*remote sensing*

MDPI

*Article*

# Multistage Interaction Network for Remote Sensing Change Detection

**Meng Zhou *, Weixian Qian and Kan Ren**

School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; qianweixian@njust.edu.cn (W.Q.); k.ren@njust.edu.cn (K.R.)
* Correspondence: zhoumeng@njust.edu.cn

**Abstract:** Change detection in remote sensing imagery is vital for Earth monitoring but faces challenges such as background complexity and pseudo-changes. Effective interaction between bitemporal images is crucial for accurate change information extraction. This paper presents a multistage interaction network designed for effective change detection, incorporating interaction at the image, feature, and decision levels. At the image level, change information is directly extracted from intensity changes, mitigating potential change information loss during feature extraction. Instead of separately extracting features from bitemporal images, the feature-level interaction jointly extracts features from bitemporal images. By enhancing relevance to spatial variant information and shared semantic channels, the network excels in overcoming background complexity and pseudo-changes. The decision-level interaction combines image-level and feature-level interactions, producing multiscale feature differences for precise change prediction. Extensive experiments demonstrate the superior performance of our method compared to existing approaches, establishing it as a robust solution for remote sensing image change detection.

**Keywords:** remote sensing images; change detection; deep learning; temporal interaction



**Citation:** Zhou, M.; Qian, W.; Ren, K. Multistage Interaction Network for Remote Sensing Change Detection. *Remote Sens.* **2024**, *16*, 1077. https://doi.org/10.3390/rs16061077

Academic Editor: Gemine Vivone

Received: 3 February 2024
Revised: 28 February 2024
Accepted: 13 March 2024
Published: 19 March 2024



## 1. Introduction

People on Earth are increasingly focused on monitoring the planet due to the heightened frequency of disasters like earthquakes and floods and their effects on ongoing human activities such as construction projects and deforestation. Through the analysis of temporal remote sensing images taken at the same location, change detection (CD) emerges as a crucial tool in monitoring Earth's status, driving a wide range of applications in environmental monitoring, resource monitoring, urban planning, and disaster assessment [1,2]. Therefore, CD has attracted extensive attention in recent years.

In the initial phases of research, researchers predominantly employed conventional algorithms, encompassing algebra-based, transform-based, classification-based, and machine learning-based techniques for change detection tasks. Algebra-based methods derive the change map through algebraic operations or transformations on temporal images, such as image differencing, image regression, image rationing, and change vector analysis (CVA) [3]. Transform-based methods utilize diverse transformations to map images into another space, highlighting the change information [4,5]. These methods then employ threshold-based and clustering-based approaches to generate change maps. Classification-based methods identify changes by comparing multiple classification maps or using a trained classifier [6]. While traditional algorithms demonstrate efficacy in specific applications, their adaptability and accuracy are often restricted due to their dependence on manual features and threshold selection. Furthermore, their performance is significantly compromised when faced with variations in atmospheric conditions, seasonal factors, and differences between various sensors.

With the advancements in remote sensing technology, different platforms have become increasingly capable of collecting a wide range of data. These large-scale data enable deep

learning to model the relationship between the image contents and the real-world geographical feature as closely as possible, greatly improving the effectiveness and robustness in change detection tasks. Differentiated by their use of temporal images, the methods can be classified as early fusion and late fusion. Early fusion involves concatenating inputs and conducting feature extraction, followed by classification [7,8]. On the other hand, late fusion methods use feature extraction networks to extract features independently from dual-temporal images and compare feature differences to detect changes [9,10]. Compared with the early fusion, late fusion generally provides higher performance.

Numerous studies have embraced the Siamese network architecture, leveraging a shared feature extractor to map temporal remote sensing images into a unified space for quantifying differences [11]. Techniques such as astronus convolution [12], large-kernel convolution [13], and feature pyramid networks [14] have been incorporated to broaden the receptive field. This augmentation strengthens the network's capability of acquiring hierarchical spatial-context representations and addressing potential disruptive factors, such as season and illumination changes. Spatial attention mechanisms [15–18] and channel attention mechanisms [19–21] play a pivotal role in guiding the network to automatically focus on important information related to images/features in channels or positions while suppressing irrelevant portions that are commonly associated with backgrounds and disruptive elements. For instance, the integration of convolutional block attention modules (CBAM) in [18] facilitates the learning of spatial-wise and channel-wise discriminative features, thereby enhancing change detection. Li et al. [22] designed a supervised attention module to reweight features, enabling more effective aggregation of multilevel features from high to low levels. Self-attention is also employed to establish long-range dependencies across images and improve overall representation. Chen et al. [15] introduced a spatial–temporal attention module and a pyramid spatial–temporal attention module to capture spatial–temporal long-range dependencies and generate multi-scale attention representations, respectively. Consequently, the network exhibits increased robustness against illumination variants, demonstrating promising performance. Transformers, with self-attention as a key component, have recently shown significant improvements in change detection [23–26]. Adopting the Swin transformer as a fundamental block, Zhang et al. [24] constructed a Siamese U-shaped structure to learn multiscale features for change detection. Merging the advantages of convolutional neural networks (CNNs) and Transformers, ref. [27] extracts local–global features for enhanced change detection. Additionally, there are also some works that attempt to integrate the prior information of the changed target for enhanced performance (in a way, incorporating the edge information) [28]. In addition, leveraging the superior visual recognition capabilities of vision foundation models, Ding et al. [29] employed the visual encoder FastSAM to extract visual representations in RS scenes, achieving promising performance.

In addition to feature extraction, understanding temporal dependencies through capturing temporal interactions is crucial for generating feature differences [30–32]. Various methods, such as feature subtraction [33,34] and concatenation [35,36], are commonly employed for temporal interaction. Multiscale interaction is also recognized as beneficial, accounting for changes at different scales [37]. When treating change detection as the process of extracting change information from multi-period sequence data, recurrent neural networks (RNNs)—particularly, Long Short-Term Memory (LSTM)—have proven effective in capturing nonlinear interactions between bitemporal data. Previous studies [38–40] have utilized LSTM for acquiring change information. To address potential misinteraction, attention mechanisms have been introduced [41–43] to guide the network's focus on critical interactions. Additionally, Fang et al. [44] emphasized the importance of temporal interaction during feature extraction. Consequently, aggregation–distribution and feature exchange were introduced to enable interaction during feature extraction. Liang et al. [31] proposed patch exchange between temporal images as a means to augment change detection. Feature exchange, although effective for aligning multimodality features in fusion scenarios [45], poses challenges in bitemporal images due to their inherent content differ-

ences. It is worth noting that existing works primarily perform interaction at the feature level, often neglecting change intensity information at the image level, which could lead to the loss of crucial details.

This paper introduces a multistage interaction network (MIN-Net) to address the aforementioned issues. MIN-Net facilitates bitemporal interaction at three stages: image-level interaction, feature-level interaction, and decision-level interaction. The image-level interaction captures information from changes in image intensity through subtraction. Feature-level interaction guides the network in extracting critical spatial features related to image variants and emphasizes alignment of critical semantic channels to overcome pseudo-changes. Finally, decision-level interaction combines these stages to produce feature differences for effective change detection. The comprehensive multistage interaction in MIN-Net enhances its capacity to accurately extract changes between bitemporal images. Extensive experiments on three datasets—LEVIR-CD, WHU-CD, and CLCD—demonstrate MIN-Net's state-of-the-art performance. The contributions of this work can be summarized in three key aspects:

1. We introduce a multistage interaction network that allows our network to leverage the advantages of both early fusion and late fusion for effective change extraction;
2. We introduce the spatial and channel interactions to overcome challenges posed by background diversity and pseudo-changes;
3. Extensive experiments on LEVIR-CD, WHU-CD, and CLCD datasets showcase promising performance with F1 (we provide the definition in Section 3.1), with scores of **91.47%**, **93.73%**, and **76.60%** , respectively.

The remainder of this paper is organized as follows. Section 2 presents the details of our MIN-Net. Section 3 shows the experimental results. This paper concludes in Section 4.

## 2. The Proposed Method

This section details the introduced MIN-Net, encompassing its overall framework along with a comprehensive explanation of its components, including image-level interaction, feature-level interaction, and decision-level interaction.

### *2.1. Overall Framework*

Figure 1 illustrates the overall framework of our MIN-Net. Given bitemporal images $X_1$ and $X_2$, MIN-Net initially extracts hierarchical features $\{F_1^j, F_2^j\}_{j=1\cdots4}$ using the shared backbone ResNet-18. These extracted features are then fed into a feature pyramid network (FPN) to leverage the combined benefits of low-level and high-level representations, producing $\{P_1^j, P_2^j\}_{j=1\cdots4}$. Distinctively, we introduce a feature-level interaction module (FIM) between the two FPNs, enabling interaction at the feature extraction stage. In addition to feature-level interaction, we incorporate image-level interaction to directly extract difference information from the given images. With both image-level and feature-level interactions, we proceed to extract feature differences using the decision-level interaction, resulting in $D$.

Using $D$, the change probability for each pixel is generated through a simple multilayer perceptron (MLP):

$$p = \text{softmax}(\text{Up}(\text{Conv}_{1\times1}(D))) \tag{1}$$

Here, $\text{Up}(\cdot)$ represents an upsampling operation.

The loss function in our MIN-Net comprises two components, pixel-wise classification loss $\mathcal{L}_{BCE}$ and the dice loss $\mathcal{L}_{Dice}$, to address the sample imbalance problem. Their definitions are given by

$$\begin{aligned} \mathcal{L}_{BCE} &= -\sum_{i=1}^{N} t_i \log p_i + (1-t_i)\log(1-p_i) \\ \mathcal{L}_{Dice} &= \sum_{i=1}^{N} 1 - \frac{2p_i t_i}{p_i + t_i} \end{aligned} \tag{2}$$

where $N$ indicates the number of pixels and $i$ indexes each pixel. Here, $p_i$ represents the predicted probability value output by the network, and $t_i$ and $1 - t_i$ correspond to the ground-truth labels. We assign equal contribution to both losses, i.e., the two losses are directly summed. In the following subsections, we elaborate on the image-level, feature-level and decision-level interactions.

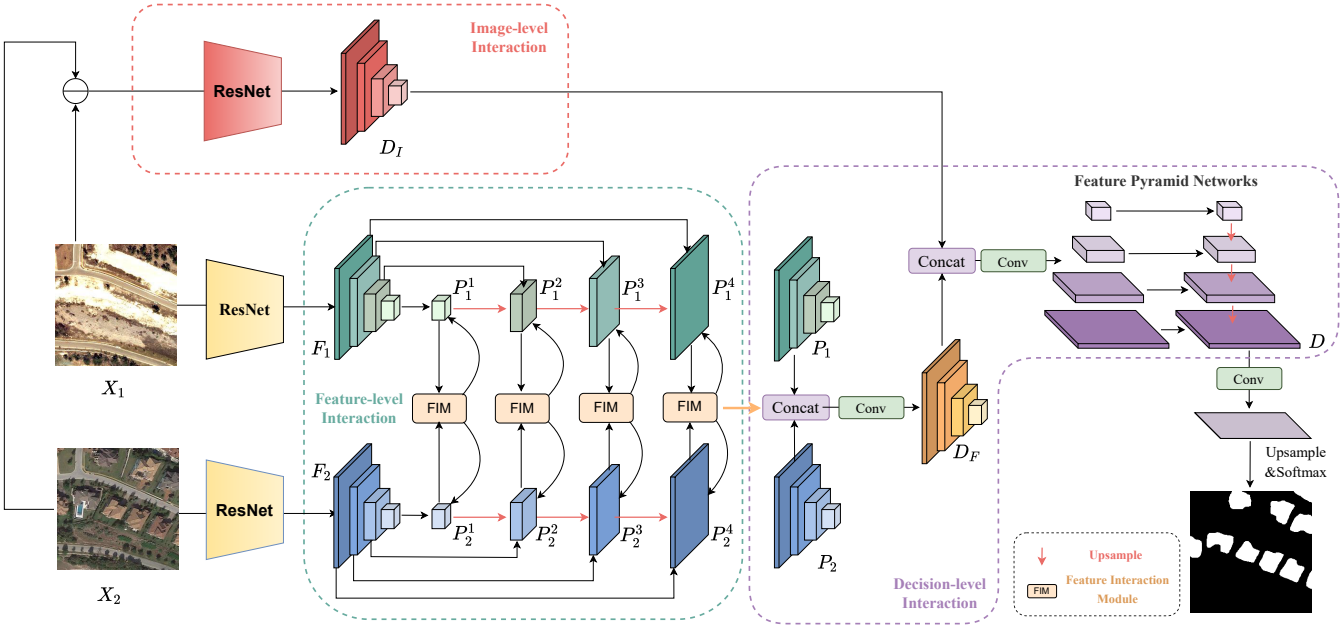



**Figure 1.** The framework of our multistage interaction network.

### *2.2. Image-Level Interaction*

The image-level interaction can be considered as an early fusion step that directly extracts change information from the given data, compensating for potential loss of change information during feature extraction. Specifically, image-level interaction initiates subtraction between $X_1$ and $X_2$, producing change intensity information. This information is then fed through the subsequent ResNet-18 backbone to extract multiscale change semantic information, resulting in $D_I^j$. Here, the index $j$ denotes the scale, ranging from the first to the fourth. In this setup, the backbone network is unshared between the feature extraction from $X_1$ and $X_2$. The primary reason for this choice is the clear information distinction between them, and we expect the network to effectively extract the change information.

### *2.3. Feature-Level Interaction*

As illustrated in Figure 2, the feature-level interaction employs a dual strategy, involving spatial interaction blocks to guide the network in jointly extracting crucial spatial features and channel interaction blocks to jointly emphasis the critical semantic channels. Change detection inherently involves extracting semantic variants from the provided temporal images. Therefore, the network should prioritize the extraction of semantic differences between bitemporal images rather than focusing on all information indiscriminately. To achieve this, we introduce the spatial interaction block to direct the network's attention towards critical spatial features related to semantic variants, preventing it from being misled by irrelevant information. Simultaneously, we acknowledge the potential variations in semantic channels arising from differences in imaging conditions and weather, possibly leading to pseudo-changes. Hence, a channel interaction block is incorporated to align images, jointly emphasizing critical semantic channels between bitemporal images.

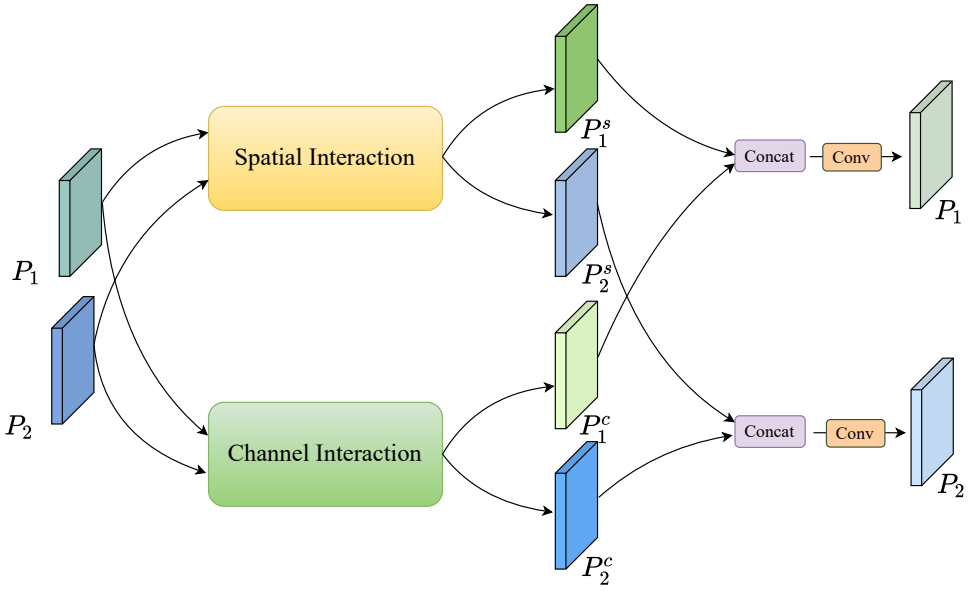



**Figure 2.** The framework of our feature-level interaction.

**Spatial Interaction Block:** As shown in Figure 3, given $P_1$ and $P_2$ with $C$ channels, the first step involves extracting their semantic differences using the following equation:

$$Q = \text{Conv}(|P_1 - P_2|) \tag{3}$$

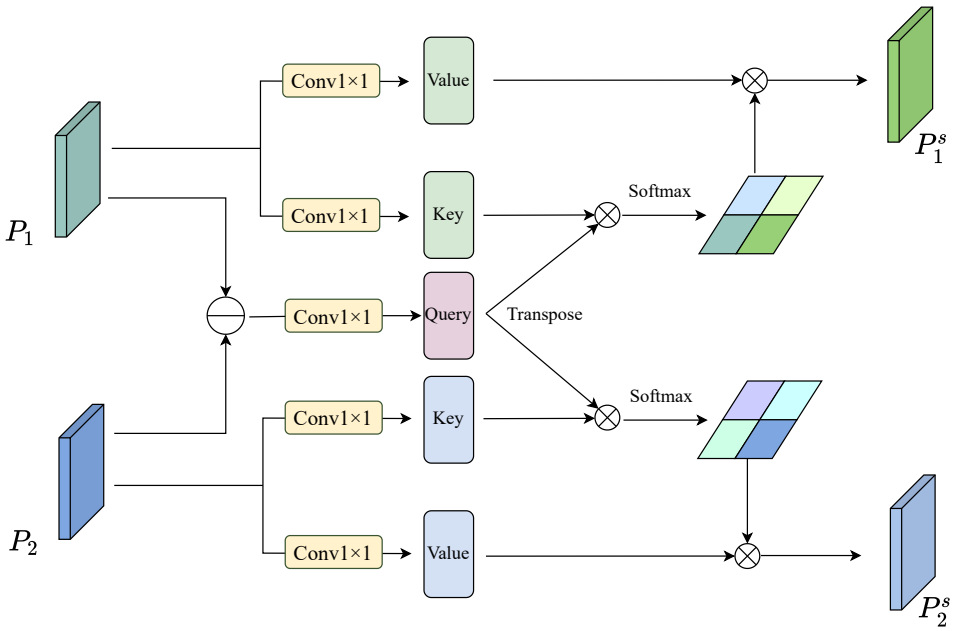



**Figure 3.** Spatial interaction block.

Using the semantic differences as the query, the features most related to these differences are obtained through cross-attention mechanism:

$$\begin{aligned} P_1^s &= \text{softmax}\left(\frac{Q\text{Conv}(P_1)}{\sqrt{C}}\right)\text{Conv}(P_1) \\ P_2^s &= \text{softmax}\left(\frac{Q\text{Conv}(P_2)}{\sqrt{C}}\right)\text{Conv}(P_2) \end{aligned} \tag{4}$$

Here, Conv represents the $1 \times 1$ convolutional operation, and softmax is applied along the channel dimension with scaling factor $\sqrt{C}$. This process aims to capture and emphasize features most relevant to the semantic differences in the images, making the semantic variations more apparent and easier to detect by the network.

**Channel Interaction Block:** As shown in Figure 4, the channel interaction block aligns the extracted features through shared channel attention. Specifically, considering $P_1$ and $P_2$ with $C$ channels, the context information is initially extracted from bitemporal images using global average pooling (GAP). The results are then concatenated, and a fully connected layer is applied to obtain the shared attention, denoted as SCA. This process can be mathematically formulated as

$$\mathrm{SCA} = \mathrm{FC}(\mathrm{Concat}(\mathrm{GAP}(P_1), \mathrm{GAP}(P_2))) \tag{5}$$

where $\mathrm{FC}(\cdot)$ is the fully connected layer. Utilizing the shared channel attention, the extracted feature maps are then calibrated as follows:

$$\begin{aligned} P_1^C &= \mathrm{SCA} \otimes P_1 \\ P_2^C &= \mathrm{SCA} \otimes P_2 \end{aligned} \tag{6}$$

where $\otimes$ is the broadcast element-wise multiplication. Consequently, the extracted features ensure a focus on the same critical semantics, proving advantageous in addressing the challenges posed by pseudo-changes.

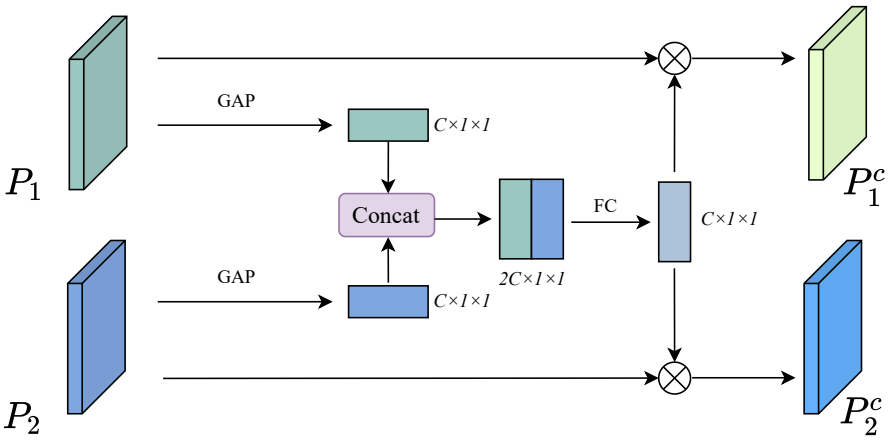



**Figure 4.** Channel interaction block.

With the extracted spatial and channel-wise interacted features, we then obtain the augmented features via

$$\begin{aligned} P_1 &= \mathrm{Conv}(\mathrm{Concat}(P_1^s, P_1^c)) \\ P_2 &= \mathrm{Conv}(\mathrm{Concat}(P_2^s, P_2^c)) \end{aligned} \tag{7}$$

### *2.4. Decision-Level Interaction*

The decision-level interaction leverages both image-level and feature-level interactions to capture difference information for subsequent change detection. Given the feature-level augmented interactions $P_1^j$ and $P_2^j$, the first step involves concatenating them along channels. The result is then passed through a convolutional layer to produce the feature differences $D_F$. Mathematically, this process is represented as

$$D_F^j = \mathrm{concat}(P_1^j, P_2^j) \tag{8}$$

Subsequently, $D_F^j$ is fused with $D_I^j$ through feature concatenation to generate multi-scale feature differences $D^j$. The resulting $D^j$ at different scales is then fed into a feature pyramid network to produce the final output $D$ for change detection.

## 3. Experimental Results

In this section, we present comprehensive experiments to demonstrate the efficacy of our proposed MIN-Net in change detection. A detailed ablation study is also provided to elucidate the effectiveness of individual modules within the network.

### *3.1. Experimental Setting*

**Datasets:** We select three datasets for evaluation, including LEVIR-CD, WHU-CD and CLCD. Their information is as follows:

- **LEVIR-CD** [15] is a large-scale dataset for building change detection, consisting of 637 pairs of high-resolution images from Google Earth. Each image is $1024 \times 1024$ pixels with a spatial resolution of 0.5 m. The dataset spans 20 different regions from 2002 to 2018. Following [15], images are segmented into non-overlapping patches of $256 \times 256$ pixels, resulting in a total of 7120/1024/2048 samples for training, validation, and testing, respectively;
- **WHU-CD** [46] is a publicly available building change detection dataset. It comprises one pair of aerial images covering the area of Christchurch, New Zealand, for the years 2012 and 2016. The image dimensions are $32,507 \times 15,354$ pixels with a spatial resolution of 0.075 m. Similar to LEVIR-CD, the dataset is divided into non-overlapping patches of $256 \times 256$ pixels. The dataset is randomly split into 6096/762/762 samples for training, validation, and testing, respectively;
- **CLCD** [47] is a dataset designed for cropland change detection, collected by Gaofen-2 in Guangdong Province, China, in 2017 and 2019. It consists of 600 pairs of cropland change samples, each with dimensions of $512 \times 512$ pixels and varying spatial resolutions from 0.5 to 2 m. Following the methodology in [47], we allocate 360 pairs for training, 120 pairs for validation, and 120 pairs for testing.

**Network Implementation:** Our network was trained using two NVIDIA 3090 GPUs, employing the AdamW optimizer. We implemented the OneCycleLR strategy for learning rate tuning, setting a maximum learning rate of 0.005 and a minimum of 0.005/500. For the LEVID-CD and WHU-CD datasets, the batch size was fixed at 32 and the learning rate was set to 0.005. In the case of the CLCD dataset, the learning rate was adjusted to 0.001 and the batch size was set to 4. All the training was conducted for 250 epochs.

**Evaluation Metrics:** The confusion matrix elements FP (false positive), FN (false negative), TP (true positive), and TN (true negative) serve as the foundation for quantitative analysis in binary change detection. These elements denote pixels that were misclassified as changed, pixels misclassified as unchanged, correctly detected changed pixels, and correctly detected unchanged pixels, respectively. Evaluation metrics, including overall accuracy (OA), precision, recall, F1 score, and Intersection over Union (IoU), are then computed using these elements:

- **OA** calculates the ratio of correctly classified pixels to the total number of pixels in the dataset, defined by

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FN+FP}} \tag{9}$$

- **Precision** measures the fraction of detections that were actually changed among all the instances predicted as changed, defined by

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{10}$$

- **Recall** measures the ability of the model to capture all the actual changes, defined by

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{11}$$

- **F1** combines recall and precision together, defined by

$$\mathrm{F1} = \frac{2}{\mathrm{R}^{-1} + \mathrm{P}^{-1}} \tag{12}$$

- **IoU** computes the overlap between the predicted and actual change regions, defined by

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP+FN+FP}} \tag{13}$$

In general, larger values indicate better prediction.

**Compared Methods:** We select 10 methods for comparison, including FC-EF [11], FC-Siam-Diff [11], FC-Siam-Conc [11], STANet [15], DTCDSCN [48], ChangeFormer [49], BIT ChangeFormer [23], ICIF-Net [30], DMINet [32], and WNet [26]. A comparison of all the methods is presented in Table 1.

**Table 1.** Comparison of different methods.

| Method | Architecture | Interaction | Params | FLOPs |
|---|---|---|---|---|
| FC-EF | CNN | Image-level | 1.351 M | 3.577 G |
| FC-Siam-Diff | CNN | Feature-level | 1.350 M | 4.727 G |
| FC-Siam-Conc | CNN | Feature-level | 1.546 M | 5.331 G |
| STANet | CNN | Feature-level | 16.892 M | 26.022 G |
| DTCDSCN | CNN | Feature-level | 31.257 M | 13.224 G |
| ChangeFormer | Transformer | Feature-level | 41.027 M | 202.788 G |
| BIT | Transformer | Feature-level | 3.496 M | 10.633 G |
| ICIF-Net | Transformer+CNN | Feature-level | 23.843 M | 25.410 G |
| DMINet | CNN | Feature-level | 6.242 M | 14.551 G |
| WNet | Transformer+CNN | Feature-level | 43.07 M | 19.20 G |
| MIN-Net (**Ours**) | CNN | Multistage | 42.12 M | 15.37 G |

By default, we utilized the original parameters as provided in the respective papers for training the comparison methods. Additionally, considering that the image input size for the CLCD dataset is 512 × 512, whereas the default batch size setting of the comparison methods is often configured for an image size of 256 × 256, we adjusted the batch size accordingly to 1/4 of the original values when training the comparison methods on the CLCD dataset. Specifically, the learning rate, batch size, and epochs for FC-EF, FC-Siam-Diff, and FC-Siam-Conc were set to 0.01, 16, and 200, respectively. For STANet, they were set to 0.001, 4, and 200; for DTCDSCN, 0.001, 16, and 200; for ChangeFormer, 0.0001, 16, and 200; for BiT, 0.01, 16, and 200; for ICIF-Net, 0.01, 8, and 200; for DMINet, 0.01, 16, and 250; and, for WNet, 0.0001, 16, and 250.

### *3.2. Comparisons with State-of-the-Art*

#### 3.2.1. Results on LEVIR-CD Dataset

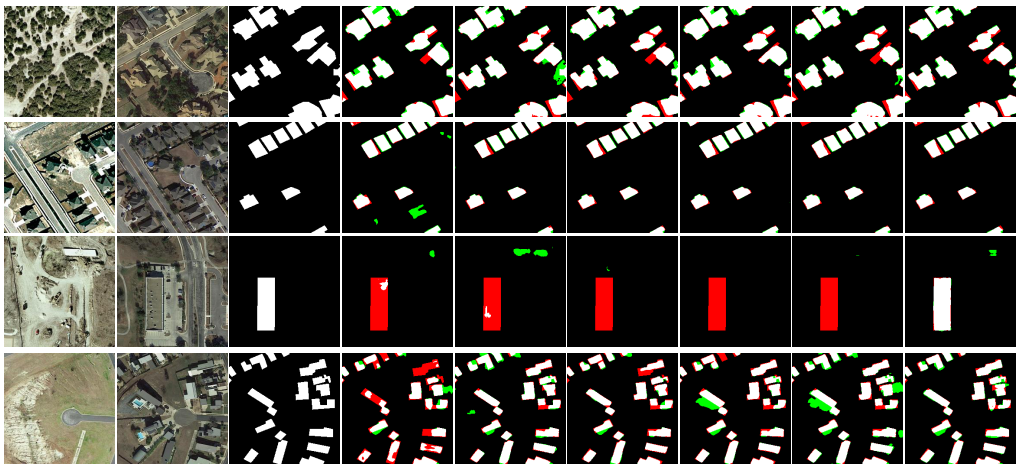
Table 2 provides a comprehensive comparison of various methods on the LEVIR-CD dataset. In general, our MIN-Net outperforms the alternative methods, particularly in terms of F1, IoU, and OA. Methods like BIT, ChangeFormer, and WNet exhibit better performance, owing to superior feature representation. The hybrid advantages of combining CNN and Transformer architecture contribute to the success of ICIF-Net, achieving more promising results. Thanks to the multistage temporal interaction, our MIN-Net demonstrates enhanced capabilities in suppressing irrelevant positions and channels, effectively addressing background complexity and pseudo-changes. As a result, MIN-Net shows

a notable improvement in performance, achieving a gain of **0.77%** over the second-best method, DMINet, with respect to the F1 score. Overall, the superior performance clearly demonstrates the effectiveness of our method in capturing temporal dependencies for enhanced change detection.

**Table 2.** Comparison of different methods on results in the LEVIR-CD dataset. (**Bold**: best; Underline: second best).

| Method | P | R | F1 | IoU | OA |
|---|---|---|---|---|---|
| FC-EF [11] | 86.91 | 80.17 | 83.40 | 71.53 | 98.39 |
| FC-Siam-Diff [11] | 89.53 | 83.31 | 86.31 | 75.92 | 98.67 |
| FC-Siam-Conc [11] | 91.99 | 76.77 | 83.69 | 71.96 | 98.49 |
| STANet [15] | 83.81 | **91.00** | 87.26 | 77.40 | 98.66 |
| DTCDSCN [48] | 88.53 | 86.83 | 87.67 | 78.05 | 98.77 |
| ChangeFormer [49] | 92.05 | 88.80 | 90.40 | 82.48 | 99.04 |
| BIT [23] | 89.24 | 89.37 | 89.31 | 80.68 | 98.92 |
| ICIF-Net [30] | 91.39 | 89.24 | 90.38 | 82.31 | 98.99 |
| DMINet [32] | **92.52** | 88.86 | 90.70 | 82.99 | 99.07 |
| WNet [26] | 91.23 | 89.62 | 90.42 | 82.51 | 99.03 |
| MIN-Net (**Ours**) | 92.04 | 90.91 | **91.47** | **84.29** | **99.14** |

We present a qualitative comparison of the LEVIR-CD dataset in Figure 5. In the figure, true positives and true negatives are denoted by white and black, while false positives and false negatives are indicated by green and red. Here, we focus on visual results of BIT, ChangeFormer, ICIF-Net, DMINet, WNet, and our MIN-Net, considering their superior performance over other methods. Overall, our method surpasses alternative methods, with fewer false positives and false negatives, providing a better match with the ground truth. This phenomenon is particularly evident in the third scene, where all other methods exhibit obvious false negatives. Augmented by the multistage interaction to address background complexity and pseudo-changes, our MIN-Net effectively extracts the actual changes.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

**Figure 5.** Visual comparison on the LEVIR-CD dataset: (**a**) T1 images, (**b**) T2 images, (**c**) Ground-truth, (**d**) BIT, (**e**) ChangeFormer, (**f**) ICIF-Net, (**g**) DMINet, (**h**) WNet, (**i**) Ours.

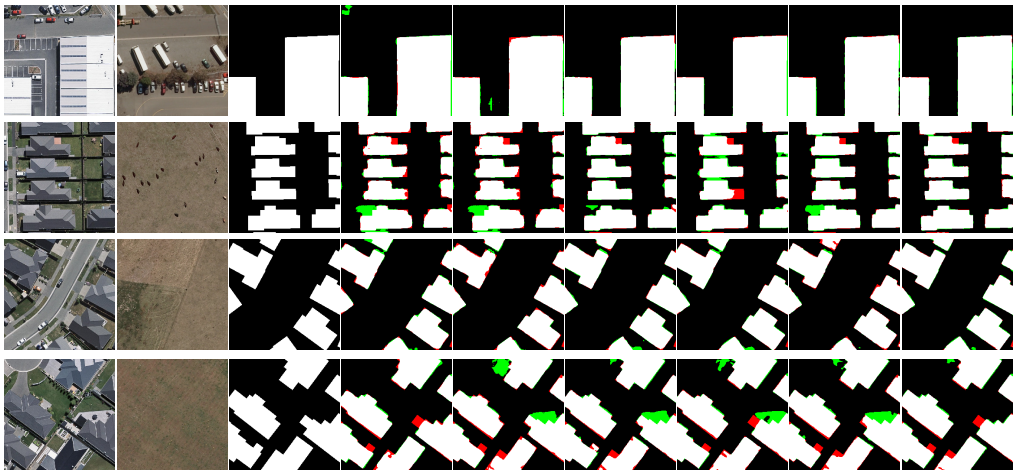
#### 3.2.2. Results on WHU-CD Dataset

We present the performance results of all methods on the WHU-CD dataset in Table 3. Notably, our method consistently outperforms other approaches, demonstrating a significant gain of **2.95%** in the F1 score over the second-best method, BIT. The visual comparison in Figure 6 highlights the effectiveness of our MIN-Net, showcasing superior performance with fewer false positives and negatives. This phenomenon underscores the

efficacy of MIN-Net in change detection, attributed to its three stages of interaction. These interactions contribute to shortening the semantic gap between bitemporal images and effectively suppressing interruptions from complex backgrounds.

**Table 3.** Comparison of different methods on the WHU-CD dataset. (**Bold**: best; Underline: second best).

| Method | P | R | F1 | IoU | OA |
|---|---|---|---|---|---|
| FC-EF [11] | 83.54 | 73.85 | 78.39 | 64.47 | 98.21 |
| FC-Siam-Diff [11] | 85.92 | 78.89 | 82.26 | 69.86 | 98.50 |
| FC-Siam-Conc [11] | 82.46 | 85.24 | 83.83 | 72.16 | 98.55 |
| STANet [15] | 85.10 | 79.40 | 82.20 | 69.70 | 98.50 |
| DTCDSCN [48] | 91.42 | 87.60 | 89.47 | 80.94 | 99.09 |
| ChangeFormer [49] | 92.06 | 83.46 | 87.55 | 77.86 | 98.96 |
| BIT [23] | 93.91 | 87.84 | 90.78 | 83.11 | 99.21 |
| ICIF-Net [30] | 91.19 | 85.92 | 88.48 | 79.34 | 99.01 |
| DMINet [32] | 82.87 | 87.54 | 85.14 | 74.12 | 98.65 |
| WNet [26] | 94.17 | 83.94 | 88.76 | 79.79 | 99.06 |
| MIN-Net (**Ours**) | **95.26** | **92.25** | **93.73** | **88.20** | **99.46** |



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

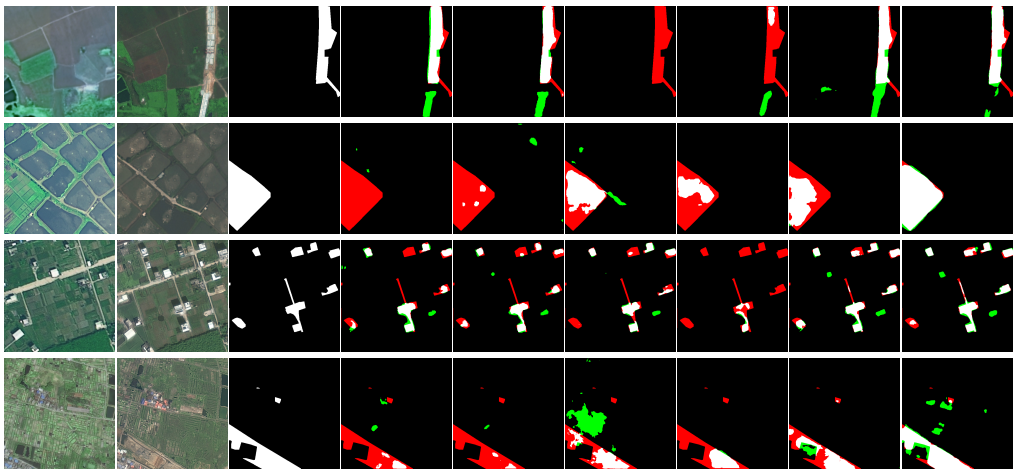
**Figure 6.** Visual comparison on the WHU-CD dataset: (**a**) T1 images, (**b**) T2 images, (**c**) Ground-truth, (**d**) BIT, (**e**) ChangeFormer, (**f**) ICIF-Net, (**g**) DMINet, (**h**) WNet, (**i**) Ours.

#### 3.2.3. Results on CLCD Dataset

The CLCD dataset presents increased complexity with multiple changes related to cropland, and the number of training samples is notably fewer than in the LEVIR-CD and WHU-CD datasets. Consequently, in Table 4, all methods exhibit a noticeable performance drop compared to results for the LEVIR-CD and WHU-CD datasets. Despite these challenges, our MIN-Net, leveraging image-level, feature-level, and decision-level interaction, demonstrates very promising performance by more accurately detecting information changes between bitemporal images. The imaging complexity is visually evident in Figure 7. Despite these challenges, our MIN-Net exhibits higher robustness, particularly noticeable in the last image with low resolution and numerous textures. The feature-level interaction plays a crucial role in enhancing change information extraction and semantic alignment, resulting in improved change boundaries.

**Table 4.** Comparison of different methods on CLCD Dataset. (**Bold**: best; Underline: second best).

| Method | P | R | F1 | IoU | OA |
|---|---|---|---|---|---|
| FC-EF [11] | 58.88 | 56.32 | 57.57 | 40.42 | 93.82 |
| FC-Siam-Diff [11] | 59.27 | 62.38 | 60.79 | 43,66 | 94.31 |
| FC-Siam-Conc [11] | 61.71 | 65.29 | 63.00 | 45.99 | 94.85 |
| STANet [15] | 55.80 | 68.00 | 61.30 | 38.40 | 93.60 |
| DTCDSCN [48] | 61.25 | 59.11 | 60.16 | 43.02 | 94.18 |
| ChangeFormer [49] | 58.29 | 47.25 | 52.19 | 35.31 | 93.56 |
| BIT [23] | 64.18 | 58.63 | 61.28 | 44.18 | 94.48 |
| ICIF-Net [30] | 66.84 | 54.02 | 58.75 | 42.60 | 94.58 |
| DMINet [32] | 70.30 | 46.40 | 55.90 | 38.79 | 94.55 |
| WNet [26] | 68.45 | 57.82 | 62.69 | 45.66 | 94.88 |
| MIN-Net (**Ours**) | **77.53** | **75.70** | **76.60** | **62.08** | **96.56** |



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

**Figure 7.** Visual comparison on the CLCD dataset: (**a**) T1 images, (**b**) T2 images, (**c**) Ground-truth, (**d**) BIT, (**e**) ChangeFormer, (**f**) ICIF-Net, (**g**) DMINet, (**h**) WNet, (**i**) Ours.

### *3.3. Ablation Study*

Here, we provide an ablation study on different modules and spatial–channel interactions to showcase their effect.

#### 3.3.1. Effectiveness of Different Modules

In this section, we conduct an ablation study, focusing on image-level and feature-level interaction. The ablation study for decision-level interaction is omitted since it is essential for producing and fusing different-level feature differences. As presented in Table 5, the inclusion of feature-level interaction proves effective in encouraging the network to extract variant information between bitemporal images, thereby reducing the semantic gap and resulting in performance improvement. Image-level interaction serves to compensate for potential loss of change information and contributes to enhanced change detection performance. The combination of image-level and feature-level interaction yields hybrid advantages, culminating in the highest performance.

**Table 5.** Ablation study on different modules.

| Dataset | Method | P | R | F1 | IoU | OA |
|---|---|---|---|---|---|---|
| LEVIR-CD | BaseLine | 90.82 | 90.80 | 90.81 | 83.16 | 99.06 |
| | w/Image | **92.17** | 90.04 | 91.09 | 83.64 | 99.10 |
| | w/Feature | 91.81 | 90.82 | 91.32 | 84.02 | 99.12 |
| | Ours | 92.04 | **90.91** | **91.47** | **84.29** | **99.14** |
| WHU-CD | BaseLine | 95.02 | 90.23 | 92.57 | 86.16 | 99.36 |
| | w/Image | 94.72 | 92.21 | 93.45 | 87.70 | 99.43 |
| | w/Feature | 95.26 | 91.64 | 93.41 | 87.64 | 99.43 |
| | Ours | **95.26** | **92.25** | **93.73** | **88.20** | **99.46** |
| CLCD | BaseLine | 75.06 | 74.45 | 74.75 | 59.69 | 96.26 |
| | w/Image | 76.20 | 74.91 | 75.54 | 60.70 | 96.39 |
| | w/Feature | 78.12 | 73.48 | 75.73 | 60.94 | 96.50 |
| | Ours | **77.52** | **75.70** | **76.60** | **62.08** | **96.56** |

#### 3.3.2. Effectiveness of Spatial and Channel Interaction Blocks

In this study, we investigate the impact of spatial and channel interaction, as presented in Table 6. The removal of spatial interaction results in the failure to guide the network in extracting change features, leading to a noticeable performance drop. Similarly, removing channel interaction prevents the network from focusing on shared semantics, resulting in decreased performance. Overall, this experiment clearly demonstrates the effectiveness of both spatial and channel interaction for robust feature extraction.

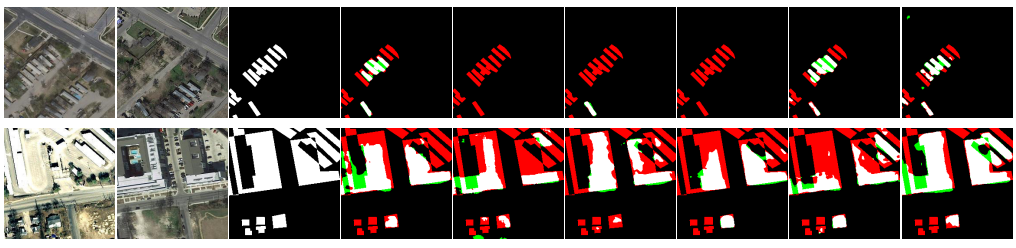
**Table 6.** The impact of the spatial and channel interaction.

| Dataset | Method | P | R | F1 | IoU | OA |
|---|---|---|---|---|---|---|
| LEVIR-CD | w/o spatial | **92.01** | 90.20 | 91.10 | 83.65 | 99.10 |
| | w/o channel | 91.60 | 90.66 | 91.13 | 83.70 | 99.10 |
| | Ours | 91.81 | **90.82** | **91.32** | **84.02** | **99.12** |
| WHU-CD | w/o spatial | 94.65 | 91.54 | 93.07 | 87.04 | 99.40 |
| | w/o channel | 94.73 | 91.15 | 92.90 | 86.75 | 99.39 |
| | Ours | **95.26** | **91.64** | **93.41** | **87.64** | **99.43** |
| CLCD | w/o spatial | 80.38 | 70.77 | 75.27 | 60.35 | **96.54** |
| | w/o channel | 78.36 | 72.77 | 75.46 | 60.59 | 96.48 |
| | Ours | **78.12** | **73.48** | **75.73** | **60.94** | 96.50 |

### *3.4. Discussion*

As evidenced by the improved performance on the WHU-CD and LEVIR-CD datasets, our method offers valuable insights into urban construction. Furthermore, the superior performance on the CLCD dataset suggests that our approach can effectively monitor cropland areas.

To provide a comprehensive assessment of our method, we present some challenging cases in Figure 8. As depicted, all methods struggle to accurately extract all changes, possibly due to buildings being obscured by trees. Incorporating global feature extraction may help overcome this limitation, which is an avenue for future research.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

**Figure 8.** Failure cases: (**a**) T1 images, (**b**) T2 images, (**c**) Ground-truth, (**d**) BIT, (**e**) ChangeFormer, (**f**) ICIF-Net, (**g**) DMINet, (**h**) WNet, (**i**) Ours.

## 4. Conclusions

In conclusion, this paper introduces a Multistage Interaction Network that has been tailored for change detection. The image-level interaction facilitates the extraction of change information from provided bitemporal images. Concurrently, the feature-level interaction directs the network to extract pertinent information from critical spatial positions associated with changes, utilizing channel interaction to consider shared semantics. Harnessing spatial and channel interaction, the decision stage adeptly extends multiscale change information, contributing to precise change detection. The presented experimental results and ablation studies robustly highlight the advantages of our method. As part of our future endeavors, we aim to explore the integration of large language models, leveraging language-level interaction to further enhance overall performance.

**Author Contributions:** All authors contributed to this manuscript: Conceptualization, M.Z. and W.Q.; methodology, M.Z.; validation, M.Z. and W.Q.; resources, K.R.; writing—original draft preparation, M.Z.; writing—review and editing, W.Q. and K.R.; supervision, W.Q. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The dataset used in our research includes LEVIR-CD, WHU-CD and CLCD. They can be publicly accessed via https://chenhao.in/LEVIR/, https://study.rsgis.whu.edu.cn/pages/download/building_dataset.html and https://github.com/liumency/CropLand-CD respectively.

**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. Jiang, H.; Peng, M.; Zhong, Y.; Xie, H.; Hao, Z.; Lin, J.; Ma, X.; Hu, X. A survey on deep learning-based change detection from high-resolution remote sensing images. *Remote Sens.* **2022**, *14*, 1552. [CrossRef]
3. Bruzzone, L.; Prieto, D. Automatic analysis of the difference image for unsupervised change detection. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 1171–1182. [CrossRef]
4. Du, B.; Wang, Y.; Wu, C.; Zhang, L. Unsupervised scene change detection via latent Dirichlet allocation and multivariate alteration detection. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 4676–4689. [CrossRef]
5. 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]
6. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sens.* **2020**, *12*, 1688. [CrossRef]
7. Jaturapitpornchai, R.; Matsuoka, M.; Kanemoto, N.; Kuzuoka, S.; Ito, R.; Nakamura, R. Newly built construction detection in SAR images using deep learning. *Remote Sens.* **2019**, *11*, 1444. [CrossRef]
8. Li, Y.; Peng, C.; Chen, Y.; Jiao, L.; Zhou, L.; Shang, R. A Deep Learning Method for Change Detection in Synthetic Aperture Radar Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5751–5763. [CrossRef]
9. Liu, J.; Gong, M.; Qin, K.; Zhang, P. A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 545–559. [CrossRef]
10. Zhang, P.; Gong, M.; Su, L.; Liu, J.; Li, Z. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *116*, 24–41. [CrossRef]
11. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
12. Wang, M.; Tan, K.; Jia, X.; Wang, X.; Chen, Y. A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images. *Remote Sens.* **2020**, *12*, 205. [CrossRef]
13. Wu, J.; Xie, C.; Zhang, Z.; Zhu, Y. A deeply supervised attentive high-resolution network for change detection in remote sensing images. *Remote Sens.* **2022**, *15*, 45. [CrossRef]
14. Xiong, F.; Li, T.; Chen, J.; Zhou, J.; Qian, Y. Mask-Guided Local–Global Attentive Network for Change Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2024**, *17*, 3366–3378. [CrossRef]
15. 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]
16. 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 Observ. Remote Sens.* **2021**, *14*, 4816–4831. [CrossRef]

17. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7296–7307. [CrossRef]
18. 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.* **2022**, *60*, 5604816. [CrossRef]
19. Liu, M.; Huang, J.; Ma, L.; Wan, L.; Guo, J.; Yao, D. A Spatial-Temporal-Channel Attention Unet++ for High Resolution Remote Sensing Image Change Detection. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 11–16 July 2021; pp. 4344–4347.
20. Eftekhari, A.; Samadzadegan, F.; Javan, F.D. Building change detection using the parallel spatial-channel attention block and edge-guided deep network. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103180. [CrossRef]
21. 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.
22. 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]
23. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607514. [CrossRef]
24. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
25. Zheng, Z.; Zhong, Y.; Tian, S.; Ma, A.; Zhang, L. ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 228–239. [CrossRef]
26. Tang, X.; Zhang, T.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. WNet: W-Shaped Hierarchical Network for Remote-Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615814. [CrossRef]
27. Li, Q.; Zhong, R.; Du, X.; Du, Y. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622519. [CrossRef]
28. Bai, B.; Fu, W.; Lu, T.; Li, S. Edge-Guided Recurrent Convolutional Neural Network for Multitemporal Remote Sensing Image Building Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5610613. [CrossRef]
29. Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Yang, K.; Bruzzone, L. Adapting Segment Anything Model for Change Detection in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611711. [CrossRef]
30. Feng, Y.; Xu, H.; Jiang, J.; Liu, H.; Zheng, J. 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*, 1–13. [CrossRef]
31. Liang, S.; Hua, Z.; Li, J. Enhanced Feature Interaction Network for Remote Sensing Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
32. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change Detection on Remote Sensing Images Using Dual-Branch Multilevel Intertemporal Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3241257. [CrossRef]
33. 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]
34. Zhang, M.; Shi, W. A feature difference convolutional neural network-based change detection method. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7232–7246. [CrossRef]
35. Xiang, S.; Wang, M.; Jiang, X.; Xie, G.; Zhang, Z.; Tang, P. Dual-task semantic change detection for remote sensing images using the generative change field module. *Remote Sens.* **2021**, *13*, 3336. [CrossRef]
36. Zheng, Z.; Wan, Y.; Zhang, Y.; Xiang, S.; Peng, D.; Zhang, B. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 247–267. [CrossRef]
37. Ma, H.; Zhao, L.; Li, B.; Niu, R.; Wang, Y. Change Detection Needs Neighborhood Interaction in Transformer. *Remote Sens.* **2023**, *15*, 5459. [CrossRef]
38. Lyu, H.; Lu, H.; Mou, L. Learning a transferable change rule from a recurrent neural network for land cover change detection. *Remote Sens.* **2016**, *8*, 506. [CrossRef]
39. Mou, L.; Bruzzone, L.; Zhu, X.X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 924–935. [CrossRef]
40. Song, A.; Choi, J.; Han, Y.; Kim, Y. Change detection in hyperspectral images using recurrent 3D fully convolutional networks. *Remote Sens.* **2018**, *10*, 1827. [CrossRef]
41. Zheng, J.; Tian, Y.; Yuan, C.; Yin, K.; Zhang, F.; Chen, F.; Chen, Q. MDESNet: Multitask Difference-Enhanced Siamese Network for Building Change Detection in High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3775. [CrossRef]
42. Zhang, L.; Hu, X.; Zhang, M.; Shu, Z.; Zhou, H. Object-level change detection with a dual correlation attention-guided detector. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 147–160. [CrossRef]
43. Cheng, G.; Wang, G.; Han, J. ISNet: Towards Improving Separability for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623811. [CrossRef]
44. 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]

45. Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; Huang, J. Deep multimodal fusion by channel exchanging. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; pp. 4835–4845.
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.* **2019**, *57*, 574–586. [CrossRef]
47. Liu, M.; Chai, Z.; Deng, H.; Liu, R. A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 4297–4306. [CrossRef]
48. 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]
49. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210.