# Graph Neural Network Based Multi-feature Fusion for Building Change Detection

W. Yuan[1,2,*], X. Yuan[2], Z. Fan[1], Z. Guo[1], X. Shi[1], J. Gong[2], R. Shibasaki [1]

[1] Center for Spatial Information Science, University of Tokyo, 5-1-5, Kashiwa, Chiba, Japan – (miloyw,fanzipei,guozhilingcc,shixiaodan,shiba)@csis.u-tokyo.ac.jp
[2] Wuhan University, School of Remote Sensing and Information Engineering, 129 Luoyu Road Wuhan, Hubei, China - (yuanxx, jygong)@whu.edu.cn

**Commission III, WG III/7**

**KEYWORDS:** Graph Neural Network, Feature extraction, Multi-feature fusion, Dense Image Matching, Building Change Detection, Node aggregation;

**ABSTRACT:**

Building Change Detection (BCD) via multi-temporal remote sensing images is essential for various applications such as urban monitoring, urban planning, and disaster assessment. However, most building change detection approaches only extract features from different kinds of remote sensing images for change index determination, which can not determine the insignificant changes of small buildings. Given co-registered multi-temporal remote sensing images, the illumination variations and misregistration errors always lead to inaccurate change detection results. This study investigates the applicability of multi-feature fusion from both directly extract 2D features from remote sensing images and 3D features extracted by the dense image matching (DIM) generated 3D point cloud for accurate building change index generation. This paper introduces a graph neural network (GNN) based end-to-end learning framework for building change detection. The proposed framework includes feature extraction, feature fusion, and change index prediction. It starts with a pre-trained VGG-16 network as a backend and uses U-net architecture with five layers for feature map extraction. The extracted 2D features and 3D features are utilized as input into GNN based feature fusion parts. In the GNN parts, we introduce a flexible context aggregation mechanism based on attention to address the illumination variations and misregistration errors, enabling the framework to reason about the image-based texture information and depth information introduced by DIM generated 3D point cloud jointly. After that, the GNN generated affinity matrix is utilized for change index determination through a Hungarian algorithm. The experiment conducted on a dataset that covered Setagaya-Ku, Tokyo area, shows that the proposed method generated change map achieved the precision of 0.762 and the F1-score of 0.68 at pixel-level. Compared to traditional image-based change detection methods, our approach learns prior over geometrical structure information from the real 3D world, which robust to the misregistration errors. Compared to CNN based methods, the proposed method learns to fuse 2D and 3D features together to represent more comprehensive information for building change index determination. The experimental comparison results demonstrated that the proposed approach outperforms the traditional methods and CNN based methods.

## 1. INTRODUCTION

Building Change Detection (BCD) via multi-temporal remote sensing images is essential for various applications such as urban monitoring, urban planning, and disaster assessment. Automated BCD technology has been a hot topic in the remote sensing field in recent years and has accelerated the development of actual industrial production and applications (Xiao, 2017).

In the past few decades, a lot of BCD methods have been proposed. Based on the difference of input data, BCD methods can be categorized as 2D BCD and 3D BCD. Traditional 2D BCD methods utilize the image information for building segmentation and conduct the post-classification methods for change detection. As the development of the sensor, the remotely sensed image has reached a finer level. Pixels in very high resolution images contain more detailed information, making the BCD results more sensitive to the pixel-based comparison (Liu, 2020).

Meanwhile, perspective distortion is a big challenge for 2D BCD. The orthophoto generated in different periods usually contained geometric distortions. The per-pixel-based image registration is generally conducted as a preprocessing for 2D BCD methods, which is difficult for images with a large difference in viewing angle (Annibale, 2009). The 2D BCD methods can only extract the planimetric changes such as appearing/disappearing, shrinking/expanding. It is insufficient for applications requiring more specific volumetric information, such as building construction progress monitoring (Qin, 2016). With additional information from the 3D points cloud, more detailed change types and accurate change detection results can be automatically generated by 3D BCD methods. Since the 3D information is free from illumination changes and perspective distortions, the co-registration is more likely to achieve. But the biggest challenge of 3D BCD is the 3D data acquisition. LiDAR data has high accuracy, but the cost is expensive. Photogrammetric dense image matching generated 3D point clouds is much denser and cheaper than LiDAR, which is more suitable for this kind of work. Based on this, in this paper, we present a 2D and 3D combined BCD method based on multi-feature fusion via graph neural network.

* Corresponding author

## 2. RELATED WORKS

Building change detection can be categorized as pixel-based change detection and object-based change detection. During the last decades, a lot of well-known machine learning technology has been utilized for pixel-based change detection, such as Markov random field (MRF)-based change detection (Moser, 2011), support vector machine (SVM)-based change detection (Volpi, 2011), k-nearest neighbor (KNN)-based change detection, and random forest (RF)-based change detection. Due to the fast development of deep learning, well-performed deep neural networks are introduced for improving the building segmentation accuracy and change detection accuracy (Chen, 2012). Such as the multi-layer perceptron neural network (Xu, 2017) and convolutional neural network-based U-net (Lian, 2020). However, a single neural network usually considers the 2D image features for instance classification and segmentation. Which is not adequate for multi-temporal images with different illumination changes. To address this issue, ensamble learning has been introduced to enhance change detection and classification accuracy. Although deep learning-based methods have achieved preferable results, the training sample preparation is still economical and time-consuming.

Despite the advantages of the deep learning based methods, pixel-based methods usually contained many noises around the building edges and boundaries. On the other hand, the image resolution becomes higher and higher. The influence of land cover changes between the multi-temporal data is sensitive for the pixel-based methods to generate robust results. And the manually selected labeled training sample can not be transferred well between the multi-temporal data. To overcome the above issues, the object-oriented approaches have become more popular than pixel-based methods. In 2018, Zhang et al proposed an SVM based uncertainty analysis for object-based change detection, which outperforms pixel-based methods (Zhang, 2018 ).

Since the 2D change detection can only provide limited information, more and more researchers are concentrated on 2D-3D combined BCD to meet the requirements of more specific applications. The most challenging thing is the 3D data acquisition and multi-data co-registration. Since the LiDAR data is usually collected with no texture information, and the data collection is economical consuming, the photogrammetric generated 3D points cloud is preferable for 3D BCD methods. The generated ortho-photo and dense 3D point cloud can be registered very well through photogrammetric processing, which makes the co-registration between multi-temporal data much easier.

## 3. METHOD

The aim of this work is to combine the 2D and 3D information to generated more comprehensive change detection results for urban monitoring and planning. Given bi-temporal raw image data, we automatically generate the whole area's dense 3D point clouds and rectify images and detect the changed buildings with semantic change types.

In Figure 1, our entire approach consists of: (a) an optical flow field based dense image matching step for dense 3D point clouds generation and ortho-image rectification; (b) a 2D and 3D feature extractor, extracted 2D and their corresponded geo-position then fuse 3D features; (c) a graph neural network that takes the fused features as input and constructs a graph with the node feature from the fused feature extract from the bi-temporal data. Then the graph neural network iteratively aggregates the node feature from the neighborhood and contributes the affinity matrix for change type determination.

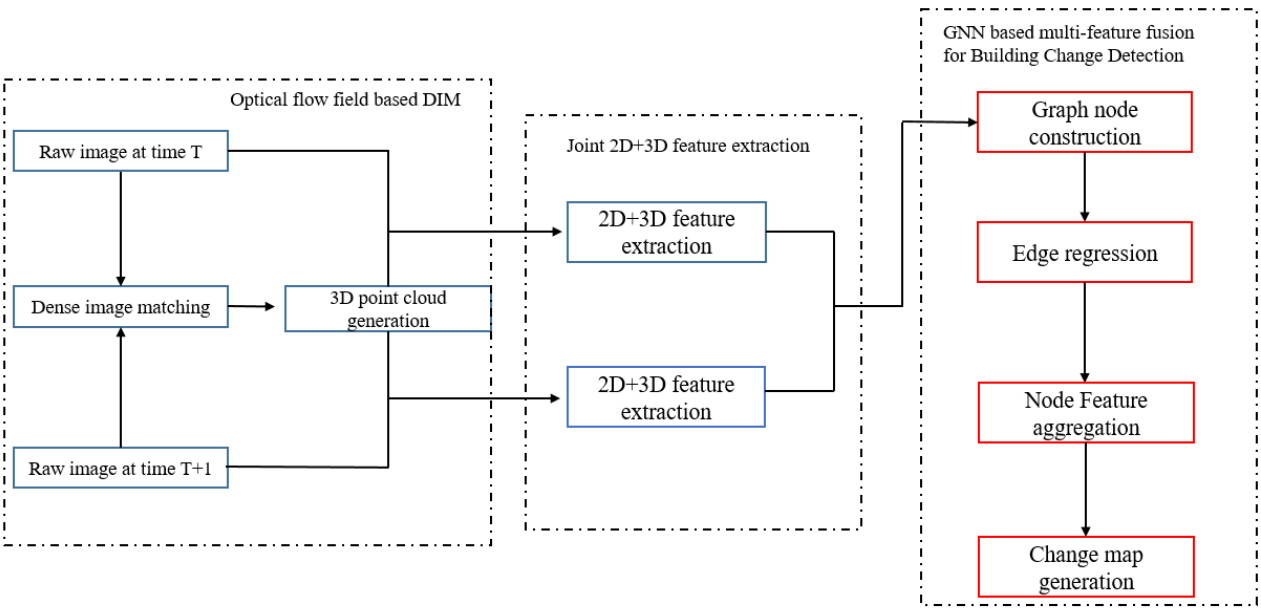



Figure 1. The workflow of the proposed GNN based multi-future fusion for building change detection.

### 3.1 Optical flow field based dense image matching and image rectification

The 3D information is more informative to illumination changes and perspective distortions. To make the pixel between two time-period data precisely aligned, georeferencing between the two images is necessary. In order to balance the processing time and image rectification quality, we use our former optical flow field based DIM methods for pixel-wise dense correspondence matching and 3D point cloud generation. The detailed matching strategy and procedures can be found in the paper (Yuan et al., 2019). After the pixel-wised matching results is generated, the photogrammetric forward intersection is conducted for the 3D points cloud generation. Then the orthophoto is generated by a pixel-wise image rectification.

### 3.2 2D and 3D feature extractor

The 2D and 3D feature extractors have different modalities and can be utilized to learn discriminative features. The commonly utilized approaches treat the 3D information as the height dimension or depth dimension in traditional RGB images while then embedding them into a 4-dimensional image to input deep neural networks. This kind of approach only utilizes the height information of the 3D data but can not be aware of the internal structure connection between the 3D data itself. In our 2D and 3D feature extractor, we input 3D point cloud and rectified 2D RGB images into separate feature extractors to leverage 2D and 3D features strength.

#### 3.2.1 2D Feature extraction

We first employ a U-net (Ronneberger, 2015) like architecture for 2D feature extraction. Unlike common CNN-based feature extractors, we use the whole raw image as inputs instead of utilizing the sliced image patches. In our case, U-net shows a strong performance in extracting appearance features on the raw image resolution, which makes it fulfill our requirements. The utilized U-net architecture contained with fifve layers, the second and third layer are downsampled layer with PReLU activation function, and the fourth to fifth layer are the upsample layer with PReLU function and batch normalization.

#### 3.2.2 3D Feature extraction

Different from the 2D feature extractor, the 3D feature extractor considering both the appearance feature and also the structure information. Given a detected building in time t, and time t+1 we want to extract the corresponding 3D feature containing both appearance and structure information. By analyzing that information, it is more informative for us to determine the corresponded building change types, such as newly built, demolished, under constructed, etc.

Since the whole 3D point cloud dataset is quite large. For appearance feature extraction, we first extract the point cloud enclosed by a certain 3D detection box and then utilizing PointNet (Qi, 2017) for appearance feature extraction. The structure information is extracted by utilizing a three-layer MLP, then the structure and appearance feature vector is forced to have the same dimension and concatenated to build the final 3D feature.

#### 3.2.3 Feature fusion

After the 2D and 3D features are extracted, we utilize an add operation instead of concatenating two features. The add operation will force the two kinds of features to have the same dimension had more feasible for operation.

### 3.3 Graph neural network based change detection

Compared with traditional CNNs, the GNN can esily find the internal connections between one node with its neighbors, which may reduce the false-positive noise from the pixel-based detection results. To leverage the strength of GNN we feed the fused features extracted from the different time period to automatically determin the building change maps.

#### 3.3.1 Graph construction

The fused features are treated as the node to build the graphs. In the graph build step, we need to define the edges between all the node to construct the whole graph. Since we aim to find the difference between the two time period, just connected every node is computationally expensive. Instead of using the simply edge construction, we selected to build the edges only between nodes from different time periods. Also, the corresponded features should have near geo-loactions, so we add a threshold to build the edges.

#### 3.3.2 Edge regression and Node feature Aggregation

To find the changes between the two-time series data, we need to compute an affinity matrix based on the pairwise similarity of the features extracted by the 2D and 3D feature extractors. For every pixel on the rectified image, it will have a label to represent the changed types. In the edge regression part we utiliz a two-layer MLP to calculate the pairwise similarity score. The detailed calculation is shown as function (1):

$$A_{ij} = S(\sigma_1(\mathrm{ReLU}(\sigma_2(n_i^t - n_j^{t+1})))) \quad (1)$$

Where $\sigma_1$ and $\sigma_2$ are two linear layers, $n_i^t$ and $n_j^{t+1}$ are two extracted features from different time period, $S$ is the sigmoid function. The node feature aggregation is utilizing the same as GraphConv (Morris, 2019).

#### 3.3.3 Loss Function

The proposed networks introduce a combined losses for traning. The first one is affinity loss and the second is a contrastive loss. The affinity loss is the calculate tby the average cross entropy between the ground truth matrix and the affinity matrix. It can be represented as :

$$L_{aff} = \frac{-1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} G_{ij}\log(A_{ij}) \quad (2)$$

Where $A$ is the predicted affinity matrix, $G$ is the ground truth matrix, $M$ is the number of extracted features in the time t, and $N$ is the number of extracted features in the time t+1.

The contrastive loss function can be represented as:

$$L_{contr} = \frac{1}{2MN}\sum_{i=1}^{M}\sum_{j=1}^{N} sim(n_i^t, n_j^{t+1}) \left\| n_i^t - n_j^{t+1} \right\|_2^2 + (1 - sim(n_i^t, n_j^{t+1}))max(\tau - \left\| n_i^t - n_j^{t+1} \right\|_2)^2 \quad (3)$$

Where *sim* is the function to calculate weather $n_i^t$ is equal to $n_j^{t+1}$ if not sim() equals to 0 otherwise it equal to 1. $\tau$ is the margin of $L_{contr}$. The combined loss is:

$$L = L_{contr} + L_{aff} \quad (4)$$

## 4. EXPERIMENT AND ANALYSIS

### 4.1 Datasets

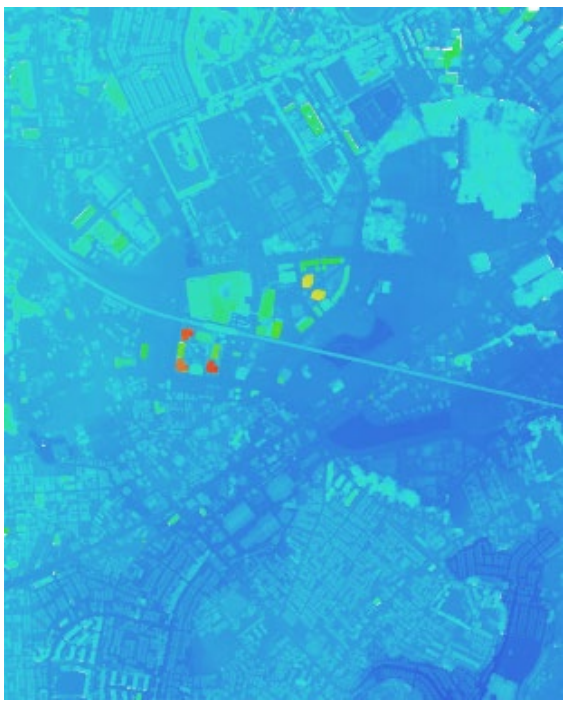
The experiment is conducted on an Aerial image dataset from Tokyo urban area. It covers the whole Setagaya district the detailed parameter for the test image is listed as Table 1. Since the raw data from 2 years are collected from the same platform and same camera, so here we only show one year's data parameter.

Table 1. Technical parameter of the test images

| Item | Tokyo |
|---|---|
| Aerial craft | Aerial |
| Camera | Intergraph/ZI DMC II |
| Principal distance (mm) | 70.50000 |
| Format (pixels) | 11310×17310 |
| Pixel size (μm) | 6.0 |
| Ground sample distance (GSD) (cm) | 16 |
| Relative flying height (m) | 1180 |
| Longitudinal overlap (%) | 80 |
| Lateral overlap (%) | 80 |
| Number of mapping strips | 4 |
| Number of control strips | 0 |
| Number of images | 48 |
| Number of ground control points | 21 |
| Number of pass points | 23317 |
| Block area (km2) | 4.0 × 3.0 |
| Maximum topographic relief (m) | 135 |
| Average terrestrial height (m) | 35 |

The automatic generated georectified image and the corresponded 3D points cloud is shown as Figure 2. The orthophoto achieved the same resolution as the raw aerial image and the corresponded 3D points cloud accuracy achieve 1.5GSD.

(a)



(b)

Figure 2. The automatic generated (a) orthophoto and (b) corresponded 3D point cloud.

### 4.2 Implementation details

The experiment is conducted on a high-performance computational server with 2 RTX 2080ti GPU. Evaluation matrics quantitatively evaluate the change detection results with

the index of precision (P), recall I, and F1 score. Let $N_{ij}$ be the number of pixels of class *i* predicted as class *j,* and there have 3 kinds of change types we detect, newly built, demolished, and unchanged. Then the precision, recall and F1 score can be calculated as below:

$$P = N_{ii} / \sum_j N_{ji}$$
$$R = N_{ii} / \sum_j N_{ij} \quad (5)$$
$$F1 = 2PR / (P + R)$$

### 4.3 Experimental results

In figure 3 we show the selected area of our proposed change detection results with the ground truth. For comparison, we choose the CNN-based method proposed by Lian et al. (2020). They were utilizing DSM and orthophoto as input for building change detection and achieved superior results than tradional image-based post-classification method and pixel-based deep learning method. Figure 3 (c) and (g) show that the GNN based methods can accurately detect the changed building with completed building structures, but some small changes are missed. Compared with CNN-based methods, the detection results have significantly improved, but it occurred with some small noises. The quantitative comparison is shown in Table 2.

Table 2. The quantitative evaluation matric of CNN based method and the proposed method.

| Methods | Precision(%) | Recall(%) | F1-Score(%) |
|---|---|---|---|
| post-classification | 57.1 | 37.1 | 44.9 |
| CNN based | 68.1 | 52.0 | 58.9 |
| Proposed | 76.2 | 62.1 | 68.3 |

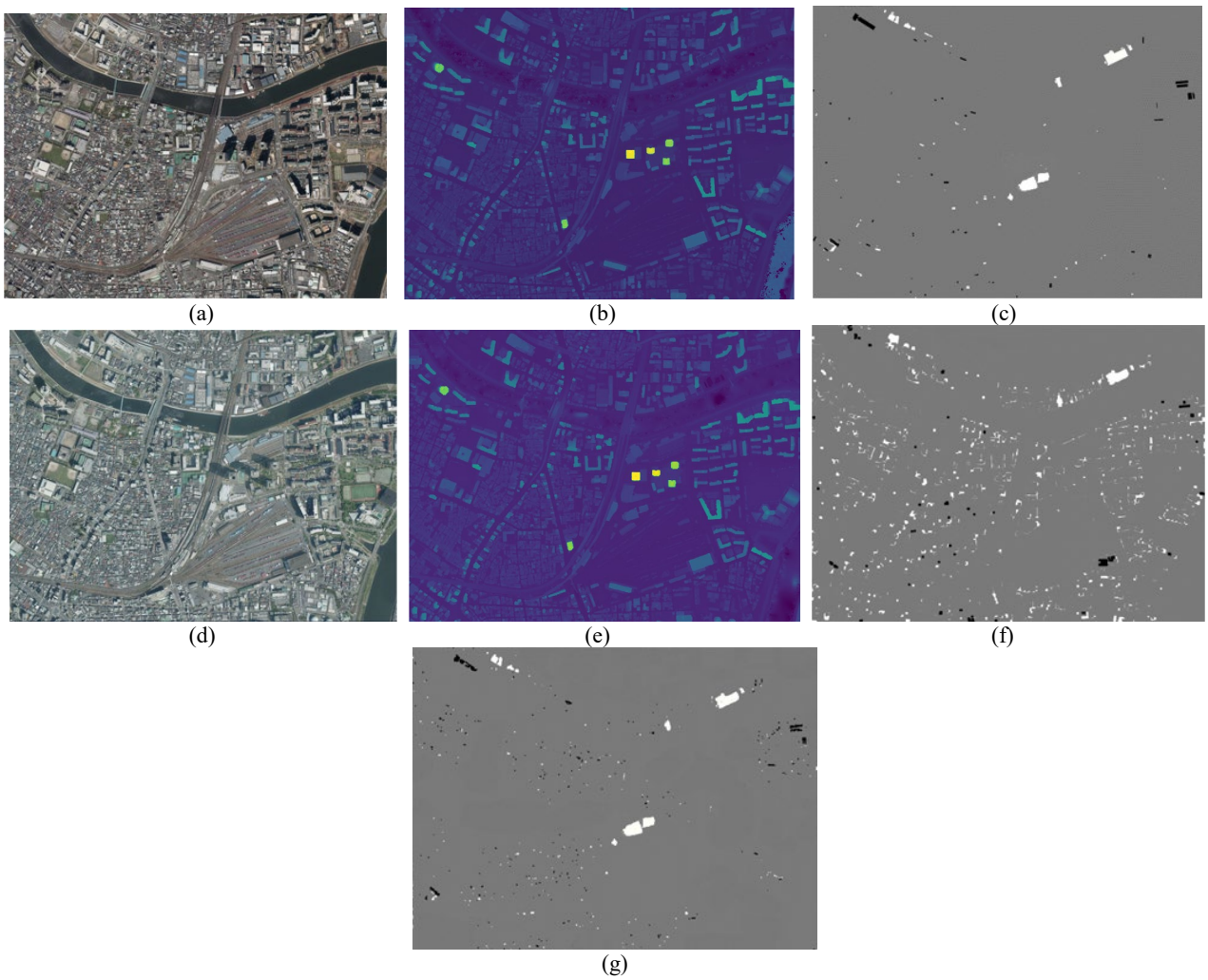


Figure 3. The experimental results of the proposed method. (a) and (d) are two time periods orthophoto, (b) and (e) are the corresponded 3D points cloud, (c) is the ground truth, black means new constructed, grey means unchanged, black means demolished, (f) is the CNN based change detection results, (g) is this paper proposed GNN based change detection results.

## 5. CONCLUSION

In this paper, we propose a GNN based multi-level feature fusion network for building change detection. The experimental results demonstrated that the proposed method is superior to the traditional 2D image-based post-classification methods and outperforms the CNN-based method, which combines the DSM and orthophoto for building change-type detection. The proposed methods extract 2D and 3D features separately and leverage the strength of those two kinds of features to generate the building change map comprehensively. However, the proposed methods still suffered from two major limitations. First, it utilizing the supervised learning manner, which requires large numbers of manually labeled samples; Second, this method lacks considering the transferability of the trained model, which will be considered in the future.

## REFERENCES

Xiao, P., Yuan, M., Zhang, X., Feng, X., Guo, Y. Cosegmentation for object-based building change detection from high-resolution remotely sensed images. IEEE Trans. Geosci. Remote Sens. 2017, 55, 1587–1603.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., and Yang, X. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. IEEE Geoscience and Remote Sensing Letters. 2020

Anniballe, R., Noto, F., Scalia, T., Bignami, C., Stramondo, S., Chini, M., Pierdicca, N. Earthquake damage mapping: An overall assessment of ground surveys and VHR image change detection after L'Aquila 2009 earthquake. Remote Sens. Environ. 2018, 210, 166 – 178.

Qin, R., Tian, J., & Reinartz, P.. 3D change detection–approaches and applications. ISPRS Journal of Photogrammetry and Remote Sensing. 2016, 122, 41-56.

Moser, G., Angiati, E., Serpico, S.B. Multiscale unsupervised change detection by Markov random fields and wavelet transforms. IEEE Geosci. Remote Sens. Lett. 2011, 8, 725–729.

Volpi, M.; Kanevski, M. Supervised change detection in VHR images using support vector machines and contextual information. Int. J. Appl. Earth Obs. Geoinf. 2010, 20, 77–85.

Chen, X., Chen, J., Shi, Y., Yamaguchi, Y. An automated approach for updating land cover maps based on integrated change detection and classification methods. ISPRS J. Photogramm. Remote Sens. 2012, 71, 86–95.

Xu, D., Ouyang, W., Ricci, E., Wang, X., and Sebe, N. 2017. Learning cross-modal deep representations for robust pedestrian detection. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Lian, X., Yuan, W., Guo, Z., Cai, Z., Song, X., and Shibasaki, R. End-to-end Building Change Detection Model in Aerial Imagery and Digital Surface Model Based On Neural Networks. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020,43, 1239-1246.

Zhang, Y., Peng, D., Huang, X. Object-based change detection for vhr images based on multiscale uncertainty analysis. IEEE Geosci. Remote Sens. Lett. 2018, 15, 13–17.

Yuan, W., Yuan, X., Xu, S., Gong, J., and Shibasaki, R. Dense image-matching via optical flow field estimation and fast-guided filter refinement. Remote Sensing, 2019. 11(20), 2410.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J.. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. 652-660.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence. 2019. 33-01, 4602-4609.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. 2008. The graph neural network model. IEEE transactions on neural networks, 20(1), 61-80.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., ... & Sun, M. 2018. Graph neural networks: A review of methods and applications. arXiv preprint arXiv:1812.08434.