

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044

HF-FCN: Hierarchical Fusion Fully Convolutional Network for Robust Building Extraction

Anonymous ACCV 2016 submission

Paper ID 97

Abstract. Currently, automatic building extraction from remote sensing images plays a critical role in a diverse range of applications. However, it is significantly challenging to extract arbitrary-size buildings with large variational appearances or occlusions. To tackle these problems, we propose a robust system employing a novel hierarchical fusion 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 proved that our proposed methodology significantly reduce time cost 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 softwares, very high resolution satellite images are 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 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 shadows [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.*

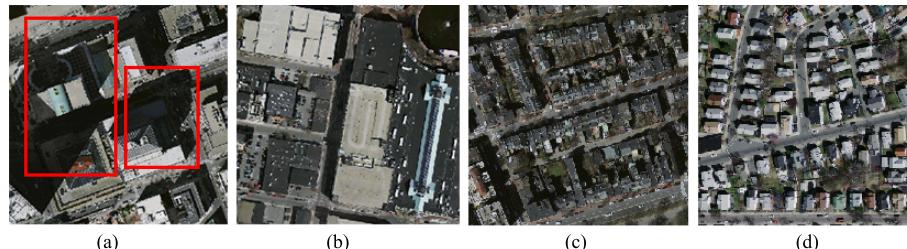
CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

2 ACCV-16 submission ID 97

[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 variational 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.* [16] firstly proposed a effective architecture for semantic image segmentation, namely, fully convolutional network (FCN). Chen *et al.* [17] 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.* [18] introduced an end-to-end network which integrates CRF models with CNNs to avoid off-line post-processing for object delineation. Noh *et al.* [19] 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[17, 18] 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 variational appearance 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 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 (eight resolution of input image) of output from [16], precise structures are sacrificed severely. Noh *et al.* [19] claimed it handles objects in multiple scales, but it only suitable to multi-classes object segmentation.

**Fig. 1.** Examples of aerial image

090 Here, we present a robust building extraction system by developing a hi-
091 erarchical fusion fully convolutional network (HF-FCN) trained on a publicly
092 available large aerial imagery dataset [12]. In our architecture (HF-FCN), we
093 design a new scheme to integrate multi-level semantic information generated
094 from convolutional layers with a group of incremental receptive fields. Incre-
095 mental sized receptive fields are able to capture context information in different
096 neighbourhood sizes. Therefore, it is more effective to handling buildings with
097 arbitrary sizes, variational appearances or occlusions. Compared with [12, 13],
098 overlapped cropping and model averaging are not required for HF-FCN. It takes
099 whole images as inputs, and directly outputs segmentation maps by one pass of
100 forward propagation. Hence, our system decreases the computation complexity
101 significantly. In conclusion, our contributions include two aspects: (1) A new
102 architecture is developed for building extraction, which has a strong ability in
103 processing appearance variations, varying sizes and occlusions. Meanwhile, the
104 overall accuracy also exceeds state-of-the-art [13]. (2) Our approach leads to a
105 notable reduction of computation cost compared with traditional solutions.
106

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

114 2 Related Works 115

116 In previous literatures, one popular way of extracting buildings is employing their
117 shape information. It is observed that rooftops have more regular shapes, which
118 usually are rectangular or combinations of several rectangles. Several studies [1–
119 4] exploited a graph-based search to establish a set of rooftop hypotheses through
120 examining the relationships of lines and line intersections, and then removed the
121 fake hypotheses using a series of criteria. Cote and Saeedi [5] generated rooftop
122 outline from selected corners in multiple color and color-invariance spaces, fur-
123 ther refined to fit the best possible boundaries through level-set curve evolution.
124 Though these geometric primitives based methods achieve good performance
125 in high contrast remote sensing imagery, they suffer from three shortcomings.
126 Firstly, they lack the ability of detecting arbitrarily shaped building rooftop.
127 Secondly, they fail to extract credible geometric features in buildings with in-
128 homogeneous color distribution or low contrast with surroundings. Thirdly, it is
129 time-consuming to process large-scale scenes because of their high computational
130 complexity.

131 Apart from using shape information, spectral information is a distinctive
132 feature for terrestrial object extraction. For instance, shadows are commonly
133 dark grey or black, vegetations are usually green or yellow with particular tex-
134 tures, and main roads are dim gray in most cases. According to these prior

CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

4 ACCV-16 submission ID 97

knowledge, Ghaffarian *et al.* [14] 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, 15] 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 variational 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 variational 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, have 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.

180 3 System Overview 180
181

182 In this section, we introduce a hierarchical fusion fully convolutional network
183 (HF-FCN) for extracting rooftops, and then formulate our problem and loss
184 function.

185

186 3.1 Network Architecture 186
187

188 We design our network based on VGG16 Net [20] and make some modification-
189 s. The reasons for choose VGG16 Net are two-fold: (1) It has great depth (16
190 convolutional layers), and multiple stages (five 2-stride down-sampling layers).
191 We can acquire enough multi-level information from different stages and convolu-
192 tional layers. (2) Network parameters pre-trained on very large image dataset
193 (ImageNet) are helpful for initializing our network because our aerial data is
194 essentially optical imagery. The modifications are listed as follows: (1) Two fully
195 connected layer $fc6$, $fc7$ and fifth pooling layer are cut, because they are at $\frac{1}{32}$
196 of input resolution. Meanwhile, the number of neurons in $fc6$, $fc7$ is too large
197 to cost intensive computation. (2) Feature maps from each convolutional layer
198 in trimmed VGG16 Net (denote as level 1) are fed into a convolutional layer
199 with a filter of 1×1 kernel and 1 neuron. The outputs of these convolutional
200 layers are upsampled and cropped to the same size of input image (denote as
201 level 2). Upsampling is implemented via deconvolution which is initialized by
202 bilinear interpolation. Finally, all the feature maps in level 2 are stacked and
203 put into a convolutional layer with a filter of 1×1 kernel and 1 neuron to yield
204 final predicted map (also denote as level 3). (3) The feature map size of the last
205 stage in level 1 is one-sixteen of input image, it is too small to use. Thus, we
206 apply a popular trick that input images are padded with all-zero band to enlarge
207 the size of feature maps. Our architecture is shown in Fig. 2.

208

209

210 **Table 1.** The receptive field and stride size in level 2 in our architecture.
211

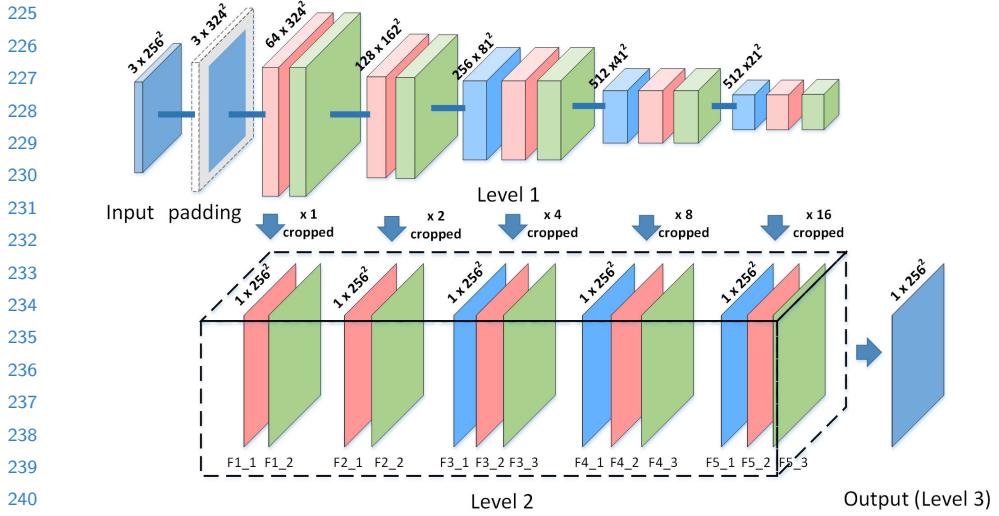
layer	F1_1	F1_2	F2_1	F2_2	F3_1	F3_2	F3_3	F4_1	F4_2	F4_3	F5_1	F5_2	F5_3
rf size	3	5	10	14	24	32	40	60	76	92	124	164	196
stride	1	1	2	2	4	4	4	8	8	8	16	16	16

212

213 In level 2 of our architecture, feature maps with increasing receptive field
214 (see Table 1) capture local information in different neighbourhood sizes and at
215 different semantic levels. Therefore, if we integrate all these information together,
216 it is helpful for extracting buildings with variational appearances or occlusions.
217 We take a concrete instance to show how HF-FCN works for such cases. In this
218 case, F1_1 with small receptive field generates fine spatial resolution and responds
219 to low level features like edges and corners (see Fig. 3(b)). F1_2 functions like
220 over-segmentation algorithm to grouping pixels with similar color or texture
221 into a subregion (see Fig. 3(c)). In F2_1, color information is disappear, shape
222

223

224

**Fig. 2.** Our network architecture.

information is augmented (see Fig. 3(d)). In F3_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 F4_2 and F5_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} , $\tilde{\mathbf{M}}$, $\hat{\mathbf{M}}$. Here we formulate our approach for building extraction. We denote our input training data set by $\mathbf{I} = \{(\mathbf{S}_n, \tilde{\mathbf{M}}_n), n = 1, \dots, |\mathbf{S}_n|\}$, where sample $\mathbf{S}_n = \{s_j^{(n)}, j = 1, \dots, |\mathbf{S}_n|\}$ denotes the raw input image and $\tilde{\mathbf{M}}_n = \{\tilde{m}_j^{(n)}, j = 1, \dots, |\mathbf{S}_n|\}$, $\tilde{m}_j^{(n)} \in \{0, 1\}$ 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 over all pixels in a training image $\mathbf{S} = \{s_j, j = 1, \dots, |\mathbf{S}|\}$ and building map $\tilde{\mathbf{M}} = \{\tilde{m}_j, j = 1, \dots, |\mathbf{S}|\}$, $\tilde{m}_j \in \{0, 1\}$. 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 $\hat{m}_j = Pr(\tilde{m}_j = 1 | \mathbf{S}; \mathbf{W})$.

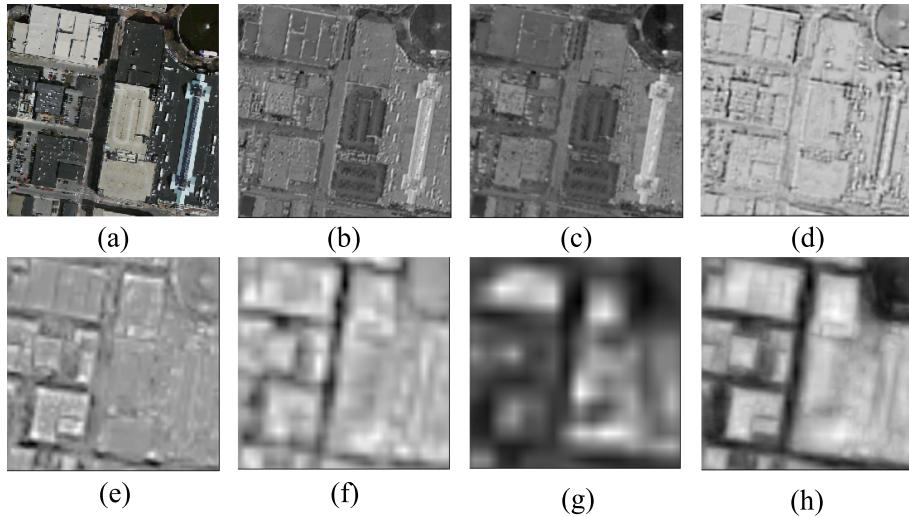


Fig. 3. (a) is input aerial image, feature maps generated from F1_1 (b), F1_2 (c), F2_1 (d), F3_3 (e), F4_2 (f), F5_2 (g), level 3 (h)

the definition of sigmoid cross-entropy loss function is shown in Eq (1).

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

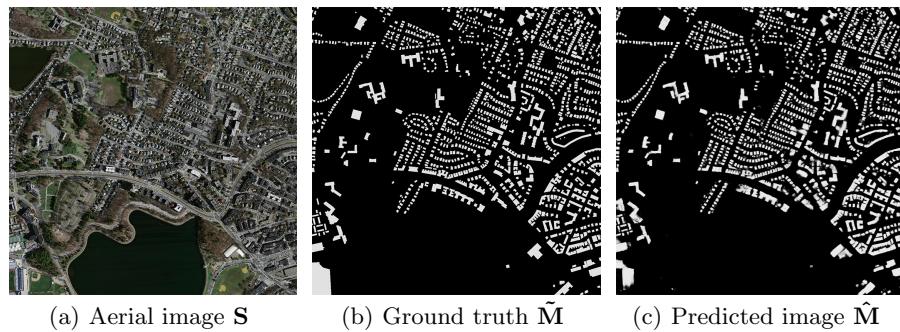


Fig. 4. An example of the resulting predicted image.

4 Experiments

In this section, we discuss 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* [21] 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.

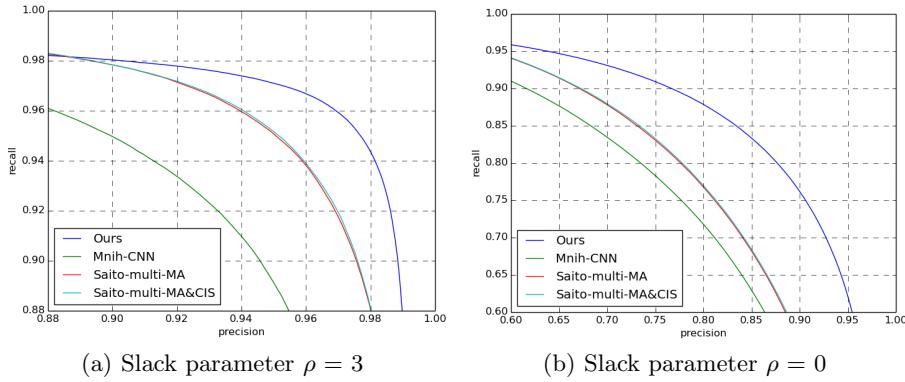
4.3 Results

To show the effectiveness of HF-FCN, we train and test our network on *MassBuildings*. 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 consuming is another

360 important index to evaluate system performance. We calculate the mean time
 361 of processing ten test images in the same computer using the same program.
 362 Table 2 shows that our method is able to not only significantly improve the
 363 performance, but dramatically decrease the time cost.

364

365



377 **Fig. 5.** Two relaxed precision-recall curves

378

379

380

381

382 To prove our network having strong ability in extracting buildings with variational
 383 appearances, arbitrary sizes, occlusions, we perform further evaluation.
 384 We crop seven 256×256 image patches that have buildings with variational ap-
 385 pearances or occlusions from test image of *Mass. Buildings*. And then, we directly
 386 crop corresponding predictions from predicted images generated by three models
 387 (Mnih-CNN+CRF [12], Saito-multi-MA&CIS [13] and ours). Here, we binarize
 388 the probability map using a threshold of 0.5. Seven groups of example are shown
 389 in Fig. 6. In addition, Table 3 shows the resulting recalls at breakeven points of
 390 standard precision recall curve for each patches.

391

392

393 **Table 2.** Performance is compared with [12, 13]. Recall here means recall at breakeven
 394 points. Time is computed in the same computer with a single NVIDIA Titan 12GB
 395 GPU.

396

397

	Recall ($\rho = 3$)	Recall ($\rho = 0$)	Time(s)
Mnih-CNN [12]	0.9150	0.7661	8.70
Mnih-CNN+CRF [12]	0.9211		
Saito-multi-MA [13]	0.9426	0.7858	67.72
Saito-multi-MA&CIS [13]	0.9488	0.7857	67.84
Ours (HF-FCN)	0.9643	0.8424	1.07

403

404

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

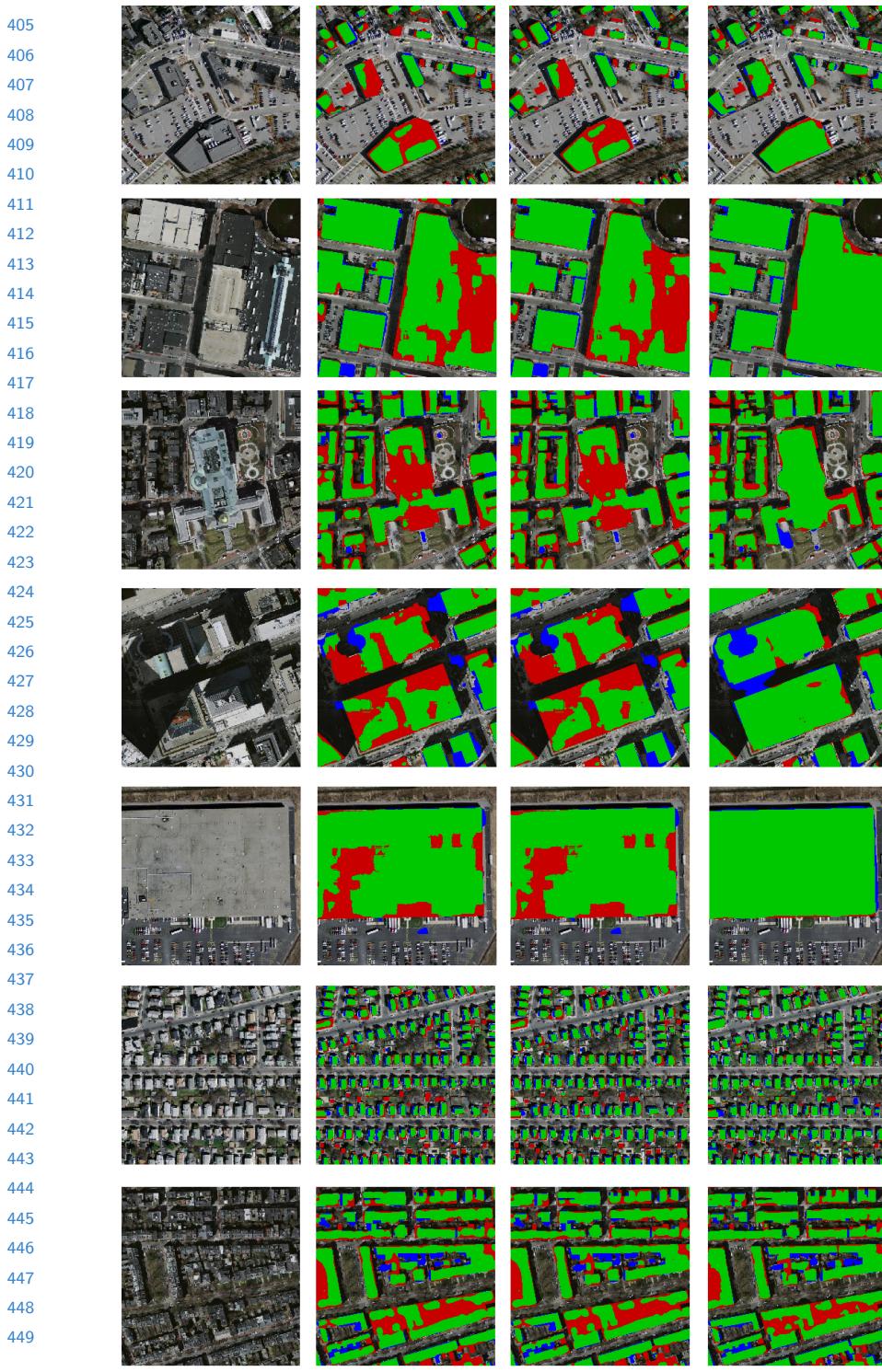


Fig. 6. (a) Input images. (b) Results of Mnih-CNN+CRF[12]. (c) Results of Saito-multi-MA&CIS[13]. (d) Our results. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

450 **Table 3.** Recall at selected region of the test images

451 Image ID	452 01	453 02	454 03	455 04	456 05	457 06	458 07	459 mean
Mnih-CNN+CRF[12]								
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

460

5 Conclusions

461 In this article, we proposed a improved fully convolutional network which is
462 strongly capable of extracting buildings with arbitrary sizes, variational appear-
463 ances or occlusions without any post-processing. Meanwhile, it further improves
464 the overall accuracy. The network can take arbitrary-size image as input as
465 long as GPU memory allowed. Compared with patch-based methods, there is
466 no need to label a whole image by cropping the image into small patches. As
467 consequence, inconsistant border caused by cropped would not occurred in our
468 system. Though a effective technique [13], namely, model averaging with spatial
469 displacement, is proposed, it is troublesome to train a network eight times and
470 predict labelling image with the same times. While in our system, time cost is
471 tremendously decreased. On the other hand, we demonstrate that our network
472 is generally adapt to various types of aerial scenes selected from real-world data.
473 Furthermore, our architecture can be easily extended to extract multi-objects in
474 remote sensing imagery. Consequently, we believe that our technique potentially
475 provides a generic solution to understand complex aerial scenes.

476

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

CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

12 ACCV-16 submission ID 97

- 495 7. Ok, A.O., Senaras, C., Yuksel, B.: Automated detection of arbitrarily shaped
496 buildings in complex environments from monocular vhr optical satellite imagery.
497 Geoscience and Remote Sensing, IEEE Transactions on **51** (2013) 1701–1717
498 8. Chen, D., Shang, S., Wu, C.: Shadow-based building detection and segmentation
499 in high-resolution remote sensing image. Journal of Multimedia **9** (2014) 181–188
500 9. Ngo, T.T., Collet, C., Mazet, V.: (Automatic rectangular building detection from
501 vhr aerial imagery using shadow and image segmentation)
502 10. Baluyan, H., Joshi, B., Al Hinai, A., Woon, W.L.: Novel approach for rooftop
503 detection using support vector machine. ISRN Machine Vision **2013** (2013)
504 11. Li, E., Femiani, J., Xu, S., Zhang, X., Wonka, P.: Robust rooftop extraction from
505 visible band images using higher order crf. Geoscience and Remote Sensing, IEEE
Transactions on **53** (2015) 4483–4495
506 12. Mnih, V.: Machine learning for aerial image labeling. Doctoral (2013)
507 13. Saito, S., Yamashita, Y., Aoki, Y.: Multiple object extraction from aerial imagery
508 with convolutional neural networks. Journal of Imaging Science & Technology **60**
509 (2016)
510 14. Ghaffarian, S., Ghaffarian, S.: Automatic building detection based on purposive
511 fastica (pfica) algorithm using monocular high resolution google earth images. IS-
512 PRS Journal of Photogrammetry and Remote Sensing **97** (2014) 152–159
513 15. Dornaika, F., Moujahid, A., Bosaghzadeh, A., El Merabet, Y., Ruichek, Y.: Object
514 classification using hybrid holistic descriptors: Application to building detection in
515 aerial orthophotos. Polibits (2015) 11–17
516 16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic
517 segmentation. Computer Science **79** (2014) 1337–1342
518 17. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic
519 image segmentation with deep convolutional nets and fully connected crfs. In:
520 ICLR. (2015)
521 18. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V.: Conditional random
522 fields as recurrent neural networks. (2015) 1529–1537
523 19. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmen-
524 tation. (2015)
525 20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale
526 image recognition. Eprint Arxiv (2014)
527 21. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama,
528 S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding.
529 Eprint Arxiv (2014) 675–678
530
531
532
533
534
535
536
537
538
539