remote sensing

MDPI

Article

A Heterogeneity-Enhancement and Homogeneity-Restraint Network (HEHRNet) for Change Detection from Very High-Resolution Remote Sensing Imagery

Biao Wang [1,†], Ao He [1,†], Chunlin Wang [2,*], Xiao Xu [3], Hui Yang [4] and Yanlan Wu [5]

[1] School of Resources and Environmental Engineering, Anhui University, Hefei 230601, China
[2] Anhui & Huaihe River Institute of Hydraulic Research, Hefei 230088, China
[3] Department of Art and Design, Jining Polytechnic, Jining 272007, China
[4] Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China
[5] School of Artificial Intelligence, Anhui University, Hefei 230601, China
[*] Correspondence: wcl@ahwrri.org.cn
[†] These authors contributed equally to this work.

Abstract: Change detection (CD), a crucial technique for observing ground-level changes over time, is a challenging research area in the remote sensing field. Deep learning methods for CD have made significant progress in remote sensing intelligent interpretation. However, with very high-resolution (VHR) satellite imagery, technical challenges such as insufficient mining of shallow-level features, complex transmission of deep-level features, and difficulties in identifying change information features have led to severe fragmentation and low completeness issues of CD targets. To reduce costs and enhance efficiency in monitoring tasks such as changes in national resources, it is crucial to promote the practical implementation of automatic change detection technology. Therefore, we propose a deep learning approach utilizing heterogeneity enhancement and homogeneity restraint for CD. In addition to comprehensively extracting multilevel features from multitemporal images, we introduce a cosine similarity-based module and a module for progressive fusion enhancement of multilevel features to enhance deep feature extraction and the change information utilization within feature associations. This ensures that the change target completeness and the independence between change targets can be further improved. Comparative experiments with six CD models on two benchmark datasets demonstrate that the proposed approach outperforms conventional CD models in various metrics, including recall (0.6868, 0.6756), precision (0.7050, 0.7570), F1 score (0.6958, 0.7140), and MIoU (0.7013, 0.7000), on the SECOND and the HRSCD datasets, respectively. According to the core principles of change detection, the proposed deep learning network effectively enhances the completeness of target vectors and the separation of individual targets in change detection with VHR remote sensing images, which has significant research and practical value.

Keywords: CNN; change detection; cosine similarity; very high-resolution image

check for updates

Citation: Wang, B.; He, A.; Wang, C.; Xu, X.; Yang, H.; Wu, Y. A Heterogeneity-Enhancement and Homogeneity-Restraint Network (HEHRNet) for Change Detection from Very High-Resolution Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 5425. https://doi.org/10.3390/rs15225425

Academic Editors: Xinzheng Zhang, Hong Huang and Ce Zhang

Received: 26 September 2023
Revised: 10 November 2023
Accepted: 13 November 2023
Published: 20 November 2023

1. Introduction

Accurate and dynamic surface spatial structure monitoring is particularly important for urbanization management [1], ecological environment monitoring [2], and emergency disaster relief [3]. Researchers are actively exploring and developing effective technical methods in this field. Furthermore, very high-resolution (VHR) remote sensing imagery has become crucial for dynamically observing the Earth, especially in urban areas, because of its capacity to provide rich information for detailed feature characterization, spatial structure analysis, and proximity relationship assessment [4]. This is significant for understanding the relationships and interactions between urban development and human activities. However, the accelerated pace of urbanization and the increasing heterogeneity in VHR images present challenges for change detection in practical applications. The observed heterogeneity in the images among the same class of targets at different times is a significant challenge in automatically interpreting change information.

In general, traditional change detection methods extract features at the pixel level [5]. Change indicators are quantified based on well-defined units to determine the target category. Although these methods are effective for change detection in low- and medium-resolution remote sensing images, they are not suitable for VHR images. As the spatial resolution of the remote sensing images increases, the dependence on neighboring units becomes more pronounced [6]. Moreover, individual objects in VHR images usually contain more pixels and show greater heterogeneity than objects in lower-resolution images, which can lead to difficulties in completely identifying objects with pixel-based change detection algorithms. Object-based change detection methods have been employed to classify spatial correspondences in VHR images and analyze segmented objects at multiple scales for image analysis rather than considering pixels at a single scale [7]. These methods mitigate the salt-and-pepper effect associated with creating classification units that represent differences and reduce false detections caused by spectral and spatial disparities [8]. Furthermore, object-based methods analyze individual target units rather than individual target pixels, which align more closely with the actual target characteristics [9,10]. However, object-based change detection methods also encounter issues, such as addressing topological relationships among different targets and capturing internal details of target objects [11].

In recent years, deep learning (DL) has become a highly representative and discriminative method known for its end-to-end, multidimensional, and multilevel feature extraction capabilities in machine learning and pattern recognition [12,13]. Based on the statistical features of the observed data, deep learning methods can automatically identify and extract complex features at various levels, enhancing image feature extraction performance. As the extraction of changing targets can be represented hierarchically through features, deep learning models have shown effective performance in extracting complex changing targets. Among deep learning techniques, the convolutional neural network (CNN) [14] has emerged as a mature and popular method. CNNs can learn high-level abstract features from raw data through multiple convolutional layers and are commonly applied in tasks such as visual recognition [15] and image classification [16]. Compared with traditional methods, CNNs use more parameters and hyperparameters that can be customized for specific tasks, resulting in increased efficiency and accuracy. Furthermore, the fully convolutional neural network (FCN) [17], in which fully connected layers are replaced with deconvolutional layers, enables pixel-level image prediction and has become a popular method in semantic segmentation tasks. Numerous FCN-like models, such as UNet [18], SegNet [19], and PSPNet [20], have been created by scholars and widely employed in applications, including land use classification [21], vehicle detection [22], and semantic segmentation [23]. Studies have demonstrated the suitability of FCNs for change detection tasks due to their ability to predict every pixel in an image. Typically, FCN-based change detection models use two consecutive images as inputs, encode multitemporal features, and fuse the features before generating change information. Depending on the timing of feature stacking within the network, these approaches can be divided into two categories: single-branch early-fusion (EF) change detection methods and two-branch late-fusion (LF) change detection methods [24]. In EF change detection methods, raw remote sensing images are directly input into the model. Deep neural networks have been employed to construct and extract features and generate change detection patches. For instance, Peng et al. [25] enhanced the UNet++ framework with a deep supervision strategy to address error accumulation issues and improve small change detection performance with complex scenes. Zheng et al. [26] introduced an end-to-end cross-layer network, CLNet, that integrated multiscale features and multilevel contextual information, mitigating side effects introduced by advanced features. However, this method does not perform feature extraction separately for two time-phase images. Currently, in change detection with bitemporal high-resolution optical remote sensing images, the prevailing approach involves a two-branch neural network with a late-fusion strategy. Here, the dual-branch approach involves two parallel subnetworks in the encoding phase, each processing one of the two input images separately. Generally, Siamese structures with shared weights have been employed to constrain feature learning and efficiently evaluate relative information [27]. For example, Daudt et al. [28] proposed FCLF-SIAM-CONC and FCLF-SIAM-DIFF based on the UNet framework, which improves detection accuracy by splicing or differentiating biphasic features. Li et al. [29] introduced a differential enhancement module based on the features extracted by a transformer and UNet to achieve precise localization of changing targets. Consequently, effectively utilizing relevant information from multitemporal images for change detection has become a prominent research focus in related fields. By utilizing a similar Siamese structure with shared weights, researchers have attempted to apply attention-based transformer models to change detection tasks [30,31]. Zhang et al. [32] adopted a U-shaped structure and processed bitemporal phase images using a pure transformer, effectively extracting features and performing change detection. Song et al. [33] combined two branches with transformers and CNNs, fused local and global information through an axial cross-attention mechanism and improved the accuracy of change detection.

However, existing DL-based methods suffer from some problems. (1) First, DL-based models generate inadequate bitemporal semantic associations. The changed regions often account for only a small portion of the data. Due to the unbalanced sample ratio, it is difficult to learn crucial features that occupy a small proportion of the data [34]. Most existing methods analyze stacked features or use balanced contrast loss functions to address data imbalance issues, with limited research on semantic correlations between bitemporal features [35]. Nevertheless, considering the semantic association between bitemporal features is crucial for effective change detection. (2) Second, DL-based models have inadequate feature fusion at different levels during the decoding process. While many works have explored the concatenation and fusion of low-level features in the encoding stage through skip connections, only a limited number of studies have addressed the incorporation of deep features in the decision-making layer [36]. However, low-level features can introduce significant noise. Therefore, fusing depth features at different stages in the upsampling process is important for change detection.

To overcome the limitations of existing change detection models and functions, this paper presents innovative solutions. The primary contributions of our work can be summarized as follows:

(1) Inspired by attention mechanisms [37], we introduce a cosine similarity (CS) module for enhancing change detection performance by emphasizing change information. Specifically, our approach involves correlating the two-phase deep features extracted by the backbone network using the CS module during the upsampling recovery process. In our approach, we emphasize capturing feature differences at various scales, enabling the construction of a deep supervised network that enhances feature variability and, consequently, improves change detection accuracy.

(2) We introduce the multilevel feature progressive fusion enhancement (MFPFE) module to refine deep semantic features. Early features are susceptible to noise, which can lead to error accumulation. Moreover, inadequate feature fusion leads to suboptimal training effects with change detection methods. In this paper, we employ a fusion method during the upsampling process. The MFPFE module effectively leverages information from the multilevel upsampling process to recover local information around target boundaries and mitigate global information loss. This approach enhances the consistency among the pixel classes by incorporating spatial information at different resolutions through progressive fusion steps.

The remainder of this study is organized as follows. Section 2 introduces the general workflow and details of the proposed method. The experimental data, accuracy evaluation methods, experimental details, results, and discussion are presented in Section 3. The conclusions are presented in Section 4.

2. Materials and Methods

2.1. Framework

In this paper, we propose a change detection model (HEHRNet). This method generates change maps based on VHR image data using an encoding–decoding structural framework, as illustrated in Figure 1. To construct multilevel deep features, we employ the HRNet network [38] during the downsampling stage. However, variations in imaging conditions between multitemporal phase images result in differences in the feature distributions of multitemporal images. To allow the network to effectively focus on genuine changes and mitigate the influence of issues related to metamerism on the extraction of change targets, we developed effective structures and methods to address these challenges. These methods primarily involve enhancing the correlations among bitemporal features, amplifying heterogeneous information that contributes to changes, and reducing attention toward similar homogeneous information. By integrating deep-level abstract features, the detected changes can be more effectively aligned with real surface data. More specifically, in the change region extraction process, we introduce the CS module to calculate the correlations among multiple time series images. Simultaneously, the MFPFE module is designed to sequentially fuse the decoded features. Finally, we employ a standard sigmoid classifier to generate the extraction results.

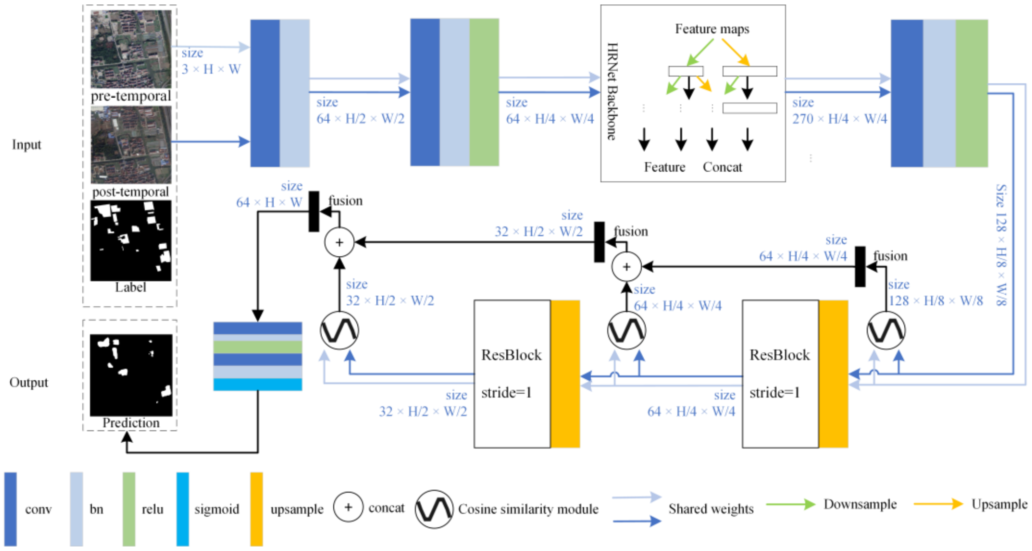

Figure 1. HEHRNet network structure.

In HEHRNet, we employ HRNet18 with multiple parallel branches as an encoder to extract multilevel features within the backbone network. To mitigate parameter redundancy and minimize the risk of overfitting during transitions, we utilize a weight-sharing concatenated network strategy to extract image features. Specifically, we input precisely georeferenced three-channel optical image pairs with dimensions (C × H × W), where C, H, and W represent the number of channels, height, and width of the original image, respectively. In the downsampling stage, we apply two convolutions with a step size of 2 to the original image. This process increases the feature dimension to 64 channels while reducing the feature size to H/4 × W/4. Subsequently, we utilize the four branches in HRNet18 to extract rich semantic information from images at various scales. The feature sizes of these four branches are H/4 × W/4, H/8 × W/8, H/16 × W/16, and H/32 × W/32. We employ a cross-fusion strategy between these branches to effectively condense information at each scale. Then, we extend the multiscale features extracted from the same image to the common dimensions of H/4 × W/4 and stack the features to integrate the multiscale information. This integration increases the feature depth and semantic expression. We then reduce the dimensionality of these stacked features and fuse them using a convolution with a step size of 2. These fused features serve as inputs in the decoding stage.

The features from the encoding stage are passed to an upsampling network with shared weights. The image size is restored, and the dimensionality is reduced using an upsampling strategy with a ratio of 2, a ResBlock module with a step size of 1, and a constant number of channels to distill the actual ground object information. The feature restoration process includes 3 stages. With this approach, the deep semantic feature information in the bitemporal phase remote sensing image is mined and utilized, and the model parameters can closely represent the actual ground objects. In addition, in the decoding stage, to enhance the inconsistencies in the changed region, the consistency of the unchanged region is constrained so that the model can better detect the changed region. A well-designed CS module is added in each stage to reconstruct the recovered features of different sizes. The use of this module for different-sized bitemporal feature metrics coincides with the number of upsampling processes, totaling three. The channel dimension of the features is reduced, and the spatial resolution is increased by the upsampling process of the network. The feature sizes are (512 × C) × H/8 × W/8, (256 × C) × H/4 × W/4, and (128 × C) × H/2 × W/2. These different feature scales contain interpretations of ground objects at different levels, which are important for accurate identification and precise localization of targets of different shapes and sizes. The multiscale features output by the CS module at different stages are fused by the MFPFE module. Finally, the bitemporal phase features are stacked and upsampled to the original image size, and the change location is detected by a simple change detection predictor head.

2.2. Cosine Similarity Module

This module is designed to highlight potential change areas in bitemporal remote sensing image pairs by grouping and calculating cosine correlations between channels at different stages in the upsampling and feature reduction processes. As illustrated in Figure 2a, the cosine similarity quantifies the semantic relationship between two entities by measuring the cosine of the angle between their respective vectors. This value is determined by the direction of the feature vectors and is independent of their magnitude. The cosine similarity metric is commonly employed in various fields, including natural language processing [39] and time series data analysis [40]. The cosine similarity formula for a multidimensional vector is

$$\cos\theta = \frac{\sum_{i=1}^{N}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{N}(A_i)^2} \times \sqrt{\sum_{i=1}^{N}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|} \tag{1}$$

where A and B are two eigenvectors, N is the dimension of the vector, and θ is the angle between the vectors.

In the BCD task, it is essential to consider not only the features from both periods but also their semantic relevance. Inspired by a study on joint multitask learning for semantic change detection [41], we adopt a two-branch upsampling structure with shared weights. We independently recover image features for each period before merging them to create bitemporal features. This approach improves the parameter expressions for the different period features without increasing parameter redundancy, which is important for the correlation calculations. In this paper, based on previous research on group convolutions [42], we group every 8 feature channels to compute the cosine similarity. The feature map is organized into groups of 8 channels, with each pixel in each group representing a high-dimensional vector composed of 8 feature values. Figure 2b illustrates the cosine distance calculation at corresponding positions in two single-layer feature maps.

The overall structure of the module is illustrated in Figure 3a. I1 and I2 represent features with identical numbers of channels, widths, and heights from two branches in the same stage during the upsampling process. In each stage, I1 and I2 are initially grouped into sets of 8 channels. The cosine similarity between the corresponding groups is then calculated, resulting in the creation of a multilayer cosine similarity feature layer that maintains the same dimensions as the original features. Following this, we apply the concept of residual networks [43] to combine the cosine similarity features with I1 and I2, enhancing the original features. This approach ensures that feature correlations are effectively leveraged from the original basis, emphasizing the expression of semantic differences within the original features across time. Finally, the merged results are passed through a convolutional layer with a kernel size of 3, producing feature maps at varying scales and stages.

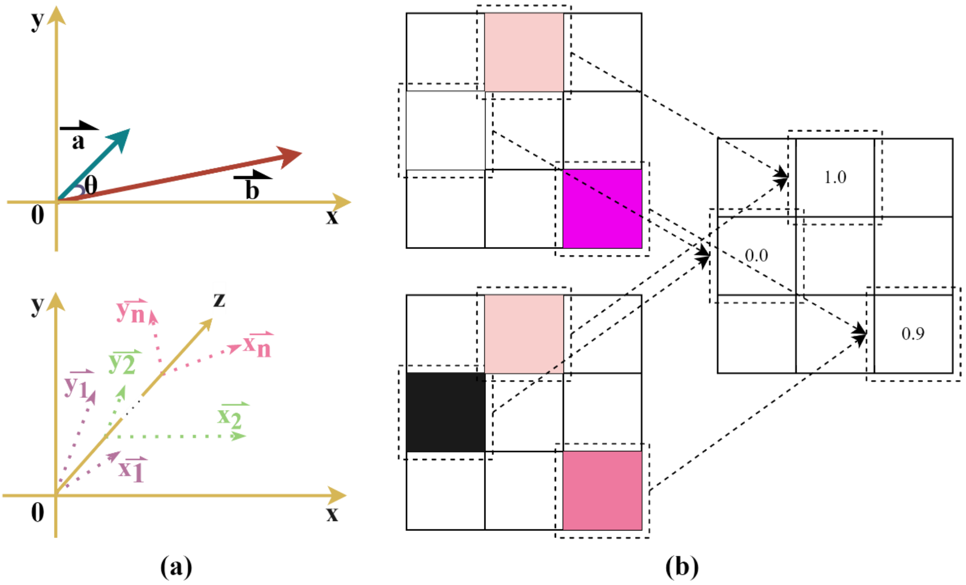

Figure 2. (**a**) Illustration of the cosine calculation. (**b**) Illustration of the cosine distance calculation for the corresponding pixel position on the feature map.

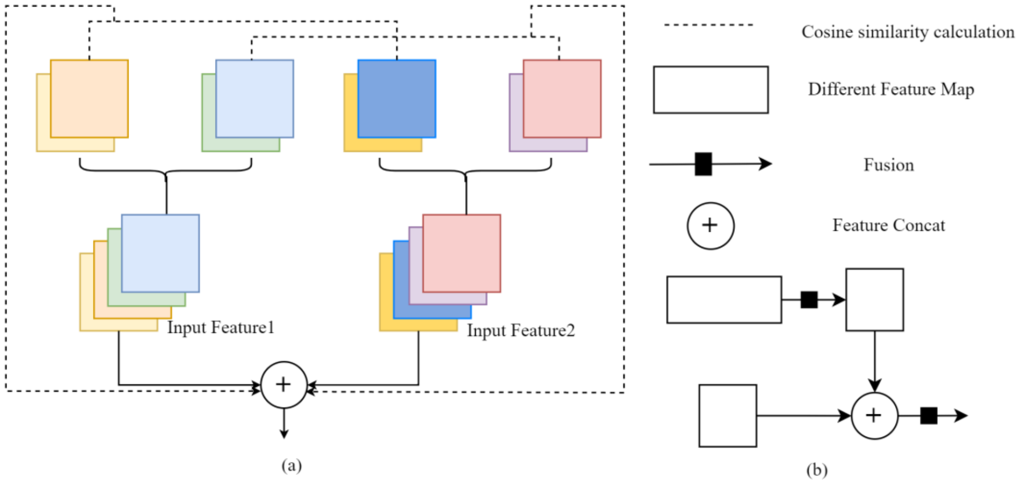

Figure 3. (**a**) Overview of the CS module. (**b**) Illustration of the two layers in the MFPFE module in the feature map block.

2.3. Multilayer Feature Progressive Fusion Enhancement Module

In each stage of the upsampling process, pixels have varying receptive fields and exhibit varying sensitivities to objects of different sizes. A larger resolution is advantageous for precisely locating small change targets, whereas a smaller resolution is beneficial for recognizing the overall change target orientation. The simplistic continuous upsampling method used in existing approaches leads to issues such as the loss of change targets and incomplete semantic information, ultimately resulting in leakage and misdetection. Hence, features at different levels must be integrated for change entities of varying sizes in the binary change detection task. To enhance the consistency in the semantic information at different scales and improve the change extraction localization accuracy, we introduce the MFPFE module. Within the decoding process, a multilayer feature fusion approach is employed for feature information from adjacent stages. This approach enhances the semantic information and mitigates the multiscale feature loss. By utilizing the features condensed by the CS module at multiple stages, we achieve greater robustness and generalization, enabling us to successfully perform the change detection task.

The structure of this module is depicted in Figure 3b. Throughout the upsampling process, we map the change detection features of adjacent stages as Fi and Fi + 1, where the dimensions (height, width) of the latter are twice those of the former. To facilitate the integration of feature representations from different stages, we first upsample Fi by a scale of 2 and apply a convolution with a kernel size of 3. Subsequently, we concatenate Fi with Fi + 1 and perform a simple convolution to fuse the features. This operation is carried out with a layer-by-layer approach across multiple feature layers, allowing us to combine global information from the low-resolution feature map with detailed information from the high-resolution feature map. As the high-level features obtained during the upsampling process include fully extracted semantic information regarding the ground features, the underlying noise is not propagated during this process. This provides valuable insights into the integrity and separability of change instance boundaries. Consequently, distinct features at different scales can be fully utilized for change detection during the upsampling process.

3. Experiments and Results

3.1. Dataset

To achieve the objectives of this study, we utilized two datasets to assess the performance and feasibility of our methodology. Figure 4 illustrates various scenarios from both datasets. In these figures, the white pixels represent the target regions that have undergone changes, while the black areas depict the background regions that remain unchanged.

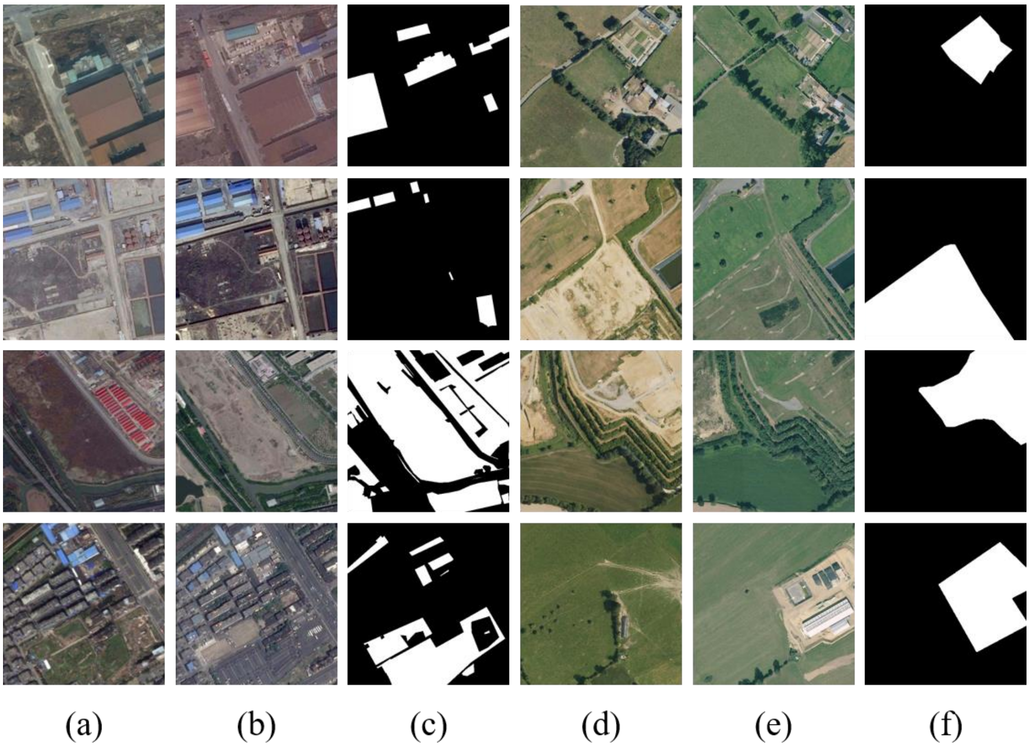

Figure 4. Sample images from the change detection datasets. (**a**–**c**) are pretemporal images, posttemporal images, and change masks for images in the SECOND dataset, respectively. (**d**–**f**) are pretemporal images, posttemporal images, and change masks for images in the HRSCD dataset, respectively.

The first dataset [44] includes data from major cities, including Hangzhou, Shanghai, and Chengdu in China. This dataset was provided by the Computational and Photogrammetric Vision team at Wuhan University and is available online as an open benchmark dataset. Additionally, the SEmantic Change detectiON Data (SECOND) dataset provides semantic labels for changed regions, classifying them into six common land cover categories: non-vegetated ground surfaces, trees, low vegetation, water, buildings, and playgrounds. These areas are typically susceptible to human disturbance in large urban areas. We adapted the SECOND dataset for our binary change detection task, preserving only the label values of 0 and 1. A label value of 0 indicates no change, while a value of 1 indicates a change. The dataset consists of 2968 tiles, each measuring 512 × 512 pixels, with ground resolutions from 0.5 m to 3 m. The tiles are composed of red–green–blue three-channel images.

The second dataset [45] was obtained in northern France, encompassing the areas around the city of Rennes in Brittany and the Caen district in Lower Normandy. The images in this dataset were acquired from 2005 to 2012, with a high spatial resolution of 0.5 m. The images include six types of feature changes: no information, artificial surfaces, agricultural areas, forests, wetlands, and water. Due to its larger coverage area, the images capture not only changes in densely populated urban regions but also complex changes in suburban and rural landscapes, including farmland. The High-Resolution Semantic Change Detection (HRSCD) dataset includes 291 tiles, each measuring 10,000 × 10,000 pixels, with mainly three-channel RGB images. To alleviate the computational demands on the hardware device, we cropped the images to a size of 512 × 512 pixels while maintaining their compatibility. Furthermore, we selected image pairs in which more than 20% of the pixels were labeled as changed, resulting in a total of 2607 tiles as samples.

In our study, we preprocessed the data to ensure that the model has an adequate amount of data for feature learning during training and sufficient samples to accurately evaluate model performance during testing. To align with the requirements of the change detection task and the dataset sizes, the SECOND dataset and HRSCD dataset were divided into training and testing sets at an 8:2 ratio. This resulted in 2374 tiles and 594 tiles allocated for training and testing in the SECOND dataset and 2086 tiles and 521 tiles for training and testing in the HRSCD dataset, respectively. Before feeding the data into the model, we applied data augmentation techniques during the training stage, primarily employing random flipping and rotating methods when loading image pairs. Additionally, to enhance the model's generalizability, we calculated the mean and standard deviation of each dataset and utilized these statistics to standardize each image before inputting the images into the model.

3.2. Experimental Settings

The proposed model and comparison methods in this study were implemented using the PyTorch-1.8.1 framework within a Python 3.7 environment. All experiments were conducted on a workstation equipped with an NVIDIA Quadro P5000 GPU. Consistent hyperparameters were applied across all experiments during the training process, including the number of training epochs (set to 60), batch size (set to 4), and initial learning rate (set to 0.1). Additionally, the stochastic gradient descent (SGD) optimizer in PyTorch was employed to optimize the model parameters. Specifically, the momentum parameter (Momentum) was set to 0.9 to expedite model convergence and ensure the model did not fall into local optima. The weight decay was set to 0.00001 to mitigate model overfitting.

3.3. Evaluation Measures

The four most common evaluation metrics were employed to evaluate the performance of our proposed method: precision (Pr), recall (completeness), F1 score (correctness), and mean intersection over union (MIoU). The recall represents the proportion of correct pixels in the detected classified pixels and is calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Here, TP and FN are the numbers of true positives and false negatives, respectively. The precision is often used to measure the proportion of true classified pixels among the detected classified pixels and is calculated as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

where FP is the number of false positives. The F1 score is a powerful evaluation metric that represents the harmonic mean of the precision and recall and is calculated as

$$\mathrm{F1} = \frac{2 \times \mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{4}$$

The MIoU is used to measure the overlap between the detected changes and labeled changes and is defined as

$$\mathrm{MIoU} = \frac{1}{\mathrm{k}} \sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP} + \mathrm{FP}} \tag{5}$$

where k and FN are the number of change categories and false negatives, respectively.

3.4. Experimental Results

The qualitative and quantitative metrics of HEHRNet based on the two datasets are presented in Figure 5 and Table 1, respectively. The detection results based on the SECOND dataset demonstrate the effectiveness of our proposed method in identifying dense and intricate urban surface changes. HEHRNet achieved precision, recall, F1, and MIoU scores of 0.7050, 0.6868, 0.6958, and 0.7013, respectively. When detecting changes in large areas with bare soil and buildings, HEHRNet successfully identified change locations and distinguished individual targets with a significant degree of separation, as shown in Figure 5(1d–3d). Additionally, HEHRNet accurately detected changes in small areas with landscapes and buildings, as depicted in Figure 5(4d,5d). These achievements are attributed to the capabilities of the MFPFE module, which effectively fuses multiscale features and discriminates between objects of varying scales and shapes. Furthermore, the CS module models feature associations differentially, enabling the identification of subtle change targets.

Table 1. The evaluation metric results of HEHRNet based on the SECOND and HRSCD datasets.

Dataset	Precision	Recall	F1 Score	MIoU
SECOND	0.7050	0.6868	0.6958	0.7013
HRSCD	0.7570	0.6756	0.7140	0.7000

The results based on the HRSCD dataset demonstrate the proficiency of the proposed model in recognizing various changes between farmland and human-created features, as well as changes between bare soil and forested land. The proposed model achieved notable precision, recall, F1, and MIoU values of 0.7570, 0.6756, 0.7140, and 0.7000, respectively. Notably, HEHRNet excels at distinguishing different objects among neighboring changing targets, as illustrated in Figure 5(4h). This indicates that the designed auxiliary structure effectively leverages the deep semantic information extracted during the encoding stage to enhance heterogeneous change information and restrict similar homogeneous information. Furthermore, our model exhibits strong detection capabilities across multiple change categories. For instance, the model accurately detects changes between roads and grass, as depicted in Figure 5(1h,2h). Additionally, the model demonstrates sensitivity to changes between buildings and grass, as observed in Figure 5(3h). This proficiency is a result of the deep feature extraction process, as the CS module analyzes semantic correlations, emphasizing the network's focus on distinctive areas. This approach ensures that comprehensive internal information about change objects can be extracted while enabling fine separation of neighboring change objects.

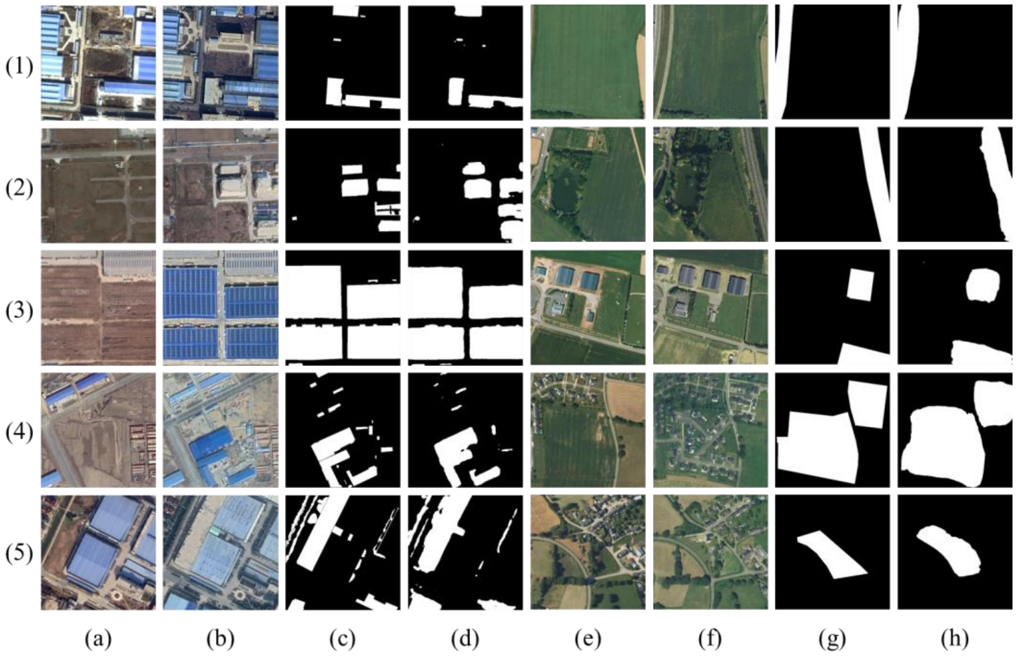

Figure 5. Visualization results of HEHRNet based on the SECOND and HRSCD datasets. (**a**,**e**) are pretemporal images in the SECOND and HRSCD datasets, respectively. (**b**,**f**) are posttemporal images in the SECOND and HRSCD datasets, respectively. (**c**,**g**) are ground-truth images in the SECOND and HRSCD datasets, respectively. (**d**,**h**) are the prediction results of HEHRNet based on the SECOND and HRSCD datasets, respectively.

Overall, the detection results based on different datasets demonstrate the good target separation and complete target extraction capabilities of HEHRNet. However, the edge regularization and fine object detection performance can be improved when compared to the ground-truth labels of actual surface changes. This limitation can be attributed to the fact that these fine objects occupy a very small portion of the VHR images and are no longer of sufficient size to be detected as distinct object-level entities following the downsampling phase of feature extraction, making them challenging to detect. In conclusion, our proposed method demonstrates robust performance across change detection datasets with various regions, types, and levels.

4. Discussion and Analysis

4.1. Comparative Experimental Results Analysis

To assess the effectiveness of our proposed deep learning method, we conducted comparisons with other typical deep learning algorithms, including a fully convolutional Siamese concatenation model (FC-Siam-Conc), a fully convolutional Siamese difference model (FC-Siam-Diff), Unet, SSCDL [24], BIT-CD [31], and ChangeFormer [46]. These typical detection algorithms have the same encoder–decoder structure as the method proposed in this paper. In the comparison experiments, we designed the UNet-based method as an early-fusion structure, utilizing stacked bitemporal phase images as the network input. UNet incorporates skip connections to capture rich information across different feature levels, enabling precise segmentation of remote sensing images for more accurate change detection. Furthermore, the FC-Siam-Conc and FC-Siam-Diff networks employ concatenated strategies to separately input images from two different periods, resulting in enhanced effectiveness. Although both models are considered late-fusion structures, their skip connections are considerably different. The former transmits feature concatenations in the two temporal images according to their corresponding upsampling stages, while the latter transmits difference maps computed based on the two temporal images. In addition, SSCDL is a novel model originally designed for multiclass change detection. It employs deep change detection units to infer semantic associations across temporal states, identifying both the locations and specific feature class changes. We retained the change detection branch of SSCDL and adapted it for binary change detection in our experiments. BIT-CD is a novel approach that integrates convolution and transformer techniques. Initially, it leverages a fully convolutional network for image feature extraction and subsequently employs a bitemporal image transformer (BIT) to capture change-related information within the images. ChangeFormer adopts a Siamese network architecture that combines transformers and multilayer perceptrons (MLPs). This model is particularly effective at capturing multiscale, long-range details with improved efficiency. We compared the performance of the proposed network (HEHRNet) with that of six conventional change detection models.

The quantitative results for the SECOND dataset are presented in Table 2. HEHRNet achieves the best results and the best overall performance when comparing the quantitative evaluation metrics with those of the other models. More specifically, compared to the UNET, FC-LF-CONC, and FC-LF-DIFF methods, which do not consider the semantic or temporal correlations of features, there are at least 1.83%, 2.89%, and 2.56% enhancements in the precision, F1 score, and MIoU metrics, respectively, with the proposed model. Compared to the SSCDL network, which considers biphasic semantic segmentation information, the precision, recall, F1 score, and MIoU are improved by 3.87%, 0.8%, 2.37%, and 2.02% with the proposed model. While the two transformer-based methods obtain good precision and MIoU values, importantly, HEHRNet still obtains the best MIoU value, surpassing the MioU values of these models by at least 2.97%. When analyzed from a network perspective, SSCDL, which optimizes the two-branch parameters using feature semantic information, performs significantly better than the other four comparison networks with feature stacking only. In addition, the two-branch late-fusion change detection methods FC-Siam-Conc and FC-Siam-Diff performed significantly better than the UNet-based early single-branch fusion model according to the performance metrics. BIT-CD and ChangeFormer outperformed the other models in terms of precision, indicating that they are less likely to classify positive samples as negative. Overall, our methods yielded better results than the comparison models. Therefore, the dual-branch structure and the deep semantic information of the bitemporal phase images can be considered beneficial for change detection information.

Table 2. Comparison between the proposed approach and the six typical methods for the SECOND and HRSCD datasets.

Dataset	Models	Precision	Recall	F1 Score	MIoU
SECOND	FC-LF-CONC	0.6867	0.6398	0.6624	0.6713
	FC-LF-DIFF	0.6505	0.6841	0.6669	0.6757
	UNET	0.6260	0.5702	0.5968	0.6308
	SSCDL	0.6663	0.6780	0.6721	0.6811
	BIT-CD	0.6999	0.6090	0.6513	0.6716
	ChangeFormer	0.6690	0.6030	0.6343	0.6578
	HEHRNet	0.7050	0.6868	0.6958	0.7013
HRSCD	FC-LF-CONC	0.7448	0.6544	0.6967	0.6861
	FC-LF-DIFF	0.6752	0.6885	0.6818	0.6662
	UNET	0.6961	0.6919	0.6940	0.6778
	SSCDL	0.7092	0.6862	0.6975	0.6822
	BIT-CD	0.7777	0.6545	0.7108	0.6998
	ChangeFormer	0.7945	0.6170	0.6946	0.6896
	HEHRNet	0.7570	0.6756	0.7140	0.7000

Figure 6 illustrates the qualitative results for several scenes from the SECOND dataset. The SECOND dataset contains mainly changes in human-created objects in large cities. Therefore, most of the change targets shown in these scenes are changes between buildings and bare soil or buildings and vegetation. The existing conventional models, which can identify the change location well, can be applied to detect multicategory changes. However, HEHRNet can extract fine-grained changes more accurately than these models without a significant increase in noise, as shown in Figure 6(1j,4j). UNet, FC-SIAM-CONC, and FC-SIAM-DIFF have many false detections in the detection results. Although SSCDL performs well in changing the contours of objects, there are still missed detections and considerable noise. BIT-CD relies on its nonlocal self-attention mechanism to enhance its ability to focus on finer details. However, it may encounter challenges when dealing with scenarios involving voids within the change target. On the other hand, ChangeFormer overlooks the differences among individual changes. However, HEHRNet accounts for the correlations between temporal sequences and recovers more accurate semantic features. In addition, HEHRNet's detection results are more complete in complex scenarios, as shown in Figure 6(2j–4j). For the changes occurring in the green vegetation on both sides of the road and between the roads, only HEHRNet extracted complete results. This may be because the proposed model utilizes the MFPFE module to consider multiple feature levels, which allows for a more complete recovery of changes in complex features. However, as observed in Figure 6(5i), only the ChangeFormer method does not misidentify changes between human-created features. This could be attributed to its robust capability to effectively capture the global context, resulting in more efficient representations of image features. As the results in Figure 6(5j,6j) show, HEHRNet maintains the separability between the changing targets. This is because HEHRNet focuses on the utilization of deep features better than the conventional algorithms. The CS module is utilized to enhance the changes in the heterogeneous regions and focus the network's attention on such changing regions. A comprehensive comparison shows that HEHRNet has better detection results than the other models based on the SECOND dataset.

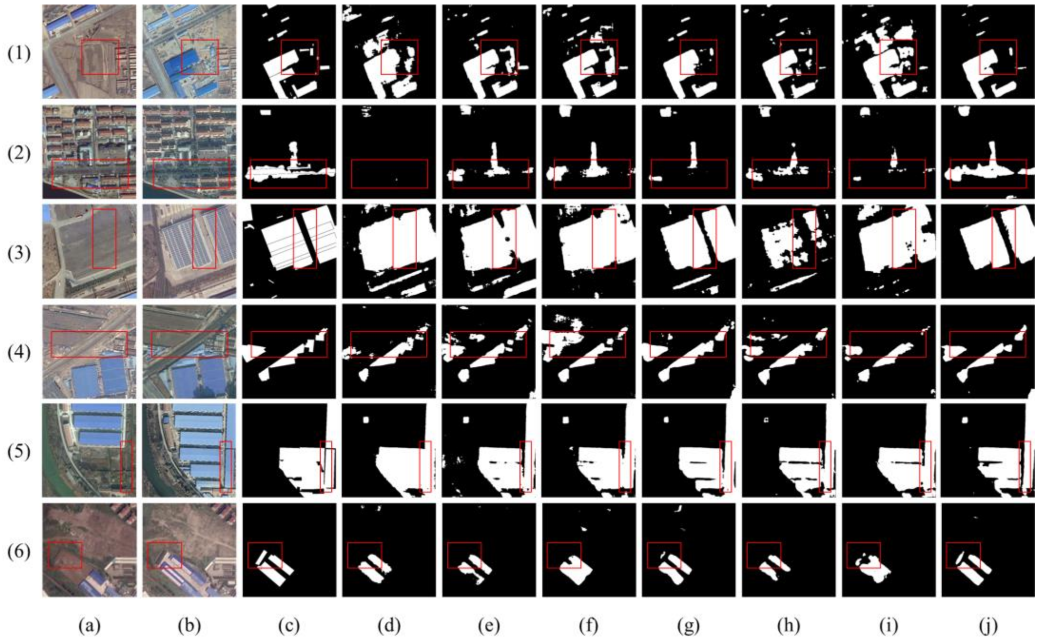

Figure 6. Visualization results of several CD methods based on the SECOND detection dataset. (**a**) Pretemporal images. (**b**) Posttemporal images. (**c**) Ground-truth images. (**d**) Results of UNet. (**e**) Results of FC-SIAM-CONC and (**f**) results of FC-SIAM-DIFF. (**g**) Results of SSCDL. (**h**) Results of BIT. (**i**) Results of ChangeFormer. (**j**) Results of the proposed model.

The quantitative results for the HRSCD dataset are presented in Table 2. Despite not achieving the best results on all metrics, HEHRNet obtained the best F1 score and MIoU. Compared to the other methods, HEHRNet achieved up to 3.22% and 3.38% increases in these two metrics. The SSCDL, which accounts for the semantic information of biphasic features, outperformed the other methods, and the two-branch late-fusion models outperformed the single-branch early-fusion models. This confirms the importance of deep semantic features and temporal associations for change detection tasks.

The qualitative comparison of the seven deep learning-based change detection methods based on the HRSCD dataset is shown in Figure 7. The change scenarios in the HRSCD dataset are complex and diverse, including changes between vegetation and bare soil and changes between human-created objects (buildings, roads) and other features. Although other methods can also identify the locations and boundaries of changes well, HEHRNet detects changes between complex ground objects better according to the above visualization results. As shown in Figure 7(1j,2j), HEHRNet obtained the most complete detection results. This is because HEHRNet considers correlation features at multiple stages and different scales. In addition, in agreement with the results based on the SECOND dataset, as shown in Figure 7(3j,5j), HEHRNet exhibits less noise and false detections in the results than the other models. Moreover, as shown in Figure 7(4j,6j), our method performs the best in terms of the separability of the results, although all the methods show irregularities in the detected boundaries. This is especially true when we compare HEHRNet with the BIT-CD and ChangeFormer methods. As illustrated in Figure 5h,i and Figure 6h,i, while both methods excel at detecting the change locations, irregular jaggedness along the edges can clearly be observed. This is because the upsampling structure of HEHRNet with shared weights in the upsampling process can optimize the semantic information for the parameters in the two periods, which also constrains the interference of similar information on the variation results to some extent. Overall, our method is equally good at fitting the HRSCD dataset with different annotation levels than the SECOND dataset.

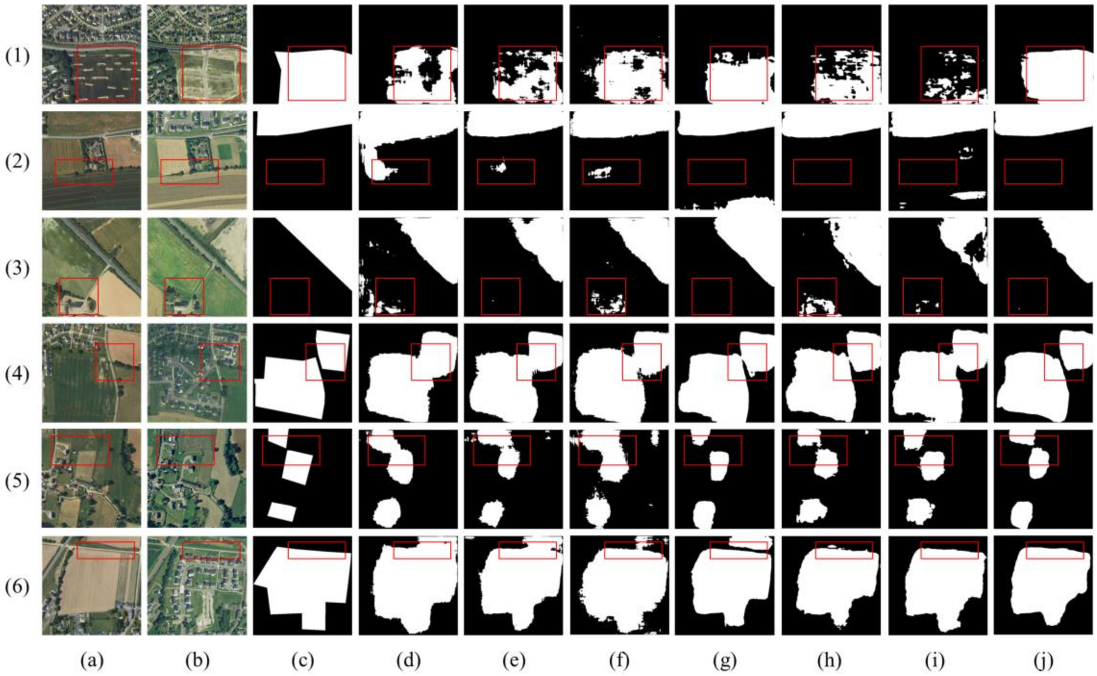

Figure 7. Visualization results of several CD methods based on the HRSCD detection dataset. (**a**) Pretemporal images. (**b**) Posttemporal images. (**c**) Ground-truth images. (**d**) Results of Unet. (**e**) Results of FC-SIAM-CONC and (**f**) results of FC-SIAM-DIFF. (**g**) Results of SSCDL. (**h**) Results of BIT. (**i**) Results of ChangeFormer. (**j**) Results of the proposed model.

Deep learning-based models have demonstrated remarkable performance in detecting changes in various features, including buildings and vegetation. Whether dealing with urban area change scenarios with the SECOND dataset or suburban changes with the HRSCD dataset, these models excel at identifying change targets. However, existing methods continue to face challenges related to the recognition of small changes and the preservation of the target area, primarily due to their inability to account for semantic associations between temporal features. In our experiments using the SECOND and HRSCD datasets, HEHRNet consistently outperforms conventional models in terms of both visual display and quantitative evaluation metrics for change detection. HEHRNet offers innovative solutions to these persistent challenges by exhibiting heightened sensitivity to changes in fine ground features and effectively mitigating issues related to "metamerism" interference. Furthermore, HEHRNet accurately identifies changing target locations and delineates smooth boundaries, significantly reducing false detections resulting from salt-and-pepper noise. These advantages can be attributed to the incorporation of the CS module in HEHRNet, which accounts for the differential relationships between changing regions in the temporal image data. Furthermore, the MFPFE module considers multilevel features, resulting in more comprehensive recognition results. Finally, we illustrate the efficiency of each model by providing details regarding the number of parameters in each model and the training time under the same configuration, as shown in Table 3. In the table, the unit "M" represents one million parameters, and "s" represents seconds. The primary distinction among the models is in the training phase. Despite its time-efficiency advantages, the transformer-based approach does not perform consistently on datasets with a limited number of samples. Because it is more difficult to focus on the local information of the image, it often requires more data for training. In contrast, our proposed method, based on homogeneity enhancement and heterogeneity constraints, is better able to extract complete change patches, which is in line with its application to the change scenarios of real projects. In practical applications, there is minimal variation in the time needed to generate change detection results when utilizing pretrained models for predicting changes in bitemporal images.

Table 3. The parameters and training time for different models.

Models	**UNet**	**FC-LF-DIFF**	**FC-LF-CONC**	**SSCDL**	**BIT-CD**	**ChangeFormer**	**Ours**
Parameters (M)	1.239	1.350	1.546	2.535	5.106	5.727	11.000
Time (s/epoch)	111	121	140	228	460	516	991

4.2. Ablation Study and Analysis

The results of extensive comparative experiments confirm the excellent performance of the proposed HEHRNet model. Furthermore, we provide experimental evidence to demonstrate that the inclusion of auxiliary modules enhances the model's performance. To assess the effectiveness and efficiency of these modules, we conducted ablation experiments using the detailed annotations available in the SECOND dataset. The quantitative results are presented in Table 4, while the qualitative results are visualized in Figure 8. We denote the basic model without the proposed auxiliary modules as 'Base'. Specifically, we directly concatenated the final feature maps from the two branches for change detection. The MIoU, recall, and F1 scores for this base model are 67.59%, 66.65%, and 67.89%, respectively, serving as the benchmark for subsequent evaluations.

Table 4. The evaluation metric results of ablation experiments. And × indicates that the module in the corresponding column is not added, and $\surd$ indicates that it is added.

Models	**CS**	**MFPFE**	**Precision**	**Recall**	**F1 Score**	**MIoU**
Base	×	×	0.6936	0.6665	0.6798	0.6759
Base + CS	√	×	0.6748	0.6852	0.6800	0.6926
Base + MFPFE	×	√	0.6817	0.6785	0.6801	0.6866
HEHRNet	√	√	0.6780	0.6958	0.6868	0.7013

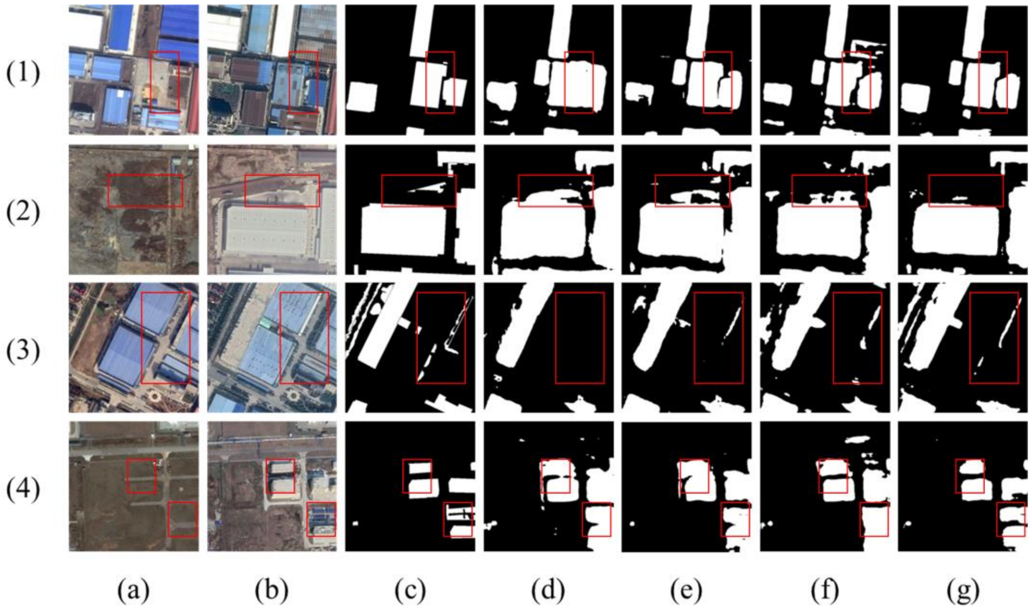

Figure 8. Visualization results of ablation experiments based on the SECOND detection dataset. (**a**) Pretemporal images. (**b**) Posttemporal images. (**c**) Ground-truth images. (**d**) Results of the benchmark model. (**e**) Results of the benchmark model with the CS module. (**f**) Results of the benchmark model with the MFPFE module. (**g**) Results of the proposed model (HEHRNet).

Subsequently, each module was incrementally incorporated into the HEHRNet structure at specific locations. Initially, we introduced the CS module in the final stage of the upsampling process to assess the effectiveness of this module. The inclusion of the CS module led to a notable increase in accuracy, with a 1.67% improvement in the MIoU and a 1.87% enhancement in the recall score. Moreover, the addition of the MFPFE module, designed to enhance the fusion of multistage features, yielded improvements of 1.07% in the MIoU and 1.2% in the recall score. These results demonstrate that the CS and MFPFE modules both enhance the network's sensitivity to changing regions. Finally, we evaluated the quantitative results of the benchmark model and HEHRNet. HEHRNet achieved the highest values in terms of the MIoU (70.13%), recall (69.58%), and F1 score (68.68%), with improvements of 2.54%, 2.93%, and 0.70% over the baseline model, respectively. These findings demonstrate that HEHRNet, incorporating all additional design elements, enhances the change recognition accuracy.

Figure 8 displays some qualitative results obtained from testing based on the SECOND dataset. Comparing these results to those of the base model, we observe gradual improvements with the introduction of the CS and MFPFE modules. These improvements are most evident in terms of the completeness and independence of the change targets. For example, in Figure 8(1d), the base model works well in the overall identification of changing targets. However, it struggles with recognizing multiple independent change targets and occasionally misidentifies nonchanging building class differences as changes. This limitation may occur due to the model's focus on the semantic difference correlations among time series features, which may lead to the model overlooking the segmentation of change targets at different scales. The addition of the CS module mitigates these misidentifications to some extent. Notably, in Figure 8(1e), the misidentification of shadows as changes among unchanged buildings no longer occurs. Similarly, the issue of recognizing two changing buildings as a single entity is alleviated to some extent. Further improvement is achieved with the inclusion of the MFPFE module. However, due to the multilevel feature approach, noise is transmitted through the different model layers, resulting in some artifacts, as shown in Figure 8(1f). This is attributed to the neglect of the semantic information reinforced by the CS module. When both auxiliary modules are combined, the optimal outcome is obtained, as shown in Figure 8(1g), with minimal noise and clear separation of different objects. Similar challenges are evident in Figure 8(2d–4d), as the models without the auxiliary modules exhibit more recognition errors and noise than the overall model, as shown in Figure 8e,f. Figure 8g shows the results when combining neighboring temporal semantic information connections and multilevel features to achieve the best results. Through these ablation experiments, we observed that the inclusion of these auxiliary modules significantly improved the change detection accuracy, as evidenced by the various evaluation metrics.

5. Conclusions

In this study, we introduced a deep learning model for the automatic extraction of comprehensive change targets within high-resolution remote sensing imagery. To enhance the model's semantic-level change recognition performance, we devised two auxiliary modules to effectively use deep nonlinear information. The CS module primarily captures temporal difference information, reinforces heterogeneous variations, suppresses homogeneous invariances, and directs the network's attention toward the change regions. Furthermore, as the network's depth increases, the MFPFE module considers more multiscale spatial information, preserving fine details. This addresses the issue of underutilized semantic information, mitigating misidentification problems in change detection tasks to a certain extent. Our experimental results, with experiments conducted based on two high-resolution datasets with resolutions ranging from 0.5 m to 2 m, unequivocally demonstrate the effectiveness of our proposed approach. While it is worth noting that some very small change targets may not be detected due to minor information loss during the feature extraction process within the deep learning framework, our method preserves the overall contours and separability of the change targets. This good performance can be attributed to the model's increased attention toward change regions. Given the challenges associated with acquiring homologous biphasic optical remote sensing images, many works have utilized heterologous data, such as optical and radar images of the same region at different times, for detection tasks. Additionally, our future work will focus on optimizing the contouring issues observed with deep learning frameworks and addressing the challenges of detecting very small targets.

Author Contributions: Methodology, B.W., A.H. and H.Y.; supervision, C.W. and Y.W.; visualization, X.X.; writing—original draft, A.H.; writing—review and editing, B.W. 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 (grant numbers 41901282, 42101381, and 41971311), the International Science and Technology Cooperation Special (grant number 202104b11020022), the Anhui Provincial Natural Science Foundation (grant number 2308085MD126), the Natural Resources Science and Technology Program of Anhui Province (grant number 2023-K-7), Shandong Province Humanities and Social Science Research Project (grant number 2023-WHLC-121), and the Hefei Municipal Natural Science Foundation (grant number 202323).

Data Availability Statement: Publicly available datasets were analyzed in this study. The datasets can be found here: SECOND: https://aistudio.baidu.com/datasetdetail/87088, accessed on 14 November 2023. And HRSCD: https://ieee-dataport.org/open-access/hrscd-high-resolution-semantic-change-detection-dataset, accessed on 14 November 2023.

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

References

1. Paul, S.; Saxena, K.G.; Nagendra, H.; Lele, N. Tracing land use and land cover change in peri-urban Delhi, India, over 1973-2017 period. *Environ. Monit. Assess.* **2021**, *193*, 52. [CrossRef]
2. Zhang, Y.J.; Wang, L.; Zhou, Q.; Tang, F.; Zhang, B.; Huang, N.; Nath, B. Continuous Change Detection and Classification-Spectral Trajectory Breakpoint Recognition for Forest Monitoring. *Land* **2022**, *11*, 504. [CrossRef]
3. Yokoya, N.; Yamanoi, K.; He, W.; Baier, G.; Adriano, B.; Miura, H.; Oishi, S. Breaking Limits of Remote Sensing by Deep Learning from Simulated Data for Flood and Debris-Flow Mapping. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4400115. [CrossRef]

4. Velumani, K.; Lopez-Lozano, R.; Madec, S.; Guo, W.; Gillet, J.; Comar, A.; Baret, F. Estimates of Maize Plant Density from UAV RGB Images Using Faster-RCNN Detection Model: Impact of the Spatial Resolution. *Plant Phenomics* **2021**, *2021*, 9824843. [CrossRef]
5. Xiao, P.F.; Zhang, X.L.; Wang, D.G.; Yuan, M.; Feng, X.Z.; Kelly, M. Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. *Isprs J. Photogramm. Remote Sens.* **2016**, *119*, 402–414. [CrossRef]
6. Xu, L.; Jing, W.P.; Song, H.B.; Chen, G.S. High-Resolution Remote Sensing Image Change Detection Combined with Pixel-Level and Object-Level. *IEEE Access* **2019**, *7*, 78909–78918. [CrossRef]
7. Zhang, L.; Hu, X.Y.; 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]
8. Wang, Z.H.; Liu, Y.L.; Ren, Y.H.; Ma, H.J. Object-Level Double Constrained Method for Land Cover Change Detection. *Sensors* **2019**, *19*, 79. [CrossRef]
9. Bansal, P.; Vaid, M.; Gupta, S. OBCD-HH: An object-based change detection approach using multi-feature non-seed-based region growing segmentation. *Multimed. Tools Appl.* **2022**, *81*, 8059–8091. [CrossRef]
10. Bai, T.; Sun, K.M.; Li, W.Z.; Li, D.R.; Chen, Y.P.; Sui, H.G. A Novel Class-Specific Object-Based Method for Urban Change Detection Using High-Resolution Remote Sensing Imagery. *Photogramm. Eng. Remote Sens.* **2021**, *87*, 249–262. [CrossRef]
11. Zhang, X.Z.; Liu, G.; Zhang, C.; Atkinson, P.M.; Tan, X.H.; Jian, X.; Zhou, X.C.; Li, Y.M. Two-Phase Object-Based Deep Learning for Multi-Temporal SAR Image Change Detection. *Remote Sens.* **2020**, *12*, 548. [CrossRef]
12. Oh, J.H.; Kim, H.G.; Lee, K.M. Developing and Evaluating Deep Learning Algorithms for Object Detection: Key Points for Achieving Superior Model Performance. *Korean J. Radiol.* **2023**, *24*, 698–714. [CrossRef]
13. Wu, M.F.; Li, C.; Yao, Z.H. Deep Active Learning for Computer Vision Tasks: Methodologies, Applications, and Challenges. *Appl. Sci.* **2022**, *12*, 8103. [CrossRef]
14. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**. [CrossRef]
15. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
17. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2014; pp. 3431–3440. [CrossRef]
18. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597. [CrossRef]
19. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2016; pp. 6230–6239. [CrossRef]
21. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S.; et al. Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci. Data* **2022**, *9*, 251. [CrossRef]
22. Berwo, M.A.; Khan, A.; Fang, Y.; Fahim, H.; Javaid, S.; Mahmood, J.; Abideen, Z.U.; Syam, M.S. Deep Learning Techniques for Vehicle Detection and Classification from Images/Videos: A Survey. *Sensors* **2023**, *23*, 4832. [CrossRef]
23. Wang, L.B.; Li, R.; Zhang, C.; Fang, S.H.; Duan, C.X.; Meng, X.L.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
24. Ding, L.; Guo, H.T.; Liu, S.C.; Mou, L.C.; Zhang, J.; Bruzzone, L. Bi-Temporal Semantic Reasoning for the Semantic Change Detection in HR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620014. [CrossRef]
25. Peng, D.; Zhang, Y.; Guan, H. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet plus. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
26. 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]
27. Chen, J.; Yuan, Z.Y.; Peng, J.; Chen, L.; Huang, H.Z.; Zhu, J.W.; Liu, Y.; Li, H.F. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1194–1206. [CrossRef]
28. Daudt, R.C.; Le Saux, B.; Boulch, A.; IEEE. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [CrossRef]
29. 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]
30. Chen, H.; Qi, Z.P.; Shi, Z.W. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
31. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]

32. Zhang, C.; Wang, L.; Cheng, S.L.; Li, Y.M. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224713. [CrossRef]
33. Song, L.; Xia, M.; Weng, L.; Lin, H.; Qian, M.; Chen, B. Axial Cross Attention Meets CNN: Bibranch Fusion Network for Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 32–43. [CrossRef]
34. Dong, J.; Zhao, W.F.; Wang, S. Multiscale Context Aggregation Network for Building Change Detection Using High Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8022605. [CrossRef]
35. Sun, Y.; Tian, Y.; Xu, Y.P. Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning. *Neurocomputing* **2019**, *330*, 297–304. [CrossRef]
36. Li, L.L.; Ma, H.B.; Jia, Z.H. Multiscale Geometric Analysis Fusion-Based Unsupervised Change Detection in Remote Sensing Images via FLICM Model. *Entropy* **2022**, *24*, 291. [CrossRef] [PubMed]
37. Shi, H.; Cao, G.; Ge, Z.X.; Zhang, Y.Q.; Fu, P. Double-Branch Network with Pyramidal Convolution and Iterative Attention for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 1403. [CrossRef]
38. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef] [PubMed]
39. Mohamed, E.H.; Shokry, E.M. QSST: A Quranic Semantic Search Tool based on word embedding. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 934–945. [CrossRef]
40. Souza, P.V.D.; Nunes, C.F.G.; Guimares, A.J.; Rezende, T.S.; Araujo, V.S.; Arajuo, V.J.S. Self-organized direction aware for regularized fuzzy neural networks. *Evol. Syst.* **2021**, *12*, 303–317. [CrossRef]
41. Zhu, Q.Q.; Guo, X.; Deng, W.H.; Shi, S.N.; Guan, Q.F.; Zhong, Y.F.; Zhang, L.P.; Li, D.R. Land-Use/Land-Cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [CrossRef]
42. Wang, X.; Kan, M.; Shan, S.; Chen, X. Fully Learnable Group Convolution for Acceleration of Deep Neural Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9041–9050. [CrossRef]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
44. Yang, K.; Xia, G.S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; Zhang, L. Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5609818. [CrossRef]
45. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Underst.* **2019**, *187*, 102783. [CrossRef]
46. Bandara, W.G.C.; Patel, V.M.; IEEE. A transformer-based siamese network for change detection. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210. [CrossRef]

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