Land Use Change Detection Using Deep Siamese Neural Networks and Weakly Supervised Learning

Indrajit Kalita[1], Savvas Karatsiolis[2], and Andreas Kamilaris[2,3]

[1] Indian Institute of Information Technology Guwahati, Assam 781015, India
indrakalita09@gmail.com
[2] CYENS Center of Excellence, Nicosia, Cyprus
{s.karatsiolis, a.kamilaris}@cyens.org.cy
[3] Dept. of Computer Science, University of Twente, Enschede, The Netherlands

Abstract. A weakly supervised change detection method is proposed for remotely sensed multi-temporal images, by utilizing a Siamese neural network architecture. The architecture of the Siamese network is a combination of two multi-filter multi-scale deep convolutional neural networks (MFMS DCNN). Initially, the Siamese network is trained by utilizing the image-level semantic labels of the image pairs in the dataset. The features of the image pairs are obtained using the trained network to generate the difference image (DI). Then, a combination of the PCA and the K-means algorithms has been used to produce the change map for the pair of images. Experiments were carried out using two remotely sensed image datasets. The weakly supervised method proposed in this paper offers better results in comparison to both weakly supervised- and unsupervised-based state-of-the-art models and techniques.

Keywords: Change detection · Weakly supervised · Siamese network · Convolutional neural network

1 Introduction

The massive volume increase of collected images from satellites and unmanned aerial vehicles, together with the remarkable success of Deep Neural Network (DNN) in computer vision applications, enabled the remote sensing community to develop a plethora of interesting earth observation-related applications. Aerial imagery contains spectral, spatial, and temporal information, which is valuable for the monitoring of different ecosystems and for planning tasks like crop surveillance, deforestation control, soil and water contamination monitoring, biogeochemical cycle monitoring, global heat mapping [18] etc. Moreover, landscape change detection (CD) is a critical task whose results are valuable for many policy-making mechanisms. For example, the outcomes of CD can be utilized to identify illegal changes, evaluate disasters [5] and set goals for mitigating climate change [19]. Manual landscape CD monitoring is expensive, time-consuming and infeasible to perform on a large scale, due to the massive volume of data required and the large size of monitored areas.

2 Related work

CD approaches are categorized into two classes: ones that apply post-classification comparison [22] and ones that apply post-comparison analysis [7, 17]. The former class focuses on classifying temporal images of the same region, followed by pixel-by-pixel comparison. The success of these methods depends on the classification strategy. The latter class of methods focuses on estimating a difference-image (DI) by considering multi-temporal images of the same area. The DI produced is used to acquire a feature map to distinguish regions of change over regions of no change. Thus, the quality of the DI is critical for achieving good performance. The features of the DI can be extracted by techniques involving image arithmetics [7] and transformations [17]. The acquired features can be evaluated using strategies such as thresholding, clustering [7], and Markov random fields [2].

DNNs extract robust features from complex input samples and thereby utilize the rich information contained in images. This is especially helpful for large-scale remotely sensed datasets [13]. Under this scenario, a DNN-based CD approach provides better performance on remotely sensed high-resolution images [1, 16, 21]. DNN based CD methods can be categorized as unsupervised [16], fully supervised [21] and weakly supervised approaches [1]. Liu *et al.* [16] developed an unsupervised CD algorithm using the pre-trained U-net architecture. Similarly, De *et al.* [12] also explored the U-net architecture under the unsupervised scenario. Moreover, Cao *et al.* [6] proposed a deep belief network (DBN) technique for improving the quality of the DI using the SPOT5 multispectral images. Due to the lack of labels, it is difficult to achieve a detailed CD map using the unsupervised scheme. As a result, the fully supervised learning approaches employ labeled information (ground truth) for enhancing the CD performance [10, 21]. Under this scenario, Zhan *et al.* [21] propose a contrastive loss-based supervised Siamese network to obtain the change and unchanged regions in an aerial image using the SZTAKI dataset [2]. Ji *et al.* [10] explore the Mask R-CNN and the U-net architecture for the identification of building changes using very high-resolution datasets [11]. Still, identifying pixel-level change patterns in the fully supervised system is tedious, inefficient, and expensive. This stresses the importance of exploring CD methods that minimize labelling costs by introducing image-level labels instead of pixel-level ones. Under this scenario, a supervised CD model can be trained based on high-level annotations that indicate whether two images depict land change or not. This is known as a *weakly supervised scheme* [1, 14]. Andermatt *et al.* [1], proposed a weakly supervised CD approach using a U-net based Siamese architecture. Similarly, Khan *et al.* [14] have used a pre-trained DNN in conjunction with a directed acyclic graph (DAG) to learn patterns of change from image-level labelled training data. The majority of the weakly supervised schemes mentioned above focus on pre-trained models. However, the datasets used to pre-train the models are very different from the targeted CD datasets. Therefore, the performance of these models is low. To address this issue, *this work explores a weakly supervised CD method learned from scratch under a post-comparison analysis framework.*

3 Methodology

The proposed CD approach is based on a multi-filter multi-scale (MFMS) [9] DNN, used as the feature extractor of a Siamese model [21]. The co-registered images are first preprocessed with histogram matching and then passed through the Siamese model. The Siamese model has been trained from scratch using the image-level labels only. After, the features of the two images (captured at different times) obtained using the trained Siamese network are used to estimate the DI. Finally, the PCA and the K-Means algorithms are applied to the DI, to produce the CD maps. Features are extracted at different resolutions and then processed to obtain the CD maps for the images captured at two different timestamps. A final stage processes the CD maps and determines which pixels in the images depict land change. Figure 1 illustrates the overall technique.

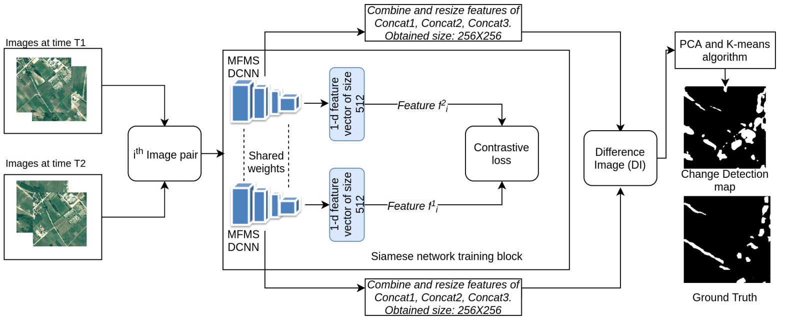
Images at time T1
Images at time T2
i^th Image pair
MFMS DCNN
Shared weights
MFMS DCNN
1-d feature vector of size 512
1-d feature vector of size 512
Feature f²ᵢ
Feature f¹ᵢ
Contrastive loss
Siamese network training block
Combine and resize features of Concat1, Concat2, Concat3. Obtained size: 256X256
Combine and resize features of Concat1, Concat2, Concat3. Obtained size: 256X256
Difference Image (DI)
PCA and K-means algorithm
Change Detection map
Ground Truth

Fig. 1: The architecture of proposed change detection model. The Concat1, Concat2, and Concat3 layers are displayed in Figure 2

3.1 Multi filter multi-scale deep convolutional neural network

The MFMS DNN combines the features learned at multiple levels of the architecture, motivated by [9]. The effectiveness of this strategy relies on the fact that the model builds representations at different resolutions. The proposed architecture applies convolution-batch normalization-activation (CBA) layers and down-sampling (max-pooling) layers to create multi-scale feature maps (see Figure 2). We discriminate between intermediate max-pooling layers and the max-pooling layers applied at the input by naming the latter down-sampling (DS) layers. This distinction stresses the purpose of each unit since the intermediate max-pooling layers aim at reducing the dimensionality of the feature maps while the down-sampling layers aim at producing multi-scale feature maps. We use three down-sampling units (DS1, DS2, DS3) as shown in Figure 2. The DS1 unit down-samples the input while the DS2 and DS3 units apply further dimensionality reduction and extend the models multi-scale processing. Similar to [9], various kernel sizes are used for the convolution operations in the CBA layers. The ReLU activation function is used to introduce non-linearities to the model. The outputs of the DS units are also processed by CBA layers to create multi-scale feature maps from various image resolutions. Concatenation operations (Concat1, Concat2, Concat3) fuse the distinct feature maps at different levels of the architecture and the average pooling (AP) operation is applied after the last CBA unit (CBA9). Finally, the generated features are combined using a dense layer. The proposed MFMS architecture constitutes the backbone of the Siamese model described in Section 3.2.

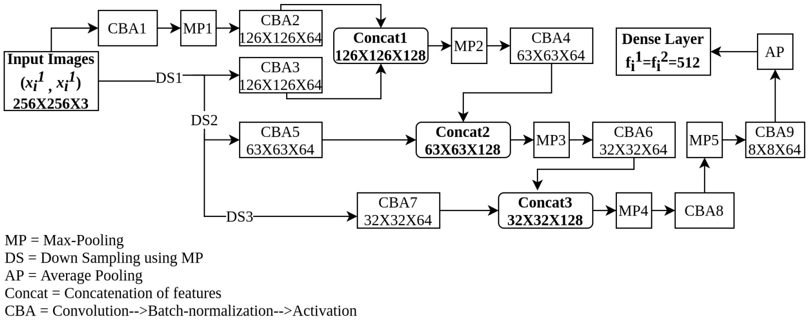

Fig. 2: The architecture of the MFMS DCNN model used in the Siamese network

3.2 Siamese Neural network

Siamese networks are models that use instances of the same DNN and share the same architecture and weights, as shown in Figure 1. The primary objective of the network is to map the features of similar samples closer to each other and the features corresponding to dissimilar samples far apart. Accordingly, a dataset used for training a Siamese model is grouped into pairs of similar or dissimilar samples: pairs of similar samples are labelled as one and the pairs of samples from different classes are labelled as zero. During training, the Siamese model extracts the features of each sample in an image pair and outputs two one-dimensional vectors. Then, the difference between the two vectors is measured using some distance metric and an optimizer minimizes this distance if the features belong to images of the same class and maximize it if the features belong to images of different classes. In the proposed approach, the datasets are prepared (i.e. paired) according to the following two conditions:

1: The paired images share the same geographical region at two different times.
2: The paired images depicting land change are labelled as zero and paired images not depicting land change are labelled as one.

This kind of labelling enables the model to operate in a weakly supervised mode, identifying the exact pixels corresponding to land change without strong supervision, i.e., without explicitly providing a label for each pixel that identifies a change or not. Instead, we just feed the model with weak supervisory information regarding which image pairs depict land change. Concretely, the proposed Siamese model tackles the CD task in a weakly supervised fashion mainly because:

1) It is trained using image-level labels instead of pixel-level labels.
2) It uses a multi-scale DNN architecture which is very efficient on the specific task.

The features of each image are computed as a 512-D vector and the contrastive loss function (L_{cons}) [8] is applied on the two extracted feature vectors (shown in Equation 2). L represents the label of image pair i, D_w is the Euclidean distance between the features of the pair (defined in Equation 1) and m is the desirable margin between image pairs that depict land change and thus have a zero label.

After training the Siamese network, the feature maps of the architecture are used to calculate the differential feature map between the images in the pair. The three concatenation layers (Concat1, Concat2, and Concat3 as shown in Figure 2) are up-scaled to form a 256 × 256 feature map for every image. Then, the Euclidean distance between the two feature maps is calculated. This provides the differential feature map (i.e. DI), which is a 2D representation of the difference between the two images in a pair. The DI holds the necessary information to identify the regions of change in an image pair (see Section 3.3).

$$D_w = \|f_i^1 - f_i^2\| \tag{1}$$

$$L_{cons} = (L)\frac{1}{2}(D_w)^2 + (1-L)\frac{1}{2}\{\max(0, m - D_w)\}^2 \tag{2}$$

3.3 Generation of change detection maps

To obtain the CD maps from a DI, we use a technique inspired by Celik *et al.* [7], based on the PCA and the K-means algorithms. Initially, the PCA is applied to every non-overlapping patch of the DI to obtain its eigenvector space. Then, the eigenvector space is projected on the overlapping patches to produce the feature vector space, which is then divided into two clusters via the K-means algorithm. The cluster with the least number of indexes (data points) is considered as the change class because the number of changed pixels in a pair of images is generally much smaller compared to the number of unchanged pixels. In contrast to Celik *et al.*, we use Euclidean distance to calculate the DI between the image representations obtained by the Siamese model, instead of the absolute difference between the two images.

4 Experimental setup

4.1 Datasets' description

The effectiveness of the proposed methodology was tested on two very high resolution (VHR) remotely sensed image datasets.

SZATAKI AirChange Benchmark Dataset : A CD dataset consisting of three parts: SZADA, TISZADOB, and ARCHIVE [2]. SZADA is used to evaluate the performance of the model, comparing with related work. We use the 43 multi-temporal images between the years 2000 and 2005. The sizes of the training images are 952×640 (30 images) and 640×952 pixels (12 images). An image of size 784×448 is used to test the model. According to the datasets description, the changed pixels annotation of the images has been provided by an expert. During training, the 42 images are divided into patches of size 256 × 256 × 3, and the *change/no change*-depicting image pairs are identified based on the annotations of the images. An image pair gets a label one if its images depict no land change and zero if the images depict land change. A total of 253 image patches are collected for training. Sample images are shown in Figure 3.

Aerial image change detection (AICD) dataset : AICD is a synthetic change detection dataset [3], used to compare the performance of the proposed model with other state-of-the-art approaches trained with weak supervision. The dataset contains 1000 images (500 image pairs) and each image is $600 \times 800 \times 3$ pixels. Each image contains only one change object (one structure). During training, images are divided into patches of size $256 \times 256 \times 3$ and the *change/no change*-depicting image pairs are identified based on the ground truth label of the corresponding images. In total, 630 image pairs depict changed structures and 2370 depict no land change.

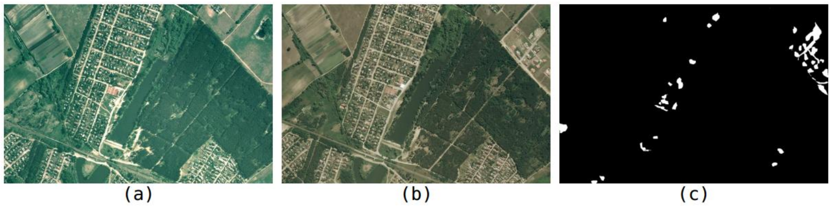

Fig. 3: Sample images from the SZADA dataset. (a,b) Training image pair, (c) corresponding ground truth information.

4.2 Model adaptation and parameter setting

The proposed MFMS CNN comprises 9 CBA layers, 8 max-pooling layers (including DS1, DS2, and DS3), 1 average pooling layer, and 1 dense layer. The number of filters in the CBA layers is 64, whereas the size and stride of each filter are 3×3 and 1 respectively. The window size and stride of the max-pooling layers are 3×3 and 2 respectively. For average pooling, the window and stride are 2×2 and 2 respectively. Both the CBA and the dense layers use the ReLU activation function. The initial images of size $256 \times 256 \times 3$ are down-sampled to $128 \times 128 \times 3$, $64 \times 64 \times 3$ and $32 \times 32 \times 3$ by the DS1, DS2 and DS3 units respectively. The Concat1 layer merges the feature maps generated by CBA2 and CBA3. Here, the size of each feature map for both cases is $126 \times 126 \times 64$. Similarly, the Concat2 and Concat3 layers combine the feature maps of size $63 \times 63 \times 64$, $63 \times 63 \times 64$ computed by CBA4 and CBA5, as well as $32 \times 32 \times 64$, $32 \times 32 \times 64$ computed by CBA6 and CBA7 respectively. The dense layer computes an output of size 512 based on a flattened input of size $8 \times 8 \times 64$. The MFMS CNN extracts a 512-D feature vector (f_i^j) for each image at the input, where i is the index of the image and $j \in 1, 2$ is the reference time of image acquisition reflecting images T1 and T2. This means that for each image pair i (consisting of images x_i^1, x_i^2), the model calculates two feature vectors (f_i^1 ,f_i^2), each having a size of 512. Moreover, the margin m of the contrastive loss is set to 6 (set empirically). Finally, the Adam [15] optimizer with a learning rate of 0.001 is used. The model was trained for 120 iterations with a batch size of 256.

5 Results

Evaluation measures : The precision (p), recall (r), and F-measure (f) (harmonic mean of p and r corresponding to the changed class) have been used to compare the result of the proposed method with state-of-the-art unsupervised approaches. Accuracy and mean intersection over union ($mIOU$) have been considered to compare with weakly supervised approaches. These measures are considered as the standard ones in literature [1, 14]. We acknowledge the unfairness of comparing our weakly supervised method with state-of-the-art unsupervised methods. However, we believe there is some value in this comparison because of the significantly less effort required to produce the weak labels compared to the effort required to produce the pixel-level change maps. We do not claim that our approach is superior to unsupervised methods nor do we suggest that we adapt the comparison as a head-to-head apposition of methods operating under the same regime. However, we note that significant performance improvement on the CD task can be achieved with minimal effort, by incorporating image-level labels.

Table 1 Comparison of the two unsupervised methods (U-A [16], U-B [4]) with the proposed scheme (WS-C). Results are in percentages.

Model	**Precision**	**Recall**	**F-Measure**
U-A	27.2	56.1	36.6
U-B	19.2	48.7	27.5
WS-C	**43.2**	**66.9**	**52.5**

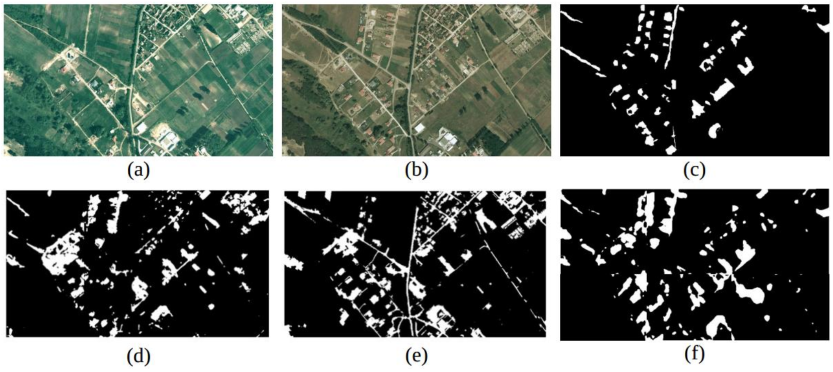

Fig. 4: Comparisons on the SZADA dataset. (a,b) Test image pair, (c) ground truth maps, (d) results obtained using U-A [16], (e) results of U-B [4], and (f) results obtained using the proposed method (WS-C).

Analysis of results using the SZADA dataset : The values for p, r, and f for the proposed approach and other state-of-the-art approaches are listed in Table 1. Here, the performance of the proposed methodology is compared with two unsupervised state-of-the-art schemes: Liu *et al.* [16] (U-A in Table 1) and S3VM [4] (U-B in Table 1). The results indicate that the performance of the proposed model (WS-C in Table 1) is significantly better than the two unsupervised schemes (U-A and U-B). In this regard, the proposed method outperforms the scheme U-A by a margin of $\approx 16\%$, $\approx 10\%$, and $\approx 15\%$ in terms of p, r, and f, respectively. Similarly, the proposed approach surpasses the scheme U-B by a margin of $\approx 24\%$, $\approx 18\%$, and $\approx 25\%$ in terms of p, r, and f respectively. Figure 4 shows example visual outputs of different methods on the SZADA dataset.

Table 2 Comparison of the two weakly supervised methods (WS-A [14], WS-B [1]) with the proposed scheme (WS-C). Results are in percentage.

Model	Accuracy	mIOU
WS-A	99.1	71
WS-B	99.2	70.3
WS-C	**99.5**	**74.3**

Analysis of results using the AICD dataset : The results obtained using the proposed model are compared with those obtained using two other weakly supervised state-of-the-art CD techniques [1, 14]. Here, the results obtained using the proposed approaches (WS-C in Table 2) and the two state-of-the-art approaches (Khan *et al.* [14] and Andermatt *et al.* [1]) are represented as WS-A, and WS-B, respectively in Table 2. It is observed that the proposed approach outperforms WS-A by a margin of 0.4% and 3.3% in terms of accuracy and *mIOU* respectively. Similarly, the proposed approach achieves a higher performance of 0.3% and 4% in terms of accuracy and *mIOU* respectively compared to WS-B. Figure 5 shows some example generated results obtained with the proposed method on the AICD dataset.

Table 3 Comparison of the two base methods (Base-1 [7], Base-2) with the proposed scheme (WS-C). Results are in percentages.

Model	Precision	Recall	F-Measure
Base-1	**46.5**	41.9	44.2
Base-2	39.6	65.3	49.3
WS-C	43.2	**66.9**	**52.5**

5.1 Ablation analysis of the proposed model

In this experiment, the proposed model is decomposed to its distinct components (Siamese network and PCA + K-means), performing the CD task separately on each component. In this way, we can assess the contribution of each component. The plain Siamese network is tested by replacing the PCA + K-means processing stage with a threshold on the computed DI. Specifically, we use the OTSU threshold [20] and report the performance by performing CD on the test set. We call this approach as *Base-2* in Table 3. Accordingly, we use the PCA + K-means on the DI computed directly from the images as suggested by [7] and not on the DI computed by the Siamese model. We call this method *Base-1* in Table 3. The experiments are carried out using the SZADA datasets. The results shown in Table 3 suggest that the proposed approach (WS-C in Table 3) outperforms the two schemes (Base-1 and Base-2) in terms of recall and F-measure scores. However, in terms of precision, the Base-1 strategy surpasses the proposed scheme. Thus, the proposed (complete) model incorporates the PCA + K-means technique on the DI and significantly improves the recall score of the results at the cost of a small decrease in the precision score.

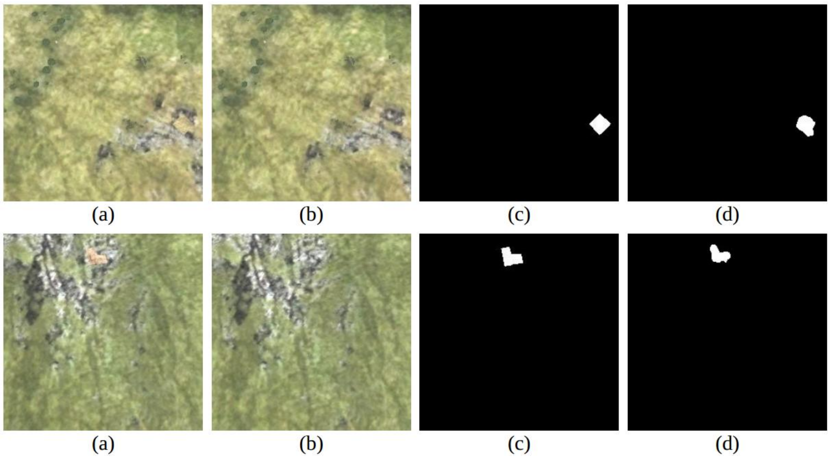

Fig. 5: Sample images of AICD dataset and their ground truth as well as the generated change map. (a,b) Test image pair, (c) corresponding ground truth information, (d) results obtained using proposed method (WS-C).

6 Conclusions

In this work, a weakly supervised change detection model is proposed for analyzing remotely sensed multi-temporal images. An MFMS CNN Siamese network is trained using the image-level labels of image pairs and not the pixel-level labels adding flexibility and removing complexity from tackling the task. The proposed model achieves huge improvements as compared to the unsupervised approaches by incorporating simple image-level labels. Moreover, it enhances the state-of-the-art weakly-supervised performance on the AICD dataset.

Acknowledgment

This project has received funding from the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

1. Andermatt, P., Timofte, R.: A weakly supervised convolutional network for change segmentation and classification. arXiv preprint arXiv:2011.03577 (2020)
2. Benedek, C., Szirányi, T.: Change detection in optical aerial images by a multilayer conditional mixed markov model. IEEE Transactions on Geoscience and Remote Sensing **47**(10), 3416–3430 (2009)
3. Bourdis, N., Marraud, D., Sahbi, H.: Constrained optical flow for aerial image change detection. In: International Geoscience and Remote Sensing Symposium. pp. 4176–4179. IEEE (2011)
4. Bovolo, F., Bruzzone, L., Marconcini, M.: A novel approach to unsupervised change detection based on a semisupervised svm and a similarity measure. IEEE Transactions on Geoscience and Remote Sensing **46**(7), 2070–2082 (2008)

5. Brunner, D., Lemoine, G., Bruzzone, L.: Earthquake damage assessment of buildings using vhr optical and sar imagery. IEEE Transactions on Geoscience and Remote Sensing **48**(5), 2403–2420 (2010)
6. Cao, G., Wang, B., Xavier, H., Yang, D., Southworth, J.: A new difference image creation method based on deep neural networks for change detection in remote-sensing images. International Journal of Remote Sensing **38**(23), 7161–7175 (2017)
7. Celik, T.: Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geoscience and Remote Sensing Letters **6**(4), 772–776 (2009)
8. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 1735–1742. IEEE (2006)
9. Hu, J., Chen, Z., Yang, M., Zhang, R., Cui, Y.: A multiscale fusion convolutional neural network for plant leaf recognition. IEEE Signal Processing Letters **25**(6), 853–857 (2018)
10. Ji, S., Shen, Y., Lu, M., Zhang, Y.: Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples. Remote Sensing **11**(11), 1343 (2019)
11. Ji, S., Wei, S., Lu, M.: 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), 574–586 (2018)
12. de Jong, K.L., Bosman, A.S.: Unsupervised change detection in satellite images using convolutional neural networks. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
13. Kalita, I., Roy, M.: Deep neural network-based heterogeneous domain adaptation using ensemble decision making in land cover classification. IEEE Transactions on Artificial Intelligence pp. 1–1 (2020). https://doi.org/10.1109/TAI.2020.3043724
14. Khan, S.H., He, X., Porikli, F., Bennamoun, M., Sohel, F., Togneri, R.: Learning deep structured network for weakly supervised change detection. arXiv preprint arXiv:1606.02009 (2016)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Liu, J., Chen, K., Xu, G., Sun, X., Yan, M., Diao, W., Han, H.: Convolutional neural network-based transfer learning for optical aerial images change detection. IEEE Geoscience and Remote Sensing Letters **17**(1), 127–131 (2019)
17. Liu, J., Gong, M., Qin, K., Zhang, P.: A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE transactions on neural networks and learning systems **29**(3), 545–559 (2016)
18. Meher, S.K., Kumar, D.A.: Ensemble of adaptive rule-based granular neural network classifiers for multispectral remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **8**(5), 2222–2231 (2015)
19. Mubea, K., Menz, G.: Monitoring land-use change in nakuru (kenya) using multisensor satellite data (2012)
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
21. 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 Geoscience and Remote Sensing Letters **14**(10), 1845–1849 (2017)
22. Zhong, P., Wang, R.: A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Transactions on Geoscience and Remote Sensing **45**(12), 3978–3988 (2007)