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

Anonymous ACCV 2016 submission

Paper ID 663

Abstract. Currently, 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 large variant appearances or occlusions. To tackle these problems, 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 can take whole aerial images as inputs without warping or cropping and output building map directly. The experiment results tested on a publicly available aerial imagery dataset prove that our proposed methodology surpass the performance of state-of-the-art method and significantly reduce time cost.

1 Introduction

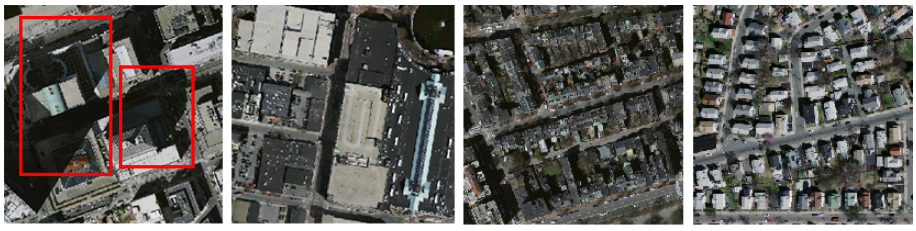
With the rapid development of remote sensing technologies and popularization of geospatial related commercial softwares, very high resolution satellite images are easily accessible. These valuable data provide a huge fuel for interpreting real terrestrial scenes. The building rooftop is one of the most important 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 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 works define 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 probability models learned from small samples are much dependent on data. For example, they assume that profile of buildings is polygon while the shape of stadiums always is circle or oval. For the sake of deploying a practical building extraction system, Mnih [12] created a huge publicly dataset including large-scale aerial images and corresponding human-labeled maps, and proposed a patch-based convolutional neural network to extract location of objects automatically. Based on Mnih's work, Saito *et al.*

[13] improved the performance further by developing two effective techniques. Though these methods achieve high performance, they still have limited ability to deal with two often appearing cases: (1) Buildings are occluded by shadows or trees. (2) Buildings possess moderately variant appearances.

Mapping 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 a 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 localize segment boundaries at a quite high level of accuracy. 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 a 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. Firstly, buildings are frequently occluded by shadows or trees (see Fig.1 (a)). It is challengable to delineate building boundaries even for human experts. Though some literatures [15, 16] achieve excellent performance in processing boundary of natural image, neither of them reported that they have strong ability in handling occlusions. Secondly, buildings have significantly variant appearances even in a single one. (see Fig.1 (b)). Moreover, a number of buildings are very close to the plot on the ground or road (see Fig.1 (c)). Based on our observation, there are few such samples emerge in PASACAL VOC 2012 [18] 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 (d)) and some ones are composed of moderate quantity of small-scale rooftops and a few large-scale rooftops. On account of low resolution (one-eighth resolution of input image) of output from [14], precise structures are sacrificed severely. Noh *et al.* [17] claimed it handles objects in multiple scales, but it only suitable to multi-classes object segmentation.



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

Fig. 1. Examples of aerial image. (a) Occlusions (shown in red rectangle). (b) Variant appearances. (c) Low contrast. (d) A large number of tiny buildings.

Here, we present a robust building extraction system by developing a hierarchically fused fully convolutional network (HF-FCN) trained on a publicly available large aerial imagery dataset [12]. In our architecture (HF-FCN), we design a new scheme to integrate multi-level semantic information generated from convolutional layers with a group of incremental receptive fields. Incremental sized receptive fields are able to capture context information in different neighbourhood sizes. Therefore, it is more effective to handling buildings with arbitrary sizes, variant appearances or occlusions. Compared with [12, 13], overlapped cropping and model averaging are not required for HF-FCN. It takes whole images as inputs, and directly outputs segmentation maps by one pass of forward propagation. Hence, our system decreases the computation complexity significantly. In conclusion, our contributions include two aspects: (1) A new architecture is developed for building extraction, which has a strong ability in processing appearance variations, varying sizes and occlusions. Meanwhile, the overall accuracy also exceeds state-of-the-art [13]. (2) Our approach leads to a notable reduction of computation cost compared with traditional solutions.

The rest of this article is organized as follows. In Section 2, we summarize main methods for building extraction. Section 3 provides details of our neural network architecture and formulation of building extraction problem. Section 4 introduces the dataset and training strategies of our proposed network, and then we compare our results with two patch-based methods using the three types of criteria. In Section 5, we discuss the experimental results and summarize whole article.

2 Related Work

In previous literatures, one popular way of extracting buildings is employing their shape information. 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 relationships of lines and line intersections, and then removed the fake hypotheses using a series of criteria. Cote and Saeedi [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 achieve 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 (PFastICA) technique is employed to separate building area from 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. A number of authors [8–10, 20] designed a similar framework. Firstly, an aerial image is divided into superpixels using a certain over-segmentation algorithm. Secondly, hand-crafted features, such as, color histograms or local binary patterns (LBP), are extracted from each over-segmented regions. Finally, each region is classified using machine learning tools and a gallery of training descriptors. Because it's inevitable for machine learning method to mislabel regions with similar appearance, additional information is utilized to refine previous results. Ngo *et al.* [9] removed false rooftops using a 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 is refined by high order conditional random field. The drawbacks of these algorithms are threefold. (1) It is problematic to recognize a over-segmented region as part of buildings because terrestrial objects have huge variant appearances in real aerial scene. (2) Hand-craft features are sensitive to input data, therefore, it is not robust to process large-scale remote sensing images. (3) Additional information is unreliable. For instance, some low buildings have no shadow in its neighbourhood, and some buildings have unique structures which are not satisfied to hand-coded criteria.

As mentioned above, traditional methods are weakly capable of adapting to real scenes with huge variant appearances, occlusions or low contrast. Our method does not design image features manually, on the contrary, building features are directly learned from a mass of real data using deep neural networks. Therefore, our algorithm is more robust to extract buildings in real scenes. 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 into 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 System Overview

In this section, we introduce a hierarchically fused fully convolutional network (HF-FCN) for extracting rooftops, and then formulate our problem and loss function.

3.1 Network Architecture

We design our network based on VGG16 Net [21] and make some modifications. The reasons for choose VGG16 Net are two-fold: (1) It has great depth (16 convolutional layers), and multiple stages (five 2-stride down-sampling layers). We can acquire enough multi-level information from different stages and convolutional layers. (2) Network parameters pre-trained on very large image dataset (ImageNet) are helpful for initializing our network because our aerial data is essentially optical imagery. The modifications are listed as follows: (1) Two fully connected layer *fc6, fc7* and fifth pooling layer are cut, because they are at 1/32 of input resolution with the consequence that the interpolated prediction map will be too fuzzy to utilize. Meanwhile, the number of neurons in *fc6, fc7* is too large to cost intensive computation. (2) Feature maps from each convolutional layer in trimmed VGG16 Net (denote as Level 1) are fed into a convolutional layer with a filter of kernel size of 1×1. The outputs of these convolutional layers are upsampled and cropped to the same size of input image (denote as Level 2). Upsampling is implemented via deconvolution which is initialized by bilinear interpolation. Finally, all the feature maps in Level 2 are stacked and put into a convolutional layer with a filter of kernel size of 1×1 to yield final predicted map (denote as Level 3). (3) The size of feature map in last stage of Level 1 is 1/16 of input image, it is too small to use. Thus, we apply a popular trick that input images are padded with all-zero band to enlarge the size of feature maps. Our architecture is shown in Fig. 2.

Table 1. The receptive field (rf) and 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 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

In Level 2 of our architecture, feature maps with increasing receptive field (see Table 1) capture local information in different neighbourhood sizes and at different semantic levels. Therefore, if we integrate all these information together, it is helpful for extracting buildings with variant appearances or occlusions. We take a concrete instance to show how HF-FCN works for such cases. In this case, U1_1 with small receptive field generates fine spatial resolution and responds to low level features like edges and corners (see Fig. 3(b)). U1_2 functions like over-segmentation algorithm to grouping pixels with similar color or texture into a subregion (see Fig. 3(c)). In U2_1, color information is disappear, shape

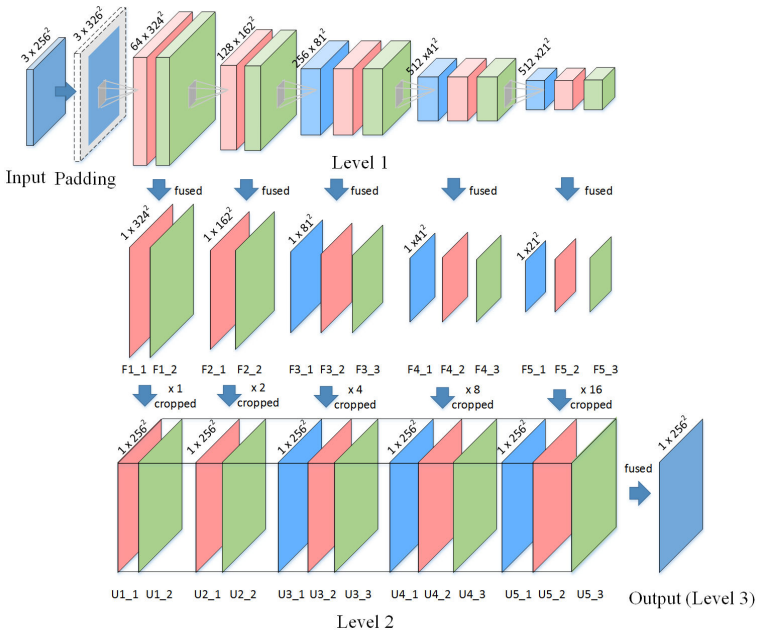


Fig. 2. Our network architecture.

information is augmented (see Fig. 3(d)). In U3_3, it is surprised that regions with significantly varying appearance are merged into a integrated building by considering some unknown high level features (see Fig. 3(e)). In U4_2 and U5_2, our network learned strong semantic knowledge to distinguish dark rooftops with dim shadows and dark-green water (see Fig. 3(f)(g)). In Level 3, we show that HF-FCN obtains reliable prediction by combining multi-level semantic information and spatial information (see Fig. 3(h)).

3.2 Formulation

Our goal is to predict labelling image $\hat{\mathbf{M}}$ from a input aerial image $\mathbf{S}$. We directly learn a mapping from raw pixels in $\mathbf{S}$ to a true label image $\tilde{\mathbf{M}}$ by training the whole network. Fig. 4 shows an example of $\mathbf{S}$, $\hat{\mathbf{M}}$, $\tilde{\mathbf{M}}$. Here we formulate our approach for building extraction. We denote our input training data set by $\mathbf{I} = \{(\mathbf{S}_n, \tilde{\mathbf{M}}_n), n = 1, \ldots, |\mathbf{S}_n|\}$, where sample $\mathbf{S}_n = \{s_j^{(n)}, j = 1, \ldots, |\mathbf{S}_n|\}$ denotes the raw input image and $\tilde{\mathbf{M}}_n = \{\tilde{m}_j^{(n)}, j = 1, \ldots, |\mathbf{S_n}|\}$, $\tilde{m}_j^{(n)} \in \{0, 1\}$ 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

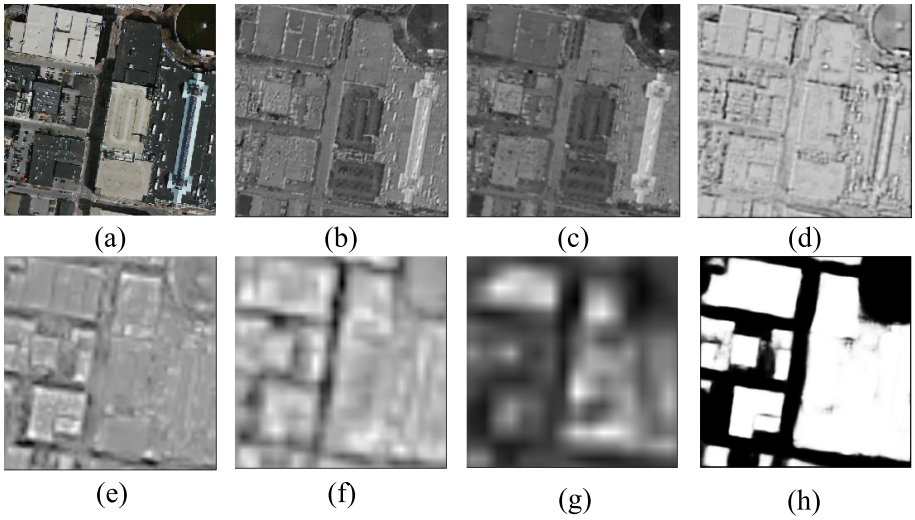


Fig. 3. (a) Input aerial image. (b - g) Feature maps generated from U1_1, U1_2, U2_1, U3_3, U4_2, U5_2, respectively. (h) Predicted labelling map.

over all pixels in a training image $\mathbf{S} = \{s_j, j = 1, \ldots, |\mathbf{S}|\}$ and building map $\tilde{\mathbf{M}} = \{\tilde{m}_j, j = 1, \ldots, |\mathbf{S}|\}$, $\tilde{m}_j \in \{0, 1\}$. 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 $\hat{m}_j = Pr(\tilde{m}_j = 1|\mathbf{S}; \mathbf{W})$. the definition of sigmoid cross-entropy loss function is shown in Eq (1).

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

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

Fig. 4. An example of the resulting 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] 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 which include more than 160 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 to improving prediction performance by means of scaling the intensity of input image into range of [0,1].

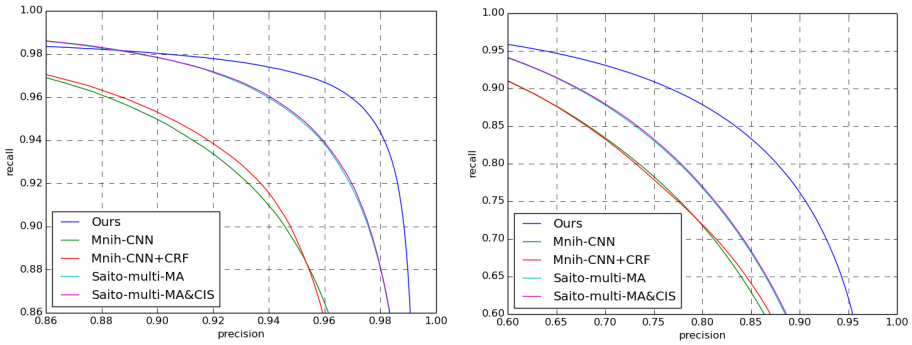
4.2 Training Settings

The implementation of our network is based on the publicly available *Caffe* [22] Library. 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 stochastic gradient descent with the following hyper-parameters, including mini-batch size (18), initial learning rate (10^{-5}), learning rate is divided by 10 for each 5000 iterations, momentum (0.9), weight decay (0.02), clip_gradients (10000), number of training iterations (12000). We find that learned deconvolutions provide no noticeable improvements in our experiments, therefore, lr_mult is set to zero for all deconvolutional layers. Additionally, except that the pad of first convolutional layer is set to 35, others are set to 1 as the same as VGG16 Net. It takes about six hours to train a network on a single NVIDIA Titan 12GB GPU.

4.3 Results

To show the effectiveness of HF-FCN, we train and test our network on *Mass. Buildings*. In order to comparing our results with previous works [12, 13], we use three metrics to evaluate our results: (1) relaxed precision and recall scores ($\rho = 3$). (2) relaxed precision and recall scores ($\rho = 0$). (3) time cost. The relaxed precision is defined as the fraction of detected pixels that are within ρ pixels of a detected pixel, while the relaxed recall is defined as the fraction of true

pixels that are within ρ pixels of a detected pixel. In one of our experiments, the slack parameter ρ is set to 3, which is the same value as used in [12, 13]. Compared relaxed precision-recall curves are shown in Fig. 5(a). In order to evaluate our results more strictly, we set slack parameter ρ as 0, that is to say, it becomes a standard precision and recall scores. Compared standard precision-recall curves are shown in Fig. 5(b). Additionally, time cost is another important index to evaluate system performance. We calculate the mean time of processing ten test images in the same computer using the same program. Table 2 shows that our method is able to not only significantly improve the performance, but dramatically decrease the time cost.



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

Fig. 5. Two relaxed precision-recall curves

Table 2. Performance is compared 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**

To prove our network having strong ability in extracting buildings with variant appearances, arbitrary sizes, occlusions, we perform further evaluation. We crop seven 256×256 image patches that have buildings with variant appearances or occlusions from test image of *Mass. Buildings*. And then, we directly crop corresponding predictions from predicted images generated by three models

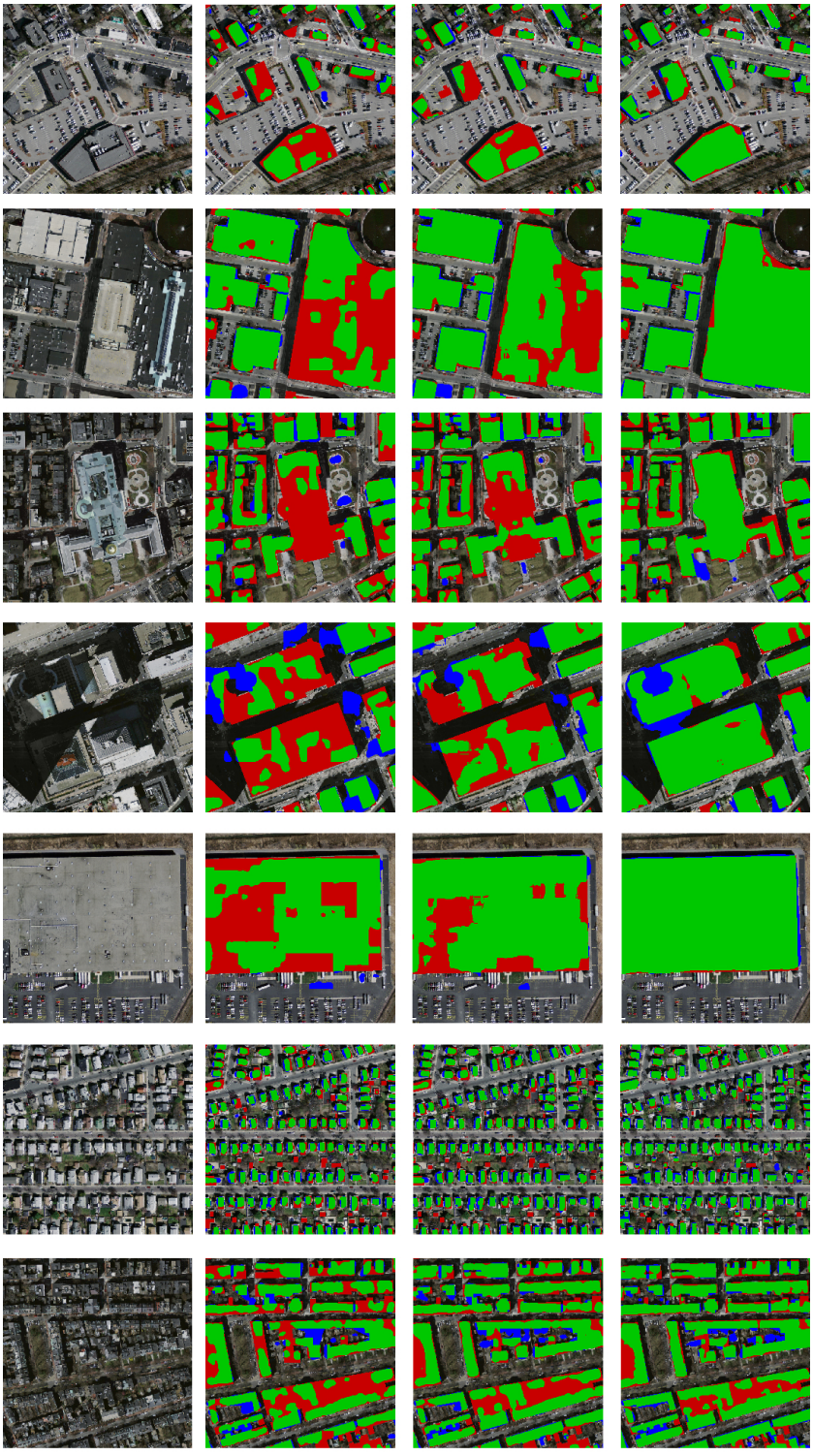


Fig. 6. (a) Input images. (b) Results of Mnih-CNN+CRF [12]. (c) Results of Saito-multi-MA&CIS [13]. (d) Our results. Correct results (TP) are shown in green, false positives (FP) are shown in blue, and false negatives (FN) are shown in red.

(Mnih-CNN+CRF [12], Saito-multi-MA&CIS [13] and ours). Here, we binarize the probability map using a threshold of 0.5. Seven groups of example are shown in Fig. 6. In addition, Table 3 shows the resulting recalls at breakeven points of standard precision recall curve for each patches.

Table 3. Recall at selected region 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 improved fully convolutional network which is strongly capable of extracting buildings with arbitrary sizes, variant appearances or occlusions without any post-processing. Meanwhile, it further improves the overall accuracy. The network can take arbitrary-size image as input as long as GPU memory allowed. Compared with patch-based methods, there is no need to label a whole image by cropping the image into small patches. As consequence, inconsistant border caused by cropped would not occurred in our system. Though some effective techniques, such as model average and conditional random field, can improve the performance further, time cost is increased by several times. While in our system, elapsed time is tremendously decreased. On the other hand, we demonstrate that our network is generally adapt 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.

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
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. Sirmacek, B., Unsalan, C.: Building detection from aerial images using invariant color features and shadow information. In: International Symposium on Computer and Information Sciences. (2008) 1–5
7. Ok, A.O., Senaras, C., Yuksel, B.: Automated detection of arbitrarily shaped buildings in complex environments from monocular vhr optical satellite imagery. Geoscience and Remote Sensing, IEEE Transactions on **51** (2013) 1701–1717
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)
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. 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. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Computer Science **79** (2014) 1337–1342
15. 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)
16. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V.: Conditional random fields as recurrent neural networks. (2015) 1529–1537
17. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. (2015)
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. Eprint Arxiv (2014)
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