High-resolution triplet network with dynamic multiscale feature for change detection on satellite images

Hou, Xuan; Bai, Yunpeng; Li, Ying; Shang, Changjing; Shen, Qiang

Published in:
ISPRS Journal of Photogrammetry and Remote Sensing

DOI:
10.1016/j.isprsjprs.2021.05.001

Publication date:
2021

Citation for published version (APA):
Hou, X., Bai, Y., Li, Y., Shang, C., & Shen, Q. (2021). High-resolution triplet network with dynamic multiscale feature for change detection on satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *177*, 103-115. https://doi.org/10.1016/j.isprsjprs.2021.05.001

HIGH-RESOLUTION TRIPLET NETWORK WITH DYNAMIC MULTISCALE FEATURE FOR CHANGE DETECTION ON SATELLITE IMAGES

Xuan Hou[a], Yunpeng Bai[b] , Ying Li[a,*], Changjing Shang[b], Qiang Shen[b]

[a] School of Computer Science, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech & Image Information Processing, Northwestern Polytechnical University, Xi'an 710129, China
[b] Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth SY23 3DB, U.K.

KEY WORDS: Change detection, triplet network, high-resolution images, dynamic convolution, remote sensing

ABSTRACT:

Change detection in remote sensing images aims to accurately determine any significant land surface changes based on acquired multi-temporal image data, being a pivotal task of remote sensing image processing. Over the past few years, owing to its powerful learning and expression ability, deep learning has been widely applied in the general field of image processing and has demonstrated remarkable potentials in performing change detection in images. However, a majority of the existing deep learning-based change detection mechanisms are modified from single-image semantic segmentation algorithms, without considering the temporal information contained within the images, thereby not always appropriate for real-world change detection. This paper proposes a High-Resolution Triplet Network (HRTNet) framework, including a dynamic inception module, to tackle such shortcomings in change detection. First, a novel triplet input network is introduced, which is capable of learning bi-temporal image features, extracting the temporal information reflecting the difference between images over time. Then, a network is employed to extract high-resolution image features, ensuring the learned features preserving high-resolution characteristics with minimal reduction of information. The paper also proposes a novel dynamic inception module, which helps improve the feature expression ability of HRTNet, enriching the multi-scale information of the features extracted. Finally, the distances between feature pairs are measured to generate a high-precision change map. The effectiveness and robustness of HRTNet are verified on three popular high-resolution remote sensing image datasets. Systematic experimental results show that the proposed approach outperforms state-of-the-art change detection methods.

1. INTRODUCTION

Change detection (CD) in remote sensing images is a technology that relies on the analysis of the spectral information provided by remote sensing data to detect and extract the information of land surface changes (Singh, 1989). While remote sensing has been utilized as a major method for obtaining information in various applied fields, CD forms an important underlying task of processing remote sensed images. Indeed, remote sensing image CD has been widely applied to resolving various problems, including: disaster assessment, land management, resource management and urban expansion research (Jin et al., 2013, Mundia and Aniya, 2005, Brunner et al., 2010, Wang and Xu, 2010). Of course, generally speaking, different applications require the identification of different types of changing target. For example, land management needs to identify changes in land use and cover, resource management needs to identify changes in forests and vegetation, and urban expansion research needs to identify changes in buildings.

With the development of satellite imaging technology, the resolution of remote sensing images is becoming higher and higher. Particularly, the surface information in high-resolution images becomes more abundant and diverse. Therefore, high-resolution remote sensing images form a useful data for CD (Bruzzone and Bovolo, 2012). One current research issue on CD is how to effectively learn the rich feature representation in remote sensing images, while reducing the interference of pseudo-changes caused by atmospheric characteristics, sunshine, seasons, etc., in order to obtain a robust high-resolution satellite image CD method.

Traditional CD techniques tend to use clustering or threshold segmentation (Hussain et al., 2013, Wu et al., 2017, Celik, 2009) to process image differences to determine any changed regions or unchanged regions. Such methods mainly rely on handcrafted features, which are inefficient and are generally of poor robustness (El Amin et al., 2017). In recent years, having recognized the excellent ability of deep learning models such as convolutional neural networks (CNNs) for capturing and expressing informative image features, they have been widely used in various tasks of computer vision and image processing (Voulodimos et al., 2018, Li et al., 2018, Li et al., 2017, Zhang et al., 2019). As a classic semantic segmentation CNN, for instance, the fully convolutional network (FCN) (Long et al., 2015) has been applied to performing the CD task of remote sensing images. In sharp contrast with traditional CD techniques, FCN-based CD exploits spatial context information and does not require manually designed features, thereby offering stronger generalization capacity and better robustness (Alcantarilla et al., 2018, Zhan et al., 2017, Lei et al., 2019) However, existing CD methods, including deep learning based ones, still have the following shortcomings: 1) As the resolution of remote sensing images becomes higher and higher, they do not make full use of the rich information contained in high-resolution images. This leads to the CD methods being unable to sensitively distinguish pseudo-changes such as angle, climate, and sunshine. 2) The information on the edges within a change map of high-resolution remote sensing images learned by a CNN is often not ideal. Unlike low-resolution remote sensing images, the changing regions of high-resolution images usually contain significantly more information on the edges, which should be exploited. 3) Temporal information contained within bi-temporal remote sensing images is not utilized, which if used would benefit the CD performance.

In order to address the aforementioned limitations, a High-Resolution

[*] Corresponding author, with School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China. E-mail address lybyp@nwpu.edu.cn

Triplet Network (HRTNet) framework for CD is proposed, acting upon high-resolution remote sensing images. It is a three-branch network with its input consisting of bi-temporal remote sensing images as well as the difference images generated by directly comparing pixel and textual values between bi-temporal images. This differs from the previous CD approaches based on traditional FCN (Alcantarilla et al., 2018, Daudt et al., 2018, Peng et al., 2019) whose input involves the difference image (DI) or the concatenate image of bi-temporal images based on channel dimension only, thereby reducing information loss at the earliest stage possible. Also, the proposed method is different from the existing change detection approach based on the use of a Siamese network (Chen et al., 2020, Liu et al., 2020b, Chen et al., 2021). The latter only has two bi-temporal images as input, while an additional DI is taken as part of the input to achieve the purpose of learning the change information. Furthermore, a high-resolution network is employed to reduce the loss of information during the process of down-sampling feature learning. In HRTNet, a dynamic inception module is utilized to enhance the ability of representing multi-scale features, making the overall model more sensitive to change regions of different sizes. Finally, the temporal features contained within the DI are exploited to ensure the model paying more attention in surface changes, reinforcing its robustness for the recognition of pseudo-changes. Based on the evaluation over three popular high-resolution remote sensing image datasets, HRTNet is shown to be able to achieve better performance than other algorithms for remote sensing image CD (Daudt et al., 2018, Peng et al., 2019, Chen et al., 2021, Zhang et al., 2020, Chen and Shi, 2020). The objective of this proposed framework is to identify the change information of interest in a specified application and to filter out irrelevant change information as interference factors. It can be applied for change detection with different purposes, as illustrated with the different tasks performed in the experimental investigations. The major contributions of this paper are as follows:

1. In view of the lack of full use of temporal information in existing CD methods, a triplet input network is proposed to learn and exploit temporal information. The robustness of the model in detecting pseudo-change and avoiding the effect of the noise in bi-temporal images is improved.
2. A high-resolution triplet network architecture is devised to capture and represent high-resolution remote sensing image features. The information loss of an input image is reduced using this three-branch network structure. High-resolution feature extraction also enables the network to reduce the loss of information when down-sampling high-resolution remote sensing images, generating a wealth of useful image features.
3. A dynamic inception module (DIM) is presented to enhance the comprehensiveness and expressiveness of the resulting model, making it capable of recognizing change regions of different scales. This in turn, helps detect objects of a different scale successfully.

The rest of this article is structured as follows: Section 2 presents an overview of related work. Section 3 describes the proposed approach. Section 4 evaluates the effectiveness of the proposed algorithm through systematic experiments. Section 5 summarizes the main work of this paper and briefly discusses important further research.

2. RELATE WORK

2.1 Change detection

In this section, existing CD methods are introduced with respect to two types of distinct approach: traditional CD and deep learning based CD.

Traditional CD algorithms can be divided into two groups: pixel-based and object-based (Hussain et al., 2013). A pixel-based CD method uses mechanisms such as clustering or threshold segmentation to segment difference images to determine the changed area and the unchanged area (Gil-Yepes et al., 2016, Cao et al., 2014, Cao et al., 2016).However, such a method only compares single pixels themselves individually without considering the correlation information between pixels. In addition, it is difficult to decide on the required change threshold and the resulting change map often contains a large amount of “salt and pepper” noise (Peng et al., 2019). In response to this problem, object-based CD methods make use of specific information concerning the different objects in an image as the analysis unit upon which to capture effective change information (Zhang et al., 2017, Qin et al., 2013, Ma et al., 2016). Essentially, they divide an remote sensing image into multiple homogeneous regions according to its underlying spectral and spatial characteristics. With an integrated use of both spectral and spatial information, the segmentation images at two different temporal moments are subtracted to obtain the change map, minimizing the effect of noise. Importantly, both aforementioned groups of traditional CD methods work by resorting to artificially designed features, which are generally not only complicated but also poor in robustness. As the amount of remote sensing image data continues to increase, traditional CD methods can work at the expense of consuming a great deal of effort for any meaningful practical applications.

Recently, CNN based on deep learning has been rapidly developed. Due to its outstanding generalization ability for feature capturing and expression, it has been applied to performing remote sensing image CD tasks (Zhang et al., 2016, Khelifi and Mignotte, 2020). In particular, FCN has been introduced as an effective CD method. In 2015, (Long et al., 2015) firstly proposed FCN, as a landmark model for image segmentation. FCN replaces a fully connected layer within the traditional CNN model with a convolutional layer, which can adapt to inputs of any size. Also, the deconvolution layer is used for up-sampling in an effort to achieve pixel-level classification.

In terms of how a CD algorithm based on deep learning manages bi-temporal images, such methods can be classified into two complementary categories: early-fusion methods and late-fusion methods. The former refers to the use of concatenate or differential images of two bi-temporal images as an input to the network. For instance, (Alcantarilla et al., 2018) fed two bi-temporal images concatenating six channels into an FCN composed of stack contraction and expansion blocks, to achieve binary classification of pixels. (Peng et al., 2019) used an image composed of two bi-temporal images as the input to a modified U-Net++ network (Zhou et al., 2018), for the purpose of detecting any underlying changed area. Similarly, (Liu et al., 2020a) proposed a modified U-Net model with depth-wise separable convolution, whose input is an image stacked the bi-temporal images along the channel. The latter, namely the late-fusion CD algorithms, refer to the approach where two bi-temporal images are taken as independent input images. In this category of approaches, features are extracted through two independent pipelines of the network and subsequently, the resulting two sets of features are fused to generate a change map.For example, (Daudt et al., 2018) proposed two models, Fully Convolutional Siamese-Concatenation (FC-Siam-conc) and Fully Convolutional Siamese-Difference (FC-Siam-diff), to implement the late-fusion approach. The encoder part of the U-Net network is used to extract features at each temporal moment and then, the decoder part is used for feature fusion to generate the change map required. Compared with an early-fusion CD method (e.g., Fully Convolutional Early Fusion (FC-EF) (Daudt et al., 2018)), FC-Siam-diff and FC-Siam-conc as late-fusion methods can produce better results Another example for late-fusion is the deeply supervised image fusion network (IFN), proposed by (Zhang et al., 2020), which fused deep features extracted in parallel through a fully convolutional two-stream architecture and fed into the difference discrimination network for change detection. (Lei et al., 2020) used a Siamese convolutional neural network to extract features and explored the fusion of channel pairs at multiple feature levels. (Jiang et al., 2020) constructed a Siamese network with an encoder-decoder structure, where the bi-temporal images are utilized as two inputs in the encoder, and then the change residual module is used to fuse the features of the bi-temporal images as the input of the decoder. (Xu et al., 2020) employed a pseudo-Siamese capsule network to extract features of the bi-temporal images, and the extracted features are directly concatenated to calculate change probability map. Additionally, there is another CD method based on deep learning that does not fuse any two bi-temporal images. Instead, FCN is exploited to extract features from the bi-temporal images T1 image and T2 image, and changes are subsequently detected by measuring the distance between feature pairs. Furthermore, (Chen et al., 2021) proposed a dual attentive fully convolutional Siamese network (DASNet) to extract features over image pairs, with the resulting features adopted to modify contrast loss, thereby improving the performance of the model.(Zhang et al., 2018) introduced a deep Siamese semantic network with a triplet loss function to improve change detection performance. (Wang et al., 2020a) presented a supervised change detection method based on the Siamese CNN, which is exploited to detect changes through the difference of extracted features. To generate more discriminating features, (Chen and Shi, 2020) proposed the spatial-temporal attention neural network (STANet), which uses a Siamese FCN to extract the bi-temporal image feature maps with a self-attention module.

Experimental results available in the literature have so far, convincingly shown that the CD methods based on deep learning can produce a change map that is superior to what is attainable using a traditional CD method. However, important issues remain in the existing deep learning-based approaches, such as insufficient utilization of information embedded in high-resolution remote sensing images and lack of temporal information of dual-time images. Inspired by this observation, this paper proposes HRTNet that can help effectively alleviate these issues.

2.2 Attention mechanism

The attention mechanism employed in a CNN-based model originates from the study of human vision (Mnih et al., 2014).Human beings can expeditiously select high-value informative content from a large amount of information with inhibitory control of attention. Recent research has demonstrated that mimicking the idea of visual attention mechanism can significantly improve the efficiency and accuracy of automated visual information processing (Hu et al., 2018, Woo et al., 2018).

According to the different domains of attention, there are three attention mechanisms: spatial domain attention, channel domain attention and mixed domain attention. (Jaderberg et al., 2015) proposed a spatial transformer module, which transforms spatial information in an image to implement the intention of noticing spatial domain information. (Hu et al., 2018) enhanced channel information by introducing a Squeeze-and-Excitation (SE) module to exploit the relationship between channels. (Wang et al., 2020b)developed a technique to reduce loss of information due to dimension compression, presenting an Efficient Channel Attention (ECA) module that pays more attention to useful information in the channel domain. Mixed domain attention refers to the application of attention in both channel and spatial domains. (Woo et al., 2018) employed a Convolutional Block Attention Module (CBAM) to infer the attention weight along the spatial and channel dimensions in turn, adaptively adjusting the original feature map. In the study of CD for remote sensing images, the attention mechanism can also be exploited to help identify any change regions. Particularly, (Chen et al., 2021) proposed DASNet using dual attention(Fu et al., 2019) to establish associations between extracted features, obtaining global context information. (Zhang et al., 2020) presented the IFN model using CBAM (Woo et al., 2018) to fuse information from different domains. Nonetheless, all these CD approaches work without considering any temporal information embedded within the original images. In sharp contrast, HRTNet is herein proposed as a CD model with attention mechanism that exploits temporal information that reflects what surfaces have changed over time. This is described below.

3. PROPOSED METHOD

In this section, the overall HRTNet framework is introduced first, which is followed by a description of the high-resolution feature extract network and that of the dynamic inception module, including its role within the framework.

3.1 Network architecture

HRTNet is an end-to-end deep network, using a bi-temporal image pair and a DI as input, and it produces a change map as the output. The main structure of the model is shown in Figure 1, where the T1 and T2 images and the DI between them are used as the parallel inputs to the model, for the extraction of the deep features of their corresponding original image, as shown in Figure 1(a). Then, the dynamic inception module, as shown in Figure 1(b), is adopted to learn the multi-scale temporal features of the input images, again respectively in response to their originals, making the model potentially more sensitive to any changes of a different scale. Next, the features of the DI are fused with the temporal features of the T1 image and those of the T2 image, respectively. Finally, the change map is obtained by computing the distances between each pair of such fused features.

In order to take advantage of the temporal information contained within the bi-temporal images, such a triplet network model is devised to learn high-resolution features from the given T1, T2 and DI. Here, DI is defined to be the result of subtracting the corresponding pixel values of two given images, in an effort to weaken any similar part of the images while highlighting any changing part between them. It is computed from bi-temporal inputs as follows (Negri et al., 2020):

$$DI = |T1 - T2| \tag{1}$$

Note that many deep learning-based change detection methods use difference images as the only input to the network in order to directly detect any change areas(Negri et al., 2020, Gong et al., 2017, Geng et al., 2019), as DI contains critical information for change detection, especially the information on land surface

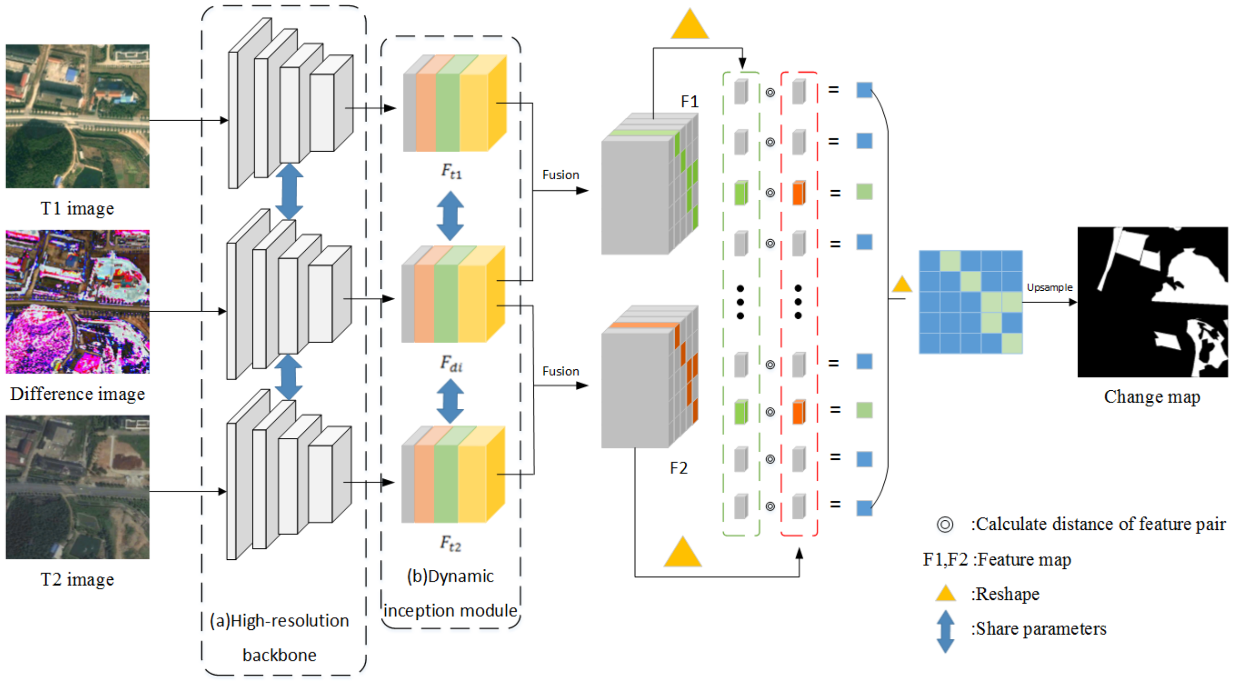

Figure 1: Architecture of high-resolution triplet network (HRTNet) with dynamic inception module, where (a) illustrates the high-resolution feature extraction backbone and (b) represents the dynamic inception module (DIM).

change over a given period. As DI is the result of subtracting the corresponding pixel values of two given images T1 and T2, it may include pseudo-changes caused by seasons, sunshine, atmospheric characteristic, shooting angle, etc. Of course, DI loses certain image details contained within the original T1 and T2 images. Therefore, complementing all three input images effectively improves the performance of the resultant model while retaining the sensitivity over pseudo-changes. There is little redundant information generated during the compution process as the three inputs (T1, T2 and DI) generally reflect different image features. They enable HRTNet not only to extract image details and features of the bi-temporal images, but also to learn the change information.

Let F$_{t1}$, F$_{t2}$ and F$_{di}$ denote the three features produced by three parallel computing steams that share network parameters. In particular, F$_{di}$ is used to weight the features of F$_{t1}$ and F$_{t2}$ using Equation 2, and assign temporal information on to the images at the time of T1 and T2. The use of weighted features makes those regional features that change over the given time period more distinct, while restraining those that do not change in the period. In so doing, it effectively improves the network to focus the attention on the region changed over time.

$$\begin{aligned} F_1 &= F_{t1} \times F_{di} + F_{t1} \\ F_2 &= F_{t2} \times F_{di} + F_{t2} \end{aligned} \tag{2}$$

In order to extract useful features regarding the changing regions at different scales, the inception module (Szegedy et al., 2015) is applied to perform convolution and re-aggregation on multiple scales of the feature map. The resulting richer features help to improve the final classification, raising the classification accuracy. In order to increase the coverage, or the comprehensiveness, of the model, dynamic convolution is applied to implementing a dynamic inception module. Dynamic convolution automatically learns the essence of different convolution kernels involved, leading to a stronger expressive power than that traditional convolution offers.

Note that existing CD algorithms commonly fuse the features extracted from the T1 and T2 images directly, and then use upsampling or deconvolution to restore the features to having a size of the original image. From there, they utilize the sigmoid function to implement pixel-level classification in an effort to achieve the detection of changed area. Different from this approach, in HRTNet, the distances between the bi-temporal features returned by DIM are measured. Given feature maps F_1 and F_2, each feature map is firstly resized to be of the same size as the input bitemporal images by bilinear interpolation. Then, the euclidean distance between the resized feature maps pixel-wise is calculated to generate the distance map $D \in R^{H\times W}$, where H and W are the height and width of the input images respectively. In the training phase, a contrastive loss (as per Equation 3) is used to learn the parameters of the network, in an effort to decrease the pixel pair distance of the unchanged area while increasing the pixel pair distance of the changed area:

$$\begin{aligned} Contrastive\ loss = &\sum_{i,j} \frac{1}{2}[(1 - y_{i,j})\, d_{i,j}^2 \\ &+ y_{i,j} \max\left(\text{margin} - d_{i,j}, 0\right)^2] \end{aligned} \tag{3}$$

where $d_{i,j}$ denotes the distance between the corresponding pixels of the feature maps F_1 and F_2 at coordinates (i, j); the margin is a set threshold, enforcing the changed feature pairs; and $y_{i,j}$ represents the label of a pixel. Particularly, when $y_{i,j} = 0$, it means that the feature pairs are unchanged and the loss function is $\sum_{i,j} \frac{1}{2} d_{i,j}^2$. For unchanged feature pairs, if the distance in the feature space is large, it indicates that the current model is not good and so, the loss is increased. When $y_{i,j} = 1$, it means that the feature pairs are changed and the loss function is

$\sum_{i,j} \frac{1}{2} \max\left(\text{margin} - d_{i,j}, 0\right)^2$. When the feature pairs are changed, while the distance within the feature space is small, the loss is in-

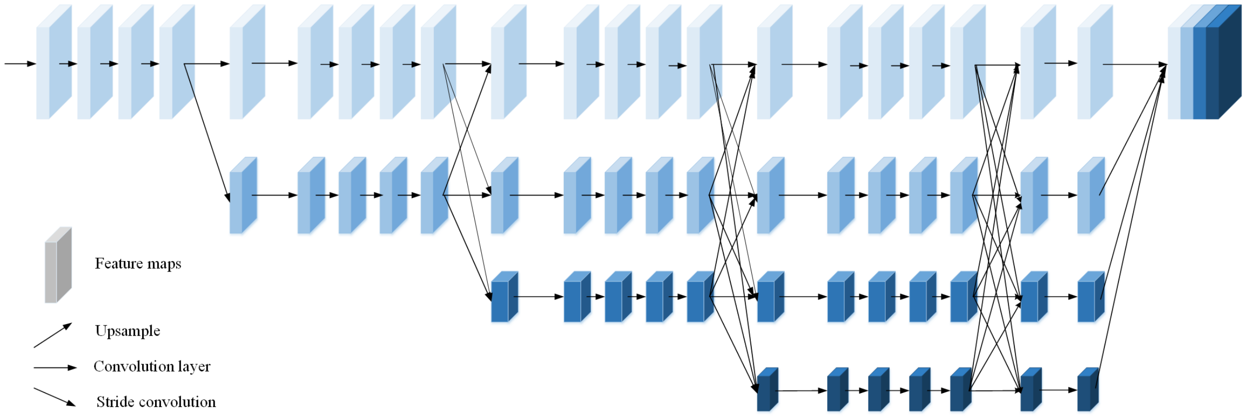

Figure 2: The architecture of high-resolution feature extract network (HRNet).

creased in response, correctly reflecting the design intuition underlying the approach.

In the testing phase, however, the change map M is obtained by a simple threshold segmentation:

$$M_{i,j} = \begin{cases} 1 & D_{i,j} > \theta \\ 0 & \text{else} \end{cases} \tag{4}$$

where the subscripts i and j $(1 \leq i \leq H, 1 \leq j \leq W)$ denote the indices of the height and width respectively; and θ is a fixed threshold to separate the change areas. In implementing the present work, θ is empirically assigned to 1.

3.2 High-resolution feature extract network

As indicated previously, with the development of remote sensing technology, the resolution of remote sensing images obtained is getting higher and higher, and the information contained in the images is becoming more and more abundant. However, traditional feature extraction networks are generally encoded using multiple convolutional layers to obtain a low-resolution feature map. As such, the mainstream network structure does not take advantage of the high-resolution characteristics of the remotely sensed images.

In this work, a high-resolution network (HRNet) is proposed, as shown in Figure 2, to act as the high-resolution feature extraction network. In so doing, the high-resolution feature map is always maintained during the entire learn process of the backbone, whilst low-resolution information is simultaneously added during the process of encoding in parallel. In conclusion, when HRNet is utilized to extract high-resolution image features, the semantic information contained in the original image can be learned, and the potential loss of pixel-level information is devised to be minimal.

It can be seen from Figure 2 that, in depth, HRNet is composed of four stages. The first stage consists of four residual units, each of which is formed by a bottleneck with a width of 64, and is followed by one convolutional layer with a kernel of size 3 × 3 changing the width of feature maps to C. The second, third, and fourth stages contain one, four, and three modularized blocks, respectively. In implementing the multi-resolution parallel convolution of the modularized block, each branch contains four residual units, with each involving two 3 × 3 convolutions per resolution, followed by batch normalization and non-linear activation function ReLu. From the perspective of width, each stage of HRNet has a different width. Stage 1 has only one branch, whilst stage 2 comprises a high-resolution branch and a branch whose resolution is doubled. At the end of these two branches, the output feature maps of the two resolutions are combined with each other to form the input for the next stage, and the feature maps of different resolutions are connected as illustrated in Figure 3. The structure of stage 3 and that of stage 4 are both similar to the structure of stage 2. Finally, the numbers of the output channels of the four branches at stage 4 are C, 2C, 4C and 8C respectively. Up-sampling is carried out to restore the low-resolution feature map to one that is of the highest resolution feature map size, and then the four resolution feature maps are concatenated along the channel dimension, serving as the required high-resolution feature map.

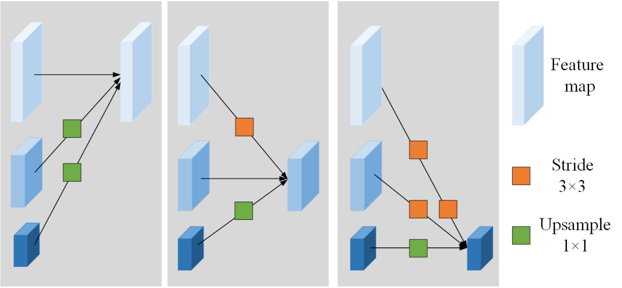

Figure 3: Feature map fusion with different resolutions. Stride 3 × 3 refers to convolution layer whose stride is 2 with a kernel of size 3 × 3, where Upsample 1 × 1 indicates the bilinear upsampling followed by a 1 × 1 convolution.

3.3 Dynamic inception module

The concept of inception module was originally introduced by (Szegedy et al., 2015), in order to make more efficient use of computing resources and to extract more informative features under the same amount of calculation, so as to improve the training results. It uses 1 × 1 convolution to adjust the dimensions of the feature map and simultaneous convolution re-aggregation at multiple scales.

Instead of taking a static approach as per the original inception module, within HRTNet, a Dynamic Inception Module is introduced as illustrated in Figure 4, which uses dynamic convolution to enable inception module having a stronger feature expression capability. Note that the static convolution is defined as follows:

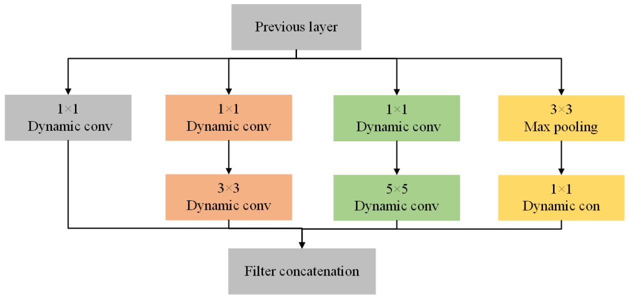

Figure 4: Structure of dynamic inception module.

$$y = g\left(W^T x + b\right) \tag{5}$$

where W, b and g represent weights, biases, and activation functions, respectively. In the present work, the dynamic convolution is specified as follows:

$$\begin{aligned} & y = g\left(\tilde{W}^T x + \tilde{b}\right) \\ & \tilde{W} = \sum_{k=1}^{K} \pi_k(x) \tilde{W}_k \\ & \tilde{b} = \sum_{k=1}^{K} \pi_k(x) \tilde{(b)}_k \\ & s.t. 0 \leq \pi_k(x) \leq 1, \sum_{k=1}^{K} \pi_k(x) = 1 \end{aligned} \tag{6}$$

where π_k is the weight of attention. The weight of attention is not fixed but varies with respect to the given input. Therefore, dynamic convolution has a stronger learning ability than its static counterpart. Figure 5 shows the structure of dynamic convolution, where the determination of the attention weight K depends on the lightweight Squeeze-and-Excitation (SE) module (Hu et al., 2018). SE is herein used to assign attention to the convolution kernel.

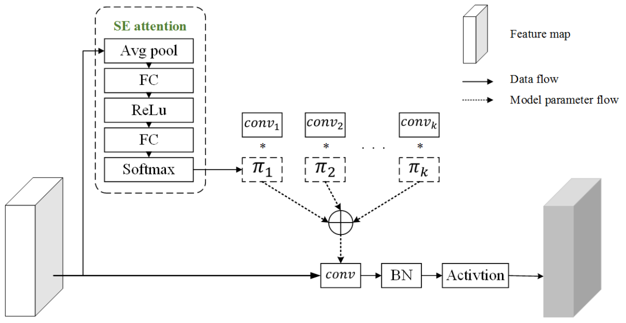

Figure 5: Structure of dynamic convolutions

4. EXPERIMENTS AND DISCUSSION

4.1 Dataset

In order to verify the effectiveness of the proposed approach, experimental studies on three popular datasets are carried out. The first as exemplified in Figure 6(a), is released by (Lebedev et al., 2018). The original image of this dataset consists of 11 pairs of multi-source remote sensing images, with resolutions ranging from 3cm to 100cm per pixel. (Ji et al., 2018) processed the original data and generated a training set of 10,000 remote sensing image pairs of 256 × 256 and a test set of 3,000 remote sensing image pairs of 256 × 256 .

The second dataset as exemplified in Figure 6(b), is the remote sensing CD data set provided by the 2020 Artificial Intelligence Remote Sensing Interpretation Competition held by SenseTime Science and Technology (SenseTime, 2020). It consists of 2968 pairs of 512 × 512 bi-temporal remote sensing images with resolutions ranging from 0.5m to 3m. The training set and the test set are divided according to the ratio of 8:2, following the common practice in the literature.

The last dataset LEVIR-CD is released by (Chen and Shi, 2020), which includes 637 very high-resolution Google Earth image patch pairs, each with a resolution of 0.5m and a size of 1024 × 1024. The original images of this dataset are collected from 20 different regions that sit in Texas of the USA, having a time span of 5 14 years. LEVIR-CD is annotated with the change information of the building and contains a total of 31333 change buildings. Compared with the previous two datasets, LEVIR-CD has labeled building changes, and most of the change areas are depicted in rectangles or polygons with clear edges. Due to hardware limitations, the original image is cropped into 256 × 256 as experimental data, as shown in Figure 6(c) for example. According to the method proposed by (Chen and Shi, 2020), the LEVIR-CD dataset is divided into a training set, a validation set and a test set.

4.2 Evaluation metrics

The performance of the proposed model is evaluated with respect to four performance metrics: Recall, Precision, F1-score and Overall Accuracy (OA). OA is the standard accuracy metric, measured in terms of the classification accuracy over positive and negative samples. Recall and Precision together indicate the effect of classification accuracy. The greater the recall value, the fewer positive samples the model misses. The higher the Precision value, the fewer false detection regarding the positive samples. Because Recall and Precision restrict each other, F1-Score is used to comprehensively consider both of them, as the metric of overall performance. Formally, these metrics are defined by

$$\begin{aligned} & OA = \frac{TP + TN}{TP + TN + FP + FN} \\ & Recall = \frac{TP}{TP + FN} \\ & Precision = \frac{TP}{TP + FP} \\ & F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \end{aligned} \tag{7}$$

where TP is the number of positive pixels correctly classified in the prediction change map, TN is the number of negative pixels correctly classified in the prediction change map, FP is the number of positive pixels wrongly predicted in the change map and FN is the number of negative pixels wrongly predicted in the change map.

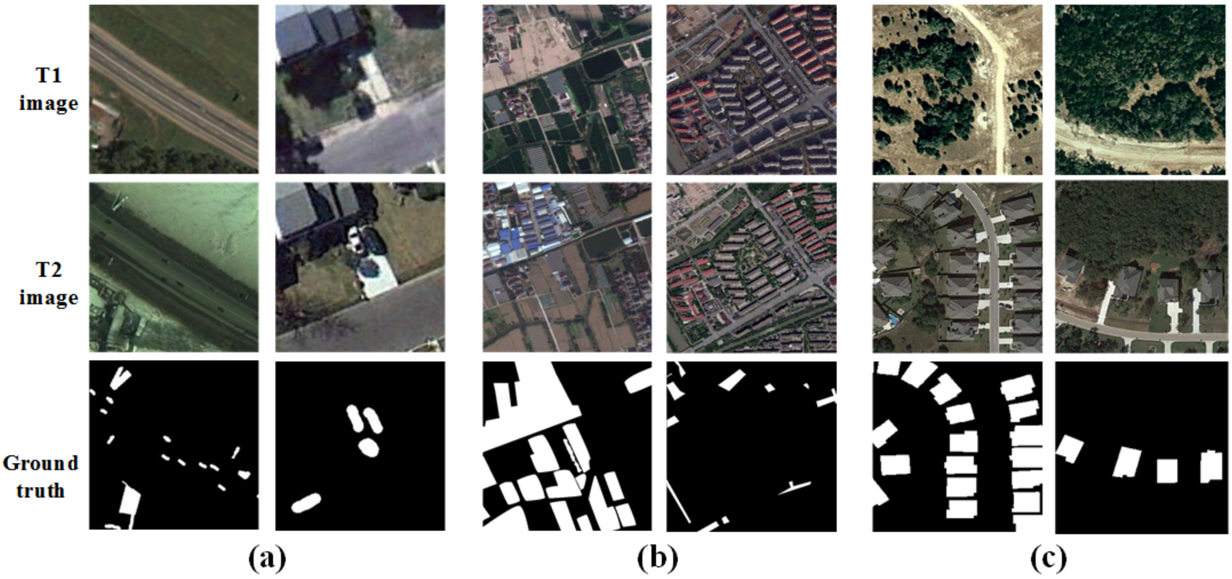

Figure 6: Bi-temporal images and ground truths that are selected from three datasets: Lebedev, SenseTime and LEVIR-CD, where (a) belongs to Lebedev, (b) is selected from SenseTime and (c) belongs to LEVIR-CD.

4.3 Implementation details

The proposed model is based on PyTorch, with the training and testing of the network implemented on the NVIDIA TITAN Xp GPU. All models evaluated are trained with Adam optimizer with an initial learning rate of 1e-4 and a weight decay of 5e-5. The same learning rate setting is also used for all models. HRNet is pre-trained on the ImageNet dataset. The entire model is trained for 60 epochs while other baseline methods compared are trained for 100 epochs.

4.4 Benchmark methods

In order to verify the effectiveness of HRTNet, The experimental results of running the following six benchmark methods are compared, on the aforementioned three datasets:

(1) **FC-EF** (Daudt et al., 2018)

Fully Convolution Early Fusion (FC-EF), which concatenates two-time images before passing them through the network, treating them as six channels images. The fused image is ”encoded-decoded” to obtain a mapping input to the change map.

(2) **FC-Sima-conc** (Daudt et al., 2018)

Fully Convolutional Siamese – Concatenation (FC-Siam-conc), which is the Siamese extension of FE-EF, changing the input of the network into two equal streams with shared weights. Each image is given to one of these equal streams. In the decoding step, concatenate is used to connect the features of the two-time images.

(3) **FC-Sima-diff** (Daudt et al., 2018)

The difference between Fully Convolution Siamese-Difference (FC-Siam-diff) and FC-Sima-conc only rests in the decoding step. The absolute value of the difference between the two-time image features is connected, instead of directly concatenating the feature pairs.

(4) **Unet++MSOF** (Peng et al., 2019)

Peng et al. proposed an end-to-end architecture inspired by UNet++ (Zhou et al., 2018), which connects two-time images as the input of the network, and uses UNet++ to learn visual feature representation. At the same time, it uses the technique of multiple side-output fusion (MSOF) to further improve the spatial details.

(5) **IFN** (Zhang et al., 2020)

Zhang et al. introduced the deeply supervised image fusion network (IFN), which uses a fully convolutional dual-stream architecture to learn representative deep features in dual-time images, with the resulting features input into a deep differential recognition network for CD. Meanwhile, the attention module is used to reconstruct the change map and to refine the CD results.

(6) **DASNet** (Chen et al., 2021)

Chen et al. proposed a dual attentive fully convolutional Siamese network (DASNet) for CD, capable of learning bi-temporal image features, with weighted double margin contrastive loss to improve the performance of change detection. Note that on the Lebedev dataset, DASNet represents the performance of state-of-the-art (SOTA).

(7) **STANet** (Chen and Shi, 2020)

Chen et al. put forward the spatial-temporal attention neural network (STANet) for CD, which uses Siamese FCN to extract the bi-temporal image feature maps with a self-attention module to generate more discriminative features. A metric module is also employed to obtain the change map. As with DASNet, on the LEVIR-CD dataset, STANet can also reach the SOTA performance.

4.5 Performance Comparison

The proposed HRTNet is compared with the existing start-of-the-art approaches on three datasets in terms of their performance.

On Lebedev dataset: Table 1 shows the experimental results on the Lebedev dataset, with the best performance highlighted in bold.

Table 1: Quantitative results of HRTNet and seven benchmark methods on Lebedev.

Method	Precision(%)	Recall(%)	F1(%)	OA(%)
FC-EF	81.56	76.13	77.11	94.13
FC-Sima-conc	84.41	82.50	82.50	95.72
FC-Sima-diff	85.78	83.64	83.73	95.75
UNet++MSOF	89.54	87.11	87.56	96.73
IFN	**94.96**	86.08	90.30	97.71
DASNet	91.45	92.52	91.93	98.07
STANet	88.24	93.43	90.75	97.84
HRTNet	93.34	**94.09**	**93.71**	**98.41**

It can be seen from Table 1 that the overall performance of HRTNet is the best, with Recall (94.09%), F1-Score (93.71%) and OA (98.41%) all being the highest. This is because the high-resolution deep features learned by the proposed method are more representative, with abundant feature information obtained using dynamic inception module. Particularly, the Recall of HRTNet is much higher than other methods (and the Precision also ranks second), which implies that HRTNet is far more robust to pseudo-change and noise than other benchmark methods. Whilst IFN has the highest precision but it has a relatively lower recall rate, indicating that it causes more false detection. Note that HRTNet is slightly inferior to IFN in Precision. One possible reason is that the number of changed pixels and that of unchanged are rather uneven. IFN alleviates the sample imbalance problem using its weighted loss function. Qualitatively, the CD results of different methods on Lebedev dataset are shown in Figure 7.

The change maps predicted by HRTNet are also presented on the Lebedev dataset under different conditions in Figure 8. From the first column of these experimental results, it can be seen that HRTNet is able to detect scattered small area changes well. The results given in the second and third column jointly reflect that the proposed approach is capable of detecting changes in both small and large areas when the land is covered by heavy snow due to seasonal changes, while the land surfaces are almost completely changed. The fourth column shows that HRTNet can recognize pseudo-changes such as tree shadows. Note that the T2 image in the last column has a large area of noise.

The above results demonstrate that the proposed algorithm has excellent robustness and has a strong ability to counter against noise. That is, HRTNet can identify change areas of different scales while at the same time, reducing the impact of pseudo-changes and noise. These observations are further confirmed by the experimental outcomes regarding the other two datasets, as presented below.

On SenseTime dataset: From a quantitative point of view, as reflected by Table 2, the overall performance of HRTNet is best with the highest Recall (71.38%), F1-score (67.90%), and OA (86.63%). FC-EF has the worst overall performance with the lowest Precision (46.73%), Recall (63.30%), F1-score (53.77%), and OA (74.86%). FC-Sima-conc shows improved performance over FC-EF, especially regarding Recall and F1-score. This is because FC-sima-conc is based on the late fusion approach, which provides deep characteristics of a single image by skip-connection to help rebuild the original image, keeping the boundary of the change map continuous, whilst FC-EF is an early fusion method. FC-sima-diff explores the difference information contained within the bi-temporal features, thereby achieving better results than FC-sima-conc. UNet++ MSOF, which is also an early fusion method, fuses the change maps at different semantic levels and obtains the change maps with each evaluation index superior to FC-EF. IFN has the highest precision because it fuses deep features and difference features to improve the accuracy by the use of an attention module. However, the algorithm is insensitive to noise and pseudo-changes on dual-time images, therefore its recall rate is low. The CD results of different methods on SenseTime dataset are qualitatively shown in Figure 9.

Table 2: Quantitative results of HRTNet and seven benchmark methods on SenseTime.

Method	Precision(%)	Recall(%)	F1(%)	OA(%)
FC-EF	46.73	63.30	53.77	74.86
FC-Sima-conc	47.45	69.79	56.49	76.89
FC-Sima-diff	55.19	68.57	61.16	81.74
UNet++MSOF	54.95	68.88	61.13	81.64
IFN	**72.89**	57.67	64.39	85.85
DASNet	59.73	68.38	63.77	83.71
SATNet	62.97	69.17	65.93	85.87
HRTNet	64.74	**71.38**	**67.90**	**86.63**

On LEVIR-CD dataset: The results of the comparison between HRTNet and other benchmark methods on this dataset are listed in Table 3. The performance of HRTNet is the best with the highest Precision (85.43%), Recall (91.77%), F1-score (88.48%), and OA (98.79%). STANet and HRTNet, both of which are based on metric learning, achieve better CD outcomes than the other methods. Figure 10 shows the examples of CD results using different methods on the LEVIR-CD dataset, qualitatively. It can be seen that HRTNet is better at predicting the edge information of the buildings.

Table 3: Quantitative results of HRTNet and other benchmark methods on LEVIR-CD.

Method	Precision(%)	Recall(%)	F1(%)	OA(%)
FC-EF	48.90	85.90	62.32	93.83
FC-Sima-conc	60.40	76.63	68.21	96.30
FC-Sima-diff	54.21	73.18	63.09	95.95
UNet++MSOF	79.00	84.14	81.49	98.05
IFN	79.55	87.99	83.57	98.20
DASNet	77.41	89.87	82.83	98.06
STANet	83.81	91.04	87.34	98.33
HRTNet	**85.43**	**91.77**	**88.48**	**98.79**

4.6 Ablation study

In this experimental study, the impact of the proposed Triplet structure and Dynamic Inception Module on model performance is examined, making quantitative comparisons on the Lebedev dataset.

Triplet structure:HRTNet takes difference images as a parallel stream input, in addition to the T1 and T2 images, to learn the temporal information of bi-temporal images. In order to evaluate the effect of extracted temporal features, a high-resolution Simaese network (HRSNet) is constructed as a model without introducing such temporal information, to compare against the standard HRTNet. The input of HRSNet involves only two inputs, namely, the T1 and T2 images. The parameter settings of the two networks are devised to be the same during training. The quantitative comparative results on the Lebedev dataset are shown in Table 4. It can be seen from this table that compared with HRSNet, HRTNet has a 0.82% increase on Recall and 0.42% increase on F1-score, while the Precision and OA values are also improved. This demonstrates that HRTNet reinforces the performance of HRSNet through the introduction of the temporal information. More importantly, by learning the characteristics of such information, the robustness of the model to pseudo-change and noise of images at bi-temporal images is strengthened also, with the recall rate significantly increased. Qualitative, Figure 11 contrasts the change maps predicted by the two models.

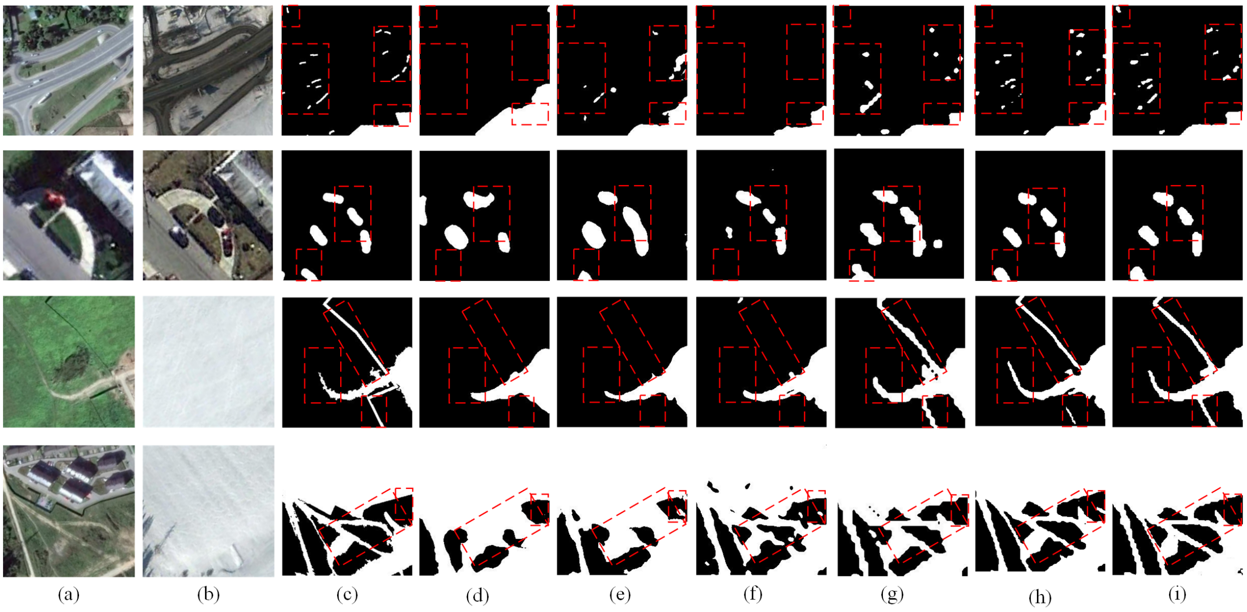

Figure 7: Qualitative performance comparison between HRTNet and other benchmark algorithms on Lebedev. (a) T1 image. (b) T2 image. (c) Ground truth map. Change maps predicted by means of (d) FC-Siam-diff, (e) UNet++MSOF, (f) IFN, (g) DASNet, (h) STANet, (i) HRTNet. The first and second rows illustrate the detection of small change areas. Compared with other SOTA methods, HRTNet can detect more changes in small areas. In addition, it can accurately separate the boundaries of independent small change areas. The land surface of bi-temporal images in the third and fourth rows is shown to have greatly changed due to seasonal variation. HRTNet can detect more details than other benchmark methods, and it is also more sensitive to areas involving multi-scale changes.

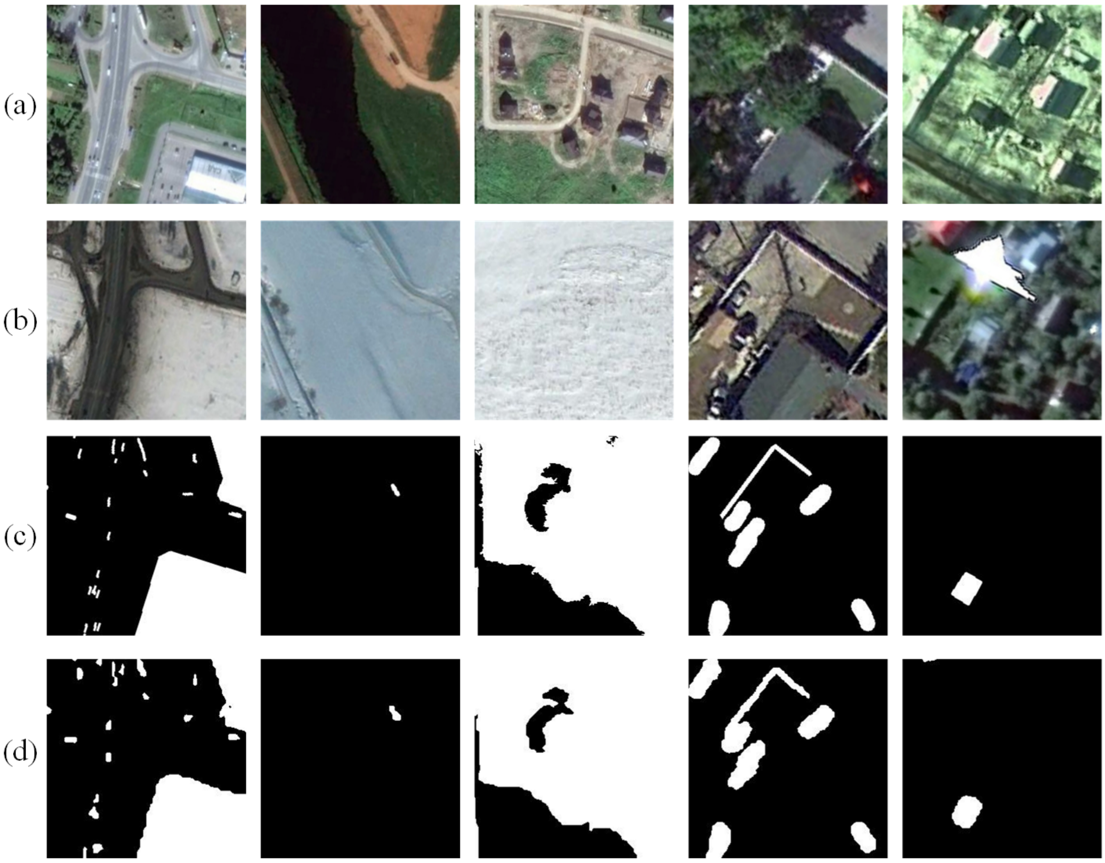

Figure 8: Under different conditions change detection results on the Lebedev dataset. (a) T1 images. (b) T2 images. (c) Ground truth maps. (d) Change detection maps predicted by HRTNet.

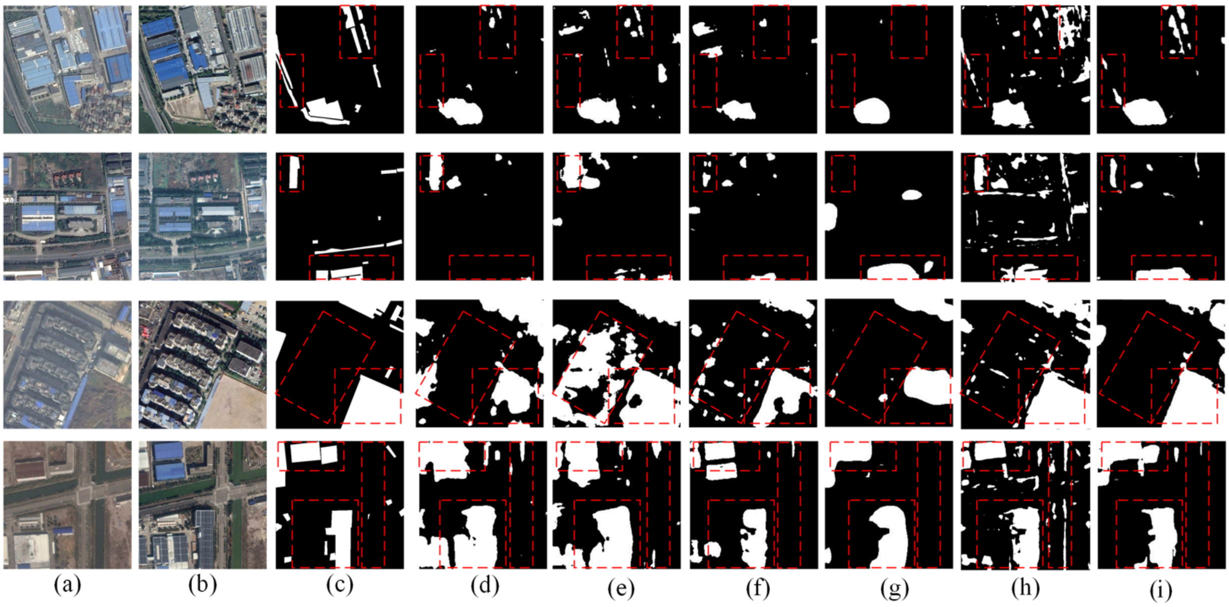

Figure 9: Qualitative comparison between HRTNet and benchmark SOTA algorithms on SenseTime. (a) T1 image. (b)T2 image. (c) Ground truth map. Change maps predicted by means of (d) FC-Siam-diff, (e) UNet++MSOF, (f) IFN, (g) DASNet, (h) STANet, (i) HRTNet. The first and second rows are the results of detecting small change areas. Compared with SOTA methods, HRTNet can detect more changes in small areas with more change details. In addition, it has fewer falsely detected pixels. The third and fourth rows show the detection of big change areas. Again, HRTNet can detect more details than other methods, with accurately detected boundaries of big change areas.

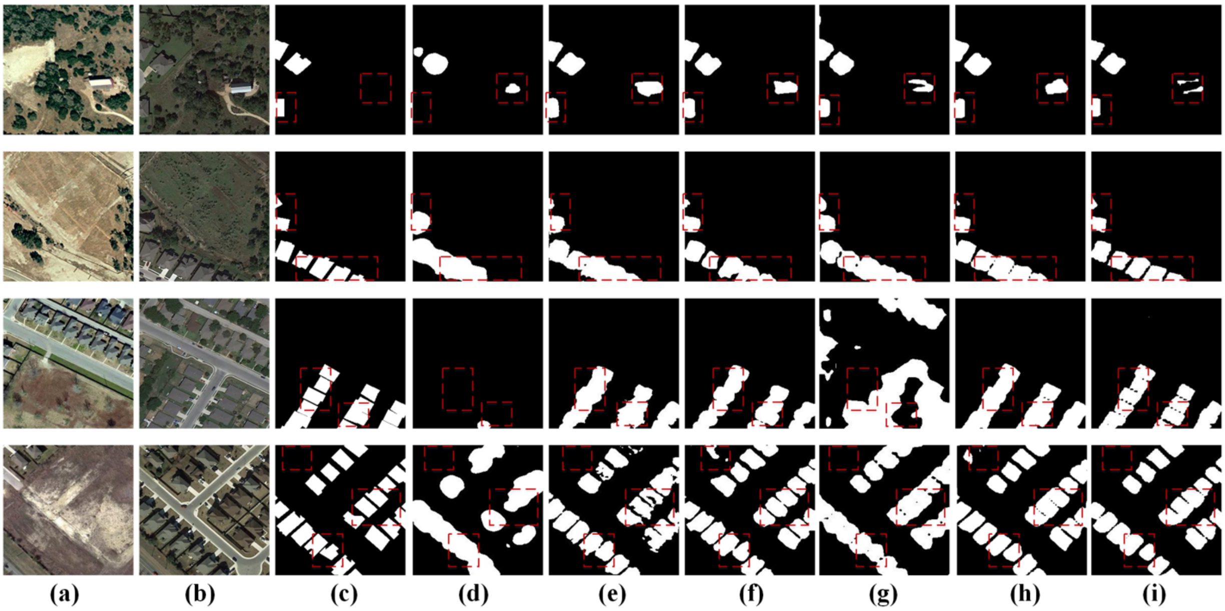

Figure 10: Qualitative comparison between HRTNet and other SOTA algorithms on LEVIR-CD. (a) T1 image. (b)T2 image. (c) Ground truth map. Change maps predicted by means of (d) FC-Siam-diff, (e) UNet++MSOF, (f) IFN, (g) DASNet, (h) STANet, (i) HRTNet. The first row shows the detection outcome of sparse change areas. Compared with SOTA methods, HRTNet has fewer falsely detected pixels. The remaining three rows are the CD outcomes of change areas of different scales, demonstrating that change maps predicted by HRTNet can detect more details and fewer conglutinations between buildings than the other methods.

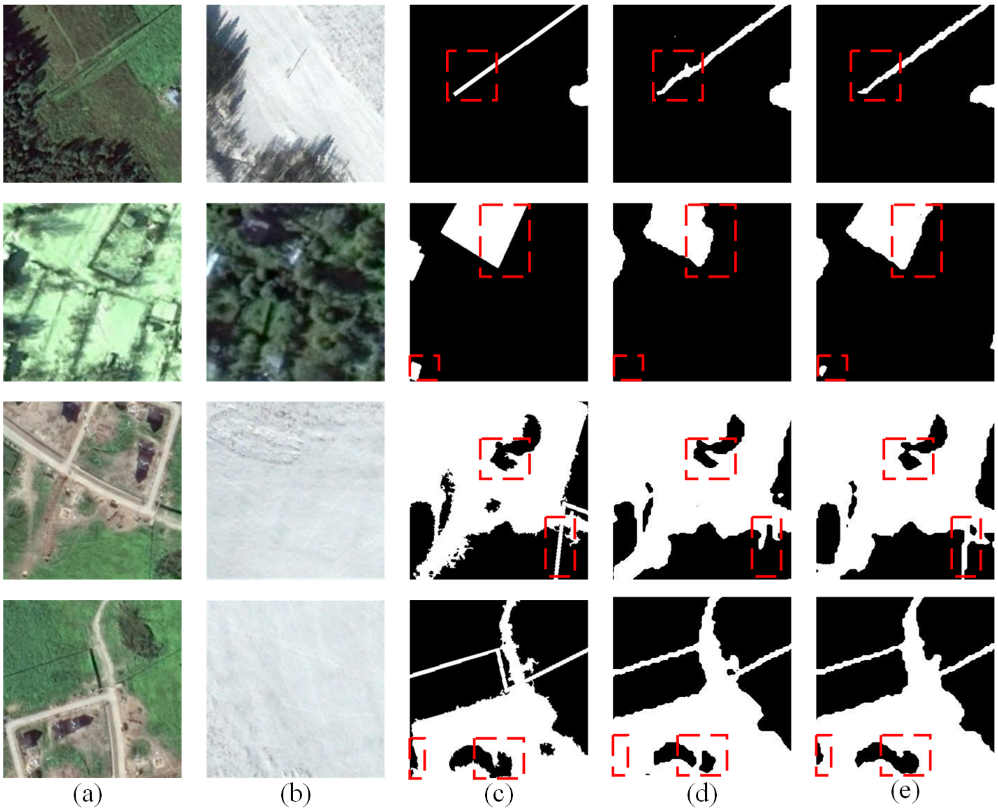

Figure 11: Qualitative comparison of HRTNet and HRSNet on Lebedev. (a) T1 image. (b)T2 image. (c) Ground truth. Change maps predicted by means of (d) HRSNet, (e) HRTNet. The illustrations involve two moment images of pseudo-changes caused by seasonal changes in weather or plantation, HRTNet is shown to be more sensitive to such pseudo-changes.

Table 4: Quantitative results of HRTNet and HRSNet on Lebedev

Method	Precision(%)	Recall(%)	F1(%)	OA(%)
HRSNet	93.28	93.27	93.29	98.31
HRTNet	**93.34**	**94.09**	**93.71**	**98.41**

Table 5: Parameters and inference time of HRSNet and HRTNet on Lebedev

Method	Parameters	Inference time(s)	F1(%)
HRSNet	66.89M	**0.110**	93.29
HRTNet	66.89M	0.134	**93.71**

The speed of HRTNet and that of HRSNet are compared in Table 5. It can be seen from this table that given the same parameters of HRTNet and HRSNet, (since the feature extraction network is shared between these models for learning the image features), for a pair of bi-temporal images, of a size 256 × 256, HRTNet requires slightly more time to run (0.134s) than HRSNet (0.110s). This is because it needs to learn the features of DI. However, the F1-score of HRTNet is higher than that of HRSNet. Trading the improved performance with the rather slight increase in computation costs is therefore worthwhile.

To reinforce the conceptual distinction between the use of correlation and that of difference, experimental investigation on ablation learning has been extended. In particular, T1, T2 and correlation images are chosen as the inputs of HRTNet, with the correlation image being T1 and T2 images stacked along the channel dimension (over six channels). Table 6 contrasts the performances of the two models.

As can be seen from Table 6, HRTNet with correlation underperforms in comparison to HRTNet with difference. This is likely attributed to the fact that a difference image explicitly guides the network to learn the differences between the bi-temporal images. In addition, a correlation image is a 6-channel image, which cannot share parameters completely with T1 and T2, and its computation is obviously more costly

Table 6: Quantitative results of different inputs on Lebedev

Method	Precision(%)	Recall(%)	F1(%)	OA(%)
HRTNet with correlation	92.90	90.69	91.78	97.96
HRTNet with difference	**93.34**	**94.09**	**93.71**	**98.41**

Dynamic inception module: The proposed HRTNet structure applies DIM to enrich the representation of feature information, extracting multi-scale temporal features. In order to evaluate the impact of introducing DIM, both HRSNet and HRTNet models with DIM included are compared to their counterparts without DIM. The parameter settings of the four networks are the same during the training process. The results of quantitative comparison on their working with the Lebedev dataset are shown in Table 7. It can be seen that compared to the models without DIM, HRSNet and HRTNet including DIM improve Precision, F1 and OA significantly, although the Recall value decreases slightly. One possible reason for this observation is that DIM introduces more feature information. While improved the precision value, the employment of DIM also introduces a small amount of redundant feature information, which may lead to false prediction of the change maps, causing the recall value to decrease. Figure 12 shows the example change maps predicted by the four models. It can be seen that the effect on detecting the boundaries for

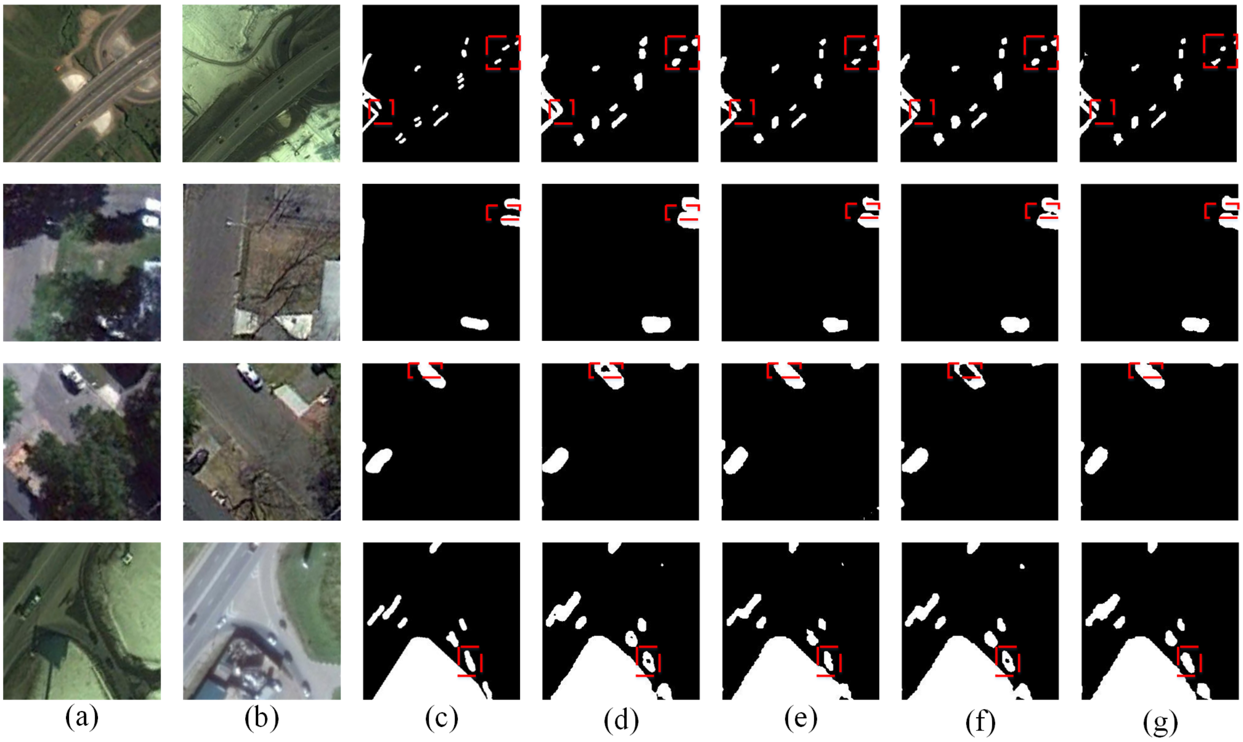

Figure 12: Qualitative comparison of different backbones on Lebedev. (a) T1 image. (b)T2 image. (c) Ground truth. Change maps predicted by means of (d) HRSNet without DIM, (e) HRSNet, (f) HRTNet without DIM, and (g) HRTNet.

objects of different scales is improved thanks to the introduction of multi-scale information. From the third and fourth rows of the images in this figure , it is observed that the models with DIM can learn the features of the objects better, with reduced incorrect detection while avoiding the phenomenon of ”hollow” in the detection areas.

Table 7: Quantitative results of using different backbones on Lebedev.

Method	Precision(%)	Recall(%)	F1(%)	OA(%)
HRSNet without DIM	89.77	**95.03**	92.33	98.02
HRSNet	93.28	93.27	93.29	98.31
HRTNet without DIM	91.61	94.25	92.91	98.19
HRTNet	**93.34**	94.09	**93.71**	**98.41**

5. CONCLUSION

This paper has proposed a High-Resolution Triplet Network (HRTNet) for change detection in remote sensing images. Different from existing methods HRTNet pays particular attention to the temporal information contained within the dual-time images. It learns the respective features of bi-temporal images and time information through three parallel streams. Targeting at high-resolution images, HRNet is used as the backbone of feature extraction to reduce the loss of image information during the learning process. A dynamic inception module is introduced to help cope with change regions of different scales, enhancing the expression ability of the model while exploiting multi-scale image features. Compared with state-of-the-art methods, the proposed HRTNet performs well on three popular datasets, showing good learning ability for change areas of different scales while being able to detect pseudo-changes and counter against noise. Systematic experimental results have demonstrated the effectiveness and robustness of the proposed method.

Whilst very promising, the proposed approach has two key limitations as follows (which are shared in general with most deep learning-based approaches): 1) The model can only be trained and tested under the same scenario concerned; it remains a significant challenge for it to be able to perform cross-domain change detection. 2) The method is based on supervised learning, which requires a large amount of annotated data to train the model in order to obtain a superior performance model. Thus, for future research, it would be very interesting to study how to use a relatively smaller number of training samples to achieve the same or even better detection performance. It would also be potentially beneficial to investigate how to use transfer learning to improve the model’s ability of performing cross-domain change detection tasks.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (61871460), the Shaanxi Provincial Key RD Program (2020KW-003, 2021KWZ-03) and the Strategic Partner Acceleration Award under Grant 80761-AU201, funded under the Ser Cymru II program, U.K. The authors are very grateful to the Editor and anonymous reviewers for their constructive comments that have helped improve this work significantly.

REFERENCES

Alcantarilla, P. F., Stent, S., Ros, G., Arroyo, R. and Gherardi, R., 2018. Street-view change detection with deconvolutional networks. Autonomous Robots 42(7), pp. 1301–1322.

Brunner, D., Lemoine, G. and Bruzzone, L., 2010. Earthquake damage assessment of buildings using vhr optical and sar imagery. IEEE Transactions on Geoscience and Remote Sensing 48(5), pp. 2403–2420.

Bruzzone, L. and Bovolo, F., 2012. A novel framework for the design of change-detection systems for very-high-resolution remote sensing images. Proceedings of the IEEE 101(3), pp. 609–630.

Cao, G., Li, Y., Liu, Y. and Shang, Y., 2014. Automatic change detection in high-resolution remote-sensing images by means of level set evolution and support vector machine classification. International journal of remote sensing 35(16), pp. 6255–6270.

Cao, G., Zhou, L. and Li, Y., 2016. A new change-detection method in high-resolution remote sensing images based on a conditional random field model. International Journal of Remote Sensing 37(5), pp. 1173–1189.

Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geoscience and Remote Sensing Letters 6(4), pp. 772–776.

Chen, H. and Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing 12(10), pp. 1662.

Chen, H., Wu, C., Du, B., Zhang, L. and Wang, L., 2020. Change detection in multisource vhr images via deep siamese convolutional multiple-layers recurrent neural network. IEEE Transactions on Geoscience and Remote Sensing 58(4), pp. 2848–2864.

Chen, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y. and Li, H., 2021. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, pp. 1194–1206.

Daudt, R. C., Le Saux, B. and Boulch, A., 2018. Fully convolutional siamese networks for change detection. In: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, pp. 4063–4067.

El Amin, A. M., Liu, Q. and Wang, Y., 2017. Zoom out cnns features for optical remote sensing change detection. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), IEEE, pp. 812–817.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.

Geng, J., Ma, X., Zhou, X. and Wang, H., 2019. Saliency-guided deep neural networks for sar image change detection. IEEE Transactions on Geoscience and Remote Sensing 57(10), pp. 7365–7377.

Gil-Yepes, J. L., Ruiz, L. A., Recio, J. A., Balaguer-Beser, Á. and Hermosilla, T., 2016. Description and validation of a new set of object-based temporal geostatistical features for land-use/land-cover change detection. ISPRS Journal of Photogrammetry and Remote Sensing 121, pp. 77–91.

Gong, M., Yang, H. and Zhang, P., 2017. Feature learning and change feature classification based on deep learning for ternary change detection in sar images. ISPRS Journal of Photogrammetry and Remote Sensing 129, pp. 212–225.

Hu, J., Shen, L. and Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

Hussain, M., Chen, D., Cheng, A., Wei, H. and Stanley, D., 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS Journal of photogrammetry and remote sensing 80, pp. 91–106.

Jaderberg, M., Simonyan, K., Zisserman, A. et al., 2015. Spatial transformer networks. Advances in neural information processing systems 28, pp. 2017–2025.

Ji, S., Wei, S. and Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing 57(1), pp. 574–586.

Jiang, H., Hu, X., Li, K., Zhang, J., Gong, J. and Zhang, M., 2020. Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection. Remote Sensing 12(3), pp. 484.

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J. and Xian, G., 2013. A comprehensive change detection method for updating the national land cover database to circa 2011. Remote Sensing of Environment 132, pp. 159–175.

Khelifi, L. and Mignotte, M., 2020. Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. arXiv preprint arXiv:2006.05612.

Lebedev, M., Vizilter, Y. V., Vygolov, O., Knyaz, V. and Rubis, A. Y., 2018. Change detection in remote sensing images using conditional adversarial networks. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

Lei, T., Zhang, Y., Lv, Z., Li, S., Liu, S. and Nandi, A. K., 2019. Landslide inventory mapping from bitemporal images using deep convolutional neural networks. IEEE Geoscience and Remote Sensing Letters 16(6), pp. 982–986.

Lei, Y., Peng, D., Zhang, P., Ke, Q. and Li, H., 2020. Hierarchical paired channel fusion network for street scene change detection. IEEE Transactions on Image Processing 30, pp. 55–67.

Li, Y., Zhang, H. and Shen, Q., 2017. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. Remote Sensing 9(1), pp. 67.

Li, Y., Zhang, H., Xue, X., Jiang, Y. and Shen, Q., 2018. Deep learning for remote sensing image classification: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(6), pp. e1264.

Liu, R., Jiang, D., Zhang, L. and Zhang, Z., 2020a. Deep depthwise separable convolutional network for change detection in optical aerial images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, pp. 1109–1118.

Liu, Y., Pang, C., Zhan, Z., Zhang, X. and Yang, X., 2020b. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geoscience and Remote Sensing Letters pp. 1–5.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Ma, L., Li, M., Blaschke, T., Ma, X., Tiede, D., Cheng, L., Chen, Z. and Chen, D., 2016. Object-based change detection in urban areas: The effects of segmentation strategy, scale, and feature space on unsupervised methods. Remote Sensing 8(9), pp. 761.

Mnih, V., Heess, N., Graves, A. et al., 2014. Recurrent models of visual attention. Advances in neural information processing systems 27, pp. 2204–2212.

Mundia, C. N. and Aniya, M., 2005. Analysis of land use/cover changes and urban expansion of nairobi city using remote sensing and gis. International journal of Remote sensing 26(13), pp. 2831–2849.

Negri, R. G., Frery, A. C., Casaca, W., Azevedo, S., Dias, M. A., Silva, E. A. and Alcântara, E. H., 2020. Spectral-spatial-aware unsupervised change detection with stochastic distances and support vector machines. IEEE Transactions on Geoscience and Remote Sensing.

Peng, D., Zhang, Y. and Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved unet++. Remote Sensing 11(11), pp. 1382.

Qin, Y., Niu, Z., Chen, F., Li, B. and Ban, Y., 2013. Object-based land cover change detection for cross-sensor images. International Journal of Remote Sensing 34(19), pp. 6723–6737.

SenseTime, 2020. Artificial intelligence remote sensing interpretation competition.

Singh, A., 1989. Review article digital change detection techniques using remotely-sensed data. International journal of remote sensing 10(6), pp. 989–1003.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Voulodimos, A., Doulamis, N., Doulamis, A. and Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience.

Wang, F. and Xu, Y. J., 2010. Comparison of remote sensing change detection techniques for assessing hurricane damage to forests. Environmental monitoring and assessment 162(1-4), pp. 311–326.

Wang, M., Tan, K., Jia, X., Wang, X. and Chen, Y., 2020a. A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images. Remote Sensing 12(2), pp. 205.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q., 2020b. Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542.

Woo, S., Park, J., Lee, J.-Y. and So Kweon, I., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Wu, C., Du, B., Cui, X. and Zhang, L., 2017. A post-classification change detection method based on iterative slow feature analysis and bayesian soft fusion. Remote Sensing of Environment 199, pp. 241–255.

Xu, Q., Chen, K., Sun, X., Zhang, Y., Li, H. and Xu, G., 2020. Pseudo-siamese capsule network for aerial remote sensing images change detection. IEEE Geoscience and Remote Sensing Letters.

Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H. and Qiu, X., 2017. Change detection based on deep siamese convolutional network for optical aerial images. IEEE Geoscience and Remote Sensing Letters 14(10), pp. 1845–1849.

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L. and Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 166, pp. 183–200.

Zhang, H., Li, Y., Jiang, Y., Wang, P., Shen, Q. and Shen, C., 2019. Hyperspectral classification based on lightweight 3-d-cnn with transfer learning. IEEE Transactions on Geoscience and Remote Sensing 57(8), pp. 5813–5828.

Zhang, L., Zhang, L. and Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geoscience and Remote Sensing Magazine 4(2), pp. 22–40.

Zhang, M., Xu, G., Chen, K., Yan, M. and Sun, X., 2018. Triplet-based semantic relation learning for aerial remote sensing image change detection. IEEE Geoscience and Remote Sensing Letters 16(2), pp. 266–270.

Zhang, Y., Peng, D. and Huang, X., 2017. Object-based change detection for vhr images based on multiscale uncertainty analysis. IEEE Geoscience and Remote Sensing Letters 15(1), pp. 13–17.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. and Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, pp. 3–11.