Automatic Building Change Detection on Aerial Images using Convolutional Neural Networks and Handcrafted Features

Diego Alonso Javier Quispe[1], Jose Sulla-Torres[2]
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—In this article, we present a new framework to solve the task of building change detection, making use of a convolutional neural network (CNN) for the building detection step, and a set of handcrafted features extraction for the change detection. The buildings are extracted using the method called Mask R-CNN which is a neural network used for object- based instance segmentation and has been tested in different case studies to segment different types of objects obtaining good results. The buildings are detected in bitemporal images, where three different comparison metrics MSE, PSNR and SSIM are used to differentiate if there are changes in buildings, we used this metrics in the Hue, Saturation and Brightness representation of the image. Finally the characteristics are classified by two algorithms, Support Vector Machine and Random Forest, so that both results can be compared. The experiments were performed in a large dataset called WHU building dataset, which contains very high-resolution (VHR) aerial images. The results obtained are comparable to those of the state of the art.**

***Keywords*—Bi-temporal images; convolutional neural network (CNN); building detection; building change detection; Mask R-CNN**

I. Introduction

Building detection and change detection is a field that has been studied for a long time and attempts to solve different problems such as urban planning, cadastral updating, damage detection by natural disasters, among many others. Building change detection consists in differentiating the changes that occur over time in a building, considering that a building could be built, could be destroyed or could be modified.

Different researchers use different types of data to carry out their experiments. In their most basic form, aerial images or RGB satellite images are used, but these images have different drawbacks, starting with low resolution, perspective, lighting changes, shadows, and various variations that a building can have depends on the country, the city and the area where the images were captured. Other authors use different sensors to obtain more information such as multispectral sensors, synthetic aperture radar (SAR), light detection and ranging (LiDAR), digital surface models (DSM) and so on [1]. Using DSM allows us to obtain a 3D model of buildings, the advantage is that having altitude information allows us to better analyze changes in buildings, although the drawback is that it is difficult to obtain such information.

The related works will be divided into two sections, the works related to the detection of buildings, and those related to the detection of changes in buildings. The Morphological building / shadow index (MBI) [2] is a long-term building detection method used in different works related to building change detection [3] [4] [5] [6] [7], which consists of representing the spatial edges of the buildings in such a way that they can be distinguished from other objects. These properties are represented by brightness, size, contrast, directionality and shape. Other ways to detect the edge of a construction is through the use of shadows, Ali Ozgun [8] created a model to determine the relationship between the buildings and their shadows using a probabilistic approach. Saman Ghaffarian and Salar Ghaffarian [9] uses a different approach, instead of using the RGB image, they use another color space, the LUV, and through the FastICA algorithm they divide the image into three different regions: vegetation and shadows, firm ground and tracks, and buildings, in this way they try to avoid confusing some other object with buildings. Fadi Dornaika et al. [10] use a segmentation technique called statistical region mergin (SRM), which segments an image into small homogeneous regions based on their similar properties, considering the spectral information of shape and scale. After applying SRM they extract information using the local binary pattern (LBP) algorithm for each region segmented in the previous step. Finally they use four classifiers which are listed below: 1-NN, 3-NN, J48 and SVM, to determine if they are buildings or not buildings. Comparing these classifiers results in SVM being better than the other classifiers.

The works described so far only use machine learning and handcrafted features extraction techniques to segment buildings, but in recent years deep neural networks have obtained very good results in the field of object segmentation. Ji et al. [11], propose a deep neural network based on a neural network known as U-net, where instead of having a single input it requires two inputs so they call it Siamese U-net (SiU-Net), their proposal is compared to other networks, which are also used for segmentation of objects at the instance level, these networks are Mask R-CNN and U-net. The authors determine that SiU-net is slightly better than U-net and Mask R-CNN.

The detection of changes in buildings can be done at two different levels, detection of changes- at the pixel level and detection of changes at the object level, most of the related works were prepared based on changes at the pixel level. The work of Wen et al. [5], consists in classifying each pixel in 4 different categories: building, vegetation, water and soil. To determine the changes, each bi-temporal image is subdivided into quadrants, for each quadrant a histogram is calculated so that it can be compared with the histogram of its corresponding bitemporal image, finally the histograms are compared to determine if the quadrant is considered with changes or without changes. In addition to the four categories mentioned, a category also considered is the shadow which is taken into account by various authors [12] [7] [13], if any object generates shadow in a location where long ago there was no shadow, it is an indication of change, the disadvantage is that the shade is conditioned at the time of day and at the atmospheric condition. Considering only RGB information can be deficient, so other authors choose to use 3D information [14] [15], one of the advantages that this type of data contains is height, so that if we compare the height of buildings, it can be known if there have been changes or there have been no changes.

Deep Learning is also currently used in detecting changes in buildings. Considering that there are bi-temporal images, Debu et al.[16], propose three Fully Convolutional Neural Netwroks (FCNs) based on the concept of Siamese networks, they consider two FCN with double input. The third FCN has a single image as input, where the bitemporal images are concatenated. The results obtained are relatively low, this is because the data set contains low resolution images. Ji el al. [17] use a simple CNN for the detection of changes, their proposal is to use as input only the binary mask of the bitemporal images, and not the complete images, so that with this information the binary mask of changes is generated as output. Their experiments were carried out on a data set with a large number of buildings and an acceptable amount of changes. Furthermore, these images have a very high resolution.

In this article we propose a new framework for detecting changes in buildings using very high resolution aerial images (VHR). For the detection of buildings we use a convolutional neural network and for the detection of changes we use comparison metrics in different color spaces of the images, so that different characteristics of the buildings can be compared. Finally we use two SVM and RF classifiers to determine the buildings in which changes have been detected. The data set used is called the WHU building data set [11], which contains more than 220,000 independent buildings which vary in color and size.

This paper is organized as follows. In Section II the methodology applied for the automatic detection of changes in buildings is detailed. Section III describes the database used and presents the results obtained in the steps of building detection and building change detection with the proposed model. Finally, in Section V details the conclusions of the present paper.

II. Methodology

The general pipeline is shown in Fig. 1. First we take the bitemporal images as input, for each image a binary mask is created where it is determined if an area is a building or background, by means of a building extraction network. Next, each building is compared with its corresponding temporal image using comparison metrics as MSE, PSNR and SSIM, we use the HSV representation of each image. Finally we use two classifiers Support Vector Machine (SVM) and Random Forest (RF) to determine if there are changes in buildings.

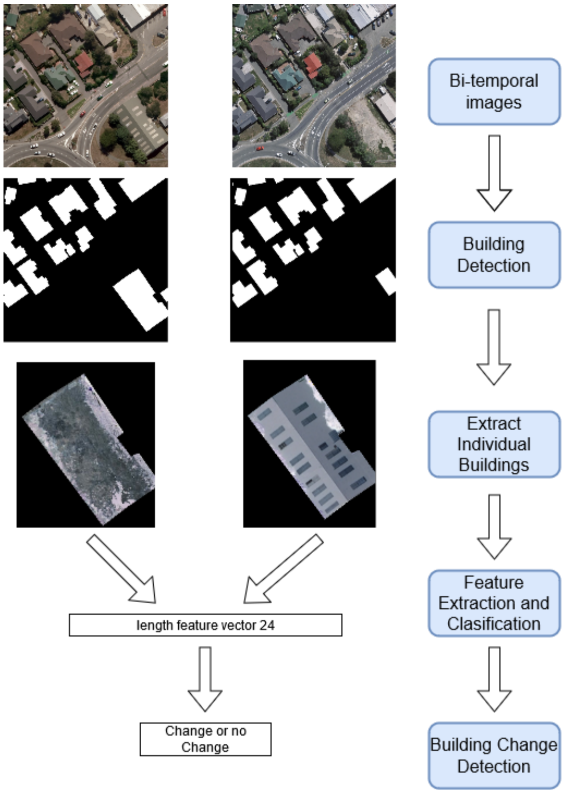
Bi-temporal images
Building Detection
Extract Individual Buildings
Feature Extraction and Clasification
Building Change Detection
length feature vector 24
Change or no Change

Fig. 1. Pipeline for Building Change Detection using a Neural Network for Building Detection and Handcrafted Features for Change Detection.

A. Building Detection

We use a convolutional neural network called Mask R-CNN [18], which is applied for object detection and we used it for detection of buildings. Based on the results obtained in [19]. This deep neural network was proposed in 2017 and is one of the most powerfull networks for object instance segmentation.

The advantages provided by this network is that it provides three different outputs for object detection, as the first output presents the classification of objects, the second output is the regression box and finally the prediction of the mask. While the classification section does not have to distinguish a large number of different objects, it must be able to differentiate a building from any other object. The main problem lies in the data because the bulding could have different sizes and shapes, also different objects can be confused as buildings, as is the case with large trucks.

In Fig. 2 shows the general architecture of Mask R-CNN, the network parameters suggested by the author were not modified.

B. Building Change Detection

1) Extract individual buildings: In an image of 512 x 512 we have different number of buildings. Each individual building is extracted, considering two cases, the first where the building has not changed and the second in the case that has changed. We obtained a total of 15005 sub - images with individual buildings.

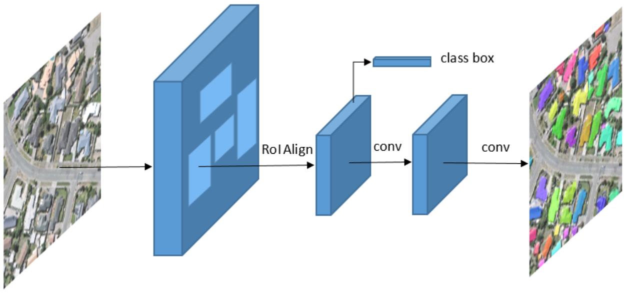

Fig. 2. Mask R-CNN Framework Arquitecture for Instance Segmentation.

2) Image Representation: To compare two images we use a different representation of the image, if we compare a RGB image with another RGB image we will get the same value for the red, green and blue bands, for that reason instead of using a RGB image we use the HSV representation where we have three different bands (Hue, Saturation, Value). In addition to using the HSV representation of the image, the grayscale representation is used.

3) Histogram of Oriented Gradient: The Histogram of Oriented Gradient (HOG) [20], is a feature descriptor that focuses on determining the shape of an object. This is determined by calculating the difference between the gradients of an image, to subsequently obtain a value and a direction for each pixel of an image until obtaining a histogram that represents its characteristics.

4) Comparison Metrics: We are going to compare the bitemporal images so that we obtain a unique value for each pair of images. We use three different comparison metrics: mse, psnr and ssim, its definition is detailed below:

a) MSE (Mean Square Error): This is the average square difference between two values. This is always no negative and if the value is very close to zero it means that there are no changes. We compare pixel by pixel the two images. The MSE formula is as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{1}$$

b) PSNR (Peak Signal to Noise Ratio): It is represented as the relationship between the maximum value that a pixel can have and the noise that affects the representation of the entire image. The PSNR formula is as follows:

$$PSNR = 10\log_{10}\frac{R^2}{MSE} \tag{2}$$

The variable R corresponds to the maximum value that an image can take, depending on how it is represented, it can take the value of 1 or 255.

c) SSIM (Structure Similarity Index Method): It is used to measure the quality between two images. Considering that MSE tries to find the differences between the pixels of an image, SSIM does the opposite, comparing the pixels to determine their similarity, based on three different terms: luminance, contrast and structure. The final value is represented by multiplication of these terms. The SSIM formula is as follows:

$$SSIM = \frac{(2U_xU_y + C_1)(2\alpha_{xy} + C_2)}{(U_x^2 + U_y^2 + C_1)(\alpha_x^2 + \alpha_y^2 + C_2)} \tag{3}$$

One drawback to consider is the difference that exists between the 2012 images versus the 2016 images, due to the difference in the atmospheric conditions in which they were captured. Fig. 3 shows that there is a difference in the contrast between these images, therefore to solve this problem, a histogram equalization is applied to each representation obtained. Equalization is used as a complement and in addition, a feature vector is extracted using HOG for the gray images. Therefore we finally have 10 different representations which are the following: hue, value, saturation, gray, equalized hue, equalized value, equalized saturation, gray equalized, HOG of a gray image and HOG of a gray equalized image. Our final vector of characteristics is conformed by the concatenation of the 3 comparison metrics described (MSE, PSNR, SSIM) for each image representations obtained, by having 10 representations, our final final vector for each pair of images will have a length of 30.

5) Classifiers:

a) Support Vector Machine: The support vector machine (SVM) is a binary classifier which separates the classes into two different spaces by means of a hyperplane which is known as the support vector [21]. The characteristics can be separated in three different ways: by means of a linear nucleus for which the Euclidean distance is used to define the hyperplane, by means of a polynomial nucleus and by means of a Gaussian nucleus which is associated with the variance.

b) Random Forest: The random forest classifier consists of a large number of independent binary trees, the end result is a majority vortex [22]. Each binary tree generates a different prediction, it is expected that a set of them tends to a wrong result but a larger set of trees tends to the correct result, causing the global result to be correct. The tree mainly depends on two values, N that represents the number of trees, and p that indicates the depth of the tree .

III. Experimental Results

A. Data Set

We use the WHU building dataset proposed by [11], this dataset contains aerial and satellite images, but we only use aerial images for their high quality. This dataset consists in aereal images captured between 2011 to 2016, the area captured cover the city of Christchurch, New Zealand; have a total of 120000 buildings and covers an area between 450m^2 and 550m^2. The main drawbacks of the database are the noise caused by plants that obstruct sections of the roofs and large cars that are confused with buildings.

We divide this dataset in three sections, the first and second sections were used in the building detection step as training and validation data respectively. The third section was used as a test to building detection and all this section was used for the building change detection step.

In Fig. 3 we can see different images of the dataset, we take the same pattern proposed by the author of segmenting the image into sub-images of 512 by 512, in each sub image we can see that there are small and large buildings, as well as few changes in one image and many changes in another.

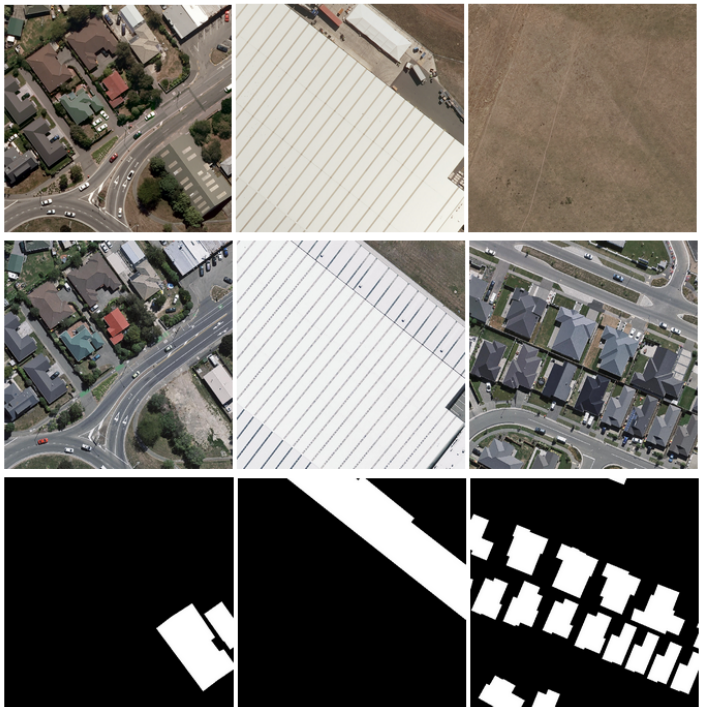

Fig. 3. Example of Database Images. First Row is Images of 2011. Second Row is Images of 2016. Third Row is Change Label.

B. Experiments on Building Detection

We used Mask R-CNN for building detection and use the same protocol used in the work of [19]. Mask R-CNN was pretrained with the COCO dataset, this converged after 30 epochs and the process took about 18 hours.

The test area consists of 1920 tiles of 512 × 512 pixels, in two different times, 2011 and 2016. Table I shows the results applying Mask R-CNN to the two bitemporal data sets to classify objects considering the buildings as objects. We compare our results with the results obtained in the work of [19].

TABLE I. Comparison of methods for building detection

Time	Method	Accuracy
2011	Mask R-CNN [19]	0.892
2011	MS-FCN [19]	0.922
2011	Our Mask R-CNN	0.866
2016	Mask R-CNN [19]	0.922
2016	MS-FCN [19]	0.939
2016	Our Mask R-CNN	0.897

Our results are similar to the results obtained in [19], although they are slightly lower, it may be because the training and validation data are generated without considering overlapping and in [19] it considers overlapping.

In Fig. 4 shows an example of building detection with their respective masks per building, it is observed that the mask is not perfect, the edges of the buildings are oval when they should be straight, this is a drawback of performing instance segmentation, due to the complexity it represents and to the different shapes and sizes that buildings can have, but this building extraction does not greatly affect the results obtained.

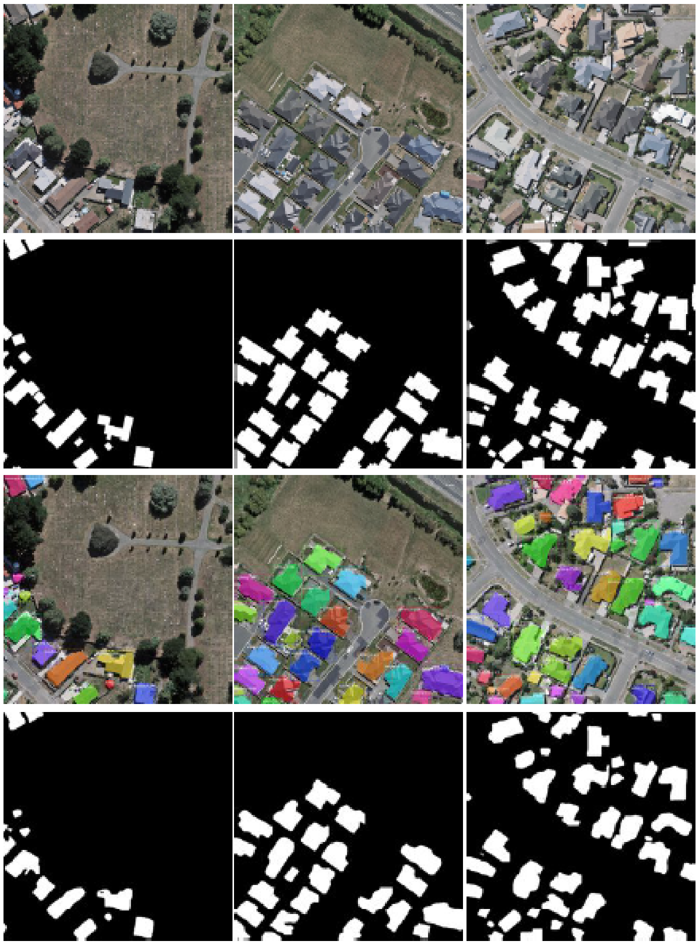

Fig. 4. Example of Building Detection. First Row is the Images of Different Zones. Second Row is the Ground Truth Masks. Third Row is the Title of the First Row with the Building Mask Detected. Fourth Row is the Detected Mask using Mask R-CNN

C. Building Change Detection Results

The binary building maps obtained in the previous step are taken and the pre-processing is carried out using the same criteria as [19], where all the buildings that are less than 500 pixels in size corresponding to a size of 4.8 x 4.8 m2 are eliminated, considering that buildings of that size are not usual, so they are considered a false detection.

Then, we extract the characteristics based on the comparison metrics MSE, PSNR and SSIM for each representation of the image, obtaining a feature vector of length 30 for each pair of images. Finally we use two classifiers, the first is Support Vector Machine for which we use a linear kernel and the second is Random Forest. For Random Forest we analyze which are the best parameters for number of trees N and the depth of each tree d, in the Fig. 5 it is observed that RF begins to converge with a value greater than 25 for N and in Fig. 6 it is observed that begins to converge with a value greater than 15 for d. For this reason we consider a value of 26 for N and a value of 16 for d.

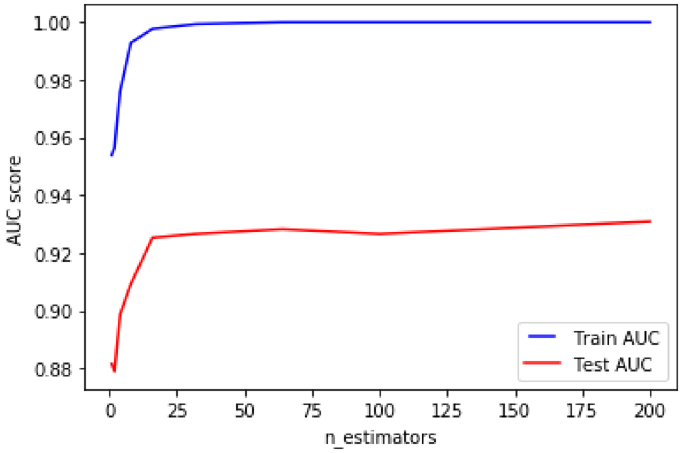

Fig. 5. Area under the Curve Considering a Number of Trees N from 1 to 200, to Random Forest.

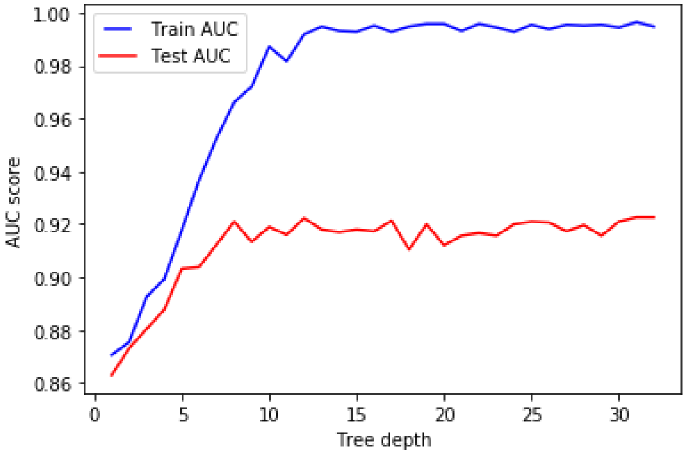

Fig. 6. Area under the Curve Considering a Depth per Tree d from 0 to 35, to Random Forest.

Table II shows a comparison of five different methods evaluated using three different metrics, for AP (counted on changed building instances), the best result is obtained with Our Mask R-CNN with SVM with a value of 0.854, in terms of recall the results obtained by almost all the methods are very similar close to 0.89 and in the same way for accuracy the results are similar close to 0.9, the FC-EF method is the only one that obtains results far below the other methods.

The method that obtains the best results in the state of the art is Mask R-CNN [17], so we can conclude that our proposal using Mask R-CNN with SVM obtains comparable results with that related work to which we obtain an accuracy of 0.911.

IV. Future Works

Explore different techniques for the detection of buildings, this step directly affects the building change detection, so an improvement in the accuracy of building detection, improve the precision of the general model proposed in this article.

Evaluate the proposed model in different data sets, in order to test its scalability.

TABLE II. Comparison of Methods for Building Change Detection.

Method	AP	Accuracy	Recall
Mask R-CNN [19]	0.814	0.910	0.883
MS-FCN [19]	0.796	0.891	0.872
FC-EF [16]	0.254	0.519	0.462
Our Mask R-CNN with SVM	0.854	0.911	0.891
Our Mask R-CNN with RF	0.852	0.891	0.899

V. Conclusions

In this article we propose a new framework for detecting changes in buildings using high-resolution images, for this we use a neural network for change detection and handcrafterd features for change detection. Experiments show that the results obtained by our proposal are comparable to those obtained in the state of the art. In this study we evaluate the detection of changes in buildings ignoring other types of objects such as bridges, tracks among others. Building detection directly affects change detection, so it is necessary to improve the precision of this step to improve the precision of the general model.

Acknowledgment

I would like to thank the Universidad Nacional de San Agustín de Arequipa for the financing provided with contract TP-015-2018, for which the development of this research work was possible.

References

[1] P. Sidike, D. Prince, A. Essa, and V. Asari, “Automatic building change detection through adaptive local textural features and sequential background removal,” 07 2016.

[2] X. Huang and L. Zhang, “Morphological building/shadow index for building extraction from high-resolution imagery over urban areas,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing - IEEE J SEL TOP APPL EARTH OBS*, vol. 5, pp. 161–172, 02 2012.

[3] Y. Tang, X. Huang, and L. Zhang, “Fault-tolerant building change detection from urban high-resolution remote sensing imagery,” *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 1060–1064, 09 2013.

[4] X. Huang, L. Zhang, and T. Zhu, “Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index,” *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, pp. 105–115, 01 2014.

[5] D. Wen, X. Huang, L. Zhang, and J. Benediktsson, “A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 1–17, 01 2015.

[6] B. Qi, Q. Kun, Z. Han, H. Wenjun, L. Zhili, and X. Kai, “Building change detection based on multi-scale filtering and grid partition,” 08 2018, pp. 1–8.

[7] C. Zhong, Q. Xu, F. Yang, and L. Hu, “Building change detection for high-resolution remotely sensed images based on a semantic dependency,” 07 2015, pp. 3345–3348.

[8] A. O. Ok, "Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, p. 21–40, 12 2013.

[9] S. Ghaffarian and S. Ghaffarian, "Automatic building detection based on purposive fastica (pfica) algorithm using monocular high resolution google earth images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 97, p. 152–159, 11 2014.

[10] F. Dornaika, A. Moujahid, Y. El Merabet, and Y. Ruichek, "Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors," *Expert Systems with Applications*, vol. 58, 03 2016.

[11] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, pp. 1–13, 08 2018.

[12] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 406–417, 01 2014.

[13] P. Sidike, D. Prince, A. Essa, and V. Asari, "Automatic building change detection through adaptive local textural features and sequential background removal," 07 2016.

[14] B. Chen, L. Deng, Y. Duan, S. Huang, and J. Zhou, "Building change detection based on 3d reconstruction," 09 2015, pp. 4126–4130.

[15] B. Chen, Z. Chen, L. Deng, Y. Duan, and Z. Jie, "Building change detection with rgb-d map generated from uav images," *Neurocomputing*, vol. 208, pp. 350 – 364, 06 2016.

[16] R. Daudt, B. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," 10 2018, pp. 4063–4067.

[17] S. Ji, Y. Shen, and M. Lu, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, p. 1343, 06 2019.

[18] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[19] S. Ji, Y. Shen, and M. Lu, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, p. 1343, 06 2019.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," vol. 1, 07 2005, pp. 886–893.

[21] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, pp. 18 – 28, 08 1998.

[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.