BUILDING OUTLINE EXTRACTION FROM AERIAL IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

F. Alidoost [1], H. Arefi [1], *, F. Tombari [2]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran - (falidoost, hossein.arefi)@ut.ac.ir
[2] Chair for Computer Aided Medical Procedures & Augmented Reality, Faculty of Informatics, Technical University of Munich, Germany - tombari@in.tum.de

KEY WORDS: Building Detection, Deep Learning, Active Contour Models, Selective Search, Depth Prediction

ABSTRACT:

Automatic detection and extraction of buildings from aerial images are considerable challenges in many applications, including disaster management, navigation, urbanization monitoring, emergency responses, 3D city mapping and reconstruction. However, the most important problem is to precisely localize buildings from single aerial images where there is no additional information such as LiDAR point cloud data or high resolution Digital Surface Models (DSMs). In this paper, a Deep Learning (DL)-based approach is proposed to localize buildings, estimate the relative height information, and extract the buildings' boundaries using a single aerial image. In order to detect buildings and extract the bounding boxes, a Fully Connected Convolutional Neural Network (FC-CNN) is trained to classify building and non-building objects. We also introduced a novel Multi-Scale Convolutional-Deconvolutional Network (MS-CDN) including skip connection layers to predict normalized DSMs (nDSMs) from a single image. The extracted bounding boxes as well as predicted nDSMs are then employed by an Active Contour Model (ACM) to provide precise boundaries of buildings. The experiments show that, even having noises in the predicted nDSMs, the proposed method performs well on single aerial images with different building shapes. The quality rate for building detection is about 86% and the RMSE for nDSM prediction is about 4 m. Also, the accuracy of boundary extraction is about 68%. Since the proposed framework is based on a single image, it could be employed for real time applications.

1. INTRODUCTION

One of the most important applications of remotely sensed data focuses on the detection/extraction, identification, localization and characterization of man-made structures including buildings. Precise and up-to-date information regarding the buildings' locations are essential and invaluable for various application such as search and rescue, monitoring, security and surveillance, navigation, and civil infrastructure inspection. On the other hand, with the advent of remote sensing technologies and artificial intelligent techniques, demands and interests of using aerial images for 3D localization and mapping are keep increasing. Compared with satellite images applied to remote sensing applications, aerial images, acquired by both aircrafts and Unmanned Aerial Vehicles (UAVs), offer an affordable, fast and effective approach for acquisition of high resolution multi-view aerial images over small areas. However, because of spatial variation of buildings, including shape, size, materials, colour, structure, and interference of building shadows, building detection and extracting building boundaries from single aerial images are often challenging and need manual works (Alidoost and Arefi, 2018).

Several methods are available for extracting buildings from a single image. Some of these recent algorithms include an energy based optimization algorithm using the Local Gradient Orientation Density (LGOD) (Benedek et al., 2012), a graph-based algorithm using shadow information of buildings (Izadi and Saeedi, 2012; Ok et al., 2013), a combination of the k-means clustering algorithm and a Purposive FastICA model (Ghaffarian and Ghaffarian, 2014), the multi label graph partitioning strategy (Manno-Kovacs and Ozgun Ok, 2015), a combination of Gaussian Mixture Model (GMM) clustering and Conditional Random Field (CRF) classification algorithms (Li et al., 2015), a self-supervised decision fusion framework (Senaras and Yarman Vural, 2015), and a supervised segmentation algorithm based on the image descriptors (Dornaika et al., 2016).

Compared to those traditional methods applied to building detection, the Deep Learning methods such as Convolution Neural Networks (CNNs) are recently employed for urban image classification (Alidoost and Arefi, 2016; Makantasis et al., 2015; Saito and Aoki, 2015; Vakalopoulou et al., 2015; Yuan, 2016; Zhang et al., 2016). Kaiser et al. (Kaiser et al., 2017) employed the Fully Convolutional Networks (FCNs) including skip connection layers to classify the buildings and roads in aerial images. Persello and Stein (Persello and Stein, 2017) developed a FCN, in which novel convolutional layers including dilated kernels were used for binary segmentation of satellite images which was resulted in two building and non-building segments. Wen et al. (Wen et al., 2019) modified the Mask Region CNN to extract the oriented bounding boxes of buildings. Srivastava et al. (Srivastava et al., 2017) proposed a pyramidal encoder-decoder CNN for both building extraction and DSM prediction to jointly estimate height and semantically label monocular aerial images.

In this study, we propose to address the problem of building outline extraction by exploiting the Convolutional Neural Networks as well as Active Contour Models. To this end, the CNNs-based classification is employed to detect buildings automatically resulting the initial boundaries of buildings (e.g. bounding boxes), while the CNNs-based regression is used to

* Corresponding author

https://doi.org/10.5194/isprs-archives-XLII-4-W18-57-2019 | © Authors 2019. CC BY 4.0 License.

provide the height information (e.g. nDSMs) from a single RGB image. The precise building outlines are then extracted by integration of extracted initial boundaries and nDSMs using an Active Contour Model. This work's contributions are as follows.

- It proposes an automatic CNN-based framework (FC-CNN) for individual building detection.
- It proposes an automatic CNN-based framework (MS-CDN) for depth prediction of single aerial images.
- Since ACMs need initial values of objects' boundaries, the proposed method uses the combination of the selective search algorithm and FC-CNN's results to provide bounding boxes of buildings.
- The ACM uses depth information instead of spectral information to provide more accurate boundaries.

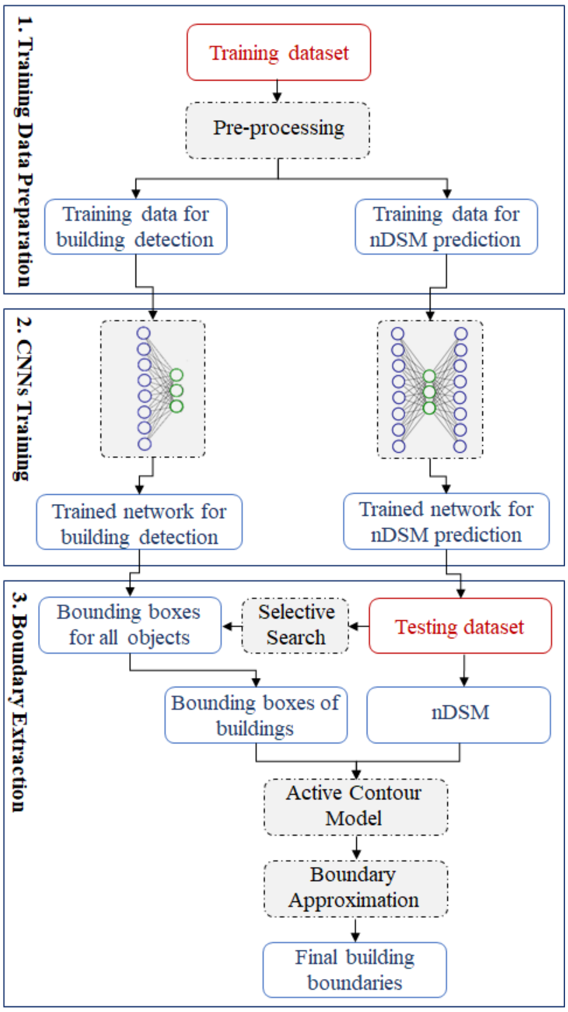

1. Training Data Preparation
Training dataset
Pre-processing
Training data for building detection
Training data for nDSM prediction
2. CNNs Training
Trained network for building detection
Trained network for nDSM prediction
3. Boundary Extraction
Bounding boxes for all objects
Selective Search
Testing dataset
Bounding boxes of buildings
nDSM
Active Contour Model
Boundary Approximation
Final building boundaries

Figure 1. The proposed method

2. PROPOSED METHOD

In this paper, a sequential framework is proposed for automatic building localization and outline extraction using supervised CNNs as well as Active Contour Models as shown in Figure 1. The main steps include training data preparation, CNNs training, and boundary extraction. First, two training datasets are generated for building detection and nDSM prediction, respectively. Next, the training data for building detection is employed to train a Fully Connected Convolutional Neural Network (FC-CNN), while the training samples for nDSM prediction is used to train a Multi-Scale Convolutional-Deconvolutional Network (MS-CDN). The third step is composed of four sub-steps. The nDSM of single aerial images are first predicted using the trained MS-CDN. Next, the selective search algorithm is employed to extract bounding boxes of all objects. The image tiles, related to the bounding boxes, are then fed into the trained FC-CNN to classify building objects and non-building objects. Finally, the bounding boxes of buildings and nDSMs are fed into an Active Contour Model (ACM) to generate the outlines of buildings. The summary of each step and their main components are given in the following sub-sections.

2.1 Training data preparation

Since the proposed method includes two different goals as building detection and height prediction, two different training datasets are required. For building detection, a dataset including building and non-building objects is generated by manually cropping the high resolution aerial images. Therefore, the building and non-building classes contain several image tiles with the same size of 224×224×3. For height prediction, the training dataset is composed of the aerial images with the size of 224×224×3 and corresponding nDSMs with the size of 224×224×1. However, the number of image tiles is not sufficient to train all of CNN's parameters. To overcome this issue the data augmentation technique is also applied to both training datasets. The data augmentation include scaling, cropping randomly, rotating randomly between [-5, 5] degrees, and flipping horizontally and vertically.

2.2 CNNs training

For building detection, a CNN with a fully connected layer at the end (e.g. FC-CNN) is utilized which is based on the ResNet-50 architecture (Kaiming et al., 2016). The input of the network includes the image tiles in two classes of building and non-building objects and the output of the network is a score matrix in [0, 1]. The dimension of the score matrix is ***S***×***C***, where ***S*** and ***C*** stand for the number of image tiles and the number of classes, respectively. For each image tile, the maximum score demonstrates the corrected label. The network is trained from the scratch using random initial values for the learnable parameters. Moreover, the mini batch Stochastic Gradient Descent (SGD) algorithm and the softmax log loss function, given by Eq. 1, are used for training the FC-CNN.

$$L(x,c) = -\log(\exp(x(c))/\sum exp(x(q))) \tag{1}$$

where ***c*** is the reference label and the ***x*** is the predicted label.
For depth prediction, a modern Multi-Scale Convolutional-Deconvolutional Network (MS-CDN) is proposed including coarse and fine prediction scales. The details of the proposed architecture is shown in Figure 2. The coarse prediction scale includes the convolutional and de-convolutional sub-networks to predict the global depth information, while the fine prediction scale is used to enhance the details of predicted coarse depth maps. To boost the performance of the network, we also added three skip connection layers for each encoder and decoder parts. To train the MS-CDN, the reverse Huber (berHu) function is employed, inspired by (Laina, et al. 2016). The berHu function considers the L_1 norm and the L_2 norm based on the Eq. (2).

$$B(x) = \begin{cases} |x| & |x| \le c \\ \frac{x^2+c^2}{2c} & |x| > c \end{cases} \tag{2}$$

where ***x*** is the pixel-wise difference between the predicted depth map and the ground truth and the ***c*** is selected as 20% of the maximum error for each training batch.

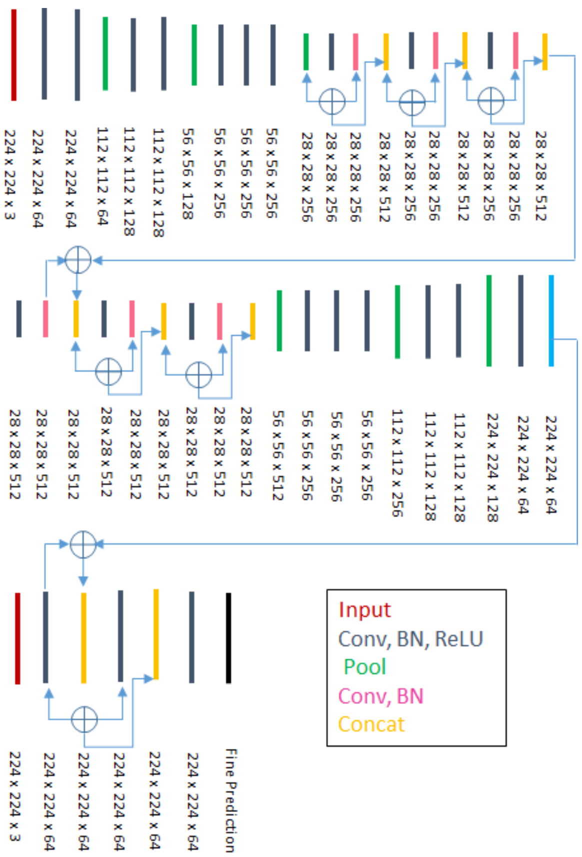


Figure 2. The proposed MS-CDN for height prediction

2.3 Boundary extraction

In the final step of the proposed method, a test dataset including aerial images is selected outside the training datasets. As shown in Figure 1, the candidate regions are first extracted from the test dataset using the selective search algorithm (Felzenszwalb and Huttenlocher, 2004), which is based on a graph segmentation method. The output of this step are bounding boxes including several objects at all scales and with the different sizes. Therefore, the candidate regions, which are image tiles, can be generated by cropping the test image for each bounding box. Next, the extracted candidate regions are fed into the trained FC-CNN and subsequently the candidate regions' score matrix is calculated. The maximum score in each class is the final label for each candidate region and consequently, the candidate regions are classified into two classes of building and non-building objects. The advantage of the selective search algorithm is to generate the bounding boxes which are appropriate initial values for the building boundaries. As a primary result, the buildings as well as the initial boundaries are extracted from a single image. On the other hand, the MS-CDN is applied to the test image to predict the nDSM. Since the ACM (Chan and Vese, 2001) needs the initial values to detect the object boundaries, the extracted bounding boxes are then used as the initial values and the ACM is applied to the predicted nDSM to detect the building boundaries. The experiments show that the ACM leads to the better results using nDSMs, instead of RGB images. The outputs of the ACM are initial polygons of building blocks which are not regular polygons. Therefore, in the next step, the Minimum Bounding Rectangle (MBR)-based technique (Arefi and Reinartz, 2013) is employed for approximation. The MBR-based technique is an iterative method based on searching the best rectangular polygon by fitting the bounding boxes to the initial polygon at each iteration. The outputs of MBR-based techniques are the final building outlines with regular and geometric shapes.

3. EXPERIMENTS AND RESULTS

Two datasets of this study includes aerial images from Stuttgart and Potsdam, Germany consisting of true ortho-images with a GSD of 20 cm and 5 cm respectively as well as corresponding nDSMs. These datasets are divided into two test and training subsets. For FC-CNN learning, the training subset from Stuttgart was selected and cropped manually into 500 image tiles per class of building and non-buildings. A sample of those aerial images as well as a sub-set of generated training dataset are shown in Figure 3. For MS-CDN learning, the training subset of the Potsdam is selected and both ortho images and nDSMs are cropped randomly (Figure 4). To increase the size of generated training datasets, the data augmentation process is employed like random cropping, rotating, scaling, and flipping. After the data augmentation process, the training datasets include 15000 and 35000 image tiles, respectively.

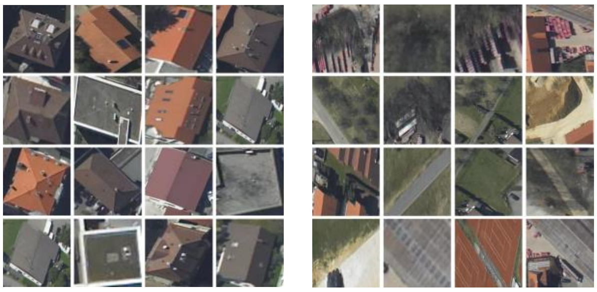
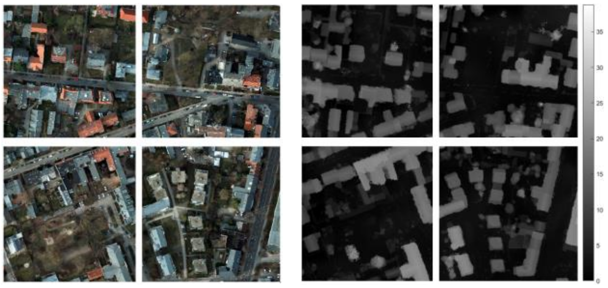

(a) (b)

Figure 3. Training data for the building detection including buildings (a); and non-buildings (b) objects



(a) (b)

Figure 4. The training samples including ortho-images (a); and the corresponding nDSMs (b)

Both FC-CNN and MS-CDN are trained using the MATLAB Deep Learning toolbox on a single NVIDIA GeForce GTX 1080 Ti and with a batch size of 16 for 100 epochs for the building detection and height prediction tasks. The learning rate and the momentum are about 0.01 and 0.9, respectively.

The trained networks are applied to a test dataset including a single RGB image (Figure 5a) from Potsdam images and outside of the training samples. The results of extracted buildings and bounding boxes (Figure 5c), predicted nDSM (Figure 5d), ACM's output (Figure 5e), and final outlines of buildings (Figure 5f) are illustrated in Figure 5.

In the building detection step, the results are evaluated using the standard quality measures of Completeness (or Recall), Correctness (or Precision), and Quality (McGlone and Shufelt, 1994; McKeown et al., 2000) as given in Eq. (3).

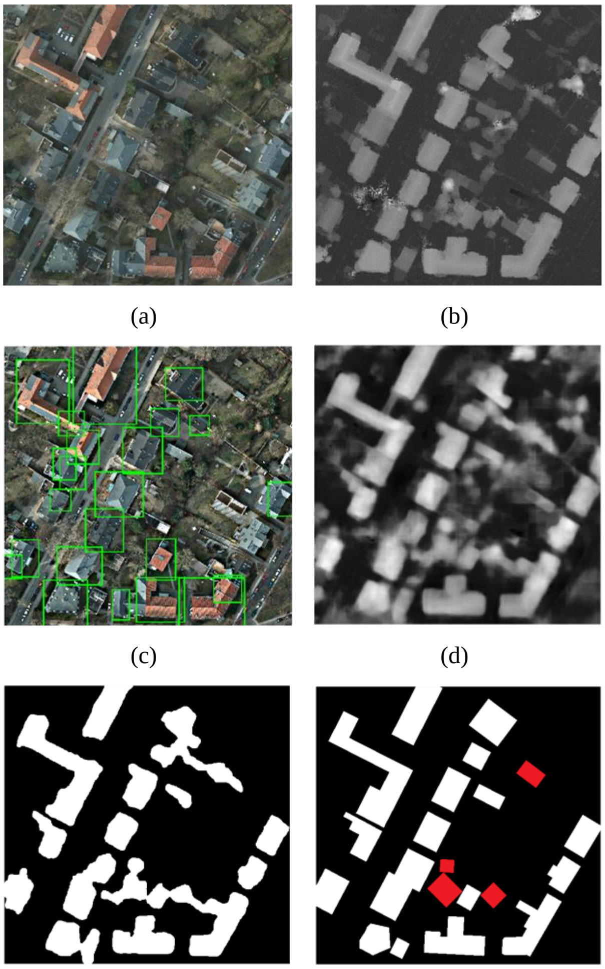
$$Comp. = \frac{TP}{TP+FN}; Corr. = \frac{TP}{TP+FP}; Qual. = \frac{TP}{TP+FN+FP} \tag{3}$$

where ***TP*** is True Positive (the number of correctly detected buildings), ***FP*** is False Positive (the number of non-building objects detected as buildings), and ***FN*** is False Negative (the number of undetected buildings). To evaluate the predicted nDSMs, the error metrics such as the Relative Error (***REL***), Root Mean Squared Error (***RMSE***) and Root Mean Squared Logarithmic Error (***RMSLE***) are used as Eq. (4).

$$REL = \tfrac{1}{T}\textstyle\sum_{i,j}|y-\tilde{y}|/y;\ RMSE = \sqrt{\tfrac{1}{T}\textstyle\sum_{i,j}|y-\tilde{y}|^2}; \tag{4}$$

$$RMSLE = \sqrt{\frac{1}{T}\sum_{i,j}|logy - log\tilde{y}|^2}$$

where ***y*** is the ground truth, $\tilde{\boldsymbol{y}}$ is the predicted nDSM, and ***T*** is the number of pixels. The quantitative values of different metrics are reported in Table 1.



(a) (b)

(c) (d)

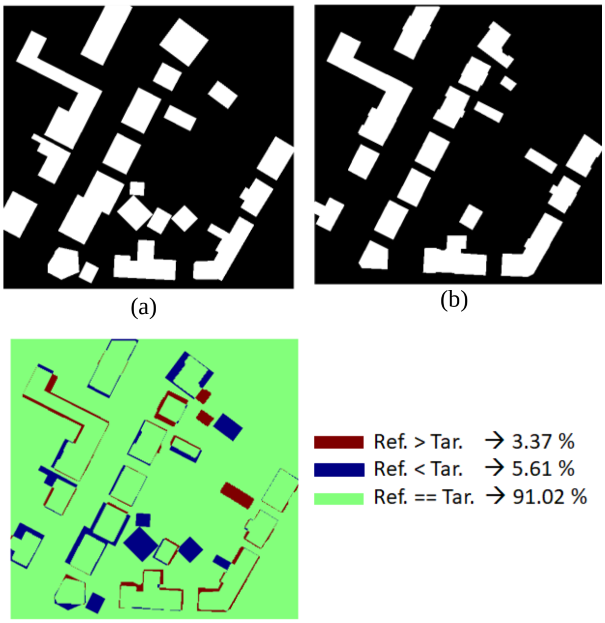
(e) (f)

Figure 5. The test data and the results: a) the test RGB image, b) the ground truth nDSM, c) the detected buildings and bounding boxes, d) the predicted nDSM, e) the ACM's output, and f) the final outlines of buildings

As shown in Table 1, the accuracy of building detection is about 86%. This accuracy is acceptable because the FC-CNN is trained on the Stuttgart dataset and applied to the Potsdam test data which shows the generalization capability of the trained network. However, most of the large building blocks are detected correctly and the errors are related to the small or ambiguous buildings. In addition, the accuracy of the predicted nDSM is about 3.57 m which is a promising results for depth prediction from a single RGB image. To evaluate the quality of final building boundaries, the Intersection over Union (IoU) metric is calculated to quantify the overlap percentage between the extracted boundaries and reference boundaries, which is obtained about 68%. The low accuracy is because of non-buildings objects such as trees (i.e. red polygons in Figure 5f) which are detected by the ACM. The difference map between the final outlines of buildings (Figure 6a) and the ground truth (Figure 6b) per-pixel level is shown in Figure 6c. The green segments are true positive pixels, the red segments are false negative pixels, and the blue segments are false positive pixels. Consequently, there is a similarity ratio of about 91% between the extracted boundaries and reference boundaries. As shown in Figure 6, the large differences are corresponding to the non-building objects such as trees.

Task	Accuracy metrics		
	Comp.	**Corr.**	**Qual.**
Building detection	86 %	100 %	86 %
	REL	**RMSE**	**RMSLE**
Height prediction	0.4%	3.57 m	0.23 m

Table 1. The quantitative results of the proposed method



(a) (b)

(c)

Figure 6. The difference map (c) between the extracted outlines (a) and the ground truth (b)

4. CONCLUSIONS

In this study, we proposed a novel ensemble approach based on supervised deep learning techniques to extract the precise outlines of buildings from a single aerial image. Unlike current methods in photogrammetry and remote sensing that require both ortho images as well as high resolution DSMs, the proposed method uses the single images and the power of CNNs to extract the valuable information like building boundaries and height values. Although we had some limitations to produce the proper training datasets, the results showed the reasonable performance of the proposed CNNs to detect buildings with the quality rate of 86%, extract the initial bounding boxes and predict the nDSMs with the RMSE of 3.57 m. Moreover, the precise outlines of buildings are extracted with the accuracy of 91% which shows the effectiveness of the proposed framework.

REFERENCES

Alidoost, F., Arefi, H., 2018. A CNN based Approach for Automatic Building Detection and Recognition of Roof Models using a Single Aerial Image, *Photogramm., Remote Sens. and Geoinf. Sci. (PFG)*, 86(5-6), 235-248, DOI 10.1007/s41064-018-0060-5.

Arefi, H., Reinartz, P., 2013. Building Reconstruction Using DSM and Orthorectified Images. *Remote Sens*. 5, 1681–1703. https://doi.org/10.3390/rs5041681

Benedek, C., Descombes, X., Zerubia, J., 2012. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Trans. Pattern Anal. Mach. Intell*. 34, 33–50. https://doi.org/10.1109/TPAMI.2011.94

Chan, T.F., and Vese, L.A., 2001. Active Contours without Edges, *IEEE Trans. on Image Processing*, 10(2).

Dornaika, F., Moujahid, A., El Merabet, Y., Ruichek, Y., 2016. Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors. *Expert Syst. Appl*., 58, 130–142. https://doi.org/10.1016/j.eswa.2016.03.024

Felzenszwalb, PF., Huttenlocher, DP., 2004. Efficient graph-based image segmentation. *Int J Comp Vision,* 59(2):167–181. https ://doi.org/10.1023/B:VISI.00000 22288 .19776 .77

Ghaffarian, Saman, Ghaffarian, Salar, 2014. Automatic building detection based on Purposive FastICA (PFICA) algorithm using monocular high resolution Google Earth images. *ISPRS J. Photogramm. Remote Sens.,* 97, 152–159. https://doi.org/10.1016/j.isprsjprs.2014.08.017

Izadi, M., Saeedi, P., 2012. Three-Dimensional Polygonal Building Model Estimation From Single Satellite Images. Geosci. *Remote Sensing, IEEE Trans.*, 50, 2254–2272. https://doi.org/10.1109/TGRS.2011.2172995

Kaiming, H., Xiangyu, Z., Shaoqing, R., Sun, J., 2016. Deep Residual Learning for Image Recognition, *CVPR 2016*, Las Vegas, USA, pp. 1–9. https://doi.org/10.1109/CVPR.2016.90

Kaiser, P., Wegner, J.D., Aurélien, L., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. *IEEE Trans. Geosci. Remote Sens.* 55, 6054–6068. https://doi.org/10.1109/TGRS.2017.2719738

Laina, I., Rupprecht, Ch., Belagiannis, V., Tombari, F., and Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. CoRR, abs/1606.00373, 2016. doi: 10.1109/3dv.2016.32. URL http://arxiv.org/abs/1606.00373.

Li, E., Femiani, J., Xu, S., Zhang, X., Wonka, P., 2015. Robust rooftop extraction from visible band images using higher order CRF. *IEEE Trans. Geosci. Remote Sens*. 53, 4483–4495. https://doi.org/10.1109/TGRS.2015.2400462

Makantasis, K., Karantzalos, K., Doulamis, A., Doulamis, N., 2015. Deep supervised learning for hyperspectral data classification through convolutional neural networks, *IGARSS 2015*, Milan, Italy, 4959–4962. https://doi.org/10.1109/IGARSS.2015.7326945

Manno-Kovacs, A., Ozgun Ok, A., 2015. Building Detection from Monocular VHR Images by Integrated Urban Area Knowledge. *IEEE Geosci. Remote Sens. Lett.* 12, 2140–2144. https://doi.org/10.1109/LGRS.2015.2452962

McGlone, J.C., Shufelt, J.A., 1994. Projective and object space geometry for monocular building extraction, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94. *IEEE Comput. Soc.*, USA, 54–61. https://doi.org/10.1109/CVPR.1994.323810

McKeown, D.M., Bulwinkle, T., Cochran, S., Harvey, W., McGlone, C., Shufelt, J.A., 2000. Performance evaluation for automatic feature extraction, *Int. Archives of the Photogramm, Remote Sens. Spatial Inf. Sci*., Amsterdam, The Netherlands, 379–394.

Ok, A.O., Senaras, C., Yuksel, B., 2013. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. *IEEE Trans. Geosci. Remote Sens.,* 51, 1701–1717. https://doi.org/10.1109/TGRS.2012.2207123

Persello, C., Stein, A., 2017. Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images. IEEE Geosci. *Remote Sens. Lett.,* 14, 2325–2329. https://doi.org/10.1109/LGRS.2017.2763738

Saito, S., Aoki, Y., 2015. Building and road detection from large aerial imagery, in: *Proceedings of SPIE - The International Society for Optical Engineering*., San Francisco, California, USA. https://doi.org/10.1117/12.2083273

Senaras, C., Yarman Vural, F. atos T., 2015. A Self-Supervised Decision Fusion Framework for Building Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 9, 1780–1791. https://doi.org/10.1109/JSTARS.2015.2463118

Srivastava, S., Volpi, M., Tuia, D., 2017. Joint Height Estimation and Semantic Labeling of Monocular Aerial Images with CNNs, *IGARSS 2017.* https://doi.org/10.1109/IGARSS.2017.8128167

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features, *IGARSS 2015,* Milan, Italy, pp. 1873–1876. https://doi.org/10.1109/IGARSS.2015.7326158

Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., Wang, P., 2019. Automatic building extraction from google earth images under complex backgrounds based on deep instance segmentation network. *Sensors,* 19, 1–16. https://doi.org/10.3390/s19020333

Yuan, J., 2016. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. Oak Ridge, Tennessee.

Zhang, Q., Wang, Y., Liu, Q., Liu, X., Wang, W., 2016. CNN based suburban building detection using monocular high resolution Google Earth images, *IGARSS 2016,* 661–664. https://doi.org/10.1109/IGARSS.2016.7729166