BUILDING EXTRACTION FROM REMOTE SENSING DATA USING FULLY CONVOLUTIONAL NETWORKS

K. Bittner[a], S. Cui[a], P. Reinartz[a]

[a] Remote Sensing Technology Institute, German Aerospace Center (DLR), D-82234 Wessling, Germany
(ksenia.bittner, shiyong.cui, peter.reinartz)@dlr.de

Commission VI, WG VI/4

KEY WORDS: deep learning, DSM, fully convolutional networks, building footprint, binary classification, fully connected CRF

ABSTRACT:

Building detection and footprint extraction are highly demanded for many remote sensing applications. Though most previous works have shown promising results, the automatic extraction of building footprints still remains a nontrivial topic, especially in complex urban areas. Recently developed extensions of the CNN framework made it possible to perform dense pixel-wise classification of input images. Based on these abilities we propose a methodology, which automatically generates a full resolution binary building mask out of a *Digital Surface Model (DSM)* using a Fully Convolution Network (FCN) architecture. The advantage of using the depth information is that it provides geometrical silhouettes and allows a better separation of buildings from background as well as through its invariance to illumination and color variations. The proposed framework has mainly two steps. Firstly, the FCN is trained on a large set of patches consisting of normalized DSM (nDSM) as inputs and available ground truth building mask as target outputs. Secondly, the generated predictions from FCN are viewed as unary terms for a Fully connected Conditional Random Fields (FCRF), which enables us to create a final binary building mask. A series of experiments demonstrate that our methodology is able to extract accurate building footprints which are close to the buildings original shapes to a high degree. The quantitative and qualitative analysis show the significant improvements of the results in contrast to the multy-layer fully connected network from our previous work.

1. INTRODUCTION

1.1 Related work

Building detection and footprint extraction are important remote sensing tasks and used in the fields of urban planning and reconstruction, infrastructure development, three-dimensional (3D) building model generation, etc. Due to the sophisticated nature of urban environments the collection of building footprints from remotely sensed data is not yet productive and time consuming, if it is manually performed. Therefore, automatic methods are needed in order to complete the efficient collection of building footprints from large urban areas comprising of numerous constructions. Recently, various approaches have been developed, which perform building extraction on the basis of high-resolution satellite imagery. Depending on the type of data employed for building extraction the existing methods can be divided into two main groups: using aerial or high-resolution satellite imagery and using three-dimensional (3D) information.

Aerial photos and high-resolution satellite images are extensively used in urban studies. The pioneering approaches proposed to extract edge, line and/or corner information, which are fundamental elements for buildings extraction (Huertas and Nevatia, 1988, Irvin and McKeown, 1989). Many studies additionally incorporate shadow information to the low-level features (Liow and Pavlidis, 1990, McGlone and Shufelt, 1994, Peng and Liu, 2005). Some methodologies formalize the building extraction problem in terms of graph theory (Kim and Muller, 1999, Krishnamachari and Chellappa, 1996, Sirmacek and Unsalan, 2009). Many researchers implemented more forward-looking methods to extract shapes of the detected buildings (Karantzalos and Paragios, 2009, Sirmacek et al., 2010). Further studies, which employ the advantages of multi-spectral information, solve the detection problem in a classification framework (Lee et al., 2003, Koc-San and Turker, 2014, Sumer and Turker, 2013). However, due to the complexity of shapes and variety of materials of human-made constructions, the image classification in urban areas is still complicated.

In order to use the advantage of height information from DSM, obtained from optical stereo images or light detection and ranging measurements (LIDAR), several works investigated the building footprint extraction from DSM alone or together with high-resolution imagery. In general, building detection from DSM is a very challenging task due to scene complexities and imperfections in the methodological steps for depth image generation such as stereo matching methods. As a result this leads to a presence of noise in the generated DSM. Although, the quality of stereo DSM concedes to the one from LIDAR data, they have become more popular in recent years due to their large coverage and lower costs as compared to LIDAR data. (Gerke et al., 2001) detects and generates building outlines from DSM by separating them from surrounding above-ground objects such as trees using Normalized Difference Vegetation Index (NDVI). Similar studies are employed in (San and Turker, 2006, Lu et al., 2002). (Krauß et al., 2012) introduced a methodology for DSM-based building mask generation by using an Advanced Rule-based Fuzzy Spectral Classification algorithm, which fuses nDSM with the classified multispectral imagery. Afterwards, the height thresholding is applied to extract buildings from other surrounding objects. The approach proposed in (Brédif et al., 2013) extracts rectangular building footprints directly from the DSM using a Marked Point Process (MPP) of rectangles and then refines them into polygonal building footprints.

In spite of the efforts put into developing methodologies for the

automatic extraction of building footprints from DSM, they are still not able to provide satisfactory results. Therefore, our goal is to implement such a methodology, which will automatically, without any assumptions on the shape and size of buildings, extract them from DSMs.

1.2 Deep Neural Networks for building extraction

With the revolutionary appearance of Convolutional Neural Networks (CNNs), which became the state-of-the-art for image recognition problems, it became possible to automatically detect buildings in remote sensing data. In (Yuan, 2016) the building footprints are automatically extracted from hight-resolution satellite images using Convolutional Network (ConvNet) framework. The authors in (Maggiori et al., 2017) propose to generate building mask out of RGB satellite imagery by using a FCN, firstly, trained on possibly inaccurate reference data, and, finally, refined on a small amount of manually labeled data. One of the first approaches for above-ground objects classification from high-resolution DSM using a deep learning technique, specifically a Multilayer Perceptron model, was demonstrated in the work of (Marmanis et al., 2015). In our previous study (Davydova et al., 2016), we developed a similar approach to create a binary building mask from DSM using a four-layer fully connected neural network. As a continuation of our previous work in this paper we present a methodology using a deep learning approach, for building footprint extraction from remote sensing data, particularly nDSM, with a focus on dense residential areas. Besides learning discriminative per-pixel classifiers, we further encode the output relationship from FCN as unary term for fully connected Conditional Random Fields (CRF) and generate a final building mask.

2. METHODOLOGY

2.1 Fully Convolutional Network

Traditional CNNs architectures were generated for image-level classification tasks, which require an input image of a fixed size $h \times w \times ch$ (h and w represent the spatial dimensions, and ch is the feature/channel dimension) and output a vector of class scores cl_i. The fully connected layers of such architectures have fixed dimensions and completely discard the spatial information. FCNs on the other hand transform fully connected layers as a large set of 1×1 convolutions allowing the network to take input of any size and output classification maps of class scores $cl_i(x, y)$. However, the maps generated by FCN per-class probability have smaller sizes and coarser spatial resolution compared to the input image due to the pooling layers along the network. The solution to this problem is to enlarge the FCN with deconvolution layers, which up-sample the previous layer. As a result, by adding several deconvolution operations at the top part of the network it allows to up-sample the coarse maps to the input image size and get the class scores for each pixel, performing an end-to-end learning by backpropagation from the pixel-wise loss. However, the output of such FCN (known as FCN-32s) cannot obtain satisfying object boundaries, because of the final deconvolution layer, which has the 32 pixel stride, restrict the scale of details. In order to refine object boundaries the high-frequency information from lower network layers is added with the help of the so-called "skip" layer. The "skip" layer combines the final prediction layer with the output of earlier convolutional layers with rich information. In this way, the FCN-16s adds "skip" connection from pool4, and FCN-8s propagates even more detailed information from the pool3 layer, in addition to the pool4 layer (see Figure 1).

	Input image	Kernel	Stride	Pooling	Feature maps
	Conv1_1 + ReLU1_1	3 x 3	1		64
	Conv1_2 + ReLU1_2	3 x 3	1		64
	Pool 1			2 x 2	
	Conv2_1 + ReLU2_1	3 x 3	1		128
	Conv2_2 + ReLU2_2	3 x 3	1		128
	Pool 2			2 x 2	
	Conv3_1 + ReLU3_1	3 x 3	1		256
	Conv3_2 + ReLU3_2	3 x 3	1		256
	Conv3_3 + ReLU3_3	3 x 3	1		256
	Pool 3			2 x 2	
	Conv4_1 + ReLU4_1	3 x 3	1		512
	Conv4_2 + ReLU4_2	3 x 3	1		512
	Conv4_3 + ReLU4_3	3 x 3	1		512
	Pool 4			2 x 2	
	Conv5_1 + ReLU5_1	3 x 3	1		512
	Conv5_2 + ReLU5_2	3 x 3	1		512
	Conv5_3 + ReLU5_3	3 x 3	1		512
	Pool 5			2 x 2	
	FCN_6 – ReLU6	1 x 1	1		4096
	Drop-out				
	FCN_7 – ReLU7	1 x 1	1		4096
	Drop-out				
	Conv_8	1 x 1	1		2
	Deconv_1	4 x 4	2		2
"Skip"	Conv_Pool 4	1 x 1	1		2
	Decon_1 + Conv_Pool 4	1 x 1	1		2
	Deconv_2	4 x 4	2		2
"Skip"	Conv_Pool 3	1 x 1	1		2
	Decon_2 + Conv_Pool 3	1 x 1	1		2
	Deconv_3	16 x 16	8		2
	Softmax + Loss				

Figure 1. Fully convolution network with "skip" connections. The last layers of the network are 2 dimensional feature maps as we have two classes: building and non-building. More detailed description is given in Section 2.1

2.2 Problem Formulation

In practice, very few people train an entire CNN from scratch (with random initialization) due to the limited amount of training data. Therefore, it is common to take a model, pre-trained on a very large dataset, and transfer its relevant knowledge as an initialization for a new task. Such models can be adapted then to the new task with relatively few training data. Nowadays, several networks exist, which have been pre-trained on huge image datasets. A FCN, proposed in (Long et al., 2015), was constructed based on VGG-16 network, for which a model pre-trained on the large public image repositories Imagenet (Deng et al., 2009) exists. We take this pre-trained model and fine-tune it for our task, but randomly initialize the last layers, because the channel dimension is changed to 2 in order to predict scores for our binary task.

However, in contrast to Imagenet, which contains RGB images as inputs, our training dataset consists of depth images, which carry completely different information compared to intensities. The main concern is whether it is applicable to use a pre-trained model. It turns out to be suitable due to the fact that RGB and depth images share common features such as edges, corners, endpoints, etc., at the low and middle level image domains.

The Imagenet database contains images which fit into the GPU size. Remote sensing images are huge and cannot be loaded into the GPU as a whole. As a result, in our work input image I, i.e., nDSM, and corresponding target image with labels M, i.e., a building mask, are tiled into patches of size $w \times h$: $Patch_i(I)_{w\times h}$ and $Patch_i(M)_{w\times h}$.

(Long et al., 2015) performed a cascaded training, starting from

the shallower network FCN-32s and gradually adding the "skip" connections to include the high-frequency information from the pool4 layer (FCN-16s) and then pool3 (FCN-8s). We apply the same procedure on our dataset: first, we fine-tune a 32 stride network, then 16 stride and finally 8 stride, and each next network uses the previous network's weights to speed up the training process. We fine-tune the models by minimizing the negative log likelihood and adjusting the weights and biases along the whole network with a backpropagation algorithm using stochastic gradient descent (SGD) with a small batch. Mathematically, we solve

$$\min_{\mathbf{W}} \sum_{i}^{N} \mathcal{L}(softmax(\mathbf{W}^{l} a(Patch_i(I)_{w\times h}, \mathbf{W}^{l-1}), y_i)), \quad (1)$$

where $\mathbf{W}^l$ are the weights of the last layer, $\mathbf{W}^{l-1}$ are the weights of the previous layer, $a(\cdot)$ is an activation function, y_i represents given true mask patches $Patch_i(M)_{w\times h}$. The softmax function is given by

$$softmax(\mathbf{z}) = \frac{\exp(\mathbf{z})}{\|\mathbf{z}\|_1} \quad (2)$$

and the loss function is computed as

$$\mathcal{L} = -\sum_{k} y_k \log p_k \quad (3)$$

where p_k is the label assignment probability at pixel k.

After training, we take the FCN-8s as a final classifier and perform predictions on a new unseen dataset. Those new data are forwarded though the network as separate patches and the predicted maps with the same size as the patches are obtained. After that, the tiles are stitched together in order to generate an image with the same size as the original DSM.

2.3 Fully connected Conditional Random Field for object boundaries enhancement

In order to generate a binary building mask we need to assign to each image pixel the best suitable label, where 1 corresponds to building and 0 to non-building/background label. At the same time we want to keep spatial correlations between neighboring pixels and accurately localize segment boundaries.

Modern Deep CNN architectures produce typically quite smooth classification results (Chen et al., 2014). Therefore, we are interested in obtaining detailed local structures (object boundaries) rather than further smooth it. It can be reached, by applying Fully connected Conditional Random Field (CRF) approach proposed by (Koltun, 2011). Fully connected CRF allows an elegant way to combine single pixel predictions and shared structure through unary and pairwise terms. It differs from standard CRF by establishing pairwise potentials on all pairs of pixels in the image and not only on neighboring pixels.

As described above, the predictions are computed by the chosen FCN-8s. These predictions can be seen as pixel-wise unary likelihoods $\phi_i(x_i)$ for the Fully connected CRF energy function shown in Equation (4).

$$E = \sum_{i}^{N} \phi_i(x_i) + \sum_{ij} \psi_{ij}(x_i, x_j) \quad (4)$$

The pairwise edge potentials is defined by a linear combination of Gaussian kernels and has the form

$$\psi_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} \omega^{(m)} k^{(m)}(f_i, f_j) \quad (5)$$

where μ is a label compatibility function, $k^{(m)}(f_i, f_j)$ is a Gaussian kernel, which depends on features (defined as f) extracted for pixel i and j and is weighted by parameter $\omega^{(m)}$. The kernels consist of two parts and are contrast-sensitive. They are defined as

$$k^{(m)}(f_i, f_j) = \omega^{(1)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}) + \quad (6)$$

$$\omega^{(2)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}) \quad (7)$$

where the first term called the *appearance kernel* and depends on both the pixel color intensities (I_i and I_j) and pixel positions (p_i and p_j). This term encourages assigning similar labels to nearby pixels with similar color. Parameter θ_α controls the degrees of nearness and θ_β of similarity. The second term called *smoothness kernel* is responsible for removing small isolated regions (Shotton et al., 2009).

As a result, applying the described methodology and minimizing the CRF energy $E(x)$ we search for the most probable label assignment for each pixel taking into account spatial correlations between them. This finally leads to a binary building mask.

3. STUDY AREA AND EXPEREMENTS

We perform experiments on datasets consisting of DSM reconstructed from WorldView-2 stereo panchromatic images with a resolution of 0.5 meter per pixel using the semi-global matching methodology proposed by (d'Angelo and Reinartz, 2011). In order to obtain a nDSM with only above-ground information a topography information was removed based on the methodology described in (Qin et al., 2016). As ground truth, a building mask covering the same region as DSM from the municipality of the city of Munich, Germany is used for learning the parameters in the neural network.

The fine-tuning was done on FCNs implemented in the $Caffe$ deep learning framework. For learning process we prepared a training data consisting of 7161 pairs of patches with size of 300×300 pixels. To avoid the artifacts and object discontinuity at tile boundaries the patches are generated with an overlap of 100 pixels. We start fine-tuning process of FCN-32s with a learning rate of 0.0001, decreasing it by a factor of 10 for each next stage of gradual learning. We used a weight decay of 0.0005 and a momentum of 0.99.

The final binary building mask was obtained using the FCRF software developed in (Koltun, 2011). We chose the smoothness θ_γ and appearance kernel parameters θ_α and θ_β after performing an experimental grid search varying the spatial and color ranges of these parameters and examining the resulting classification accuracy. As a result, we found that θ_γ = 3, θ_α = 3 and θ_β = 11 work well in practice. The weight parameters $\omega^{(1)}$ and $\omega^{(2)}$ were set to 1.

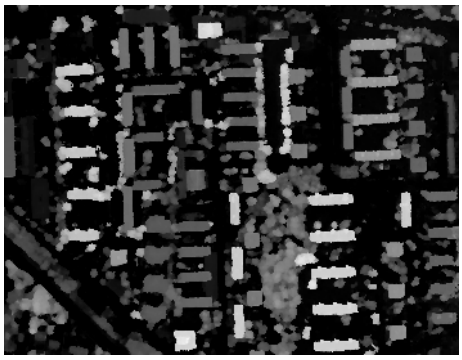


Figure 2. Normalized Digital Surface Model. Test region.

4. RESULTS AND DISCUSSIONS

In this section we present and discuss the results obtained from the methodology described above for binary building mask generation. After gradual training of FCNs, we use the final FCN-8 model for our binary classification task. To demonstrate the performance of our model we present to the network a new test dataset, which was used neither for training nor for validation (see Figure 2).

4.1 Qualitative Results

The building mask generated directly from the last layer of FCN-8 network is presented in Figure 3(b). As can be seen from the results, the FCN model is able to extract only the buildings from the nDSM without any influence of other above-ground objects. However, some noise could be noticed in the form of small regions near building boundaries together with irregularities of the boundaries itself. Therefore, the Fully connected CRF is applied as a post processing step in order to remove this noise and improve the boundaries (see Figure 3(c)). In Figure 3(d) we present the building mask extracted from the four-layer fully connected network from our previous work. The configuration of the network stayed the same as described in (Davydova et al., 2016), but in order to perform the comparison, we train it on the same dataset. In comparison to the four-layer network FCN can manage to extract complex building structures, without missing some of its parts (for example see red marks in left and down parts of Figure 3(c)) and with the size of footprints, which is closer to original one (see red mark on the upper part of Figure 3(c)). This can be explained as deep neural network learns to take into account the context within a 300×300 pixel window in contrast to 32×32 pixel window of four-layer network. Of course, one can increase the input patch size to a bigger one for multy-layer network, but it will heavily influence the computation time. Besides, FCN does not over-smooth the object boundaries and can identify more detailed building structures. The absence of some buildings on defined mask is due to the low sensitivity of our approach to the recognition of low-rise buildings, which are surrounded by higher buildings. Another possible reason is their small amount in the training dataset that limited our network in learning this kind of buildings. Besides, the low-raised buildings can be totally covered by the trees that makes their detection impossible.

4.2 Quantitative Results

To assess the quality of the proposed methodology on the selected test dataset in comparison to the ground truth shown in Figure 3(a) we used metrics commonly used for semantic segmentation and useful for binary classification task. Let TP, FP, FN denote the total number of true positive, false positive and false negative, respectively. Then we can define those metrics as following:

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where the $Precision$ is the fraction of predicted positives which are actual positive, the $Recall$ is the proportion of actual positives which are predicted positive. The higher these metrics, the better the performance of model. Besides, we employ the $F-measure$ (see Equation (10)) derived from the precision and recall values in Equations (8) - (9) for the pixel-based evaluation. For simplicity, we set β = 1. It reaches its best value at 100% and worst at 0%.

$$F_{measure} = \frac{(1 + \beta^2)TP}{(1 + \beta)^2 TP + \beta^2 FN + FP} \tag{10}$$

Another useful metric is Intersection over Union (IoU), which is an average value of the intersection of the prediction and ground truth regions over the union of them. Here we adapted this metric to the binary case, because in our data there are many more pixels which belong to the background than those belonging to the buildings areas. Therefore, in our case IoU is defined as the number of pixels labeled as building on both in the ground truth and predicted mask, divided by the total number of pixels labeled as buildings in each of them (Maggiori et al., 2017).

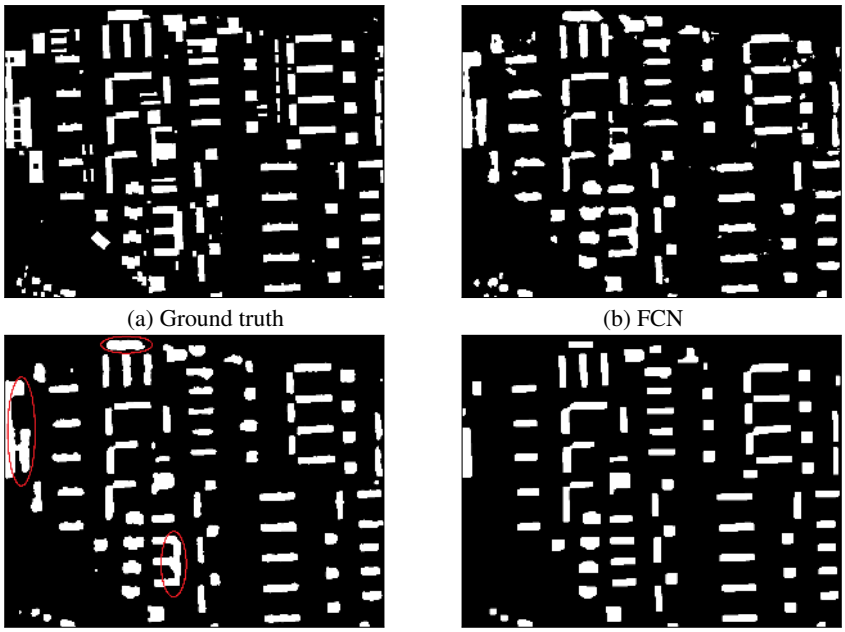
$$IoU = \frac{TP}{n_{pred} + n_{gt}}, \tag{11}$$

where n_{pred} is the number of pixels labeled as buildings in predicted mask and n_{gt} is the one in ground truth. The results are presented in Table 1.

The results show that doing a binary classification of remote sensing data by using a deep convolutional network, in our case the FCN-8, outperforms the binary mask generated by four-layer fully connected network. The FCN followed by a dense fully connected CRF refinement significantly improves the mask quality. This statement is also confirmed from the quantitative point of view (see second line of Table 1). The Intersection over Union (IoU) was not improved after applying a post processing step. This can be explained as for some buildings in DSM boundaries cannot be clearly seen due to their possible overlap with trees or inaccuracies of DSM itself. As a result the presence of inaccuracies in buildings outlines can be observed.

Table 1. Binary classification results on the test dataset in comparison to the four-layer network.

	$F_{measure}$	IoU	$Precision$	$Recall$
FNC-8	67%	41%	82%	79%
FNC-8+FCRF	70%	41%	86%	78%
4-layer net + MRF	66%	39%	84%	71%



(a) Ground truth (b) FCN

(c) FCN + FCRF (d) 4-layer NN + Markov Random Field (MRF)

Figure 3. Results of extraction of building footprints from test region.

5. CONCLUSION AND FUTURE WORK

Recent developments in neural network approaches have greatly advanced the performance of visual recognition tasks such as image classification, localization and detection. We proposed to use a fully convolutional network architecture for automatic building footprint extraction from remote sensing data, specifically the nDSM. The main advantage of using DSM data is that they provide the elevation information of the objects, which is crucial for the tasks such as buildings extraction in urban area. Because the satellite images are huge, we tile the nDSM and available reference building mask into patches. In the first step, the FCNs were trained on the prepared set of patches to extract the building footprints. The predictions, generated afterwords, are presented as unary terms to the Fully connected CRF and, finally, the binary building mask is obtained. Experimental results show that the deep neural network approaches are suitable for the remote sensing applications such as building footprints extraction. The proposed methodology can generalize various shapes of urban and industrial buildings and is robust to their complexity and orientation. Some undetected buildings can be explained as they are totally covered by trees or they exhibit noisy representations of the DSM itself. In our further work, we will refine building outlines directly during the learning process including additional input data such as panchromatic or RGB images and will reorganize the network structure. The extracted building outlines then will be used for 3D model reconstruction.

REFERENCES

Brédif, M., Tournaire, O., Vallet, B. and Champion, N., 2013. Extracting polygonal building footprints from digital surface models: a fully-automatic global optimization framework. *ISPRS journal of photogrammetry and remote sensing* 77, pp. 57–65.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

d'Angelo, P. and Reinartz, P., 2011. Semiglobal matching results on the isprs stereo matching benchmark.

Davydova, K., Cui, S. and Reinartz, P., 2016. Building footprint extraction from digital surface models using neural networks. In: *SPIE Remote Sensing*, International Society for Optics and Photonics, pp. 100040J–100040J.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, pp. 248–255.

Gerke, M., Heipke, C. and Straub, B.-M., 2001. Building extraction from aerial imagery using a generic scene model and invariant geometric moments. In: *Remote Sensing and Data Fusion over Urban Areas, IEEE/ISPRS Joint Workshop 2001*, IEEE, pp. 85–89.

Huertas, A. and Nevatia, R., 1988. Detecting buildings in aerial images. *Computer Vision, Graphics, and Image Processing* 41(2), pp. 131–152.

Irvin, R. B. and McKeown, D. M., 1989. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. In: *OE/LASE'89, 15-20 Jan., Los Angeles. CA*, International Society for Optics and Photonics, pp. 156–164.

Karantzalos, K. and Paragios, N., 2009. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. *IEEE Transactions on Geoscience and Remote Sensing* 47(1), pp. 133–144.

Kim, T. and Muller, J.-P., 1999. Development of a graph-based approach for building detection. *Image and Vision Computing* 17(1), pp. 3–14.

Koc-San, D. and Turker, M., 2014. Support vector machines classification for finding building patches from ikonos imagery: the effect of additional bands. *Journal of Applied Remote Sensing* 8(1), pp. 083694–083694.

Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst* 2(3), pp. 4.

Krauß, T., Sirmacek, B., Arefi, H. and Reinartz, P., 2012. Fusing stereo and multispectral data from worldview-2 for urban modeling. In: *SPIE Defense, Security, and Sensing*, International Society for Optics and Photonics, pp. 83901X–83901X.

Krishnamachari, S. and Chellappa, R., 1996. Delineating buildings by grouping lines with mrfs. *IEEE Transactions on image processing* 5(1), pp. 164–168.

Lee, D. S., Shan, J. and Bethel, J. S., 2003. Class-guided building extraction from ikonos imagery. *Photogrammetric Engineering & Remote Sensing* 69(2), pp. 143–150.

Liow, Y.-T. and Pavlidis, T., 1990. Use of shadows for extracting buildings in aerial images. *Computer Vision, Graphics, and Image Processing* 49(2), pp. 242–277.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.

Lu, Y. H., Trunder, J. and Kubik, K., 2002. Automatic building extraction for 3d terrain reconstruction using interpretation techniques. *School of Surveying and Spatial Information Systems, University of New South Wales, NSW*.

Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* 55(2), pp. 645–657.

Marmanis, D., Adam, F., Datcu, M., Esch, T. and Stilla, U., 2015. Deep neural networks for above-ground detection in very high spatial resolution digital elevation models. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2(3), pp. 103.

McGlone, J. C. and Shufelt, J. A., 1994. Projective and object space geometry for monocular building extraction. Technical report, DTIC Document.

Peng, J. and Liu, Y., 2005. Model and context-driven building extraction in dense urban aerial images. *International Journal of Remote Sensing* 26(7), pp. 1289–1307.

Qin, R., Tian, J. and Reinartz, P., 2016. Spatiotemporal inferences for use in building detection using series of very-high-resolution space-borne stereo images. *International Journal of Remote Sensing* 37(15), pp. 3455–3476.

San, D. K. and Turker, M., 2006. Automatic building detection and delineation from high resolution space images using model-based approach. In: *Proceedings of the ISPRS Workshop on Topographic Mapping from Space*.

Shotton, J., Winn, J., Rother, C. and Criminisi, A., 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision* 81(1), pp. 2–23.

Sirmacek, B. and Unsalan, C., 2009. Urban-area and building detection using sift keypoints and graph theory. *IEEE Transactions on Geoscience and Remote Sensing* 47(4), pp. 1156–1167.

Sirmacek, B., d'Angelo, P. and Reinartz, P., 2010. Detecting complex building shapes in panchromatic satellite images for digital elevation model enhancement.

Sumer, E. and Turker, M., 2013. An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images. *Computers, Environment and Urban Systems* 39, pp. 48–62.

Yuan, J., 2016. Automatic building extraction in aerial scenes using convolutional networks. *arXiv preprint arXiv:1602.06564*.