

Article

PGA-SiamNet: Pyramid Feature-Based Attention-Guided Siamese Network for Remote Sensing Orthoimagery Building Change Detection

Huiwei Jiang ¹ , Xiangyun Hu ^{1,2,*}, Kun Li ¹, Jinming Zhang ¹, Jinqi Gong ¹ and Mi Zhang ¹

¹ School of Remote Sensing and Information Engineering, 129 Luoyu Road, Wuhan University, Wuhan 430079, China; huiwei_jiang@whu.edu.cn (H.J.); petrick_lee@whu.edu.cn (K.L.); nicnyzjm@whu.edu.cn (J.Z.); jinqigong@whu.edu.cn (J.G.); mizhang@whu.edu.cn (M.Z.)

² Collaborative Innovation Center of Geospatial Technology, 129 Luoyu Road, Wuhan University, Wuhan 430079, China

* Correspondence: huxy@whu.edu.cn; Tel.: +86-27-6877-1528; Fax: +86-27-6877-8086

Received: 10 January 2020; Accepted: 1 February 2020; Published: 3 February 2020



Abstract: In recent years, building change detection has made remarkable progress through using deep learning. The core problems of this technique are the need for additional data (e.g., Lidar or semantic labels) and the difficulty in extracting sufficient features. In this paper, we propose an end-to-end network, called the pyramid feature-based attention-guided Siamese network (PGA-SiamNet), to solve these problems. The network is trained to capture possible changes using a convolutional neural network in a pyramid. It emphasizes the importance of correlation among the input feature pairs by introducing a global co-attention mechanism. Furthermore, we effectively improved the long-range dependencies of the features by utilizing various attention mechanisms and then aggregating the features of the low-level and co-attention level; this helps to obtain richer object information. Finally, we evaluated our method with a publicly available dataset (WHU) building dataset and a new dataset (EV-CD) building dataset. The experiments demonstrate that the proposed method is effective for building change detection and outperforms the existing state-of-the-art methods on high-resolution remote sensing orthoimages in various metrics.

Keywords: building change detection; remote sensing orthoimagery; attention mechanism; Siamese convolutional neural network

1. Introduction

1.1. Background

Remote sensing imagery has found a wide range of applications because it can obtain change information occurring around the world, both in densely populated cities and in hard-to-reach areas. Meanwhile, change detection (CD), as a hot topic in the field of remote sensing analysis, has been studied for several decades. Because of its unique characteristics, many CD studies have been dedicated to solving large-scale and complicated problems using remote sensing images, for example, for the monitoring of forests and urban sprawl, and earthquake assessment over long periods. A lot of research institutions have made many intensive studies on CD, such as the seasonal and annual change monitoring (SATChMo) project [1] of Poland, earth watching [2] of European space agency (ESA) and Onera satellite change detection [3] of the IEEE geosciences and remote sensing association (IEEE GRSS).

Recently, the high-resolution (HR) or very-high-resolution (VHR) images have received a lot of attention because they can reveal more detailed information about the land surface, thereby increasing

the possibility of monitoring small but important objects such as buildings. Driven by this, building change detection (BCD) has attracted substantial attention in applications such as urbanization monitoring, illegal or unauthorized building identification, and disaster evaluation. In addition, automatic building change detection for remote sensing images has become a topical issue because carrying out the task manually is time consuming and tedious [4]. Therefore, there is a crucial need to investigate efficient building change detection algorithms for remote sensing images.

In general, the traditional change detection process consists of three major steps: preprocessing, change detection technique selection, and accuracy assessment. Moreover, these methods typically follow one of two approaches [5]: pixel-based or object-based methods [6–10]. Pixel-based methods are mainly used in large-scale change detection with low or medium resolution images (e.g., MODIS), while object-based methods are more popular for HR or VHR images (e.g., QuickBird, GeoEye-1, WorldView-1/2, Aerial imagery), because the high-frequency components in the VR/VHR images cannot be fully represented by pixel-based methods. These methods have been fully developed in various scenarios in the past years; however, the features used by change detection algorithms are almost hand crafted and so are weak in image representation [11]. The features are also susceptible to the process in the preprocessing stage, such as those involved in radiometric correction, geometric correction, and image registration. Besides this, the effects of changes in the appearance of objects caused by different photographic angles can be alleviated by orthorectification, but this also brings new problems affecting accurate detection. In high-resolution orthorectified images, the displacement of buildings (especially high-rise buildings) mainly caused by rectifications with a digital elevation model (DEM) and the poor alignment caused by the displacement can lead to many false positive changes. In addition, as the spatial resolution of satellite images increases, the accuracy of image registration tends to worsen [5,8]. Generally, the overall framework of the change detection technique can be summarized as feature extractions and change decisions, which is depicted in Figure 1.

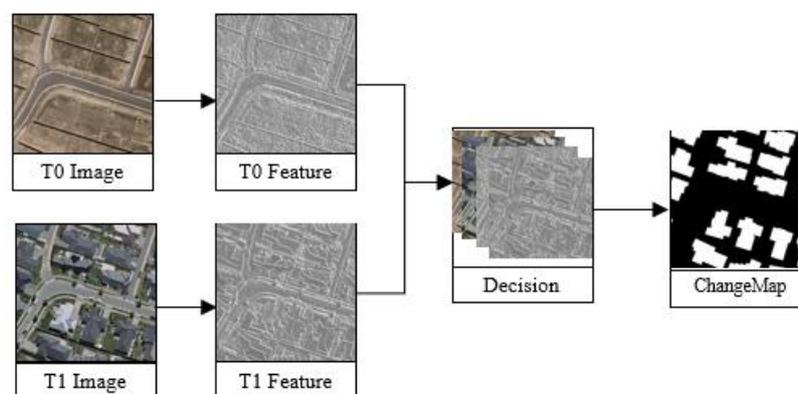


Figure 1. The basic framework for change detection.

Considering the facts described above, it is not sufficient to obtain the building changes with the 2D information delivered by satellite images, in which many irrelevant changes are mixed with the desired changes. Recently, a lot of research has focused on improving the precision of building change detection. For example, when taking the features extended into 3D space, i.e., the height information, which is free of illumination variations and perspective distortions [12], building extraction and change detection accuracy is improved [13,14]. Moreover, with the Lidar system expansion, some approaches using laser scanning data have been making headway [15]. However, when it comes to large and remote areas, the data is either low frequency, or hard or even impossible to acquire. In addition, owing to breakthroughs in dense image matching (DIM) techniques, the availability of image-based 3D information [16] has greatly increased; this field being known as DSM (digital surface model)-assisted building change detection [17]. However, the relatively low quality of the DSMs from

satellite data, which is strongly dependent on image matching technology, is still a major obstacle for detection accuracy.

Recently, the excitement around deep convolutional neural networks (CNN) has been tremendous, with successful applications in many real-world vision tasks, such as object detection [18], image classification [19], semantic segmentation [20], as well as change detection [21]. However, remote sensing applications present some new challenges for deep learning regarding multimodal and multisource data [22]. Thanks to learned feature representations, which are more robust to appearance and shape variations, significant performance improvements have been achieved. The researches in the field of building change detection can be divided into two: (1) post-classification-based methods; and (2) direct-detection-based methods. The first methods involve first extracting the buildings from the images at a different time in the same area, and then obtaining the changed buildings by comparing the extracted building maps. Furthermore, if applying the post-classification-based methods, the displacement of the objects becomes less important, and changes are found by verifying whether the two images contain the same object or not [23]. However, a high accuracy of building extraction is required, which is a hard task in itself and can lead to accumulated errors. The second set of methods are based on an end-to-end framework and have been successfully used to identify building changes. They avoid some of the weaknesses of classic methods, especially in dense urban areas. Rodrigo has trained an end-to-end Siamese architecture from scratch using a fully convolutional network, and the result surpassed the state-of-the-art methods in change detection, both in accuracy and in inference speed without any post-processing [24]. As mentioned above, the displacement of the buildings in the orthorectified images is a major challenge for most end-to-end building change detection methods. However, most current methods do not take the building displacement into account and ignore the correlation between the image pairs. Lebedev proposed a specially modified generative adversarial network (GAN) architecture based on pix2pix for automatic change detection in season-varying remote sensing images. In this research, object shift was considered, which is crucial for the buildings in orthorectified images [25].

Despite the wide availability of CNN, it lacks large amounts of available corresponding change annotations to provide training data, which is necessary to train a reliable change detector in a supervised approach. Focusing on this issue, many recent researches explore alternatives, such as training a weak supervised network [26–28], applying an unsupervised approach [29–31], or even focusing on noisy data [28]. However, studies on building change detection mainly concentrate on either two-stage detection accompanied with building detection or one-stage without taking the displacement of the buildings into account.

Recently, weakly supervised approaches have been proposed in an attempt to compensate for the dependence on the large annotated data, such as training with synthetic data obtained by a given geometric transformation [32–34]. However, detecting building changes with these methods has some restrictions: either an accurate position of the change cannot be given, or information of an independent building is required.

1.2. Related Work

1.2.1. Attention Mechanism

The so-called attention is a way to observe the world and imitate the mechanism of humans [35]. Recently, it has been demonstrated to be a simple but effective tool for improving the representation ability of CNN through reweighting of the feature maps; this is achieved using spatial attention and channel attention to scale the features which are meaningful or useless [36–43].

Channelwise attention. In order to improve the discriminant of the abstract feature, the channels of high-level features may number in the thousands (in which there inevitably exists redundant information), and each of them can be regarded as a class-specific response [37]. To exploit the interdependencies between these maps, channelwise attention was created to emphasize the channels

that are relatively informative and focus on the meaningful input. Through obtaining the channel relationship matrix, the self-attention channel weight of the original feature map can be further calculated, helping to boost feature discriminability [36,41,44,45].

Spatialwise attention. In view of the fact that the low-level features usually contain a large number of details and the receptive field of the convolution layer in traditional FCN (full convolutional network) is limited, if the network only establishes a pixel relationship in the local neighborhood, it can easily lead to unsatisfactory results. In addition, the studies on how effectively obtained long-range dependencies without using deeply stacking convolution layers are more attractive. Instead of considering all positions, spatial attention is introduced to find the relationships of the position and highlight the meaningful region. To take full advantage of the information of the features, more and more researchers are preferring to combine the spatialwise attention and channelwise attention [36–38,45].

Co-Attention mechanism. More recently, in order to understand the fine-grained relationship and mine the underlying correlations between different modalities, co-attention mechanisms have been widely studied in vision-and-language tasks, such as in visual question answering (VQA) [46–49]. In the computer vision area, Lu was inspired by the above-mentioned works and built the co-attention module to capture the coherence between video frames and effectively suppress the current alternatives [40].

1.2.2. Semantic Correspondence Mechanism

In general, the issue of change detection can be attributed to finding and matching pairs of images [50–52]. The rule of the method is based on the idea that by determining whether a similar semantic object in the second image exists or not to express unchanged and changed. Therefore, finding the correspondence of the point has long been one of the fundamental problems in the field of computer vision and photogrammetry. Traditional methods have been quite successful; they usually employed hand-crafted descriptors (e.g., Scale Invariant Feature Transform (SIFT) [53], Histogram of Oriented Gradients (HOG) [54], Oriented FAST and Rotated BRIEF (ORB) [55]) to find the key point correspondences based on minimizing an empirical matching strategy and then rejecting the outliers with a geometric match model. Afterwards, several studies began to consider trainable descriptors [56,57] with CNN. However, these methods heavily depended on predefined features of sparse points and none of the models could be directly used for semantic alignment [32,58]. Given the recent success of using end-to-end CNN in various tasks, many approaches for semantic matching have been proposed with promising results [56,59]. However, these methods also suffer from the same limitations as many other machine learning tasks. Therefore, to achieve satisfactory results, large-scale and diverse training data are required, which is labor intensive and time consuming.

In this paper, we propose a novel framework for building change detection, using satellite orthorectified images to model the complex change relationships of the buildings with displacement in the scene. The main contributions of this paper can be summarized as follows:

- (1) We introduce a co-attention module which can deal with the displacement of buildings in orthoimages to enhance the feature representations and further mine the correlations therein. Meanwhile, we fuse the semantic and context information of the feature using a context fusion strategy;
- (2) We provide a new satellite dataset for building change detection frameworks covering various sensors, and verify its effectiveness by conducting extensive experiments;
- (3) We propose an effective Siamese building change detection framework and make some improvements. Moreover, we train our model using two different datasets. The proposed method shows superior performance: it can directly obtain pixel-level predictions without any other post-processing techniques.

The structure of this paper is organized as follows: Section 2 describes the datasets and the proposed method of this paper. The experimental results and accuracy assessments are presented in

Section 3. Section 4 presents the discussion. Finally, the conclusion of this paper is summarized in Section 5.

2. Materials and Methods

In this section, we provide the formulation of our method to detect building changes. Firstly, we introduce the datasets used in our study in Section 2.1. The description of our network, pyramid feature-based attention-guided Siamese network (PGA-SiamNet), is presented in Section 2.2. Finally, in Section 2.3, the implementation of the experiment is described in detail.

2.1. Datasets

In this paper, in order to train the proposed network and evaluate its performance, we adopted two different building change detection datasets, namely dataset I (DI) and dataset II (DII). The first dataset (DI) is the Wuhan University (WHU) building change detection dataset [60] which covers Christchurch, New Zealand, and contains two scenes acquired at the same location in 2012 and 2016, with the semantic labels of the buildings and the change detection labels. The dataset is made up of aerial imagery data with a 0.075 m spatial resolution.

We named the DII dataset the earth vision-change detection (EV-CD building) dataset. The dataset was labeled by us and is extremely challenging. Moreover, soon it will be available to the public. The dataset is much more complex than the DI dataset and is made up of satellite imagery data instead of aerial imagery data. The dataset consists of data from a variety of sensors with different spatial resolution ranges from 0.2 to 2 m and contains several cities in the south of China. In addition, there are many high-rise buildings with large displacements. Figure 2 shows part of the DI and DII datasets with the building changes labeled by vectorized polygons. Figure 2a is the DI dataset, Figure 2b,c belong to the DII dataset. The zoomed-in images in the right of Figure 2 show that the buildings in dataset DII are more diverse than those in dataset DI. In addition, dataset DII has more high-rise buildings with unavoidable displacement, which is what we were focusing on all along.



(a)

Figure 2. Cont.

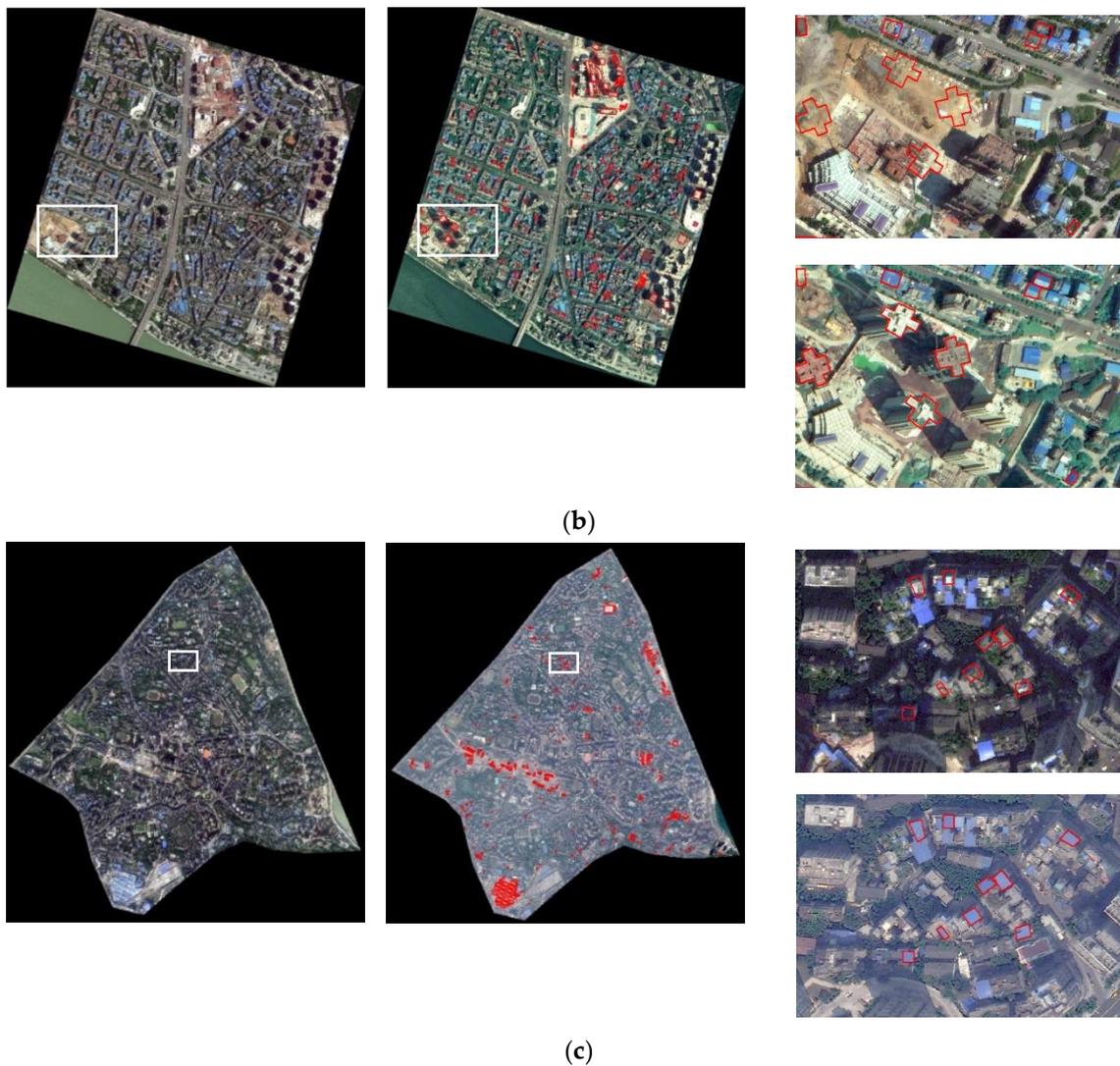


Figure 2. The images in dataset DI and dataset DII. In (a), the first two rows show the two scenes in the same area. In (b,c), the first two columns show the two scenes in the same area. The changed buildings are marked with a red polygon in the right images. The last column is the zoomed-in view of area selected by the white box.

For the two datasets, we divided all the images into tiles of 512×512 pixels, with both the overlay of the width and height being 200 pixels, and finally split each of the two datasets into training sets, validation sets, and test sets randomly in a ratio of 7:1:2. Table 1 shows the general information of the two datasets used in our experiments, including the ground sample distance (GSD), source, pixel size of the tile, and the image number to training, validation, and test sets.

Table 1. Information of our datasets for experiments.

Datasets	GSD (Meters)	Source	Size (Pixels)	Number (Tiles) (Training/Validation/Test)
DI(WHU)	0.075	Aerial	512×512	691/97/199
DII(EV-CD)	0.2–2	Satellite	512×512	1225/175/350

2.2. Methods

2.2.1. Problem Description

PGA-SiamNet was constructed as a Siamese network, with an encoder–decoder structure. The co-attention module setting at the end of the encoder learns the correlation between the deep features of the input image pair; this enables the PGA-SiamNet to find the objects with displacement in other images, which is vital to building change detection. The pyramid change module helps the network to discover the object changes of various sizes and so give better results. Specifically, context information is important for the object in complex scenes, thus aggregating long-range contextual information is useful to improve the feature representation for building change detection.

2.2.2. Architecture Overview

The proposed building change detection network is a Siamese network, following the encoder–decoder architecture as shown in Figure 3. In particular, we employed the well-known VGG16 as the backbone to encode the features of the image pairs to be detected, with each branch sharing the weight. We built a network with the change residual (CR) module for the two-input feature but without any attention mechanism as the baseline; this can be seen in the blue dashed box; the yellow box is the change residual (CR) module.

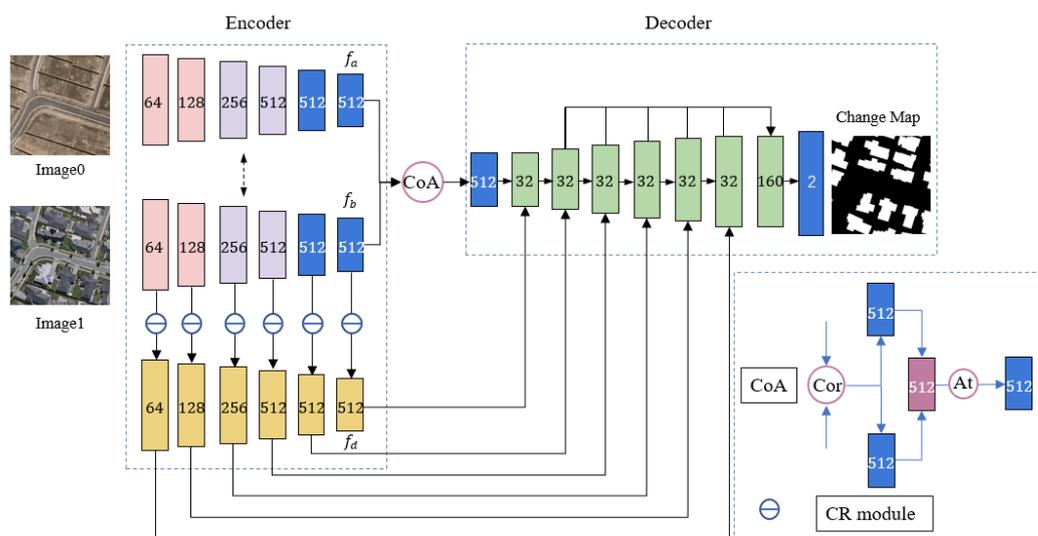


Figure 3. Overview of pyramid feature-based attention-guided Siamese network (PGA-SiamNet Network). CoA indicates the co-attention module.

Thereafter, we introduced attention modules to enrich the CR module. For example, by increasing the receptive field to extract a different scale of feature information, we applied an atrous spatial pyramid pooling (ASPP) module for the deepest level feature of the encoder. We conducted ablation studies for comparison by modifying our network with proposed modules, as is discussed in Section 3.1. To emphasize the useful information of the deep features with 512 channels, the channelwise attention is used for layer 4, 5, and 6. Similarly, the shallow features with rich position information are optimized with spatial attention. Finally, a co-layer aggregation (CLA) module was used to aggregate the low-level and high-level features, thus fusing the semantic and the context information.

2.2.3. Co-Attention Module

The first module of PGA-SiamNet is a co-attention block with an elegant differentiable attention based on a correlation network, which takes in deep feature representations from an image pair as inputs and outputs for a correlation map. If the image pairs contain common objects and therefore

belong to the unchanged category, the features at the locations of the shared objects exhibit similar characteristics. Therefore, inspired by the co-attention mechanism, which discriminates objects from video, the co-attention block was added to the proposed network to identify the changes.

The neighborhood consensus module [61] is used to obtain correlations of the two given features f_a and f_b , because it has already been applied and achieved superior performance in previous research [59]. COSNet [62] proposed another method which uses an affinity matrix to denote co-attention, and so to mine the correlations through adding a weight matrix and verifying the proposed three matrix styles by experiments. In the paper, the co-attention style is exploited to obtain the correlation map like COSNet, which is showed in Figure 4 with the blue dashed box. The correlation map, referred to as affinity matrix $S \in R^{(h \times w) \times (h \times w)}$ between f_a and f_b , is derived from

$$S = f_b^T W f_a \quad (1)$$

$$W = P^{-1} D P \quad (2)$$

where $f_a \in R^{C \times (h \times w)}$ and $f_b \in R^{C \times (h \times w)}$ are the features of the input image pair, the two features are obtained by the encoder of the network, $W \in R^{C \times C}$ is a weight matrix, P is an invertible matrix, and D is a diagonal matrix, and h and w indicate the height and width of the input features, respectively. Then, a softmax function was used to normalize S by column-wise and row-wise.

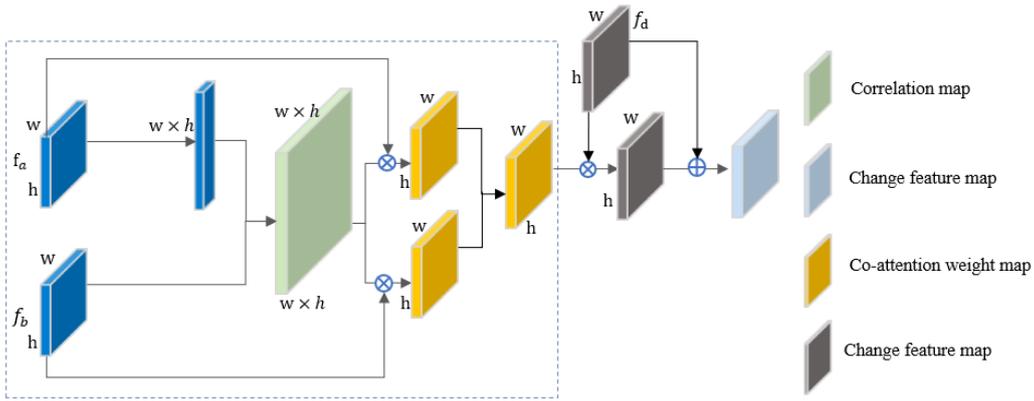


Figure 4. Overview of changed feature map produced by co-attention module.

$$S^c = \text{softmax}(S), \quad S^r = \text{softmax}(S^T) \quad (3)$$

$$\text{softmax}(S^i) = \frac{e^{S^i}}{\sum_i e^{S^i}} \quad (4)$$

where $\text{softmax}(\cdot)$ is to normalize the correlation map S ; $S^c \in R^{h \times w}$ and $S^r \in R^{h \times w}$ are the normalization of S by column-wise and row-wise, respectively. S^c represents the relevance of each feature in f_a to the feature in f_b ; Similarly, S^r is the relevance of each feature in f_b to the feature in f_a . S^i is the i -th feature of the S .

Then the input feature f_a and f_b can be computed as follows:

$$f'_a = f_a \otimes S^c = [f_a^1, f_a^2, \dots, f_a^i, f_a^{h \times w}] \in R^c \times (h \times w) \quad (5)$$

$$f'_b = f_b \otimes S^r = [f_b^1, f_b^2, \dots, f_b^i, f_b^{h \times w}] \in R^c \times (h \times w) \quad (6)$$

where f_a^i, f_b^i denote the i -th column of f'_a and f'_b , and the operator \otimes represents elementwise multiplication.

Furthermore, an attention gate was followed to weight the information of the pair features. The gate is composed of one convolution layer with the kernel size being 1.

In the end, the features are concatenated and fed to a multi-layer perceptron (MLP) to obtain a new representation about the correlation map. To avoid parameter excessive, the MLP is composed of three convolution layers, and the kernel size of each layer is 1, 3, and 1, respectively. After obtaining the common object from the two inputs of the correlation map by MLP, a linear transformation is used to compute the change information. In short, the changed feature f'_d is calculated as follows:

$$f = \text{concat}(f'_a, f'_b) \tag{7}$$

$$A(f) = \sigma(\text{MLP}(f)) = \sigma(W_2(W_1(W_0(f)))) \tag{8}$$

$$f'_d = (1 + A(f)) \times f_d \tag{9}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10}$$

where σ denotes the sigmoid function, $W_0 \in R^{1 \times 1 \times C \times C/r}$, $W_1 \in R^{3 \times 3 \times C/r \times C/r}$, and $W_2 \in R^{1 \times 1 \times C/r \times C/r}$, r is the reduction ratio of the channel and equals two in the paper, f_d represents the output of the change residual (CR) module. Note that, before input to the MLP, the feature f should be normalized to [0,1] by a sigmoid function. Figure 4 depicts the computation process of the changed feature with co-attention module and the module is showed in the blue dashed box. Detailed experiments were conducted to compare the effects of the module during the ablation studies.

2.2.4. Co-Layer Aggregation Module

Recent studies have shown that the high-level layers encode abundant contextual and global information but lost fine spatial information, while the low-level layers are the opposite. Therefore, by adopting layer aggregation to merge the different level features with various details, a good performance may be obtained [36,63,64]. In this paper, we added a co-layer aggregation (CLA) module to the proposed network to weight the low-level features in order to enhance the change information.

The encoder of our model contains six layers, we chose the first three layers: $f_l \in \{f_1, f_2, f_3\}$ as the low-level features, and the last layer weighted by co-attention: $f_h \in \{f_6\}$ as the high-level feature to perform the operation. As show in Figure 5, to merge both spatial and channelwise information, the SE block [38] was firstly applied to both the shallow and the deep features. After being given the transformed features, we forwarded the transformed high-level feature to a global pooling and two convolutions to get a global attention, which can be used to enhance the context representation of the low-level feature.

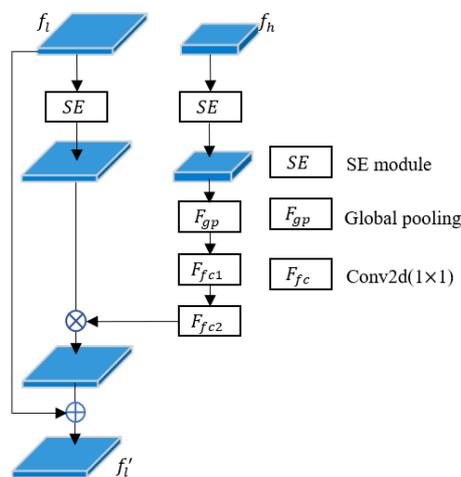


Figure 5. Co-layer aggregation module.

Finally, the origin low feature was added into the enhanced one as a residual block. In this way, the shallow features are refined by correlation if they merge with the changed feature produced by co-attention module.

2.2.5. Pyramid Change Module

To make full use of the effective receptive field at each level, the decoder consists of a pyramid of features $\{f_c^1, f_c^2, \dots, f_c^N\}$, as show in Figure 6, which is designed to find the building changes at different scales in the images. At each scale, the feature from the previous scale is unsampled and added to the changed feature f_d generated from change residual (CR) module. Then, the result is fed into a convolution layer, with the kernel size being 1. After performing the same steps for all the scales, the results from each scale were concatenated together, and then fed into a convolution layer; the output is the change map. Referring to the classic feature pyramid method, FPN (feature pyramid network) [65] iteratively merges features with a top-down mechanism until the resolution of the last layer recover to the original input. We fused the changed feature in a top-down pathway in order to catch the change information with different sizes.

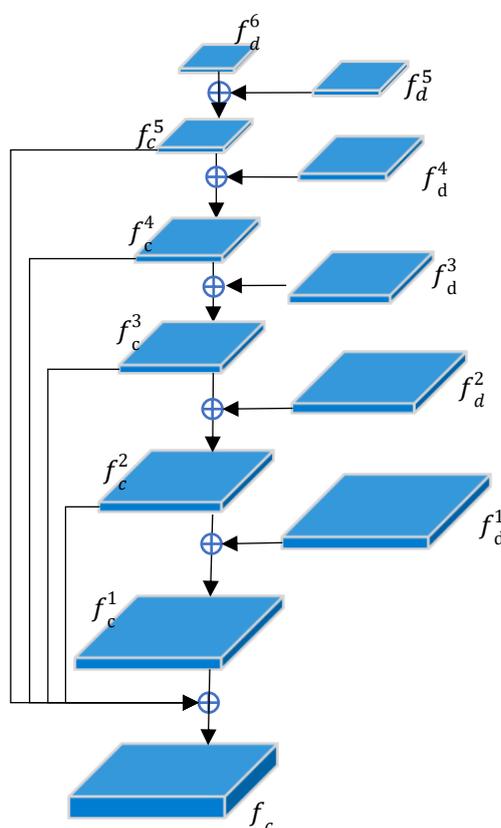


Figure 6. Pyramid change feature decoder.

The CR module is shown in Figure 7. The objective of the module is to obtain the distinctive and discriminative features for the two inputs. As shown in Figure 7c, the generation starts with the input of the two image features f_a and f_b , and learns to produce a difference map for the input features. The module merges with two kinds of fusion strategy: elementwise difference and elementwise addition. The elementwise difference is to get the absolute value of their difference (see Figure 7a) while the elementwise addition is to add the two input features (see Figure 7b). The CR module learns the addition features (see Figure 7b) as the residual counterparts, which are added by the difference feature (see Figure 7a), making the information refinement task easier.

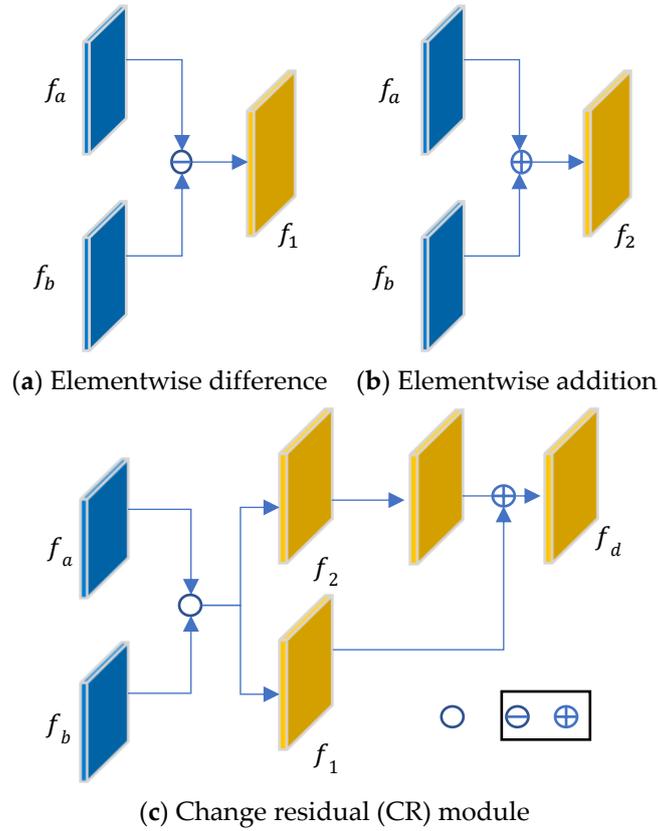


Figure 7. Single scale changed feature generation.

2.3. Implementation Details

The proposed PGA-SiamNet was implemented using Pytorch; the training procedure used a single NVIDIA RTX 2080 Ti GPU with 11 GB memory. We used a mini-batch of two and the initial learning rate was 10^{-4} for the two datasets and decreased linearly according to the number of iteration times. The optimization algorithm to train the network was the adaptive moment estimation (Adam) algorithm [66]. We regarded the task as a binary segmentation for the final output of the network, which is change or no-change. To measure the performance of the proposed network, the metrics intersection over union (IoU), F1 score, precision, recall, and overall accuracy (OA) were used. Generally, the most meaningful metric was IoU in our research. The imbalance in the two classes, changed and un-changed, resulted in a large value of OA. A large number of unchanged pixels have severely made the calculated results obviously too high. Just as OA shows in Section 3, the value is too high when regarding the unchanged pixels, so it is not a good metric to reflect the accuracy of the results. Taking our focus into consideration, the precision, recall, and F1 were only calculated on the changed pixels. The metrics are defined as follows:

$$IoU = \frac{TP}{TP + FP + FN} \quad (11)$$

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

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (15)$$

$$Kappa = \frac{p_0 - p_c}{1 - p_c} \quad (16)$$

$$p_0 = OA \quad (17)$$

$$p_c = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + FP + TN + FN)^2} \quad (18)$$

where true positive (TP) indicates the number of pixels correctly classified as changed buildings, true negative (TN) denotes the number of pixels correctly classified as unchanged buildings, false positive (FP) represents the number of pixels misclassified as changed buildings, and false negative (FN) is the number of pixels misclassified as unchanged buildings. In Equation (15), P and R denotes precision and recall, respectively. During the training period, Z-score standardization was firstly used for the multitemporal image pairs. In addition, we trained a convolution neural network to obtain better results by multiple iterations. The binary cross-entropy loss function is a popular and effective solution, so we minimized it to optimize the network. The loss function is calculated as follows:

$$l = -y \log y' - (1 - y) \log(1 - y') = \begin{cases} -\log y', & y = 1 \\ -\log(1 - y'), & y = 0 \end{cases} \quad (19)$$

where y refers to the ground truth and y' refers to the predicted result.

3. Results

In this section, some ablation studies are provided in Section 3.1. Then, we compared the results of the proposed method with other methods in Section 3.2. Furthermore, the robustness of the proposed algorithm is proved in Section 3.3.

3.1. Ablation Study

In this part, we focus on exploration studies to assess the different components of the network with the two datasets, DI and DII. We trained the proposed baseline network and obtained a satisfactory result, which confirms that our proposed base network is effective. Furthermore, we introduced some attention modules to improve the performance. However, it is difficult to balance the performance on two datasets because of the completely different sensors involved. After a lot of experiments, we verified that the described modules can improve the performance of the network; the results of the metrics are showed in Table 2.

Table 2. Ablation experiments of the methods with different modules (the shadow represents the best results of the basic network with all the proposed modules).

Network	IoU (%)		OA (%)		Recall (%)		Precision (%)		F1(%)		Kappa (%)	
	DI	DII	DI	DII	DI	DII	DI	DII	DI	DII	DI	DII
Baseline	96.52	92.02	99.75	99.63	96.24	90.23	96.89	92.65	96.25	90.67	96.12	90.48
+CS+CLA	97.15	92.13	99.77	99.65	96.62	89.42	97.79	93.55	97.09	90.83	96.97	90.65
+ASPP	97.11	92.52	99.78	99.66	96.91	90.25	97.38	93.83	97.0	91.38	96.88	91.21
+CoA	97.38	92.73	99.79	99.68	97.01	90.59	97.84	94.01	97.29	91.74	97.17	91.57

Line 2 of Table 2 shows that, by adding channelwise and spatialwise attention (CS) and a co-layer aggregation (CLA) module to the base network, we obtained a slight increase in accuracy for the two datasets. After adding the ASPP module to the network, there was a slight decline regarding dataset DI, but a great improvement regarding dataset DII in all areas. The reason for this is that dataset DII consists of results from various sensors and more variable information with multi-scales, while dataset DI is relatively unvaried as regards scales information. Finally, we introduced a co-attention (CoA) module, which is important for the performance when using orthoimages with building displacement. In addition, Line 4 in Table 2 further demonstrates the efficacy.

3.2. Comparisons with Other Methods

To evaluate the performance of the proposed architecture, we further compared our method with other recent change detection methods: a deep architecture for detecting changes (ChangeNet) [67], a correlated Siamese change detection network (CSCDNet) [68], multiple side-output fusion (MSOF) [69], dual task constrained deep Siamese convolutional network (DTCDCN) [70], multi-scale fully convolutional early fusion (MSFC-EF) [31], deep Siamese multi-scale fully convolutional network (DSMS-FCN) [31], fully convolutional early fusion (FC-EF) [24], fully convolutional Siamese-difference (FC-Siam-Diff) [24], and fully convolutional Siamese-concatenation (FC-Siam-Conc) [24]. CSCDNet was proposed to train a semantic change detection network with a Streetview dataset; by inserting correlation layers into the network, it can overcome the limitation caused by camera viewpoints, which is a major problem to the end-to-end building change detection task. In this paper, to validate the effect of the layer, we give the comparison of the network with and without the correlation layer, denoted CSCDNet/w and CSCDNet/wo, respectively. ‘FC-’ is a series of full convolution network and the performance is improved through extracting multiscale features in the decoder, such as MSFC-EF and DSMS-FCN. For a fair comparison, we trained and tested our PGA-SiamNet and other methods with the two available datasets mentioned above and the same parameter settings. The results are shown in Table 3. One thing to be mentioned is that some of the comparison methods are completed with the semantic task, we only took the change detection network as the comparison due to a lack of semantic labels. The results show that PGA-SiamNet easily outperforms other approaches. Simultaneously, the visualized results of the proposed method and the other methods were also compared, as shown in Figure 8, which contains the first three image pairs are from the DI dataset and the last four pairs are from DII dataset.

Table 3. Comparison with other related methods (the lighter shadow represents the best results of other related methods; the darker shadow represents the results of our proposed method).

Network	IoU (%)		OA (%)		Recall (%)		Precision (%)		F1(%)		Kappa (%)	
	DI	DII	DI	DII	DI	DII	DI	DII	DI	DII	DI	DII
ChangeNet	70.80	56.19	96.88	97.44	52.97	21.39	66.99	32.48	57.48	23.46	55.79	22.41
MSOF	90.84	82.66	99.08	99.20	88.68	71.55	92.45	89.12	89.40	78.20	88.92	77.81
DTCDCN	83.55	78.67	98.61	99.11	80.44	64.74	78.28	84.55	78.22	71.2	77.45	70.77
CSCDNet/w	95.04	87.91	99.63	99.49	94.03	83.15	95.96	90.19	94.66	85.69	94.45	85.43
CSCDNet/wo	94.68	87.53	99.63	99.45	93.74	81.38	95.63	91.19	94.09	85.13	93.89	84.85
FC-EF	78.70	67.24	97.98	98.36	71.24	47.71	74.81	57.67	71.43	50.03	70.33	49.26
FC-Siam-Diff	88.66	80.5	99.0	99.1	85.67	71.73	88.89	79.95	86.11	74.15	85.58	73.71
FC-Siam-Con	82.08	68.02	98.4	98.43	74.67	47.76	88.72	60.85	76.57	51.47	75.69	50.73
MSFC-EF	90.72	83.65	99.26	99.29	88.54	79.97	90.31	87.51	88.72	79.69	88.30	79.33
DSMS-FCN	88.61	83.37	99.12	99.25	88.29	73.01	86.32	89.35	86.09	79.18	85.62	78.81
Baseline(ours)	96.52	92.02	99.75	99.63	96.24	90.23	96.89	92.65	96.25	90.67	96.12	90.48
PGA-SiamNet	97.38	92.73	99.79	99.68	97.01	90.59	97.84	94.01	97.29	91.74	97.17	91.57

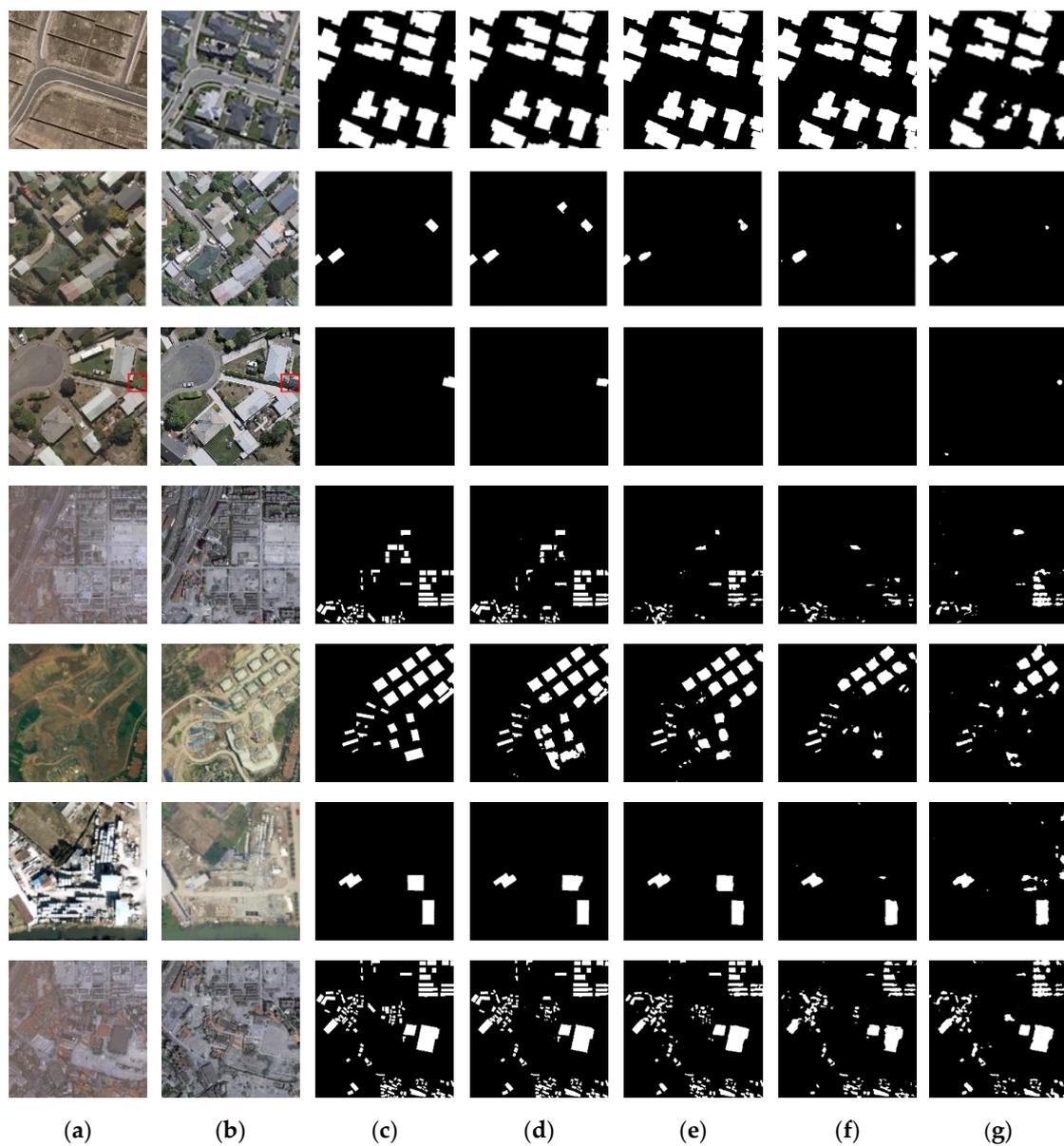


Figure 8. Results by our proposed method, and comparisons with others. (a) Before images; (b) after images; (c) ground truth; (d) PGA-SiamNet; (e) CSCDNet; (f) MSOF; (g) FC-Siam-Diff.

3.3. Robustness of the Method

To prove the robustness of the proposed algorithm, we tested on other orthoimages with the model trained on EV-CD building dataset. The image pairs are different with our training samples because they are located in the north of China. As shown in Figure 9, the acceptable result shows that the proposed method has a great potential for the high-resolution remote sensing orthoimages from various sensors with more training samples.

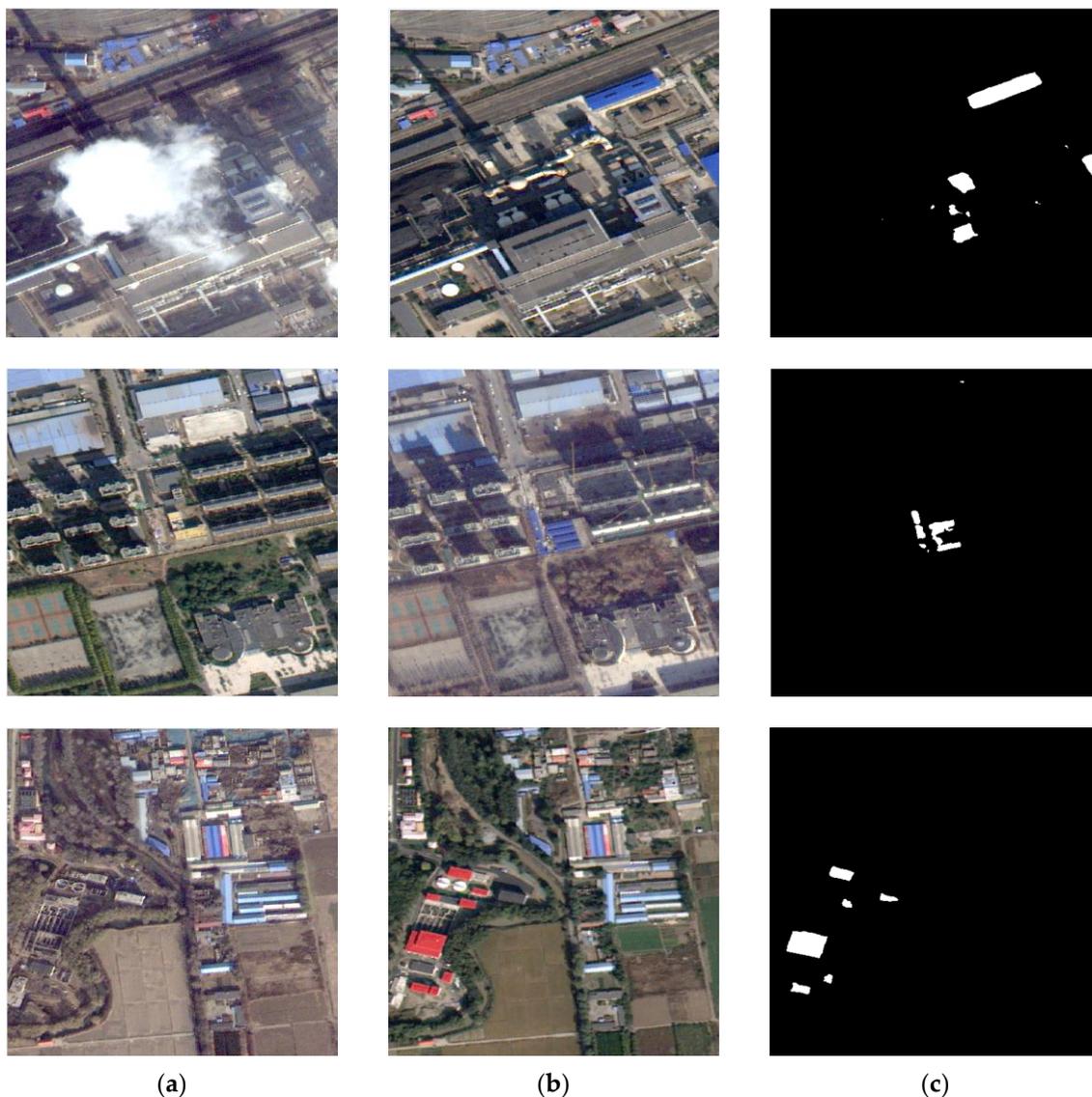


Figure 9. Results on other image pairs. (a) Before images; (b) after images; (c) results of the proposed method.

4. Discussion

The main goal of the study was to find changes of the buildings on high-resolution remote sensing orthoimages automatically. We first trained an end-to-end framework on the two available datasets. Then, representation of the features was enhanced by attention modules by fusing global context features and local semantic features. Meanwhile, the correlation of the image pairs was considered in the network by co-attention module and co-layer aggregation. Finally, the proposed method obtained better results in our studies.

4.1. Importance of the Proposed Dataset

Change detection, as a hot topic in the field of remote sensing, has attracted extensive attention and emerged many related datasets. Buildings, as an important man-made object, are often in the spotlight. However, only the public WHU building datasets provide changes for buildings at present. As for satellite imagery, there are almost no available datasets for building change detection. The difficulty of building change detection is the displacement of high-rise buildings. Therefore, we built a building change detection dataset (EV-CD building datasets) with the existing satellite images.

In WHU building datasets, the buildings are low with little displacement. Besides this, the buildings are mostly independent in WHU building datasets. However, in the complex cities, the buildings are often distributed densely. Since the focus of our study is on the complex cities, it is necessary to build a relevant dataset to promote the research. The experiments show that the dataset is effective. In addition, we will publish the dataset and enlarge it in the future.

4.2. Advantages of the Proposed Baseline

The results of the experiments show that the proposed baseline network is effective and surpasses other methods with the two datasets. There are several advantages of the proposed baseline. We found that adjusting the learning rate according to the number of iteration times is a better way after performing many experiments, and the networks obtain a better performance with a pretrained weight. Given that there are a variety of buildings with different size in the datasets, we took the pyramid feature into consideration to discover the changes with various sizes in the proposed baseline network. Inspired by ResNet [71], the proposed change residual (CR) module is also a key part to get a changed feature using a residual structure. The module can fuse the features from different sources without degradation. In the decoder, the features from each scale are concatenated together. In this way, our network detects the building area accurately.

4.3. Experimental Results Compared with Other Methods

Some of the networks mentioned above are proposed for street view change detection, such as ChangeNet and CSCDNet. In ChangeNet, the branch of the Siamese contains convolutional neural networks and deconvolutional neural networks. The weights of the two branches are shared and fixed in convolutional neural networks. In the deconvolutional neural networks, the weights of the layers are not fixed. The ChangeNet only concatenates three changed features produced by the Siamese feature extraction network to incorporate both coarse and finer details. However, only combining some outputs of the decoder, the relevance of the two channels is not enough to detect changes with remote sensing images. CSCDNet achieved a better result with a slightly lower accuracy than our baseline, especially with the correlation layer. The correlation layer was utilized to deal with difference in camera viewpoints like the displacement on remote sensing images, but the correlation layer is very time consuming. The architecture of our model is somewhat similar to CSCDNet. We got more change information by CR module and applied a co-attention module to obtain the correlation of the two input features instead of the time-consuming correlation layer. Then a co-layer aggregation module was used to fuse changed features extracted from co-attention module to the shallow features. The aggregated features further improved the representation of the features in the pyramid. DTCDSN was proposed to complete both change detection and semantic segmentation at the same time, while we only employed the subnetwork of change detection. The model has shallower convolution layers than other models, and it may ignore the multi-scale change. In addition, the proposed improved network increases the receptive field to extract a different scale changed feature using the ASPP module, and it is helpful for building change detection in complex cities such as the EV-CD building dataset. Overall, the models trained with pretrained weight are superior to those without pretrained weight, such as FC-Siam-Diff and DSMS-FCN, in our studies. Finally, an addition experiment demonstrated that the proposed method has a strong robustness by testing on other orthoimages with the model trained on EV-CD building dataset.

5. Conclusions

In this paper, we proposed an end-to-end attention-guided Siamese network based on a pyramid feature (PGA-SiamNet) network. It performed excellently on remote sensing orthoimagery for building change detection and yielded better results on complex urban environments when compared with other methods. By using a co-attention mechanism, the method learns to discriminate feature change by capturing the correlation between image pairs. To obtain long-range dependencies effectively, we

adopted an attention-guided method. Our experiments on the two available datasets show that our method gives comparable results to other state-of-the-art techniques. Therefore, the modules added to this framework are both independent and can be adopted for building change detection in a convenient way. Meanwhile, the experimental results with the WHU datasets show a better performance than the results with the satellite imagery EV-CD dataset; the complexity and diversity of the scenes contribute to this result, while this type of data is more in our research. Owing to the machine-learning boom, building extraction, which used to be the central problem of traditional building change detection, has become unnecessary. However, the need for large and accurate sample data is still the main concern for deep learning, thus data-independent research is increasingly important, since most data at hand is inadequate and noisy. As regards future studies, on the one hand, we may place more attention on noisy data and one-shot/few-shot learning; on the other hand, it is possible to imagine involving more diverse information, such as using auxiliary DSM information as an object guide, as well as mining more information from the current data.

6. Patents

At present, we are applying for the patent based on the research results of this paper and the application material has been submitted to China National Intellectual Property Administration (the patent application number is 2020100445918). Moreover, we are waiting for the examination and grant of this patent.

Author Contributions: H.J. conceived, conducted, and improved the experiments, and they also wrote the manuscript. X.H. directed and revised this manuscript. K.L. assisted in the experimental verification and revised the manuscript. J.Z. and J.G. labeled the datasets and reviewed the manuscript. M.Z. revised the manuscript. All authors have read and agreed to publish the manuscript.

Funding: This research was partially supported by National Natural Science Foundation of China, grant number 41771363; the research funding was from Guangzhou Science, Technology and Innovation Commission (201802030008).

Acknowledgments: The authors sincerely appreciate that academic editors and reviewers give their helpful comments and constructive suggestions.

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

References

1. Aleksandrowicz, S.; Turlej, K.; Lewiński, S.; Bochenek, Z. Change Detection Algorithm for the Production of Land Cover Change Maps over the European Union Countries. *Remote Sens.* **2014**, *6*, 5976–5994. [CrossRef]
2. Earth Watching. Available online: <https://earth.esa.int/web/earth-watching/change-detection> (accessed on 25 January 2019).
3. Onera Satellite Change Detection. Available online: <http://dase.grss-ieee.org> (accessed on 10 May 2019).
4. Champion, N. 2D Building Change Detection from High Resolution Aerial Images and Correlation Digital Surface Models. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2007**, *36*, 197–202.
5. Cleve, C.; Kelly, M.; Kearns, F.R.; Moritz, M. Classification of the wildland–urban interface: A comparison of pixel- and object-based classifications using high-resolution aerial photography. *Comput. Environ. Urban Syst.* **2008**, *32*, 317–326. [CrossRef]
6. Hussain, M.; Chen, D.; Cheng, A.; Wei, H.; Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS J. Photogramm.* **2013**, *80*, 91–106. [CrossRef]
7. Huang, X.; Zhang, L.; Zhu, T. Building Change Detection from Multitemporal High-Resolution Remotely Sensed Images Based on a Morphological Building Index. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 105–115. [CrossRef]
8. Im, J.; Jensen, J.R.; Tullis, J.A. Object-based change detection using correlation image analysis and image segmentation. *Int. J. Remote Sens.* **2008**, *29*, 399–423. [CrossRef]
9. Bouziani, M.; Goita, K.; He, D.-C. Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge. *ISPRS J. Photogramm.* **2010**, *65*, 143–153. [CrossRef]

10. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis-Towards a new paradigm. *ISPRS J. Photogramm.* **2014**, *87*, 180–191. [[CrossRef](#)]
11. Zhan, Y.; Fu, K.; Yan, M.; Sun, X.; Wang, H.; Qiu, X. Change Detection Based on Deep Siamese Convolutional Network for Optical Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1845–1849. [[CrossRef](#)]
12. Qin, R.; Tian, J.; Reinartz, P. 3D change detection—Approaches and applications. *ISPRS J. Photogramm.* **2016**, *122*, 41–56. [[CrossRef](#)]
13. Tian, J.; Qin, R.; Cerra, D.; Reinartz, P. Building Change Detection in Very High Resolution Satellite Stereo Image Time Series. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *III-7*, 149–155. [[CrossRef](#)]
14. Malpica, J.A.; Alonso, M.C.; Papí, F.; Arozarena, A.; Martínez De Agirre, A. Change detection of buildings from satellite imagery and lidar data. *Int. J. Remote Sens.* **2012**, *34*, 1652–1675. [[CrossRef](#)]
15. Peng, D.; Zhang, Y. Building Change Detection by Combining Lidar data and Ortho Image. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *XLI-B3*, 669–676. [[CrossRef](#)]
16. Remondino, F.; Spera, M.G.; Nocerino, E.; Menna, F.; Nex, F. State of the art in high density image matching. *Photogramm. Rec.* **2014**, *29*, 144–166. [[CrossRef](#)]
17. Tian, J.; Cui, S.; Reinartz, P. Building Change Detection Based on Satellite Stereo Imagery and Digital Surface Models. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 406–417. [[CrossRef](#)]
18. Yang, J.; Price, B.; Cohen, S. Object contour detection with a fully convolutional encoder-decoder network. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
19. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C. Residual Attention Network for Image Classification. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
20. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
21. Zhang, Z.; Vosselman, G.; Gerke, M.; Tuia, D.; Yang, M.Y. Change Detection between Multimodal Remote Sensing Data Using Siamese CNN. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
22. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [[CrossRef](#)]
23. Lim, K.S.; Jin, D.K.; Kim, C.S. Change Detection in High Resolution Satellite Images Using an Ensemble of Convolutional Neural Networks. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018.
24. Daudt, R.C.; Saux, B.L.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
25. Lebedev, M.A.; Vizilter, Y.V.; Vygolov, O.V.; Knyaz, V.A.; Rubis, A.Y. Change Detection in Remote Sensing Images Using Conditional Adversarial Networks. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *XLII-2*, 565–571. [[CrossRef](#)]
26. Khan, S.H.; He, X.; Bennamoun, M.; Porikli, F.; Sohel, F.; Togneri, R. Weakly Supervised Change Detection in a Pair of Images. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
27. Khan, S.H.; He, X.; Porikli, F.; Bennamoun, M.; Sohel, F.; Togneri, R. Learning deep structured network for weakly supervised change detection. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2016.
28. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Guided Anisotropic Diffusion and Iterative Learning for Weakly Supervised Change Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
29. Jong, K.L.D.; Bosman, A.S. Unsupervised Change Detection in Satellite Images Using Convolutional Neural Networks. Available online: <https://arxiv.org/abs/1812.05815?context=cs.NE> (accessed on 22 February 2019).
30. Yang, M.; Jiao, L.; Liu, F.; Hou, B.; Yang, S. Transferred Deep Learning-Based Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6960–6973. [[CrossRef](#)]

31. Chen, H.; Wu, C.; Du, B.; Zhang, L. Deep Siamese Multi-scale Convolutional Network for Change Detection in Multi-temporal VHR Images. Available online: <https://arxiv.org/abs/1906.11479> (accessed on 1 July 2019).
32. Rocco, I.; Arandjelović, R.; Sivic, J. End-to-end weakly-supervised semantic alignment. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
33. Kanazawa, A.; Jacobs, D.W.; Chandraker, M. WarpNet: Weakly Supervised Matching for Single-View Reconstruction. Available online: <https://arxiv.org/abs/1604.05592> (accessed on 18 September 2019).
34. Huang, S.; Wang, Q.; Zhang, S.; Yan, S.; He, X. Dynamic Context Correspondence Network for Semantic Alignment. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
35. Wang, F.; Tax, D.M.J. Survey on the Attention Based RNN Model and Its Applications in Computer Vision. Available online: <https://arxiv.org/abs/1601.06823> (accessed on 18 October 2019).
36. Woo, S.; Park, J.; Lee, J.; Kweon, I. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (In European Conference on Computer Vision (ECCV)), Munich, Germany, 8–14 August 2018.
37. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
38. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. PSANet: Point-wise Spatial Attention Network for Scene Parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 August 2018.
39. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. Available online: <https://arxiv.org/abs/1805.10180> (accessed on 18 April 2019).
40. Lu, X.; Wang, W.; Ma, C.; Shen, J.; Shao, L.; Porikli, F.M. See More, Know More: Unsupervised Video Object Segmentation with Co-Attention Siamese Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
41. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
42. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018.
43. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
44. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context Encoding for Semantic Segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
45. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
46. Xiong, C.; Zhong, V.; Socher, R. Dynamic Coattention Networks for Question Answering. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2016.
47. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep Modular Co-Attention Networks for Visual Question Answering. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
48. Nguyen, D.-K.; Okatani, T. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
49. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016.
50. Xiao, P.; Yuan, M.; Zhang, X.; Feng, X.; Guo, Y. Cosegmentation for Object-Based Building Change Detection from High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 1587–1603. [[CrossRef](#)]

51. Rahman, F.; Vasu, B.; Cor, J.V.; Kerekes, J.; Savakis, A. Siamese Network with Multi-Level Features for Patch-Based Change Detection in Satellite Imagery. In Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, 26–29 November 2018; pp. 958–962.
52. Daudt, R.C.; Saux, B.L.; Boulch, A.; Gousseau, Y. Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 2115–2118. [[CrossRef](#)]
53. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
54. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; pp. 886–893. [[CrossRef](#)]
55. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
56. Choy, C.B.; Gwak, J.; Savarese, S.; Chandraker, M. Universal Correspondence Network. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016.
57. Moo Yi, K.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; Fua, P. Learning to Find Good Correspondences. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
58. Chen, Y.-C.; Huang, P.-H.; Yu, L.-Y.; Huang, J.-B.; Yang, M.-H.; Lin, Y.-Y. Deep Semantic Matching with Foreground Detection and Cycle-Consistency. In Proceedings of the 14th Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018; pp. 347–362.
59. Rocco, I.; Arandjelović, R.; Sivic, J. Convolutional neural network architecture for geometric matching. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
60. 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](#)]
61. Rocco, I.; Cimpoi, M.; Arandjelović, R.; Torii, A.; Pajdla, T.; Sivic, J. Neighbourhood Consensus Networks. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–5 December 2018.
62. Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; Huang, J.-B. Show, Match and Segment: Joint Learning of Semantic Matching and Object Co-Segmentation. Available online: <https://arxiv.org/abs/1906.05857?context=cs.CV> (accessed on 15 September 2019).
63. Zhang, C.; Cao, Z.-G.; Xiong, X.; Xian, K.; Qi, X. Salient Object Detection via Deep Hierarchical Context Aggregation and Multi-Layer Supervision. In Proceedings of the IEEE International Conference on Image Processing (ICIP) 2019, Taiwan, China, 22–25 September 2019.
64. Liu, Y.; Qiu, Y.; Zhang, L.; Bian, J.; Nie, G.-Y.; Cheng, M.-M. Salient Object Detection via High-to-Low Hierarchical Context Aggregation. Available online: <https://arxiv.org/abs/1812.10956> (accessed on 25 May 2019).
65. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017.
66. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AL, Canada, 14–16 April 2014.
67. Varghese, A.; Gubbi, J.; Ramaswamy, A.; Balamuralidhar, P. ChangeNet: A Deep Learning Architecture for Visual Change Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
68. Sakurada, K. Weakly Supervised Silhouette-based Semantic Change Detection. Available online: <https://arxiv.org/abs/1811.11985v1> (accessed on 25 June 2019).
69. Peng, D.; Zhang, M.; Wanbing, G. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [[CrossRef](#)]

70. 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. Available online: <https://arxiv.org/abs/1909.07726?context=cs.CV> (accessed on 18 October 2019).
71. He, K.; Zhang, J.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).