

Article

Remote Sensing Image-Based Building Change Detection: A Case Study of the Qinling Mountains in China

Lei Fu ^{1,2}, Yunfeng Zhang ^{2,3}, Keyun Zhao ¹, Lulu Zhang ¹, Ying Li ^{1,*}, Changjing Shang ⁴  and Qiang Shen ⁴ 

¹ School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China; bobyfly@mail.nwpu.edu.cn (L.F.); kyzhao@mail.nwpu.edu.cn (K.Z.); zzhanglull@mail.nwpu.edu.cn (L.Z.)

² Shaanxi Satellite Application Center for Natural Resources, Xi'an 710065, China; yunfengzhang@mail.nwpu.edu.cn

³ School of Software, Northwestern Polytechnical University, Xi'an 710129, China

⁴ Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK; cns@aber.ac.uk (C.S.); qqs@aber.ac.uk (Q.S.)

* Correspondence: lybyp@nwpu.edu.cn

Abstract

With the widespread application of deep learning in Earth observation, remote sensing image-based building change detection has achieved numerous groundbreaking advancements. However, differences across time periods caused by temporal variations in land cover, as well as the complex spatial structures in remote sensing scenes, significantly constrain the performance of change detection. To address these challenges, a change detection algorithm based on spatio-spectral information aggregation is proposed, which consists of two key modules: the Cross-Scale Heterogeneous Convolution module (CSH-Conv) and the Spatio-Spectral Information Fusion module (SSIF). CSHConv mitigates information loss caused by scale heterogeneity, thereby enhancing the effective utilization of multi-scale features. Meanwhile, SSIF models spatial and spectral information jointly, capturing interactions across different spatial scales and spectral domains. This investigation is illustrated with a case study conducted with the real-world dataset QL-CD (Qinling change detection), acquired in the Qinling region of China. The work includes the construction of QL-CD, which includes 12,724 pairs of images captured by the Gaofen-1 satellite. Experimental results demonstrate that the proposed approach outperforms a wide range of state-of-the-art algorithms.

Keywords: building change detection; spatio-spectral information fusion; cross-scale heterogeneous convolution; building change detection dataset



Academic Editor: Zhenwei Shi

Received: 6 May 2025

Revised: 18 June 2025

Accepted: 26 June 2025

Published: 30 June 2025

Citation: Fu, L.; Zhang, Y.; Zhao, K.; Zhang, L.; Li, Y.; Shang, C.; Shen, Q. Remote Sensing Image-Based Building Change Detection: A Case Study of the Qinling Mountains in China. *Remote Sens.* **2025**, *17*, 2249. <https://doi.org/10.3390/rs17132249>

Copyright: © 2025 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Building change detection involves conducting analysis on remote sensing data acquired at different time points over the same location, to determine whether new buildings have been constructed or existing ones have been demolished. With continuous advancements in Earth observation technologies, state-of-the-art remote sensing sensors can now provide high-resolution imagery at meter- or even sub-meter-level precision. As a result, utilizing remote sensing imagery for large-scale building change detection has become a crucial Earth observation approach. Widely utilized in practice [1], it supports applications such as urban planning [2,3], natural resource conservation [4], disaster evaluation [5], and tracking land use and land cover changes [6].

Traditional change detection methods can be broadly categorized into two types: pixel-based methods and object-based methods. Pixel-based methods use individual pixels as detection units, extracting change information by analyzing spectral differences on a per-pixel basis. Common approaches in this category such as image differencing [7], statistical regression modeling [8], change vector analysis (CVA) [9,10], and principal component analysis (PCA) [11]. Object-based methods, on the other hand, analyze objects as fundamental units, allowing them to capture both spectral information and spatial context. Representative methods include those based on conditional random fields (CRFs) [12] and Markov random fields (MRFs) [13,14]. While these traditional methods offer advantages in detection efficiency and perform well in specific application scenarios, they involve significant dependence on human-designed feature representations. This dependency limits their effectiveness in complex environments.

The swift advancement of deep learning has attracted widespread interest, achieving remarkable success in many high-level interpretation tasks [15–20]. Due to the translation invariance of convolutional operations, convolutional neural networks (CNNs) exhibit robust feature representation capacity when processing image data. Thus, numerous CNN-based methods targeting change detection have been presented, including single-stream and dual-stream structures. The former merge bi-temporal images before processing, treating them as a unified entity. Daudt et al. [21], for instance, introduced a method grounded in a fully convolutional network architecture (FC-EF) [17], while Peng et al. [22] introduced a UNet++-based [23] algorithm that employs dense skip connections to capture features at multiple scales. However, these single-stream methods lack deep modeling of land cover features, which can introduce prediction bias and limit change detection accuracy. To address this issue, researchers have developed dual-stream structures, which extract bi-temporal features separately before computing their differences. Chen et al. [24] employed a Siamese architecture for feature extraction and measured feature differences using Euclidean distance. Liu et al. [25] introduced a Siamese network constrained by dual tasks, while Li et al. [26] improved upon UNet++ and introduced the Siam-NestedUNet model. Jiang et al. [27] developed PGA-SiamNet, a Siamese network utilizing pyramid feature-based attention. These dual-stream architectures enable independent modeling of features from images, while maintaining the same parameter efficiency as single-stream networks through shared weights. This significantly enhances change detection performance.

Different from CNNs that treat all regions of an image with equal importance, the attention mechanism dynamically adjusts the weights of different regions [28,29]. Various attention-based approaches have been developed. Zhang et al. [30] employed spatial attention mechanisms [31] and channel attention mechanisms [32] to integrate deep hierarchical features and bi-temporal difference features. Song et al. [33] proposed an attention-guided network, which enhances the distinction by leveraging both spatial and channel information. Chen et al. [34] introduced DASNet, which captures long-range dependencies using dual attention mechanisms. Fang et al. [35] designed an integrated attention module to refine multi-level semantic feature information, extracting the most representative features. In addition, many attention-based methods have also achieved certain performance [36–39].

Given the vast land cover information and the complex structural in remote sensing images, single-level feature fusion often fails to effectively model the intricate relationships among different land cover types. To overcome this, multi-level feature fusion strategies [40–42] are developed. An alternative approach involves employing spatial or channel attention mechanisms [34,36,38,43–45] to highlight key information. However, multi-scale feature integration may cause loss of fine details, producing smoothed features. Although many deep learning change detection methods apply spatial and channel attention—sometimes combined—they often neglect the synergy between spatial and spectral features. Moreover,

these methods usually introduce many parameters, increasing computational cost and limiting practical use.

The scale, quality, and completeness of datasets influence deep learning detection performance by enhancing model generalization and representation. Thus, large-scale, high-quality datasets are vital for remote sensing progress. Public datasets like LEVIR-CD [24], WHU-Building [46], S2Looking [47], and CDD [48] exist but mainly cover single-scene scenarios with limited geographic and environmental diversity, restricting model adaptability and representation of complex land cover.

To address the aforementioned challenges, SSA-Net is herein proposed on the basis of spatio-spectral information aggregation, and a diverse large-scale dataset is established. First, a Cross-Scale Heterogeneous Convolution module is developed to effectively utilize multi-scale information and mitigate information loss caused by scale differences. Second, a Spatio-Spectral Information Aggregation module is developed, which efficiently captures and integrates spatio-spectral information across different scales. Finally, a change detection study is conducted in the Qinling region, constructing a large-scale dataset that encompasses diverse scenarios, including mountain ranges, forests, rural areas, and nature reserves.

The primary contributions are outlined below:

- (1) A Cross-Scale Heterogeneous Convolution (CSHConv) module is introduced to precisely capture key change information across multiple scales.
- (2) A Spatio-Spectral Information Aggregation (SSIF) module is designed to comprehensively model the complex spatial-spectral relationships between land cover features.
- (3) An extensive experimental study is conducted in the real-world Qinling region, resulting in a new change detection dataset, consisting of 12,724 pairs of images captured by the Gaofen-1 satellite. This dataset covers diverse landscapes, including mountains, forests, rural areas, and nature reserves, providing a valuable resource for future research.

The remainder of this paper is organized as follows: Section 2 provides a detailed introduction to the dataset. Section 3 presents an in-depth explanation of the proposed SSA-Net. Section 4 describes the experiments and analysis. Section 5 discusses the computational complexity of different methods. Finally, Section 6 summarizes the study and discusses future research directions.

2. Dataset

As deep learning is inherently data-driven, its performance largely depends on the scale, quality, and completeness of the training dataset. Consequently, there is an increasing demand for large-scale, high-quality change detection datasets. To address the challenges posed by existing building change detection datasets, a well-annotated dataset regarding the region over the Qinling Mountains, QL-CD, is constructed. This dataset consists of 12,724 pairs of satellite images with a spatial resolution of 2 m and a patch size of 256×256 pixels. Below is a detailed introduction to QL-CD, including the area covered, annotation process, and preprocessing methods, followed by a comprehensive statistical analysis.

2.1. Study Regions

The Qinling Mountains are located in central China, extending across southern Shaanxi Province. They serve as a critical climatic transition zone between northern and southern China and form the watershed between the Yangtze River and the Yellow River basins. The QL-CD dataset covers the central segment of the Qinling Mountains within Shaanxi Province, spanning a geographical range of $106^{\circ}03' - 110^{\circ}00'E$, $32^{\circ}4' - 34^{\circ}33'N$, with a total

area of approximately 58,000 km². As illustrated in Figure 1, the dataset encompasses 39 districts and counties across Baoji, Xi'an, Hanzhong, Ankang, and Shangluo.

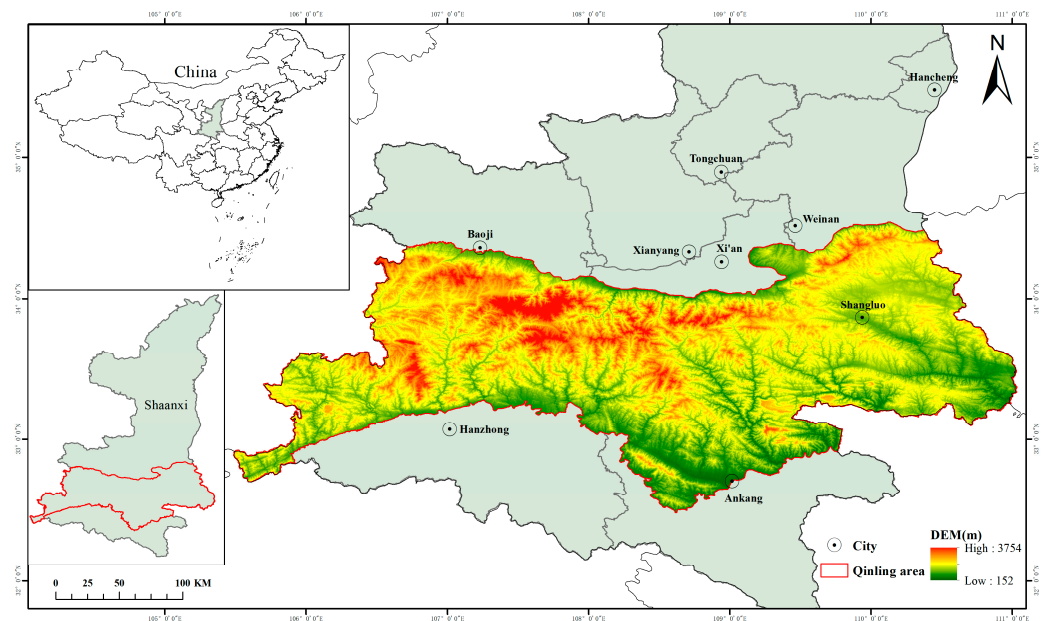


Figure 1. Location of study area.

The topography of the central Qinling region is highly rugged. The northern slopes are steep and characterized by deep valleys, while the southern slopes are more gradual, exhibiting a distinct north-steep–south-gentle mountain morphology. The region has an average elevation exceeding 1000 m, with certain peaks surpassing 3000 m. This complex terrain structure introduces significant spatial heterogeneity in human settlement distribution and building change dynamics. The selected study areas feature diverse and unique landscapes with a wide range of land cover types, including mountains, forests, rural settlements, and nature reserves. These geographical and environmental characteristics present both challenges and opportunities for building change detection systems to work in mountainous regions.

In this study, the dataset is derived from very-high-resolution (VHR) satellite imagery captured by Gaofen-1 (GF-1). The bi-temporal images were acquired in 2018 and 2022, covering the Qinling region and its surrounding areas. The imagery has a spatial resolution of 2 m and consists of three visible spectral bands (red, green, and blue).

2.2. Data Annotation and Preprocessing

Annotation: Compared to single-temporal image annotation, labeling multi-temporal remote sensing datasets is a significantly more complex task. Not only does it require annotating a larger number of change targets, but it also involves extensive cross-temporal region comparisons, increasing the overall workload. Moreover, the complex topography of mountainous regions and variations in imaging conditions across different acquisition times introduce additional challenges. In the Qinling region, the appearance and geographic characteristics of buildings can undergo substantial changes due to seasonal and weather variations, further complicating the annotation process.

To address these challenges, a refined annotation workflow is designed that ensures both high accuracy and efficiency. The annotation process carried out consists of multiple stages, each incorporating strict quality control measures to maintain the trustworthiness of the resulting dataset.

To manage the complexity of multi-temporal dataset annotation, a phased annotation strategy was adopted. Particularly, a 15-member professional annotation team was assembled, each with extensive experience in remote sensing image interpretation. The team received specialized training focused on mountainous terrain characteristics and building change patterns to enhance their expertise. Using professional GIS tools such as ArcGIS, annotators carefully delineated building change areas, based on their training and domain knowledge. In cases where images exhibited partial occlusion or distortion, the team leveraged external references, such as Google Maps and other Geographic Information Systems (GIS) resources, for cross-validation.

To further enhance annotation accuracy and consistency, a rigorous quality control framework was implemented, consisting of three key verification steps: (1) Each annotation result was reviewed by a second annotator, ensuring the detection of potential errors or omissions and maintaining high inter-annotator consistency. (2) Upon completion of the annotation process, domain experts conducted random spot checks to verify the correctness and completeness of the labeled data. (3) By integrating cross-validation, expert auditing, and iterative refinements, errors were significantly minimized, ensuring the high reliability of the final dataset.

Through such a systematic and meticulous annotation process, a highly accurate dataset tailored for building change detection is produced in mountainous environments. It lays a strong groundwork for future remote sensing image interpretation and model training over complex landscapes.

Preprocessing: To ensure consistency and usability of the dataset, a structured preprocessing pipeline was applied, which includes vector-to-raster conversion, image tiling, geographic metadata preservation, and secondary quality checks. Specifically, the precisely annotated vector files were converted into binary raster masks, where changed areas were assigned a value of 255, and unchanged areas were set to 0. The original high-resolution remote sensing images were segmented into separate 256×256 GeoTIFF blocks, ensuring compatibility with modern GPUs and deep learning frameworks. No overlap was introduced between image patches, facilitating efficient processing while maintaining spatial integrity. Geographic coordinates were preserved for each image patch, allowing them to be reassembled into a full reference map when needed. Each image block was sequentially numbered, providing a structured format for large-scale mapping and further applications. To enhance dataset relevance and precision, irrelevant regions were manually filtered out, ensuring that the final dataset focuses strictly on meaningful change areas.

Despite the rigorous quality control implemented during annotation, minor errors might still be present. To address this issue, a secondary review was conducted through cross-validation by multiple reviewers. This process effectively identified and removed images with unclear or incorrect annotations, ensuring high accuracy and reliability in the final dataset. By applying the aforementioned meticulous preprocessing steps, a well-structured, high-quality dataset optimized for building change detection is created for remote sensing applications.

Through this systematic processing workflow, the work implemented has not only ensured the high quality of the dataset but also provided an efficient and standardized input for subsequent building change detection models. The meticulous operations applied during the image tiling and filtering stages significantly enhance their effectiveness. Moreover, the systematic approach adopted in building this dataset lays a strong groundwork for remote sensing image analysis and change detection, promoting progress in both scholarly research and practical applications.

2.3. Dataset Analysis

Compared to existing building change detection datasets, QL-CD offers significant advantages in terms of coverage area, scene diversity, background complexity, and illumination variations. The key advantages of QL-CD are further emphasized as follows:

- (1) **Extensive Geographic Coverage:** The QL-CD dataset encompasses 12,724 image pairs collected over a vast 58,000 km² area. Compared to existing datasets, QL-CD covers a significantly larger geographic region, making it one of the most comprehensive datasets in this domain. Specifically, the dataset represents a ground area of over 3300 km², with change regions covering approximately 367 km². Each image pair captures rich land cover variations, posing a more challenging benchmark for evaluating model performance in detecting change regions. This extensive coverage enhances the dataset's utility for performing real-world change detection tasks across diverse environments.
- (2) **Diverse Scene Coverage:** As shown in Figure 2, the QL-CD dataset includes rich scene types, such as urban regions, suburban regions, rural settlements, hills, and rivers. This scene diversity poses a greater challenge for change detection algorithms, as it requires strong adaptability to model variations in complex environments. Additionally, it can be expected to help enhance the generalization capability of models trained on this dataset. This is because compared to existing datasets like LEVIR-CD and CDD, QL-CD not only provides a more comprehensive diversity of scenes but also serves as a multi-layered data resource for in-depth research and analysis. Such an extended scene coverage ensures that models developed using QL-CD are adaptable to more real-world scenarios.

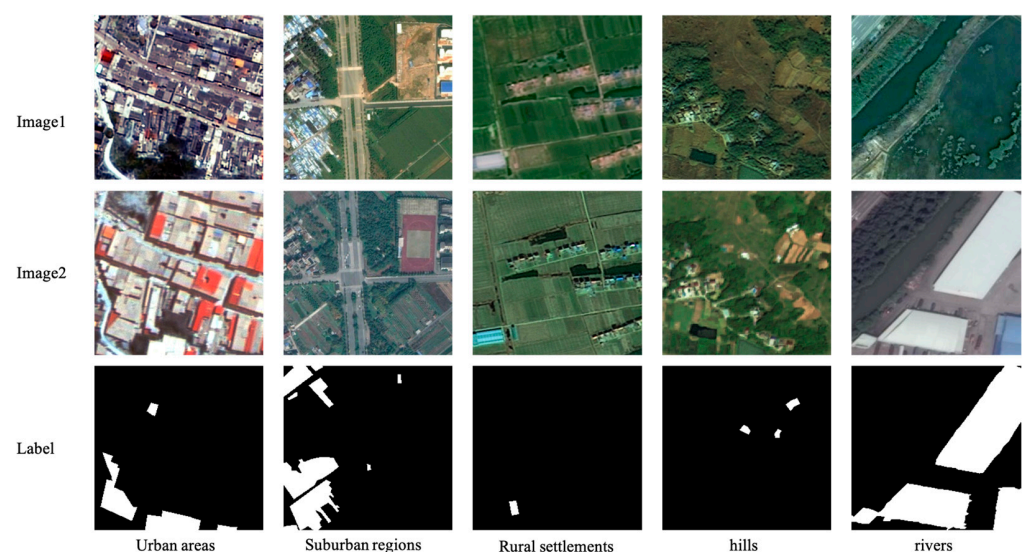


Figure 2. Diverse scene coverage of QL-CD.

- (3) **High Background Complexity:** Most existing change detection datasets primarily focus on specific urban areas, where buildings are typically present against simplistic backgrounds such as streets and roads. In contrast, the QL-CD dataset not only retains these urban elements but also significantly expands the variety of background types, including lakes, grasslands, farmland, low vegetation, and bare land. This diverse background complexity, as illustrated in Figure 3, introduces additional challenges for change detection algorithms, requiring them to distinguish between building-related changes and natural environmental variations. The inclusion of such varied backgrounds enhances the dataset's practical value, making it a more realistic and robust benchmark for real-world applications.

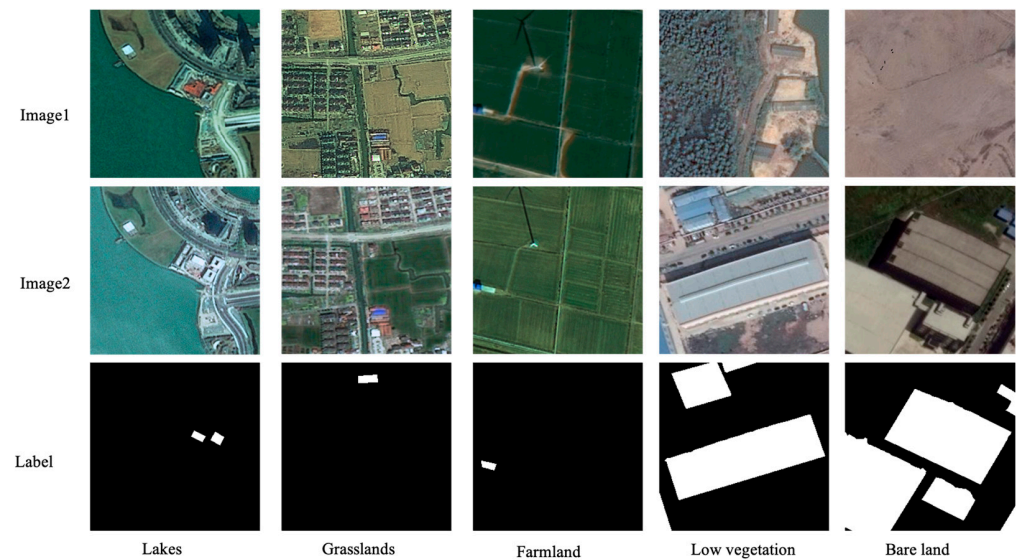


Figure 3. High background complexity of QL-CD.

- (4) **Illumination Heterogeneity:** As shown in Figure 4, the QL-CD dataset exhibits significant illumination heterogeneity, with noticeable variations in brightness, saturation, contrast, and overall image style between the two temporal images. Unlike conventional datasets captured under uniform lighting conditions, QL-CD introduces a greater degree of illumination variability, making it more representative of real-world remote sensing scenarios. This heterogeneity enables models to better capture dynamic surface changes, including seasonal transitions, meteorological variations, and natural events that impact land cover. Additionally, illumination-induced pseudo-changes present an extra challenge for algorithms, requiring them to distinguish actual building changes from lighting variations. As a result, models trained on QL-CD can be expected to achieve greater robustness with improved generalization.

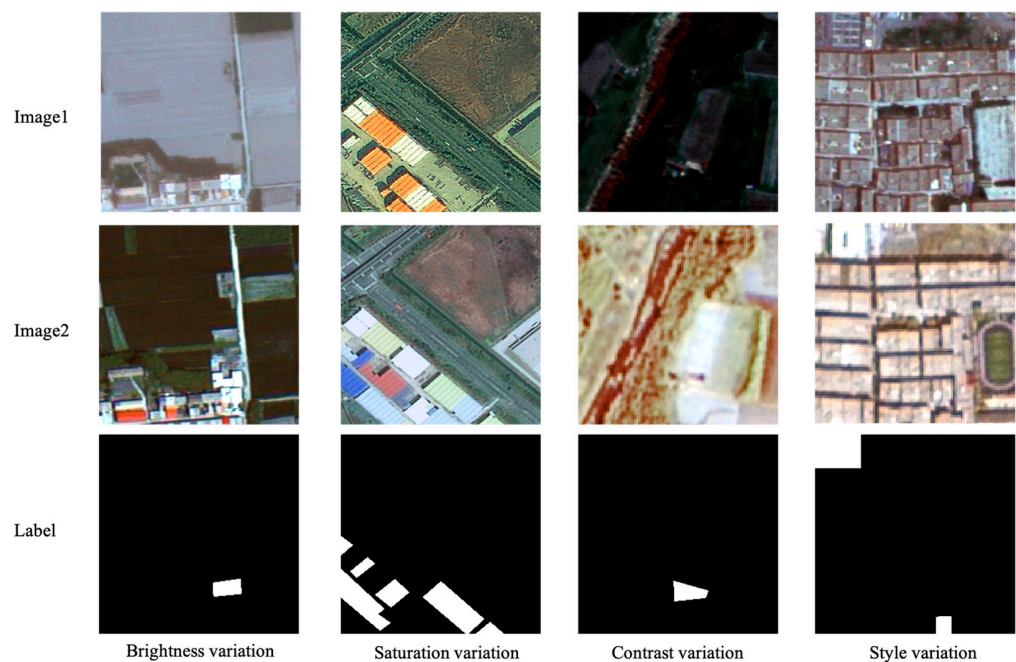


Figure 4. Illumination heterogeneity of QL-CD.

3. Methodology

To enhance the performance, a Spatio-Spectral Information Aggregation Change Detection Network (SSA-Net) is proposed, as illustrated in Figure 5. Unlike traditional networks, SSA-Net incorporates two novel modules:

1. Cross-Scale Heterogeneous Convolution (CSHConv) module
2. Spatio-Spectral Information Fusion (SSIF) module

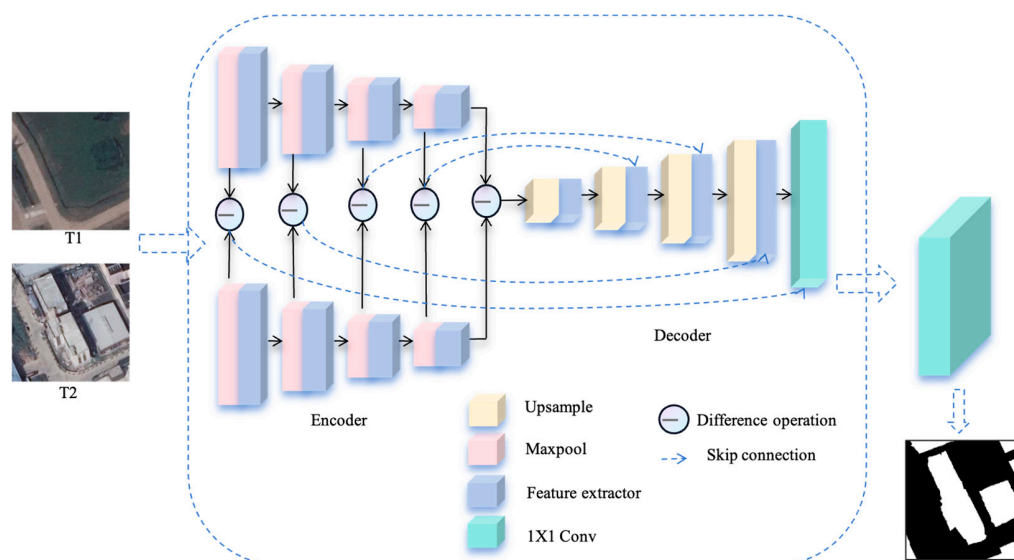


Figure 5. The architecture schematic of SSA-Net.

The CSHConv module mitigates information degradation caused by scale heterogeneity in standard convolutional kernels when processing land cover changes in remote sensing images. Meanwhile, recognizing the interdependencies between spatial and spectral channel information, the SSIF module is employed to facilitate feature fusion, effectively capturing cross-scale and cross-channel interactions. This enhanced information aggregation significantly improves model accuracy in detecting building changes.

3.1. Overview

A U-shaped network with a non-shared pseudo-Siamese structure [49] is adopted as the backbone of SSA-Net. The introduction of a non-weight-sharing encoder enhances flexibility in learning feature representations, while skip connections facilitate the interaction of temporal difference information between bi-temporal images. Bi-temporal images are first fed into the encoder, where four successive downsampling steps are performed to progressively extract multi-level features.

Instead of traditional convolutions, the Cross-Scale Heterogeneous Convolution (CSHConv) module is employed, which integrates different receptive fields to effectively capture multi-scale change information. Meanwhile, the Spatio-Spectral Information Fusion (SSIF) module is utilized to aggregate spatial and spectral information, ensuring a comprehensive representation of image features and an efficient encoding of semantic information. Using different convolutional operations at each stage, SSA-Net gradually extracts rich semantic, spatial, and spectral channel information, facilitating improved recognition and representation of change objects by the network in bi-temporal images. The decoder is designed to be approximately symmetric to the encoder, supporting effective upsampling of feature maps. This balanced structure helps keep the high-level multi-scale features acquired during encoding, preserving fine object details and structural integrity. The decoder applies four successive transposed convolutions to restore spatial resolution, followed by a final

refinement step using a 1×1 convolution combined with an activation function to fine-tune the feature representations and generate the final change map. This architecture allows SSA-Net to effectively capture, analyze, and reconstruct change information, entailing high accuracy in building change detection.

3.2. Cross-Scale Heterogeneous Convolution Module

Traditional convolutional neural networks (CNNs) utilize single-scale convolutional kernels, meaning that kernels of the same size compute feature information within the same spatial region. Due to this design limitation, the extracted features may lack comprehensiveness and completeness. As a result, CNNs may struggle to effectively model complex multi-scale variations present in remote sensing images. To overcome this limitation, we propose the Cross-Scale Heterogeneous Convolution (CSHConv) module, which addresses the information loss caused by scale heterogeneity in traditional convolutional kernels when processing land surface changes in remote sensing imagery. By incorporating multiple receptive fields, CSHConv enables more effective capture of multi-scale changes, which improves both accuracy and robustness in detection models.

Figure 6 illustrates the basic architecture of CSHConv. In CSHConv, the input feature f is first processed using a 1×1 convolution, which reduces the channel dimension by half, thereby obtaining the initial feature representation. This operation can be mathematically expressed as follows:

$$O_1 = \sigma(\text{Conv}_{1 \times 1}(f)) \quad (1)$$

where $\sigma(\cdot)$ represents the ReLU activation function, $\text{Conv}_{1 \times 1}$ denotes the 1×1 convolution operation, and O_1 is the feature map obtained after convolution. Subsequently, O_1 passes through heterogeneous convolution to obtain multi-scale feature information. The design of heterogeneous convolution involves the operation with different kernel sizes, where a portion of the kernels are 3×3 , and the rest are 1×1 . This design allows the network to extract local fine-grained details (with 3×3 kernels) while preserving global semantic information (with 1×1 kernels), effectively addressing the scale heterogeneity issue in remote sensing images.

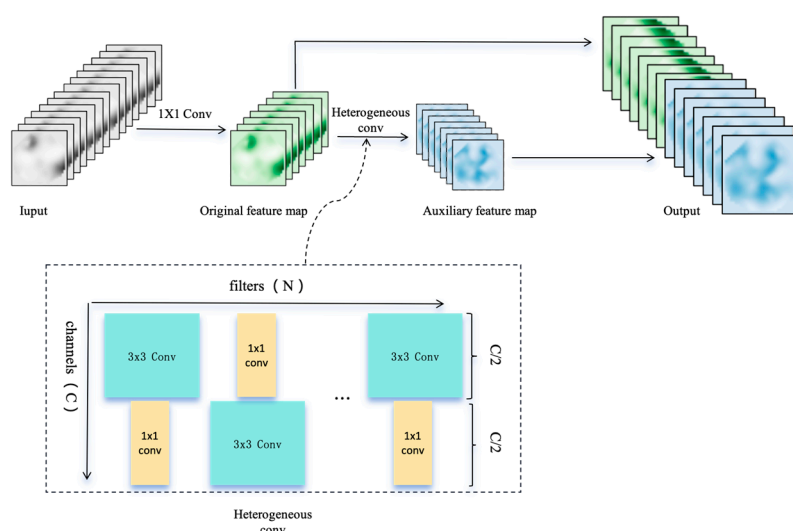


Figure 6. Cross-Scale Heterogeneous Convolution module.

Formally, the heterogeneous convolution operation can be mathematically expressed as follows:

$$O_2 = \text{Conv}_{3 \times 3}(O_1) + \text{Conv}_{1 \times 1}(O_1) \quad (2)$$

Also, the 3×3 convolution operation in the CSHConv module can be expressed by

$$\text{Conv}_{3 \times 3}(x, y, o) = \sum_{i=0}^{I/G-1} \sum_{j=0}^{K-1} \sum_{k=0}^{K-1} X(x+j, y+k, i+o \times I/G) \times \text{kernel}(j, k, i, o) \quad (3)$$

where X represents the input tensor, which corresponds to O_1 ; (x, y) denotes the spatial coordinates of the feature; K is the kernel size, defining the spatial extent of the convolution; I is the number of input channels in the feature map; G indicates how many groups are in the grouped convolution; I/G represents the input channel number per group in grouped convolution; $\text{kernel}(j, k, i, o)$ represents the convolution kernel's weight parameters, with j, k indicating the spatial offsets of the kernel, representing its position in the convolution window (i.e., row and column offsets), i is the index of the channel group, with $(i+o \times I/G)$ being the input channel index, denoting which input channel the kernel operates on, and o being the output channel index, specifying the feature map's corresponding output channel.

Each convolution kernel consists of a set of trainable weights that perform a weighted summation over different input channels and spatial positions. This operation generates an element of the output tensor, by computing a local weighted sum across the corresponding receptive field. The obtained feature information is then concatenated with the original feature map to produce the final prediction. This process can be formally summarized as follows:

$$O_{ans} = \text{concat}(\sigma(O_2), O_1) \quad (4)$$

The CSHConv module serves as a compact feature extraction unit, replacing the standard convolution operations in the encoder–decoder structure. It aims to obtain multi-scale representations of change objects, helping deep learning models effectively detect variations at multiple scales. The design of CSHConv leverages heterogeneous convolution, which incorporates various kernel sizes (e.g., some kernels are 3×3 while others are 1×1). The core idea behind this heterogeneous convolution is to leverage multiple receptive fields, enabling the network to extract multi-scale contextual information and enhance feature learning for change detection. By integrating varied convolutional kernels, CSHConv preserves fine information details while capturing broader spatial structures, making it particularly suitable for detecting complex and multi-scale changes.

3.3. Spatial and Spectral Information Fusion Module

As previously noted, straightforward fusion of auxiliary and original feature maps might cause information loss in complex scenes. To resolve this, the Spatio-Spectral Information Fusion (SSIF) module is proposed, which refines the feature fusion process by explicitly modeling interactions across different spatial scales and spectral channels. This design ensures that the network effectively captures and integrates complementary information.

Typically, attention mechanisms include three main components: Query (Q), Key (K), and Value (V). These elements work together to establish strong correlations between intent (query) and target (key). In this context, the query represents the system's intent, while the key represents the regions of interest in the image. The attention mechanism enhances feature representation by calculating the interaction between query and key, then applying this to the corresponding value, allowing the system to concentrate on important image regions.

To compute the correlation between query and key, following the Nadaraya–Watson kernel regression [50], a novel spatio-spectral information aggregation strategy is proposed.

Specifically, it utilizes Gaussian kernel-based spatio-spectral information aggregation, implementing the strategy below:

$$Y = \text{sigmoid}\left(\sum_{i=1}^{n=2} \frac{(Q_i - K)^2}{2\sigma_i^2} + \frac{1}{2}\right) \times V \quad (5)$$

where the Gaussian kernel function is used to measure the similarity between the query Q_i and the key K , with σ_i controlling the sensitivity of similarity computation. By performing a weighted summation across all similarity scores, the final similarity measure is obtained. This formulation is commonly applied, particularly in Nadaraya–Watson kernel regression, which is often used for non-parametric estimation of relationships between variables. In machine learning and statistics, employing such kernel functions helps capture complex dependencies between input variables, thereby increasing the model’s effectiveness in learning complex patterns [51].

Figure 7 illustrates the detailed process of the Spatio-Spectral Information Fusion (SSIF) module. In this framework, both K (Key) and V (Value) originate from the input feature map $X \in \mathbb{R}^{C \times H \times W}$, where C , H , and W represent the channel dimension, height, and width of the feature map, respectively. Channel-wise Query Q_1 is extracted from the channel information, represented as the mean of X along the channel axis, denoted as $X \in \mathbb{R}^{C \times 1 \times 1}$. Spatial-wise Query Q_2 is extracted from the spatial dimension, represented as the mean of X along the spatial axis, denoted as $X \in \mathbb{R}^{1 \times H \times W}$. Additionally, the variances σ_1^2 and σ_2^2 are computed for the channel and spatial information, respectively. These variance values play a crucial role in controlling the richness of feature representations: a greater variance σ_i^2 indicates greater variance, meaning that the feature map contains richer contextual information. This formulation allows the model to capture attention relationships across different dimensions, leveraging both spatial and channel statistics.

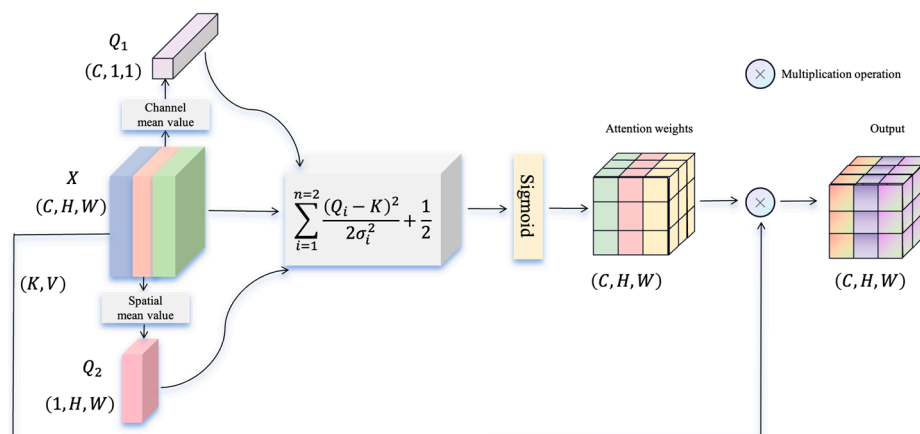


Figure 7. Spatial and Spectral Information Fusion module.

To ensure that attention weights positively contribute to feature learning, the following regularization steps are applied: (1) A bias term $\frac{1}{2}$ is added to the raw attention scores to stabilize the attention values. (2) To normalize attention scores, a Sigmoid activation is used, allowing weights to reflect the relevance of diverse regions in the feature map. The computed attention weights are then multiplied with the input feature map X to generate the final output feature map Y .

Figure 8 illustrates the structural diagram of integrating SSIF into CSHConv, demonstrating their synergistic effect. CSHConv effectively captures multi-scale features, refining the representation of changed objects with greater precision. Simultaneously, SSIF efficiently aggregates spectral information across different channels, leveraging spectral dependencies to provide a more accurate depiction of surface changes. Combining SSIF with CSHConv

strengthens the model's capacity to represent and capture input data, which is especially beneficial for complex scenes and multi-scale information.

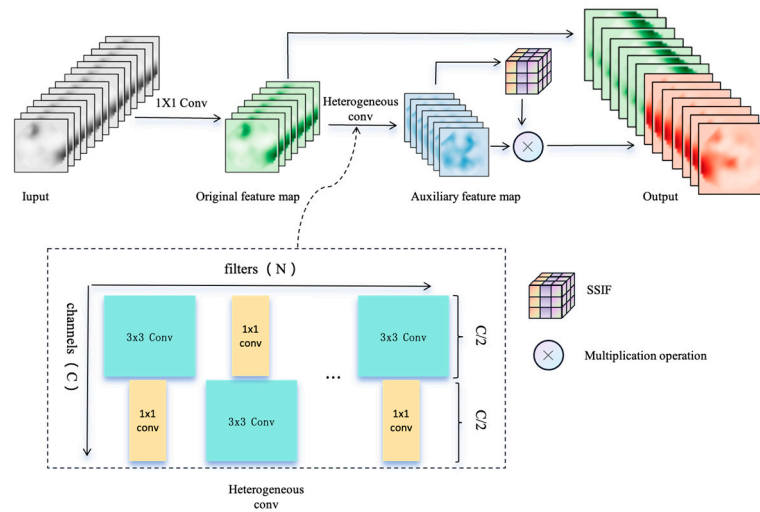


Figure 8. Architecture schematic of integrating SSIF into CSHConv.

3.4. Loss Function of SSA-Net

Change detection is commonly formulated as a pixel-wise binary classification problem, where the Binary Cross-Entropy (BCE) loss function is used to measure the discrepancy between the model's predictions and ground truth labels. However, due to the class imbalance issue (where the proportion of changed regions is significantly smaller than that of unchanged regions), BCE loss alone may lead to biased learning, favoring the majority class (unchanged pixels). To address this, a mixed loss function is used, inspired by A2Net [52], which combines BCE [53] and Dice Loss [54]. This hybrid approach balances the impact of both loss components, ensuring that the model learns effectively from both changed and unchanged regions.

For each sample, the BCE Loss and the Dice Loss are, respectively, computed as follows:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (6)$$

$$L_{dice}(y, \hat{y}) = 1 - \frac{2 \sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i^2 + \sum_{i=1}^N \hat{y}_i^2 + \epsilon} \quad (7)$$

where N is the total number of pixels in the image, y_i is the ground truth label for the i -th pixel, $\hat{y}_i$ is the predicted probability of change for the i -th pixel output by the network, and ϵ is a small constant to prevent division by zero.

The overall loss for SSA-Net is computed by summing BCE and Dice losses, defined as follows:

$$L_{loss}(y, \hat{y}) = L(y, \hat{y}) + L_{dice} \quad (8)$$

4. Experimental Studies

4.1. Implementation Setup

SSA-Net is fully implemented using PyTorch 2.1.0 and all experiments are conducted on two American Nvidia RTX 4090 GPUs (Santa Clara, CA, USA). The Adam optimizer is used for training, with momentum set to 0.9 and weight decay set to 0.0005 as typically performed in the relevant literature. A polynomial decay learning rate scheduler (lr_scheduler) is applied, with a decay cycle of 50 epochs. The initial learning rate is set to 1×10^{-4} . The max epoch number is set to 200. The batch size is set to 8 for all methods.

4.2. Evaluation Metrics

To quantitatively assess different models, researchers typically rely on a set of commonly used evaluation metrics. These metrics not only help in understanding the accuracy of an algorithm but also serve as objective standards for comparing different methods across various datasets.

As stated before, change detection is considered as a binary classification task. For evaluation, performance metrics are derived from the confusion matrix. Table 1 shows the standard binary classification confusion matrix, dividing all pixels into four categories based on the comparison between model predictions and ground truth labels: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

Table 1. Confusion matrix.

Prediction\Ground Truth	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Based on confusion matrices, four commonly used evaluation metrics are herein adopted to assess the accuracy of the change detection results: Precision, Recall, F1-score, and Intersection over Union (IoU). Their mathematical formulations are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (12)$$

4.3. Methods Compared

To comprehensively evaluate SSA-Net, 12 change detection methods are compared, including three widely applied classic algorithms: FC-EF [21], FC-Siam-Conc [21], and FC-Siam-Diff [21]; six high-performing recent algorithms evaluated on public datasets: STANet and its two derivatives (STANet_BAM and STANet_PAM) [24], SNUNet [35], A2Net [52], and HANet [55]; and three Transformer-based methods: BIT [56], Changer [57], and ChangeFormer [58]. These methods are briefly introduced below for academic completeness.

FC-EF merges bi-temporal images early and utilizes a Fully Convolutional Network for detecting changes; FC-Siam-Conc integrates features from both times through skip connections in a Siamese FCN for long-range mapping; FC-Siam-Diff detects changes by computing differences between features in a Siamese network setup. STANet series [24] introduces a spatiotemporal attention module into a Siamese network and employs a pyramid attention mechanism to leverage spatiotemporal dependencies, generating more expressive region change features. SNUNet [35] designs a densely connected U-shaped Siamese network, incorporating a channel attention module to optimize different semantic features. A2Net [52] utilizes a lightweight backbone and enhances change information through attention mechanisms. HANet [55] adopts a progressive sampling approach balanced toward the foreground, which improves the model's early-stage change identification and detection accuracy. BIT [56] presents a Transformer encoder that captures spatiotemporal relationships using context-aware token representations to extract high-level semantic features. Changer [57] presents a general change detection framework, improving detec-

tion performance through feature interaction and fusion strategies. ChangeFormer [58] integrates a hierarchical Transformer in Siamese architecture, enabling stronger multi-scale information modeling capability.

4.4. Experimental Results on QL-CD

Table 2 presents the quantitative results of running different methods across five evaluation metrics: Precision, Recall, F1, IoU, and Kappa, with the highest scores highlighted in bold. The experimental results demonstrate that FC-EF, FC-Siam-Conc, and FC-Siam-Diff exhibit relatively poor performance. In contrast, algorithms such as STANet, Changer, and ChangeFormer achieve better performance by employing distinct feature fusion approaches and attention mechanisms. Compared with the 12 existing change detection methods, the proposed SSA-Net attains the best performance on the two comprehensive metrics of F1 and IoU, achieving scores of 84.15% and 74.64%, respectively, surpassing the second-best algorithm (STANet_PAM) by 2.65% and 3.44%. Furthermore, SSA-Net also achieves the highest Kappa coefficient of 82.43%, significantly outperforming STANet_PAM's 79.54% by 2.89%. This superior Kappa score reinforces the model's overall effectiveness in achieving high agreement beyond chance, aligning with its leading performance in F1 and IoU. Notably, from an application perspective, the F1 metric holds greater significance as it requires a balanced performance between Recall and Precision to ensure robust and effective detection results.

Table 2. Performance of different methods on QL-CD.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)	Kappa (%)
FC-EF	47.64	46.36	46.99	30.71	40.99
FC-Siam-Conc	68.43	37.03	48.06	31.63	44.00
FC-Siam-Diff	82.49	26.08	39.63	24.72	36.49
STANet	65.14	76.97	70.56	54.51	66.86
STANet_BAM	75.58	85.21	79.22	68.28	77.67
STANet_PAM	79.25	84.34	81.50	71.20	79.54
SNUNet	87.23	74.64	79.27	69.04	78.39
HANet	77.09	55.06	64.23	47.31	60.88
A2Net	88.96	72.81	80.08	66.78	78.04
BIT	70.93	69.25	70.04	59.23	66.69
Changer	71.92	70.35	71.09	60.22	67.85
ChangeFormer	79.39	69.21	72.86	62.34	71.19
SSA-Net	88.70	80.04	84.15	72.64	82.43

Figure 9 qualitatively illustrates the prediction results of the proposed method and other approaches on the QL-CD test set images. Here, T1 and T2 represent a pair of multi-temporal images to be analyzed. "GT" denotes the ground truth. As shown in the figure, for scenarios where the multi-temporal images exhibit consistent styles and simple backgrounds, all methods can roughly localize changed buildings. However, the proposed method demonstrates highly consistent edge details between predictions and ground truth labels. In complex scenes with cluttered backgrounds, other algorithms show limitations in handling the continuity of change regions and edge details, accompanied by missed or false detections. In contrast, SSA-Net effectively addresses such incoherent cases, accurately localizes changed objects, and robustly suppresses background interference across bi-temporal images. For scenarios with inconsistent imaging styles between multi-temporal phases, SSA-Net significantly outperforms all comparison methods, achieving the best performance by precisely localizing change regions, which benefits from its domain consistency constraints.

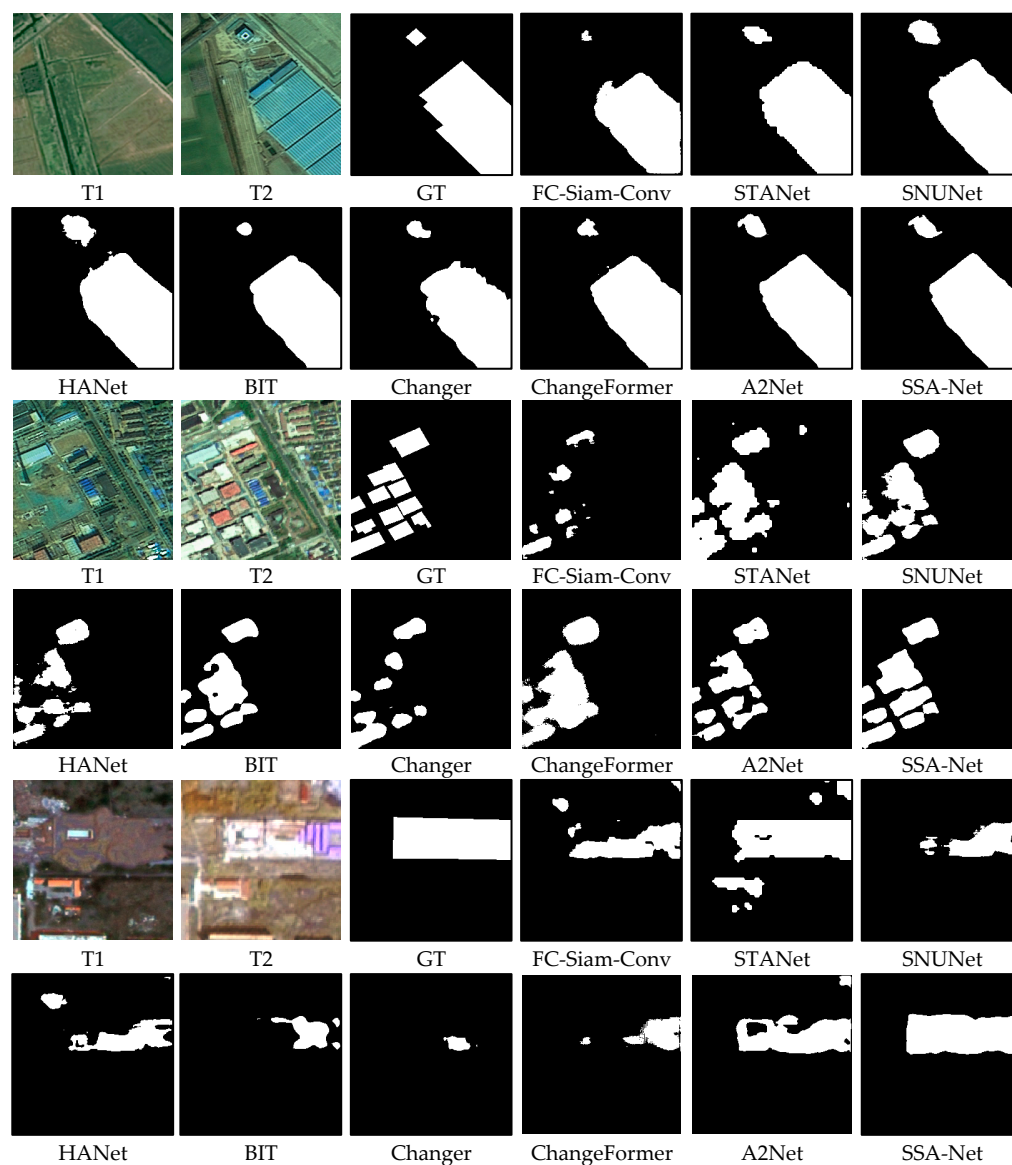


Figure 9. Change detection results on QL-CD.

4.5. Ablation Investigation

To further investigate the value of the proposed CSHConv and SSIF modules in SSA-Net, ablation studies are carried out on the QL-CD dataset. The experimental protocol involves incrementally integrating each module into the backbone and evaluating the resulting accuracy improvements. Table 3 shows the outcomes.

Table 3. Ablation experiment of SSA-Net on QL-CD.

Methods	Precision (%)	Recall (%)	F1 (%)	IoU (%)	Kappa (%)
Backbone	86.11	75.32	80.35	70.62	78.26
Backbone + CSHConv	86.93	76.07	81.13	71.50	79.13
Backbone + SSIF	83.85	79.14	81.14	68.87	79.36
SSA-Net	88.70	80.04	84.15	72.64	82.43

The first step is to verify whether the design of CSHConv enhances building change detection performance. This module addresses the issue of information loss caused by scale heterogeneity when using traditional designs. By replacing standard convolutions with heterogeneous convolutions, CSHConv captures multi-scale feature representations of change objects, enabling the deep learning model to more effectively learn multi-scale

change information. Experimental results demonstrate that utilizing CSHConv leads to a modest overall improvement in network performance.

The second step is to show that the SSIF module aggregates spectral information from different bands without introducing any additional parameters. Experimental results indicate that SSIF significantly boosts the model's change detection capability, with the Recall metric increasing by nearly 4%.

Moreover, to intuitively demonstrate the roles of CSHConv and SSIF within the network, we visualized the intermediate feature activation maps from the encoder, as shown in Figure 10. As observed, when only the baseline is used, the model mistakenly focuses on many unchanged regions. With the introduction of CSHConv, the attention becomes more aligned with the actual change areas, though some discontinuities along class boundaries and missed detections remain. Finally, SSA-Net, incorporating both CSHConv and SSIF, effectively focuses on the truly changed regions in the images.

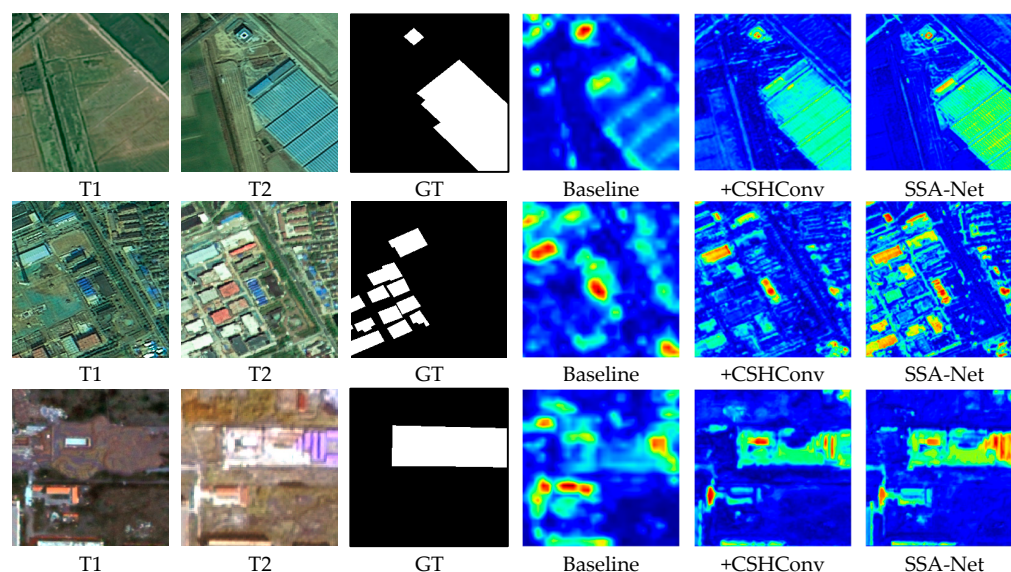


Figure 10. Feature activation maps of different models.

These ablation results demonstrate that the proposed CSHConv and SSIF modules improve various network metrics, leading to superior detection performance. The integration of both modules enhances the network's IoU and F1-scores. This indicates that, for the change detection task, these two innovative modules enable the network to more effectively extract change features from bi-temporal images, focusing on regions with potential changes and fusing multi-scale features efficiently.

5. Discussion

Parameter and computation efficiency are critical for real-world deployment. As shown in Table 4, SSA-Net achieves exceptional efficiency with only 3.54M parameters and 6.65G FLOPs—close to lightweight FC variants (1.29–1.93M Params) but surpassing them by >44% in F1. This demonstrates that our architecture eliminates redundant parameters without sacrificing accuracy.

Notably, SSA-Net outperforms all high-accuracy competitors in efficiency; it reduces parameters by 79% versus similarly accurate STANet_PAM while maintaining equivalent computation (6.65G vs. 6.58G). Compared to BIT, SSA-Net uses nearly identical parameters but cuts computation by 37% and improves F1 by 13.41%. In terms of inference time, SSA-Net also achieves competitive performance. This optimal balance establishes SSA-

Net as a practical solution for resource-constrained scenarios requiring high-precision change detection.

Table 4. Complexity comparison of different methods.

Method	Params (M)	Flops (G)	Times (S)
FC-EF	1.93	4.55	60
FC-Siam-Conc	1.75	3.99	57
FC-Siam-Diff	1.29	2.92	52
STANet	12.28	25.69	165
STANet_BAM	16.93	14.4	154
STANet_PAM	16.93	6.58	159
SNUNet	28.34	97.87	191
HANet	3.03	14.07	102
A2Net	3.78	6.02	120
BIT	3.55	10.6	244
Changer	11.39	11.89	184
ChangeFormer	20.75	11.35	80
SSA-Net	3.54	6.65	84

6. Conclusions

This paper has presented a novel system, SSA-Net, for remote sensing change detection. It leverages the CSHConv module to integrate multi-scale receptive field information through heterogeneous convolutions, thereby precisely capturing critical features of change objects across varying scales. Furthermore, the SSIF module is employed to deeply explore complex interdependencies between spatial and channel-wise spectral features in the input data. The system enables efficient global context modeling without introducing additional parameters, refining features to suppress interference from irrelevant regions and extract more accurate change information. Additionally, a case study has been carried out with data regarding China's Qinling Mountains region, including the creation of a comprehensive QL-CD change detection dataset. Experimental results demonstrate that SSA-Net achieves competitive performance in change detection. For future work, it would be interesting to prioritize lightweight network design and incorporate weakly supervised learning, thereby reducing the approach's reliance on extensive labeled samples, alleviating annotation complexity. It is also worth exploring multi-modal data fusion to enhance building change detection by leveraging diverse data sources.

Author Contributions: Conceptualization, L.F. and L.Z.; methodology, L.F. and Y.Z.; validation, L.F. and Y.Z.; writing—original draft preparation, L.F.; writing—review and editing, C.S. and Q.S.; visualization, Y.Z. and K.Z.; supervision, Y.L. and Q.S.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62271400).

Data Availability Statement: Data will be made available on request.

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

References

1. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sens.* **2022**, *14*, 871. [[CrossRef](#)]
2. Bouziani, M.; Goita, 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](#)]
3. Huang, X.; Han, X.; Ma, S.; Lin, T.; Gong, J. Monitoring ecosystem service change in the City of Shenzhen by the use of high-resolution remotely sensed imagery and deep learning. *Land Degrad. Dev.* **2019**, *30*, 1490–1501. [[CrossRef](#)]

4. Wang, J.; Yang, D.; Detto, M.; Nelson, B.W.; Chen, M.; Guan, K.; Wu, S.; Yan, Z.; Wu, J. Multi-scale integration of satellite remote sensing improves characterization of dry-season green-up in an Amazon tropical evergreen forest. *Remote Sens. Environ.* **2020**, *246*, 111865. [\[CrossRef\]](#)
5. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sens. Environ.* **2021**, *265*, 112636. [\[CrossRef\]](#)
6. Lv, Z.; Zhong, P.; Wang, W.; You, Z.; Falco, N. Multi-scale attention network guided with change gradient image for land cover change detection using remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2501805.
7. Turner, H. A comparison of some methods of slope measurement from large-scale air photos. *Photogrammetria* **1977**, *32*, 209–237. [\[CrossRef\]](#)
8. Ludeke, A.K.; Maggio, R.C.; Reid, L.M. An analysis of anthropogenic deforestation using logistic regression and GIS. *J. Environ. Manag.* **1990**, *31*, 247–259. [\[CrossRef\]](#)
9. Chen, J.; Gong, P.; He, C.; Pu, R.; Shi, P. Land-use/land-cover change detection using improved change-vector analysis. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 369–379. [\[CrossRef\]](#)
10. Bayarjargal, Y.; Karnieli, A.; Bayasgalan, M.; Khudulmur, S.; Gandush, C.; Tucker, C. A comparative study of NOAA-AVHRR derived drought indices using change vector analysis. *Remote Sens. Environ.* **2006**, *105*, 9–22. [\[CrossRef\]](#)
11. Deng, J.; Wang, K.; Deng, Y.; Qi, G. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens.* **2008**, *29*, 4823–4838. [\[CrossRef\]](#)
12. Kasetkasem, T.; Varshney, P.K. An image change detection algorithm based on Markov random field models. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1815–1823. [\[CrossRef\]](#)
13. Benedek, C.; Szirányi, T. Change detection in optical aerial images by a multilayer conditional mixed Markov model. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3416–3430. [\[CrossRef\]](#)
14. Bruzzone, L.; Prieto, D.F. An MRF approach to unsupervised change detection. In Proceedings of the 1999 International Conference on Image Processing (Cat. 99CH36348), Kobe, Japan, 24–28 October 1999; pp. 143–147.
15. Li, Y.; Zhang, H.; Xue, X.; Jiang, Y.; Shen, Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1264. [\[CrossRef\]](#)
16. Li, Y.; Zhang, H.; Shen, Q. Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.* **2017**, *9*, 67. [\[CrossRef\]](#)
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [\[CrossRef\]](#)
19. Hou, X.; Bai, Y.; Xie, Y.; Zhang, Y.; Fu, L.; Li, Y.; Shang, C.; Shen, Q. Self-supervised multimodal change detection based on difference contrast learning for remote sensing imagery. *Pattern Recognit.* **2025**, *159*, 111148. [\[CrossRef\]](#)
20. Hou, X.; Bai, Y.; Li, Y.; Shang, C.; Shen, Q. High-resolution triplet network with dynamic multiscale feature for change detection on satellite images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 103–115. [\[CrossRef\]](#)
21. 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.
22. Peng, D.; Zhang, Y.; Guan, H. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [\[CrossRef\]](#)
23. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; pp. 3–11.
24. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [\[CrossRef\]](#)
25. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 811–815. [\[CrossRef\]](#)
26. Li, K.; Li, Z.; Fang, S. Siamese NestedUNet networks for change detection of high resolution satellite image. In Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System, Xiamen, China, 27–29 October 2020; pp. 42–48.
27. Jiang, H.; Hu, X.; Li, K.; Zhang, J.; Gong, J.; Zhang, M. PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection. *Remote Sens.* **2020**, *12*, 484. [\[CrossRef\]](#)
28. 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**, 6000–6010.

29. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [\[CrossRef\]](#)
30. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. 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. [\[CrossRef\]](#)
31. Wang, H.; Fan, Y.; Wang, Z.; Jiao, L.; Schiele, B. Parameter-free spatial attention network for person re-identification. *arXiv* **2018**, arXiv:1811.12150.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
33. Song, K.; Jiang, J. AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4816–4831. [\[CrossRef\]](#)
34. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1194–1206. [\[CrossRef\]](#)
35. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8007805. [\[CrossRef\]](#)
36. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604816. [\[CrossRef\]](#)
37. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Wang, D.; Chen, X.; Jiang, M.; Du, S.; Xu, B.; Wang, J. ADS-Net: An Attention-Based deeply supervised network for remote sensing image change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *101*, 102348.
39. Zheng, H.; Gong, M.; Liu, T.; Jiang, F.; Zhan, T.; Lu, D.; Zhang, M. HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images. *Pattern Recognit.* **2022**, *129*, 108717. [\[CrossRef\]](#)
40. Ren, W.; Wang, Z.; Xia, M.; Lin, H. MFINet: Multi-scale feature interaction network for change detection of high-resolution remote sensing images. *Remote Sens.* **2024**, *16*, 1269. [\[CrossRef\]](#)
41. Yu, X.; Fan, J.; Zhang, P.; Han, L.; Zhang, D.; Sun, G. Multi-scale convolutional neural network for remote sensing image change detection. In Proceedings of the Geoinformatics in Sustainable Ecosystem and Society: 7th International Conference, GSES 2019, and First International Conference, GeoAI 2019, Guangzhou, China, 21–25 November 2019; pp. 234–242.
42. Yu, X.; Fan, J.; Chen, J.; Zhang, P.; Zhou, Y.; Han, L. NestNet: A multiscale convolutional neural network for remote sensing image change detection. *Int. J. Remote Sens.* **2021**, *42*, 4898–4921. [\[CrossRef\]](#)
43. Ren, H.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual attention-guided multiscale feature aggregation network for remote sensing image change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4899–4916. [\[CrossRef\]](#)
44. Yin, H.; Weng, L.; Li, Y.; Xia, M.; Hu, K.; Lin, H.; Qian, M. Attention-guided siamese networks for change detection in high resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103206. [\[CrossRef\]](#)
45. Ding, Q.; Shao, Z.; Huang, X.; Altan, O. DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102591. [\[CrossRef\]](#)
46. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [\[CrossRef\]](#)
47. Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; Jiang, B. S2Looking: A Satellite Side-Looking Dataset for Building Change Detection. *Remote Sens.* **2021**, *13*, 5094. [\[CrossRef\]](#)
48. Lebedev, M.; Vizilter, Y.V.; Vygolov, O.; Knyaz, V.A.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 565–571. [\[CrossRef\]](#)
49. Lei, T.; Geng, X.; Ning, H.; Lv, Z.; Gong, M.; Jin, Y.; Nandi, A.K. Ultralightweight Spatial-Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4402114. [\[CrossRef\]](#)
50. Demir, S.; Toktamış, Ö. On the adaptive Nadaraya-Watson kernel regression estimators. *Hacet. J. Math. Stat.* **2010**, *39*, 429–437.
51. Hofmann, T.; Schölkopf, B.; Smola, A.J. Kernel methods in machine learning. *Ann. Statist.* **2008**, *36*, 1171–1220. [\[CrossRef\]](#)
52. Li, Z.; Tang, C.; Liu, X.; Zhang, W.; Dou, J.; Wang, L.; Zomaya, A.Y. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602812. [\[CrossRef\]](#)
53. Ruby, U.; Yendapalli, V. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*. [\[CrossRef\]](#)
54. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
55. Han, C.; Wu, C.; Guo, H.; Hu, M.; Chen, H. HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3867–3878. [\[CrossRef\]](#)

56. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5900318. [[CrossRef](#)]
57. Fang, S.; Li, K.; Li, Z. Changer: Feature interaction is what you need for change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610111. [[CrossRef](#)]
58. 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.

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.