Downloaded from orbit.dtu.dk on: Apr 30, 2024

DTU Library

A Difference Enhanced Neural Network for Semantic Change Detection of Remote Sensing Images

Wang, Renfang; Wu, Hucheng; Qiu, Hong; Liu, Xiufeng; Wang, Feng; Cheng, Xu

Published in:
IEEE Geoscience and Remote Sensing Letters

Link to article, DOI:
10.1109/LGRS.2023.3310676

Publication date:
2023

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Wang, R., Wu, H., Qiu, H., Liu, X., Wang, F., & Cheng, X. (2023). A Difference Enhanced Neural Network for Semantic Change Detection of Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, *20*. https://doi.org/10.1109/LGRS.2023.3310676

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

A Difference Enhanced Neural Network for Semantic Change Detection of Remote Sensing Images

Renfang Wang, Hucheng Wu, Hong Qiu, Feng Wang, Xiufeng Liu, and Xu Cheng

***Abstract*—Deep learning techniques have been widely used for semantic change detection (SCD) of remote sensing images (RSIs) and have shown encouraging performance. In this paper, we propose a novel neural network by embedding the difference enhancement (DE) module into the adjacent layers of ResNet for SCD of RSIs (DESNet), which can pay more attention to the changes of bi-temporal RSIs. Furthermore, we deploy the module of multi-scale parallel sampling spatial-spectral non-local (SSN) after feature extraction, which can effectively improve the robustness to large-scale changes and the integrity of the changed objects by fusing global features that sampled from the multi-scale feature space. The experimental tests demonstrate that our DESNet can achieve state-of-the-art accuracy on the SECOND dataset and the LandSat-SCD dataset.**

***Index Terms*—Remote sensing image; Semantic change detection; Difference enhancement; Deep learning**

I. Introduction

CHANGE detection (CD) is to obtain the changed objects by joint analysis of two (or more) RSIs that obtained in the same area and at different times[1, 2, 3], and it has been applied to various kinds of real-world applications including land and resource survey, environmental monitoring, and urban management[4, 5].

Deep learning is a type of powerful numerical tools for extracting features and has been very popular in the community of CD. Zhang and Lu [6] proposed a spectral-spatial joint learning network using a Siamese CNN (convolutional neural network) to extract a dual-temporal spectral-spatial joint representation. To address the lack of resilience to pseudo-variation information, Chen et al. [1] introduced a dual-attentive fully convolutional siamese neural network (DASNet) that employs weighted double marginal contrast loss (WDMC) to solve the sample imbalance issue. Chen et al. [7] and Dong et al. [8] proposed to improve the capacity to describe contextual information in the network with a dual-time spatial and temporal domain-based Transformer.

Although binary change detection (BCD) can provide the information about the location and the geometry of changes, the information is usual coarse-grained, and it can not describe the types of changes. On the contrary, SCD approaches can provide the location and the geometry of changes as well as the types of changes. Yang et al. [9], Ding et al. [10], and Mou et al. [11] introduced a triple-branch change detection paradigm in which two semantic segmentation branches divide the dual-temporal pictures into LCLU (Land Cover and Land Use) mappings, respectively, and the third one is used to identify the changes. Yang et al. [9] developed an asymmetric siamese network (ASN) for locating and identifying semantic changes by incorporating gating and weighting schemes into the decoder. Ding et al. [10] discussed the possible network architecture for semantic change detection and demonstrated that the late fusion method of separating semantic segmentation tasks and change detection tasks (SSCD1) is appropriate for semantic change detection. Yang et al. [9], Ding et al. [10], and Caye Daudt et al. [12] designed deep neural networks with three CNNs that can extract the semantic information and changes individually.

The above-mentioned SCD approaches have shown outstanding performance, however, two main issues are still remained: 1) For large-scale variation in RSIs, existing models are not sensitive enough to the edge of the changed objects. False alarms and missed alarms often occur at the edge of changed targets; 2) Many of them failed to capture the tiny discontinuous changes (e.g., vegetation degradation) in localized objects. Additionally, the non-local block [13] can obtain the global correlation using a self-attention mechanism, which benefits to capture long-range dependencies among inputs. Lei et al. [2] and Yuan et al. [14] improved the feature extraction by increasing the size of receptive field.

In remainder of this paper, we firstly report how to construct DESNet by embedding the DE module into the adjacent layers of ResNet to ensure the network can focus on the changes of bi-temporal RSIs, and we then introduce the SSN module, combines multi-scale spatial global features to simulate large-scale variations, to enhance the integrity of the changed objects. Finally, the performance of the experimental tests on the SECOND dataset and the LandSat-SCD dataset demonstrate the superiority of DESNet in terms of SCD accuracy and preservation of the integrity of changed objects.

This work was supported in part by the National Natural Science Foundation of China under Grant 61906170, in part by the Project of the Science and Technology Plan for Zhejiang Province under Grant LGF21F020023, and in part by the Plan Project of Ningbo Municipal Science and Technology under Grant 2021Z050, 2022Z233, 2022S002 and 2023J403 (*Corresponding author: Hong Qiu*).

Renfang Wang, Hong Qiu, and Feng Wang are with the College of big data and software engineering, Zhejiang Wanli University, Ningbo, 315200, China (e-mail: renfang_wangac@126.com; qiuhong@zwu.edu.cn; wangf_721@zju.edu.cn).

Hucheng Wu is with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China (e-mail: wuhucheng1@163.com).

Xiufeng Liu is with the Department of Technology, Management and Economics, Technical University of Denmark, Produktionstorvet, Denmark (e-mail: xiuli@dtu.dk).

Xu Cheng is with the Section of Energy Markets Smart Innovation Norway Halden, Norway (e-mail: xu.cheng@ieee.org).

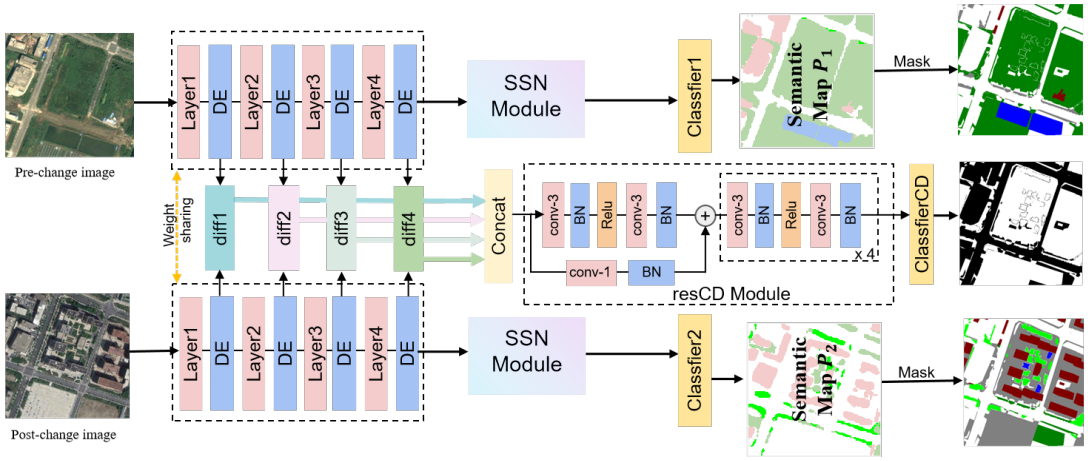

Fig. 1. Illustration of DESNet. Layer-n ($n = 1, 2, 3, 4$) represents the n-residual block of the ResNet34, conv-1 denotes the convolution with kernel size of 1×1, and conv-3 denotes the convolution with kernel size of 3×3.

II. Methodology

Given a pair of multi-temporal image I_1 and I_2, the goal of SCD is to find the mapping function that can present the changed areas and their semantic categories.

$$F_{scd}(I_1, I_2) = \begin{matrix} (0,0), & L_1^P = L_2^P \\ (L_1^P, L_2^P), & L_1^P \neq L_2^P \end{matrix}, \forall P \in I, \quad (1)$$

where P represents the same spatial location of the multi-temporal image I_1 and I_2, and L_1^P, L^P are the semantic classes of bi-temporal at P.

Inspired by [10] that utilizes a siamese CNN encoder to extract features from dual-temporal images. We develop a modified siamese ResNet34 as the backbone for extracting features, which embeds DE module and SSN module into ResNet34 to improve the SCD accuracy (Fig. 1). The neural network utilizes ResNet34 with an embedded DE module (weight sharing) for dual time-phase remote sensing image processing. This extracts semantic and multiscale variation features from the dual time-phase data. The multiscale variation features merge into the binary variation detection branch, creating a binary variation detection map. Concurrently, the SSN module enhances semantic features by fusing spatial and spectral information. The resulting semantic segmentation output undergoes a mask multiplication operation with the binary change detection map, ultimately producing the dual time-phase semantic change map.

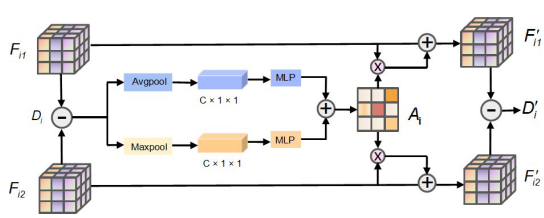

Fig. 2. Architecture of the DE module.

1) DE module: In SSCDl[10], the semantic features extracted by the backbone network lack attention to changes, which can hinder the detection of objects that have undergone a change. Therefore, we embed the DE module between adjacent layers of the feature extraction backbone. This allows the network to learn the different information of dual-time images and establish semantic interaction among the dual-time branches. Hence, the network can filter out irrelevant changes and focus on the objects that have truly changed. The structure of the DE module is shown in Fig. 2. The original bi-temporal images are fed into the feature encoder followed by inputting the extracted feature maps to the DE module to enhance the bi-temporal difference features. The encoder consists of four feature extraction layers. The difference feature D_i can be written as

$$D_i = |F_{i1} - F_{i2}|, \quad (2)$$

where F_{i1} and F_{i2} represent the outputted bi-temporal feature maps of the i-th layer from the encoder, and $|\cdot|$ represents the absolute value operation to ensure D_i is meaningful.

We obtain the attention maps by

$$A_i = \sigma(MLP(Maxpool(D_i)) + MLP(Avgpool(D_i))), \quad (3)$$

where A_i represents the attention map after the channel attention operation. MLP denotes the multilayer perceptron network, Maxpool and Avgpool represent the max-pooling operation and average-pooling operation, respectively, and σ is the sigmoid function.

Once the attention map achieved, we can enhance the original features by

$$F_i' = A_i \times F_{in} + F_{in}, \quad (4)$$

where $i = \{1, 2, 3, 4\}$ indexes the layer for feature extraction, $n = \{1, 2\}$ indexes the dual-branch path and F_{in} denotes the output of the i-th layer for feature extraction on the n-th branch of the encoder.

The enhanced difference feature maps obtained by performing subtraction operations on F_{i1}' and F_{i2}' can be expressed as

$$D_i' = |F_{i1}' - F'|. \quad (5)$$

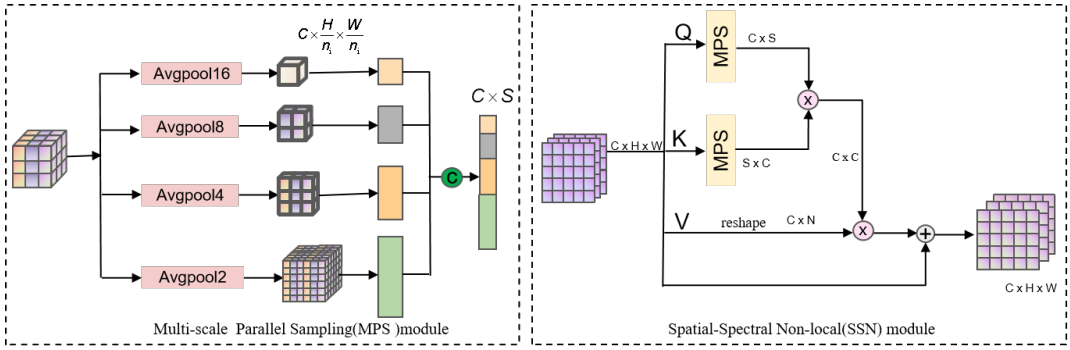

Fig. 3. Architecture of the SSN module.

2) SSN module: To enhance the modeling capability of CNN in space-time domain, the non-local mechanism is introduced into CD and can improve the detection accuracy by capturing the long-range correlation of pixels. However, the conventional non-local module usually failed to detect the small-scale changes and to keep the good integrity of the changed objects. As a result, we introduce SSN module after the feature encoder (Fig. 3).

In non-local module, features output from the CNN encoder will be fed into three branches Q, K, and V individually, where $\{Q,K,V\} \in \mathrm{R}^{C\times H\times W}$. We can obtain the sampled features at multi-scales by

$$MS_n = Avgpool_n(x), \tag{6}$$

where $x \in \mathrm{R}^{C\times H\times W}$ denotes the features, n indicates the scale of pooling, and MS_n stands for the sampled map. Four sampling scales ($n = 2, 4, 8, 16$) are used in DESNet. We reshape these feature maps to $\mathrm{R}^{C\times M_n}$, with $M_n = (H/n) \times (W/n)$. Thus, we define MPS as

$$MPS(x) = Cat(R_2, R_4, R_8, R_{16}), \tag{7}$$

where Cat denotes the concatenation operation, and $R_n(n = 2, 4, 8, 16)$ stands for the feature maps at four different sampling scales. The shape of the output of MPS is $C \times S$, with $S = \sum_{n=2,4,8,16} M_n$.

We then obtain the interrelationship using the global features,

$$M_c = MPS(Q) \times MPS(K)^T, \tag{8}$$

where T is transpose for matrix, and $M_c \in \mathrm{R}^{C\times C}$ is the channel attention matrix.

Thus, the augmented feature map F_{out} can be obtained by

$$F_{out} = softmax(M_c) \times V + X, \tag{9}$$

where $softmax$ is the softmax function, X and F_{out} are the input and out of SSN, respectively.

The loss function used for training DESNet consisted of semantic loss, binary change loss, and consistency loss. The semantic loss L_{seg} can be written as

$$L_{seg} = -\frac{1}{n}\sum_{i=1}^{N} y_i log(p_i), \tag{10}$$

where N is the number of semantic classes, and y_i and p_i denote the GT (Ground Truth) label and the predicted probability of the i-th class, respectively. The binary change loss is defined as

$$L_{change} = -y_c log(p_c) - (1 - y_c) log(1 - p_c), \tag{11}$$

where y_c and p_c denote the GT label and the predicted probability of change, respectively. The semantic consistency loss L_{sc} is expressed as

$$L_{sc} = \begin{cases} 1 - cos(x_1, x_2), & y_c = 1 \\ cos(x_1, x_2), & y_c = 0 \end{cases}, \tag{12}$$

where x_1 and x_2 are feature vectors of a pixel on predicted semantic maps P_1 and P_2, respectively, and y_c is the value at the same position on GT semantic change maps L_c. Therefore, we can obtain the loss function L_{scd} by

$$L_{scd} = L_{seg} + L_{change} + L_{sc}. \tag{13}$$

III. Experimental Results

We use two benchmark dataset to test the effectiveness of DESNet, including the SECOND dataset [9] and the LandSat-SCD dataset [14]. In the SECOND dataset, all RSIs have a size of 512 × 512 pixels and are annotated at pixel level. In the annotated labels, there are 1 class without changes and 5 LC classes, namely non-vegetated land surface, trees, low vegetation, water, buildings, and playgrounds. For the test on the SECOND dataset, we split this dataset into a training set and a test set with the numeric ratio of 4:1 (i.e., 2375 RSIs for training and 593 RSIs for testing). The RSIs in Landsat-SCD have been annotated into the class without changes and 4 LC classes including farmland, desert, building and water (only the changed areas are annotated). The Landsat-SCD datset has 8468 RSIs, with the spatial resolution of 416 × 416. We split them into training, validation and test sets with 7468, 560 and 560 RSIs, respectively, following the numeric ratio of 8:1:1.

We conduct the network training and testing on a workstation with NVIDIA GeForce RTX 3060. We set the same hyperparameters for the two experimental tests. We set the batch size to 4, the epoch of training to 100 and the initial learning rate to 0.01, respectively. We adopt stochastic gradient descent (SGD) method to optimize the weights. Also, we augment the training data by flipping and/or rotating the RSIs. We use the

overall accuracy (OA), mean Intersection over Union (mIou), Separated Kappa (Sek) coefficient and SCD-targeted F1 Score (F_{scd})[10] to quantify the effectiveness of DESNet.

Assume Q = $\{q_{i,j}\}$ is the confusion matrix, where $q_{i,j}$ represents the number of pixels that are classified into class i while their ground truth index is j (i, j $\in \{0, 1, ..., N\}$ (0 represents no-change). The OA represents the numeric ratio between the correctly classified pixels and the total image pixels, which can be defined as

$$OA = \sum_{i=0}^{N} q_{ii} / \sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij}. \tag{14}$$

Also, we can obatin mIou by averaging the Ious between the non-changed and changed classes,

$$mIou = \frac{(Iou_{nc} + Iou_{c})}{2}, \tag{15}$$

where $Iou_{nc} = q_{00} / (\sum_{i=0}^{N} q_{i0} + \sum_{j=0}^{N} q_{0j} - q_{00})$, and $Iou_{c} = \sum_{i=1}^{N} \sum_{j=1}^{N} q_{ij} / \sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij} - q_{00}$.

Sek evaluates the segmentation of semantic classes, especially in the changed areas. It is calculated based on the confusion matrix $Q' = \{q'_{ij}\}$, where $q'_{ij} = q_{ij}$ except that $q'_{0} = 0$. The formula of Sek can be express as

$$Sek = Kappa * e^{Iou_{c}^{-1}}, \tag{16}$$

where $Kappa = (p_0 - p_e)/(1 - p_e)$ with $p_0 = \sum_{i=0}^{N} q'_{i} / \sum_{i=0}^{N} \sum_{j=0}^{N} q'_{ij}$ and $p_e = \sum_{i=0}^{N} (\sum_{j=0}^{N} q'_{i} * \sum_{j=0}^{N} q'_{ji}) / (\sum_{i=0}^{N} \sum_{j=0}^{N} q'_{ij})^2$.

F_{scd} is used to evaluate the segmentation accuracy of the changed objects. It can be written as

$$F_{scd} = \frac{2 \times P_{scd} \times R_{scd}}{P_{scd} + R_{scd}}, \tag{17}$$

where $P_{scd} = \sum_{i=1}^{N} q_{ij} / \sum_{i=1}^{N} \sum_{j=0}^{N} q_{ij}$, and $R_{scd} = \sum_{i=1}^{N} q_{ij} / \sum_{i=0}^{N} \sum_{j=1}^{N} q_{ij}$.

Table. I and Table. II list the average OA, F_{scd}, Sek, and mIou results of different methods for the SECOND dataset and the LandSat-SCD dataset. As one can see, DESNet outperforms the competed methods. For Landsat-SCD, DESNet outperforms BiSNet by about 0.94 in mIoU, 1.95 in Sek, and 1.24 in F_{scd}. Fig. 4 and Fig. 5 show the visual comparisons of different methods for the SECOND dataset and the LandSat-SCD dataset, respectively. It can be seen that DESNet is more sensitive to the small-scale changes and can produce good integrity. Certainly, we demonstrate the computational efficiency of the networks by calculating their inference times on a pair of input images, as shown in Table. I.

Furthermore, we test the effectiveness of the DE module by introducing it into the baseline to train the DE-base. From Tab. III, we can obviously observe that DESNet outperforms DE-base about 0.7% in SeK, 0.62% in mIou and 0.8% in F_{scd}, which may indicate that the difference information has received more attention. In addition, DESNet outperforms SSN module by about 0.86% in SeK, 0.59% in mIou and 0.82% in F_{scd}.

TABLE I
COMPARISON WITH THE SOTA METHODS FOR SCD ON THE SECOND DATASET

Method	OA(%)	F_{scd}(%)	Sek(%)	mIou(%)	time(s)
HRSCD-Str3[12]	82.97	49.11	8.52	63.91	0.15
HRSCD-Str4[12]	85.10	54.12	13.09	67.69	0.16
SSCDl[10]	86.86	58.34	17.33	70.19	0.14
BiSRNet[10]	86.65	58.86	17.96	70.56	0.15
DESNet	**86.82**	**58.75**	**17.97**	**70.73**	0.22

TABLE II
COMPARISON WITH THE SOTA METHODS FOR SCD ON THE LANDSAT-SCD DATASET

Method	OA(%)	F_{scd}(%)	Sek(%)	mIou(%)
HRSCD-Str3[12]	84.47	57.47	19.24	79.08
HRSCD-Str4[12]	83.04	51.24	13.57	78.54
SSCDl[10]	89.20	74.54	37.12	80.25
BiSRNet[10]	89.72	76.24	38.82	80.32
DESNet	**90.43**	**77.18**	**40.77**	**81.56**

IV. CONCLUSION

In this paper, we propose a neural work for SCD of RSIs, where the DE module is embedded into the adjacent layers of ResNet and the SSN module is deployed after the module of feature extraction. DESNet can enhance the difference features when establish the connection among the bi-temporal semantic branches. Also, our DESNet can capture the multi-scale and long-range contexts to enhance the integrity of the changed objects, which beneficial to improve its robustness to the small-scale changes. Experimental results on on the SECOND dataset and the LandSat-SCD dataset demonstrate that DESNet can effectively enhance the integrity of the changed objects and can achieve higher SCD accuracy compared with the competitive methods.

REFERENCES

[1] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "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*, vol. 14, pp. 1194–1206, 2021.

[2] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference Enhancement and Spatial–Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

TABLE III
QUANTITATVE RESULTS OF THE ABLATION STUDY. THE BEST VALUES ARE IN BOLD

Method	Components			OA(%)	F_{scd}(%)	Sek(%)	mIou(%)
	DE	SSN	L_{sc}				
base				84.84	52.19	11.52	66.83
L_{sc}-base			✓	86.77	57.46	16.40	69.57
DE-base	✓		✓	**86.99**	58.26	17.10	70.19
SSN-base		✓	✓	86.98	58.28	17.26	70.16
DESNet	✓	✓		84.88	53.13	12.41	67.27
DESNet	✓	✓	✓	86.82	**58.75**	**17.97**	**70.73**

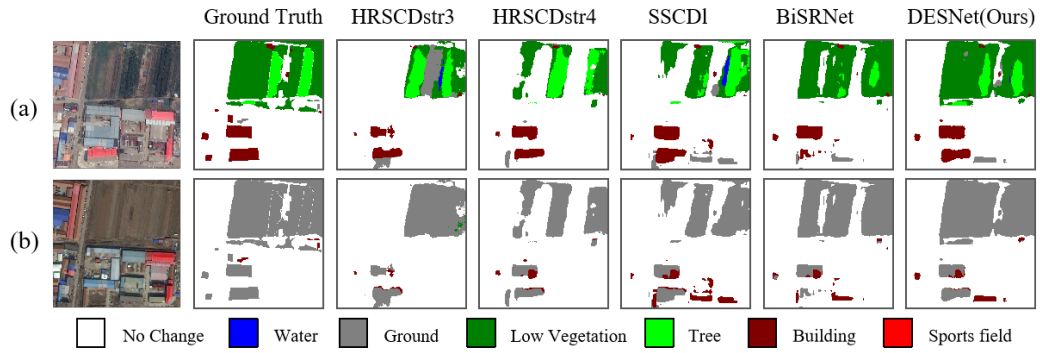

Fig. 4. Visual comparisons of DESNet and the state-of-the-art approaches on the SECOND dataset. (a) Pre-change Image. (b) Post-change Image.

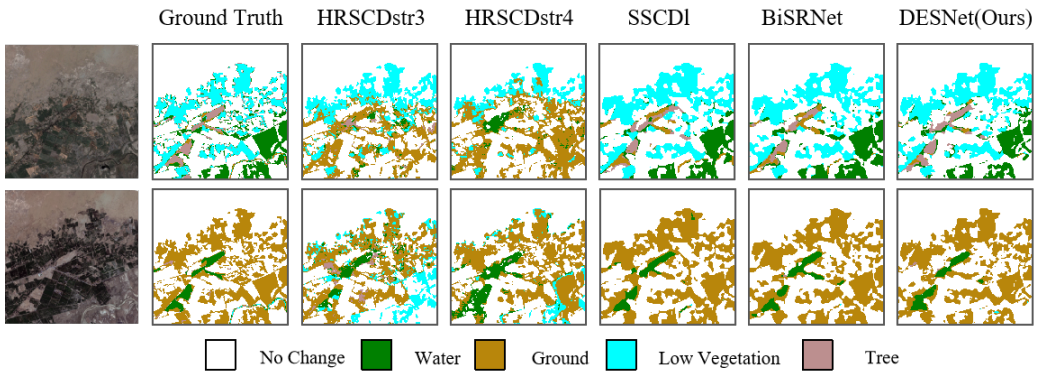

Fig. 5. Visual comparisons of DESNet and the state-of-the-art approaches on the LandSat-SCD dataset. (a) Pre-change Image. (b) Post-change Image.

[3] Y. Zhang, Y. Zhao, Y. Dong, and B. Du, "Self-Supervised Pretraining via Multimodality Images With Transformer for Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[4] E. Ghaderpour, P. Mazzanti, G. S. Mugnozza, and F. Bozzano, "Coherency and phase delay analyses between land cover and climate across Italy via the least-squares wavelet software," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103241, Apr. 2023.

[5] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep Learning-Based Change Detection in Remote Sensing Images: A Review," *Remote Sensing*, vol. 14, no. 4, p. 871, Feb. 2022.

[6] W. Zhang and X. Lu, "The Spectral-Spatial Joint Learning for Change Detection in Multispectral Imagery," *Remote Sensing*, vol. 11, no. 3, p. 240, Jan. 2019.

[7] H. Chen, Z. Qi, and Z. Shi, "Remote Sensing Image Change Detection With Transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[8] W. Dong, Y. Yang, J. Qu, S. Xiao, and Y. Li, "Local Information-Enhanced Graph-Transformer for Hyperspectral Image Change Detection With Limited Training Samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[9] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[10] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-Temporal Semantic Reasoning for the Semantic Change Detection in HR Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[11] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[12] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, p. 102783, Oct. 2019.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018, pp. 7794–7803, place: Salt Lake City, UT, USA.

[14] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images," *International Journal of Digital Earth*, vol. 15, no. 1, pp. 1506–1525, Dec. 2022.