Article

MDESNet: Multitask Difference-Enhanced Siamese Network for Building Change Detection in High-Resolution Remote Sensing Images

Jiaxiang Zheng 1,2, Yichen Tian 1, Chao Yuan 1, Kai Yin 1,*, Feifei Zhang 1, Fangmiao Chen 1 and Qiang Chen 1

1 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
2 University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: yinkai@aircas.ac.cn; Tel.: +86-010-64101439

Abstract: Building change detection is a prominent topic in remote sensing applications. Scholars have proposed a variety of fully-convolutional-network-based change detection methods for high-resolution remote sensing images, achieving impressive results on several building datasets. However, existing methods cannot solve the problem of pseudo-changes caused by factors such as “same object with different spectrums” and “different objects with same spectrums” in high-resolution remote sensing images because their networks are constructed using simple similarity measures. To increase the ability of the model to resist pseudo-changes and improve detection accuracy, we propose an improved method based on fully convolutional network, called multitask difference-enhanced Siamese network (MDESNet) for building change detection in high-resolution remote sensing images. We improved its feature extraction ability by adding semantic constraints and effectively utilized features while improving its recognition performance. Furthermore, we proposed a similarity measure combining concatenation and difference, called the feature difference enhancement (FDE) module, and designed comparative experiments to demonstrate its effectiveness in resisting pseudo-changes. Using the building change detection dataset (BCDD), we demonstrate that our method outperforms other state-of-the-art change detection methods, achieving the highest F1-score (0.9124) and OA (0.9874), indicating its advantages for high-resolution remote sensing image building change detection tasks.

Keywords: remote sensing; building change detection; Siamese network; difference-enhanced; multitask

1. Introduction

Numerous Earth observation studies have been conducted using remote sensing technology [1-5], among which building change detection is currently a research hotspot in remote sensing applications [6]. Building change detection, which uses remote sensing images, is important for urban management and supervision, urban development and planning, and city feature renewal [6,7]. With the continuous improvement of aerospace and imaging technologies, the ability to obtain high-resolution remote sensing images has been gradually improved. These images are characterized by more abundant spectral features of various ground objects and clearer shape, texture, and color features [8]. However, pseudo-changes caused by background and noise information in images might result in increased false detections. Many studies have shown that the improvement in image spatial resolution is not accompanied by an improvement in interpretation accuracy or reduction in detection difficulty [9-11]. Therefore, building change detection in high-resolution remote sensing images remains a great challenge, and improving the accuracy and intelligent interpretation is still an important issue in remote sensing building change detection.

Since the 1970s, scholars have conducted various studies on remote sensing change detection [12]. After approximately half a century of development, many change detection methods have emerged. With the improvement in the spatial resolution of remote sensing
images, automatic building change detection technology has gradually matured. In terms of granularity analysis, traditional change detection methods are mainly divided into pixel-based change detection (PBCD) and object-based change detection (OBCD) [13]. PBCD includes pre- and post-classification methods. The former analyzes the spectral characteristics of each pixel and uses methods such as feature index difference [14], change vector analysis (CVA) [15], and principal component analysis (PCA) [16], while the latter is mainly based on machine learning classification algorithms represented by support vector machines [17], random forest [18], and artificial neural networks [19] to extract buildings in different phases, and applies fuzzy matching or threshold methods to obtain change results. In 1994, Hay et al. [20] first proposed the concept of objects in remote sensing images. Since then, several studies have been conducted on object-based building change detection [21–24]. In fact, due to their own applicable scenarios and specific conditions, the generalization performance of these two types of methods is unsatisfactory [25]. In addition, the performance of change detection is limited by its weak ability to extract features and it cannot adequately satisfy the requirements of building change detection in high-resolution remote sensing images.

In recent years, deep learning has become one of the most rapidly developing branches in computer science, achieving great success in fields such as image classification, image segmentation, object detection, and natural language processing [26]. Convolutional neural networks play an important role in image analysis, and neural networks with fully convolutional architectures are widely used for building change detection. Daudt et al. [27] proposed three fully convolutional networks for remote sensing change detection: fully convolutional early fusion (FC-EF), fully convolutional Siamese concatenation (FC-Siam-conc), and fully convolutional Siamese difference (FC-Siam-diff). They showed that the feature difference method of the FC-Siam-diff network guided the network in comparing the differences between images more clearly than the feature connection method of the FC-Siam-conc network. Papadomanolaki et al. [28] believed that the above three methods failed to sufficiently learn the time-varying pattern of multitemporal remote sensing data and proposed embedding the long short-term memory (LSTM) layer into the U-shaped network to complete spatiotemporal modeling, hence achieving a 57.78% F1-score on the Onera Satellite Change Detection dataset. FC-Siam-diff-PA [29] added a pyramid attention layer based on FC-Siam-diff to deeply mine multiscale and multilevel features, and verified that this method was better than the FC-EF network model and FC-Siam-diff model. The encoders of the above networks do not apply an advanced backbone network, whereas an excellent backbone network is beneficial for improving the feature extraction ability of the model. ChangeNet [30] used a pretrained resnet50 to extract features, and the increase in network depth and width markedly improved the performance of the model. In addition to increasing the ability of feature extraction, increasing the robustness of the model to various pseudo-changes is also one of the strengths of methods in improving the accuracy of change detection. DASNet [25] combines spatial and channel attention to capture long-range dependencies and obtain more discriminative features. This method achieves F1-scores of more than 90% on the change detection dataset (CDD) and the BCDD dataset, reducing the number of detected pseudo-changes. Lee et al. [31] proposed a local similarity attention module based on cosine distance because they considered that the simple feature similarity measure had a limited ability to resist pseudo-changes. This method achieved state-of-the-art performance on CDD, BCDD, and urban change detection (UCD) datasets. The MFCN [32] applies multiscale convolution kernels to extract detailed features of ground objects, which can effectively detect detailed change information and achieve excellent results on a digital globe dataset. However, these models are all single-task change detection models that cannot efficiently utilize the bitemporal features extracted by the encoder. Liu et al. [33] argued that a multitask building change detection model with auxiliary tasks for semantic segmentation can provide object-level building semantic constraints for the change detection branch. Increasing the change detection performance can result in obtaining the building semantic segmentation features of bitemporal images,
which can then be efficiently used. However, the complex decoder structure in this model might lead to a large number of model parameters, which is not conducive to deployment in practical projects.

Although scholars have proposed many methods, the problems in building change detection have not been completely solved. The main reasons are as follows: First, when measuring the similarity between feature pairs, the existing measurement methods cannot distinguish between real changes and pseudo-changes caused by registration, imaging angle, and shadow, leading to serious false detection phenomena. Second, frequent down- and upsampling in the network causes errors in the positioning information. This results in extremely difficult restoration of the regular and neat boundaries of buildings, and a less ideal prediction effect of small targets. Finally, the number of changes in the real world is far fewer than the constant number of samples, leading to a serious imbalance in the sample distribution and increasing the difficulty of model training.

In this study, we propose a novel deep learning model, MDESNet, which provides a more effective solution for building change detection in high-resolution remote sensing images. MDESNet has an encoder–decoder structure. In the encoding stage, the shared-weight Siamese network is used as the encoder to extract the features of bitemporal inputs, while a feature pyramid network (FPN) [34] is introduced to fuse the multiscale feature maps top-down. The encoded features eventually pass through two semantic decoders and a change detection branch to obtain segmentation and change results. In contrast to the single-task change detection model, MDESNet has two semantic segmentation branches, which can provide semantic constraints for the Siamese encoder and increase the resistance of the model to pseudo-changes. More importantly, we propose a method, called the FDE module, to measure the similarity between multiscale bitemporal features. It focuses more on the real changing area in the model training process and improves the comprehensive performance of the model. To obtain the real boundary information of buildings, we propose a multiscale feature fusion method called the MSFF module. By fusing the features of adjacent scales level by level, the network learns the importance of each scale feature. In the MSFF module, we introduce the scSE [35] dual attention mechanism to recalibrate the spatial and channel dimensions of feature maps and obtain more discriminative features. Finally, we propose a weighted loss function combining binary cross-entropy loss and focal loss to alleviate the problem of imbalance between changing and unchanged samples.

The main contributions of this study are as follows:
1. We propose a multitask difference-enhanced Siamese network based on a fully convolutional structure, which consists of a main task branch for change detection and two auxiliary branches for extracting bitemporal buildings. The introduction of semantic constraints enabled the model to learn the features of targets, facilitating the avoidance of pseudo-changes. The MSFF module was designed as a decoder in three branches, and the scSE algorithm was introduced to improve the ability of the model to recover spatial details.
2. We propose an FDE module that combines concatenation and differences. This module enhanced the differences in bitemporal features at different scales, increasing the distance between pairs of real-changing pixels and thus enlarging the interclass disparity. This improved its ability to detect real changes and its robustness to pseudo-changes.
3. We verify the performance of the proposed method on BCDD and achieve the best F1-score (0.9124) compared with other baseline methods.

Our manuscript is organized as follows. Section 2 introduces our proposed method in detail. Section 3 describes the BCDD dataset and evaluation metrics, and conducts ablation and comparative experiments. Section 4 discusses and analyzes the experimental results. Finally, Section 5 summarizes our research and suggests future work.
2. Materials and Methods

In this section, we detail the proposed multitask change detection network and its application. The general structure of the proposed network is described after enumerating several related studies. Then, the FDE and MSFF modules are illustrated. Finally, the loss function is introduced.

2.1. Related Work

- **Siamese Network**

  The Siamese network [36], which was originally used to compare whether the signature on the check is consistent with the signature reserved by the bank, was proposed in 1993. The Siamese network has two inputs that are extracted by the same encoders, whose weights are shared. This dual-stream structure is beneficial for measuring the similarity between the two inputs. It has been widely used in signature verification [36], face verification [37], change detection [25,30,31], and semantic similarity analysis [38,39].

- **ResNeSt**

  ResNeSt [40] is a variant of ResNet [41], which introduces the split-attention module of ResNet. It employs multibranch convolution in the same layer to extract more discriminative features, whereas application of the channel-attention algorithm highlights the importance of different feature channels and assigns them different weights. Compared with ResNet, ResNeXt [42], and SENet [43], ResNeSt performs better in image classification, semantic segmentation, instance segmentation, and object detection tasks [40].

- **FPN**

  The FPN structure was originally designed for object detection tasks [34] and was later applied to a panoptic semantic segmentation task [44] by He et al. When there are objects of various sizes in the image, an FPN ensures that these objects are represented on feature maps of different scales. Similar to UNet, the FPN network applies skip connections to fuse the features extracted by the encoder in the top-down process. Unlike UNet, the original FPN upsamples the low-resolution feature maps by a factor of 2, and after the features introduced by the skip connections pass through a 1 × 1 convolutional layer, they are finally added together. We used this FPN in our study, which can fuse multiscale information without increasing the number of parameters.

- **scSE**

  The concurrent spatial and channel squeeze and channel excitation (scSE) module [35], which is a synthesis of channel attention and spatial attention, is similar to a plugin. The scSE module compresses and activates features in the channel and spatial dimensions, respectively, and can explicitly model the interdependence between feature map channels. After using the scSE module, the feature maps undergo spatial and channel recalibration, which increases the ability of the model to learn more discriminative features; more importantly, it only adds a few parameters.

2.2. Network Architecture

Figure 1 shows the network structure of the proposed MDESNet. The input of the model was bitemporal remote sensing images of the same area. The bitemporal multiscale features were obtained after passing through the Siamese encoder composed of ResNeSt50 and the neck layer composed of FPN. The decoder, composed of FDE and MSFF modules, decoded the bitemporal multiscale features. Finally, the prediction results of semantic segmentation and change detection of buildings in bitemporal images were obtained.
In the encoding stage, we extracted bitemporal image features using a Siamese network that shares weight. We used ResNeSt-50 as the backbone of the model to enable it to extract more discriminative features. ResNeSt, which introduces a split-attention module, realizes the weight allocation of feature channels through group convolution and channel-wise attention. Therefore, it extracts more diversified features, achieving remarkable performance in image classification, object detection, and semantic segmentation [40]. The feature maps obtained from the Siamese network were divided into four different spatial scales with resolutions of 1/4, 1/8, 1/16, and 1/32 of the input image, and numbers of channels of 256, 512, 1024, and 2048, respectively. To fuse the context information, multiscale feature maps were inputted into the FPN structure, reducing the number of channels to 256. The top-down summation operation and convolution layer facilitated multiscale information aggregation and also reduced the aliasing effect that results from the superimposition of feature maps of different scales. In addition, reducing the number of channels further reduced the amount of computation and memory usage in the decoding stage.

In the decoding stage, we used two branches for semantic segmentation and one branch for change detection. For the two semantic segmentation branches, the bitemporal multiscale feature maps were decoded using the MSFF module, and the semantic prediction results of bitemporal buildings were obtained after four rounds of upsampling. For the change detection branch, the bitemporal feature maps of the corresponding scales were inputted into the FDE module to finally obtain the multiscale difference feature maps, which were decoded using the MSFF module and then upsampled four times to obtain the prediction results of the changed buildings. The proposed FDE and MSFF modules are shown in Figures 2 and 3 and are introduced in detail in Sections 2.2 and 2.3, respectively.
modules are shown in Figures 2 and 3 and are introduced in detail in Section 2.2 and Section 2.3, respectively.

Figure 2. Overview of the FDE module. The proposed FDE module has difference and concatenation branches. The former applies the sigmoid activation function to obtain the feature difference attention map after making a difference on bitemporal feature maps, whereas the latter element-wise multiplies the attention map and the concatenated features to enhance the difference. The feature maps of the four scales are operated in the same way.

Figure 3. A diagram of the MSFF module. Its basic unit is FusionBlock composed of residual block and the scSE module. The feature maps of adjacent scales were successively passed through FusionBlock to output high-resolution fused feature maps. The MSFF module contained six FusionBlocks, and finally applied a 1 × 1 convolutional layer and softmax function to obtain the classification results.
2.3. Feature Difference Enhancement Module

A suitable similarity measure facilitates the extraction of more discriminative features from the network, which is very meaningful for increasing the robustness of the model and reducing the impact of pseudo-changes. Daudt et al. [27] proposed a difference or concatenation method to derive changed information from bitemporal feature maps. However, although the difference method more clearly indicates the change area, it might contain more erroneous information than the concatenation method when bitemporal images have problems such as inconsistent imaging perspectives and large registration errors. Conversely, despite containing the complete features of bitemporal images, the concatenation method cannot clearly specify the changed information as efficiently as the difference method. As shown in Figure 2, we propose the use of an FDE module to integrate these two methods.

The bitemporal multiscale feature maps were inputted in the FDE module, and went through two branches on the same scale: difference and concatenation. Initially, the distance of the bitemporal feature maps was calculated in the difference branch, and we chose the element-wise subtract because it consumed less time and space complexity than cosine distance. Thus, the feature difference maps were obtained as follows:

\[
\text{FDM}(i) = \left| F_{i1} - F_{i0} \right|, \quad i = 1, \ldots, N
\]  

where FDM denotes the feature difference maps, \( F_{i1} \) is the feature map with T1 as the input, \( F_{i0} \) is the feature map with T0 as the input, and \( N \) is the number of channels in the input feature maps.

Subsequently, we applied a \( 1 \times 1 \) convolutional layer to compress the channels of the feature difference maps from 256 dimensions to 1 dimension and obtained the feature difference attention maps after activating them through the sigmoid function. The feature difference attention maps of the four scales were obtained as follows:

\[
\text{FDAM}(j) = \frac{1}{1 + \exp(-\eta(\text{FDM}(j)))}, \quad j = 1, 2, 3, 4
\]  

where FDAM represents the feature difference attention map, \( \eta \) is the \( 1 \times 1 \) convolutional layer, and \( j \) represents the four scales.

In the concatenation branch, we first passed the concatenated feature maps through a \( 3 \times 3 \) convolutional layer followed by batch normalization and the ReLU function, which reduced the number of channels from 512 to 256 dimensions and prepared the feature difference enhancement. Subsequently, the concatenated feature maps were multiplied element-wise with the feature difference attention map, which represents the fusion of the differential and concatenated information. Finally, \( 3 \times 3 \) and \( 1 \times 1 \) convolutional layers were used to obtain the changed features of this scale. The difference indicated the features of the changed area, with the concatenation calibrating the error in the difference information. The FDE module combined the advantages of the two modules to enhance the feature difference. Utilizing the FDE module at four scales allowed the network to focus on regions with semantic and spatial differences.

2.4. Multiscale Feature Fusion Module

The MSFF module was designed to recover the spatial resolution of semantic feature maps in multiscale feature maps. The detailed architecture of the MSFF module is illustrated in Figure 3.

Regarding feature maps at four scales, their channels were 256-dimensional for both the semantic segmentation and change detection branches. Feature maps of adjacent scales were inputted into FusionBlock for fusion. FusionBlock contained two residual blocks and one scSE module. The former is a layer composed of two \( 3 \times 3 \) convolutional layers, followed by batch normalization and ReLU, which reduced the channels of the
input feature maps to 128. The scSE module is usually applied to learn the importance of each feature channel; hence, the feature maps undergo spatial and channel recalibration, resulting in the enhancement of more meaningful feature maps by the network. Moreover, the addition of the scSE module did not markedly increase the parameters of the model. All feature maps of adjacent scales went through the residual block, and the low-resolution feature maps were concatenated with high-resolution after upsampling. The output of FusionBlock was obtained through the scSE module after the channel was restored to 256 dimensions. Therefore, we applied six FusionBlocks to restore 1/4 of the size of the input image, and finally used a 1 × 1 convolutional layer and softmax function to obtain the classification results.

The proposed MSFF module played the role of decoding multiscale features. It did not utilize skip connections to introduce the features of the encoder because the multiscale feature maps were fused once by the FPN structure. The difference from the lightweight decoder proposed in [44] and [45] is that the MSFF module learns the importance of feature maps at different scales owing to the cascade structure and the scSE module.

2.5. Loss Function

Our network is a multitask deep learning model for extracting pixel-level semantic and changed information. The cross-entropy loss function is usually used for training pixel-level tasks [46]. However, in both semantic segmentation and change detection tasks, the sample quantities of positive and negative categories are often unbalanced [47]. Consequently, training with only the cross-entropy loss function typically results in a large number of misjudgments. To solve this problem, we used a focus loss function to alleviate the imbalance and classification difficulty between positive and negative sample sizes [48]. The binary cross-entropy loss function $L_{BCE}$ was given as follows:

$$L_{BCE} = -t_i \log(y_i) - (1 - t_i) \log(1 - y_i) = \begin{cases} -\log(y_i), & t_i = 1 \\ -\log(1 - y_i), & t_i = 0 \end{cases}$$

whereas the focal loss function $L_{Focal}$ was

$$L_{Focal} = \begin{cases} -\alpha(1 - y_i)^\gamma \log(y_i), & t_i = 1 \\ -(1 - \alpha)y_i^{\gamma} \log(1 - y_i), & t_i = 0 \end{cases}$$

where $t_i$ represents a binary label value; if pixel $i$ belongs to the ground truth, then $t_i = 1$, otherwise $t_i = 0$. $y_i$ represents the probability of the predicted pixel $i$ belonging to the ground truth. Parameter $\alpha$ alleviates the imbalance of positive and negative sample categories in each task, while parameter $\gamma$ is used to adjust the importance of the model to samples that are difficult to classify.

We combined the binary cross-entropy loss and focal loss functions, and found that their weights were $\beta$ and $1 - \beta$. Finally, the total loss function of the proposed network was

$$L_{Total} = \beta L_{BCE} + (1 - \beta) L_{Focal}$$

3. Experiments and Results

We used the BCDD dataset to verify the advanced outcome of our proposed method. We first describe the change detection dataset used. Next, we explain the evaluation metrics and parameter settings used in the experiments. Finally, we design ablation and comparison experiments and visualize and analyze the experimental results.

3.1. Dataset

In this study, we needed to use a dataset that contained both pre- and post-temporal building labels and changed building labels; therefore, we used the BCDD dataset released by Ji et al. [49] from Wuhan University. This dataset has been used by many scholars to
study various change detection methods \cite{25,27,31,33,46,47}. It includes aerial imagery, building labels, and changing building labels in Christchurch, New Zealand, before the 2012 earthquake and after the 2016 reconstruction. The spatial resolution of the dataset was 0.075 m/pixel and the image size was 32,507 × 15,324 pixels. We cropped the size of the BCDD dataset to 512 × 512 pixels without overlay and obtained 2400 pairs of available images after data cleaning. We provided statistics on the ratio of building pixels to other pixels and the ratio of changed pixels to unchanged areas before and after cleaning the data, as shown in Table 1.

Table 1. Statistics of changed pixels and unchanged pixels before and after cleaning data.

| Clear Data | Changed Pixels | Unchanged Pixels | C/UC  |
|------------|----------------|------------------|-------|
| No         | 21,188,729     | 450,670,471      | 21.2693 |
| Yes        | 21,188,729     | 240,958,226      | 11.3720 |

Finally, we randomly divided them into training and validation image pairs according to a ratio of 9:1; an example of the dataset is shown in Figure 4. When training the network, we used flips, rotations, brightness, and contrast enhancements with random probability.

![Figure 4](image-url)  
**Figure 4.** The images of 512 × 512 pixels are obtained by cropping the BCDD dataset. Each column represents a group of samples, containing five images of prechange, postchange, and ground truth. In T0 and T1 labels, white and black represent buildings and background, respectively. In change labels, white and black represent changed and unchanged areas, respectively. Panels (a–e) show changes in buildings, whereas (f,g) show no change.

3.2. Experimental Details
3.2.1. Evaluation Metrics

We used evaluation metrics to evaluate the performance of the various change detection algorithms. We evaluated the proposed model using precision, recall, F1-score, and total accuracy (OA). A higher precision indicated fewer false detections, whereas a higher recall indicated fewer missed detections. The F1-score is a metric used to measure...
the performance of a binary classification algorithm, considering both precision and recall. The F1-score and OA can comprehensively evaluate the performance of the model, with higher values representing improved performance. The evaluation metrics were defined as follows:

\[
\text{precision} = \frac{TP}{TP + FP} \quad (6)
\]

\[
\text{recall} = \frac{TP}{TP + FN} \quad (7)
\]

\[
F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (8)
\]

\[
OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)
\]

In the change detection task, true positive (TP) and true negative (TN) denote the number of changed and unchanged pixels detected correctly, whereas false positive (FP) and false negative (FN) denote the number of changed and unchanged pixels detected incorrectly. In the semantic segmentation task, TP and TN represent the number of buildings and background pixels detected correctly, respectively, whereas FP and FN represent the number of buildings and background pixels detected incorrectly, respectively.

### 3.2.2. Parameter Settings

We implemented distributed data parallel (DDP) and synchronized cross-GPU batch normalization (SyncBN) [50] on PyTorch using the NVIDIA NCCL ToolKit, which ensures high-performance computing of the model. In addition, we chose the Adam [51] optimizer and used the cosine annealing [52] algorithm as the learning rate decay method, which makes the network jump out of the saddle point faster by oscillating. The initial learning rate was set to $1 \times 10^{-3}$, while the minimum learning rate was set to $1 \times 10^{-7}$ to prevent a small learning rate. Considering the memory of the GPU, we set the batch size to 4. Furthermore, we set parameter $\alpha$ and parameter $\gamma$ in focal loss to 0.25 and 2, respectively. Finally, we ran all experiments on two NVIDIA GeForce RTX 2080 Ti GPUs with 11 GB memory.

### 3.3. Ablation Study

To investigate the effect of adding semantic segmentation branches, the FPN structure, and the scSE module on model performance, we conducted an ablation study on the BCDD dataset; the results are shown in Table 2.

#### Table 2. Ablation study of segmentation branch, FPN, and scSE module on the BCDD dataset.

| Method          | Seg | FPN | scSE | F1 (cd) | F1 (seg) |
|-----------------|-----|-----|------|---------|----------|
| Baseline        |     |     |      | 0.7791  | -        |
| Baseline + FPN  | √   |     |      | 0.8857  | -        |
| Baseline + scSE |     | √   |      | 0.8395  | -        |
| Baseline + FPN + scSE | √   | √   |      | 0.8863  | -        |
| Baseline + Seg  |     | √   |      | 0.8149  | 0.9059   |
| Baseline + Seg + FPN | √   | √   |      | 0.9032  | 0.9195   |
| Baseline + Seg + scSE | √   | √   |      | 0.8774  | 0.9297   |
| Baseline + Seg + FPN + scSE | √   | √   | √    | 0.9124  | 0.9441   |

It has been reported that without adding a semantic segmentation branch, adding an FPN or scSE module alone improves the performance of change detection by 10.66% and 6.04%, respectively, compared with the F1-score of the baseline model change detection of 0.7791. When the FPN and scSE modules were added simultaneously, the performance of the model was further improved, and the F1-score reached 0.8857, which was an increase of 10.72% compared with baseline. Interestingly, we found that the change detection F1-score of the model was increased by 3.58%, and we also obtained a semantic segmentation F1-score of 0.9059 after introducing the semantic segmentation branch. On this basis,
we assumed that the introduction of an FPN or scSE module further improved model performance. The former achieved a change detection F1-score of 0.9032 and a semantic segmentation F1-score of 0.9195, whereas the latter two tasks exhibited F1-scores of 0.8774 and 0.9297, respectively. Finally, if we applied all three to the baseline model, the F1-scores for change detection and semantic segmentation reached their highest values of 0.9124 and 0.9441, respectively.

Figure 5 illustrates the results of the ablation study. The first, third, and fifth lines represent new factories, subtle changes, and new dense buildings, respectively. The second, fourth, and sixth rows are the magnified details, respectively. Red and blue pixels in the results represent false and missed detections, respectively, while white and black represent changed and unchanged pixels, respectively. Before adding the semantic segmentation branch (Figure 5d–g), we observed many false and missed detections in the baseline model. However, adding FPN and scSE reduced false and missed detections, respectively. Further, adding FPN and scSE simultaneously greatly reduced false and missed detections. After adding the semantic segmentation branch (Figure 5h–k), we found that combining the three methods (Figure 5k) on the baseline model yielded the smallest false and missed detection rates, maximizing the ability to discriminate between building features of different scales. This implementation obtained more complete and regular boundaries and ultimately improved the comprehensive performance of the model. Through the comparison of performance and visual effects, we confirmed the easy application and effectiveness of our proposed model and method.

Figure 5. Visualization of ablation study on the BCDD dataset: (a,b) are bitemporal images; (c) the ground truth; (d) baseline; (e) baseline + FPN; (f) baseline + scSE; (g) baseline + FPN + scSE; (h) baseline + Seg; (i) baseline + Seg + FPN; (j) baseline + Seg + scSE; (k) baseline + Seg + FPN + scSE. Changed pixels are indicated by white, whereas unchanged areas are shown in black. Red and blue represent false and missed detections, respectively.
3.4. Comparative Study of Similarity Measures

A suitable similarity measure facilitates the identification of pseudo-changes by the network. To verify the effectiveness of the proposed FDE module, we only changed the feature similarity measure method in MDESNet and chose the following four similarity measures:

(a) Concatenation [27]: We concatenated the bitemporal features, and then used three consecutive $3 \times 3$ convolutional layers, which reduced the channels, to extract the change information from the connected features according to the FC-Siam-conc.

(b) Difference [27]: We subtracted the bitemporal features from the corresponding channel dimension and used the absolute value of the difference as the changed feature.

(c) Normalized difference [53]: Based on this difference, we performed a further normalization operation.

(d) Local similarity attention module [31]: In this module, we extracted the similarity attention (SA) value from the input feature maps using the cosine distance and then multiplied the SA element by element with the bitemporal feature maps. Finally, we concatenated the bitemporal feature maps and applied a $3 \times 3$ convolutional layer to adjust the number of channels as the changed features.

We calculated the OA, precision, recall, and F1-score for the five methods, as shown in Table 3. We observed that the proposed FDE module performed the best in terms of accuracy (0.9874), precision (0.9264), and F1-score (0.9124). Compared with the second case, it was better by 0.26%, 4.54%, and 1.40%, respectively. We found that the best performance in recall was that of local similarity attention, which reached 0.9571, whereas that of our method was only 0.8988. The F1-scores of the difference and normalized difference were not markedly different at 0.8984 and 0.8974, respectively, which were 1.40% and 1.50% lower than that of the FDE module, respectively. We also detected that the F1-scores of the concatenation and local similarity attention modules were 0.8767 and 0.8781, which were 3.57% and 3.43% lower than that of the FDE module, respectively. This confirmed that the FDE module achieved the best F1-score among the five methods and improved the comprehensive performance of the model. The highest precision also indicated that the FDE method achieved the fewest false detections and was more resistant to pseudo-changes.

Table 3. Results generated by several similarity measures on the BCDD dataset.

| Method                        | OA    | Pre.  | Rec.  | F1    |
|-------------------------------|-------|-------|-------|-------|
| Concatenate                   | 0.9814| 0.8458| 0.9100| 0.8767|
| Difference                    | 0.9848| 0.8750| 0.9231| 0.8984|
| Normalized difference         | 0.9772| 0.8810| 0.9559| 0.8974|
| Local similarity attention    | 0.9807| 0.8111| 0.9571| 0.8781|
| FDE (ours)                    | 0.9874| 0.9264| 0.8988| 0.9124|

In the FDE module, inspired by the scSE module, we compressed the channels after subtracting the bitemporal features and used the sigmoid activation function to obtain the feature difference attention maps. We visualized the feature-difference attention maps on four scales to analyze the principle of the FDE module. Figure 6 shows differences in the emphasis of the feature-difference attention map at different scales. We observed that the attention map at a resolution of $16 \times 16$ pixels contained rich low-frequency semantic information, whereas the FDE module focused on enhancing semantic differences. In the process of increasing the resolution, we detected that the enhanced center of gravity gradually shifted to spatial information owing to the abundance of the high-frequency spatial details of the attention maps. Finally, we found that the FDE module enhanced the differences in features at multiple scales to achieve a comprehensive performance of the model.
Figure 6. (1) and (2) are two different examples; the difference being that buildings were also present in the pretemporal image in (2)(a). In the two groups, (a) and (b) show pre- and post-temporal images, respectively; (c) and (d) show the change ground truth and predicted result, respectively, where white represents the buildings and black represents the background. (e–h) Feature difference attention maps at 4 scales (16 × 16, 32 × 32, 64 × 64, and 128 × 128 pixels), in which blue to red represents enhancement from weak to strong.

3.5. Comparative Study Using Other Methods

To evaluate the performance and competitiveness of MDESNet, we selected several popular change detection models for comparative experiments, as follows:

(a) FC-EF [27]: This model concatenated two temporal images and formed an image with skip connections and a U-shaped structure, using six channels as input. We extracted
the change feature from the fused image, and finally obtained the change result using the softmax function.

(b) FC-Siam-conc [27]: This model was an extension of FC-EF that used a Siamese network with the same structure and shared weights as the encoder. We concatenated the extracted bitemporal features and then inputted them into the decoder with skip connections to obtain the change results.

(c) FC-Siam-diff [27]: This model was very similar to FC-Siam-conc; the difference was that we subtracted and obtained the absolute value of the extracted bitemporal features and then input them into the decoder with skip connections to obtain the change results.

(d) ChangeNet [30]: This was proposed to detect changes in street scenes. We sampled the change features of the three scales to the same scale and used the softmax function to obtain the change results after summation, which located and identified the changes between image pairs.

(e) DASNet [25]: Its core was set to use the spatial attention and channel attention algorithms to obtain more abundant discriminative features. Unlike other methods, this method gave a distance map as output and used a threshold algorithm to obtain the final changed results.

(f) SNUNet-CD [54]: This method used a dense connection Siamese network, similar to UNet++, which is known to mitigate the effects of deep location information loss in neural networks. We employed the ensemble channel attention module (ECAM) to extract the representative features of different semantic levels and obtain the change results.

(g) DTCDSCN [33]: This model was also a multitask Siamese network with two semantic segmentation branches and a change detection branch, similar to the proposed MDESNet. Its decoder structure was similar to that of FC-Siam-diff, except that it added the scSE module to improve the feature representation capabilities.

To ensure fairness and effectiveness of the comparative study and consistency with MDESNet, we replaced the backbone network used by ChangeNet, DASNet, and DTCDSCN with ResNeSt-50. Of note, FC-EF, FC-Siam-conc, FC-Siam-diff, and SNUNet-CD do not apply backbone networks such as VGG or ResNet; therefore, their models remain unchanged. The other parameters were consistent with those in the respective studies. In addition, we trained them on the divided BCDD dataset under the same experimental environment. Table 4 shows that the proposed network outperformed the other models in terms of accuracy, precision, and F1-score, reaching 0.9874, 0.9264, and 0.9124, respectively, which corresponded to improvements of 0.72%, 5.89%, and 2.96%, respectively. These results also verified the validity of the FDE and MSFF modules. We found that DASNet achieved the highest recall rate of 0.9266, which was 2.78% higher than that of the proposed network. The proposed method did not achieve the best recall because it was more focused on solving false detections owing to pseudo-changes; however, the highest F1-score verified that our proposed method had the best comprehensive performance.

Table 4. Comparative study of several change detection methods on the BCDD dataset.

| Method        | OA   | Pre. | Rec.  | F1   |
|---------------|------|------|-------|------|
| FC-EF         | 0.9747 | 0.8553 | 0.7943 | 0.8237 |
| FC-Siam-conc  | 0.9662 | 0.7199 | 0.8932 | 0.7972 |
| FC-Siam-diff  | 0.9555 | 0.6425 | 0.9056 | 0.7517 |
| ChangeNet     | 0.9378 | 0.5560 | 0.8119 | 0.6600 |
| DASNet        | 0.9802 | 0.8430 | **0.9266** | 0.8828 |
| SNUNet-CD     | 0.9792 | 0.8675 | 0.9494 | 0.8584 |
| DTCDSCN       | 0.9917 | 0.7469 | 0.9233 | 0.8258 |
| MDESNet (ours)| **0.9874** | **0.9264** | 0.8988 | **0.9124** |

Figure 7 shows a part of the experimental results that more intuitively demonstrate the superior performance of our proposed network. The first and second rows demonstrate that our model recovered more regular boundaries and tighter interiors when recognizing
large or small buildings, respectively. In the third and fourth rows, although the bitemporal images contained impermeable pavements and containers that were similar to roof material, our proposed network recognized them as unchanged pixels, indicating its ability to resist pseudo-changes or noise. The last two rows show that our model identified and located imperceptible changes far better than other models.

![Figure 7](image_url)

**Figure 7.** Visualized comparison of the results of several change detection methods on the BCDD dataset: (a,b) are bitemporal images; (c) is the ground truth; (d) FC-EF; (e) FC-Siam-conc; (f) FC-Siam-diff; (g) ChangeNet; (h) DASNet; (i) SNUNet-CD; (j) DTCDSCN(CD); (k, ours) MDESNet. Changed pixels are represented by white, whereas unchanged areas are shown in black.

4. **Discussion**

4.1. **The Effect of Semantic Segmentation Branches**

Table 2 shows the effect of adding the semantic segmentation branch, FPN structure, and scSE algorithm on the performance of the baseline model. To more clearly compare the changes in the performance of the model before and after adding the semantic segmentation branch, we made a histogram, as shown in Figure 8. We found that adding the semantic segmentation branch on different baseline models improved the F1-score of change detection.
Remote Sens. 2022, 14, x FOR PEER REVIEW

16 of 21

the changes in the performance of the model before and after adding the semantic segmentation branch, we made a histogram, as shown in Figure 8. We found that adding the semantic segmentation branch on different baseline models improved the F1-score of change detection.

![Figure 8](https://via.placeholder.com/150)

**Figure 8.** Comparison of F1-scores with and without semantic segmentation branches under different baseline models.

We also selected several semantic segmentation models for dense prediction for comparison with MDESNet: UNet [55], UNet++ [56], PSPNet [57], DeepLabv3+ [58], DTCDSCN [33], and FarSeg [45]. The average experimental results for the two branches in the same environment are shown in Table 5. The accuracy, precision, recall, and F1-score of the semantic segmentation branch of the proposed model were 0.9792, 0.9485, 0.9397, and 0.9441, respectively. Compared with other models, the performance of MDESNet in the semantic segmentation task was at a moderate level. This was consistent with the failure of the semantic segmentation task in the DTCDSCN to achieve a good F1-score. Although adding a semantic segmentation branch adds constraints to the encoder for extracting building features and achieving better change detection results, it also restricts the accuracy of the semantic segmentation model to a certain extent. This was a limitation of the present study. In the future, further research should be conducted on the means by which to achieve satisfactory results for both the main task of change detection and the auxiliary task of semantic segmentation.

| Method          | OA     | Pre.   | Rec.   | F1     |
|-----------------|--------|--------|--------|--------|
| UNet            | 0.9801 | 0.9546 | 0.9410 | 0.9478 |
| UNet++          | 0.9814 | 0.9469 |        |        |
| PSPNet          | 0.9730 | 0.9258 | 0.9341 | 0.9299 |
| DeepLabv3+      | 0.9808 | 0.9519 | 0.9477 | 0.9498 |
| FarSeg          | 0.9816 | 0.9587 | 0.9447 | 0.9516 |
| DTCDSCN         | 0.9711 | 0.9158 | 0.9352 | 0.9255 |
| MDESNet (ours)  | 0.9792 | 0.9485 | 0.9397 | 0.9441 |

**Table 5.** Comparative study of several semantic segmentation methods on the BCDD dataset.

### 4.2. The Effect of the Value of $\beta$ in Loss Function

Several experiments were designed to determine the value of $\beta$ in the loss function with the best performance. Figure 9 shows the effect of different $\beta$ values on the overall performance of the network. When $\beta$ was 5/6, the performance of the network was optimal, achieving the highest change detection F1-score (0.9124) and the highest semantic
segmentation F1-score (0.9441). If only the focal loss function or the binary cross-entropy loss function was applied, the performance of the network was slightly lower, indicating that this combination was beneficial to the improvement of the network performance.

![Figure 9. The influence of $\beta$ value on the performance of MDESNet.](image)

**4.3. Comparison of the Number of Parameters and Prediction Efficiency**

Figure 10 shows the number of parameters of the proposed model. Compared with the multitask detection model, DTCDSN, of the same type, our model reduced the number of parameters by 4.278 M. In addition, compared with other single-task change detection models and the best-performing semantic segmentation model (UNet++), our method had the highest change detection F1-score and the lowest number of parameters, indicating the advantage of the multitask model in terms of the number of parameters. One of the future research directions will be to further reduce the number of parameters of the multitask change detection model while retaining the network performance.

![Figure 10. Comparison of the number of parameters between MDESNet, multitask model DTCDSN, and 2 single-task models (change detection models and UNet++).](image)
5. Conclusions

In this study, we proposed MDESNet, a multitask difference-enhanced Siamese network that can output semantic information and change information end-to-end, for building change detection in high-resolution remote sensing images. By combining the Siamese network and FPN structure, we generated bitemporal multiscale feature maps that fully incorporated contextual information. The semantic segmentation branches added constraining information to the encoder for the extraction of ground features, having a positive impact on the change detection branch. To enhance the changed regions in the feature maps and improve the interclass separability, we proposed the FDE module as a similarity measure for bitemporal multiscale features. Our proposed network achieved the highest F1-score (0.9124) and OA (0.9874) on the BCDD dataset, with better overall performance and resistance to spurious changes compared with other state-of-the-art methods. Likewise, compared with the same type of multitask model, our method had higher accuracy and contained fewer parameters. However, its disadvantage was that it had an intermediate level of performance in the semantic segmentation branch.

In the future, we will validate the proposed method using other public datasets and conduct further research on the relationship and influence between semantic segmentation and change detection.

Author Contributions: Conceptualization, J.Z. and Y.T.; methodology, J.Z.; validation, K.Y. and C.Y.; formal analysis, K.Y. and C.Y.; investigation, J.Z. and F.C.; resources, J.Z. and F.Z.; data curation, J.Z. and K.Y.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z. and K.Y.; visualization, J.Z.; supervision, Y.T., F.Z. and Q.C.; project administration, J.Z., Y.T., C.Y., K.Y., F.Z., F.C. and Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Major Project of High Resolution Earth Observation System of the State Administration of Science, Technology and Industry for National. (06-Y30F04-9001-2022).

Data Availability Statement: Publicly available datasets were analyzed in this study. The dataset can be found here: BCDD: http://study.rsgis.whu.edu.cn/pages/download/building_dataset.html, accessed on 25 July 2022.

Acknowledgments: Our team is grateful to those who have published data and classic models; it is their selflessness that provides an open research environment for the community.

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

Abbreviations

The following abbreviations are used in this manuscript:

- MDESNet: Multitask difference-enhanced Siamese network
- FDE: Feature difference enhancement
- BCDD: Building change detection dataset
- PBCD: Pixel-based change detection
- OBCD: Object-based change detection
- CVA: Change vector analysis
- PCA: Principal component analysis
- FC-EF: Fully convolutional early fusion
- FC-Siam-conc: Fully convolutional Siamese-concatenation
- FC-Siam-diff: Fully convolutional Siamese-difference
- LSTM: Long short-term memory
- DASNet: Dual attentive fully convolutional Siamese network
- CDD: Change detection dataset
- UCD: Urban change detection dataset
- MSFF: Multiscale feature fusion
- ResNeSt: Split-attention networks
- FDM: Feature difference map
- FDAM: Feature difference attention map
- scSE: Concurrent spatial and channel squeeze and channel excitation
OA Overall accuracy
TN True negative
TP True positive
FN False negative
FP False positive
DDP Distributed data parallel
SyncBN Synchronized cross-GPU batch normalization
Adam Adaptive moment estimation
SNUNet-CD Siamese nested-Unet network for change detection
ECAM Ensemble channel attention module
DTCDSN Dual-task constrained deep Siamese convolutional network
PSPNet Pyramid scene parsing network

References

1. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep Learning-Based Classification of Hyperspectral Data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2014, 7, 2094–2107. [CrossRef]

2. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. 2004, 42, 1778–1790. [CrossRef]

3. Tian, Y.; Wu, B.; Zhang, L.; Li, Q.; Jia, K.; Wen, M. Opium poppy monitoring with remote sensing in North Myanmar. Int. J. Drug Policy 2011, 22, 278–284. [CrossRef]

4. Tian, Y.; Yin, K.; Lu, D.; Hua, L.; Zhao, Q.; Wen, M. Examining Land Use and Land Cover Spatiotemporal Change and Driving Forces in Beijing from 1978 to 2010. Remote Sens. 2014, 6, 10593–10611. [CrossRef]

5. Yin, K.; Lu, D.; Tian, Y.; Zhao, Q.; Yuan, C. Evaluation of Carbon and Oxygen Balances in Urban Ecosystems Using Land Use/Land Cover and Statistical Data. Sustainability 2014, 7, 195–221. [CrossRef]

6. Xue, J.; Xu, H.; Yang, H.; Wang, B.; Wu, P.; Choi, J.; Cai, L.; Wu, Y. Multi-Feature Enhanced Building Change Detection Based on Semantic Information Guidance. Remote Sens. 2021, 13, 4171. [CrossRef]

7. Wang, H.; Lv, X.; Zhang, K.; Guo, B. Building Change Detection Based on 3D Co-Segmentation Using Satellite Stereo Imagery. Remote Sens. 2022, 14, 628. [CrossRef]

8. Bruzzone, L.; Bovolo, F. A Novel Framework for the Design of Change-Detection Systems for Very-High-Resolution Remote Sensing Images. Proc. IEEE 2013, 101, 609–630. [CrossRef]

9. Dell’Acqua, F.; Gamba, P.; Ferrari, A.; Palmason, J.A.; Arnason, K. Exploiting spectral and spatial information in hyperspectral urban data with high resolution. IEEE Geosci. Remote Sens. Lett. 2004, 1, 322–326. [CrossRef]

10. Gong, P.; Li, X.; Xu, B. Interpretation Theory and Application Method Development for Information Extraction from High Resolution Remotely Sensed Data. J. Remote Sens. 2006, 10, 1–5.

11. Myint, S.W.; Lam, N.S.N.; Tyler, J.M. Wavelets for urban spatial feature discrimination: Comparisons with fractal, spatial autocorrelation, and spatial co-occurrence approaches. Photogramm. Eng. Remote Sens. 2004, 70, 803–812. [CrossRef]

12. Weismüller, R.A.; Kristof, S.J.; Scholz, D.K.; Anuta, P.E.; Momín, S.A. Change detection in coastal zone environment. Photogramm. Eng. Remote Sens. 1978, 12, 1533–1539.

13. Zhang, Z.; Li, Z.; Tian, X. Vegetation change detection research of Dunhuang city based on GF-1 data. Int. J. Coal Sci. Technol. 2018, 5, 105–111. [CrossRef]

14. Singh, A. Review Article Digital change detection techniques using remotely-sensed data. Int. J. Remote Sens. 1989, 10, 989–1003. [CrossRef]

15. Bruzzone, L.; Prieto, D.F. Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. 2000, 38, 1171–1182. [CrossRef]

16. Celik, T. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and K-Means Clustering. IEEE Geosci. Remote Sens. Lett. 2009, 6, 772–776. [CrossRef]

17. Li, P.; Xu, H. Land-cover change detection using one-class support vector machine. Photogramm. Eng. Remote Sens. 2010, 76, 255–263. [CrossRef]

18. Seo, D.; Kim, Y.; Eo, Y.; Park, W.; Park, H. Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection. Remote Sens. 2017, 9, 1163. [CrossRef]

19. 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]

20. Hay, G.J.; Niemann, K.O. Visualizing 3-D texture: A three-dimensional structural approach to model forest texture. Can. J. Remote Sens. 1994, 20, 90–101.

21. Lefebvre, A.; Corpetti, T.; Hubert-Moy, L. Object-Oriented Approach and Texture Analysis for Change Detection in Very High Resolution Images. In Proceedings of the 2008 IEEE International Geoscience and Remote Sensing Symposium, Boston, MA, USA, 6–11 July 2008; pp. 663–666.

22. Yu, W.; Zhou, W.; Qian, Y.; Yan, J. A new approach for land cover classification and change analysis: Integrating backdating and an object-based method. Remote Sens. Environ. 2016, 177, 37–47. [CrossRef]
23. Zhou, W.; Troy, A.; Grove, M.J.S. Object-based Land Cover Classification and Change Analysis in the Baltimore Metropolitan Area Using Multitemporal High Resolution Remote Sensing Data. *Sensors* 2008, 8, 1613–1636. [CrossRef] [PubMed]

24. Jung, S.; Lee, W.H.; Han, Y. Change Detection of Building Objects in High-Resolution Single-Sensor and Multi-Sensor Imagery Considering the Sun and Sensor’s Elevation and Azimuth Angles. *Remote Sens.* 2021, 13, 3660. [CrossRef]

25. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2021, 14, 1194–1206. [CrossRef]

26. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444. [CrossRef]

27. Daudt, R.C.; Saux, B.L.; Boulech, A. Fully Convolutional Siamese Networks for Change Detection. *arXiv* 2018, arXiv:1810.08462.

28. Papadomanolaki, M.; Verma, S.; Vakalopoulou, M.; Gupta, S.; Karantzalos, K. Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–1 August 2019.

29. Li, S.; Huo, L. Remote Sensing Image Change Detection Based on Fully Convolutional Network With Pyramid Attention. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 16–17 July 2021; pp. 4352–4355.

30. Varghese, A.; Gubbi, J.; Ramaswamy, A.; Balamuralidhar, P. ChangeNet: A Deep Learning Architecture for Visual Change Detection. In *Proceedings of the Computer Vision—ECCV 2018 Workshops*; Springer: Cham, Switzerland, 2019; pp. 129–145.

31. Lee, H.; Lee, K.; Kim, J.H.; Na, Y.; Park, J.; Choi, J.P.; Hwang, J.Y. Local Similarity Siamese Network for Urban Land Change Detection on Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2021, 14, 4139–4149. [CrossRef]

32. Li, X.; He, M.; Li, H.; Shen, H. A Combined Loss-Based Multiscale Fully Convolutional Network for High-Resolution Remote Sensing Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* 2022, 19, 1–5. [CrossRef]

33. 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. *arXiv* 2019, arXiv:1909.07726. [CrossRef]

34. Lin, T.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.

35. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel ‘Squeeze & Excitation’ in Fully Convolutional Networks. *arXiv* 2018, arXiv:1803.02579.

36. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a “Siamese” time delay neural network. In Proceedings of the 6th International Conference on Neural Information Processing Systems, Denver, CO, USA, 29 November–2 December 1993; pp. 737–744.

37. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 539–546.

38. Benajiba, Y.; Sun, J.; Zhang, Y.; Jiang, L.; Weng, Z.; Biran, O. Siamese Networks for Semantic Pattern Similarity. *arXiv* 2018, arXiv:1812.06604.

39. Ranasinghe, T.; Oraee, C.; Mitkov, R. Semantic Textual Similarity with Siamese Neural Networks. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, Varna, Bulgaria, 2–4 September 2019; pp. 1004–1011.

40. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.-L.; Lin, H.; Sun, Y.; He, T.; Mueller, J.; Mannmatha, R.; et al. ResNeSt: Split-Attention Networks. *arXiv* 2020, arXiv:2004.08955.

41. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

42. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv* 2016, arXiv:1611.05431.

43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *arXiv* 2017, arXiv:1709.01507.

44. Kirillov, A.; Girshick, R.; He, K. Panoptic Feature Pyramid Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6392–6401.

45. Zheng, Z.; Zhong, Y.; Wang, J. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4095–4104.

46. Zhang, X.; He, L.; Qin, K.; Dang, Q.; Si, H.; Tang, X.; Jiao, L. SMD-Net: Siamese Multi-Scale Difference-Enhancement Network for Change Detection in Remote Sensing. *Remote Sens.* 2022, 14, 1580. [CrossRef]

47. Zheng, D.; Wei, Z.; Wu, Z.; Liu, J. Learning Pairwise Potential CRFs in Deep Siamese Network for Change Detection. *Remote Sens.* 2022, 14, 841. [CrossRef]

48. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, 42, 318–327. [CrossRef]

49. 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]
50. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context Encoding for Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.

51. Kingma, D.P.; Ba, J.J.A. Adam: A Method for Stochastic Optimization. arXiv 2014, arXiv:1412.6980.

52. Loshchilov, I.; ArXiv, F.H.J. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv 2017, arXiv:1608.03983.

53. Zhang, M.; Shi, W. A Feature Difference Convolutional Neural Network-Based Change Detection Method. IEEE Trans. Geosci. Remote Sens. 2020, 58, 7232–7246. [CrossRef]

54. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. IEEE Geosci. Remote Sens. Lett. 2022, 19, 1–5. [CrossRef]

55. Olaf, R.; Philipp, F.; Thomas, B. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Munich, Germany, 5–9 October 2015.

56. Zhou, Z.; Md Mahfuzur Rahman, S.; Nima, T.; Jianming, L. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the DLMIA: International Workshop on Deep Learning in Medical Image Analysis, Granada, Spain, 20 September 2018; Springer: Berlin/Heidelberg, Germany, 2018.

57. 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, Honolulu, HI, USA, 21–16 July 2017; pp. 6230–6239.

58. Liang-Chieh, C.; Yukun, Z.; George, P.; Florian, S.; Hartwig, A. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.