DOI: 10.1109/TGRS.2018.2858817

Fully Convolutional Networks for Multi-Source Building Extraction from An Open Aerial and Satellite Imagery Dataset

Shunping Ji, Shiqing Wei, Meng Lu

Abstract—**The application of the convolutional neural network (CNN) has shown to greatly improve the accuracy of building extraction from remote sensing imagery. In this study, we created and made-open a high quality multi-source dataset for building detection, evaluated the accuracy obtained in most recent studies on the dataset, demonstrated the use of our dataset, and proposed a Siamese FCN model that obtained better segmentation accuracy. The building dataset 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 dataset outperformed several other open building datasets. On the aerial dataset, 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 multi-source building extraction, the generalization ability is further evaluated and extended by applying a radiometric augmentation strategy to transfer pre-trained models on the aerial dataset to the satellite dataset. The designed experiments indicate our dataset 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; remote sensing building dataset; full convolutional network; deep learning**

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 studies for more than thirty 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, semantic [11], etc., could vary under different circumstances of light, atmospheric conditions, sensor quality, scale, surroundings and building architectures, etc. 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 labelling and classification [12, 13]. CNN automatically learns multi-level 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., labelling 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], 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].

The most recent studies on building extraction exclusively utilized the FCN-based methods. [14] designed a two-scale neuron module in a FCN to reduce the trade-off between recognition and precise localization. [15, 28] integrated multiple layers of activation into pixel level prediction based on FCN. [29] designed a multi-constraint FCN that utilizes multi-layer outputs. Among these studies, only [28] utilized open-

This paper was submitted on February 7, 2018.

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

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

source dataset (and opened the dataset at the same time). As the current deep learning is data driven, the accuracy of deep learning technique depends heavily on the training dataset. Several open, crowdsource datasets, such as ImageNet [30], Coco [31], have dramatically stimulated the development of deep-learning methods; however, such large, high-quality datasets 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 datasets. In addition, using different private datasets brings difficulties to quantitatively compare studies, and may hinder improving algorithms. Both [14] and [15] reported the undesirable accuracy of the used datasets. [29] used an accurate but small-size aerial building dataset. [28] provides an open-source aerial building dataset (named Inria dataset) 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 dataset is a necessary supplement to aerial data for its large spatio-temporal coverage. However, there is no large open-source satellite building dataset available and no relevant studies yet to evaluate the generalization from aerial data to satellite data and vice versa.

Besides the Inria dataset that has been proposed in a most recent study [28], there are only two open-source datasets that can be used for building extraction. One is a dataset of 1 m ground resolution and contains 151 aerial image tiles of 1500×1500 pixels [32] (referred to as Massachusetts dataset). The other is provided by the ISPRS society (referred to as the ISPRS dataset) consists of two aerial subsets, the Vaihingen and Potsdam datasets [33]. The Vaihingen dataset has a 0.05m resolution, with 24 image tiles of 6000×6000 pixels and the Potsdam dataset has a 0.09 resolution with 16 11500×7500 images. The Massachusetts dataset has low quality and resolution, and has not been applied to the current building extraction studies. Whereas the ISPRS dataset 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 datasets, we created and made-open a large, accurate and open building dataset collection that contains both aerial and satellite images covering 450 km^2 and 550 km^2 area respectively.

In addition to the need of large and accurate sample datasets, 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. [14] utilized a two-scale neuron module; [15] recovered every down-sampled layer to full-resolution; [29] leveraged the multi-scale outputs of multi-layers 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. [28] discussed the problem of learning to extract buildings from different cities, however the article only applied a pre-trained model on source datasets directly to target datasets. [35] found a pre-trained 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 dataset and hence improves the segmentation accuracy on satellite dataset.

The main contributions of the paper are, 1) introducing and providing a large, accurate and open-source datasets collection which consists of an aerial image dataset with 220,000 samples of buildings from 0.075 m resolution images, and two satellite image datasets 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 dataset. The following sections are arranged as follows: Section II provides a detailed description of the dataset. Section III describes the novel variant of FCN. In section IV, experiments are designed to thoroughly compare our dataset to other open datasets and to compare our FCN structure to most recent studies. A discussion is provided in Section V that especially address the transfer learning from aerial dataset to satellite dataset and evaluate the generalization ability of FCN; further prospects of using our dataset as building instance segmentation and change detection are also discussed. Section VI finishes with a conclusion.

II. The Aerial and Satellite Datasets

We manually edited an aerial and a satellite imagery dataset of building samples and named it a WHU building dataset. The aerial dataset 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 extraction algorithm. In addition, as the other open-source building datasets collects data from Europe (the Inria dataset and the ISPRS dataset) or America (the Massachusetts dataset), our dataset 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, non-existing, 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 6 months to complete the whole manual work, among which discriminating manmade 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 judgement. 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.

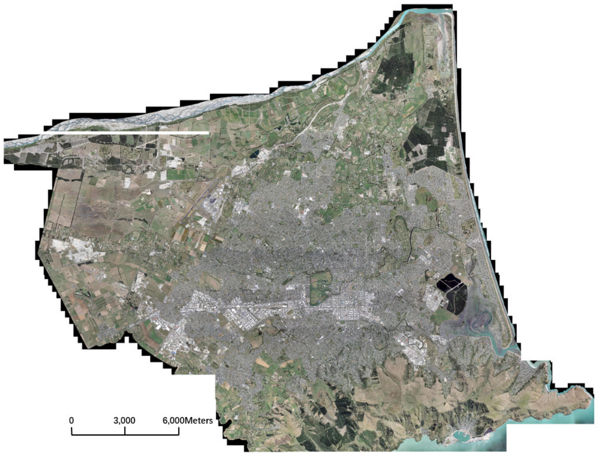


Fig. 1. The area covered by the aerial dataset.

Besides providing the accurate shape file of the whole area, we edited a large sub-dataset containing 18,7000 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 0.3m. The down-sampled aerial images are seamlessly cropped into 8,189 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.

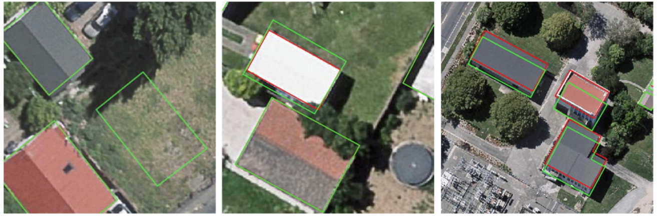


Fig. 2. Errors in the original vector data. Green polygons show the vectorized buildings of the original. We manually edited all these polygons (red polygon).

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 dataset 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, ZY-3, etc. We manually delineated all the buildings. It contains 204 images (512 × 512 tiles with resolutions varying from 0.3 m 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).

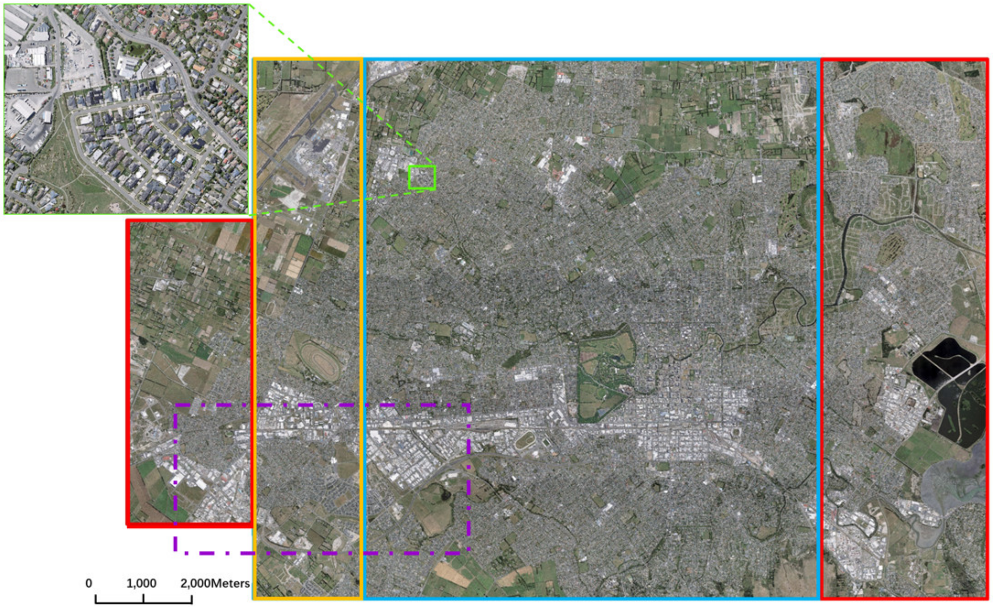


Fig. 3. The image covers most of the building area in the middle of the aerial dataset. It was seamlessly cropped into 8189 512×512 tiles with 0.3 m ground resolution. The area in the blue box contains 130,000 buildings and is used for training, the area in the yellow box containing 14,500 buildings 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 5.4)

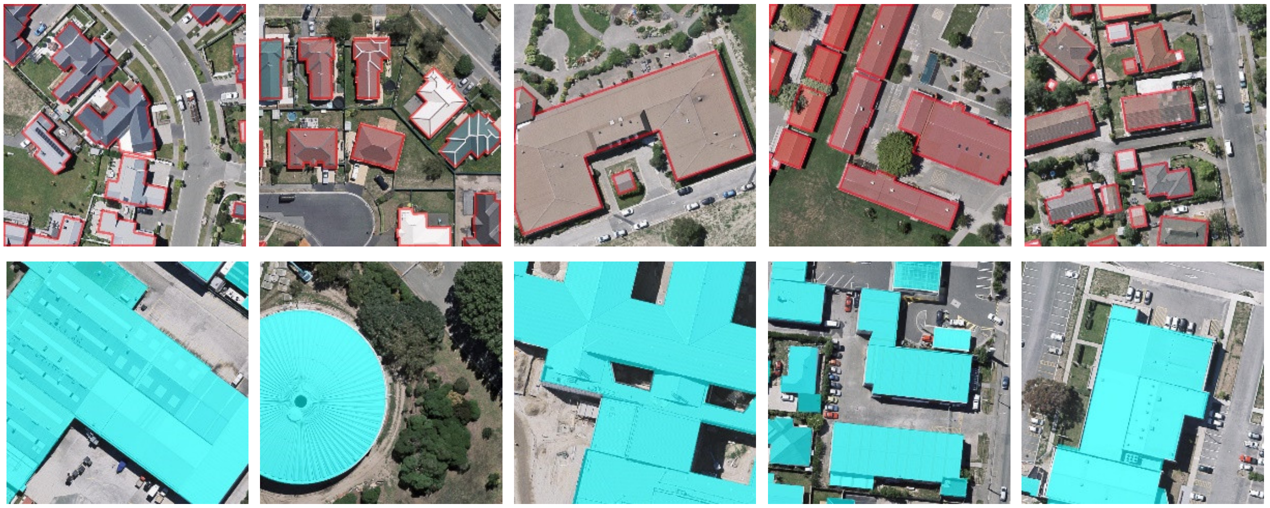
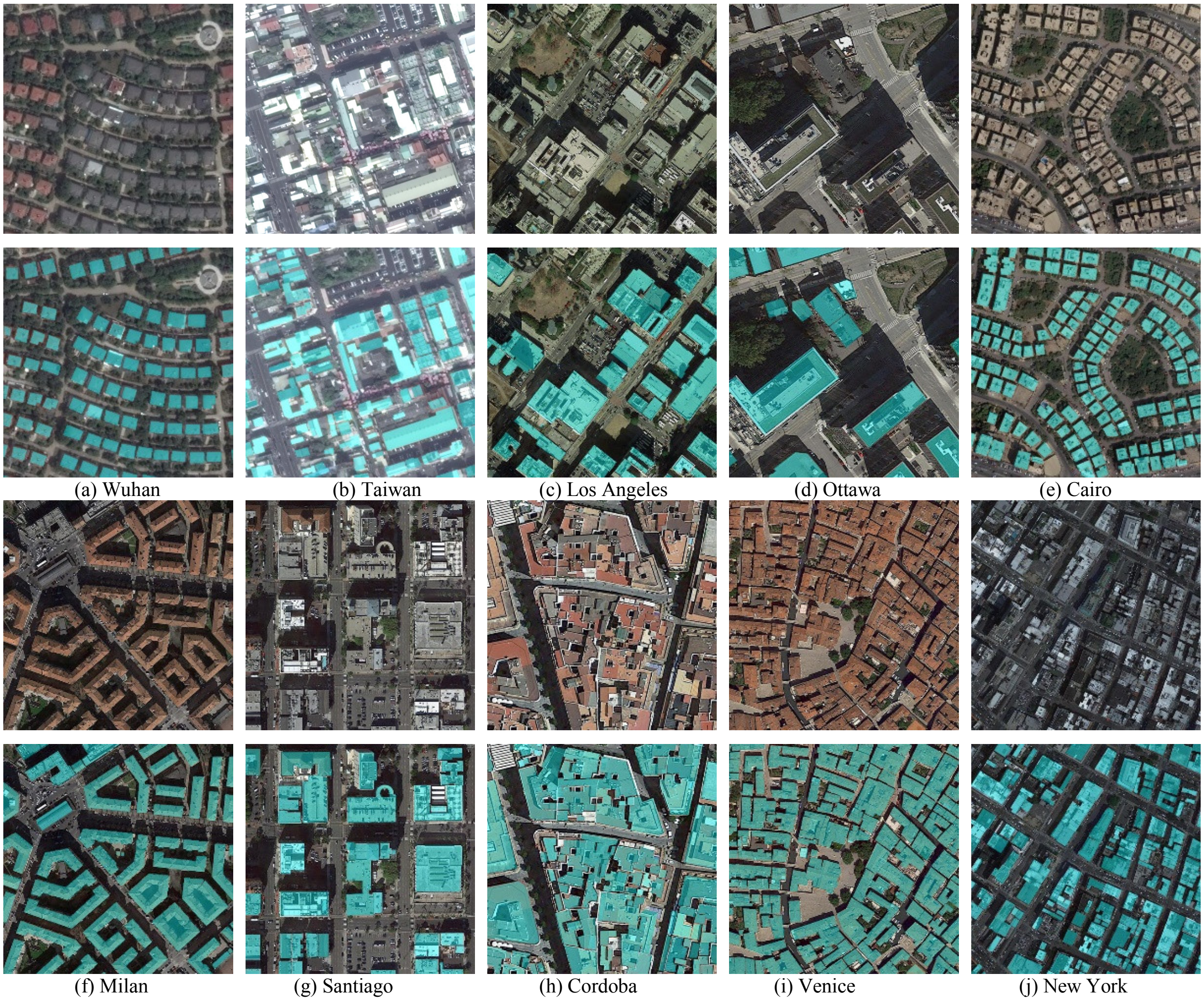


Fig. 4. Examples of our aerial dataset with different architectures, purposes, scales and colors. The label format of the first row is with red vector shapes and the second row is with blue masks.



(a) Wuhan (b) Taiwan (c) Los Angeles (d) Ottawa (e) Cairo

(f) Milan (g) Santiago (h) Cordoba (i) Venice (j) New York

Fig. 5. Examples of the satellite dataset I with different architectures from cities over the world.

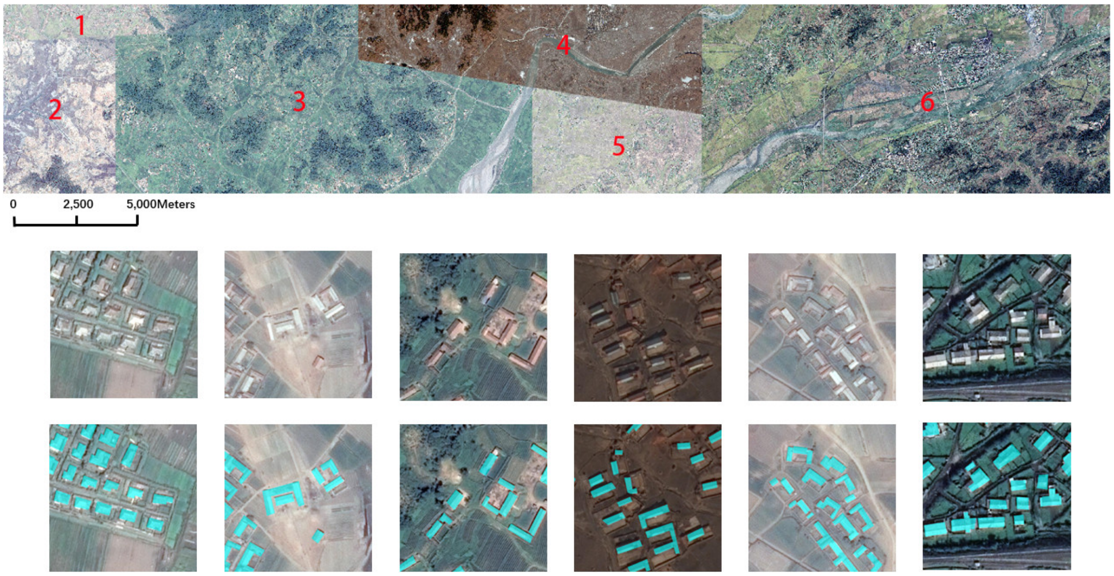


Fig. 6. Satellite dataset II. An area of 550 km^2 covered by six satellite images in East Asia. The image tiles below are retrieved from the numbered areas and displayed sequentially.

The other satellite building sub-dataset consists of 6 neighboring satellite images covering 550 km^2 on East Asia with 2.7 m ground resolution (Fig. 7). 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 datasets 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 29085 buildings. The whole image is seamlessly cropped into 17388 512×512 tiles for convenient training and testing with the same processing as in our aerial dataset. Among them 21556 buildings (13662 tiles) are separated for training and the rest 7529 buildings (3726 tiles) are used for testing.

The WHU dataset including both the aerial and satellite sub-datasets with corresponding shape files and raster masks are freely available[1].

Besides our dataset, there are three datasets: the ISPRS dataset [33], Massachusetts dataset [32] and Inria dataset [28], openly available in building extraction. Table 1 shows the ground resolution, area coverage, source, number of image tiles, and label format of these datasets. The ISPRS Vaihingen dataset and Potsdam datasets provide labels for semantic segmentation, consisting of high resolution ortho-photos and the corresponding digital surface models (DSMs). However, the Vaihingen and Potsdam datasets only cover a very small ground range (2km and 11 km respectively). Other datasets are much larger for representing the diversity of buildings. The Massachusetts dataset covers 340 km but has a relatively low resolution. The spatial resolution and covering area of the Inria dataset are similar to our dataset. 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 datasets, only the WHU dataset provides satellite image sources and building vector maps, which are useful supplements to the current open datasets. In section III, we will carefully evaluate the accuracy of these datasets with the same FCN model.

TABLE I
GENERAL COMPARISON BETWEEN OUR DATASET AND OTHER OPEN SOURCE DATASETS

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.

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.

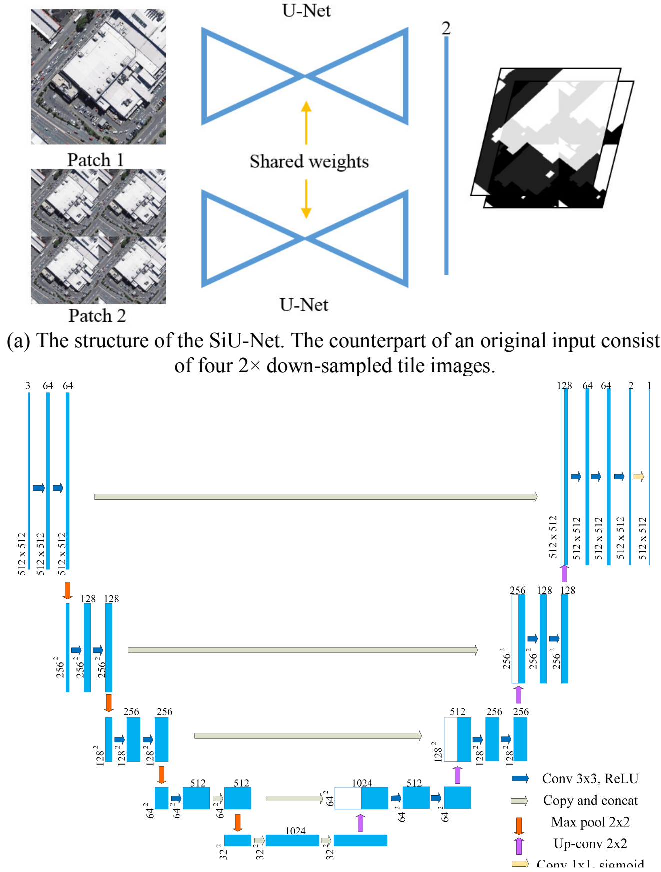
The SiU-Net is developed on the backbone of the U-Net structure. The improvement is mainly on the network input. In current stage, cropping the large-size high-resolution remote sensing image into tiles is unavoidable for a deep learning based

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

DOI: 10.1109/TGRS.2018.2858817

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 2-channel map, which corresponds to the 2-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 the paper. The inputs are firstly 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 (by a transposed convolution operator) and concatenated with the same-layer features of the down-sample stage, till the original scale.



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

(b) The U-net structure

Fig. 7. The SiU-Net structure (a) and its main body, the U-Net structure (b).

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 6 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 datasets

We compare our aerial dataset with the Massachusetts and Inria datasets 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. 7b) is used for the comparison. From the aerial dataset, 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 dataset, we used three-quarters of samples (110 Out of 151) for training and the rest for testing. For the Inria dataset, 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 dataset, the training of 130,000 building samples (4736 512×512 image tiles) stopped after 12 epochs. The process took about 3 hours with a single NVIDIA Titan Xp GPU.

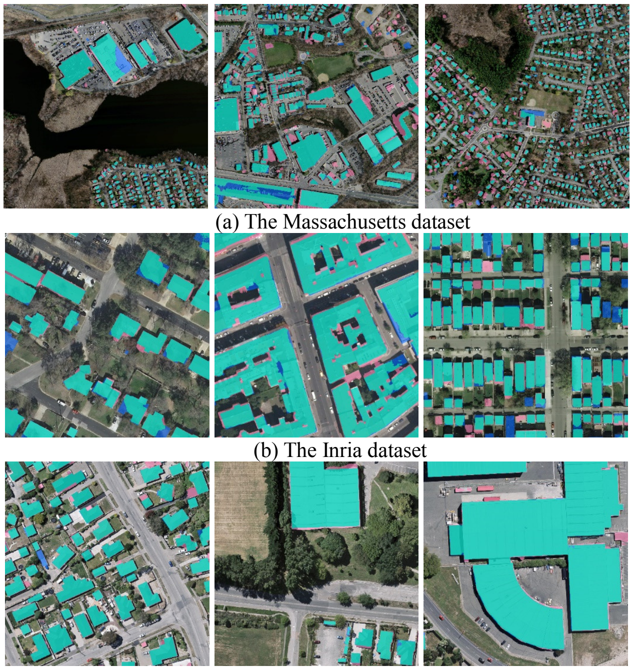
Three indicators are used to evaluate the accuracy of the detection results. The first one is the intersection on union (IoU), the ratio between the intersection of the building pixels detected by the algorithm and the true positive pixels and the result of their union. The second is the precision, the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

The comparison results are shown in Table 2 and Fig. 8. Table 2 shows the IoU and precision/recall of the Massachusetts dataset 30% and 20% lower than ours, respectively. The Massachusetts dataset 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 dataset. 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.

TABLE II
The Comparison of the WHU Dataset, the Massachusetts Dataset and the Inria Dataset 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

DOI: 10.1109/TGRS.2018.2858817



(a) The Massachusetts dataset

(b) The Inria dataset

(c) The WHU aerial dataset.

Fig. 8. Examples of segmentation results using the U-Net on the three datasets. Blue: reference; green: predicted; pink: wrongly classified.

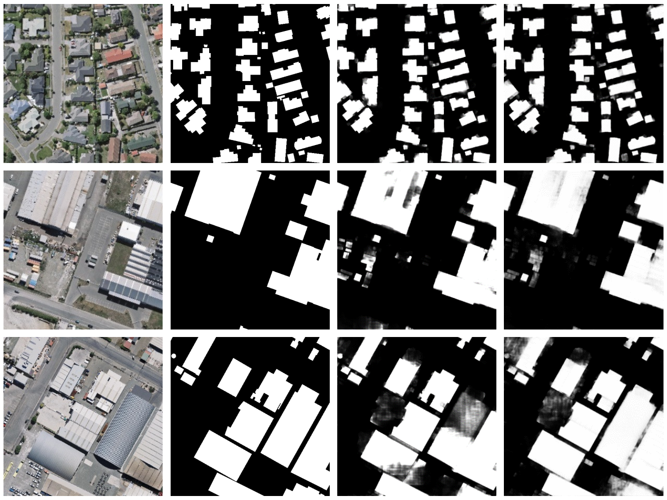
The Inria dataset obtained much better results than the Massachusetts dataset. It is also comparable to our dataset as they have similar spatial resolution. Our dataset outperformed the Inria dataset 14% in IoU and 20% in recall, and they showed almost the same score in precision. We reviewed the images from the Inria dataset 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 dataset. For example, the right image of Figure 8(b) shows six correctly predicted buildings that were wrongly taken as false positive (pink) as the labels are missing. As for our dataset, we spent plenty of time in cross-checking to guarantee the best labelling accuracy. Although the Inria dataset shows to obtain a lower performance compared to our WHU dataset, it is valuable for evaluating the generalization ability of a deep learning based method as it contains scenes from multiple cities.

B. Experiments on aerial dataset

Using the same network and input settings as was described in the Section A, Table 3 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 multi-scale structure and has some ability to learn multi-scale features, our simple strategy using different scale inputs could further improve the accuracy. Figure 9 shows some qualitative results. The first image in Figure 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 multi-scale input strategy.

TABLE III
COMPARISON BETWEEN THE U-NET AND SIU-NET ON THE AERIAL DATASET

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



(a) Image (b) Label (c) U-Net (d) SiU-Net

Fig. 9. Examples of segmentation results with the U-Net and SiU-Net respectively on the aerial dataset.

C. Experiments on satellite datasets

With the same settings as the aerial dataset, the experiments result in Table 4 on the satellite dataset I and II showed the SiU-Net obtains 1.7% IoU improvement compared to the U-Net. In the test of the dataset 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 judgement.

TABLE IV
COMPARISON BETWEEN THE U-NET AND SIU-NET ON THE SATELLITE DATASET 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

In the test of the dataset 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.

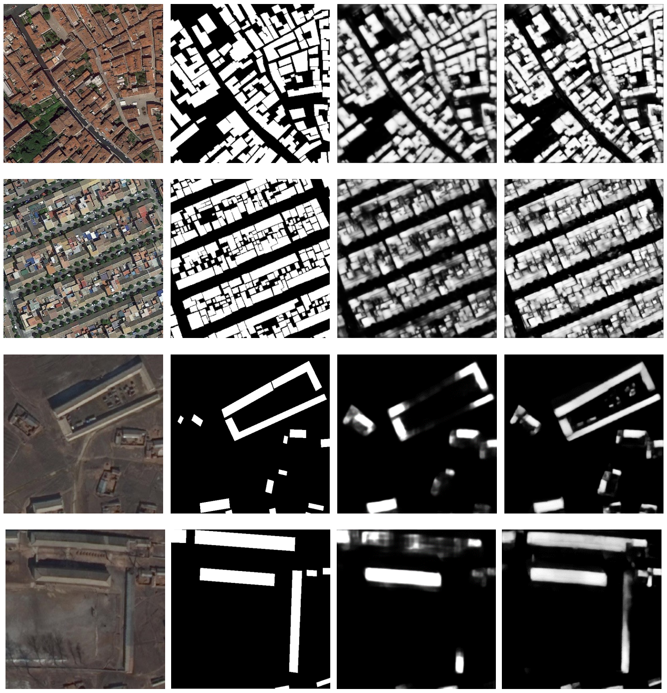


Fig. 10. Examples of segmentation results with the U-Net and SiU-Net respectively on the satellite dataset.

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 most recent studies [14, 15, 28, 29]. The [15] and [28] used an MLP upon a FCN structure (short as MLP). The [15] utilized a two-scale FCN and the [29] leveraged multi-constraint U-Net (short as CU-Net). From Table 5, we see the methods based on the U-Net structure performed significantly better than the 2-scale FCN and MLP with 15% IoU improvement. For 2-scale FCN, we checked the method and the corresponding code provided by [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 dataset. However, after introducing the 2-scale strategy upon it, the IoU dropped 1%. The results are compatible with [14] that reported the 2-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.

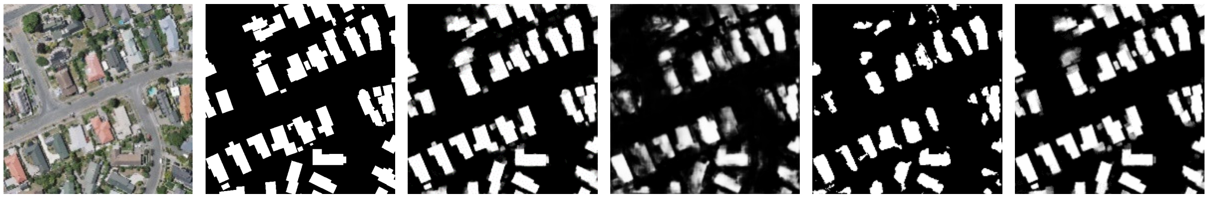
TABLE V
THE COMPARISON OF MOST RECENT STUDIES ON OUR AERIAL DATASET

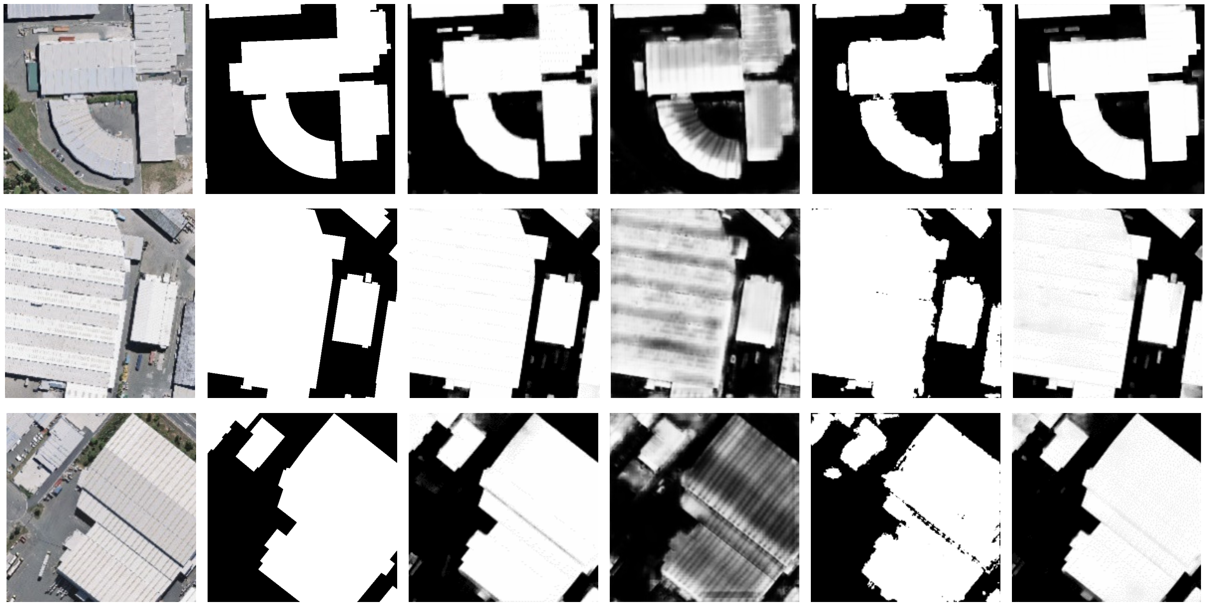
Methods	IoU	Recall	Precision
SiU-Net (Ours)	0.884	0.939	0.938
2-scale FCN [14]	0.701	0.758	0.903
MLP [15,28]	0.713	0.785	0.887
CU-Net [29]	0.871	0.917	0.946
U-Net [25]	0.868	0.945	0.914
FCN [20]	0.854	0.892	0.953

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, resulting in very low efficiency. 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 counter-intuitive. In our test, the MLP run 55000 times in 20 hours without complete convergence. The experiment of [28] took more than 50 hours to run. On the contrary, the other methods in Table 5 all converged within 6 hours. 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 multi-scale outputs of a U-Net structure, the improvement is 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 2-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 [28] utilized softmax for binary labelling and provide only binary labels here.





(a) Image (b) Label (c) SiU-Net (d) 2-scale FCN (e) MLP (f) CU-Net

Fig. 11. Comparison of the prediction results from the most recent studies on the WHU aerial dataset.

V. Discussion

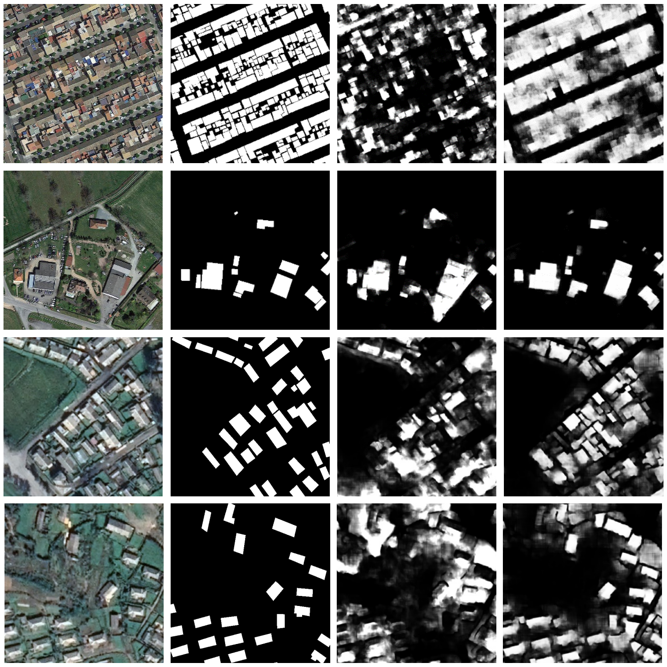
A. Direct transfer learning from aerial dataset to satellite dataset via radiometric augmentation

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 dataset varies significantly from a target dataset. In this section, we evaluate this ability via the transfer learning strategy from our aerial dataset to the satellite datasets. We firstly trained the U-net parameters according to the 14,5000 aerial building samples, and then apply them directly on satellite dataset I and II. From Table 6, all of the indicators are very low comparing to the test on the aerial dataset. The IoU of the dataset I only reach to 27.3%. It is even worse when applying the pretrained model on the dataset II as it bears almost no resemblance to the aerial dataset. In this case the deep learning method lacks the extrapolation ability of a direct model transfer.

As spectral distortion between multi-source remote sensing datasets 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 firstly randomly draw samples from the distributions of the given parameters. Then, these samples are used to resample the original image to a new input sample. The result in Table 6 shows that with the radiometric enhancement, the metrics obtained significant improvement: about 12% and 25% IoU improvement on dataset I and II respectively. Fig. 12 shows four satellite samples with the first two images from the dataset I and the rest from the dataset II. It could be observed that with radiometric augmentation the performance is improved. However, the 39.4% and 28.8% IoU of the satellite datasets indicate the generalization ability need to be further improved.

TABLE VI
DIRECT PREDICTION ON THE SATELLITE DATASETS BY THE U-NET AND THE SPECTRALLY AUGMENTED U-NET PRETRAINED ON THE AERIAL DATASET

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



(a) Image (b) Label (c) U-Net (d) Spectrally Enhanced U-Net

Fig. 12. Segmentation results with the U-net and the spectrally enhanced U-net on the WHU satellite datasets.

B. Fine tuning on target satellite datasets

We applied a transfer learning strategy with fine tuning on the satellite datasets. 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 dataset. From Table 7, 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 dataset and the target dataset 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.

TABLE Ⅶ
FINE TUNING ON THE SATELLITE DATASETS WITH THE AUGMENTED U-NET PRETRAINED ON THE AERIAL DATASET 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

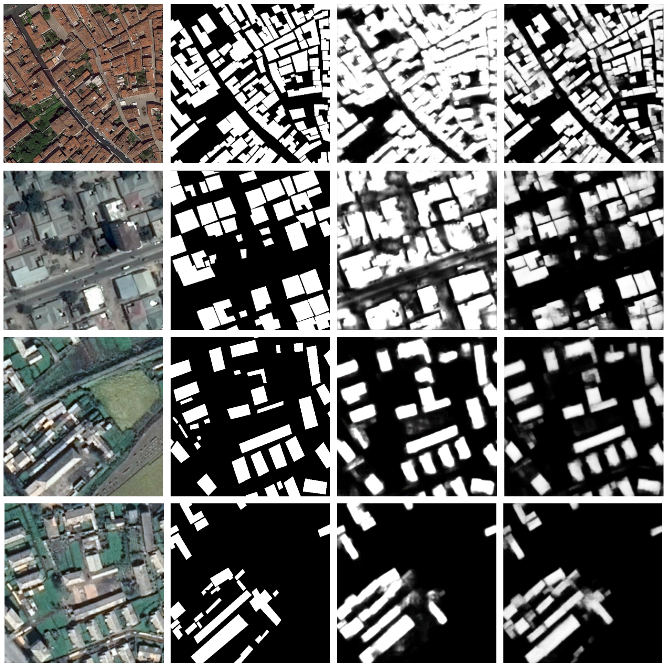

(a) Image (b) Label (c) Direct Training (d) Fine Tuning

Fig. 13. Segmentation results with direct training on the satellite dataset and fine tuning based on the pretrained model on the aerial dataset.

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 previous figures as Fig. 9 ~ 11, it has been observed that the fractured objects in the margin is 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 dataset by seamlessly stitching the 512×512 tiles. Fig. 14 shows two examples with small residential buildings and large industrial buildings 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.

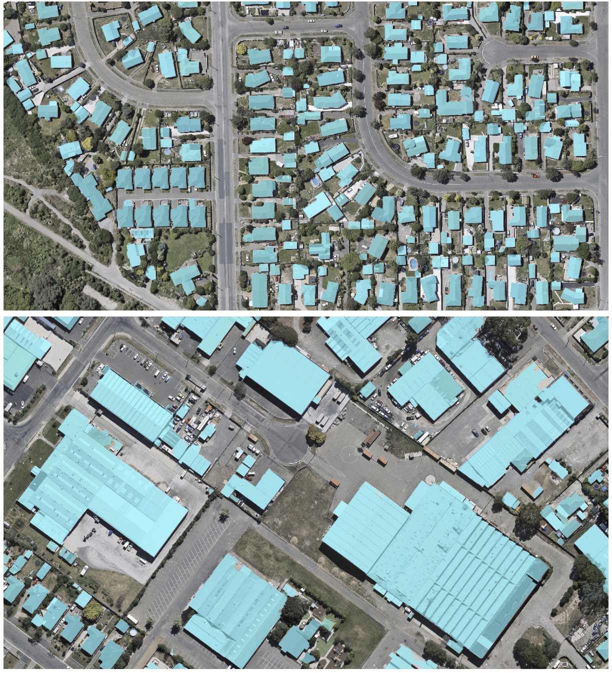


Fig. 14. Large images (with predicted mask) that recovered from 512×512 tiles. No stitching trace could be found when using FCN based methods.

D. Further prospects of our dataset

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 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 dataset, 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 14,5000 buildings and checked the model on the 4,2000 buildings. We kept all the settings of the original Mask R-CNN unchanged and run 22 hours in a single GPU. From Table 8, 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 box 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

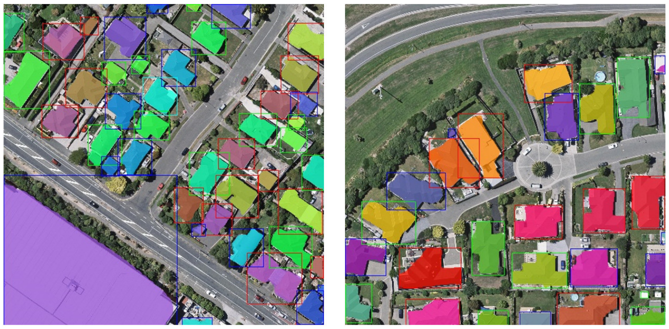


Fig. 15. Building instance segmentation using Mask R-CNN on the aerial dataset.

The second important application of our dataset is building change detection and updating. Our dataset covers an area where a 6.3-magnitude earthquake has occurred in February 2011 and rebuilt in the following years. The original aerial dataset consists of aerial images acquaint in 2016. We additionally provide a sub-dataset that consists of aerial images obtained in April 2012 that contains 12796 buildings in 20.5 km^2 (16077 buildings in the same area in 2016 dataset). By manually selecting 30 GCPs on ground surface, the sub-dataset was geo-rectified to the aerial dataset with 1.6-pixel accuracy. Fig. 16 shows two images covering the same area, where many buildings appeared or were rebuilt. This sub-dataset and the corresponding images from the original dataset are now openly provided along with building vector and raster maps.

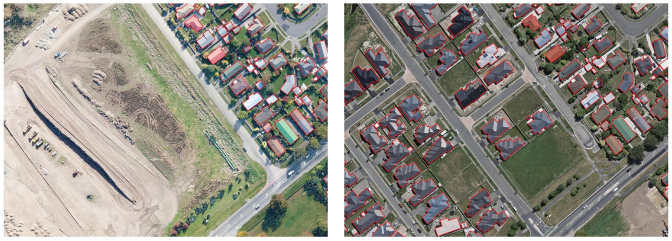


Fig. 16. Aerial images (with vector shapes) acquaint in 2012 and 2016 respectively consist of an ideal area for studying building change detection.

VI. CONCLUSION

A large sample size, accurate and multi-source dataset plays an indispensable role in developing and applying deep neural network to remote sensing applications. First, we provide an aerial and satellite building dataset, which is expected to contribute to developing and evaluating novel methods such as pixelwise segmentation, multi-source transfer learning, instance segmentation and change detection. The experiments show our aerial dataset achieved the best accuracy compared to using other existing datasets with the same FCN method. Second, we thoroughly evaluate the performance of recent studies in building extraction on the same aerial dataset 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 2-scale FCN and MLP, and our SiU-Net achieved the best accuracy. Third, as an attempt to address multi-source learning and generalization ability of deep learning, we applied radiometric augmentation in aerial dataset for pretraining, which significantly improved the prediction accuracy of applying the pre-trained model to satellite images. However, different from the satisfactory results that could be achieved in building extraction on homogenous datasets, the generalization ability of deep learning for multi-source datasets is still limited and requires to be further studied.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (41471288). The authors thank Siqi Tian, Zijie Qin, Ruojin Zhu, Chi Zhang, Yanyun Shen, Yameng Wang, Jin Liu, Dawen Yu and Shu Hu from Wuhan University, and Qi Chen from China University of Geosciences to help with preparing the dataset.

References

[1] Y. T. Liow and T. Pavlidis, "Use of Shadows for Extracting Buildings in Aerial Images," *Computer Vision Graphics & Image Processing,* 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 *International Symposium on Computer and Information Sciences*, 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 *International Conference on Signal Processing*, 2008, pp. 1219-1222.

[4] Y. Zhang, "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering," *Isprs Journal of Photogrammetry & Remote Sensing,* 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," 2008.

[6] G. Ferraioli, "Multichannel InSAR Building Edge Detection," *IEEE Transactions on Geoscience & Remote Sensing,* vol. 48, no. 3, pp. 1224-1231, 2010.

[7] A. V. Dunaeva and F. A. Kornilov, "Specific shape building detection from aerial imagery in infrared range," p. 84&ndash;100, 2017.

[8] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Improved Building Detection Using Texture Information," *ISPRS - International Archives of the Photogrammetry,* vol. XXXVIII-3/W22, no. XXXVIII-3/W22, pp. 143-148, 2011.

[9] H. E. Chunyang, "Incorporation of Texture and Structure Information for Urban Building Detection by Using Landsat7 ETM~+ Panchromatic Image," *Editorial Board of Geomatics & Information Science of Wuhan University,* vol. 29, no. 9, pp. 800-804, 2004.

[10] D. Chen, S. Shang, and C. Wu, "Shadow-based Building Detection and Segmentation in High-resolution Remote Sensing Image," *Journal of Multimedia,9,1(2014-01-01),* vol. 9, no. 1, 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 *IGARSS 2015 - 2015 IEEE International Geoscience and Remote Sensing Symposium*, 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 Sensing,* 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 *IGARSS 2017 - 2017 IEEE International Geoscience and Remote Sensing Symposium*, 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 Transactions on Geoscience & Remote Sensing,* vol. 55, no. 2, pp. 645-657, 2016.

[15] J. Yuan, "Learning Building Extraction in Aerial Scenes with Convolutional Networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. PP, no. 99, pp. 1-1, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097-1105.

[17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science,* 2014.

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," pp. 770-778, 2015.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[21] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition*, 2010, pp. 2528-2535.

[22] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. PP, no. 99, pp. 1-1, 2017.

[24] H. Noh, S. Hong, and B. Han, "Learning Deconvolution Network for Semantic Segmentation," pp. 1520-1528, 2015.

[25] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing, 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 Sensing,* 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 Transactions on Geoscience & Remote Sensing,* vol. PP, no. 99, pp. 1-13, 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 *IGARSS 2017 - 2017 IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 3226-3229.

[29] G. Wu *et al.*, "Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks," *Remote Sensing,* 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 *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255.

[31] T. Lin *et al.*, "Microsoft COCO: Common Objects in Context," vol. 8693, pp. 740-755, 2014.

[32] V. Mnih, "Machine Learning for Aerial Image Labeling," *Doctoral,* 2013.

[33] *ISPRS WG III/4. ISPRS 2D Semantic Labeling Contest.* Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[34] B. L. Saux, N. Yokoya, R. Hansch, and S. Prasad, "2018 IEEE GRSS Data Fusion Contest: Multimodal Land Use Classification [Technical Committees]," *IEEE Geoscience & Remote Sensing Magazine,* vol. 6, no. 1, pp. 52-54, 2018.

[35] J. Sherrah, "Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery," 2016.

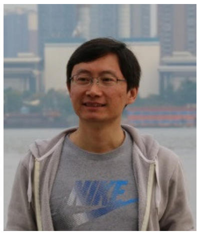
[36] *LINZ Data Service*. Available: https://data.linz.govt.nz/

[37] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Computer Vision and Pattern Recognition*, 2015, pp. 4353-4361.

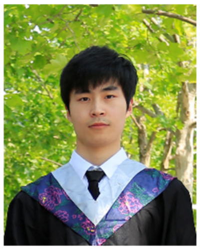
[38] Y. Lecun, *Stereo matching by training a convolutional neural network to compare image patches*. JMLR.org, 2016, pp. 2287-2318.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," pp. 1026-1034, 2015.

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017.



Shunping Ji received the Ph.D degree in photogrammetry and remote sensing from Wuhan University, China in 2007. He is currently a Professor in the School of Remote Sensing and Information Engineering, Wuhan University, China. He has co-authored more than 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 China University of Petroleum, China in 2017. He is currently pursuing the M.Sc. degree in the School of Remote Sensing and Information Engineering, Wuhan University, China. His current research interests include remote sensing, machine learning.

Meng Lu received the M.Sc. degree in earth science system from University of Buffalo, SUNY, USA, and the Ph.D. degree in Geoinformatics in University of Muenster, Germany. She joined the department of Physical Geography, Utrecht University, Utrecht, the Netherlands, as a research associate specializing in spatial data analysis, environmental modelling and geocomputation. Her research interests include geoscientific data analysis, spatiotemporal statistics, machine learning, remote sensing, environmental modelling, and health geography.