# EFFECT OF TRANSFERRING PRE-TRAINED WEIGHTS ON A SIAMESE CHANGE DETECTION NETWORK

M. Aghdami-Nia [1], R. Shah-Hosseini [1, *], M. Salmani [1]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran - (aghdami.niya, rshahosseini, salmani.mohammad)@ut.ac.ir

**Commission IV, WG IV/3**

**KEY WORDS:** Remote sensing, Deep learning, Change detection, Transfer learning, Siamese network, Autoencoder.

## ABSTRACT:

Change Detection (CD) is one of the most crucial applications in remote sensing which identifies meaningful changes from bitemporal images taken from the same location. Enhancing the temporal efficiency and accuracy of this task is of great importance and one way to achieve this is through transfer learning. In this study, we investigate the influence of transferring pre-trained weights on the performance of a Siamese CD network using a benchmark dataset. For this purpose, an autoencoder with the same encoder architecture as in the Siamese model is trained on the whole dataset. Then, the encoder weights are transferred from the autoencoder and the Siamese model is trained in two modes. In the first mode, the transferred weights are frozen and only the decoder section of the Siamese models is trained while the second mode trains the whole model without freezing any part of the model. Moreover, the Siamese model is also trained without using the pre-trained weights to set the basis for comparisons. The results indicate that freezing the encoder results in a relatively lower performance but offers a considerable amount of temporal efficiency in the training phase. On the other hand, training the whole model after the weight transfer acquires the best result with an improvement of 12.43% in the Intersection over Union (IoU) metric.

## 1. INTRODUCTION

The capability of remote sensing (RS) in acquiring recurrent images from a given spatial location has enabled the study of land changes. The global coverage of these images facilitates the detection of variations in the land cover types around the world. That’s why change detection (CD) is one of the main topics in the RS community. CD is the process of identifying meaningful differences in the optical spectrum of a given object over a period of time. In the case of satellite imagery, this involves the processing of at least two distinct coregistered images taken from the same area. Many phenomena can cause changes on the ground surface including deforestation, climate change, and drought. Binary CD tries to identify whether any change has happened, disregarding the change type. This method for CD is usually considered a pre-processing step for more complex frameworks where change classes are also detected.

Deep learning (Aghdami-Nia et al., 2022; Ansari et al., 2021; Rostami et al., 2022b) and machine learning (Ranjbar et al., 2021; Rostami et al., 2022a; Zarei et al., 2021) methods have dominated many RS fields and CD is no exception. Siamese networks have attracted special attention for CD purposes among researchers in recent years. Caye Daudt et al. (Caye Daudt et al., 2018) were among the first studies to make use of end-to-end Siamese networks for CD purposes. They developed three Fully Convolutional Neural Network (FCNN) architectures namely FC-EF, FC-Siam-conc, and FC-Siam-diff, achieving better performance than the previous state-of-the-art methods. The lack of large-scale datasets was a limiting factor for developing more complicated CD architectures. In the following study, Caye Daudt et al. (Daudt et al., 2019) tackled this issue by presenting the first large-scale VHR semantic CD dataset, enabling them to perform CD and land cover mapping simultaneously. Zhang et al. (Zhang et al., 2020) developed the deeply supervised image fusion network IFN for CD in high-resolution bi-temporal remote sensing images. Their model differed from the other deep feature-based methods that were modified forms of architectures originally proposed for single-image semantic segmentation and outperformed the state-of-the-art methods.

Chen and Shi (Chen and Shi, 2020) introduced the use of spatial information for the sake of CD in VHR satellite imagery. The novel Siamese-based spatial-temporal attention neural network in their study was able to model the spatial-temporal relationships employing a CD self-attention module for feature extraction. They also released the benchmark dataset LEVIR-CD, which was two orders of magnitude larger than other public datasets at the time. Yang et al. (Yang et al., 2020) presented the asymmetric Siamese network ASN to locate and identify semantic changes through feature pairs obtained from modules of widely different structures. They created the large-scale well-annotated semantic CD dataset (SECOND) and developed an adaptive threshold learning module to better train and evaluate the model. Papadomanolaki et al. (Papadomanolaki et al., 2021) proposed a deep multi-task learning framework called L-Unet which was able to couple semantic segmentation and change detection using fully convolutional long short-term memory (LSTM) networks. This network resulted in a significant decrease in false-positive detections, with the F1-score outperforming other state-of-the-art methods in a few benchmark datasets.

This paper aims to study the effect of transferring pre-trained weights on the performance of a Siamese CD network, with the aim of improving segmentation accuracy. Firstly, a normal Siamese network is trained on a benchmark CD dataset to set

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-X-4-W1-2022-19-2023 |

the basis for comparisons. Subsequently, the whole dataset is fed into an autoencoder network that has the exact same encoder architecture as the Siamese network. Finally, the weights in the encoder part of the autoencoder are transferred to the encoder part of the Siamese network after the training.

## 2. DATASET

Yang et.al (Yang et al., 2020) presented the CD dataset SECOND with the aim of setting up a benchmark dataset that has a proper size and enough categories. They generated 4662 image pairs of size 512×512 from various VHR sensors that were collected from multiple locations such as Hangzhou and Shanghai. Moreover, the images were manually annotated at the pixel level by a group of experts. The change categories were comprised of 6 natural and man-made classes that are frequently prone to change including non-vegetated ground surface, tree, low vegetation, water, buildings, and playgrounds. Overall, 30 change categories can be extracted from these 6 classes including no-change. Several image pairs from the SECOND dataset can be seen in Figure 1.

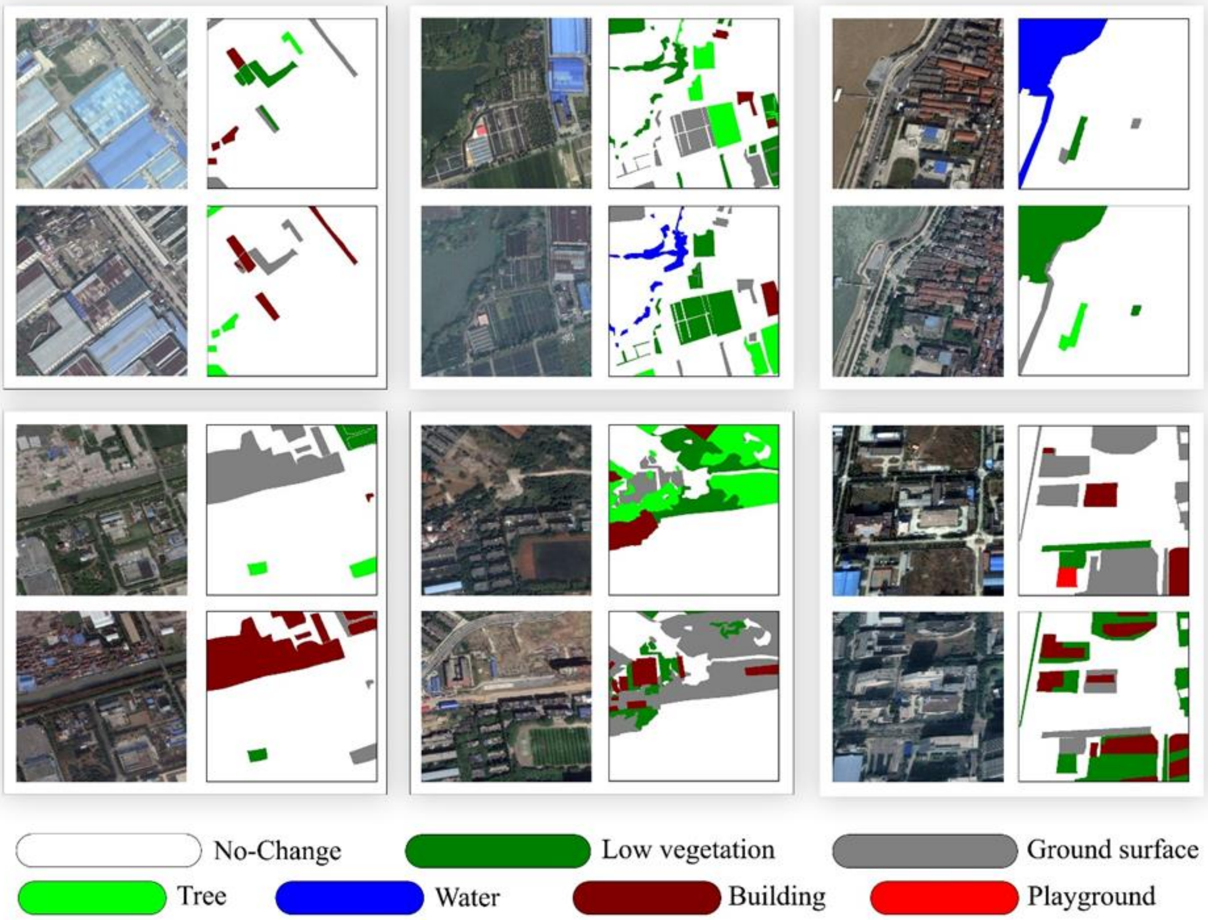

**Figure 1.** Multiple samples from the SECOND benchmark dataset. The white region is associated with no-change and other colors depict different land-cover classes.

## 3. METHODOLOGY

There is usually a high correlation between neighboring pixels in VHR images causing challenges in CD. Therefore, many conventional methods lead to poor results on these images. It has been proven that deep convolutional networks (CNNs) are more robust to this problem compared to conventional methods. Siamese networks are a branch of CNNs that are more suited for CD problems. These networks are based on the encoder-decoder architecture similar to networks such as U-Net. They have two encoder branches, each one taking an image from a pair as input. These branches have the same architecture and share the same weights. While training, the weight parameters in each branch are updated together simultaneously. These two encoder branches are in fact a single network that does the encoding on two different images at the same time. The encoded feature maps from two images are fed to the decoder branch which determines the similarity of two corresponding pixels in the two images. The output of the network is a pixel-wise probability map in the range of 0 to 1, representing the probability of change.

The Siamese network used in this study is based on the FC-Siam-conc proposed by Caye Daudt et.al (Caye Daudt et al., 2018). It takes image pairs of size 512×512 as input and passes each one through the Siamese encoder branches. The overall architecture is similar to U-Net with 4 encoder and decoder blocks individually. The number of convolutional filters in the first convolution block is set to 16, doubling every next block. After the bottleneck block, the number of convolutional filters halves until it reaches 16 in the final decoder block. In the end, the extracted feature vector of size 16 is mapped to 2 classes representing change and no-change areas. The change detection process happens when the lower level features from the two encoder branches are concatenated with the corresponding layer in the decoder branch.

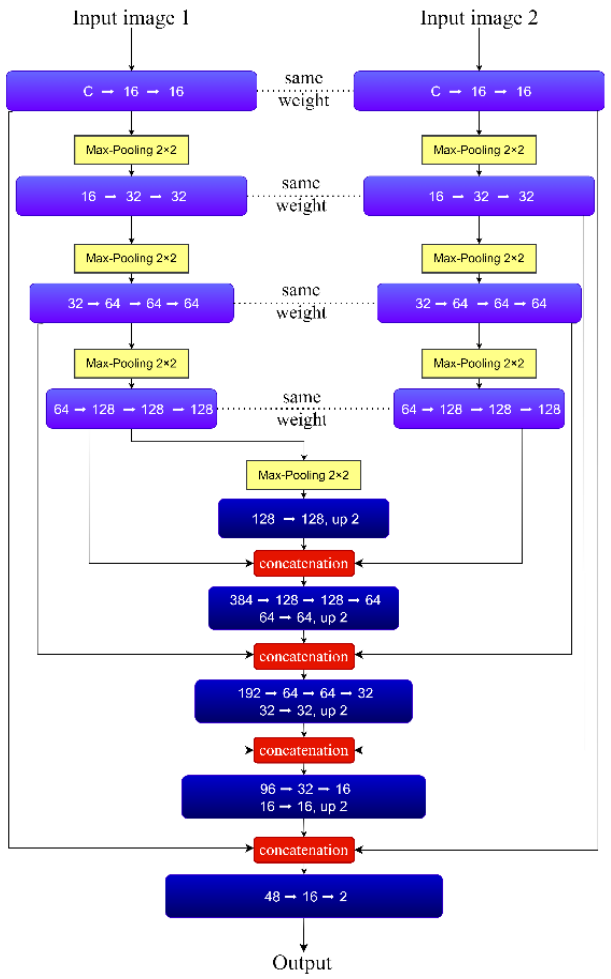

**Figure 2.** FC-Siam-conc network architecture. Dark blue and purple rectangles represent convolution and transpose convolution respectively.

In this study, we have developed an autoencoder that features the exact same encoder architecture as FC-Siam-conc. The idea is to first train the autoencoder on the whole dataset and then transfer the weights of the encoder into FC-Siam-conc. After the transfer, FC-Siam-conc can be trained in two modes: 1) freezing the encoder block and training the rest of the model and 2) training the whole model including the transferred weights. Mode 1 reduces the trainable parameters dramatically, leading to a shorter training time. On the other hand, mode 2 does not affect the number of trainable parameters but sets the weights in the decoder block to optimum values which could improve the convergence time.

## 4. IMPLEMENTATION

The main goal in this study was the binary CD however the downloaded benchmark dataset only featured semantic ground truth images. The no-change class in these images was used to create the binary ground truth images as depicted in Figure 3. The prepared data was then uploaded to the google drive account which was then accessed within the google colab notebook. All the aforementioned models in section 3 were designed using the TensorFlow API in Python. The Siamese network had the Softmax activation function in the final layer and the ReLU in all the other layers. Moreover, the model was compiled using the Adam optimizer with a learning rate of 0.001 and the binary cross-entropy loss function. The autoencoder model was designed in the same manner as the Siamese model but with a few differences. It featured the Sigmoid activation function in the final layer and the loss function was set to Mean Square Error (MSE). In the google colab environment, an Nvidia K-80 GPU was selected as the accelerator and the training was carried out for 100 epochs in all the scenarios.

Three training scenarios were considered in this study as can be seen in Table 1. Firstly, the Siamese model was trained on the prepared data. From the whole dataset, 20% was selected for testing and validation and 80% for training. Then, the autoencoder was trained on the whole dataset until reaching an MSE value of above 99%. For the sake of familiarizing the autoencoder with the whole dataset in the training phase, testing and validation sets were not separated from the dataset. At the next step, the encoder weights in the autoencoder were transferred to the Siamese network and set to freeze mode. Thus only the decoder was trained in this scenario. Finally, the whole Siamese model was trained after the weight transferring, setting no layers as frozen.

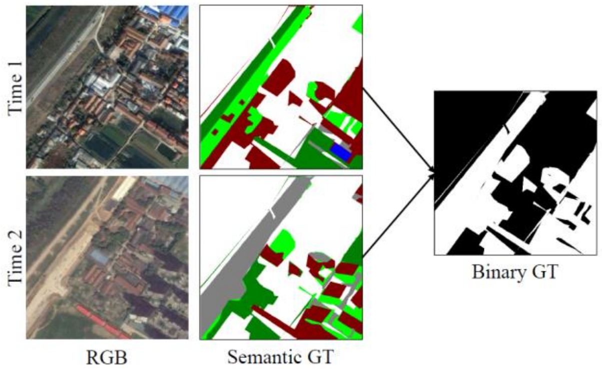

**Figure 3.** Converting semantic ground truth to binary ground truth.

## 5. RESULTS AND DISCUSSION

The testing dataset consisted of 593 image pairs that were not used in the training step. All of the three Siamese networks had the Softmax activation function in the final layer, outputting two channels. These channels correspond to the no-change and change classes, representing the probability. To achieve the final binary change map, the class with the maximum probability is assigned to each pixel. With the ground truth and prediction images at hand, a confusion matrix can be calculated which consists of the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. These values were used to calculate various metrics in order to compare the performance of the three scenarios. In this study, five evaluation metrics were defined including accuracy, sensitivity, precision, F1-score, and Intersection over Union (IoU). Among these five metrics, IoU is the most strict, offering a more realistic report of the models' performance. The quantitative results of the testing scenarios are represented in Table 2.

The benchmark dataset used in this study was very imbalanced and quite challenging to segment. This issue is manifested by the high accuracy and low IoU scores in all the scenarios. Sc.1 achieved accuracy and IoU scores of 88.87% and 48.87% and Sc.2 reached the accuracy and IoU scores of 87.42% and 45.31% respectively. Freezing the encoder block after the weight transfer in Sc.2 may have hindered the network's proper generalization, reducing the performance. However, the main advantage of Sc.2 was the considerable reduction in the training time while offering a comparable performance compared to

Sc.1 which is a valuable asset when doing an ablation study on parameters like different loss functions. Finally, Sc.3 achieved the best performance with accuracy and IoU scores of 91.95% and 61.30% respectively.

Figure 4 depicts the visual results of the three testing scenarios in a few samples which are in line with the quantitative results. Sc.1 had acceptable outputs but there were some noises at most parts. On the other hand, Sc.2 introduced the most artifacts compared to the other two scenarios. These artifacts occurred mainly in geometrically complex regions and sharp edges. Finally, Sc.3 had the cleanest outputs with the least artifacts and sharp edges.

| Scenarios | Models | Number of trainable parameters | Train time per epoch (seconds) |
|---|---|---|---|
| Scenario 1 (Sc.1) | Siamese | 1,437,186 | 364 |
| Scenario 2 (Sc.2) | Transferred Siamese (Encoder Frozen) | 957,554 | 253 |
| Scenario 3 (Sc.3) | Transferred Siamese (Encoder not Frozen) | 1,437,186 | 364 |

**Table 1.** Different training scenarios.

| Scenarios | Mean Accuracy (%) | Mean Sensitivity (%) | Mean Precision (%) | Mean F1-Score (%) | Mean IoU (%) |
|---|---|---|---|---|---|
| Scenario 1 (Sc.1) | 88.87 | 75.94 | 58.70 | 63.27 | 48.87 |
| Scenario 2 (Sc.2) | 87.42 | 68.91 | 58.19 | 59.59 | 45.31 |
| Scenario 3 (Sc.3) | **91.95** | **80.28** | **70.86** | **73.94** | **61.30** |

**Table 2.** Quantitative results of the testing scenarios (max values are bold).

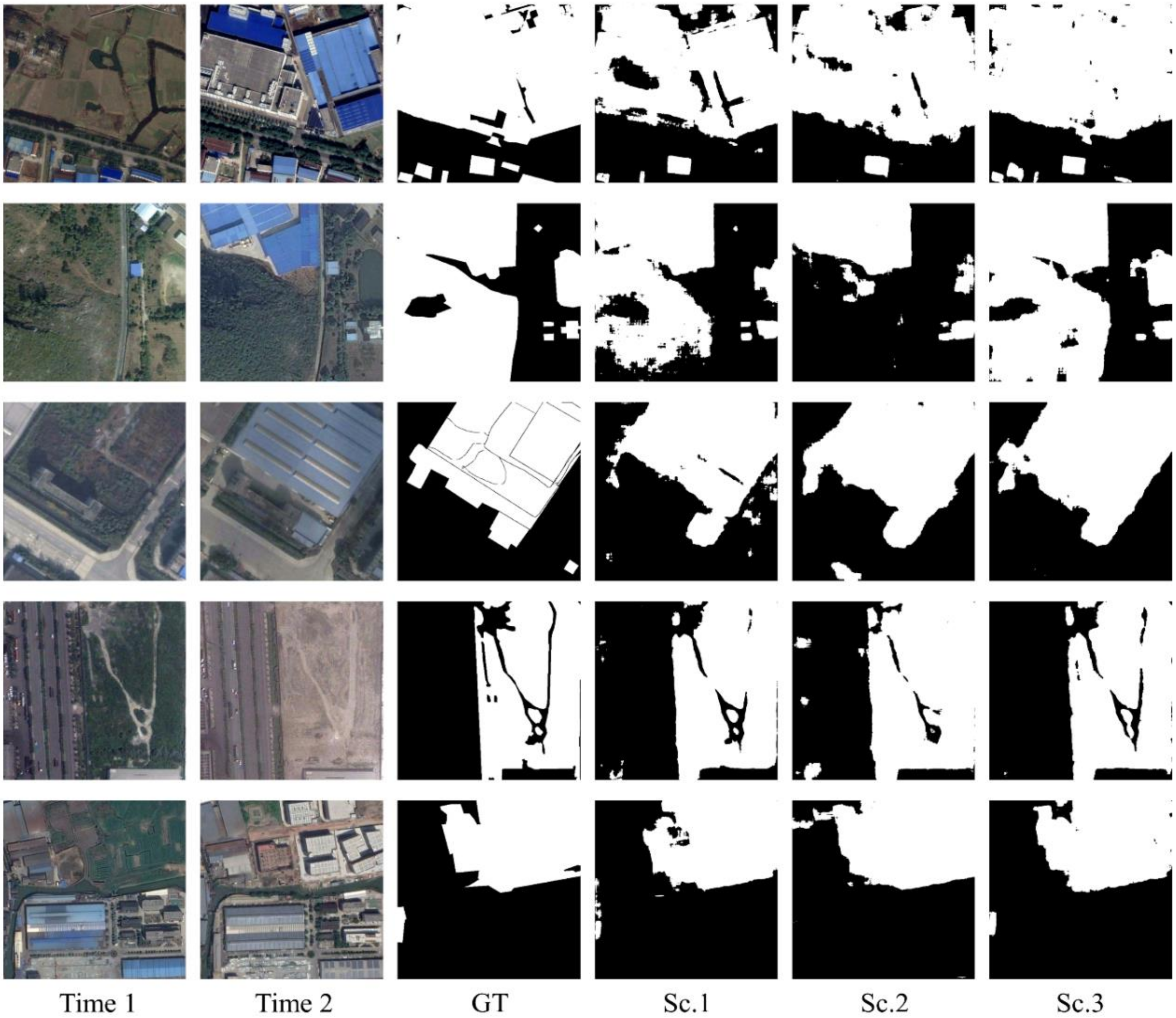

**Figure 4.** Visual results of the testing scenarios.

## 6. CONCLUSION

In this paper, we implemented a standard Siamese network on a benchmark CD dataset entitled SECOND. To study the effects of transferring pre-trained weights on enhancing accuracy and obtaining faster run times, we first trained an autoencoder that had the same encoder architecture as the Siamese model. Then the encoder weights were transferred from the autoencoder and the Siamese network was trained in two modes: 1) freezing the encoder and only training the decoder, and 2) training the whole network without freezing anything. The quantitative results proved that transferring the pre-trained encoder weights is helpful when the encoder is not frozen. The pre-trained weights help the network converge smoothly and faster while freezing the encoder interferes with the network's flexibility in generalizing. Albeit the relatively lower performance of the model when freezing the encoder, the enhancement in the training time is considerable and is a major benefit in ablation studies that deal with parameters like the loss function.

## REFERENCES

Aghdami-Nia, M., Shah-Hosseini, R., Rostami, A., Homayouni, S., 2022. Automatic coastline extraction through enhanced sea-land segmentation by modifying Standard U-Net. *Int. J. Appl. Earth Obs. Geoinf.* 109, 102785. https://doi.org/10.1016/j.jag.2022.102785

Ansari, M., Homayouni, S., Safari, A., Niazmardi, S., 2021. A New Convolutional Kernel Classifier for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 14, 11240–11256. https://doi.org/10.1109/JSTARS.2021.3123087

Caye Daudt, R., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. Proc. - *Int. Conf. Image Process. ICIP* 4063–4067. https://doi.org/10.1109/ICIP.2018.8451652

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens*. 12. https://doi.org/10.3390/rs12101662

Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y., 2019. Computer Vision and Image Understanding Multitask learning for large-scale semantic change detection 1–12.

Papadomanolaki, M., Vakalopoulou, M., Karantzalos, K., 2021. A Deep Multitask Learning Framework Coupling Semantic Segmentation and Fully Convolutional LSTM Networks for Urban Change Detection. *IEEE Trans. Geosci. Remote Sens*. https://doi.org/10.1109/TGRS.2021.3055584

Ranjbar, S., Zarei, A., Hasanlou, M., Akhoondzadeh, M., Amini, J., Amani, M., 2021. Machine learning inversion approach for soil parameters estimation over vegetated agricultural areas using a combination of water cloud model and calibrated integral equation model. https://doi.org/10.1117/1.JRS.15.018503

Rostami, A., Akhoondzadeh, M., Amani, M., 2022a. A Fuzzy-based Flood Warning System using 19-Year Remote Sensing Time Series Data in the Google Earth Engine Cloud Platform. *Adv. Sp. Res*. https://doi.org/10.1016/J.ASR.2022.06.008

Rostami, A., Shah-Hosseini, R., Asgari, S., Zarei, A., Aghdami-Nia, M., Homayouni, S., 2022b. Active Fire Detection from Landsat-8 Imagery Using Deep Multiple Kernel Learning. *Remote Sens*. 14. https://doi.org/10.3390/rs14040992

Yang, K., Xia, G.-S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L., 2020. Semantic Change Detection with Asymmetric Siamese Networks 1–14.

Zarei, A., Hasanlou, M., Mahdianpari, M., 2021. A comparison of machine learning models for soil salinity estimation using multi-spectral earth observation data. *ISPRS Ann. Photogramm. Remote Sens*. Spat. Inf. Sci. 5, 257–263. https://doi.org/10.5194/isprs-annals-V-3-2021-257-2021

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens*. 166, 183–200. https://doi.org/10.1016/j.isprsjprs.2020.06.003