# HF-FCN: Hierarchically Fused Fully Convolutional Network for Robust Building Extraction

Tongchun Zuo, Juntao Feng, Xuejin Chen

CAS Key Laboratory of Technology in Geo-spatial Information Processing and
Application System
University of Science and Technology of China

**Abstract.** Automatic building extraction from remote sensing images plays an important role in a diverse range of applications. However, it is significantly challenging to extract arbitrary-size buildings with largely variant appearances or occlusions. In this paper, we propose a robust system employing a novel hierarchically fused fully convolutional network (HF-FCN), which effectively integrates the information generated from a group of neurons with multi-scale receptive fields. Our architecture takes an aerial image as the input without warping or cropping it and directly generates the building map. The experiment results tested on a public aerial imagery dataset demonstrate that our method surpasses state-of-the-art methods in the building detection accuracy and significantly reduces the time cost.

## 1 Introduction

With the rapid development of remote sensing technologies and popularization of geospatial related commercial software, high resolution satellite images are easily accessible. These valuable data provide a huge fuel for interpreting real terrestrial scenes. The building rooftop is one of the most important types 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 to manually delineate the footprint of buildings even for human experts.

In recent decades, many researchers have made massive attempts to extract buildings automatically. Much of the past work defines criteria according to the particular characteristics of rooftop, such as, polygonal boundary [1–4], homogeneous color or texture [5], surrounding shadow [6–9], and their combinations [10, 11]. However, such approaches are weakly capable of handling real-world data because hand-coded rules or probabilistic models learned from a small set of samples are heavily dependent on data. For example, they usually assume that the building rooftop is a polygon. However, stadiums typically have circle or oval shapes. Mnih [12] proposed a patch-based convolutional neural network to extract location of objects automatically and provided a huge public dataset including large-scale aerial images and their corresponding human-labeled maps.

Based on Mnih's work, Saito *et al.* improved the extraction accuracy further by developing new cost function and model averaging techniques [13]. Though these methods achieve high performance, they still have limited ability to deal with two frequently appearing cases: (1) buildings are occluded by shadows or trees and (2) buildings possess moderately variant appearances.

Extracting buildings from aerial image is essentially a problem of semantic segmentation. Recent work suggests a number of methods in processing natural images. Long *et al.* [14] firstly proposed an effective architecture for semantic image segmentation, namely, fully convolutional network (FCN). Chen *et al.* [15] presented a system which combines the responses at the final convolutional layer with a fully connected conditional random field (CRF). The system is able to accurately segment semantic objects. Zheng *et al.* [16] introduced an end-to-end network which integrates CRF with CNNs to avoid off-line post-processing for object delineation. Noh *et al.* [17] applied a deconvolution network to each proposal in an input image, and constructed the final semantic segmentation map by combining the results from all proposals.

Although these methods show good performance in natural image segmentation, they have components not suited for building extraction in aerial image in three aspects. Firstly, each image in the PASCAL VOC 2012 dataset [18] has a handful of targets, while our source image is a complex scene, which has a number of targets with significant occlusions, variant appearances, and low contrast, as shown in Fig. 1(a)(b)(c), respectively. We directly integrate the coarse but strong semantic response into the output, instead of using CRF post-processing [12, 15]. Secondly, the image in our remote sensing dataset contains many tiny buildings, as Fig. 1(d) shows. Noh *et al.* [17] indicated that FCNs [14] have less abilities in processing small objects. Thirdly, building extraction has much higher demand in precision of structure. The output of FCNs [14] has lower resolution which sacrifices precise structures severely.

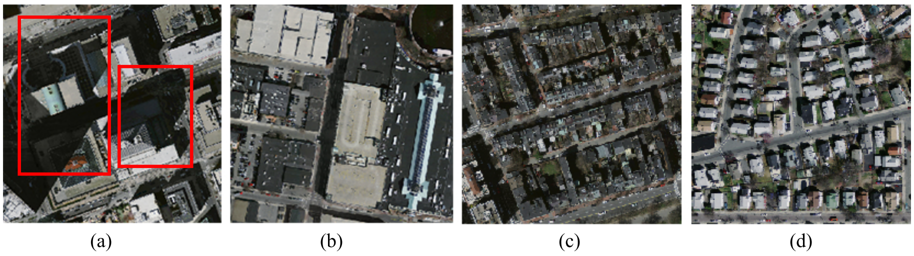


**Fig. 1.** Examples of aerial images with different type of challenges. (a) Occlusions in red boxes. (b) Variant appearances. (c) Low contrast. (d) A large number of tiny buildings.

In this paper, we present a robust building extraction system by developing a hierarchically fused fully convolutional network (HF-FCN). We trained our network on the large aerial image dataset [12]. In our architecture (HF-FCN), we design a new scheme to integrate multi-level semantic information generated from the convolutional layers with a group of increasing receptive fields, which capture context information of neighborhoods in different size. Therefore, it is more effective to handling buildings with arbitrary sizes, variant appearances or occlusions. Compared with the previous methods using convolution neural network [12, 13], our HF-FCN does not require overlapped cropping and model averaging. Taking the whole image as input, it directly outputs the segmentation map by one pass of forward propagation. Therefore, the computational complexity is reduced significantly. In conclusion, our contributions include:

1. A new architecture is developed for building extraction, which has a strong ability in processing appearance variations, varying building sizes and occlusions. The overall accuracy exceeds the state-of-the-art algorithms.
2. Our approach leads to a notable reduction of computation cost compared with previous solutions.

The rest of this article is organized as follows. In Section 2, we summarize the related work for building extraction. Section 3 provides details of our neural network architecture. Section 4 introduces the dataset and training strategies of our proposed network, and experimental results while comparing our results to two state-of-the-art methods.

## 2 Related Work

In previous studies, extracting buildings by employing their shape information is a dominant method. It is observed that rooftops have more regular shapes, which usually are rectangular or combinations of several rectangles. Several studies [1–4] exploited a graph-based search to establish a set of rooftop hypotheses through examining the relationship of lines and line intersections, and then removing the fake hypotheses using a series of manually designed criteria. Cote and Saeedi [5] generated the rooftop outline from selected corners in multiple color and color-invariance spaces, further refine the outline by the level-set curve evolution algorithm. Though these methods based on geometric primitives achieved good performance in high contract remote sensing imagery, they suffer from three 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 time-consuming to process large-scale scenes because of their high computational complexity.

Apart from using shape information, spectral information is a distinctive feature for terrestrial object extraction. For instance, shadows are commonly dark grey or black, vegetations are usually green or yellow with particular textures, and main roads are dim gray in most cases. According to these prior knowledge, Ghaffarian et al. [19] split aerial scenes into three components (respectively, shadows and the vegetation, roads and the bare soil, buildings) using a group of manually established rules. Afterwards, a purposive fast independent component analysis technique is employed to separate building area in remote sensing image. However, their results are significantly sensitive to parameter choice. A feasible alternative strategy is to learn the appearance representation using supervised learning algorithm [8–10, 20]. Firstly, an aerial image is divided into superpixels. Secondly, hand-crafted features, such as color histograms or local binary patterns, are extracted from each over-segmented region. Finally, each region is classified using machine learning tools and a gallery of training descriptors. Since it is inevitable for machine learning methods to mislabel regions with similar appearance, additional information is utilized to refine previous results. Ngo et al. [9] removed false rooftops using the assumption that buildings are surrounded by shadows because of illumination. Baluyan et al. [10] devised a "histogram method" to detect missed rooftops. Li et al. [11] selected probable rooftops after pruning out blobs using shadows, light direction, a series of shape criteria, and then these rooftops are refined by high order conditional random field. The drawbacks of these algorithms are threefold. (1) It is problematic to recognize an over-segmented region as building because terrestrial objects have hugely variant appearances in real scene. (2) Hand-craft features are less expressive to tremendous shape or appearance difference of buildings. Therefore, it is not robust to process large-scale remote sensing images. (3) Additional information is unreliable in many cases. For instance, some low buildings have no shadow in its neighborhood, and many buildings have unique structures that do not satisfy the hand-coded criteria.

As mentioned above, traditional methods are weakly capable of adapting to real scenes with huge variant appearances, occlusions or low contrast. Mnih, a pioneer, presented a patch-based framework for learning to label aerial images [12]. A neural network architecture is carefully designed for predicting buildings in aerial imagery, and the output of this network is processed by conditional random fields (CRFs). Satito *et al.* [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, these methods need to crop test image to a fixed size, which not only increases the time cost, but also breaks the integrity of buildings. Our system takes whole images as inputs without overlapped cropping or wrapping and directly outputs labelling images. It is much beneficial to preserve the whole structure of buildings and shorten computation time.

## 3 Algorithm

In this section, we introduce our hierarchically fused fully convolutional network (HF-FCN) for extracting rooftops, and the implementation in the training stage.

### 3.1 Network Architecture

Given an input aerial image $\mathbf{S}$, our goal is to predict a label image $\hat{\mathbf{M}}$ where 1 for the pixel belonging to a building and 0 otherwise. We use similar strategy with semantic segmentation. We modify the VGG16 Net [21] by hierarchically fusing the response of all layers together, as shown in Fig. 2. The VGG16 Net [21] has 16 convolutional layers and five 2-stride down-sampling layers, from which we can acquire enough multi-level information. Its network parameters pre-trained on ImageNet are helpful for initializing our network because our aerial data are essentially optical imagery. We made the following modifications to detect buildings more effectively. Firstly, the sixth and seventh fully connected layers and the fifth pooling layer in VGG16 Net are cut, because they are at 1/32 of the resolution of the input image. As a result, the interpolated prediction map will be too fuzzy to utilize. Meanwhile, the number of neurons in the sixth and seventh convolutional layers is too large to cost intensive computation. The trimmed VGG16 Net is denoted as Level 1 in our HF-FCN. Secondly, the feature map from each convolutional layer in Level 1 are fed into a convolutional layer with a filter of $1 \times 1$ kernel. The outputs of these convolutional layers are upsampled and cropped to the size of input image. Upsampling is implemented via deconvolution which is initialized by bilinear interpolation. These upsampled feature maps compose the Level 2 in our HF-FCN. Finally, all the feature maps in Level 2 are stacked and put into a convolutional layer with a filter of kernel size of $1 \times 1$ to yield final predicted map, denoted as Level 3 in our HF-FCN. The size of the feature map in last stage of Level 1 is 1/16 of input image, which is too small to use. Thus, the input images are padded with all-zero band to enlarge the size of feature maps, similar as [14].

**Table 1.** The receptive field (RF) and the stride size of Level 2 in our architecture.

| layer | F1_1 | F1_2 | F2_1 | F2_2 | F3_1 | F3_2 | F3_3 | F4_1 | F4_2 | F4_3 | F5_1 | F5_2 | F5_3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RF | 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 |

In Level 2, the feature maps with increasing receptive field (see Table 1) capture local information in larger neighbourhood sizes at higher semantic levels. The shallow layers generate feature maps with fine spatial resolution but low level semantic information. In contrast, the deep layers generate coarse feature maps with high-level semantic information. The feature maps at middle layers correspond to certain intermediate-level features. Integrating all these feature maps, buildings with variant appearances or occlusions are effectively extracted. An example is shown in Fig. 3. Given an aerial image, the U1_1 in Fig. 3(b) with small receptive field extracts low-level features like edges and corners. In Fig. 3(c), the U1_2 functions like an over-segmentation which groups pixels with similar color or texture into a subregion. In the U2_1 as Fig. 3(d) shows, shape information is augmented. From the U3_3 as Fig. 3(e) shows, we can see that

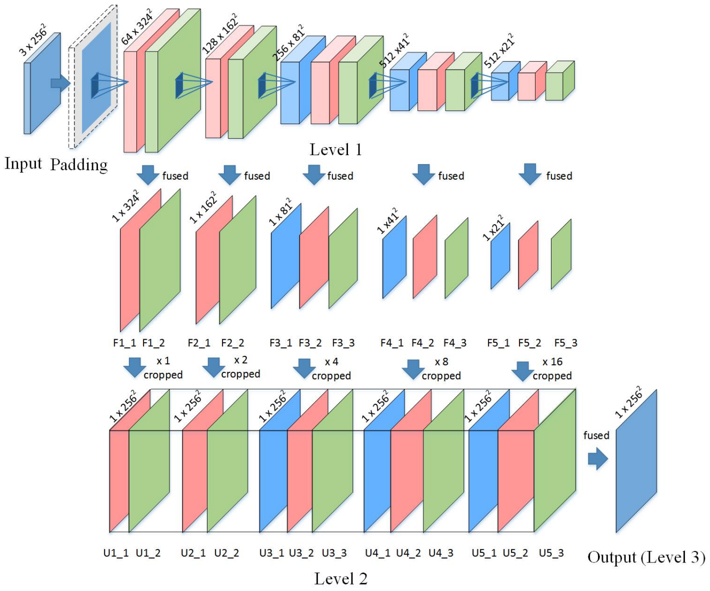


**Fig. 2.** Our network architecture. F1_1 means the fusion of feature maps generated from its corresponding convolutional layer conv1_1, U1_1 means the upsampling of F1_1, and so forth.

regions with significantly varying appearances are merged into an integrated building by considering high-level features. In U4_2 and U5_2 (see Fig. 3(f)(g)), our network learns strong semantic knowledge to distinguish dark rooftops with dim shadows and dark-green water area. In Level 3, we show that HF-FCN obtains a reliable prediction by combining multi-level semantic information and spatial information, as Fig. 3(h) shows.

### 3.2 Network Training

In the training stage, we train our network to directly generate a prediction map $\hat{\mathbf{M}}$ from raw pixels in the input aerial image $\mathbf{S}$ to approach a true label image $\tilde{\mathbf{M}}$. Fig. 4 shows an example of $\mathbf{S}$, $\tilde{\mathbf{M}}$, $\hat{\mathbf{M}}$. We denote our input training data set as $\mathbf{I} = \{(\mathbf{S}_i, \tilde{\mathbf{M}}_i), i = 1, \ldots, N\}$, $N$ is the number of aerial image and labeled map pairs.

Taking account of each input image holistically and independently, the subscript $i$ is ignored for notational simplicity in the following definition. In our image-to-image training stage, the loss function is computed over all pixels in a training image $\mathbf{S} = \{s_j, j = 1, \ldots, |\mathbf{S}|\}$ and building map $\tilde{\mathbf{M}} = \{\tilde{m}_j, j = 1, \ldots, |\mathbf{S}|\}$, $\tilde{m}_j \in \{0, 1\}$, where $|S|$ is the number of pixels in $\mathbf{S}$. For simplicity, we denote the collection of all standard network layer parameters as $\mathbf{W}$. For each pixel $j$ in a training image, the probability that assigns it to building is

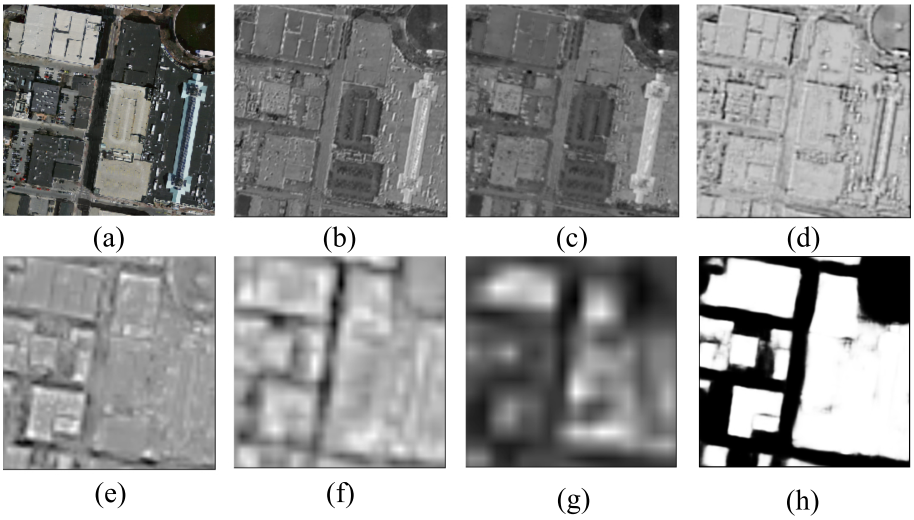
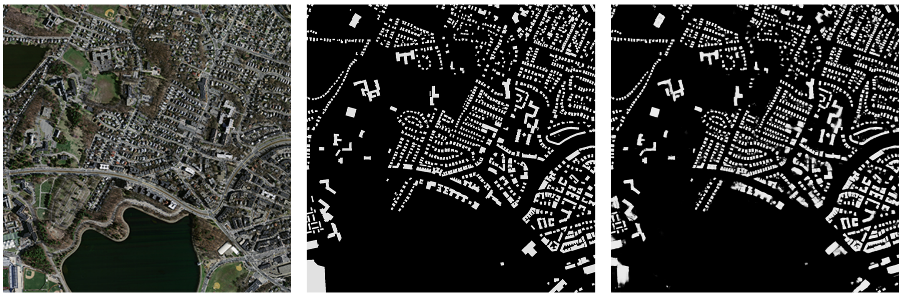


**Fig. 3.** (a) Input aerial image. (b - g) Feature maps generated from U1_1, U1_2, U2_1, U3_3, U4_2, U5_2, respectively. (h) Predicted labelling map.

denoted as its probability as a building $\hat{m}_j$. We use the sigmoid cross-entropy loss function defined as

$$\mathcal{L} = -\frac{1}{|\mathbf{S}|}\sum_{s_j \in \mathbf{S}} \left[\tilde{m}_j \log \hat{m}_j + (1-\tilde{m}_j)\log(1-\hat{m}_j)\right]. \tag{1}$$



(a) Aerial image $\mathbf{S}$ (b) Ground truth $\tilde{\mathbf{M}}$ (c) Predicted image $\hat{\mathbf{M}}$

**Fig. 4.** An example of the predicted image.

## 4 Experiments

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

### 4.1 Dataset

In our experiments, we use Massachusetts Buildings Dataset (*Mass. Buildings*) proposed by Mnih [12]. The dataset consists of 151 aerial images of the Boston area, with each image being 1500×1500 pixels for an area of 2.25 square kilometers. The entire dataset covers roughly 340 square kilometers. The intensity of each aerial image is scaled into the range of $[0, 1]$. 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 a set of image tiles for training and validation by cropping each aerial image using a sliding window with size of 256×256 pixels and stride of 64 pixels. After scanning, the training and validation datasets include 75938 tiles and 2500 tiles respectively, with their corresponding building masks. For testing, we use ten 1500×1500 images excluded from the training images.

### 4.2 Training Settings

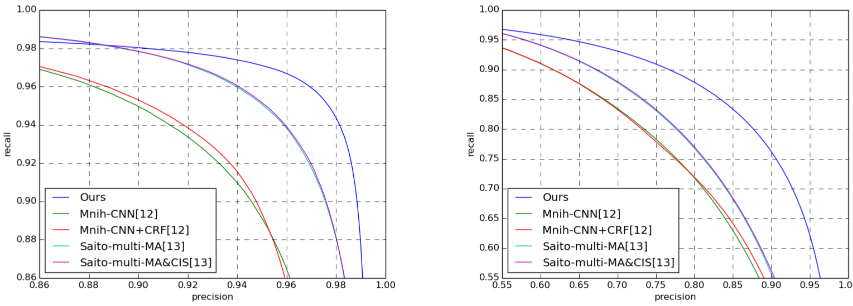
The implementation of our network is based on the *Caffe* Library [22]. Our HF-FCN is fine-tuned from an initialization with the pre-trained VGG16 Net model and trained in an end-to-end manner. It is trained using the stochastic gradient descent algorithm, with the hyper-parameters listed in Table 2. The learning rate is divided by 10 for each 8000 iterations. We find that the learned deconvolutions provide no noticeable improvements in our experiments, similar as [14, 23]. Therefore, lr_mult is set to zero for all deconvolutional layers. Except that the pad of first convolutional layer is set to 35, the others are set to 1, same as VGG16 Net. It takes about six hours to train our network on a single NVIDIA Titan 12GB GPU.

**Table 2.** Parameters for network training.

| | |
|---:|:---:|
| mini-batch size | 18 |
| initial learning rate | $10^{-5}$ |
| momentum | 0.9 |
| weight decay | 0.02 |
| clip_gradients | 10000 |
| the number of training iterations | 16000 |

### 4.3 Results

To show the effectiveness of HF-FCN, we compare our method with two state-of-the-art approaches [12, 13]. Three common metrics are used to evaluate the performance of our algorithm: (1) the relaxed precision and recall scores with $\rho = 3$; (2) the standard precision and recall scores ($\rho = 0$); (3) the time cost. The relaxed precision is defined as the fraction of detected pixels that are within $\rho$ pixels of a true pixel, while the relaxed recall is defined as the fraction of the true pixels that are within $\rho$ pixels of a detected pixel. The slack parameter $\rho$ is set to 3, which is the same value as used in [12, 13]. The relaxed precision-recall curves generated from different methods are shown in Fig. 5(a). As can be seen, all curves of ours are located above others in building prediction obviously. More strictly, we set slack parameter $\rho$ as 0, that is to say, it becomes a standard precision and recall scores. The precision-recall curves generated from different methods are shown in Fig. 5(b). We can see that our approach is more appropriate for detecting rooftops in complex scene, which significantly outperforms [12, 13]. To compare the system efficiency, we calculate the average time of processing ten test images in the same computer using different methods. Table 3 shows that our method is able to not only significantly improve the performance, but also dramatically reduces the time cost.



(a) Slack parameter $\rho = 3$ (b) Slack parameter $\rho = 0$

**Fig. 5.** The relaxed precision-recall curves from different methods with two slack parameters.

To prove that our network has strong ability in extracting buildings with variant appearances, arbitrary sizes, occlusions, a new experiment is designed. We crop seven 256×256 image patches that have buildings with variant appearances or occlusions from the test images. Corresponding predictions are directly cropped from the predicted images generated by three approaches, including Mnih-CNN+CRF [12], Saito-multi-MA&CIS [13] and ours. We binarize the probability map using a threshold of 0.5. A series of examples is shown in Fig. 6. In addition, Table 4 shows the resulting recalls at the breakeven points of

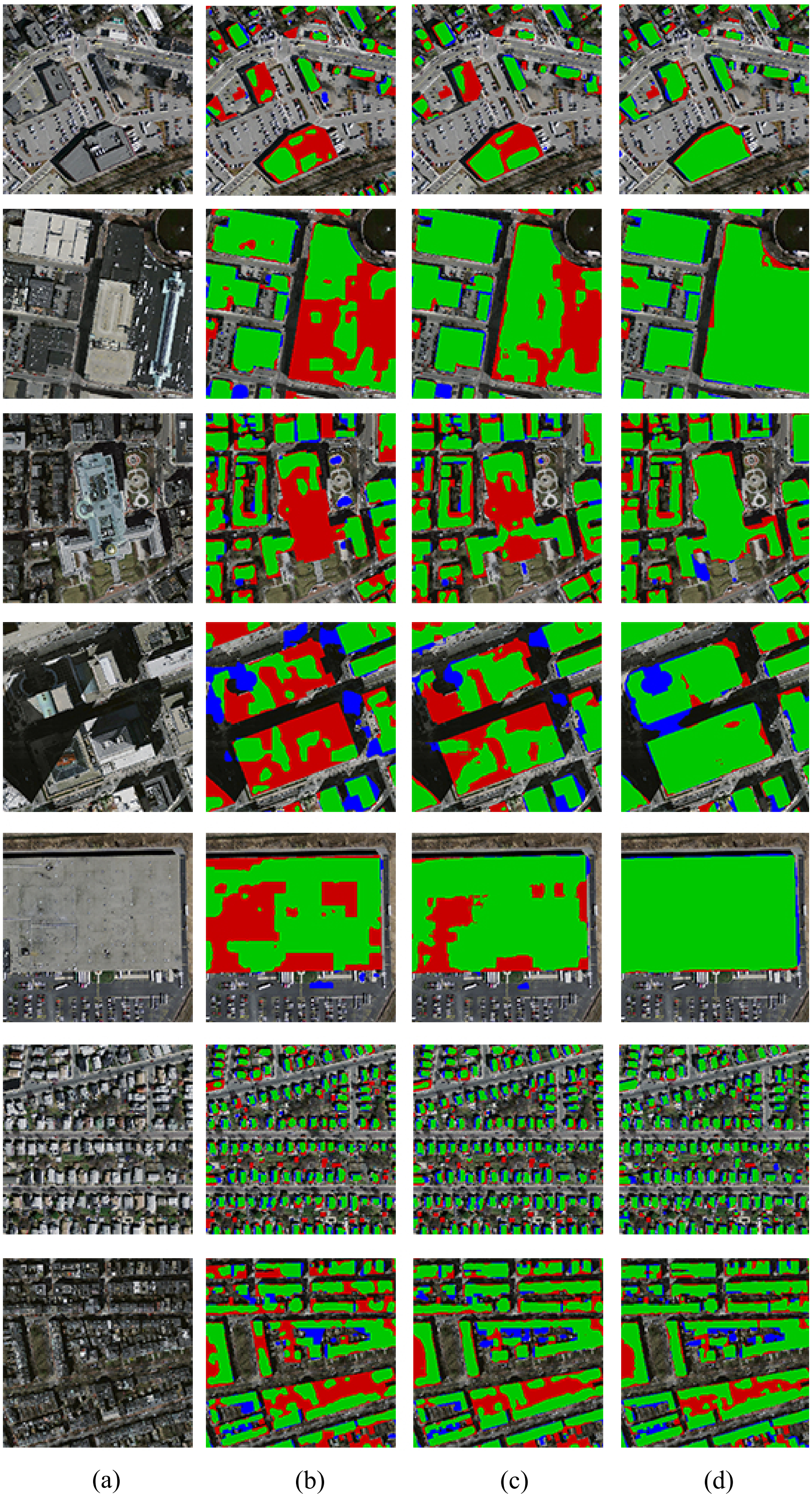


**Fig. 6.** (a) Input images. (b) Results of Mnih-CNN+CRF [12]. (c) Results of Saito-multi-MA&CIS [13]. (d) Our results. Correct results (TP) are shown in green, false positives (FP) are shown in blue, and false negatives (FN) are shown in red.

**Table 3.** Performance comparison with [12, 13]. Recall here means recall at breakeven points. Time is computed in the same computer with a single NVIDIA Titan 12GB GPU.

| | Recall ($\rho = 3$) | Recall ($\rho = 0$) | Time (s) |
|---|---|---|---|
| Mnih-CNN [12] | 0.9271 | 0.7661 | 8.70 |
| Mnih-CNN+CRF [12] | 0.9282 | 0.7638 | 26.60 |
| Saito-multi-MA [13] | 0.9503 | 0.7873 | 67.72 |
| Saito-multi-MA&CIS [13] | 0.9509 | 0.7872 | 67.84 |
| Ours (HF-FCN) | **0.9643** | **0.8424** | **1.07** |

the standard precision recall curve for each patch. The accuracy of our approach is 12.7% , 6.0% higher than [12],[13], receptively.

**Table 4.** Recall at the selected regions of the test images.

| Image ID | 01 | 02 | 03 | 04 | 05 | 06 | 07 | mean |
|---|---|---|---|---|---|---|---|---|
| Mnih-CNN+CRF [12] | 0.784 | 0.869 | 0.769 | 0.653 | 0.893 | 0.764 | 0.800 | 0.784 |
| Saito-multi-MA&CIS [13] | 0.773 | 0.915 | 0.857 | 0.789 | 0.945 | 0.773 | 0.830 | 0.851 |
| Ours (HF-FCN) | **0.874** | **0.964** | **0.899** | **0.901** | **0.986** | **0.840** | **0.851** | **0.911** |

## 5 Conclusions

In this article, we propose a novel fully convolutional network which is strongly capable of extracting buildings of arbitrary sizes, variant appearances or occlusions without any post-processing. Meanwhile, it further improves the overall accuracy. The proposed network can take arbitrary-size image as the input as long as the GPU memory allows. Compared with patch-based methods, there is no need to label a whole image by cropping the image into small patches. As consequence, inconsistent border caused by cropping would not occur in our system. Moreover, the time cost is tremendously reduced using our HF-FCN. The proposed method is demonstrated robust to various types of aerial scenes selected from real-world data. Furthermore, our architecture can be easily extended to extract multi-objects in remote sensing imagery. Consequently, we believe that our technique potentially provides a generic solution to understand complex aerial scenes.

**Acknowledgement**. We would like to thank the anonymous reviewers. This work was supported by the National Natural Science Foundation of China (NSFC) under Nos. 61472377 and 61331017, and the Fundamental Research Funds for the Central Universities under No.WK2100060011.

## References

1. Noronha, S., Nevatia, R.: Detection and modeling of buildings from multiple aerial images. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 501–518
2. Nosrati, M.S., Saeedi, P.: A novel approach for polygonal rooftop detection in satellite/aerial imageries. In: 2009 16th IEEE International Conference on Image Processing (ICIP). (2009) 1709–1712
3. Izadi, M., Saeedi, P.: Three-dimensional polygonal building model estimation from single satellite images. IEEE Transactions on Geoscience and Remote Sensing **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. IEEE Geoscience and Remote Sensing Letters **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. IEEE Transactions on Geoscience and Remote Sensing **51** (2013) 313–328
6. Sirmacek, B., Unsalan, C.: Building detection from aerial images using invariant color features and shadow information. In: Computer and Information Sciences, 2008. ISCIS '08. 23rd International Symposium on. (2008) 1–5
7. Manno-Kovcs, A., Ok, A.O.: Building detection from monocular vhr images by integrated urban area knowledge. IEEE Geoscience and Remote Sensing Letters **12** (2015) 2140–2144
8. 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
9. Ngo, T.T., Collet, C., Mazet, V.: Automatic rectangular building detection from vhr aerial imagery using shadow and image segmentation. In: Image Processing (ICIP), 2015 IEEE International Conference on. (2015) 1483–1487
10. Baluyan, H., Joshi, B., Al Hinai, A., Woon, W.L.: Novel approach for rooftop detection using support vector machine. ISRN Machine Vision **2013** (2013)
11. Li, E., Femiani, J., Xu, S., Zhang, X., Wonka, P.: Robust rooftop extraction from visible band images using higher order crf. IEEE Transactions on Geoscience and Remote Sensing **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. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 3431–3440
15. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015) 1529–1537
16. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015) 1529–1537
17. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015) 1520–1528

18. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. (http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html)
19. 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
20. 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
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2015)
22. 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
23. Xie, S., Tu, Z.: Holistically-nested edge detection. In: The IEEE International Conference on Computer Vision (ICCV). (2015)