Robust Building Extraction from Large-scale Aerial Scene with Multi-scale Fully Convolutional Network

Anonymous ACCV 2016 submission

Paper ID ***

Abstract. Automatic building extraction from remote sensing image plays a critical role in a diverse range of applications. However, it is a huge challenge to extract any-size buildings with large variational appearances and occlusions. To tack these problems, we propose a robust system employing a novel multi-scale fully convolutional networks(MSFCNs), which effectively integrates the information generated from a group of neurons with multi-scale receptive fields. In this article, three versions of architectures are deployed to adapt to different applications. All architectures can take a whole aerial images as inputs without warping or cropping and output building map directly. The experiment results test on a publicly available aerial imagery dataset proved that our proposed methodology significantly shorten time-consuming and surpass the performance of state-of-the-art.

1 Introduction

With the rapid development of remote sensing technologies and popularization of geospatial related commercial software, very high resolution satellite images are very easily accessible, these valuable data provides a huge fuel for interpreting real terrestrial scenes. Building rooftops is one of the most important type of terrestrial objects because it is essential for a wide range of technologies, such as, urban planning, automated map making, 3D city modelling, disaster assessment, military reconnaissance, etc. However, it is very costly and time-consuming for human experts to complete this task.

In recent decades, many researchers have made massive attempts to extract buildings automatically. According to our experience of life, we learn that the most of building rooftops have more regular shapes, which usually are rectangular or combinations of several rectangles. Hereby, [1][2][3][4] exploited a graph-based search to establish a set of rooftop hypotheses through examining the relationships of lines and line intersections, and then removed the fake hypotheses using a series of criteria. [5] generated rooftop outline from selected corners in multiple color and color-invariance spaces, further refined to fit the best possible boundaries through level-set curve evolution. Though these geometric primitives

based methods methods achieve good performance in high contract remote sensing imagery , they suffer from three serious shortcomings. Firstly, they lack the ability of detecting arbitrarily shaped building rooftop. Secondly, they fail to extract credible geometric features in buildings with inhomogeneous color distribution or low contrast with surroundings. Thirdly, it is hardly possible to process large-scale scenes because of their high computational-complexity.

Apart from using shape information, spectral features is a distinctive feature for terrestrial object detection. For instance, shadows are commonly dark grey or black, vegetations are usually green or yellow with particular textures, and main roads are dim gray with different road marks in most case. According to these prior knowledge, Ghaffarian *et al.* [6] split aerial scene into three components(respectively, shadows and the vegetation, roads and the bare soil, building) using a group of manually set rules. Afterwards, a purposive fast independent component analysis (PFastICA) technique is employed to separate building area from remote sensing image. However, the results is significantly sensitive to parameter choice. A feasible alternative strategy is to learn the appearance representation using supervised learning algorithm. [7][8][9][10] designed a similar framework. Firstly, aerial image is divided into superpixels using a certain over-segmentation algorithm. Secondly, hand-craft features, such as, color histograms or local binary patterns, are extracted from each over-segmented regions. Finally, each region was classified using machine learning tools and a gallery of training descriptors. **Because it's inevitable for machine learning method to mislabel regions with close appearance, additional information is utilized to refine previous result.** [10] removed false rooftops using a assumption that buildings are surrounded by shadows because of illumination. [9] devised a "histogram method" to detect missed rooftops. [11] select probable rooftops after pruning out blobs using shadows, light direction, a series of shape criteria, and then these rooftops is refined by high order conditional random field. The drawbacks are threefold. (1) It is problematic to recognize a over-segmented region as building because terrestrial objects have huge variational appearance in real aerial scene. (2) Hand-craft features are not robust to large-scale remote sensing image. (3) Additional information is unreliable. For instance, some buildings have no shadow in its neighbourhood, and some buildings have unique structures which are not satisfied to hand-coded criteria.

1.1 Related Works

As mentioned above, traditional methods are not capable of adapting to real scenes with huge variational appearance, occlusion and low contrast. Recently, deep neural networks have been widely deployed in general image segmentation or scene labelling tasks. Mnih, a pioneer, have presented a patch-based framework in [12] for learning to label aerial images. A neural network architecture is carefully designed for predicting buildings in aerial imagery, and then the output of this network is processed by conditional random fields(CRFs). Satito [13] improved Mnih's networks for extracting multiple kinds of objects simultaneously, two techniques consisting of model averaging with spatial displacement(MA)

and channel-wise inhibited softmax (CIS) are introduced to enhance the performance. However, both two methods need to crop test image into a fixed size, which not only increase the time loss, but also break the integrity of building. For example, they obtain bad performance for large-size or occluded buildings (see Fig. 4) and cost about 9s for 1500 $\times$ 1500 image without model averaging.

Mapping buildings from aerial image is essentially a problem of semantic segmentation. Recent work suggests a number of methods in processing natural images. Chen et al. [14] present a system which combine the responses at the final convulotional layer with a fully connected conditional random field (CRF). The system is able to localize segment boundaries at a quite high level of accuracy. An end-to-end network [15] which integrates CRF modelling with CNNs avoids off-line post-processing methods for object delineation. Noh et al. [16] apply a deconvolution network to each proposal in an input image, and construct the final semantic segmentation map by combining the results from all proposals in a simple manner.

Although these methods show promise in segmenting natural images, they have components not suited for building extraction. First, buildings are frequently occluded by shadows or trees (see Fig.1.1 (a)). It is challengable to delineate building boundaries even for human experts. **Though [14] [15] achieve excellent performance in processing boundary of natural image, neither of them reported that they have strong ability to handle occlusions.** Second, buildings have significantly variational appearance even in a single one. (see Fig.1.1 (b)). Moreover, a number of buildings are very close to the plot on the ground or road (see Fig.1.1 (c)). Based on our observation, there are few samples emerged in PASACAL VOC 2012 dataset. Last but not the least, the size of objects in a remote sensing image is in a wide range. For example, some images include a large number of tiny buildings (see Fig.1.1 (d)) and some ones are composed of moderate quantity of small-scale rooftops and a few of large-scale rooftops. [16] claimed that it handles objects in multiple scales, but it only suitable to multi-class object segmentation.

(a) (b) (c) (d)

Fig. 1. Examples of aerial image

In this article, we aim at deploying a robust building extraction system that can handle buildings with different sizes, variational ap-

pearances and occlusions. The rest of this article is organized as follows. The section 2.1 outlines some key ideas of fully convolutional network(FCN), the section 2.2 provides details of our neural network architecture. In section 2.3 the formulation of our system is proposed. The section 4.1 presents the datasets which is used for training and testing in our experiments. The section 3.2 introduces training settings and strategies of our proposed network. The section 3.3, we compare our results with two patch based methods using same dataset with us. In section 4, we discuss the experimental results and summarize whole article.

2 Proposed Method

In this section, we firstly introduce a well-known semantic segmentation network, called fully convolutional network(FCN)[17]. We attempt to extract footprints of building using architecture proposed by author and two modified versions. Our experiments show that these networks are not enough to accomplish building extraction task. **Inspired by [19], we introduce a multi-scale fully convolutional network(MSFCN) for extracting rooftops, it turns out to be achieve state-of-the-art performance in a public aerial image dataset.** Finally, we formulate our problem and loss function.

2.1 Fully Convolutional Network

As mentioned in introduction, patch-based building extraction methods [12] [13] suffer from shortcoming in processing big buildings because of the fixed-size inputs. As far as we know, fully convolutional network(FCN) takes whole image as inputs and directly outputs building rooftops by one pass of forward propagation. Because it can process input images of any sizes without warping or cropping, integrality of object is protected much better.

Long *et al.* put forward the concept of fully convolutional network(FCN) in [17] for the first time. **It said: "While a general deep net computes a general nonlinear function, a net with only layers of this form computes a nonlinear filter, which we call a deep filter or fully convolutional network."** In practice, the fully connected layers in traditional CNNs are transformed to convolutional layers with 1×1 kernels. Due to the mechanism of pooling, the output of the network is a coarse heat map. For pixelwise prediction, skip-net technique is used to connect various coarse outputs back to pixels. For example, the framework of FCN-8s is described as following. We firstly obtain the 16 stride predictions by fusing the predictions of pooling4 with the $2 \times$ upsampling of predictions from *conv7*(convolutionalized *fc7*). Then we continue in this fashion by fusing predictions from *pool3* with a $2 \times$ upsampling of the 16 stride predictions, denoted as 8 stride predictions. Finally, the 8 stride predictions are upsampled back to image with original size.

In our experiments, we first directly apply the FCN-8s to extract building rooftop by replacing the loss function with sigmoid cross-entropy loss. In order

to achieving better performance, the network is extended to FCN-4s, FCN-2s. Our experiments show that FCN-4s and FCN-2s get the best performance with overall recall of 70.19 % in precision recall breakeven point. According to our results, the first row of Fig. 2 indicates that FCN-4s has less ability to handle objects with small size. The second row of Fig. 2 shows that it is hard for FCN to discriminate buildings with ground if their color is similar. To overcome these drawbacks, a improved FCN is introduced in next section.

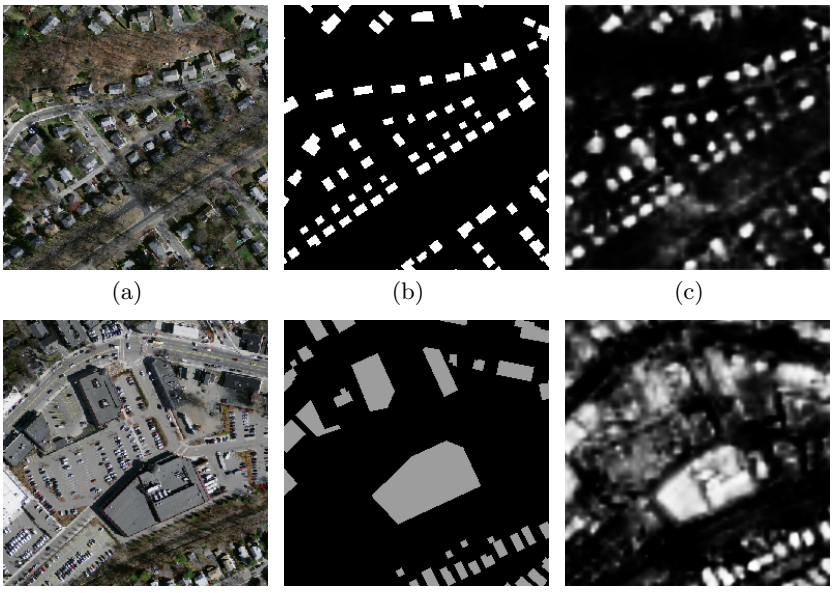

Fig. 2. (a)(d) Input image. (b)(e) Ground truth. (c)(f) FCN-4s prediction.

2.2 Our Architecture

There are three reasons that FCN-8s is not appropriate methods for building extraction. (1) **The output of 32 stride layers, including convolutionalized *fc6, fc7* and fifth pooling layer, is too fuzzy to utilize. Meanwhile, the neurons numbers of convolutionalized *fc6, fc7* are too large to cost intensive computation.** (2) The output from FCN-8s is at one-eight of the input resolution. It is not dense enough for very high resolution remote sensing image. Though FCN4s and FCN2s can increase the resolution, they have little help for improving performance. Defective fusion strategy is the most likely reason. On the one hand, in [17], the coarse predictions from a late stage are upsampled and combined with prediction from its preceding stage using simple sum operation. (3) **In ConvNet, the latter the layer is, the less the spatial information**

is and the richer the semantic information is. For instance, in FCN-8s, the firth stage generate low level semantic information, such as, regions with similar color or texture, and fine spatial resolution. On the contrary, the last stage outputs strong semantic features but coarse map. FCN-8s takes no account of combining semantic information and spatial information effectively.

Based on the above analysis, we therefore make the following modifications: (1) Convolutionalized *fc6, fc7* and fifth pooling layer are cut. (2) We design three selection modes to select part or all convolutional layers with incremental receptive fields (see Table 2.2) in VGG16 net. In first mode (denote as 5-convs), for each stage, the convolutional layer with largest receptive field is chosen, receptively, *conv1_2, conv2_2, conv3_1, conv3_3, conv4_1, conv4_3, conv5_1, conv5_3*. In second mode (denote as 7-convs), for the sake of preserving spatial information, *conv1_1* and *conv2_1* are added. In third mode (denote as 16-convs), all convolutional layer are selected. The performance of these modes will be evaluated in experiment part. These three networks are shown in Fig.3. For each selected convolutional layer, the top of which add a additional convolutional layer with 1×1 kernel, denoted as side outputs. (3) All side outputs are upsampled to the same size of input. Upsampling is implemented via deconvolution which is initialized by bilinear interpolation. Upsampled features maps are staked and fed into a convolutional layer with 1×1 kernel to yield final output.

Table 1. The receptive field and stride size in VGG16 net. Bolded layer is used in our architecture.

layer	*c1_1*	*c1_2*	*c2_1*	*c2_2*	*c3_1*	*c3_2*	*c3_3*	*c4_1*	*c4_2*	*c4_3*	*c5_1*	*c5_2*	*c5_3*
rf size	3	5	10	14	24	32	40	60	76	92	124	164	196
stride	1	1	2	2	4	4	4	8	8	8	16	16	16

2.3 Formulation

Our goal is to predict probabilistic label image $\mathbf{M}$ from an input aerial image $\mathbf{S}$. Fig **??** shows an example of $\mathbf{S}$ and $\mathbf{M}$. We directly learn a mapping from raw pixels in $\mathbf{S}$ to a true label image $\mathbf{M}$ by training the whole network. Here we formulate our approach for building extraction. We denote our input training data set by $\mathbf{I} = \{(\mathbf{S}_n, \mathbf{M}_n), n = 1, \ldots, |\mathbf{S}_n|\}$, where sample $\mathbf{S}_n = \{s_j^{(n)}, j = 1, \ldots, |\mathbf{S}_n|\}$ denotes the raw input image and $\mathbf{M}_n = \{m_j^{(n)}, j = 1, \ldots, |\mathbf{S_n}|\}$, $m_j^{(n)} \in \{0, 1\}$ denotes the corresponding ground truth binary labelling map for satellite image $\mathbf{S}_n$. Taking account of each image holistically and independently, thus, we adopt the subscript n for notational simplicity. Our goal is to have a network that learns features from which it is possible to produce building maps approaching the ground truth. In our image-to-image training, the loss function is computed over all pixels in a training image $\mathbf{S} = \{s_j, j = 1, \ldots, |\mathbf{S}|\}$ and

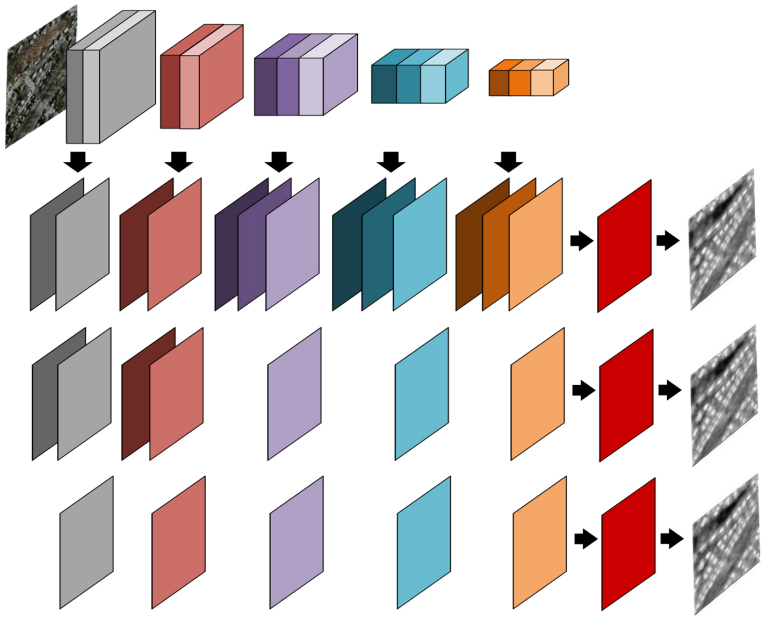

Fig. 3. Our three kinds of network architectures. The first row is five stages of VGG16 net, the second row is 16-convs, the third row is 7-convs, the fourth row is 5-convs.

building map $\mathbf{M} = \{m_j, j = 1, \ldots, |\mathbf{S}|\}$, $m_j \in \{0, 1\}$. For simplicity, we denote the collection of all standard network layer parameters as $\mathbf{W}$. For each pixel j in a training image, the possibility that assigns it to building is denoted as $\hat{m}_j = Pr(m_j = 1|\mathbf{S}; \mathbf{W})$. the definition of sigmoid cross-entropy loss function is shown in Eq (1).

$$\mathcal{L} = -\frac{1}{|\mathbf{S}|} \sum_{j \in \mathbf{S}} [m_j \log \hat{m}_j + (1 - m_j) \log (1 - \hat{m}_j)] \tag{1}$$

3 Experiments

In this section, we discuss our detailed implementation and report the performance of our proposed algorithm.

3.1 Dataset

In our experiments, we use *Massachusetts Buildings Dataset* (Mass. Buildings) proposed by Mnih [12] and publicly available on website http://www.cs.toronto.edu/ vmnih/data/. The dataset consists of 151 aerial images of the Boston area, with each of the images being 1500 × 1500 pixels for an area of 2.25 square kilometers. Hence, the entire dataset covers roughly 340 square kilometers. The

data is split into a training set of 137 images, a test set of 10 images and a validation set of 4 images. To train the network, we create image tiles for train and validation by means of cropping entire image using a sliding window with size of 256×256 pixels and stride of 64 pixels. When scanning the whole dataset, image tiles with too many white pixels are removed. **After scanning, train and validation dataset include 75938 tiles and 2500 tiles with corresponding building masks**. For testing, we use ten 1500×1500 entire images covering area excluded from the training data. **In our experiments, we find that it is benefit for prediction to scale input image into range of [0,1].**

3.2 Implementation

The implementation of our networks are based on the publicly available *Caffe* [20] Library. These three networks are fine-tuned from an initialization with the pre-trained VGG16 net model and trained in an end-to-end manner. All of our networks are trained using stochastic gradient descent with same hyper-parameters, including mini-batch size (18), fixed learning rate (10^{-5}), momentum (0.9), weight decay (5×10^{-3}), clip_gradients (10000), number of training iterations (16000, divide learning rate by 10 after 6000). We find that learned deconvolutions provide no noticeable improvements in our experiments, therefore, lr_mult is set to zero for all deconvolutional layers. It takes about six hours to train a network on a single NVIDIA Titan 12GB GPU.

3.3 Results of Three Networks

3.4 Performance Comparison

Table 2. Performance comparision with another methods. Recall here means recall at breakeven. Time is computed in a single computer.

	Recall(relax = 3)	Recall(relax = 0)	Time(s)
Mnih-CNN	0.9150		
Mnih-CNN + CRF	0.9211		
Saito-single-channel-MA	0.9426		
Saito-multi-channel-MA-CIS	0.9488		
Ours(16-convs)			

4 Conclusions

References

1. Noronha, S., Nevatia, R.: Detection and modeling of buildings from multiple aerial images. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23** (2001) 501–518

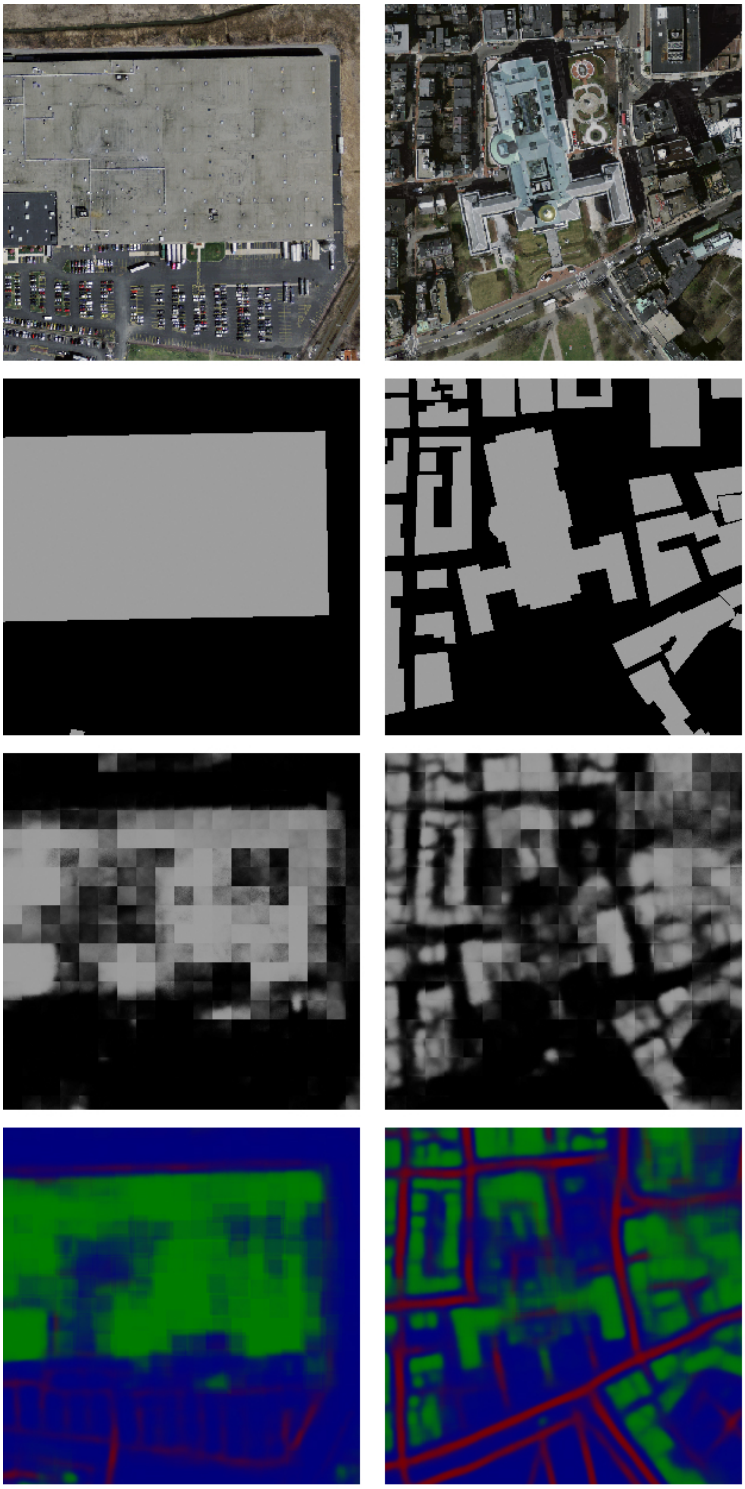

Fig. 4. (a) Input image. (b) Ground Truth. (c) Mnih's result. (d) Saito's result

2. Nosrati, M.S., Saeedi, P.: A novel approach for polygonal rooftop detection in satellite/aerial imageries. In: Image Processing (ICIP), 2009 16th IEEE International Conference on, IEEE (2009) 1709–1712
3. Izadi, M., Saeedi, P.: Three-dimensional polygonal building model estimation from single satellite images. Geoscience and Remote Sensing, IEEE Transactions on **50** (2012) 2254–2272
4. Wang, J., Yang, X., Qin, X., Ye, X., Qin, Q.: An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery. Geoscience and Remote Sensing Letters, IEEE **12** (2015) 487–491
5. Cote, M., Saeedi, P.: Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution. Geoscience and Remote Sensing, IEEE Transactions on **51** (2013) 313–328
6. Ghaffarian, S., Ghaffarian, S.: Automatic building detection based on purposive fastica (pfica) algorithm using monocular high resolution google earth images. ISPRS Journal of Photogrammetry and Remote Sensing **97** (2014) 152–159
7. Chen, D., Shang, S., Wu, C.: Shadow-based building detection and segmentation in high-resolution remote sensing image. Journal of Multimedia **9** (2014) 181–188
8. Dornaika, F., Moujahid, A., Bosaghzadeh, A., El Merabet, Y., Ruichek, Y.: Object classification using hybrid holistic descriptors: Application to building detection in aerial orthophotos. Polibits (2015) 11–17
9. Baluyan, H., Joshi, B., Al Hinai, A., Woon, W.L.: Novel approach for rooftop detection using support vector machine. ISRN Machine Vision **2013** (2013)
10. Ngo, T.T., Collet, C., Mazet, V.: (Automatic rectangular building detection from vhr aerial imagery using shadow and image segmentation)
11. Li, E., Femiani, J., Xu, S., Zhang, X., Wonka, P.: Robust rooftop extraction from visible band images using higher order crf. Geoscience and Remote Sensing, IEEE Transactions on **53** (2015) 4483–4495
12. Mnih, V.: Machine learning for aerial image labeling. Doctoral (2013)
13. Saito, S., Yamashita, Y., Aoki, Y.: Multiple object extraction from aerial imagery with convolutional neural networks. Journal of Imaging Science & Technology **60** (2016)
14. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR. (2015)
15. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V.: Conditional random fields as recurrent neural networks. (2015) 1529–1537
16. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. (2015)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Computer Science **79** (2014) 1337–1342
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Eprint Arxiv (2014)
19. Xie, S., Tu, Z.: Holistically-nested edge detection. Computer Science (2015) 1395–1403
20. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. Eprint Arxiv (2014) 675–678