Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set

Shunping Ji, Shiqing Wei, and Meng Lu

***Abstract*— The application of the convolutional neural network has shown to greatly improve the accuracy of building extraction from remote sensing imagery. In this paper, we created and made open a high-quality multisource data set for building detection, evaluated the accuracy obtained in most recent studies on the data set, demonstrated the use of our data set, and proposed a Siamese fully convolutional network model that obtained better segmentation accuracy. The building data set that we created contains not only aerial images but also satellite images covering 1000 km^2 with both raster labels and vector maps. The accuracy of applying the same methodology to our aerial data set outperformed several other open building data sets. On the aerial data set, we gave a thorough evaluation and comparison of most recent deep learning-based methods, and proposed a Siamese U-Net with shared weights in two branches, and original images and their down-sampled counterparts as inputs, which significantly improves the segmentation accuracy, especially for large buildings. For multisource building extraction, the generalization ability is further evaluated and extended by applying a radiometric augmentation strategy to transfer pretrained models on the aerial data set to the satellite data set. The designed experiments indicate our data set is accurate and can serve multiple purposes including building instance segmentation and change detection; our result shows the Siamese U-Net outperforms current building extraction methods and could provide valuable reference.**

***Index Terms*— Building extraction, deep learning, full convolutional network, remote sensing building data set.**

I. Introduction

BUILDING detection from remote sensing imagery has important implications in urban planning, population estimation, and topographic map making. The building detection has been studied for more than 30 years [1]. Novel data science and remote sensing technologies provide opportunities to automatically detect buildings, which could reduce tremendously manual works and contribute to urban dynamic monitoring. However, automatic building detection has been a long-term challenge in remote sensing due to the complex and heterogeneous appearance of buildings in mixed backgrounds.

Traditionally, the major work to detect buildings from aerial or satellite imagery is to design features that could best represent a building. The commonly used metrics such as color [2], spectrum [3], [4], length, edge [5], [6], shape [7], texture [4], [8], [9], shadow [1], [2], [10], height, and semantic [11] could vary under different circumstances of light, atmospheric conditions, sensor quality, scale, surroundings, and building architectures. The empirical feature design has shown to solve only specific problems with specific data, and is far from a general automatic building detection procedure.

Recently, convolutional neural network (CNN) has extended its application in remote sensing and shown important implications in labeling and classification [12], [13]. CNN automatically learns multilevel representations that map the original input to the designated binary or multiple labels (a classification problem), or to consecutive vectors (a regression problem). The powerful "representation learning" ability of CNN has made it gradually replacing the conventional feature handcrafting in a detection or classification application. Notably, the application of CNN on building detection greatly eases the feature design and has shown promising results [14], [15].

CNN has been extensively applied to image classification and segmentation. The commonly used CNN structures include AlexNet [16], VGGNet [17], GoogLeNet [18], and ResNet [19]. The output of these CNNs in image classification is typically a single class label. From 2015, special CNN structures are developed and contribute greatly to semantic segmentation, i.e., labeling every pixel of an image a category. Long *et al.* [20] extended the original CNN structure to enable dense prediction by a pixels-to-pixels fully convolutional network (FCN). In a FCN, feature maps are down sampled by levels of convolutions, and then transposed convolutions [21], [22] are typically applied to up-sample low-resolution features up to the original scale. Since then, a variety of FCNs have been proposed, such as SegNet [23], DeconvNet [24], and U-net [25]. In semantic segmentation of remote sensing images, earlier methods that applied non-FCN-based models are memory and computationally intensive [26]. Recent methods mostly leveraged FCN-based models [27].

Manuscript received April 27, 2018; revised June 25, 2018; accepted July 18, 2018. Date of publication August 22, 2018; date of current version December 24, 2018. This work was supported by the National Natural Science Foundation of China under Grant 41471288. *(Corresponding author: Shunping Ji.)*

S. Ji and S. Wei are with the School of Remote sensing and Information Engineering, Wuhan University, Wuhan, 430079, China (e-mail: jishunping@whu.edu.cn; wei_sq@whu.edu.cn).

M. Lu is with the Department of Physical Geography, Utrecht University, 3584 CE Utrecht, The Netherlands (e-mail: m.lu@uu.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2858817

The most recent studies on building extraction exclusively utilized the FCN-based methods. Maggiori *et al.* [14] designed a two-scale neuron module in an FCN to reduce the tradeoff between recognition and precise localization. Yuan [15] and Maggiori *et al.* [28] integrated multiple layers of activation into pixel level prediction based on FCN. Wu *et al.* [29] designed a multiconstraint FCN that utilizes multilayer outputs. Among these studies, only [28] utilized open-source data set (and opened the data set at the same time). As the current deep learning is data driven, the accuracy of deep learning technique depends heavily on the training data set. Several open, crowdsource data sets, such as ImageNet [30] and Coco [31], have dramatically stimulated the development of deep learning methods; however, such large, high-quality data sets generated from aerial, satellite imagery, or both, are scarce. As a result, researchers have to spend a huge amount of time on finding and constructing data sets. In addition, using different private data sets brings difficulties to quantitatively compare studies, and may hinder improving algorithms. Maggiori *et al.* [14] and Yuan [15] reported the undesirable accuracy of the used data sets. Wu *et al.* [29] used an accurate but small-size aerial building data set. Maggiori *et al.* [28] provide an open-source aerial building data set (named Inria data set) that contains scenes from five cities with 0.3-m spatial resolution. It can be used to test the extrapolation and generalization ability of deep learning methods. Satellite data set is a necessary supplement to aerial data for its large spatio-temporal coverage. However, there is no large open-source satellite building data set available and no relevant studies yet to evaluate the generalization from aerial data to satellite data and vice versa.

Besides the Inria data set that has been proposed in [28], there are only two open-source data sets that can be used for building extraction. One is a data set of 1-m ground resolution and contains 151 aerial image tiles of 1500×1500 pixels [32] (referred to as Massachusetts data set). The other is provided by the ISPRS society (referred to as the ISPRS data set) consists of two aerial subsets, the Vaihingen and Potsdam data sets [33]. The Vaihingen data set has a 0.05-m resolution, with 24 image tiles of 6000×6000 pixels and the Potsdam data set has a 0.09 resolution with 16 $11\,500 \times 7500$ images. The Massachusetts data set has low quality and resolution, and has not been applied to the current building extraction studies. Whereas the ISPRS data set covers 13 km^2 and few building instances to reflect the diversity in a building extraction problem. The 2018 IEEE GRSS Data Fusion Contest [34] also offers some high-resolution images for urban land cover classification, but all of them only cover a geographic area up to 4 km^2. Facing the current situation of limitation in open data sets, we created and made open a large, accurate building data set collection that contains both aerial and satellite images covering 450 and 550 km^2 area, respectively.

In addition to the need of large and accurate sample data sets, the design of special neural networks for remote sensing data plays an important role. As images are all captured from the same orthogonal bird-eye sight, scale may be the largest geometric issue that affects the performance of extracting different size of building instances, as FCN methods have shown limited ability to extract objects of very small or large sizes [20]. Many of the current building extraction studies, therefore, have focused on the scale deformation. Maggiori *et al.* [14] utilized a two-scale neuron module; Yuan [15] recovered every down-sampled layer to full resolution; Wu *et al.* [29] leveraged the multiscale outputs of multilayers in the U-Net structure. However, we empirically found all of these methods did not solve the scale problem well especially for those large buildings. Many points on a large roof are often wrongly classified to background even when the roof has the same color and texture.

Another issue we concern is the generalization and extrapolation ability of deep learning methods for building extraction from different remote sensor measurements. Maggiori *et al.* [28] discussed the problem of learning to extract buildings from different cities; however, the article only applied a pretrained model on source data sets directly to target data sets. Sherrah [35] found a pretrained CNN fine-tuned on remote sensing data can lead to better results compared to a network trained from scratch. In our study, a focus is on applying CNN model that is pretrained on aerial imagery to satellite imagery. Due to the long-distance atmospheric radiation transmission, the information contained in satellite imagery is more contaminated comparing to aerial imagery. We applied a radiometric augmentation strategy that enlarges the sample space of the source aerial data set, and hence improves the segmentation accuracy on satellite data set.

The main contributions of this paper are: 1) introducing and providing a large, accurate, and open-source data sets collection which consists of an aerial image data set with 220 000 samples of buildings from 0.075-m resolution images, and two satellite image data sets covering some scenes over the world and 2) evaluating the most recent methods thoroughly on the same benchmark and propose a novel variant of FCN specially designed for large-size building segmentation to address the scale problem of the most recent studies on the aerial data set. The following sections are arranged as follows. Section II provides a detailed description of the data set. Section III describes the novel variant of FCN. In Section IV, experiments are designed to thoroughly compare our data set to other open data sets and to compare our FCN structure to most recent studies. A discussion is provided in Section V, which especially addresses the transfer learning from aerial data set to satellite data set and evaluates the generalization ability of FCN; further prospects of using our data set as building instance segmentation and change detection are also discussed. Section VI finishes with the conclusion.

II. Aerial and Satellite Data Sets

We manually edited an aerial and a satellite imagery data set of building samples and named it a WHU building data set. The aerial data set consists of more than 220 000 independent buildings extracted from aerial images with 0.075-m spatial resolution and 450 km^2 covering in Christchurch, New Zealand (Fig. 1). This area contains countryside, residential, culture, and industrial area. Various and versatile architecture types of buildings with different color, size, and usage make it an ideal study area to evaluate the potential of a building

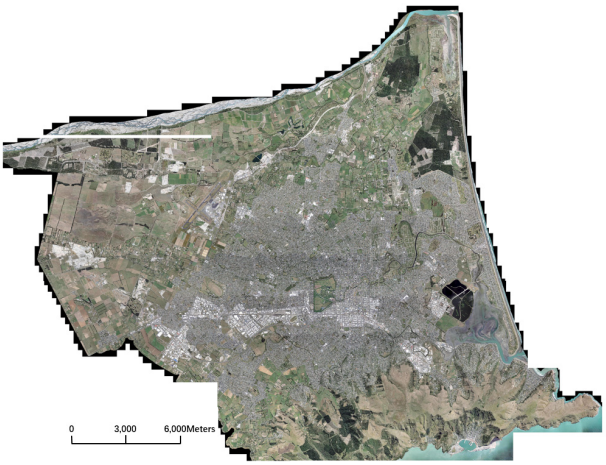


Fig. 1. Area covered by the aerial data set.

Fig. 2. Errors in the original vector data. Green polygons show the vectorized buildings of the original. We manually edited all these polygons (red polygon).

extraction algorithm. In addition, as the other open-source building data sets collects data from Europe (the Inria data set and the ISPRS data set) or America (the Massachusetts data set), our data set that collected from the southern hemisphere would be a beneficial supplement.

Although the original vector data of buildings and aerial images are openly provided by the land information service of New Zealand [36], the original data contains significant errors, such as missing, nonexisting, displaced buildings, and buildings that are not accurately delineated (Fig. 2). We edited and checked all the building samples of the original vector file using the ArcGIS software to produce a high-quality map. It took approximately six months to complete the whole manual work, among which discriminating man-made structures as large cars, containers, and greenhouses from buildings are the biggest challenges. Triple cross checking has been carefully carried out to minimize the risk of false judgment. The other small errors come from the buildings under the shades of trees. We have delineated the complete building shapes when the buildings are shaded by trees (as the middle image of Fig. 2). In our experiment, we found trees and buildings can be clearly discriminated as they are very different types. Hence the prediction accuracy could be underestimated. However, the bias is trivial as tree shading is not common in this area.

Besides providing the accurate shape file of the whole area, we edited a large subdata set containing 187 000 buildings (Fig. 3) which is ready to use for a CNN-based method. We down sampled the 0.075-m resolution aerial image to 0.3-m ground resolution as it has been experimentally proofed that the performance of an FCN method does not increase obviously with a resolution higher than 0.3 m. The down-sampled aerial images are seamlessly cropped into 8189 tiles with 512×512 pixels without overlapping, which are in proper size for a current mainstream Nvidia 1080 or Titan X GPU video card. The image tiles are numbered sequentially and can be easily reconverted to the whole georeferenced image.

Correspondingly, a Boolean raster map is derived from the building vector map, and then seamlessly cropped into 512×512 tiles as labels for CNN training. Fig. 4 shows examples of various building architectures and usages on 512×512 image tiles with both raster masks (blue) and vector shapes (red) available.

The satellite imagery data set consists of two subsets. One of them is collected from cities over the world and from various remote sensing resources including QuickBird, Worldview series, IKONOS, and ZY-3. We manually delineated all the buildings. It contains 204 images (512×512 tiles with resolutions varying from 0.3 to 2.5 m). Besides the differences in satellite sensors, the variations in atmospheric conditions, panchromatic and multispectral fusion algorithms, atmospheric and radiometric corrections, and season made the samples suitable yet challenging for testing robustness of building extraction algorithms (Fig. 5).

The other satellite building subdata set consists of six neighboring satellite images covering 550 km^2 on East Asia with 2.7-m ground resolution (Fig. 6). This test area is mainly designed to evaluate and to develop the generalization ability of a deep learning method on different data sources but with similar building styles in the same geographical area. It is also a useful compliment to other data sets that collected from Europe, America, and New Zealand and supplies regional diversity. The vector building map is also fully manually delineated in ArcGIS software and contains 29 085 buildings. The whole image is seamlessly cropped into 17 388 512×512 tiles for convenient training and testing with the same processing as in our aerial data set. Among them 21 556 buildings (13 662 tiles) are separated for training and the rest 7529 buildings (3726 tiles) are used for testing.

The WHU data set including both the aerial and satellite subdata sets with corresponding shape files and raster masks are freely available.[1]

Besides our data set, there are three data sets: the ISPRS data set [33], Massachusetts data set [32], and Inria data set [28], openly available in building extraction. Table I shows the ground resolution, area coverage, source, number of image tiles, and label format of these data sets. The ISPRS Vaihingen data set and Potsdam data sets provide labels for semantic segmentation, consisting of high-resolution orthophotographs and the corresponding digital surface models. However, the Vaihingen and Potsdam data sets only cover a very small ground range (2 and 11 km, respectively). Other data sets are much larger for representing the diversity of buildings. The Massachusetts data set covers 340 km but has a relatively low resolution. The spatial resolution and covering

[1] http://study.rsgis.whu.edu.cn/pages/download/

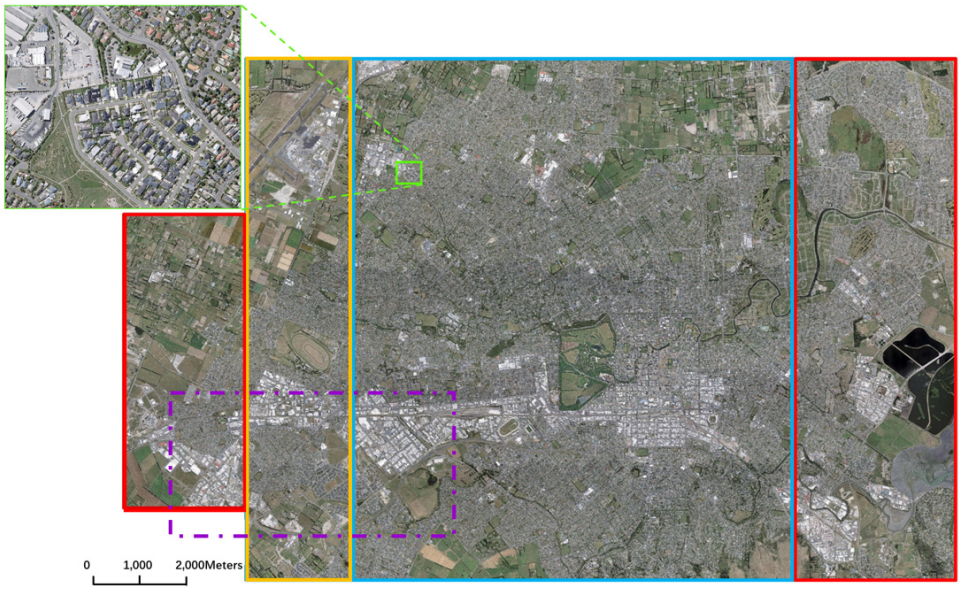


Fig. 3. Image covers most of the building area in the middle of the aerial data set. It was seamlessly cropped into 8189 512 × 512 tiles with 0.3-m ground resolution. The area in the blue box contains 130 000 buildings and is used for training, the area in the yellow box containing 14 500 buildings is used for validation and the rest in red box containing 42 000 buildings is used for testing. The area in dotted purple box provides two-period images for building change detection (see Section V-D).

Fig. 4. Examples of our aerial data set with different architectures, purposes, scales, and colors. The label format of the first row is with red vector shapes and the second row is with blue masks.

TABLE I

GENERAL COMPARISON BETWEEN OUR DATA SET AND OTHER OPEN-SOURCE DATA SETS

Datasets	GCD (m)	Area (km^2)	Source	Tiles	Pixels	Label Format
WHU (Ours)	0.075/2.7	450/550	aerial/sat	8189/17388	512×512	vector/raster
ISPRS	0.05/0.09	2/11	aerial	24/16	6000×6000/11500×7500	raster
Massachusetts	1.00	340	aerial	151	1500×1500	raster
Inria	0.3	405[1]	aerial	180	5000×5000	raster

[1] another test dataset covering 405 km^2 is used for evaluating submitted algorithm with unpublished labels.

area of the Inria data set are similar to our data set. It also contains scenes from five cities and could be used to evaluate the generalization ability of a building extraction algorithm.

However, among these open-source data sets, only the WHU data set provides satellite image sources and building vector maps, which are useful supplements to the current open

Fig. 5. Examples of the satellite data set I with different architectures from cities over the world. (a) Wuhan. (b) Taiwan. (c) Los Angeles. (d) Ottawa. (e) Cairo. (f) Milan. (g) Santiago. (h) Cordoba. (i) Venice. (j) New York.

data sets. In Section III, we will carefully evaluate the accuracy of these data sets with the same FCN model.

III. Network

FCN and its variants are the most commonly used architecture for semantic segmentation and building detection. We propose a new variant of FCN, which mainly consists of a Siamese U-Net structure and is called as SiU-Net, to improve the scale invariance of the algorithm for extracting buildings with different sizes from remote sensing data, as we found large buildings hinder a high performance of FCN-based methods on remote sensing building detection.

The SiU-Net is developed on the backbone of the U-Net structure. The improvement is mainly on the network input. In current stage, cropping the large-size high-resolution remote sensing image into tiles is unavoidable for a deep learning-based method. A large object covering the most of the scene leaves very small space for background, while the background plays usually an important role in object recognition both for computer and human. In the building extraction case, it has been empirically discovered that large buildings could be segmented more precisely in a coarser scale. Inspired by the study area of stereo matching [37], [38], we introduce a Siamese network that takes the original image tile and its down-sampled counterpart as inputs. The two branches for the two inputs in the network share the same U-Net structure and the same set of weights. The outputs of the branches are then concatenated for the final output.

Fig. 7(a) shows the structure of our Siamese network for building segmentation. 512×512 RGB image tiles and their down-sampled counterparts separately processed by the U-Net branches with shared weights. The two outputs of the U-Net are concatenated to produce a two-channel map, which corresponds to the two-channel labels (by concatenating the original label and the down-sampled label). The concatenated labels are utilized for training and weight updating; however, only the original label is used for evaluating the accuracy of model prediction. Fig. 7(b) shows the specific U-Net structure used in this paper. The inputs are first convoluted with 3×3 kernels and down sampled with max pooling layer-by-layer until 1024 32×32 feature maps are obtained. In the expanding stage, the lower layer features are up-convoluted

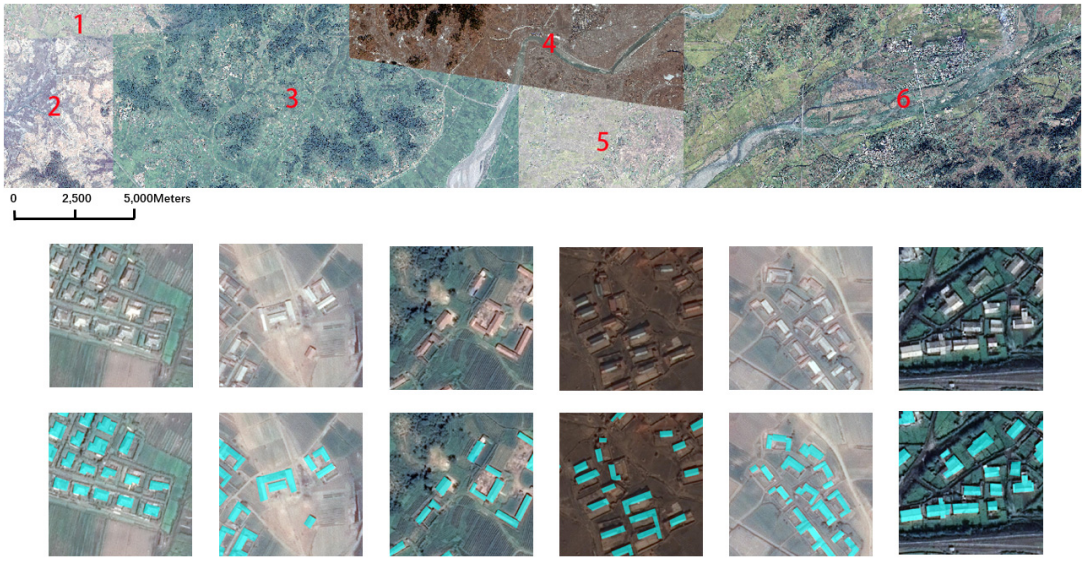


Fig. 6. Satellite data set II. An area of 550 km^2 covered by six satellite images in East Asia. The image tiles below are retrieved from the numbered areas and displayed sequentially.

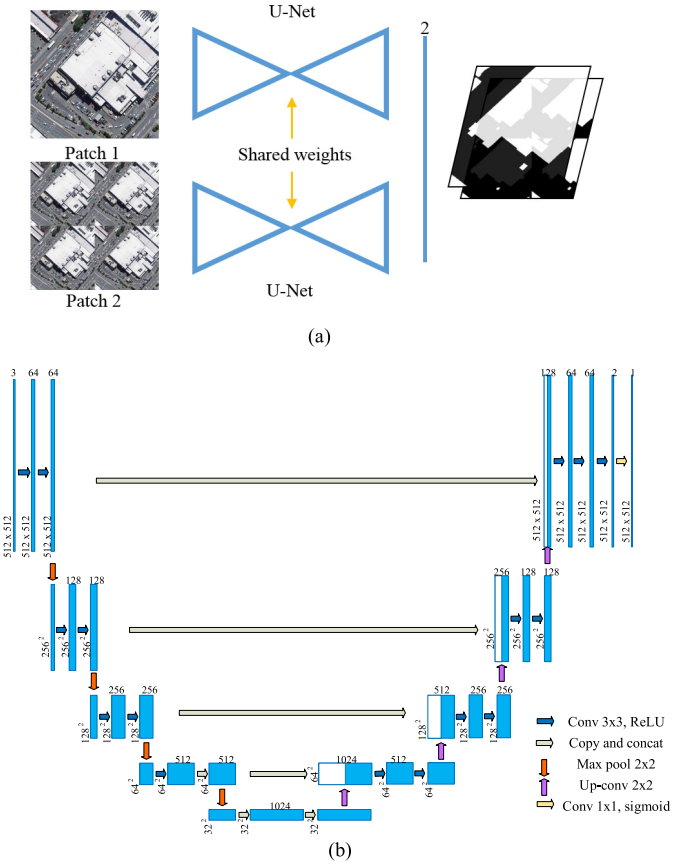


Fig. 7. (a) Structure of the SiU-Net. The counterpart of an original input consists of four 2× down-sampled tile images. (b) U-Net structure.

(by a transposed convolution operator) and concatenated with the same-layer features of the down-sampled stage, till the original scale.

In the end-to-end training, the rectified linear unit (ReLU) activation is used in all convolutional layers. An Adam (adaptive moment estimation) algorithm is used as a random gradient descent optimization with six image tiles as a mini batch. The learning rate is set to 0.0001. The weights of all filters are initialized according to a normal distribution initialization method [39], and all of the biases are initialized to zeros. The implementation is based on the Keras using a TensorFlow backend.

IV. Experiments and Results

A. Comparison to Open-Source Data Sets

We compare our aerial data set with the Massachusetts and Inria data sets using the U-Net as it has been shown to have obtained almost the best performance in building extraction [29]. The U-net architecture [Fig. 7(b)] is used for the comparison. From the aerial data set, we select 145 000 building for training (from which 14 500 buildings are used for validation) and 42 000 buildings for testing (Fig. 3). For the Massachusetts data set, we used three-quarters of samples (110 out of 151) for training and the rest for testing. For the Inria data set, we also used three-quarters of samples (27 out of 36 images) for training and the rest samples for testing. All the images (and the corresponding label maps) were seamlessly cropped to 512 × 512 tiles as network inputs for the limited GPU capacity. Basically, on our data set, the training of 130 000 building samples (4736 512 × 512 image tiles) stopped after 12 epochs. The process took about 3 h with a single NVIDIA Titan Xp GPU.

Three indicators are used to evaluate the accuracy of the detection results. The first one is the intersection on union (IoU), the ratio between the intersection of the building pixels detected by the algorithm and the true positive pixels and the result of their union. The second is the precision, the percentage of the true positive pixels among building pixels detected by the algorithm. The third is the recall, the percentage of the true positive pixels among building pixels in ground truth.

TABLE II

COMPARISON OF THE WHU DATA SET, THE MASSACHUSETTS DATA SET AND THE INRIA DATA SET USING THE U-NET

Dataset	IoU	Recall	Precision
WHU (ours)	0.858	0.945	0.903
Massachusetts	0.552	0.746	0.681
Inria	0.714	0.821	0.846

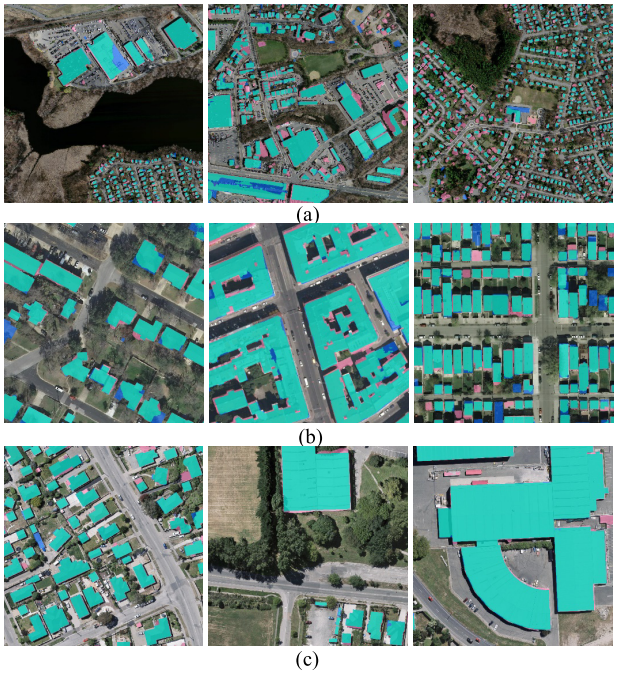


Fig. 8. Examples of segmentation results using the U-Net on the three data sets. Blue: reference; green: predicted; and pink: wrongly classified. (a) Massachusetts data set. (b) Inria data set. (c) WHU aerial data set.

The comparison results are shown in Table II and Fig. 8. Table II shows the IoU and precision/recall of the Massachusetts data set 30% and 20% lower than ours, respectively. The Massachusetts data set has a lower quality and resolution, which negatively affect the U-Net model to accurately detect buildings. Some obvious wrong labels can be found from the data set. In Fig. 8, labels are indicated in blue, predictions in green, and false positive in pink. The middle image of Fig. 8(a) shows that some blue labels (on the top left corner) do not have the corresponding buildings.

The Inria data set obtained much better results than the Massachusetts data set. It is also comparable to our data set as they have similar spatial resolution. Our data set outperformed the Inria data set 14% in IoU and 20% in recall, and they showed almost the same score in precision. We reviewed the images from the Inria data set and discover the main reason for its relatively lower accuracy might be due to some challenging cases such as with higher buildings and shadows. Another reason could also be that a few wrong labels exist in the data set. For example, Fig. 8(b) (right) shows six correctly predicted buildings that were wrongly taken as false positive (pink) as the labels are missing. As for our data set, we spent plenty of time in cross checking to guarantee the best labeling accuracy. Although the Inria data set shows to obtain a lower performance compared to our WHU data set, it is valuable for evaluating the generalization ability of a deep learning-based method as it contains scenes from multiple cities.

TABLE III

COMPARISON BETWEEN THE U-NET AND SIU-NET ON THE AERIAL DATA SET

Methods	IoU	Recall	Precision
U-Net	0.868	0.945	0.903
SiU-Net	0.884	0.939	0.938

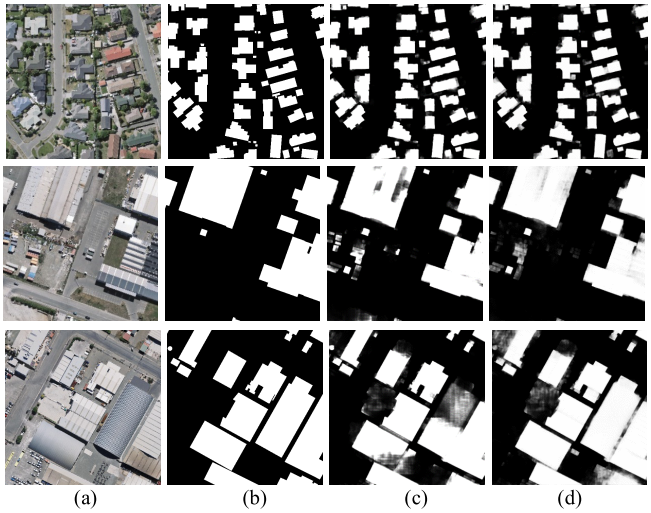


Fig. 9. Examples of segmentation results with the U-Net and SiU-Net, respectively, on the aerial data set. (a) Image. (b) Label. (c) U-Net. (d) SiU-Net.

B. Experiments on Aerial Data Set

Using the same network and input settings as was described in the Section IV-A, Table III shows the results of our proposed SiU-Net. After introducing a Siamese structure to U-net, the IoU improved 1.6% and the precision improved 3.5%. We ran the SiU-Net 5 times and the deviation of the IoU, recall, and precision is 0.00084, 0.0040, and 0.0039, respectively, indicating the IoU being nearly invariant. Although the U-Net itself is a multiscale structure and has some ability to learn multiscale features, our simple strategy using different scale inputs could further improve the accuracy. Fig. 9 shows some qualitative results. The first image in Fig. 9 contains small buildings on which the U-Net and SiU-Net perform almost the same. The images in the second and third rows consist of much larger buildings and SiU-Net performed obviously better than U-Net. From the upper building in the second row image and the two buildings with semicircular roof in the third row image, it could be observed that although the roofs share the same texture and color, they were not fully segmented by the U-Net. However, the segmentation problem on large-scale buildings could be significantly alleviated using our simple multiscale input strategy.

TABLE IV
COMPARISON BETWEEN THE U-NET AND SIU-NET ON THE SATELLITE DATA SET I AND II, RESPECTIVELY

Datasets	Methods	IoU	Recall	Precision
I	U-Net	0.577	0.733	0.731
	SiU-Net	0.595	0.780	0.716
II	U-Net	0.594	0.869	0.653
	SiU-Net	0.611	0.796	0.725

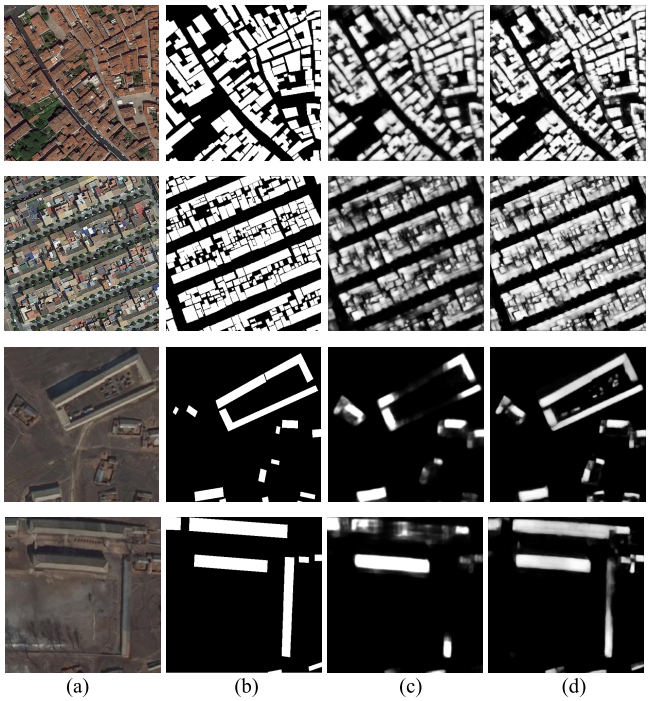


Fig. 10. Examples of segmentation results with the U-Net and SiU-Net, respectively, on the satellite data set. (a) Image. (b) Label. (c) U-Ne. (d) SiU-Net.

C. Experiments on Satellite Data Sets

With the same settings as the aerial data set, the experiments result in Table IV on the satellite data set I and II showed the SiU-Net obtains 1.7% IoU improvement compared to the U-Net. In the test of the data set I that consists of 204 images acquired from over the world, the recall was increased 4.7% and the precision dropped 1.5% when the SiU-Net is applied. The images of the first two rows in Fig. 10 are two examples. The shapes of the predicted region by the two methods are similar; however, the SiU-Net seems obviously clearer, indicating the method shows better confidence to its judgment.

In the test of the data set II, which consists of six adjacent satellite images and covers 550 km^2 with 2.7-m ground resolution, the recall dropped 7.3% and the precision improved 7.2%. The significant drop of recall could mainly be due to the image quality and the low resolution. After the additional constraint was added, i.e., the half-resolution inputs and their labels, the recall rate dropped, especially due to small buildings. However, on large buildings as in the third row and fourth row images of Fig. 10, the SiU-Net also performed better than the U-Net.

TABLE V
COMPARISON OF MOST RECENT STUDIES ON OUR AERIAL DATA SET

Methods	IoU	Recall	Precision
SiU-Net (Ours)	0.884	0.939	0.938
2-scale FCN [14]	0.701	0.758	0.903
MLP [15,28]	0.713	0.785	0.887
CU-Net [29]	0.871	0.917	0.946
U-Net [25]	0.868	0.945	0.914
FCN [20]	0.854	0.892	0.953

D. Comparison of Most Recent Studies

We then evaluate the performances of different building extraction methods under the same settings. We compare our methods to [14], [15], [28], and [29]. Reference [15] and [28] used an MLP upon an FCN structure (short as MLP). Reference [15] utilized a two-scale FCN and the [29] leveraged multiconstraint U-Net (short as CU-Net). From Table V, we see the methods based on the U-Net structure performed significantly better than the two-scale FCN and MLP with 15% IoU improvement. For two-scale FCN, we checked the method and the corresponding code provided in [28], and found the backbone structure of the FCN contains some problems. For example, the randomly sampled inputs with 64×64 pixels contain less information and could confuse the CNN classifier (e.g., a negative sample on a road has the same texture as a positive sample on a roof); only two scales are used other than popular four scales as in FCN [20] and U-Net [25]; there is only one feature map (other than 32 or more maps typically) before up-convolution. We introduced the FCN network proposed in [20] and got 0.854 IoU on the same data set. However, after introducing the two-scale strategy upon it, the IoU dropped 1%. The results are compatible with [14] that reported the two-scale strategy has no effect for a standard training–testing procedure and [29] that reported the IoU of the FCN was about 2% lower than that of the U-Net.

The reason that the accuracy of the MLP is much lower than the U-Net is also due to some problems existed in the FCN backbone that is used in [28]. A theoretical problem might also exist in the MLP. Although an FCN that aims to segment image in pixel level can be achieved by a typical ladder structure as in Fig. 7(b) or a series of convolution with full-resolution layers, the later has not been considered in current variants of FCNs as it requires more GPU capacity and is much more computationally intensive. An MLP algorithm that aims at recovering every lower spatial resolution layer in a common FCN structure to a layer combination of full resolution therefore seems lacking of efficiency. In our test, the MLP run 55 000 times in 20 h without complete convergence. The experiment of [28] took more than 50 h to run. On the contrary, the other methods in Table V all converged within 6 h. It could be concluded the low efficiency of the MLP limits its potential applications.

Our method outperformed the latest CU-Net 1.3% in IoU. Although CU-Net achieved some scale invariance by utilizing multiscale outputs of a U-Net structure, the improvement is

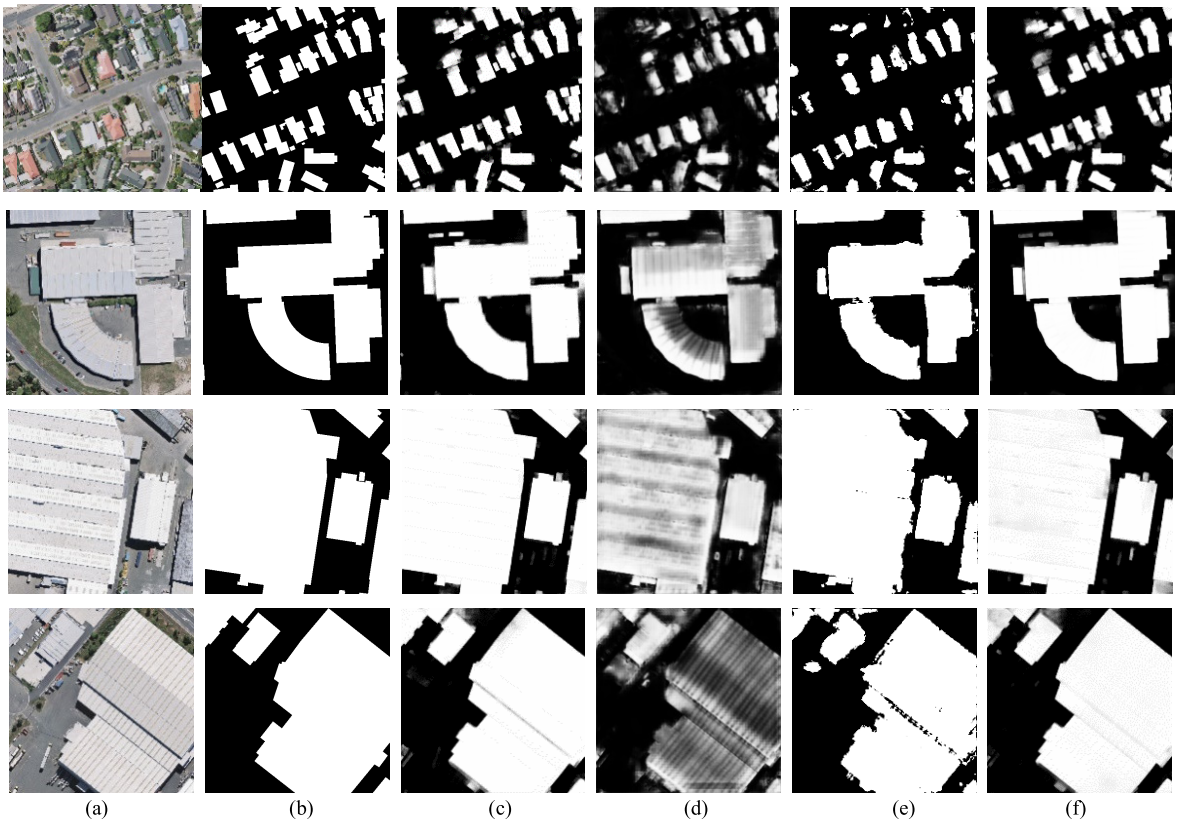


Fig. 11. Comparison of the prediction results from the most recent studies on the WHU aerial data set. (a) Image. (b) Label. (c) SiU-Net. (d) Two-scale FCN. (e) MLP. (f) CU-Net.

modest (0.3%). The simple intuition of our method that utilizes the different resolutions of input achieved better results. As both the recall and precision indexes are already higher than 93% in our method, the 1.3% improvement is not trivial.

Fig. 11 shows four examples predicted by different methods. The two-scale FCN and MLP perform worse than the SiU-Net and CU-Net. In the first two images, the CU-Net and SiU-Net almost perform the same; in the last two images, the SiU-Net shows better confidence on the predicted pixels on the large buildings and many more darker points (with lower score) appear on the buildings predicted by the CU-Net. The MLP provided by [28] utilized softmax for binary labeling and provide only binary labels here.

V. Discussion

A. Direct Transfer Learning From Aerial Data Set to Satellite Data Set via Radiometric Augmentation

The extrapolation and generalization ability of deep learning is crucial for automation but have remained unsatisfactory in computer vision and remote sensing applications when a source data set varies significantly from a target data set. In this section, we evaluate this ability via the transfer learning strategy from our aerial data set to the satellite data sets. We first trained the U-net parameters according to the 145 000 aerial building samples, and then apply them directly on the satellite data set I and II. From Table VI, all of the indicators are very low comparing to the test on the aerial data set. The IoU of the data set I only reach to 27.3%. It is even worse when applying the pretrained model on the data set II as it bears almost no resemblance to the aerial data set. In this case, the deep learning method lacks the extrapolation ability of a direct model transfer.

TABLE VI
Direct Prediction on the Satellite Data Sets by the U-Net and the Spectrally Augmented U-Net Pretrained on the Aerial Data Set

Dataset	Method	IoU	Recall	Precision
I	U-Net	0.273	0.359	0.531
	Augmented U-Net	0.394	0.565	0.566
II	U-Net	0.037	0.207	0.044
	Augmented U-Net	0.288	0.530	0.387

As spectral distortion between multisource remote sensing data sets could be a key factor for algorithm degeneration considering the long-distance atmospheric radiometric transmission, we further evaluate the performance of a spectral augmented U-Net, which samples original inputs with different virtual radiometric situations and expands the sample space in the spectral dimension. The radiometric parameter set consists of linear stretching, histogram equalization (binomial distribution), blurs, and salt noise (discrete Gaussian). A counterpart generator is used to first randomly draw samples from the distributions of the given parameters. Then, these samples

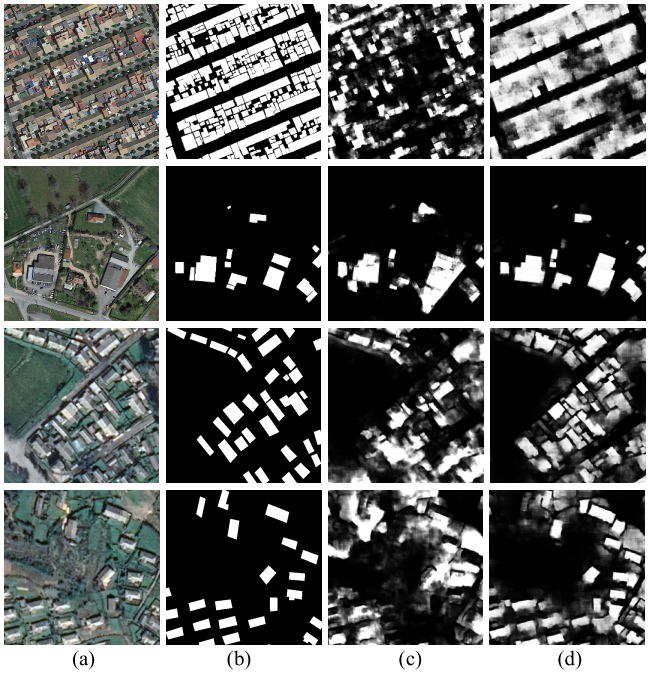


Fig. 12. Segmentation results with the U-net and the spectrally enhanced U-net on the WHU satellite data sets. (a) Image. (b) Label. (c) U-Net. (d) Spectrally enhanced U-Net.

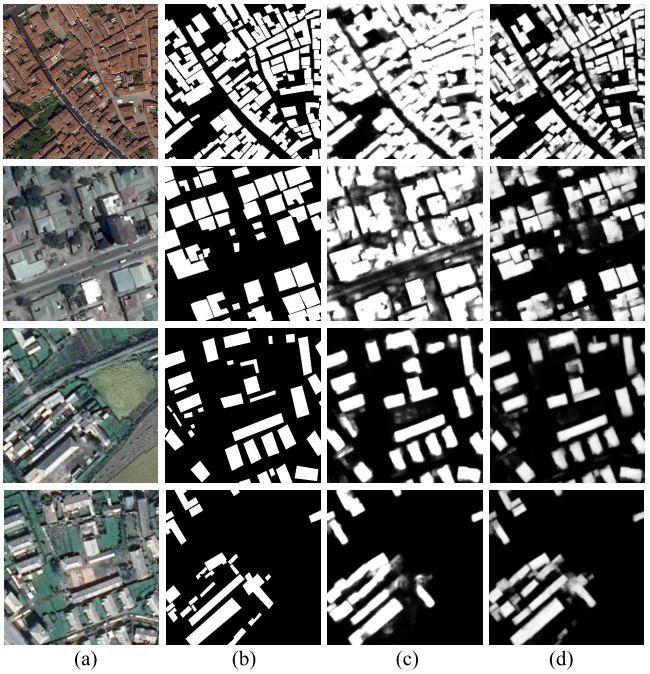


Fig. 13. Segmentation results with direct training on the satellite data set and fine tuning based on the pretrained model on the aerial data set. (a) Image. (b) Label. (c) Direct training. (d) Fine tuning.

TABLE VII
FINE TUNING ON THE SATELLITE DATA SETS WITH THE AUGMENTED U-NET PRETRAINED ON THE AERIAL DATA SET OUTPERFORMED DIRECT TRAINING BOTH ON EFFICIENCY AND ACCURACY

Datasets	Methods	Epoch	IoU	Recall	Precision
I	Direct training	12	0.577	0.733	0.731
	Pretrained	6	0.659	0.842	0.752
II	Direct training	10	0.594	0.869	0.653
	Pretrained	4	0.640	0.850	0.721

are used to resample the original image to a new input sample. The result in Table VI shows that with the radiometric enhancement, the metrics obtained significant improvement: about 12% and 25% IoU improvement on data set I and II, respectively. Fig. 12 shows four satellite samples with the first two images from the data set I and the rest from the data set II. It could be observed that with radiometric augmentation the performance is improved. However, the 39.4% and 28.8% IoU of the satellite data sets indicate the generalization ability need to be further improved.

B. Fine Tuning on Target Satellite Data Sets

We applied a transfer learning strategy with fine tuning on the satellite data sets. We select three-quarters of satellite images for model fine tuning and the rest for prediction. The network parameters are initialized by the pretrained augmented U-Net on the aerial data set. From Table VII, compared to direct training with random initial weights on the satellite images, the transfer learning with fine tuning shows better convergence in epoch iteration that saves more computational time and has obtained a higher IoU (8.2% and 4.6% improvements, respectively). Therefore, it might be a good choice utilizing available pretrained models in building extraction even if the source data set and the target data set are very different. Fig. 13 also shows the predicted maps of fine tuning on pretrained model are clearer and more accurate comparing to that of a direct training.

C. Recovering Image From Cropped Tiles

Due to limited GPU memory, cropping remote sensing images is currently unavoidable when using a deep learning method. Image cropping creates marginal effects, which poses a problem to most conventional classification methods. Our experiments show that FCN is robust against the marginal effect. From Figs. 9–11, it has been observed that the fractured objects in the margin are precisely detected using the FCN-based method. It could be explained that FCN has learned this pattern from a large amount of training samples with building parts. We then recover larger predicted building maps of the aerial data set by seamlessly stitching the 512×512 tiles. Fig. 14 shows two examples with small residential buildings and large industrial buildings respectively where no stitching trace could be observed. Hence, it is not necessary to crop images into overlapped tiles, or to draw patch inputs randomly and dynamically in training, the latter may require more iterations and time to converge.

D. Further Prospects of Our Data Set

As we provide vector maps of buildings, the current FCN-based pixel-wise segmentation can be easily extended to individual building instance segmentation that not only

Fig. 14. Large images (with predicted mask) that recovered from 512×512 tiles. No stitching trace could be found when using FCN-based methods.

Fig. 15. Building instance segmentation using mask R-CNN on the aerial data set.

segments pixels with a building mask but also recognizes single buildings via bounding box. Most recent region-based CNN methods could be introduced, such as mask R-CNN [40]. Although pixel-wise FCN methods can be further processed to retrieve building instances, it is not end-to-end and cannot separate buildings from adjacent pixels. Benefiting from the vector maps of building shapes provided by our data set, we can easily retrieve the bounding box of each building as a new type of label. As an initial experiment, we trained a mask R-CNN model on the aerial 145 000 buildings and checked the model on the 42 000 buildings. We kept all the settings of the original mask R-CNN unchanged and run 22 h in a single GPU. From Table VIII, we can see the AP50 (precision that obtained on 50% IoU) of bounding box reaches 83.6%, and the IoU of mask is 84.8%, slightly lower than that of the U-Net. In Fig. 15, all of the bounding boxes are correctly predicted. The mask of buildings is also accurate however it could be further improved as some building edges in the right image were not very accurate.

TABLE VIII

BUILDING INSTANCES (BOUNDING BOX AND MASK) RETRIEVED FROM MASK R-CNN

Method	Bounding box			Mask		
	AP_{50}	Recall	Precision	IoU	Recall	Precision
MASK R-CNN	0.836	0.887	0.846	0.848	0.938	0.898
U-Net	/	/	/	0.868	0.945	0.903
SiU-Net	/	/	/	0.884	0.939	0.938

Fig. 16. Aerial images (with vector shapes) acquaint in 2012 and 2016, respectively, consist of an ideal area for studying building change detection.

The second important application of our data set is building change detection and updating. Our data set covers an area where a 6.3-magnitude earthquake has occurred in February 2011 and rebuilt in the following years. The original aerial data set consists of aerial images acquaint in 2016. We additionally provide a sub-data set that consists of aerial images obtained in April 2012 that contains 12 796 buildings in 20.5 km^2 (16 077 buildings in the same area in 2016 data set). By manually selecting 30 GCPs on ground surface, the subdata set was geo-rectified to the aerial data set with 1.6-pixel accuracy. Fig. 16 shows two images covering the same area, where many buildings appeared or were rebuilt. This subdata set and the corresponding images from the original data set are now openly provided along with building vector and raster maps.

VI. CONCLUSION

A large sample size, accurate, and multisource data set plays an indispensable role in developing and applying deep neural network to remote sensing applications. First, we provide an aerial and satellite building data set, which is expected to contribute to developing and evaluating novel methods such as pixel-wise segmentation, multisource transfer learning, instance segmentation and change detection. The experiments show our aerial data set achieved the best accuracy compared to using other existing data sets with the same FCN method. Second, we thoroughly evaluate the performance of recent studies in building extraction on the same aerial data set and introduced a novel Siamese FCN model. It is shown that among these FCN-based architectures, U-Net-based methods performed better than older methods such as two-scale FCN and MLP, and our SiU-Net achieved the best accuracy. Third, as an attempt to address multisource learning and generalization ability of deep learning, we applied radiometric augmentation in aerial data set for pretraining, which significantly improved the prediction accuracy of applying the

pretrained model to satellite images. However, different from the satisfactory results that could be achieved in building extraction on homogenous data sets, the generalization ability of deep learning for multisource data sets is still limited and requires to be further studied.

ACKNOWLEDGMENT

The authors would to thank S. Tian, Z. Qin, R. Zhu, C. Zhang, Y. Shen, Y. Wang, J. Liu, D. Yu, and S. Hu from Wuhan University, Wuhan, China, and Q. Chen from the China University of Geosciences, Wuhan, China, to help with preparing the data set.

REFERENCES

[1] Y.-T. Liow and T. Pavlidis, "Use of shadows for extracting buildings in aerial images," *Comput. Vis. Graph. Image Process.*, vol. 48, no. 2, pp. 242–277, 1989.

[2] B. Sirmacek and C. Unsalan, "Building detection from aerial images using invariant color features and shadow information," in *Proc. Int. Symp. Comput. Inf. Sci.*, Oct. 2008, pp. 1–5.

[3] S.-H. Zhong, J.-J. Huang, and W.-X. Xie, "A new method of building detection from a single aerial photograph," in *Proc. Int. Conf. Signal Process.*, Oct. 2008, pp. 1219–1222.

[4] Y. Zhang, "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering," *ISPRS J. Photogramm. Remote Sens.*, vol. 54, no. 1, pp. 50–60, 1999.

[5] Y. Li and H. Wu, "Adaptive building edge detection by combining LiDAR data and aerial images," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 37, pp. 197–202, Jul. 2008.

[6] G. Ferraioli, "Multichannel InSAR building edge detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1224–1231, Mar. 2010.

[7] A. V. Dunaeva and F. A. Kornilov, "Specific shape building detection from aerial imagery in infrared range," *Vychislitelnaya Matematika Inform.*, vol. 6, no. 3, pp. 84–100, 2017.

[8] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Improved building detection using texture information," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 143–148, Apr. 2011.

[9] P. S. Tiwari and H. Pande, "Use of laser range and height texture cues for building identification," *J. Indian Soc. Remote Sens.*, vol. 36, no. 3, pp. 227–234, 2008.

[10] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image," *J. Multimedia*, vol. 9, no. 1, pp. 181–188, 2014.

[11] C. Zhong, Q. Xu, F. Yang, and L. Hu, "Building change detection for high-resolution remotely sensed images based on a semantic dependency," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 3345–3348.

[12] J. Guo, Z. Pan, B. Lei, and C. Ding, "Automatic color correction for multisource remote sensing images with wasserstein CNN," *Remote Sens.*, vol. 9, no. 5, p. 483, 2017.

[13] Y. Yao, Z. Jiang, H. Zhang, B. Cai, G. Meng, and D. Zuo, "Chimney and condensing tower detection based on faster R-CNN in high resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3329–3332.

[14] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[15] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[17] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[21] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.

[22] V. Dumoulin and F. Visin. (2016). "A guide to convolution arithmetic for deep learning." [Online]. Available: https://arxiv.org/abs/1603.07285

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[24] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[26] Z. Guo, X. Shao, Y. Xu, H. Miyazaki, W. Ohira, and R. Shibasaki, "Identification of village building via Google Earth images and supervised machine learning methods," *Remote Sens.*, vol. 8, no. 4, p. 271, 2016.

[27] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.

[28] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[29] G. Wu *et al.*, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, 2018.

[30] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[31] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[32] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[33] *ISPRS 2D Semantic Labeling Contest*. Accessed: Jul. 1, 2018. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[34] B. Le Saux, N. Yokoya, R. Hansch, and S. Prasad, "2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 52–54, Mar. 2018.

[35] J. Sherrah. (2016). "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery." [Online]. Available: https://arxiv.org/abs/1606.02585

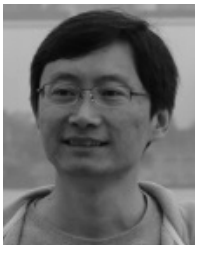
[36] *LINZ Data Service*. Accessed: Jul. 1, 2018. [Online]. Available: https://data.linz.govt.nz/

[37] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4353–4361.

[38] J. Zbontar and Y. Lecun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, pp. 1–32, Apr. 2016.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.



Shunping Ji received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. He has co-authored over 40 papers. His research interests include photogrammetry, remote sensing image processing, mobile mapping system, and machine learning.

Shiqing Wei received the B.Sc. degree in geographic information science from the China University of Petroleum, China, in 2017. He is currently pursuing the M.Sc. degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include remote sensing, machine learning.

Meng Lu received the M.Sc. degree in earth science system from the University of Buffalo, Buffalo, NY, USA, and the Ph.D. degree in geoinformatics from the University of Muenster, Muenster, Germany.

She was a Research Associate with the Department of Physical Geography, Utrecht University, Utrecht, The Netherlands, where she was involved in spatial data analysis, environmental modeling, and geocomputation. Her research interests include geoscientific data analysis, spatiotemporal statistics, machine learning, remote sensing, environmental modeling, and health geography.