MITIGATING SPATIAL AND SPECTRAL DIFFERENCES FOR CHANGE DETECTION USING SUPER-RESOLUTION AND UNSUPERVISED LEARNING

Jonathan Prexl[1]*, Sudipan Saha*[1]*, Xiao Xiang Zhu*[1,2]

Data Science in Earth Observation, Technical University of Munich, Taufkirchen/Ottobrunn, Germany[1]
Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany[2]

ABSTRACT

Change detection (CD) is one of the most researched areas in remote sensing. However, most CD methods assume that the pre-change and post-change images are acquired by the same sensor, having the same set of spectral bands and same spatial resolution. This severely limits the applicability of CD methods. It is not trivial to apply the existing CD methods in multi-sensor scenario. Towards this direction, we propose an unsupervised CD method that can handle large differences in spatial resolution and can work with completely different set of spectral bands. The proposed method uses a self-supervised super-resolution strategy to upsample the lower resolution image, thus mitigating differences in spatial resolution. To mitigate spectral differences, a self-supervised learning strategy is used that ingests both images as input and trains a network using self-supervised loss accounting for the spectral differences in both images. Once trained this network is used in deep change vector analysis framework for change detection. We validated the proposed method in an experimental setup where the pre-change and post-change images have different spatial resolution (10 m and 20 m/pixel) and completely disjoint set of spectral bands.

***Index Terms*—** Change detection, Multisensor images, Multi-spatial resolution, Deep Change Vector Analysis, Deep learning.

1. INTRODUCTION

Change detection (CD) is an important topic in remote sensing. Owing to the difficulty of collecting labeled multi-temporal data, unsupervised CD methods are preferred in the literature [1]. However, most CD methods assume that pre-change and post-change images are acquired by the same sensor. While it is convenient to perform CD using same sensor for pre-/post-change images, temporal resolution of the most sensors are limited by revisit period of the satellite. This is a hindrance in use of optical multi-temporal analysis in time-bound applications, e.g., precision agriculture or disaster management. Using different sensors to form multi-temporal sequences may allow us to obtain temporal sequences with desirable sampling rate. However, it is not trivial to process multi-sensor multi-temporal images as they are affected by differences in the spatial resolution, differences in the spectral characteristics of the sensors.

There are only few works that can work in the setting where pre-change and post-change images have different spatial resolution [2, 3] or bands with different spectral characteristics [4]. Moreover, those works are designed to deal with only minor variations in spatial or spectral characteristics. Saha *et. al.* [2] proposed a cycle-consistent generative adversarial network (CycleGAN)-based method to learn transcoding between multi-sensor multi-temporal domain. However, their work deals with Quickbird (0.6 meter/pixel) and Pleiades (0.5 meter/pixel) images having similar spatial resolution and same set of spectral bands. Moreover, [2] assumes that large (unlabeled) scenes corresponding to both sensors are available apriori that can be used to train the CycleGAN. In practice, such large areas may not be always available for training model in real applications.

We design an unsupervised CD method that is based on popular Deep Change Vector Analysis (DCVA) framework [1], however we extend it to ingest images with large difference in spatial resolution and disjoint set of spectral bands (see Figure 1). This is a stark difference in comparison to the previous works in unsupervised multi-sensor CD [2, 4]. The proposed method exploits super-resolution [5] to mitigate differences in the spatial resolution. Following this a self-supervised learning strategy is used to train a network mitigating spectral differences. Subsequently the trained network is used as deep feature extractor in the DCVA framework. To summarize, the proposed method can be considered as an extension of the DCVA for the scenario where pre-change and post-change images are acquired with different spatial resolution and completely different spectral bands. While we do not claim any novelty in the super-resolution technique, our novelty lies in how we exploit super-resolution and self-supervised learning to process bi-temporal images with strong above-mentioned differences.

The rest of this paper is organized as follows. Section 2 outlines the proposed method. Datasets and experimental results are detailed in Section 3. Finally, we conclude the paper in Section 4.

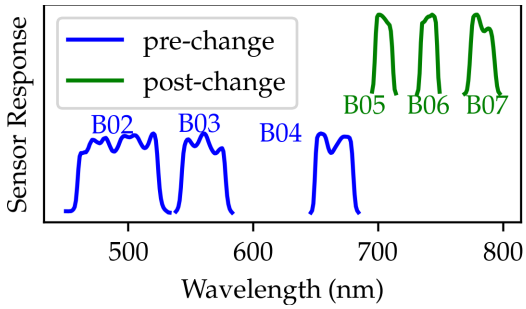

Fig. 1. The response function of the pre- and post-change imagery considered in this work. We considered pre-change and post-change images with completely different spectral bands.

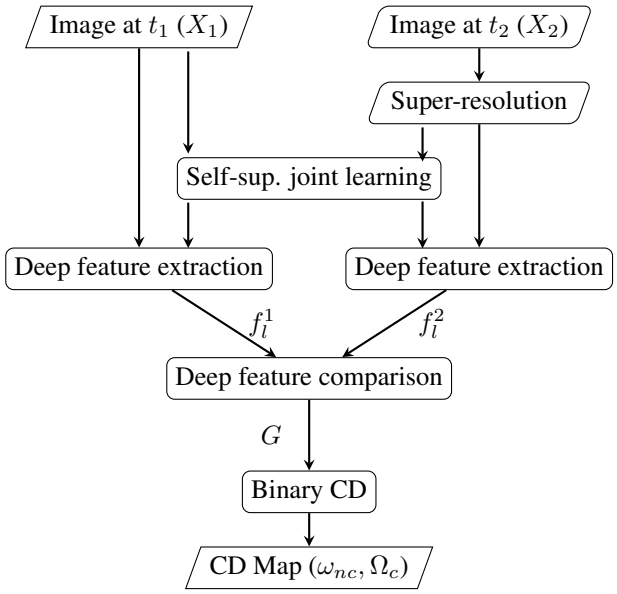

Fig. 2. Proposed CD framework

2. PROPOSED APPROACH

Let us consider two images X_1 (pre-change) and X_2 (post-change), acquired over the same geographical region at time t_1 and t_2. We assume that X_1 and X_2 show strong difference in spatial resolution. Moreover, their spectral bands also do not overlap or only slightly overlap. Our goal is to distinguish unchanged pixels (ω_{nc}) from the changed ones (Ω_c). For sake of simplicity and without loss of any generalization, let us assume that X_2 has lower resolution compared to X_1. We assume another set of unlabeled patches $\mathbf{Z_2} = \{Z_{2i}, \forall i = 1, ..., I\}$ are available drawn from the same distribution as X_2. We use the patches in $\mathbf{Z_2}$ to train a self-supervised upsampling network. Once trained, this network is used to upsample patches from X_2 to a higher resolution noted as X_2'. After mitigating difference in spatial resolution, we focus on mitigating spectral difference. This is done by learning a deep network in self-supervised fashion that takes into the both input images [6]. After self-supervised training, this network is employed using DCVA framework for CD. Figure 2 shows a block diagram of the proposed method.

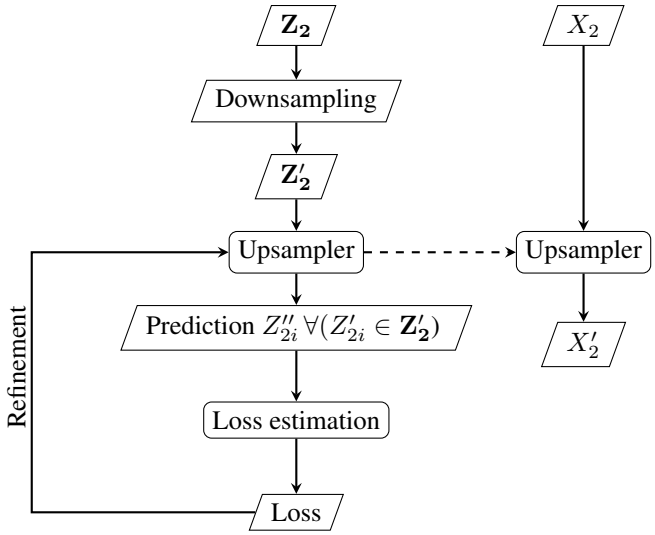

Fig. 3. The super-resolution method for the lower-resolution image X_2.

Spatial up-sampling: In multi spatial-resolution analysis, often it is easier to collect additional patches for the low resolution scene than the higher resolution one. Keeping this in mind and inpired from [7, 8], we design a superresolution technique which does not require X_1 (higher resolution in our case) or any patch similar to X_1. X_2 is upsampled entirely using X_2 and $\mathbf{Z_2}$, drawn from the same distribution as X_2. We define the desired upsampling α as the ratio between spatial resolution of X_1 and X_2. We use this α value to downsample images in $\mathbf{Z_2}$ to $\mathbf{Z_2}'$. Subsequently we process the images in $\mathbf{Z_2}'$ through a deep upsampling trainable model to obtain upsampled images $\mathbf{Z_2}''$. Loss is computed between the images in $\mathbf{Z_2}''$ and the original samples $\mathbf{Z_2}$ and is used to iteratively refine the upsampling model. Once the model is trained, it is used to upsample the lower resolution image X_2 to X_2' (Figure 3).

Self-supervised spectral differene mitigation and change detection: Once the pre-change and post-change images are projected to the same spatial resolution, we focus on mitigating spectral differences. This is achieved by learning a network jointly from X_1 and X_2' using self-supervised multi-temporal learning mechanism inspired from [6]. The self-supervised learning process exploits X_1 and X_2' in unsupervised way without using any label. In more details, X_1 and X_2' are processed through a series of trainable convolution layers. For pixels $x_{n,1}$ and $x_{n,2}$, we obtain features $y_{n,1}$ and $y_{n,2}$ in the last layer. Semantically similar pixels produce high activation in the same deep feature in the last layer. Using this hypothesis, arg-max classification is applied on the features in last layer to obtain the pseudo-labels $c_{n,1}$ and $c_{n,2}$. Following

this, cross-entropy loss $\mathcal{L}_1$ is computed using $c_{n,1}$ and $y_{n,1}$ to regulate the semantic consistency of X_1. Similarly, $\mathcal{L}_2$ is computed. Spectral difference between X_1 and X_2' is mitigated by forcing the network to achieve consistency between the features from one image and prediction from the other ($c_{n,1}$ and $y_{n,2}$ and vice-versa). In more details, for each input pixel $x_{n,1}$ and $x_{n,2}$,:

$$\ell_{n,1,2} = \text{crossentropy}(y_{n,1}, c'_{n,2}) \tag{1}$$

$$\ell_{n,2,1} = \text{crossentropy}(y_{n,2}, c'_{n,1}) \tag{2}$$

$\mathcal{L}_{1,2}$ is computed by averaging $\ell_{n,1,2}$ and $\ell_{n,2,1}$ over all pixels. $\mathcal{L}_{1,2}$ pushes the network to assign same label to same spatial location in X_1 and X_2', thus reducing gap between them in the featurespace of the trained network.

Once trained, the network is used in a in DCVA framework [1] to distinguish the unchanged pixels (ω_{nc}) from the changed ones (Ω_c). X_1 and X_2 are separately processed through the trained network and deep features are extracted from a set of layers in the network to form deep feature hypervector. Deep feature hypervectors are subtracted to obtain the deep change hypervector. Deep change hypervector (G) captures the multi-scale semantic information related to change. G is further processed to obtain deep magnitude ρ that encodes the hyper-dimensional G in just one dimension [1]. Assuming that unchanged pixels (ω_{nc}) generate similar deep features, however the changed pixels (Ω_c) generate dissimilar deep features, we distinguish changed pixels (Ω_c) from the unchanged ones (ω_{nc}) by using a threshold applied to ρ.

3. EXPERIMENTAL RESULTS

Owing to difficulty of acquiring datasets (with reference data) that are multi-sensor or adhere to the properties described previously in Section 1, we use the existing Onera Satellite Change Detection (OSCD) dataset [9] on the Cupertino city. To adhere to the characteristics outlined in Section 1, for the pre-change image we only take the band 4 (R), 3 (G), 2(B) which show 10 meter/pixel resolution. For post-change image, we only take the vegetation red edge bands (5, 6, 7) which show 20 meter/pixel. Thus in our experimental setup, pre-change and post-change images show completely different spatial resolution and mutually exclusive spectral bands. The very high difference between the chosen bands is evident in Figure 4.

We compare the proposed method to deep feature based DCVA [1] by using an ImageNet based feature extractor [10]. For fair comparison, this is also fed with the super-resolved and similarly pre-processed images as in the proposed method. We also design a domain-adaptation based variant of it by fine-tuning the ImageNet based feature extractor using difference between its output on pre-change and post-change images as loss.

Table 1. Quantitative binary CD result. Sensitivity and specificity are accuracy computed over the reference changed pixels and unchanged pixels, respectively [1]. They are shown in $\%$

Method	Sensitivity	Specificity
Proposed	67.34	92.00
Proposed (bilinear interpolation)	66.66	90.67
DCVA	48.87	88.26
DCVA with adaptation	49.83	88.38

Reference CD map is shown in Figure 5(a). Figure 5(b) shows the result obtained by the proposed method. For better visualization, a false color composition between the reference map and the obtained result is shown in Figure 5(c). Even though there are some false alarms, the number is fewer than in the compared methods. Moroever, this dataset was originally proposed in context of supervised CD and unsupervised CD performs suboptimally even when using same spectral bands in pre and post-change images [9]. Result obtained by the DCVA's domain adaptation based variant is shown in Figures 5(d). Quantitative result is shown in Table 1. It is evident that the proposed method clearly benefits from the super-resolution and the self-supervised learning. Furthermore, quantitative evaluation clearly shows the superiority of the proposed method over state-of-the-art unsupervised methods. This can be attributed to superior capability of the proposed method to ingest multi-sensor multi-temporal images.

(a) (b)

Fig. 4. Visualisation of the difference in resolution of (a) band 2 (prechange) and (b) band 5 (postchange) . Many smaller components in the urban area is distinguishable in (a), they are blurry in (b). Furthermore two spectral bands emphasize different areas, thus performing CD on them is challenging.

4. CONCLUSIONS

In this work we proposed an unsupervised binary CD method under the setting that pre-change and post-change images have completely different set of spectral bands and different spatial resolution. Our work extends DCVA framework to this challenging setting by effectively exploiting super-resolution and self-supervised learning. While the differences in the spatial resolution is mitigated by applying a self-supervised

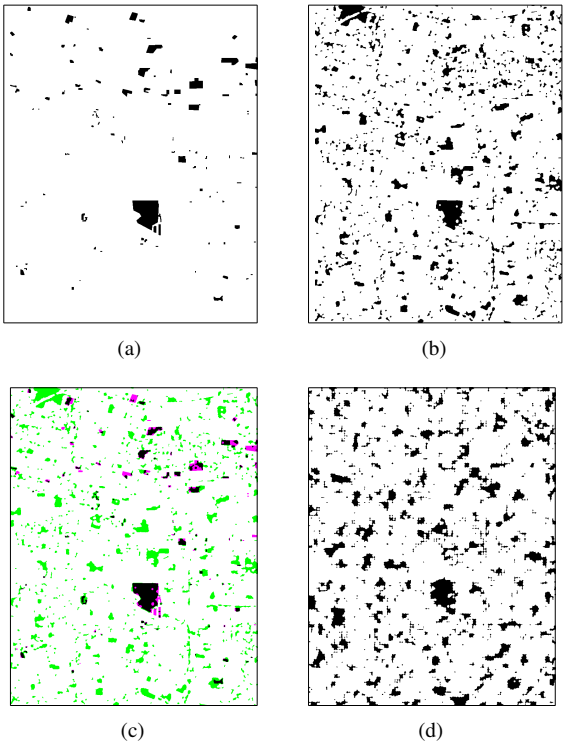

Fig. 5. Qualitative CD results for the Cupertino city. Change detection maps: (a) Reference, (b) Proposed, (c) FCC between reference and proposed (the correctly detected region are in black, false alarms are in green), (d) DCVA with adaptation.

super-resolution on the lower-resolution image, the differences in spectral bands are mitigated by joint self-supervised learning. The results show that the proposed method is capable to mitigate spatial and spectral differences. Our future development will be towards: 1) extending the method under the setting where pre-change and post-change images have different number of spectral bands, 2) extending the proposed method to distinguish between different kinds of changes (multiple CD), and 3) extending the proposed method for time-series analysis.

Acknowledgement

The work is funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

5. REFERENCES

[1] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[2] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised multiple-change detection in VHR multisensor images via deep-learning based adaptation," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5033–5036, IEEE, 2019.

[3] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 24–41, 2016.

[4] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS J Photogramm Remote Sens.*, vol. 107, pp. 50–63, 2015.

[5] S. Pendurkar, B. Banerjee, S. Saha, and F. Bovolo, "Single image super-resolution for optical satellite scenes using deep deconvolutional network," in *International Conference on Image Analysis and Processing*, pp. 410–420, Springer, 2019.

[6] S. Saha, L. Mou, C. Qiu, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Unsupervised deep joint segmentation of multitemporal high-resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[8] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[9] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2115–2118, IEEE, 2018.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.