

Article

A Differential-Based Siamese Network Integrating the CSWin Transformer for Rural Land Cover Semantic Change Detection

Bo Si ^{1,2} , Baiyu Dong ^{1,2} and Ke Wang ^{1,2,*}

¹ College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China; bo_si@zju.edu.cn (B.S.); dongbaiyu@zju.edu.cn (B.D.)

² Zhejiang Key Laboratory of Agricultural Remote Sensing and Information Technology, Zhejiang University, Hangzhou 310058, China

* Correspondence: kwang@zju.edu.cn

Highlights

What are the main findings?

- A Siamese network framework integrating CNNs and Transformers is proposed. Residual learning modules for differential structures and integrating the CSWin Transformer enhance the model's ability to extract local features and global dependencies, respectively.
- We propose a rural land cover semantic change detection dataset comprising 2000 pairs of pixelwise annotated samples, which includes 6 main rural land cover types.

What are the implications of the main findings?

- The proposed method increases the accuracy of rural land cover change detection, which adds an effective monitoring method for achieving accurate surveys of land cover change in rural areas.
- The proposed method can provide more accurate land cover change information in rural areas of developed regions, which is of great significance for rural land use planning and ecological environment monitoring.

Abstract

Deep learning-based methods for land cover semantic change detection utilizing high-resolution, multi-temporal remote sensing imagery have emerged as a research hotspot. However, traditional CNN methods often struggle to preserve long-range spatial context information and face challenges in detecting land cover types with complex semantic change patterns in natural scenes. To address these issues, this study proposes a novel network architecture that integrates a Siamese network with differential structures and a Transformer. First, we introduce residual learning modules to improve the extraction of differential features and strengthen the representation of local features. Second, we integrate the Cross-Shaped Window (CSWin) Transformer into a differential-based Siamese network to enhance global feature extraction. To promote model training and evaluation, we propose a rural land cover change detection dataset—a high-precision dataset comprising 6 main rural land cover types. Ablation and comparative experiments were conducted on the publicly available SECOND datasets and the self-built RLCD dataset. Ablation studies on the RLCD dataset demonstrate that DSTNet achieves significant improvements over the baseline, with increases of 1.77%, 1.95%, 2.57%, and 0.92% in mIoU, Sek, Fscd, and OA. Comparative experiments on the SECOND datasets reveal that the mIoU, Sek, Fsd, and OA scores of DSTNet surpassed the second-best accuracy by 1.04%, 2.15%, 2.28%, and 0.72%.



Academic Editor: Ana Cláudia Teodoro

Received: 6 January 2026

Revised: 31 January 2026

Accepted: 4 February 2026

Published: 10 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: rural land cover; change detection dataset; semantic change detection; Siamese network; Transformer; differential architecture

1. Introduction

Land cover change detection (LCCD) identifies the types and spatial changes in land cover occurring across different time periods [1]. It can provide crucial geospatial data support for decision-making and monitoring in various fields [2], including environmental monitoring [3], urban planning [4], natural damage assessment [5], and precision agriculture [6]. High-resolution, multi-temporal remote sensing images and rapidly advancing data processing methods have created new opportunities for land cover change detection [7]. However, the complex and diverse spatial information and temporal characteristics of land cover across multiple periods in natural scenes make land cover change detection highly challenging [8]. Therefore, achieving land cover change detection based on high-precision remote sensing data and high-performance remote sensing models warrants further in-depth research.

With the rapid rise in deep learning technology, data-driven deep learning models for land cover change detection have gradually become a mainstream research direction [9,10]. Numerous well-annotated semantic change detection datasets have been extensively used for training and validating semantic change detection models [11–13]. However, most datasets focus on urban areas, while datasets for land cover change detection in vast rural areas remain scarce. Unlike urban areas, rural regions possess extensive croplands and forests. Moreover, in developed regions, rural areas are experiencing rapid industrialization and urbanization, leading to the continuous expansion of facility agricultural land and construction. Currently, there is not only a lack of change detection datasets that are comprehensive in terms of rural land cover types, but also a shortage of holistic depictions of unique semantic information changes in rural land cover types. The absence of a detailed and annotated CD dataset focused on rural areas severely limits research on semantic changes in rural land cover. Therefore, constructing a dataset that encompasses multi-category changes in rural areas provides the necessary data support for research on methods for detecting semantic changes in rural land cover.

In recent years, land cover semantic change detection models based on convolutional neural networks (CNNs) have made great progress [14,15], particularly Siamese networks with differential structures [16,17]. For example, the difference enhancement module [18] extracted difference features by subtracting bitemporal encoded features and enhanced them through channel attention. This approach aims to fully leverage dual-temporal images to learn different information instead of complementary information, thereby reducing the impact of irrelevant changes on detection results. Nonetheless, CNNs are constrained by their propensity to lose long-range contextual information, whereas Transformer-based semantic change detection models exhibit stronger capability in modeling global dependency [19]. For instance, SMART [20] developed the multilayer Transformer encoders and decoders to analyze global semantic relationships across different levels, thereby enhancing features. Additionally, many SCD methods integrating Siamese networks with Transformer architectures have been developed [21–23]. For instance, MSTDSNet [24] proposed a wider and deeper layer aggregation (WDLA) to improve the distinguishability of multiscale features. Then, a Multiscale Swin Transformer (MST) is used to make the available spatial information in the refined multiscale features. Although some methods integrating Siamese networks with Transformer architectures are common, the spatial contextual information of different features generated by the Siamese networks remains

underutilized. Simultaneously, the Transformer architecture possesses unique capabilities for capturing global dependencies. Therefore, it is worth exploring how to leverage the Transformer architectures to extract spatial contextual information of differential features. Despite significant advances in existing methods for detecting land cover changes through remote sensing, achieving effective detection of rural land cover changes remains fraught with multiple challenges. Influenced by sensors, vegetation phenology, and seasonal variations, the spatial characteristics of rural land cover across different time phases exhibit complex diversity. Moreover, multi-category semantic information transformation types are difficult to identify without sufficient contextual information. Therefore, developing a semantic change detection model that simultaneously captures local features and spatial contextual features is an essential issue for improving the accuracy of rural land cover change detection.

To address the aforementioned issues, this study proposes a network framework (DSTNet) that combines a Siamese network with differential structures and a Transformer, leveraging the respective feature extraction strengths of CNNs and Transformers. Additionally, this research introduces a rural land cover change detection (RLCD) dataset for studying rural land cover changes and developing semantic change detection models. The main contributions of this study are as follows:

- (1) We propose a CNN–Transformer fusion framework, designed to learn semantic information of change categories. The Siamese network with differential structures to learn local features and the Transformers to capture spatial contextual features.
- (2) We enhance the detail extraction capability by incorporating residual learning modules to augment the bitemporal and differential features. Integrating the CSWin Transformer into the differential Siamese network enhances the model’s ability to capture global dependency.
- (3) This study introduces an RLCD dataset comprising 2000 pairs of pixel-labeled samples. This dataset includes 6 main land cover types with a spatial resolution of 1 m, providing a benchmark dataset for semantic change detection models in rural areas.

The remainder of this paper is organized as follows. Section 2 introduces related work about CNN–Transformer fusion methods and publicly available semantic change detection datasets. Section 3 details data processing, semantic label creation, and an overview of the rural land cover change detection dataset. Section 4 presents the specific network architecture of the proposed method. Section 5 demonstrates ablation and comparative experiments. Section 6 discusses the impact of the proposed model. Section 7 concludes the paper.

2. Related Work

2.1. CNN and Transformer Integrated Methods

The CNN semantic change detection models based on differential structures were first applied to remote sensing image change detection tasks [25]. However, following the innovative application of Vision Transformers to image classification in 2020 [26], the models combining CNNs and Transformers were also rapidly introduced into remote sensing image change detection tasks [27]. At the encoder stage, ICIF-Net [28] extracted local features and global features in parallel using CNN and Transformer architectures. It enabled interactive communication between CNN and Transformer features through a linearized convolutional attention module. MTSCD-Net [29] first extracted multi-scale features using the Siamese semantic-aware encoder based on Swin Transformer, then designed a feature fusion module to combine features. At the decoder stage, the SCAD-Net [30] decoder comprised a Siamese Cross-Attention (SCA) module and a Multi-Scale Feature Fusion (MFF) module. The SCA module extracted unchanged and changed feature information through

a channel Transformer based on a multi-head cross-attention mechanism, while the MFF module integrated the extracted multi-scale feature information. ICT-Net [31] introduced a Transformer decoder based on a novel Cross-Gate Attention (CGA) module to filter key multi-scale discriminative features, thereby enhancing change description performance. MDFENet [32] first inputs multi-scale encoded features into a difference enhancement module to generate refined difference features. It then employed the Transformer decoders to process semantic features at different scales, establishing long-range correlations of pixel semantic changes to reduce “pseudo changes” in the change map. At both the encoder and decoder stages, CTD-Former [33] constructed a CTD Transformer encoder based on multi-head cross-temporal difference (CTD) attention to extract features from changed regions. It further refined the coarse-scale features to fine scales using multihead cross-attention mechanisms. MCTNet [34] employed a Transformer module with an encoder–decoder architecture to capture long-range dependency among these visual tokens. Additionally, the visual general model segment anything model (SAM) based on Transformers has begun to be applied to remote sensing change detection tasks due to its robust visual recognition capabilities [35–37]. For example, VFM-ReSCD [38] fine-tunes the fast segment anything model (FastSAM) network by a Side Adapter (SA). It significantly enhances the ability to extract spatial features from very high resolution (VHR) remote sensing imagery.

Semantic models combining CNNs and Transformers for land cover change detection have become increasingly diverse. However, different CNN and Transformer architectures exhibit distinct capabilities in feature extraction, fusion, and classification. Therefore, further exploration is needed for CNN–Transformer fusion models tackling land cover change detection tasks in complex scenarios.

2.2. Land Cover Semantic Change Detection Datasets

Land cover change detection datasets are crucial for training semantic change detection models that quantitatively analyze surface changes [39]. Currently, land cover change detection datasets have evolved to include diverse types such as optical (RGB), point cloud, multispectral, hyperspectral, SAR data, and multimodal data [40,41]. Table 1 presents several publicly available optical datasets built for remote sensing land cover change detection tasks. Detailed information that contains release year, image resolution, image pairs, image size, and land-cover classes is summarized. HRSCD [11] is one of the earliest published large-scale semantic change detection datasets, designed to train and evaluate supervised deep learning models for semantic change detection tasks. It comprised 291 pairs of $10,000 \times 10,000$ RGB images from urban and countryside areas in France, serving as a classic benchmark dataset. However, the HRSCD lacks a detailed classification for facility agricultural land used for agricultural production. Hi-UCD [42], SECOND [12], CNAM-CD [43], WUSU dataset [44], ChangeNet [45], and FZ-SCD [46] were constructed from high-resolution urban remote sensing images with a resolution of 1 m or less. These datasets are used to detect land cover changes that reflect the spatial evolution characteristics of urban areas, not rural regions. Landsat-SCD [13] was created from 30 m resolution Landsat images that cover the Tumushuke City of Xinjiang from 1990 to 2020 to monitor land cover changes in ecologically fragile regions. Although Landsat-SCD can provide medium-to-high-resolution datasets, it still lacks high-precision remote sensing data. The DynamicEarthNet [47] dataset comprised daily multispectral satellite observations from 75 selected global regions provided by Planet Labs, designed to monitor changes in seven land cover categories across specific areas of interest at the global scale. Nevertheless, the DynamicEarthNet struggles to depict rural land surfaces with sufficient detail at the township scale.

It is evident from the aforementioned datasets that land cover change detection datasets for urban areas are already relatively abundant at the global and regional scales. However, the rural land cover change detection dataset annotated with pixel-wise semantic labels at the township scale remains scarce, and the remote sensing datasets encompassing primary land cover types in rural areas still require further refinement.

Table 1. The publicly available optical land cover semantic change detection datasets.

Dataset	Year	Resolution	Image Pairs	Image Size	Classes
HRSCD	2019	0.5 m	291	10,000 × 10,000	Artificial surfaces, Agricultural areas, Forests, Wetlands, Water
Hi-UCD	2020	0.1 m	1293	1024 × 1024	Water, Grassland, Woodland, Bare land, Building, Greenhouse, Road, Bridge, Others
SECOND	2020	-	4662	512 × 512	Low vegetation, N.v.g surface, Tree, Water, Building, Playground
Landsat-SCD	2022	30 m	8468	416 × 416	Farmland, Desert, Building, Water
DynamicEarthNet	2022	3 m	600	1024 × 1024	Impervious surfaces, Agriculture, Forest and Other vegetation, Wetlands, Soil, Water, Snow and Ice
CNAM-CD	2023	0.5 m	2503	512 × 512	Impervious surfaces, Bare land, Vegetation, Water, Others
WUSU dataset	2023	1 m	3	6358 × 6382/ 7025 × 5500	Road, Low building, High building, Arable land, Woodland, Grassland, River, Lake, Structure, Excavation, Bare surface
ChangeNet	2024	0.3 m	31,000	1900 × 1200	Building, Farmland, Bare land, Water, Road
FZ-SCD	2024	0.8 m	4480	512 × 512	Bare ground, Building, Vegetable, Water, Road

3. Rural Land Cover Semantic Change Detection Dataset

3.1. Research Area and Images

Tongxiang City is situated in the central region of the Hangjiahu Plain in the lower part of the Yangtze River basin. It features a subtropical monsoon climate with flat terrain and an intricate system of rivers, making it a quintessential water town of southern China. As one of Zhejiang Province's economically strong counties, Tongxiang exhibits balanced urban-rural development and has maintained the smallest urban-rural income gap in Jiaxing City for 11 consecutive years. Due to Tongxiang's exceptional agricultural resources and dynamic economic growth, significant changes in land cover types and areas have occurred in recent years. Therefore, we select Tongxiang as the research area for detecting rural land cover changes. Data collection and processing were conducted in the towns governed by Tongxiang: Wuzhen Town, Chongfu Town, Shimen Town, Puyuan Town, Gaoqiao Town, Tudian Town, and Heshan Town, as shown in Figure 1.

High-precision land cover change detection labels constitute the basis for achieving semantic land cover change detection. We utilized high-precision remote sensing images and the land change survey data from Tongxiang City to manually create rural land cover change detection labels. The before-and-after remote sensing images were 1 m resolution GF-2 imagery from 2018 and 0.3 m resolution DOM imagery from 2023, with the DOM resolution resampled to 1 m. Utilizing the land use vector data from the 2018 land change survey as a foundation, we manually annotated land use categories to correspond with the land cover categories in the GF-2 imagery. Similarly, we manually annotated the land cover categories for the 2023 DOM imagery. Detailed descriptions of the land cover semantic categories are provided in Table 2. Subsequently, we generated land cover change detection

labels containing both categorical and spatial change information using relevant methods from the ArcGIS 10.7 software toolbox.

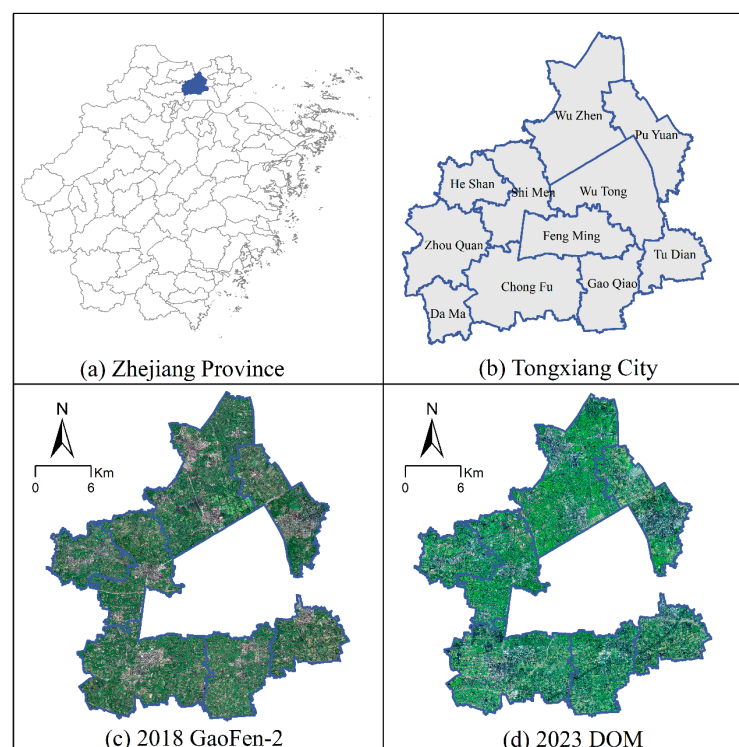


Figure 1. Research area and image data: (a) administrative boundaries of Zhejiang Province; (b) administrative boundaries of Tongxiang City; (c) Tongxiang City 2018 GaoFen-2 (GF-2) satellite imagery; and (d) Tongxiang City 2023 Digital orthophoto map (DOM) imagery.

Table 2. The land cover classes in the CLCD dataset.

Semantic Classes	Description
Cropland	Paddy field and dryland.
Forest	Forest, shrubs, and landscaping seedlings.
Facility Agricultural Land	Greenhouse and livestock, poultry, and aquaculture facilities.
Construction land	Rural dwellings, public buildings, production facilities, and other structures.
Road	Asphalt road and rural road.
Water	River, lake, pond, aquaculture pond, and artificial lake, etc.

3.2. Dataset Description

3.2.1. Overview

With the continuous advancement of rural revitalization policies, China's rural land use has experienced significant transformations. To better promote rural spatial governance and land use management, detecting land cover changes in rural regions has become increasingly important. However, existing datasets for detecting the rural land cover change across all elements at the township level remain scarce. Therefore, we propose a rural land cover change detection dataset using Tongxiang City in Zhejiang Province as the study area, providing a foundational dataset for research on refined rural land cover changes. We employed a sliding window cropping strategy to perform non-overlapping cropping on images and labels, generating the original dataset samples without applying random data augmentation. This dataset comprises 2000 pixel-labeled samples of 512×512 size, with a spatial resolution of 1 m, making it a high-precision dataset for rural land cover

change detection. To precisely represent authentic rural spatial features, we delineated six main semantic categories: water, road, construction land, cropland, forest, and facility agricultural land, as illustrated in Figure 2.

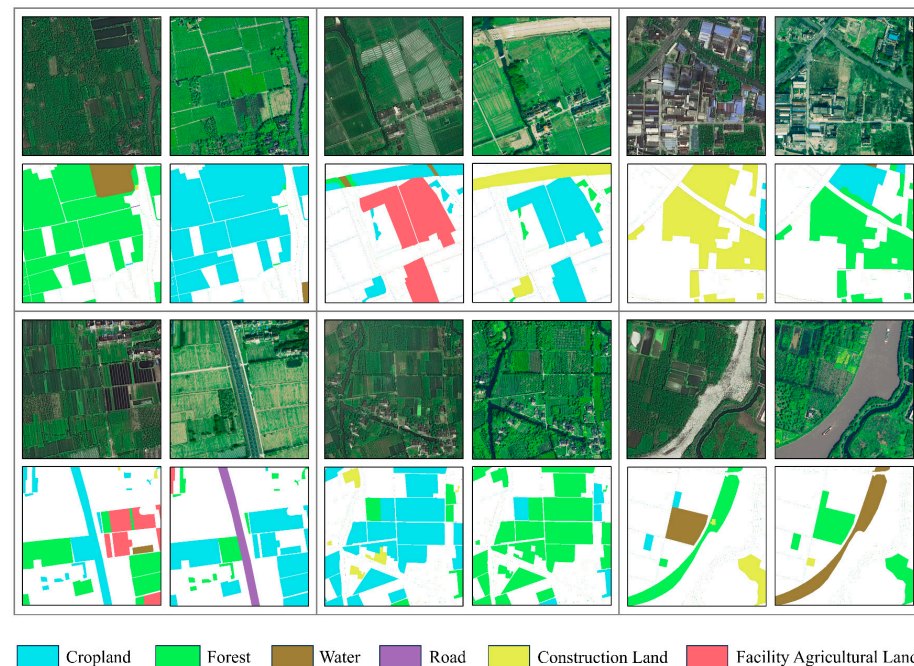


Figure 2. Examples of the rural landcover change detection (RLCD) dataset and legends.

3.2.2. Categories Distribution

To illustrate changes in land cover categories across rural areas, we calculated the proportion of area change for each category in 2018 and 2023 based on six predefined semantic categories, as shown in Figure 3. It was evident that forest, cropland, and construction land represent the primary change types, characterized by a significant reduction in forest and increases in cropland and construction land. Furthermore, water showed a slight decrease trend, while facility agricultural land and road exhibited a modest upward trend. Considering the mutual conversion between land cover categories across the two periods, this dataset encompassed 30 types of land cover change. Among these, the conversion of forest to other land types and the conversion of other land types to cropland and construction land represented the primary change patterns. Roads, due to their minimal change area, exhibited relatively infrequent conversions with other land types.

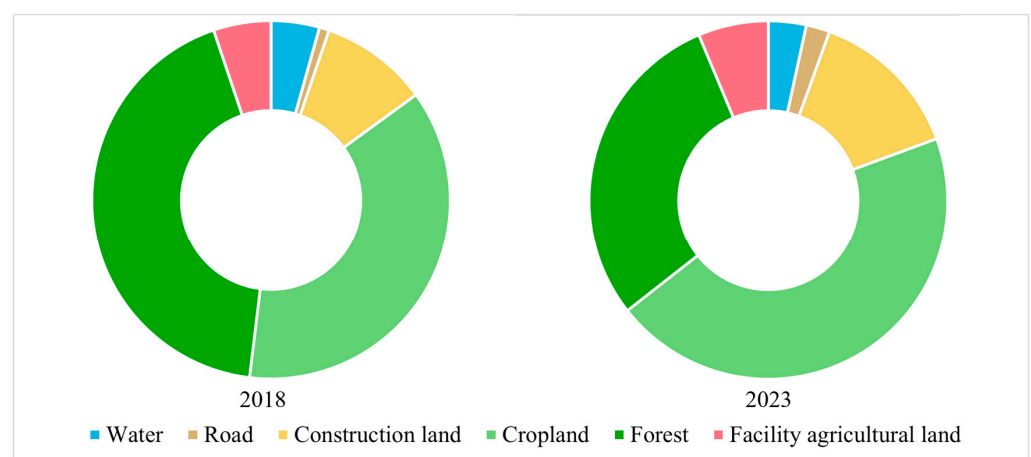


Figure 3. Proportion of area change by land cover categories.

4. Methods

The overall architecture of the proposed model is illustrated in Figure 4. This model primarily consists of three main components: the Siamese network framework with differential structures, the CSWin Transformer embedded within the Siamese network, and residual learning modules. First, we adopt a CNN change detection framework with differential features to obtain multiscale differential features from bitemporal remote sensing images. Then, we introduce the CSWin Transformer to process a sequence of tokens that are derived from multiscale feature transformations. Additionally, we enhance the learning of bitemporal features and differential features by introducing the residual learning module in the encoder, thereby strengthening the differential features representation. Furthermore, we incorporate residual layers in the decoder stage to bolster feature recovery capabilities, making it easier to preserve more feature details. Finally, we employ a classic three-branch change detection classifier, comprising two semantic segmentation (SS) classifiers and one binary change detection (BCD) classifier. The SS classifiers generate land cover maps (LCM), while the BCD classifier produces the binary change map (BCM). BCM and LCMs generate predicted bitemporal semantic change detection maps through dot product operations.

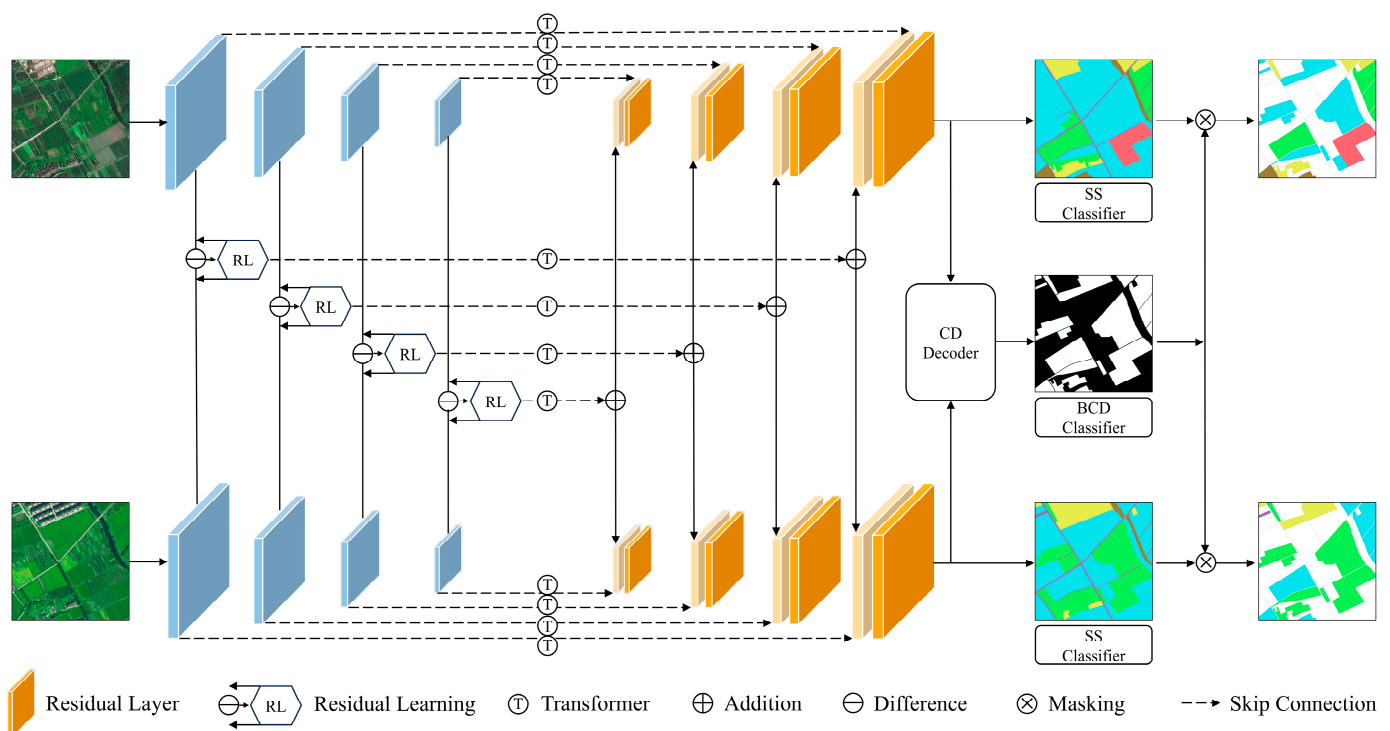


Figure 4. The proposed model architecture.

4.1. The Siamese Network Framework with Differential Structures

CNN frameworks with differential layers can effectively extract change information between bitemporal images in semantic change detection tasks [48]. The difference features generated by Siamese network encoders can highlight the differences in deep features of bitemporal images to improve CD accuracy [49]. Therefore, we construct a Siamese network architecture incorporating different layers as the primary framework for land cover change detection tasks.

The network extracts bitemporal features X_1^i, X_2^i ($i = 1, 2, 3, 4$) at stage i through a series of convolutional and pooling layers within the Siamese network encoders. The difference feature X_d^i is obtained by the subtraction of X_1^i and X_2^i . The formula is as follows:

$$X_d^i = |X_1^i - X_2^i| \quad (i = 1, 2, 3, 4) \quad (1)$$

where $|\cdot|$ denotes the absolute value, and i represents different stages. Subsequently, skip connections are introduced to effectively compensate for the loss of detailed information in the decoder. Furthermore, the difference features are added to the deep features at corresponding stages in both decoders to enhance the diversity of semantic information. The decoders adopt transposed convolutional layers to perform upsampling operations, gradually restoring the resolution of the feature maps.

4.2. The CSWin Transformer Embedded Within the Siamese Network

Recently, hybrid models that combine CNNs and Transformers have demonstrated strong performance on semantic change detection tasks [50]. CNNs may efficiently extract local features through convolution operations but struggle to capture the long-range dependency. However, although the Transformer can capture the global dependency through the self-attention mechanism, it exhibits weaker capability in capturing local features. Therefore, this study introduces the Cross-Shaped Window Transformer [51] with cross-shaped self-attention and locally enhanced position encoding to further improve the Siamese network with differential structures. The horizontal and vertical self-attention mechanisms, combined with locally enhanced positional encoding, expand the model's receptive field and its ability to capture global dependency. The integrated schematic diagram is shown in Figure 5.

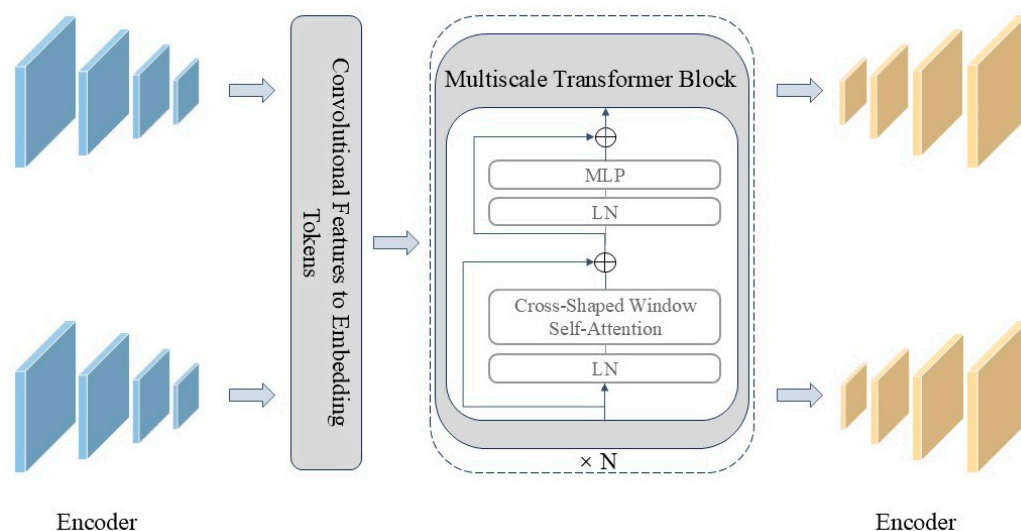


Figure 5. The Integration of CNN and Transformer.

The multi-scale features X_1^i, X_2^i , and X_d^i ($i = 1, 2, 3, 4$) at stage i extracted by the CNN encoders are divided into many subgraphs and mapped to semantic tokens x . The Transformer encoder takes these tokens as input and generates encoded feature maps. The Transformer encoder consists of encoding blocks at different scales, each containing N stacked identical layers. Each layer contains a cross-shaped self-attention mechanism block and a multi-layer perceptron block. Each layer contains a cross-shaped window self-attention mechanism block and a multilayer perceptron (MLP) block. MLP handles the nonlinear transformation and high-dimensional feature combination for each token,

endowing the model with enhanced expressive power and compensating for the linear limitations of the attention layer. The cross-shaped window self-attention mechanism block is used to process feature maps generated by tokens. When feature maps from each layer are input into the self-attention mechanism block, they are divided into multiple feature subspaces by N cross-shaped window self-attention heads. Each self-attention head transforms the input vector into queries (Q), keys (K), and values (V). The formula for calculating attention is as follows:

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

The cross-shaped window self-attention mechanism is implemented by executing self-attention operations in parallel across horizontal and vertical strips that collectively form a cross-shaped window. Its calculation formula is as follows:

$$\begin{aligned} X &= [X^1, X^2, \dots, X^M], \\ Y_k^i &= \text{Attention}(X^i W_k^Q, X^i W_k^K, X^i W_k^V), \\ H - \text{Attention}_k(X) &= [Y_k^1, Y_k^2, \dots, Y_k^M] \end{aligned} \quad (3)$$

$$\begin{aligned} \text{CSWin} - \text{Attention}(X) &= \text{Concat}(\text{head}_1, \dots, \text{head}_K) W^O \\ \text{head}_k &= \begin{cases} H - \text{Attention}_k(X) & k = 1, \dots, K/2 \\ V - \text{Attention}_k(X) & k = K/2 + 1, \dots, K \end{cases} \end{aligned} \quad (4)$$

where M denotes the number of horizontal or vertical stripes, k represents the k th detection head, and W^O denotes the projection matrix of the self-attention results. The projection matrix is used to fuse multiple head results. Due to the permutation invariance of self-attention operations, positional encoding or biased attention weights based on distance must be employed [52]. To efficiently restore positional information lost by the self-attention mechanism, locally enhanced positional encoding (LePE) introduces the per-channel bias, imposing the positional information upon the linearly projected values and integrating positional information within each Transformer module. The output of the cross-shaped window self-attention mechanism is subsequently fed into a multilayer perceptron (MLP) module, which possesses activation functions enabling it to learn nonlinear features.

When the output of the Transformer encoder is fed into the decoder, the Transformer decoder converts attention results into pixel features. The Transformer decoder also consists of N stacked identical layers. Each layer incorporates a masked self-attention mechanism, an encoder–decoder attention mechanism, and a multilayer perceptron block. The semantic information input is mapped to the pixel space through the decoder's attention mechanism to generate pixel features.

4.3. Residual Learning

To further enhance the expression of differential features, this study incorporates a residual learning module into the Siamese network with differential structures to perform enhanced learning on bitemporal and differential features. More details of the residual learning module are shown in Figure 6. The residual learning module first performs residual learning on the bitemporal features extracted by the Siamese encoders, then conducts residual learning again on the difference features obtained by subtracting the bitemporal features. The residual learning layer primarily consists of 3 convolutional layers, 3 batch normalization layers, 2 Relu layers, and 1 skip connection, representing common operations for enhancing feature learning capabilities. Residual learning modules are added at all four scales to extract multiscale features, and then these multiscale features are concatenated

into the decoder features of the corresponding stage through skip connections. Upsampling operations are performed through residual layers to help restore more feature details.

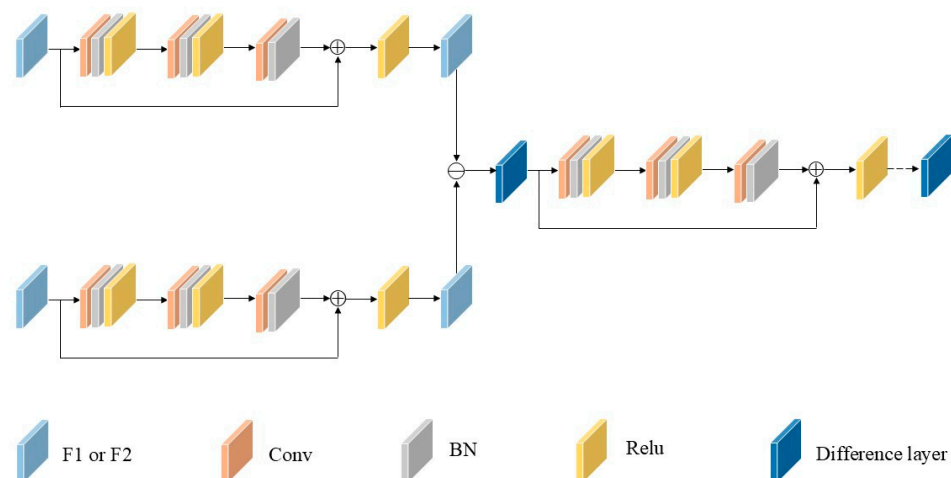


Figure 6. Residual learning modules for differential structures.

4.4. Loss Function

This study employs a composite loss function to train the proposed network. The loss comprises three components: L_{bcd} , L_{ss} , and L_{sc} . L_{bcd} represents the loss function for the binary change detection network, L_{ss} denotes the loss function for the segmentation networks, and L_{sc} refers to the semantic consistency loss function. The formula for calculating L is as follows:

$$L = L_{bcd} + L_{ss} + L_{sc} \quad (5)$$

The binary change map is primarily used to refine semantic change results in semantic change detection tasks. As the loss function for the binary change detection task, the L_{bcd} serves to quantitatively measure the deviation between the true binary change map y and the predicted binary change map $\hat{y}$. The calculation formula for L_{bcd} is as follows:

$$L_{bcd} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (6)$$

To mitigate the impact of sample imbalance, we selected a weighted class balance loss as the loss function L_{ss} for the segmentation network. The calculation formula for L_{ss} is as follows:

$$L_{ss} = -\frac{1}{n} \sum_{i=1}^n w_i y_i \log \hat{y}_i \quad (7)$$

where i represents the category i , n represents the number of categories. y_i and $\hat{y}_i$ represent the ground truth and the predicted value of category i . w_i represents the weight of category i .

Considering the semantic correlation between temporal images X_1 and X_2 , we incorporate a semantic consistency loss function L_{sc} [45]. The calculation formula for L_{sc} is as follows:

$$L_{sc} = \begin{cases} 1 - \cos(X_1, X_2), & y = 0 \\ \cos(X_1, X_2), & y = 1 \end{cases} \quad (8)$$

where y represents the change label (setting the annotation of changed regions to 0 and no-changed regions to 1).

5. Experiments and Results

5.1. Evaluation Metrics and Experimental Settings

To assess the segmentation accuracy of the proposed method, we employed four metrics commonly used in semantic change detection tasks: Overall Accuracy (OA), Mean Intersection over Union (mIoU), Separation Kappa (SeK), and F1 score based on the semantic change detection task (Fscd). Before calculating the evaluation metrics, it is essential to compute the confusion matrix $Q = \{q_{ij}\}$ based on the prediction results and labels, where q_{ij} represents the number of pixels that are classified into class i while their ground truth index is j ($i, j \in \{0, 1, \dots, N\}$). (The unchanged class is set as class 0.) OA denotes the proportion of correctly classified pixels out of the total number of pixels. Its calculation formula is as follows:

$$OA = \sum_{i=0}^N q_{ii} / \sum_{i=0}^N \sum_{j=0}^N q_{ij} \quad (9)$$

OA is a common evaluation metric in remote sensing image classification and change detection tasks [53], measuring a model's overall classification performance. However, semantic change detection tasks contain a large number of pixels belonging to unchanged categories. Therefore, additional evaluation metrics are needed to balance the segmentation performance of unchanged and changed categories. To balance the change detection performance between changed and unchanged categories, this study employs mIoU and SeK to discriminate semantic information across multiple change types. The mIoU calculation formula is as follows:

$$mIoU = \frac{IoU_1 + IoU_2}{2} \quad (10)$$

$$IoU_1 = \frac{q_{00}}{\sum_{i=0}^N q_{i0} + \sum_{j=0}^N q_{0j} - q_{00}} \quad (11)$$

$$IoU_2 = \frac{\sum_{i=1}^N \sum_{j=1}^N q_{ij}}{\sum_{i=0}^N \sum_{j=0}^N q_{ij} - q_{00}} \quad (12)$$

The mIoU is calculated as the average of the Intersection over Union for unchanged categories (IoU1) and the Intersection over Union for changed categories (IoU2). It comprehensively evaluates classification accuracy for both changed and unchanged categories. To further evaluate segmentation performance across multiple change categories, we select SeK to assess the classification of change categories, with its calculation formula as follows:

$$SeK = e^{IoU_2 - 1} * Kappa \quad (13)$$

$$Kappa = \frac{\rho - \eta}{1 - \eta} \quad (14)$$

$$\rho = \frac{\sum_{i=0}^N \hat{q}_{ii}}{\sum_{i=0}^N \sum_{j=0}^N \hat{q}_{ij}} \quad (15)$$

$$\eta = \frac{\sum_{i=0}^N \left(\sum_{i=0}^N \hat{q}_{ij} * \sum_{j=0}^N \hat{q}_{ij} \right)}{\left(\sum_{i=0}^N \sum_{j=0}^N \hat{q}_{ij} \right)^2} \quad (16)$$

Sek is derived from the intersection over union for changed categories (IoU2) and the Kappa coefficient. Before calculating the Kappa coefficient, a confusion matrix $\hat{Q} = \{\hat{q}_{ij}\}$ must be computed, where $\hat{q}_{ij} = q_{ij}$ except that $\hat{q}_{00} = 0$. The unchanged category is excluded from Sek calculations to address class imbalance. To evaluate the accuracy of land cover classification in change areas, we introduce Fscd to measure the semantic information transformation of change categories. More details about Fscd are as follows:

$$F_{\text{scd}} = \frac{2 * P_{\text{scd}} * R_{\text{scd}}}{P_{\text{scd}} + R_{\text{scd}}} \quad (17)$$

$$P_{\text{scd}} = \frac{\sum_{i=1}^N q_{ii}}{\sum_{i=1}^N \sum_{j=0}^N q_{ij}} \quad (18)$$

$$R_{\text{scd}} = \frac{\sum_{i=1}^N q_{ii}}{\sum_{i=0}^N \sum_{j=1}^N q_{ij}} \quad (19)$$

Fscd is derived from the Precision (Pscd) and Recall (Rscd) of the change categories, enabling a more balanced quantification of classification accuracy for the change categories. The four metrics above can provide a comprehensive assessment of the model's performance and the classification accuracy of SCD results.

Based on the aforementioned rural land cover change detection dataset, we randomly split the dataset into a training set and a testing set based on the ratio train: test = 9:1. Subsequently, we performed image normalization and applied random flipping and rotation techniques for image enhancement. For model training settings, we employed consistent experimental parameters, including a batch size of 8, training epochs of 50, and an initial learning rate of 0.1. The optimizer utilized a stochastic gradient descent (SGD) algorithm with a momentum value of 0.9 and weight decay of 0.0005. All experiments were implemented through Python 3.8 and the PyTorch 1.10.0 framework and trained on an NVIDIA RTX A6000 GPU equipped with 48 GB of memory.

We selected the publicly available SECOND dataset and the self-built RLCD dataset for ablation studies and comparative experiments. The SECOND dataset [12] is a high-resolution SCD dataset comprising 4668 pairs of 512×512 images, with 2968 pairs publicly available. These image pairs were collected from multiple Chinese cities, including Hangzhou, Chengdu, and Shanghai, featuring spatial resolutions ranging from 0.5 to 3 m. The SECOND dataset encompasses six land cover categories: water, non-vegetated ground surface, low vegetation, trees, buildings, and playgrounds. During training, the SECOND and RLCD datasets were randomly split into training, validation, and test sets based on the ratio train:validation:test = 8:1:1.

5.2. Comparison Methods

We selected nine different networks based on CNN or Transformer for comparison with the proposed method, comprehensively evaluating its performance on the rural land cover change detection dataset. The compared methods are as follows:

- (1) FC-Siam-diff [25]: This network initially derives difference features by subtracting the encoder features of the bitemporal images, then concatenates difference features to the Siamese decoders by skip connections.
- (2) FC-Siam-conc [25]: This network concatenates the encoder features of the bitemporal images to the Siamese decoders by skip connections.
- (3) SCDNet [48]: This method utilizes a Siamese network with differential features. It incorporates encoder features and differential features into two Siamese decoders by skip connections, achieving improvements through attention mechanisms and a deep supervision strategy.
- (4) SSD-I [54]: This method first extracts bitemporal features based on the Siamese encoders, then feeds them into three decoder branches that generate a single change map and two temporal semantic change maps.
- (5) Bi-SRNet [54]: This method enhances semantic information extraction by incorporating two Siamese semantic reasoning (Siam-SR) blocks and a cross-temporal semantic reasoning (Cot-SR) block on top of the SSD-I. It coordinates semantic representations with change representations through a semantic consistency loss (SCLoss).
- (6) MTSCD [29]: This method first extracts multi-scale features based on the Siamese semantic-aware encoder originating from Swin Transformer. Then deeply fuses the two-level differential features through an information exchange module. Finally, it fully leverages the correlation between the two subtasks with a spatial feature enhancement module.
- (7) SCanNet [55]: This method develops a semantic change Transformer (SCanFormer) to explicitly model the spatial dependency in semantic transitions between bitemporal images. It then leverages temporal consistency as a prior constraint to extract semantic information from bitemporal images.
- (8) SSCLNet [56]: This method extracts contextual information using HRNet and change information through an absolute difference to form the baseline model. It then incorporates a semi-supervised contrastive learning module for semantic segmentation to enhance class discriminability, employing a self-training (ST) method to achieve semi-supervised semantic segmentation.
- (9) STS-FINet [57]: This method extracts multilevel features through a Multi-Scale Feature Extraction Encoder (MS-FEE) equipped with Mixed Spatial Reasoning Convolution blocks (MixSrc). It then leverages a Transformer-based Multilevel Feature Interaction module (TML-FI) to capture long-range dependencies and spatial information within multi-level features. Finally, a Multilevel Feature Fusion Decoder (MLFFD) integrates multilevel features to generate semantic change maps.

5.3. Ablation Experiment

To evaluate the effectiveness of the proposed method, we conducted ablation experiments on its three core components: Residual Learning (RL), CSWin Transformer (hereafter referred to as Transformer), and the RL–Transformer Fusion Framework. This resulted in four variables for the ablation experiment: the Baseline model, Baseline–RL, Baseline–Transformer, and Baseline–RL–Transformer.

Table 3 presents the evaluation results of the ablation experiments on the SECOND dataset. The Baseline–RL–Transformer achieved the greatest improvement over the baseline, followed by the Baseline–Transformer, while the Baseline–RL model showed the smallest gain over the Baseline. The Baseline–RL–Transformer improved mIoU, Sek, Fsd, and OA by 3.91%, 8.21%, 7.37%, and 3.60%, respectively. This demonstrates that the fusion of the Siamese network based on residual learning modules and the Transformer effectively enhances network performance. The Baseline–Transformer achieves great improvements

over the baseline model, with increases of 2.62%, 5.79%, 4.58%, and 2.79% in mIoU, Sek, Fsd, and OA, respectively. The Baseline-RL achieved improvements of 1.04%, 1.52%, 1.36%, and 1.56% over Baseline in mIoU, Sek, Fsd, and OA, respectively.

To visually demonstrate the performance of the aforementioned network on the SECOND dataset, we present visualizations of several sample pairs, as shown in Figure 7. Experiments demonstrate that the change detection results of the Baseline-RL-Transformer are significantly superior to those of the Baseline-Transformer and Baseline-RL, while the Baseline-Transformer outperforms Baseline-RL.

Table 3. Evaluation results for ablation experiments on the SECOND dataset.

Models	mIoU	Sek	Fscd	OA
Baseline	71.19	20.51	60.50	85.11
Baseline-RL	72.23	22.03	61.86	86.67
Baseline-Transformer	73.81	26.30	65.08	87.90
Baseline-RL-Transformer	75.10	28.72	67.87	88.71

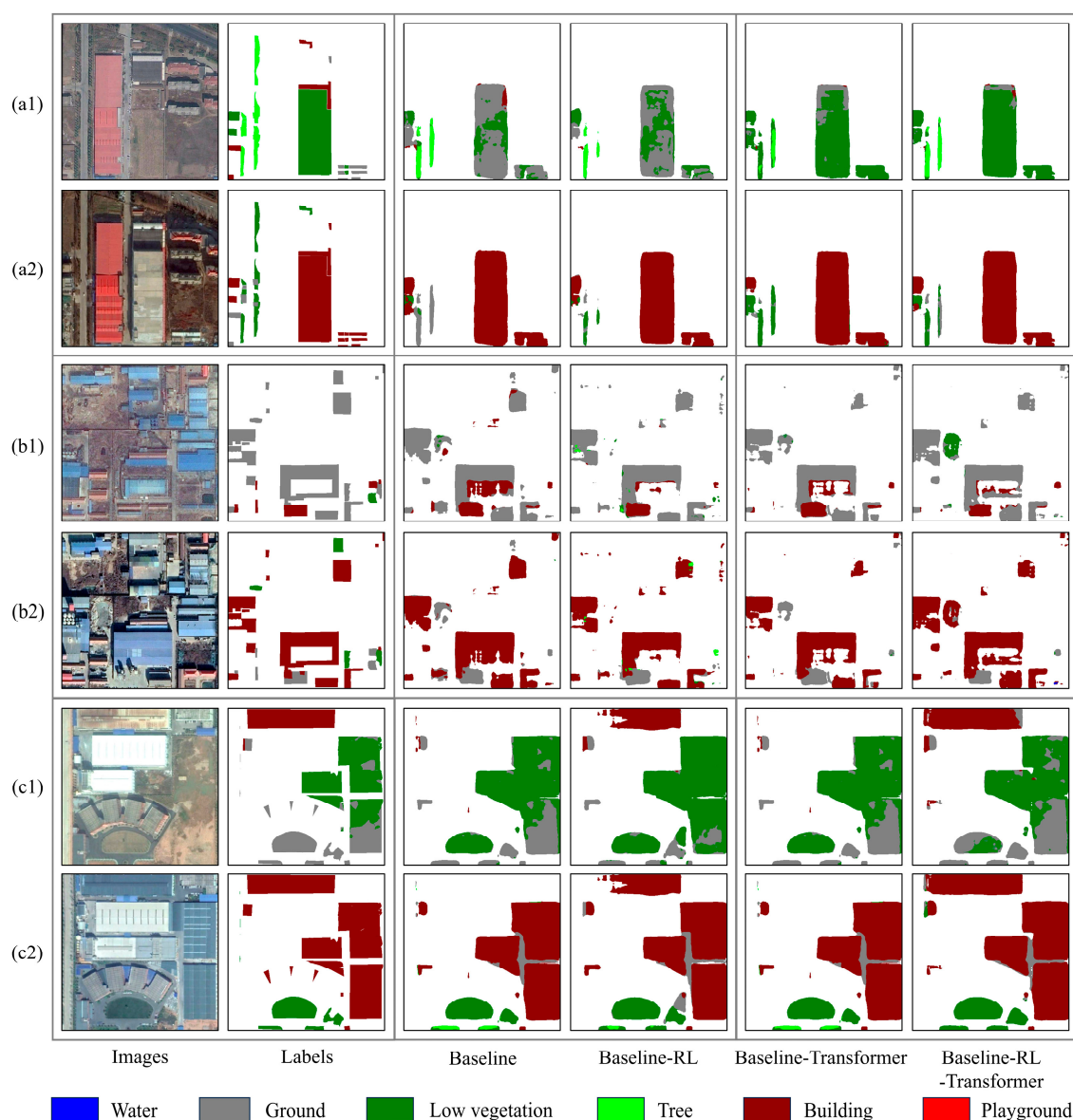


Figure 7. Ablation experiment results on the SECOND dataset. (a1–c2) represent three different pairs of images.

The Baseline–RL–Transformer enhances the detection capability of semantic change information. For instance, it can accurately identify low vegetation in Figure 7(a1,a2). The Baseline–Transformer captures contextual information more effectively than the Baseline model. It more readily identifies the contours of the ground in Figure 7(b1,b2). The Baseline–RL captures more local features compared to Baseline. It can identify more buildings in Figure 7(c1,c2).

Table 4 presents the quantitative outcomes of the ablation experiments on the RLCD dataset. The experiments demonstrate that the RL module, the embedded Transformer, and the CNN–Transformer fusion framework all exhibit differing levels of performance enhancement compared to the baseline model. Among them, Baseline–RL–Transformer achieves the highest accuracy, followed by Baseline–Transformer, while Baseline–RL only slightly outperforms the baseline model. Specifically, compared to the Baseline–RL model, Baseline–Transformer achieves improvements of 0.72%, 0.87%, 1.11%, and 0.23% in mIoU, Sek, Fscd, and OA, respectively. This indicates that Baseline–Transformer demonstrates significantly enhanced segmentation performance in semantic change detection compared to Baseline–RL. Meanwhile, Baseline–RL achieved improvements of 0.21%, 0.41%, 0.58%, and 0.15% over the Baseline model in mIoU, Sek, Fscd, and OA, respectively, demonstrating that the RL module can be helpful in detecting semantic change information.

Table 4. Evaluation results for ablation experiments on the RLCD dataset.

Models	mIoU	Sek	Fscd	OA
Baseline	67.02	15.77	58.78	85.67
Baseline-RL	67.23	16.18	59.36	85.82
Baseline-Transformer	67.95	17.05	60.47	86.05
Baseline-RL-Transformer	68.79	17.72	61.35	86.59

Baseline–RL–Transformer achieves improvements of 1.77%, 1.95%, 2.57%, and 0.92% over Baseline in mIoU, Sek, Fscd, and OA, respectively, demonstrating that Baseline–RL–Transformer significantly enhances change detection precision. However, Baseline–Transformer improved mIoU, Sek, Fscd, and OA by 0.93%, 1.28%, 1.69%, and 0.38%, respectively, compared to Baseline. Therefore, this reveals that the accuracy improvement of the Baseline–RL–Transformer model compared to the Baseline is significantly greater than that of the Baseline–Transformer model compared to the Baseline, exceeding the accuracy improvement of the Baseline–RL model compared to the Baseline. This situation indicates that the fusion model of CNN and Transformer showcases the effectiveness of the complementary advantages of CNN and Transformer, producing a synergistic effect where the whole exceeds the sum of its components.

To compare the segmentation performance of the aforementioned models on the RLCD dataset, we visualized several image pairs, as shown in Figure 8. It is evident that the Baseline model exhibits the least satisfactory segmentation results. Both Baseline–Transformer and Baseline–RL significantly reduce misclassifications compared to the Baseline model. Furthermore, the segmentation quality of Baseline–RL–Transformer is markedly superior to that of both Baseline–Transformer and Baseline–RL.

Baseline–RL–Transformer first significantly reduces misclassified areas, such as water, cropland, and forest misclassified by Baseline–RL in Figure 8(b1,b2), and forest and construction land misclassified by Baseline–Transformer in Figure 8(c1,c2). Moreover, Baseline–RL–Transformer enhances segmentation accuracy, enabling more precise identification of change areas. For example, it accurately identifies facility agricultural land in Figure 8(a1,a2) and the detected roads in Figure 8(c1,c2). Baseline–RL outperforms Baseline–Transformer in identifying unchanged areas, such as cropland and forest in

Figure 8(b1,b2) and forest and construction land in Figure 8(c1,c2). Baseline–Transformer outperforms Baseline–RL in capturing details and contours, such as well-identified crop-land, forest, and facility agricultural land in Figure 8(a1,a2).

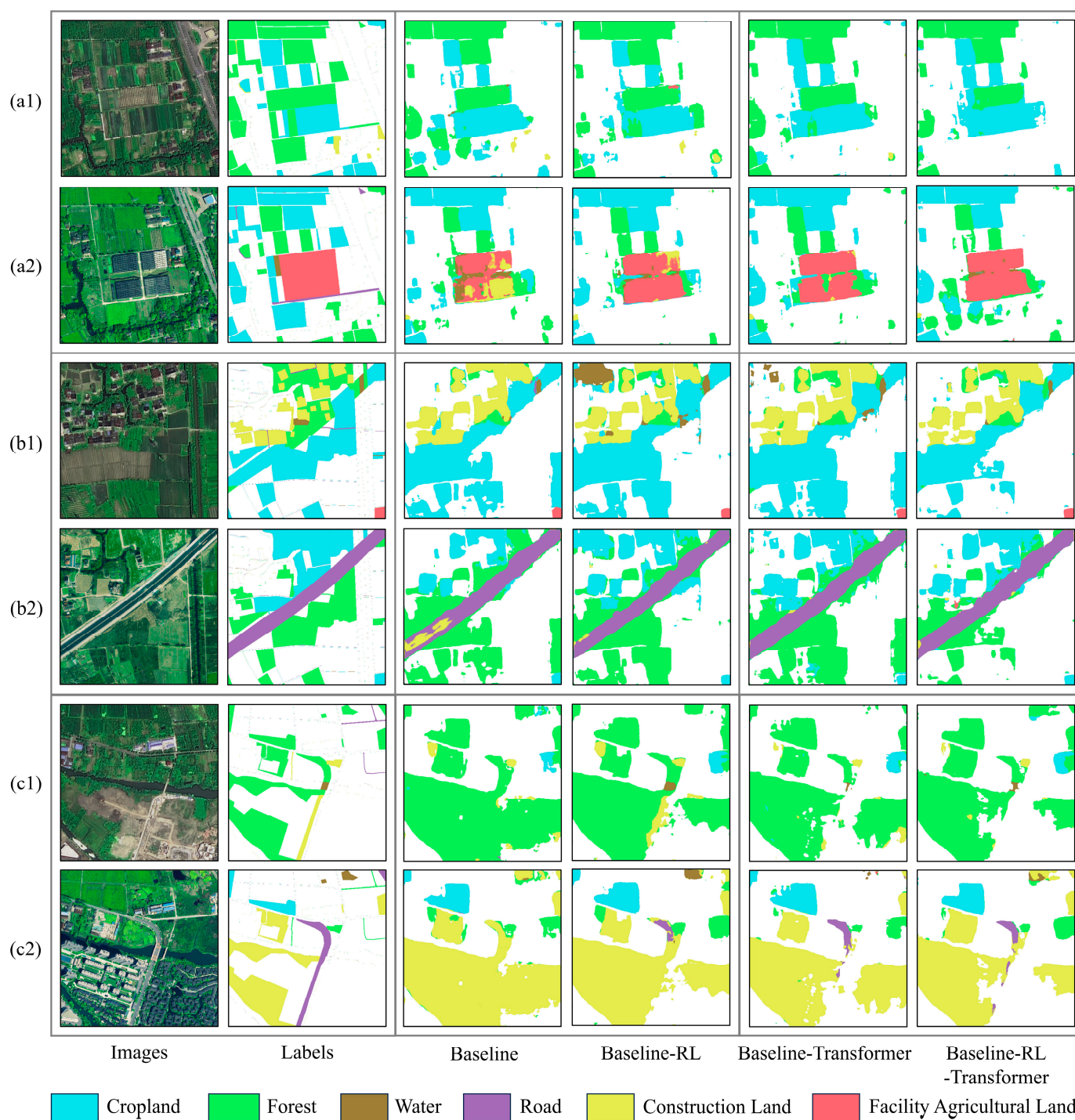


Figure 8. Ablation experiment results on the RLCD dataset. (a1–c2) represent three different pairs of images.

5.4. Comparison Results with Different Methods

The evaluation results compared with advanced methods on the SECOND dataset are shown in Table 5. Compared to other methods, the model structures of FC-Siam-diff and FC-Siam-conc are relatively simple; therefore, they failed to outperform other approaches. SCDNet, SSD-I, and Bi-SRNet are all improved models of the Siamese network

architecture, achieving excellent change detection accuracy. MTSCD, SCanNet, and STS-FINet are all models that integrate Siamese networks with Transformer architectures, achieving higher accuracy. DSTNET achieved the highest accuracy results, outperforming the second-best SSCLNET by 1.04%, 2.15%, 2.28%, and 0.72% in terms of Miou, Sek, Fscd, and OA, respectively.

Table 5. Evaluation metrics compared to different methods on the SECOND dataset.

Methods	mIoU	Sek	Fscd	OA
FC-Siam-diff	67.78	15.05	56.83	84.21
FC-Siam-conc	68.06	15.19	56.89	84.44
SCDNet	71.29	20.70	60.71	85.69
SSD-I	72.41	22.47	61.93	86.87
Bi-SRNet	72.50	22.63	62.21	87.09
MTSCD	72.61	22.72	62.51	87.14
SCanNet	73.35	23.67	63.53	87.80
SSCLNet	74.06	26.57	65.59	87.99
STS-FINet	73.02	22.79	63.08	87.26
DSTNet	75.10	28.72	67.87	88.71

To intuitively demonstrate the change detection performance of the compared methods on the SECOND dataset, we present visualization results for several sample pairs, as depicted in Figure 9. DSTNET demonstrated superior performance compared to other methods. DSTNET effectively extracts change information of different categories. For example, it more accurately classifies ground and trees in Figure 9(a1,a2), as well as ground and low vegetation in Figure 9(c1,c2). Additionally, DSTNet possesses a stronger ability to extract details. It can distinguish the complex contours of ground, trees, and buildings in Figure 9(b1,b2). SSCLNet also demonstrates excellent model performance, but exhibits some misclassifications in unchanged regions. STS-FINet, SCanNet, MTSCD, Bi-SRNet, SSD-I, and SCDNet demonstrate strong classification performance but exhibit more misclassifications in changed categories.

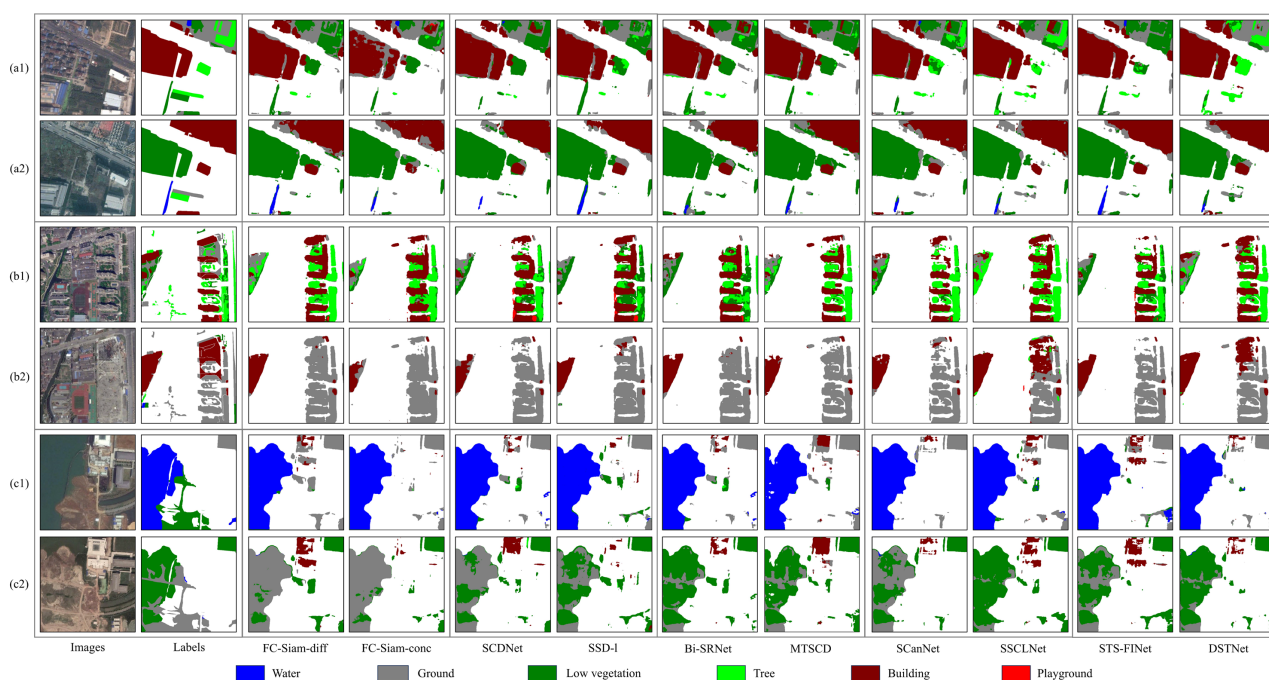


Figure 9. Comparison results of different methods on the SECOND dataset. (a1–c2) represent three different pairs of images.

The evaluation results compared with advanced methods on the RLCD dataset are shown in Table 6. FC-Siam-diff and FC-Siam-conc are traditional Siamese network architectures applied to land cover change detection, achieving impressive accuracy metrics. SCDNet and SSD-I demonstrated comparable accuracy, while Bi-SRNet achieved higher accuracy than SSD-I by incorporating two semantic reasoning modules and semantic consistency loss. MTSCD and STS-FINet are both semantic change detection models based on Transformer–CNN fusion, achieving commendable accuracy. SCanNet attained the third-best accuracy by incorporating SCanFormer and a semantic learning scheme. The proposed method achieved the highest accuracy, with improvements of 0.47%, 0.15%, 0.44%, and 0.39% over SSCLNet in mIoU, Sek, Fscd, and OA, respectively.

Table 6. Evaluation metrics compared to different methods on the RLCD dataset.

Methods	mIoU	Sek	Fscd	OA
FC-Siam-diff	63.82	12.29	53.52	84.17
FC-Siam-conc	63.90	12.45	53.69	84.21
SCDNet	66.87	15.57	58.51	85.60
SSD-I	66.95	15.68	58.63	85.65
Bi-SRNet	67.57	16.65	59.85	85.92
MTSCD	67.99	17.20	60.53	86.08
SCanNet	68.22	17.51	60.88	86.15
SSCLNet	68.32	17.57	60.91	86.20
STS-FINet	68.08	17.39	60.70	86.12
DSTNet	68.79	17.72	61.35	86.59

To intuitively demonstrate segmentation performance, we present visualization results for several sample pairs, as depicted in Figure 10. The proposed method outperforms other approaches in both segmentation accuracy and change detection capability. SCanNet and MTSCD also achieve satisfactory segmentation performance, while Bi-SRNet exhibits strong change detection ability, showing significant improvement over SSD-I and SCDNet. FC-Siam-diff and FC-Siam-conc yield less satisfactory segmentation results.



Figure 10. Comparison results of different methods on the RLCD dataset. (a1–c2) represent three different pairs of images.

The proposed method can identify changes in smaller features, such as cropland and forest in Figure 10(a1,a2), and construction land and cropland in Figure 10(b1,b2). Simultaneously, the proposed method leverages spatial contextual information to more accurately distinguish between changed and unchanged areas, such as cropland and forest in Figure 10(c1,c2) and construction land and cropland in Figure 10(b1,b2). STS-FINet, SCanNet, and MTSCD demonstrate excellent change detection capabilities, but they show a slight disadvantage in extracting complex category change information. For instance, it failed to accurately identify changes in forest and cropland in Figure 10(a1,a2). SSCLNet achieves segmentation performance nearly equivalent to the proposed method but exhibits slight deficiencies in detecting fine contours and small-object changes, such as the undetected cropland and forest in Figure 10(a1,a2). Bi-SRNet shows significant segmentation improvements over SSD-I, SCDNet, FC-Siam-diff, and FC-Siam-conc. SSD-I and SCDNet outperform only FC-Siam-diff and FC-Siam-conc, exhibiting certain misclassifications. FC-Siam-diff and FC-Siam-conc demonstrate relatively basic segmentation performance.

6. Discussion

6.1. Advantages of Our Proposed Model

The proposed framework that integrates a Siamese network with differential structures and the CSWin Transformer achieved superior segmentation performance on the rural land cover change detection dataset. To illustrate the positive impact of this framework, we visualized several representative sample pairs, as shown in Figure 11. The proposed method effectively captures local features and their spatial contextual information, reducing misclassifications to some extent. For instance, in Figure 11(a1,a2), it produces fewer misclassifications of cropland and forest, whereas other methods exhibit more misclassifications in these areas. Moreover, the proposed method excels at learning semantic transformation information in changing regions. For instance, it achieves more accurate segmentation of cropland and forest in Figure 11(b1,b2), whereas other methods fail to distinguish between cropland and forest effectively. These positive effects are due to the differential features extracted from the differential structure enhanced by the residual module, as well as the stronger modeling capability of the Transformer for global dependency.

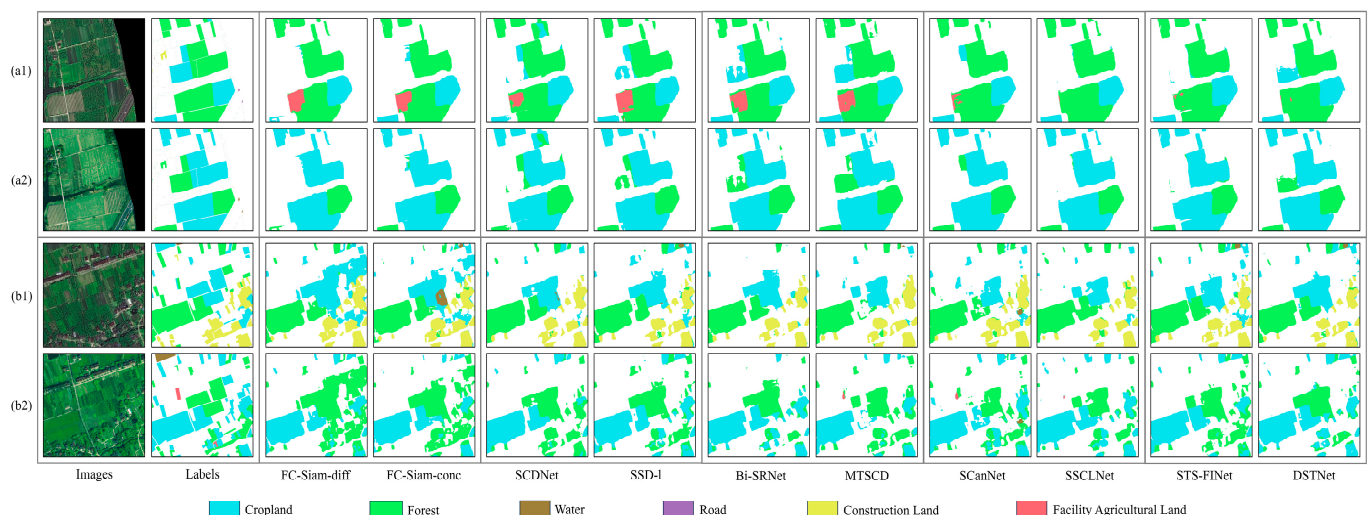


Figure 11. Benefits of DSTNet. (a1–b2) represent two different pairs of images.

6.2. Implications

The proposed method can provide more accurate land cover change information in rural areas of developed regions, which is of great significance for optimizing the alloca-

tion of rural land resources. Rural land cover change contains semantic transformation information of various land cover types, which can accurately explore the spatiotemporal characteristics and evolution patterns of rural land cover patterns and is the basic data for optimizing national land spatial layouts. The land cover change information detected by this method can strengthen rural land use planning, such as exploring the spatiotemporal evolution characteristics of blue-green space [58]. In addition, this method can also help with ecological environment monitoring and biodiversity conservation, such as the identification of mangroves [59]. The proposed method reveals the effectiveness of the fusion framework of CNNs and Transformers [60]. This method not only helps to improve the performance of semantic change detection models and increase the accuracy of rural land cover change detection but also adds an effective monitoring method for achieving accurate surveys of land cover change in rural areas.

6.3. Limitations

Although the proposed method achieves excellent classification accuracy on the rural land cover change detection dataset, its change detection capability in complex scenarios still requires further improvement. To illustrate the negative impact of this structure, we visualized several representative sample pairs, as shown in Figure 12. When the dual-temporal semantic information of a certain category is relatively complex, the segmentation capability of the proposed method decreases. For example, in Figure 12(b1,b2), cropland and forest were either unrecognized or misclassified because it is difficult to distinguish similar local features and spatial contextual information in dual-temporal features. When image features of a certain category are easily confused with those of other categories, the recognition capability of the proposed method decreases. For example, water and cropland in Figure 12(a1,a2) were not fully recognized, and cropland and construction land in Figure 12(b1,b2) were similarly overlooked. This may be due to the proposed method failing to correctly classify their semantic categories. The substantial intra-class variations in rural land cover categories contribute to the suboptimal segmentation results of the proposed method, indicating that accurate and efficient remote sensing change detection models for rural land cover still require further refinement.

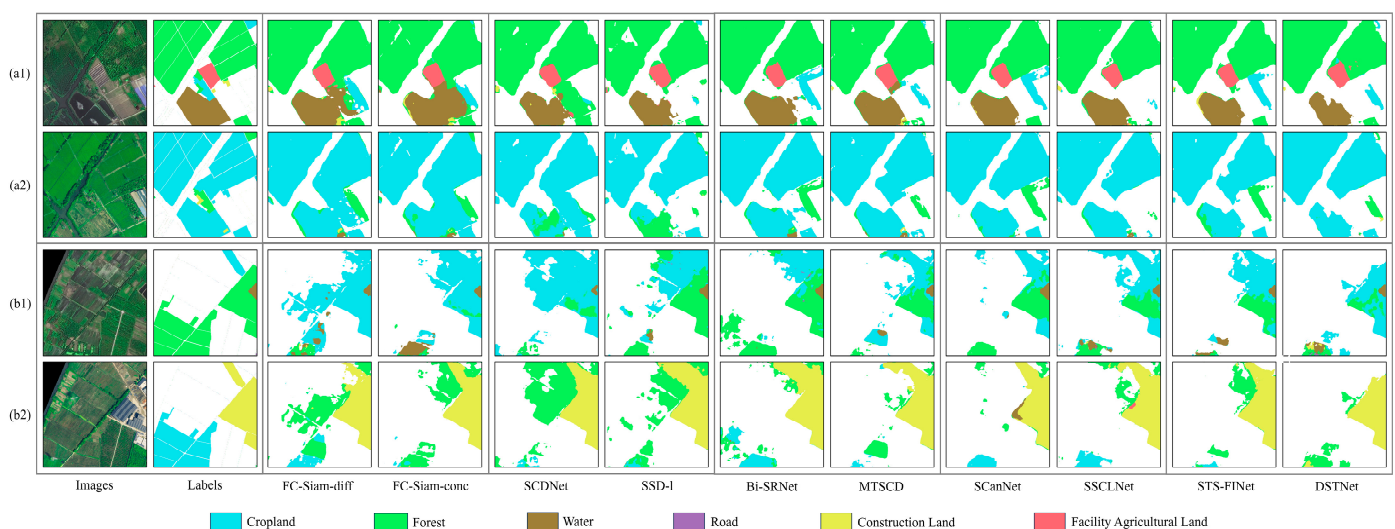


Figure 12. Limitations of DSTNet. (a1–b2) represent two different pairs of images.

7. Conclusions

In this study, we propose a network architecture that integrates a Siamese network with differential structures and the CSWin Transformer. This architecture enhances the

extraction of differential features through residual learning modules and integrates the CSWin Transformer by sequential fusion, thereby strengthening the ability to capture global dependency. In addition, we propose a high-precision rural land cover change detection dataset, encompassing six main semantic categories: cropland, forest, facility agricultural land, construction land, water, and road. Ablation experiments conducted on the RLCD dataset demonstrate that the proposed method achieves improvements of 1.77%, 1.95%, 2.57%, and 0.92% over the baseline in mIoU, Sek, Fscd, and OA, respectively, indicating a significant enhancement in change detection accuracy. In comparison to other methods, the proposed approach achieves the highest change detection accuracy, demonstrating the model's superiority. This study represents only an initial exploration of rural land cover change detection. Given the significant intra-class variations in rural land cover categories, high-precision and efficient semantic change detection models warrant further investigation.

Author Contributions: Conceptualization, B.S.; methodology, B.S.; software, B.S.; validation, B.S.; formal analysis, B.S.; investigation, B.S.; resources, K.W.; data curation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, B.S., B.D., and K.W.; visualization, B.S.; supervision, K.W.; project administration, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: DSTNet code and RLCD dataset will be announced soon. If any researcher needs them, they can contact us via email to obtain the right to use the data.

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

References

1. Singh, A. Review Article Digital change detection techniques using remotely-sensed data. *Int. J. Remote Sens.* **1989**, *10*, 989–1003. [\[CrossRef\]](#)
2. Zhang, C.X.; Yue, P.; Tapete, D.; Jiang, L.C.; Shangguan, B.Y.; Huang, L.; Liu, G.C. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [\[CrossRef\]](#)
3. Woodcock, C.E.; Loveland, T.R.; Herold, M.; Bauer, M.E. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sens. Environ.* **2020**, *238*, 111558. [\[CrossRef\]](#)
4. Zhang, J.D.; Shao, Z.F.; Ding, Q.; Huang, X.; Wang, Y.; Zhou, X.C.; Li, D.R. AERNet: An Attention-Guided Edge Refinement Network and a Dataset for Remote Sensing Building Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5617116. [\[CrossRef\]](#)
5. Liu, Z.H.; Li, J.H.; Syam, M.S.; Ashraf, M.; Asif, M.; Awwad, E.M.; Al-Razgan, M.; Bhatti, U.A. Remote sensing-enhanced transfer learning approach for agricultural damage and change detection: A deep learning perspective. *Big Data Res.* **2024**, *36*, 100449. [\[CrossRef\]](#)
6. Dai, A.J.; Yang, J.Y.; Zhang, Y.X.; Zhang, T.T.; Tang, K.X.; Xiao, X.Y.; Zhang, S.J. A difference enhancement and class-aware rebalancing semi-supervised network for cropland semantic change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *137*, 104415. [\[CrossRef\]](#)
7. Tang, X.; Zhang, T.X.; Ma, J.J.; Zhang, X.R.; Liu, F.; Jiao, L.C. WNet: W-Shaped Hierarchical Network for Remote-Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615814. [\[CrossRef\]](#)
8. Wang, Y.H.; Gao, L.R.; Hong, D.F.; Sha, J.J.; Liu, L.; Zhang, B.; Rong, X.H.; Zhang, Y.G. Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *104*, 102582. [\[CrossRef\]](#)
9. Zheng, Z.; Zhong, Y.F.; Zhao, J.; Ma, A.L.; Zhang, L.P. Unifying remote sensing change detection via deep probabilistic change models: From principles, models to applications. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 239–255. [\[CrossRef\]](#)
10. Peng, D.F.; Liu, X.L.; Zhang, Y.J.; Guan, H.Y.; Li, Y.S.; Bruzzone, L. Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *136*, 104282. [\[CrossRef\]](#)

11. Daudt, R.C.; Le Saux, B.; Boulch, A.; Gousseau, Y. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Underst.* **2019**, *187*, 102783. [[CrossRef](#)]
12. Yang, K.P.; Xia, G.S.; Liu, Z.C.; Du, B.; Yang, W.; Pelillo, M.; Zhang, L.P. Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5609818. [[CrossRef](#)]
13. Yuan, P.L.; Zhao, Q.Z.; Zhao, X.B.; Wang, X.W.; Long, X.F.; Zheng, Y.C. A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images. *Int. J. Digit. Earth* **2022**, *15*, 1506–1525. [[CrossRef](#)]
14. Liu, F.; An, J.Q.; Liu, J.; Yang, J.X.; Tang, X.; Xiao, L. Conjoint Cross-Attention Modeling and Joint Feature Calibrating for Remote Sensing Image Change Detection via a Triple-Double Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622616. [[CrossRef](#)]
15. El Amin, A.M.; Liu, Q.J.; Wang, Y.H. Zoom Out CNNs Features for Optical Remote Sensing Change Detection. In Proceedings of the 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, 2–4 June 2017.
16. Zhang, M.; Shi, W.Z. A Feature Difference Convolutional Neural Network-Based Change Detection Method. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7232–7246. [[CrossRef](#)]
17. Yuan, Y.; Chen, X.; Tang, K.; Chen, J. A “Difference-in-Differences”-Based Method for Unsupervised Change Detection in Season-Varying Images. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [[CrossRef](#)]
18. Lei, T.; Wang, J.; Ning, H.L.; Wang, X.W.; Xue, D.H.; Wang, Q.; Nandi, A.K. Difference Enhancement and Spatial-Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4507013. [[CrossRef](#)]
19. Liu, M.X.; Chai, Z.Q.; Deng, H.J.; Liu, R. A CNN-Transformer Network with Multiscale Context Aggregation for Fine-Grained Cropland Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4297–4306. [[CrossRef](#)]
20. Tang, W.J.; Wu, K.; Zhang, Y.X.; Zhan, Y.T. A Siamese Network Based on Multiple Attention and Multilayer Transformers for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5219015. [[CrossRef](#)]
21. He, F.C.; Chen, H.; Yang, S.T.; Guo, Z.X. A Hierarchical Local-Sparse Model for Semantic Change Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 3144–3159. [[CrossRef](#)]
22. Ding, Q.; Wang, F.Y.; Wang, M.C.; Zhang, Y.; Cheng, G. GLAI-Net: Global-Local Awareness Integrated Network for Semantic Change Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 14291–14307. [[CrossRef](#)]
23. Yu, W.T.; Zhuo, L.; Li, J.F. GCFormer: Global Context-Aware Transformer for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4703212. [[CrossRef](#)]
24. Song, F.; Zhang, S.X.; Lei, T.; Song, Y.X.; Peng, Z.M. MSTDSNet-CD: Multiscale Swin Transformer and Deeply Supervised Network for Change Detection of the Fast-Growing Urban Regions. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6508505. [[CrossRef](#)]
25. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional Siamese networks for change detection. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
27. Chen, H.; Qi, Z.P.; Shi, Z.W. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607514. [[CrossRef](#)]
28. Feng, Y.C.; Xu, H.H.; Jiang, J.W.; Liu, H.; Zheng, J.W. ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4410213. [[CrossRef](#)]
29. Cui, F.Z.; Jiang, J. MTSCD-Net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103294. [[CrossRef](#)]
30. Xu, C.; Ye, Z.Y.; Mei, L.Y.; Shen, S.; Zhang, Q.; Sui, H.G.; Yang, W.; Sun, S.H. SCAD: A Siamese Cross-Attention Discrimination Network for Bitemporal Building Change Detection. *Remote Sens.* **2022**, *14*, 6213. [[CrossRef](#)]
31. Cai, C.; Wang, Y.; Yap, K.H. Interactive Change-Aware Transformer Network for Remote Sensing Image Change Captioning. *Remote Sens.* **2023**, *15*, 5611. [[CrossRef](#)]
32. Li, H.; Liu, X.Y.; Li, H.H.; Dong, Z.Y.; Xiao, X.L. MDFENet: A Multiscale Difference Feature Enhancement Network for Remote Sensing Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3104–3115. [[CrossRef](#)]
33. Zhang, K.; Zhao, X.; Zhang, F.; Ding, L.; Sun, J.D.; Bruzzone, L. Relation Changes Matter: Cross-Temporal Difference Transformer for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5611615. [[CrossRef](#)]
34. Liu, W.; Kang, Z.W.; Liu, J.W.; Lin, Y.Y.; Yu, Y.T.; Li, J.A.T. A Multitask CNN-Transformer Network for Semantic Change Detection from Bitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5647215. [[CrossRef](#)]
35. Mei, L.Y.; Ye, Z.Y.; Xu, C.; Wang, H.Z.; Wang, Y.; Lei, C.; Yang, W.; Li, Y.S. SCD-SAM: Adapting Segment Anything Model for Semantic Change Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5626713. [[CrossRef](#)]
36. Jiang, Z.H.; Wang, B.; Zhang, P.; Wu, Y.L.; Ye, Z.Y.; Yang, H. Semantic enhancement and change consistency network for semantic change detection in remote sensing images. *Int. J. Digit. Earth* **2025**, *18*, 2496790. [[CrossRef](#)]

37. Zhang, D.; Wang, F.Y.; Ning, L.C.; Zhao, Z.Y.; Gao, J.Y.; Li, X.L. Integrating SAM with Feature Interaction for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4513011. [\[CrossRef\]](#)
38. Zhang, J.; Ding, L.; Zhou, T.Y.; Wang, J.; Atkinson, P.M.; Bruzzone, L. Recurrent Semantic Change Detection in VHR Remote Sensing Images Using Visual Foundation Models. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5402314. [\[CrossRef\]](#)
39. Schmitt, M.; Ahmadi, S.A.; Xu, Y.H.; Taskin, G.; Verma, U.; Sica, F.; Haensch, R. There Are No Data Like More Data: Datasets for Deep Learning in Earth Observation. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 63–97. [\[CrossRef\]](#)
40. Zan, Y.J.; Ji, S.P.; Chao, S.T.; Luo, M.Y. Open-vocabulary generative vision-language models for creating a large-scale remote sensing change detection dataset. *ISPRS J. Photogramm. Remote Sens.* **2025**, *225*, 275–290. [\[CrossRef\]](#)
41. Xiong, Z.T.; Zhang, F.H.; Wang, Y.; Shi, Y.L.; Zhu, X.X. EarthNets: Empowering artificial intelligence for Earth observation. *IEEE Geosci. Remote Sens. Mag.* **2024**, *13*, 45–78. [\[CrossRef\]](#)
42. Tian, S.Q.; Ma, A.L.; Zheng, Z.; Zhong, Y.F. Hi-UCD: A Large-scale Dataset for Urban Semantic Change Detection in Remote Sensing Imagery. *arXiv* **2020**, arXiv:2011.03247.
43. Zhou, Y.P.; Wang, J.J.; Ding, J.L.; Liu, B.H.; Weng, N.; Xiao, H.Z. SIGNet: A Siamese Graph Convolutional Network for Multi-Class Urban Change Detection. *Remote Sens.* **2023**, *15*, 2464. [\[CrossRef\]](#)
44. Shi, S.A.; Zhong, Y.F.; Liu, Y.H.; Wang, J.; Wan, Y.T.; Zhao, J.; Lv, P.Y.; Zhang, L.P.; Li, D.R. Multi-temporal urban semantic understanding based on GF-2 remote sensing imagery: From tri-temporal datasets to multi-task mapping. *Int. J. Digit. Earth* **2023**, *16*, 3321–3347. [\[CrossRef\]](#)
45. Ji, D.; Gao, S.; Tao, M.Y.; Lu, H.T.; Zhao, F. Changenet: Multi-temporal asymmetric change detection dataset. In Proceedings of the 49th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 2725–2729.
46. Liu, X.G.; Dai, C.G.; Zhang, Z.C.; Li, M.M.; Wang, H.Y.; Ji, H.L.; Li, Y.J. TBSCD-Net: A Siamese Multitask Network Integrating Transformers and Boundary Regularization for Semantic Change Detection From VHR Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6008305. [\[CrossRef\]](#)
47. Toker, A.; Kondmann, L.; Weber, M.; Eisenberger, M.; Camero, A.; Hu, J.L.; Hoderlein, A.P.; Senaras, C.; Davis, T.; Cremers, D.; et al. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 21126–21135.
48. Peng, D.F.; Bruzzone, L.; Zhang, Y.J.; Guan, H.Y.; He, P.F. SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102465. [\[CrossRef\]](#)
49. Zhang, X.R.; He, L.; Qin, K.; Dang, Q.; Si, H.J.; Tang, X.; Jiao, L.C. SMD-Net: Siamese Multi-Scale Difference-Enhancement Network for Change Detection in Remote Sensing. *Remote Sens.* **2022**, *14*, 1580. [\[CrossRef\]](#)
50. Li, W.M.; Xue, L.H.; Wang, X.Q.; Li, G. ConvTransNet: A CNN-Transformer Network for Change Detection with Multiscale Global-Local Representations. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610315. [\[CrossRef\]](#)
51. Dong, X.Y.; Bao, J.M.; Chen, D.D.; Zhang, W.M.; Yu, N.H.; Yuan, L.; Chen, D.; Guo, B.N. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12114–12124.
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
53. Ding, L.; Tang, H.; Liu, Y.H.; Shi, Y.L.; Zhu, X.X.; Bruzzone, L. Adversarial Shape Learning for Building Extraction in VHR Remote Sensing Images. *IEEE Trans. Image Process.* **2022**, *31*, 678–690. [\[CrossRef\]](#)
54. Ding, L.; Guo, H.T.; Liu, S.C.; Mou, L.C.; Zhang, J.; Bruzzone, L. Bi-Temporal Semantic Reasoning for the Semantic Change Detection in HR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620014. [\[CrossRef\]](#)
55. Ding, L.; Zhang, J.; Guo, H.T.; Zhang, K.; Liu, B.; Bruzzone, L. Joint Spatio-Temporal Modeling for Semantic Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5610814. [\[CrossRef\]](#)
56. Zhang, X.W.; Yang, Y.Z.; Ran, L.Y.; Chen, L.; Wang, K.W.; Yu, L.; Wang, P.; Zhang, Y.N. Remote Sensing Image Semantic Change Detection Boosted by Semi-Supervised Contrastive Learning of Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5624113. [\[CrossRef\]](#)
57. Zhang, Y.H.; Zhang, W.X.; Ding, S.T.; Wu, S.Y.; Lu, X.Q. Spatial-Temporal Semantic Feature Interaction Net-Work for Semantic Change Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 12090–12102. [\[CrossRef\]](#)
58. Zhang, Z.X.; Du, S.J.; Qian, L.; Qian, G.Y.; Shi, Z.W.; Yan, C. Analysis of spatial and temporal characteristics and influence mechanisms of blue-green spaces in China's, 2000–2020. *Ecol. Indic.* **2025**, *178*, 113903. [\[CrossRef\]](#)

59. Ming, X.Y.; Tian, Y.C.; Zhang, Q.; Zhang, Y.L.; Tao, J.; Lin, J.L. Coupling ICESat-2 and Sentinel-2 data for inversion of mangrove tidal flat to predict future distribution pattern of mangroves. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104398. [[CrossRef](#)]
60. Chen, M.; Zhang, Q.J.; Ge, X.M.; Xu, B.; Hu, H.; Zhu, Q.; Zhang, X. A Full-Scale Connected CNN-Transformer Network for Remote Sensing Image Change Detection. *Remote Sens.* **2023**, *15*, 5383. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.