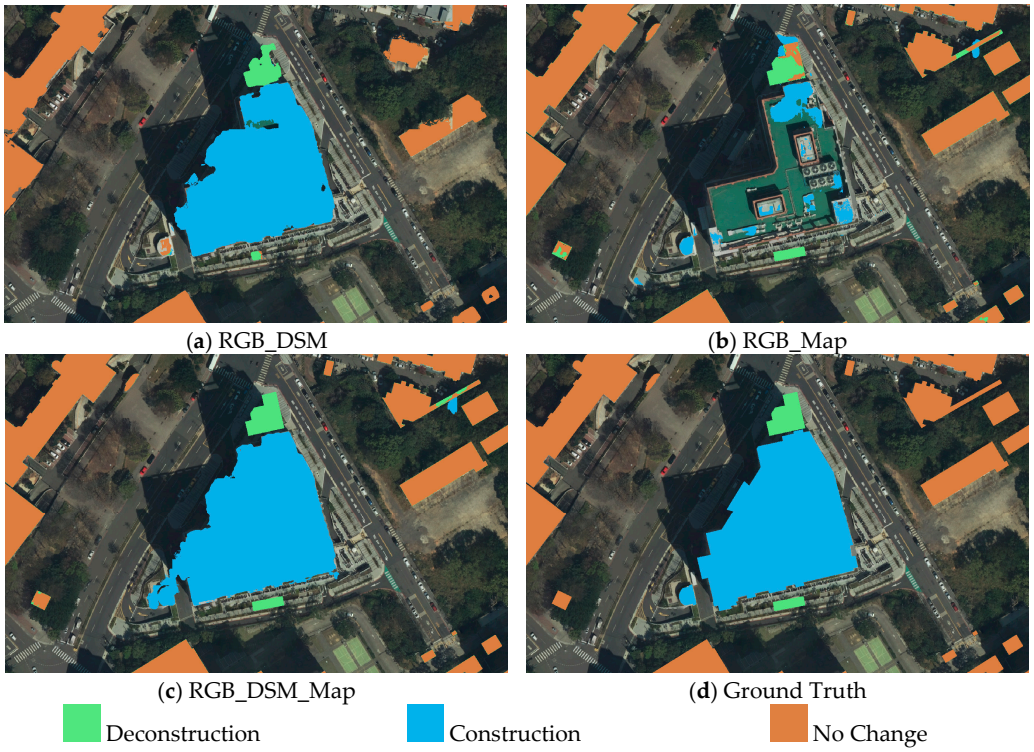

Figure 8. Close look of Figure 7 top "Construction" area: (**a**) results of RGB_DSM; (**b**) results of RGB_Map; (**c**) results of RGB_DSM_Map; and (**d**) ground truth.

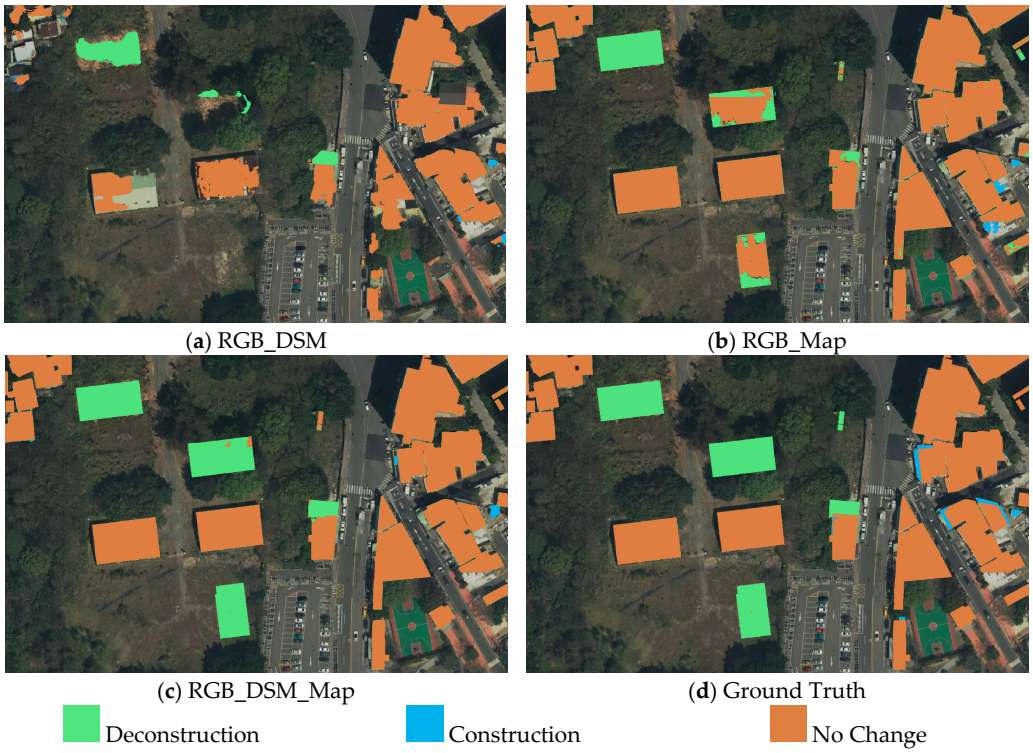

Figure 9. Close look of Figure 7 left "Deconstruction" area: (**a**) results of RGB_DSM; (**b**) results of RGB_Map; (**c**) results of RGB_DSM_Map; and (**d**) ground truth.

3.2. Pixel-Based Evaluation

The performance of the three models through pixel-based evaluation was assessed, and the results will be detailed in this section. The confusion matrix and statistic metrics of the three different input data combination models were evaluated, such as accuracy, precision,

recall, and F1 score in Table 9, Table 10, and Table 11. By analyzing how each model predicted different change categories, including non-building, deconstruction, construction, and no change, we gained insights into the models' strengths and limitations in building change detection.

Table 9. Pixel-based evaluation: confusion matrix and metrics for the RGB_DSM model.

RGB_DSM		Ground Truth Label			
		Non-Building	**Deconstruction**	**Construction**	**No Change**
Prediction	Non-building	1,619,265,182	1,345,845	1,224,553	22,211,203
	Deconstruction	7,249,205	3,864,419	77,089	1,330,942
	Construction	7,874,732	50,889	6,923,744	1,205,508
	No change	105,352,114	662,825	1,768,669	536,774,162
Accuracy		93.73%	99.54%	99.47%	94.28%
Precision		93.08%	65.23%	69.28%	95.59%
Recall		98.49%	30.86%	43.13%	83.28%
F1 score		95.71%	41.90%	53.16%	89.01%

Table 10. Pixel-based evaluation: confusion matrix and metrics for the RGB_Map model.

RGB_Map		Ground Truth Label			
		Non-Building	**Deconstruction**	**Construction**	**No Change**
Prediction	Non-building	1,642,076,980	5,749,434	2,727,417	695,460
	Deconstruction	390,948	6,869,497	1021	5,270,367
	Construction	12,363,996	88,746	3,601,456	675
	No change	698,607	7,953,613	16,328	636,514,232
Accuracy		99.03%	99.16%	99.35%	99.37%
Precision		99.19%	33.25%	56.75%	99.07%
Recall		99.45%	54.82%	22.43%	98.66%
F1 score		99.32%	41.39%	32.15%	98.86%

Table 11. Pixel-based evaluation: confusion matrix and metrics for the RGB_DSM_Map model.

RGB_DSM_Map		Ground Truth Label			
		Non-Building	**Deconstruction**	**Construction**	**No Change**
Prediction	Non-building	1,642,078,698	2,022,068	2,681,525	1,366,925
	Deconstruction	27,711	9,093,617	10,797	3,399,708
	Construction	6,857,917	26,255	9,164,715	5986
	No change	470,411	2,959,945	475,486	940,903,605
Accuracy		99.42%	99.64%	99.57%	99.63%
Precision		99.55%	64.49%	74.31%	99.26%
Recall		99.63%	72.56%	57.08%	99.39%
F1 score		99.59%	68.29%	64.57%	99.33%

Upon comparing RGB_DSM and RGB_Map, we found that both models performed well in non-building and no change categories. The RGB_Map model achieved higher precision and recall because of its reliance on building maps. Moreover, it exhibited better deconstruction change identification. The RGB_DSM model performed better in detecting construction changes because of its elevation information. When evaluating FPs and false negatives (FNs), the RGB_Map model tended to misclassify the deconstruction areas as either non-building or no change. As for the RGB_DSM model, it performed better in

recognizing both deconstruction and construction changes. However, when it came to the no change category, the RGB_Map model outperformed RGB_DSM. RGB_DSM_Map, integrating the three input data, combined the strengths of both approaches. Adding DSM data to the building map enhanced the model's ability to distinguish elevation-related changes, while the map provided semantic context for improved localization.

Among the three models, the RGB_DSM_Map model showed the highest accuracy in all classes. Integrating DSMs and building maps effectively addressed the weaknesses in the other models, particularly for detecting construction and deconstruction areas. The results reveal that fusing multiple data sources significantly enhanced the model's ability to identify building changes and improve reliability. In this section, it was demonstrated that the RGB_DSM_Map combination outperformed the other two models. Consequently, in the next section, we will focus on evaluating the changed units from the RGB_DSM_Map model using object-based analysis.

3.3. Object-Based Evaluation

Object-based evaluation involves grouping the detected changes and comparing them with the ground truth. This method provided a more detailed analysis of the model's performance by focusing on individual changes, allowing us to identify strengths and weaknesses in specific scenarios. By examining the number of instances of TPs, FPs, and FNs, the evaluation highlighted areas where the model accurately identified changes, including areas where improvements were necessary, particularly when distinguishing between subtle or overlapping change types. Figure 8 outlines the details of the evaluation process, showcasing how the ground truth labels (Figure 10a) were compared with the predicted results (Figure 10b). This study classified the evaluation results (Figure 10c) into three categories:

1. A prediction was classified as a TP when the detected change aligned correctly with the ground truth label. For instance, if the model predicted "deconstruction" for a region where the ground truth also indicated "deconstruction", this was considered a correct detection, a TP.
2. Conversely, an FN occurred when the model failed to detect a change present in the ground truth. For example, if the ground truth indicated "construction" in a specific area but the model predicted "non-building", the change was overlooked, resulting in an FN.
3. Finally, an FP instance occurred when the model incorrectly predicted a change that did not exist in the ground truth. Examples of this included predicting "construction" for a region labeled as "non-building" or predicting "deconstruction" for an area marked as "no change".

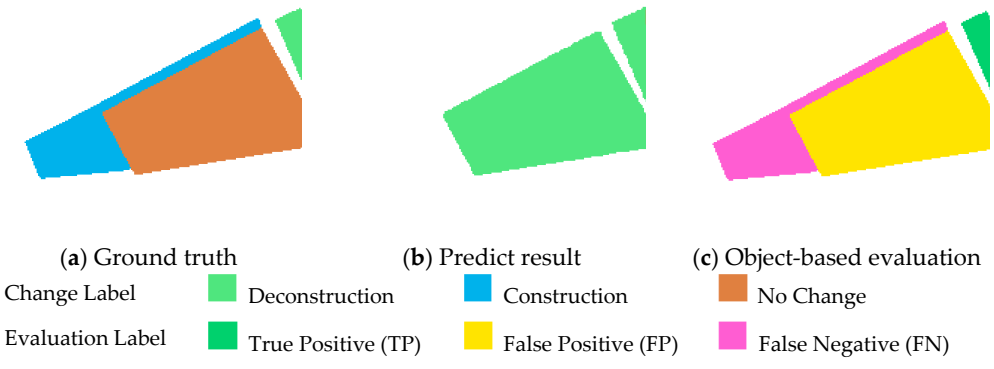

Figure 10. An example of object-based evaluation: (**a**) ground truth; (**b**) results of prediction; and (**c**) a comparison of the ground truth with the predicted results for object-based evaluation.

An intersection over union (IoU) threshold was applied to evaluate building change detection. The IoU measures the ratio of the intersection area to the union area between the predicted and ground truth regions, serving as a metric to determine whether a predicted result is a valid indication of change. In this study, the minimum IoU threshold was set to 0.3. The predicted result with an IoU value equal to or greater than this threshold was treated as the major category for the object. It can be seen from Figure 11 that a deconstruction object was successfully detected where the IoU value was 0.7, demonstrating a significant overlap between the predicted results and ground truth. To emphasize the change areas, we did not plot "no change" to "no change" TP in Figure 11c.

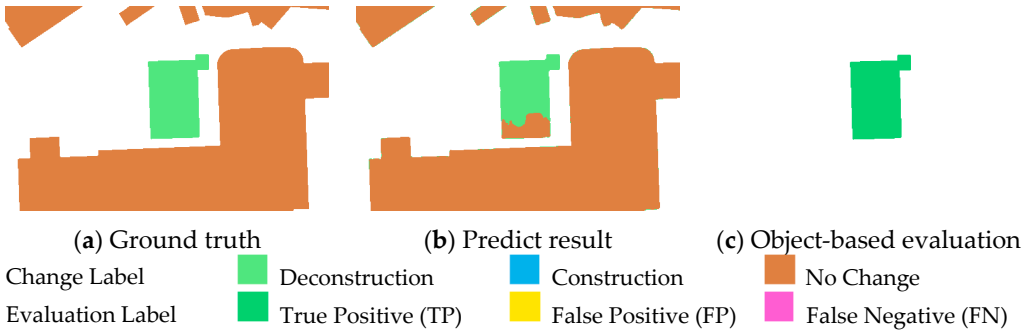

Figure 11. An example of a partially detected case: (**a**) ground truth; (**b**) results of prediction; and (**c**) a comparison of the ground truth with the predicted results for object-based evaluation.

The object-based evaluation results in Table 10 reveal the model performance. As for the no change label, the model performed very well. The RGB_DSM_Map model performed exceptionally well for the no change category, with a precision of 98.61%, a recall of 99.60%, and an F1 score of 99.11%. In contrast, a higher number of FP instances in the deconstruction category, with a lower precision of 62.49% and a substantial number of FNs in the construction category, resulted in a recall of only 43.26%. A comparison of the results of pixel-based (Table 11) and object-based (Table 12) evaluations revealed the accuracy of object-based evaluation as relatively lower. Object-based evaluation treats each changed unit as a single entity, regardless of the number of pixels within the unit. As a result, object-based evaluation better represents the requirements for map updating.

Table 12. Confusion matrix and the evaluation metrics for object-based evaluation.

Object-Based Evaluation	Ground Truth Label		
	Deconstruction	**Construction**	**No Change**
TP_object	1611	1075	9528
FP_object	967	392	134
FN_object	5	1410	38
Precision	62.49%	73.28%	98.61%
Recall	99.69%	43.26%	99.60%
F1 Score	76.82%	54.40%	99.11%

The accuracy of the changed unit is highly related to the size of the changed unit. Figure 12 shows the histogram of the changed unit area for different classes. The Q25, Q50, and Q75 represent the 25^{th} percentile (lower quartile), 50^{th} percentile (median), and 75^{th} percentile (upper quartile), respectively, providing a summary of the data distribution. Table 13 provides a statistical analysis of the median area of each class of the changed units. The deconstruction FPs had a median area of 4.89 m^2, indicating that the model sometimes misclassified small insignificant changes as deconstruction events. On the other hand, the construction FNs had a median area of 7.46 m^2, suggesting that the model struggled

to effectively detect small-scale construction events. As for the no change category, the high precision and recall, coupled with a median area of 195.14 m^2 for TPs, indicate the robustness of the model in identifying stable regions.

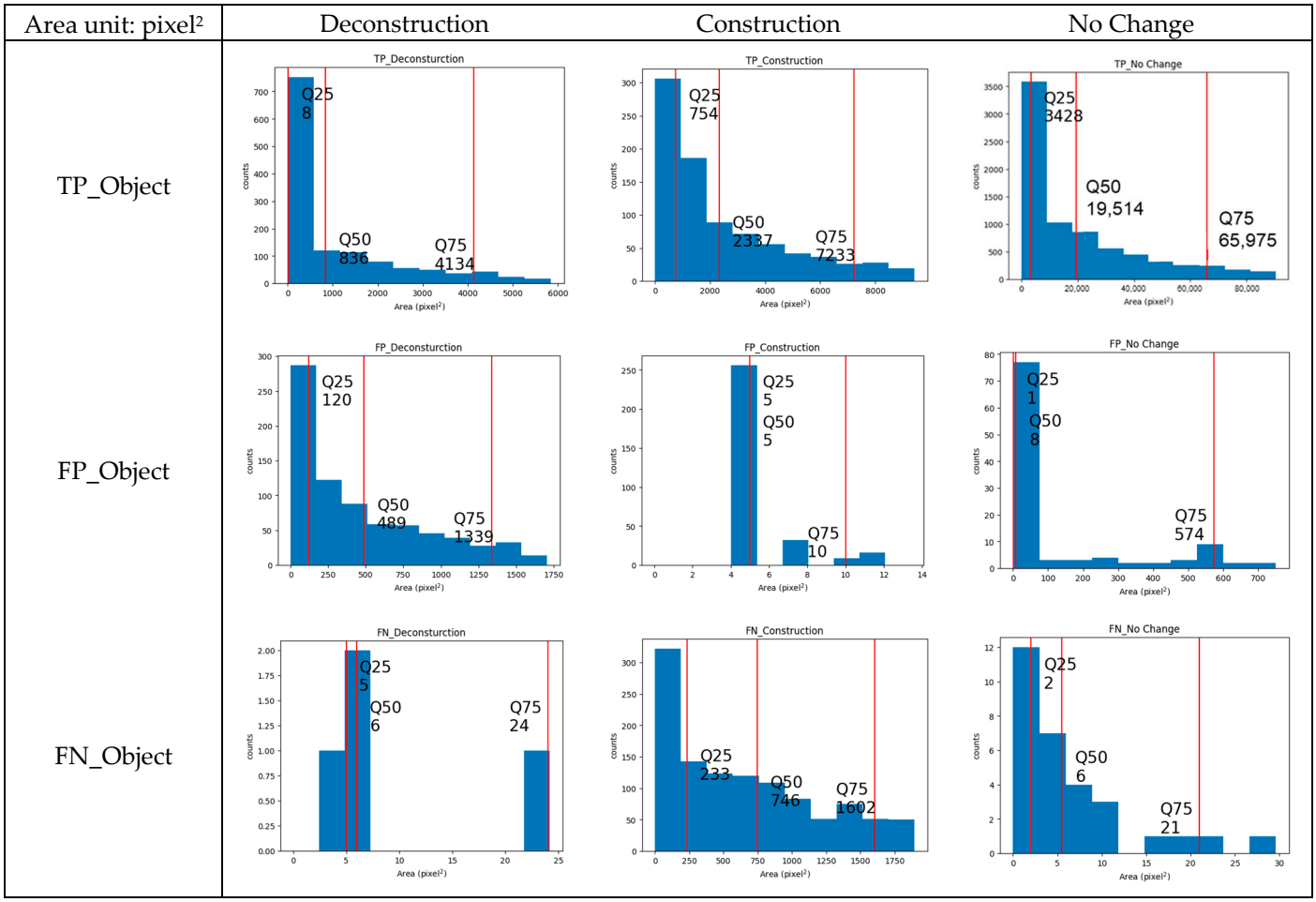

Figure 12. The histogram (area vs. count) for the area of changed units across the three classes. The red lines represent the three-quartile values of the distribution.

Table 13. The median value of the changed units over the evaluation results of the three labels.

Object Area (m^2)	Ground Truth Label		
	Deconstruction	**Construction**	**No Change**
TP_median	8.36	23.37	195.14
FP_median	4.89	0.05	0.08
FN_median	0.06	7.46	0.06

These findings highlight the challenges of detecting small-scale changes, particularly in complex urban environments where subtle variations in features such as elevation and spectral signatures can complicate classification. The excellent performance for the no change label underscores the reliability of the model for stable areas, but the commission errors for deconstruction and omission errors for construction suggest that further refinement is needed to enhance the model's sensitivity to smaller objects.

4. Discussion

In this section, the following three unusual phenomena observed in the change detection results, as shown in Figure 10, will be discussed: the large median area for decon-

struction FPs, the high median area for construction FNs, and the abnormal Q75 value for no change FPs. The specific areas were identified, and their properties were analyzed, as detailed in the following section. These three cases highlight potential factors that may affect the accuracy.

4.1. Deconstruction: False Positives

This dataset exhibited a common issue where the "deconstruction" label was often applied before the work was actually completed. In other words, there was no visible change between the aerial photos captured in 2022 and 2023, yet the ground truth labeled these areas as "changed" prior to actual deconstruction. For instance, Figure 13a shows a residential building that had not yet been demolished but was labeled as "deconstruction". Similarly, Figure 13b illustrates a case where a building area was acquired for road construction, presenting the same labeling discrepancy. We hypothesize that, once a demolition project is submitted to government authorities, the entire area is preemptively labeled as "deconstruction" during mapping, which potentially explains the observed FPs.

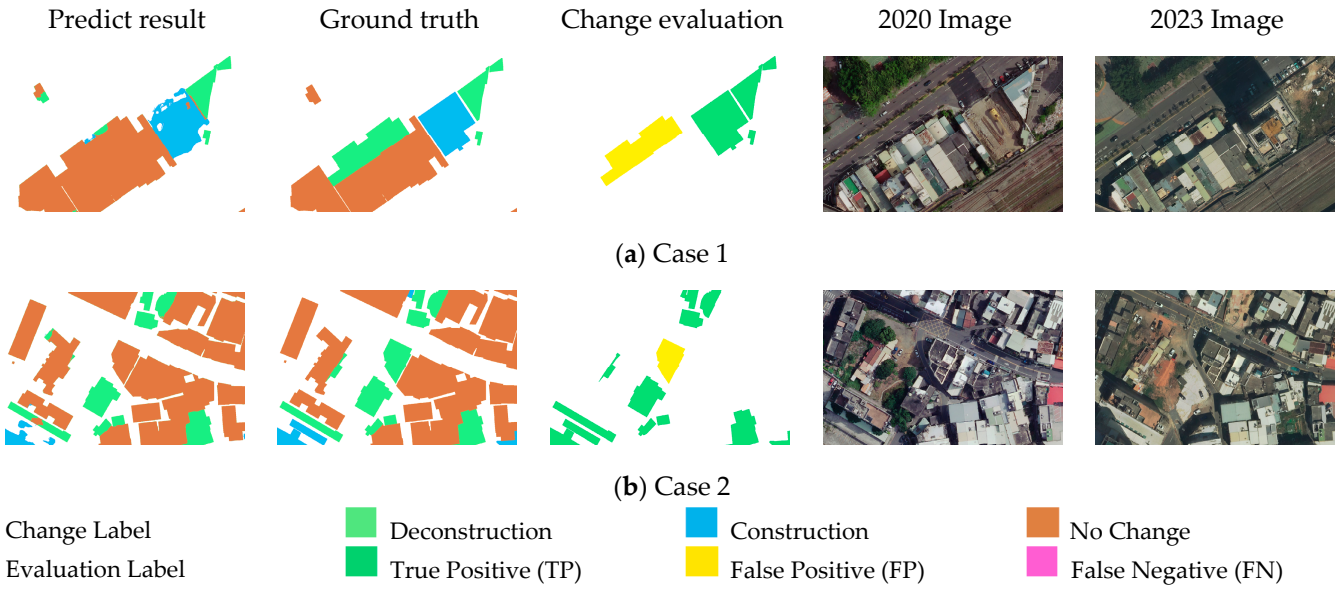

Figure 13. The cases of structures under demolition but labeled as deconstruction: (**a**) Case 1 and (**b**) Case 2.

4.2. Construction: False Negatives

The discrepancy between the building polygons in the map and the aerial true photo significantly affected the change detection accuracy. Figure 14a shows that an area under construction was already digitized as a building, which led the model to falsely classify it as "no change". Conversely, Figure 14b shows a completed construction that was not digitized as a building, resulting in an erroneous classification of "construction". The time discrepancy between the data sources introduced ambiguous or incorrect information that disrupted the model's classification logic.

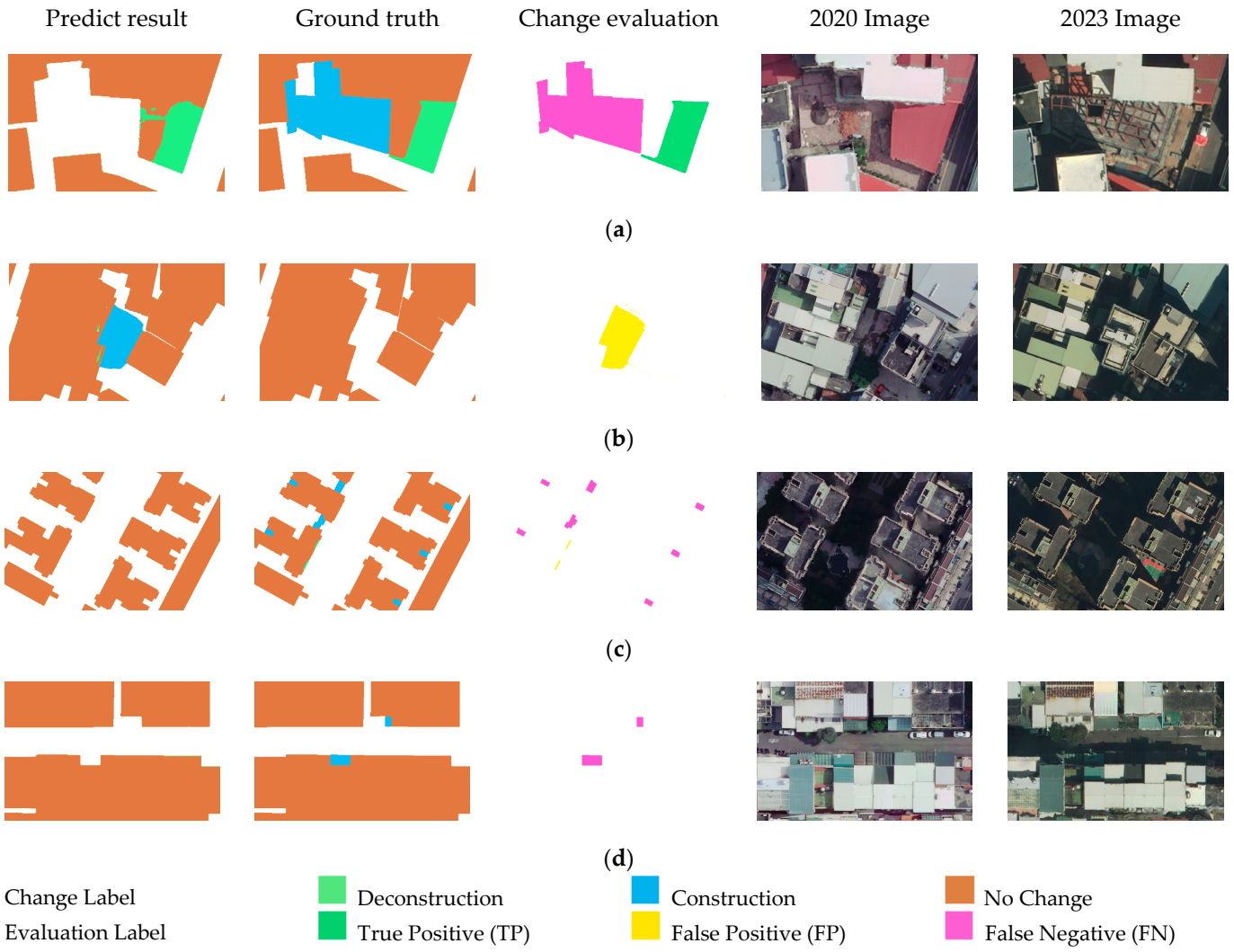

Figure 14. The building map and aerial images were not created at the same time, leading to inaccurate results: (**a**) under construction but labeled as a building; (**b**) already constructed but not labeled as a building; (**c**) building objects that existed but were not digitized at the previous map; and (**d**) rain shelters that existed but were not digitized at the previous map.

Another issue arose from the differing mapping criteria, as illustrated in Figure 14c,d. In these cases, the "construction" label was not associated with an actual change in the building structure but rather with a change in the mapping criteria itself. Specifically, areas that were not previously digitized as buildings were now classified as such, simply due to the change in the criteria used for labeling. This discrepancy in labeling standards led to FNs, where actual building changes were not derived because the mapping process failed to recognize the newly designated areas as part of the construction category.

4.3. No Change: False Positive

In the "no change" FP analysis, a notable group of areas showed distributions around the Q75. Two primary scenarios contributed to this: areas with partially demolished buildings and areas obscured by dense vegetation, such as trees. In the first case, as illustrated in Figure 15a, the model incorrectly classified regions as "no change" because the buildings in these areas were not fully demolished. Small remnants of structures remained, creating a visual ambiguity. While these remnants might not qualify as significant enough for certain categories, their presence confused the model into thinking that no change had occurred. The second scenario, shown in Figure 15b, involved a bus stop that was

almost entirely covered by trees. The dense tree cover obscured the underlying structure, making it difficult for the model to recognize changes accurately. This type of visual obstruction demonstrates the challenges of relying solely on spectral or elevation features when the target objects are hidden from view. Vegetation, in particular, introduces noise into the model's decision-making process, increasing the likelihood of FPs in areas where no meaningful changes have occurred.

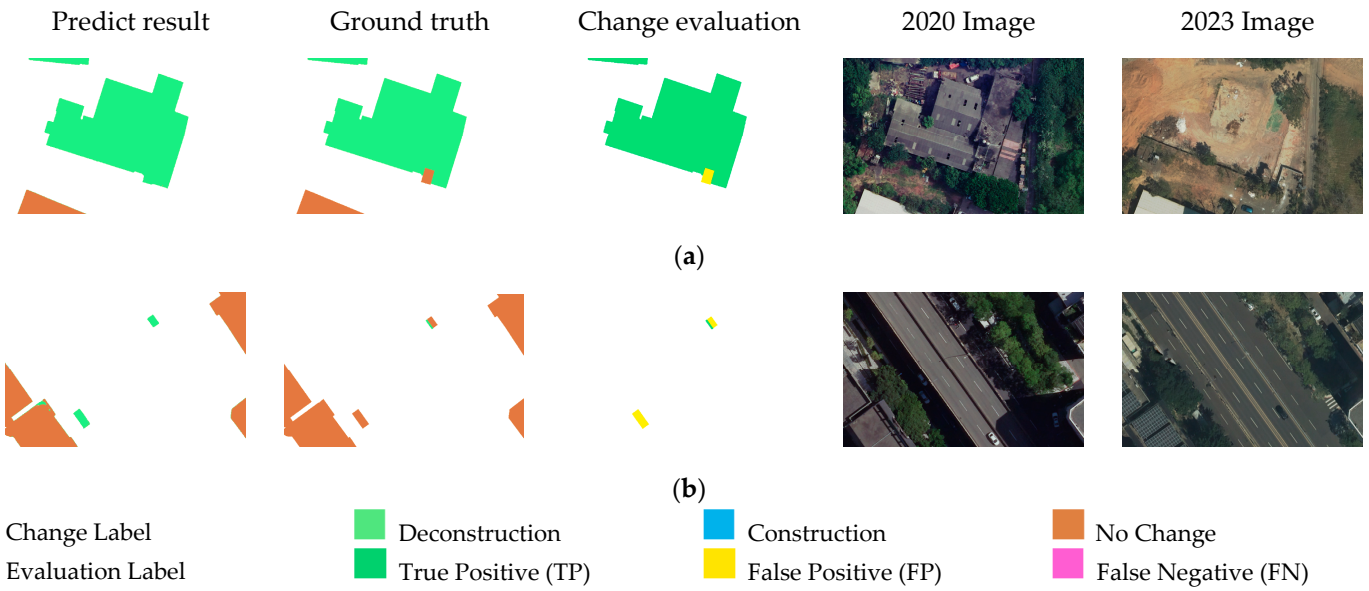

Figure 15. FP caused by a small object: (**a**) a remnant of a building under demolition and (**b**) bus stops covered fully by trees.

4.4. Two-Stage vs. One-Stage Change Detections

This section evaluates the effectiveness of one-stage (i.e., end-to-end) and two-stage (i.e., building detection then building change detection) approaches. The two-stage approach detected the building first, then compared the buildings in two periods to detect the changed areas. The two-stage approach also used the FT-UNetFormer DL model to train on the combined temporal dataset (i.e., 2017 and 2020) as a building detection model. Then, the model was applied to predict building regions in the present dataset (i.e., 2023). These detected building regions were subsequently compared with the DBM (i.e., building polygons) in 2020 to produce change maps. Table 14 summarizes the training/validation datasets and change detection approaches used in the experiments.

Table 14. A summary of the one-stage and two-stage change detection.

	Training/Validation Dataset	**Change Detection Approach**
Two-Stage (RGB_Map)	2017, 2020 (RGB images)	Predict 2023 buildings—2020 DBMs
One-Stage (RGB_Map)	2017, 2020 (RGB images, DBMs)	Predict 2020_2023 building change
Two-Stage (RGB_DSM_Map)	2017, 2020 (RGB images, DSMs)	Predict 2023 buildings—2020 DBMs
One-Stage (RGB_DSM_Map)	2017, 2020 (RGB image, DSMs, DBMs)	Predict 2020_2023 building change

Pixel-based evaluation results in Figure 16 and Table 15 reveal that the one-stage model achieved higher F1 scores (67.42% and 82.94%) compared with the two-stage approach (58.04% and 63.52%). While the overall pixel-based evaluation provided a general comparison, it did not fully capture the performance differences within specific categories.

To make a more detailed analysis, confusion matrices and statistic metrics in Tables 16 and 17 demonstrate that the two-stage approach showed lower accuracy than the one-stage model in the "construction" and "deconstruction" categories. Although the recall scores for the two-stage approach appeared relatively high, they were offset by significantly higher FP pixel counts compared with TP pixels. This imbalance resulted in poor precision scores, negatively affecting the F1 scores for these categories.

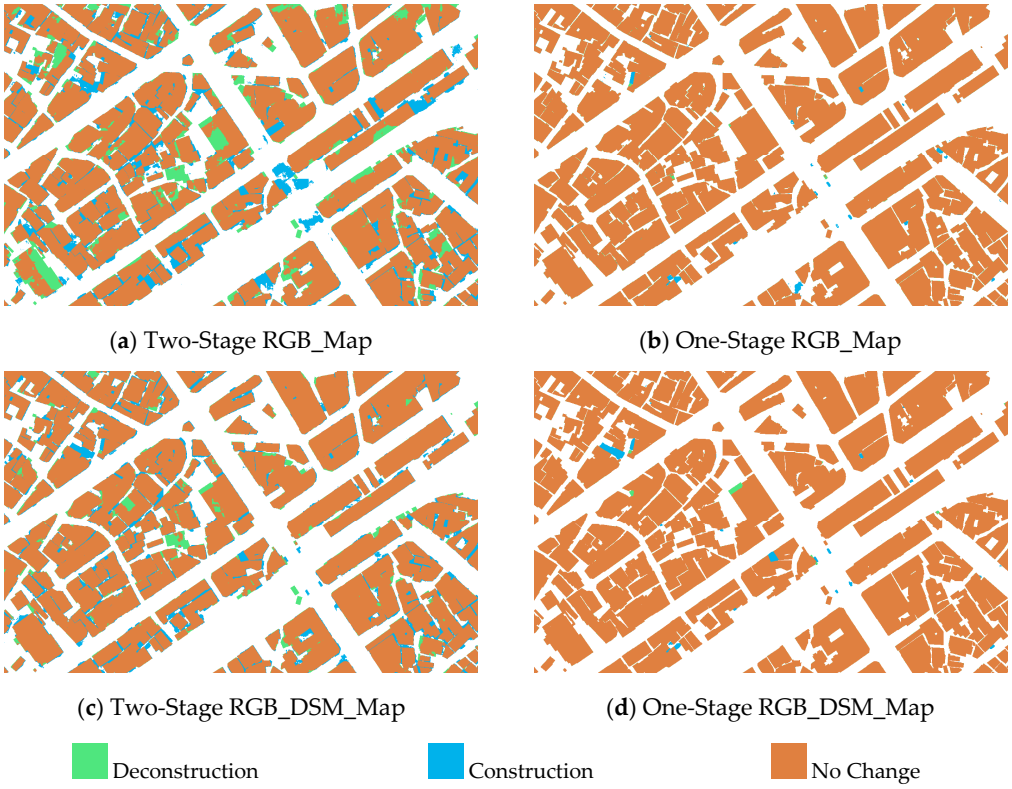

Figure 16. Comparison of change detection results for the different approaches: (**a**) results of two-stage RGB_Map; (**b**) results of one-stage RGB_Map; (**c**) results of two-stage RGB_DSM_Map; and (**d**) results of one-stage RGB_DSM_Map.

Table 15. Overall evaluation of two-stage and one-stage building change detection.

Statistic Metric	Two-Stage (RGB_Map)	One-Stage (RGB_Map)	Two-Stage (RGB_DSM_Map)	One-Stage (RGB_DSM_Map)
Accuracy	96.40%	99.25%	97.63%	99.56%
Precision	56.49%	72.21%	59.72%	84.40%
Recall	81.04%	67.99%	85.96%	82.17%
F1 score	58.04%	67.42%	63.52%	82.94%

In contrast, the proposed one-stage model effectively reduced FPs and achieved a better balance between precision and recall, leading to a superior performance in identifying building changes. These findings highlight the advantages of the one-stage approach in integrating building detection and change detection tasks into a unified workflow, improving both efficiency and quality.

Table 16. Pixel-based evaluation: confusion matrix and metrics for the two-stage RGB_Map model.

Two_Stage (RGB_Map)		Ground Truth Label			
		Non-Building	**Deconstruction**	**Construction**	**No Change**
Prediction	Non-building	1,603,790,373	1,207,153	42,804,843	966,732
	Deconstruction	0	10,476,328	0	2,045,327
	Construction	6,285,321	0	9,769,552	0
	No change	18,698	113,746,732	64,377	531,352,973
Accuracy		97.79%	94.96%	97.88%	94.97%
Precision		99.61%	8.35%	18.56%	99.44%
Recall		97.27%	83.66%	60.85%	82.36%
F1 score		98.43%	15.19%	28.44%	91.09%

Table 17. Pixel-based evaluation: confusion matrix and metrics for the two-stage RGB_DSM_Map model.

Two_Stage (RGB_DSM_Map)		Ground Truth Label			
		Non-Building	**Deconstruction**	**Construction**	**No Change**
Prediction	Non-building	1,615,721,627	1,110,250	30,873,589	1,063,635
	Deconstruction	0	10,987,456	0	1,534,199
	Construction	4,967,949	0	11,086,924	0
	No change	15,685	70,695,306	67,390	574,404,399
Accuracy		98.36%	96.84%	98.45%	96.84%
Precision		99.69%	13.27%	26.38%	99.55%
Recall		98.00%	87.75%	69.06%	89.03%
F1 score		98.84%	23.06%	38.18%	94.00%

4.5. Limitations

Despite the overall effectiveness of the proposed deep learning model, several limitations were observed during the experiment. These challenges primarily stemmed from environmental factors such as shadows, occlusions, and densely packed urban structures, which can lead to misclassification in certain scenarios:

One notable issue was occlusion and shadow. As shown in Figure 17a, a FN detection occurred in an area between buildings due to mixed environmental factors. Shadows cast by adjacent structures and low contrast—caused by insufficient valid image coverage—led to misclassification. These extreme conditions reduced the model's ability to accurately distinguish changes, highlighting the challenges posed by shadow in aerial imagery.

Another key challenge was vegetation occlusion. Figure 17b presents a case where a single-story building under tree cover was partially detected. While the change evaluation identified a true positive (TP) at the lower-left block, the overall detection was significantly impacted by the tree canopy. The building, approximately 3 m in height, was nearly indistinguishable from the pre-existing tree cover, demonstrating the difficulty of detecting structures occluded by vegetation.

Finally, challenges in dense urban areas also affected the model's performance. In highly populated residential zones, buildings were closely spaced, with minimal structural changes over time. However, when such areas were free from the aforementioned challenges (e.g., occlusions and vegetation interference), the model performed significantly better. Figure 17c illustrates an area with fewer obstructions, where the change detection results exhibited high accuracy and reliability, demonstrating the model's potential under optimal conditions.

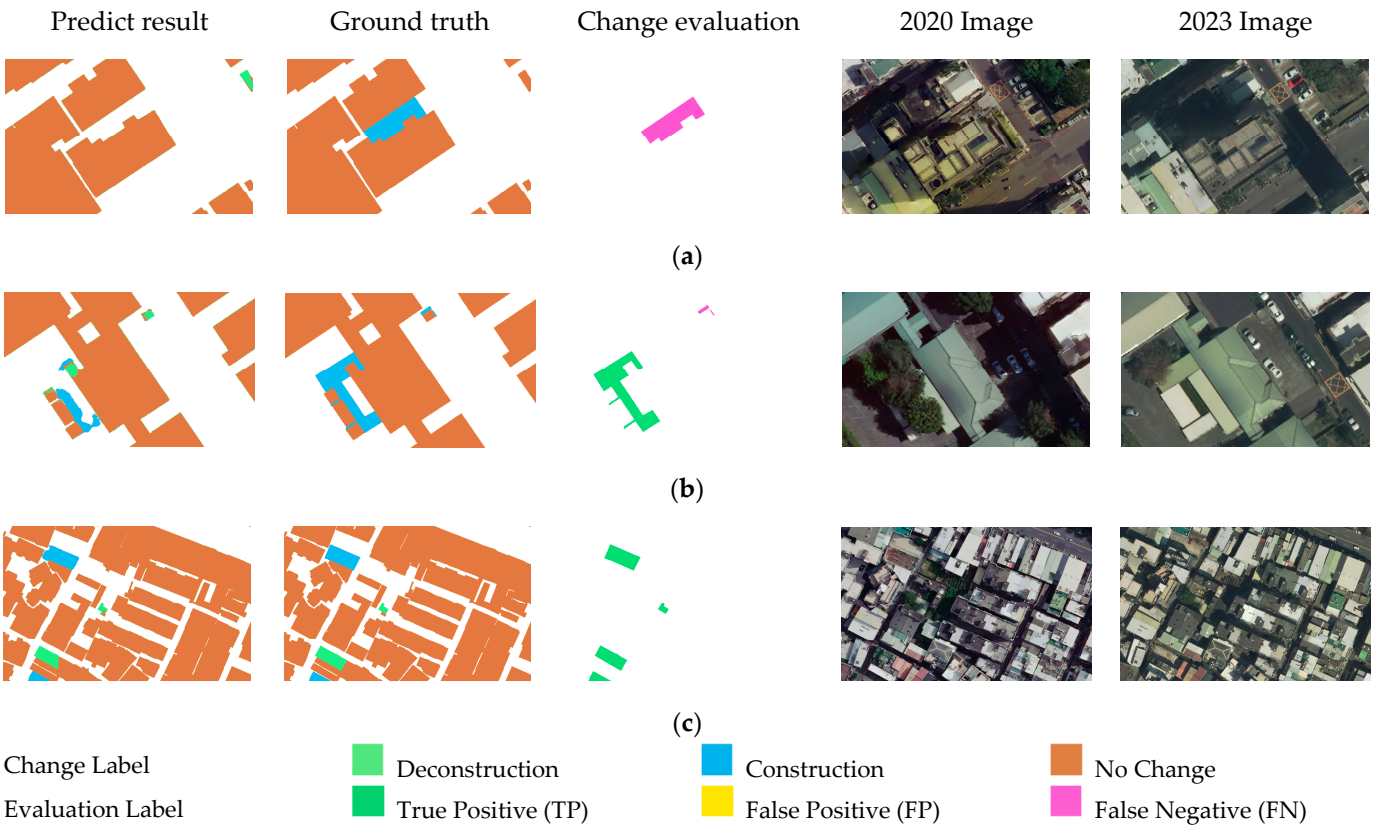

Figure 17. Several case studies about DL model limitations: (**a**) color and shadow; (**b**) low-level building under the same height tree; (**c**) dense building area.

5. Conclusions and Future Works

This study leveraged the advanced FT-UNetFormer architecture and multi-source data fusion and developed and evaluated an end-to-end deep learning-based framework for building change detection. Bi-temporal DSMs play a crucial role in building change detection, as they directly capture elevation differences over time, making them a reliable data source for identifying structural changes [24]. The RGB_DSM model can easily distinguish building change areas due to the data fusion of spectral information and height information.

Meanwhile, the building map provides a historical reference of building distribution, helping to distinguish between newly constructed, demolished, and unchanged structures [25]. The experiment with the RGB_Map model exhibited similar characteristics, where unchanged areas performed significantly better compared with change categories. The F1 score comparison between RGB_DSM and RGB_Map further highlighted this trend, with the non-building category achieving 99.32% in RGB_DSM compared with 95.71% in RGB_Map, and the no change category scoring 98.86% in RGB_DSM versus 89.01% in RGB_Map.

The RGB_DSM_Map model achieved superior performance in detecting building changes across various scenarios by integrating RGB imagery, DSMs, and building maps. The DSM component was particularly valuable in detecting changes that may not be easily distinguishable in spectral imagery alone, such as subtle height variations associated with partial deconstruction or new construction. The combination of RGB imagery for texture and color, DSM for structural elevation, and building maps for historical context enabled a more comprehensive and accurate change detection approach.

The pixel-based evaluation demonstrated the effectiveness of this approach, with the RGB_DSM_Map model consistently outperforming other configurations in terms of

accuracy, precision, recall, and F1 score. The overall pixel-based evaluation achieved an F1 score exceeding 80%, and the two categories, construction and deconstruction, attained F1 scores above 60%. Most missing objects were small ones less than 5 m^2. These results underscore the model's robust performance in detecting building changes.

The object-based evaluations further emphasized the model's robustness in identifying subtle changes. The deconstruction category achieved an exceptionally high recall value (>99%), largely due to the availability of previous building map information, which provided critical context for identifying areas designated for demolition. Meanwhile, the construction category demonstrated precision (>70%), which was attributed to the inclusion of height information from the DSMs, thus effectively highlighting newly constructed areas.

In the comparison of the two-stage and proposed one-stage approaches, the proposed one-stage model was an end-to-end approach that achieved higher F1 scores (67.42% and 82.94%) compared with the two-stage approach (58.04% and 63.52%). Using RGB images and building vector maps, the one-stage method achieved an approximate 10% increase in F1-score compared to the two-stage method. Furthermore, when employing RGB images, DSMs and building vector map, the one-stage approach demonstrated an improvement of about 20% in F1-score relative to the two-stage method. The effectiveness of the one-stage method can be attributed to the integration of multi-source input data, such as RGB imagery, DSMs, or previous DBMs simultaneously, which provided a comprehensive representation of temporal and spatial relationships. This multi-source integration allowed the model to capture subtle changes more accurately while maintaining efficiency in processing large-scale datasets. The proposed one-stage model effectively reduced false positives and achieved a better balance between precision and recall, leading to a superior performance in identifying building changes.

In conclusion, the proposed framework streamlines the process of building change detection, offering a reliable and efficient solution for map updating and urban planning. Future research will focus on three directions: enhancing the model's sensitivity to minor changes, integrating additional data sources, and exploring advanced transformer-based architectures to further improve accuracy and scalability.

In this study, DSM was generated using dense image matching techniques. However, airborne LiDAR-derived DSMs offer a higher level of detail, particularly in capturing building edges, which could improve building segmentation in densely built urban areas. Incorporating LiDAR-based elevation data may enhance the accuracy of building change detection by reducing errors in boundary delineation.

Additionally, this study currently relied solely on RGB imagery. The integration of near-infrared (NIR) data might further refine building classification, as the NIR band is highly effective in distinguishing vegetation from building structures. Exploring multi-spectral data sources could improve model performance, particularly in cases where RGB imagery is insufficient for precise feature differentiation.

Finally, DL model selection plays a crucial role in improving change detection accuracy. ChangeFormer [18] is a transformer-based DL model designed for Siamese change detection. It can be easily modified to accommodate multiple input data sources. Adapting this model to a multi-dataset may provide further improvements in feature extraction and classification, enhancing the overall robustness of the building change detection process.

Author Contributions: Conceptualization, T.-A.T.; methodology, T.-A.T. and P.-C.C.; validation, T.-A.T. and P.-C.C.; formal analysis, P.-C.C.; project supervision, T.-A.T.; writing—original draft preparation, T.-A.T. and P.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the Ministry of Interior (MOI) of Taiwan.

Data Availability Statement: The data presented in this study are available on request from the corresponding author on reasonable request.

Acknowledgments: The authors would like to thank the National Land Surveying and Mapping Center (NLSC), Taiwan and the Hsinchu City Government, for providing the necessary data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bouziani, M.; Goïta, K.; He, D.C. Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 143–153. [CrossRef]
2. Guo, H.; Shi, Q.; Marinoni, A.; Du, B.; Zhang, L. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sens. Environ.* **2021**, *264*, 112589. [CrossRef]
3. Awrangjeb, M. Effective generation and update of a building map database through automatic building change detection from LiDAR point cloud data. *Remote Sens.* **2015**, *7*, 14119–14150. [CrossRef]
4. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
6. Si Salah, H.; Goldin, S.E.; Rezgui, A.; Nour El Islam, B.; Ait-Aoudia, S. What is a remote sensing change detection technique? Towards a conceptual framework. *Int. J. Remote Sens.* **2020**, *41*, 1788–1812. [CrossRef]
7. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *54*, 6232–6251. [CrossRef]
8. Jiang, W.; Sun, Y.; Lei, L.; Kuang, G.; Ji, K. Change detection of multisource remote sensing images: A review. *Int. J. Digit. Earth* **2024**, *17*, 2398051. [CrossRef]
9. Liu, W.; Yang, M.; Xie, M.; Guo, Z.; Li, E.; Zhang, L.; Pei, T.; Wang, D. Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network. *Remote Sens.* **2019**, *11*, 2912. [CrossRef]
10. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [CrossRef]
11. Marsocci, V.; Coletta, V.; Ravanelli, R.; Scardapane, S.; Crespi, M. Inferring 3D change detection from bitemporal optical images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 325–339. [CrossRef]
12. Coletta, V.; Marsocci, V.; Ravanelli, R. 3DCD: A new dataset for 2D and 3D change detection using deep learning techniques. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *43*, 1349–1354. [CrossRef]
13. Qin, R.; Tian, J.; Reinartz, P. 3D change detection–approaches and applications. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 41–56. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 261–271. [CrossRef]
15. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
16. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.S.; Khan, F.S. Transformers in remote sensing: A survey. *Remote Sens.* **2023**, *15*, 1860. [CrossRef]
17. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
18. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210. [CrossRef]
19. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure transformer network for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]
21. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
22. Habib, A.F.; Kim, E.M.; Kim, C.J. New methodologies for true orthophoto generation. *ASPRS Photogramm. Eng. Remote Sens.* **2007**, *73*, 25–36. [CrossRef]

23. Ostrowski, W.; Gulli, V.D.; Bakula, K.; Kurczyński, Z. Quality aspects of true orthophoto in urban areas. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 191–198. [CrossRef]
24. Dos Santos, R.C.; Galo, M.; Carrilho, A.C.; Pessoa, G.G.; De Oliveira, R.A.R. Automatic building change detection using multi-temporal airborne LiDAR data. In Proceedings of the 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), Santiago, Chile, 22–26 March 2020; pp. 54–59. [CrossRef]
25. Liao, C.; Hu, H.; Yuan, X.; Li, H.; Liu, C.; Liu, C.; Fu, G.; Ding, Y.; Zhu, Q. BCE-Net: Reliable building footprints change extraction based on historical map and up-to-date images using contrastive learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *201*, 138–152. [CrossRef]