sensors

MDPI

Article

One View Per City for Buildings Segmentation in Remote-Sensing Images via Fully Convolutional Networks: A Proof-of-Concept Study

Jianguang Li [1,2,†], Wen Li [3,4,†], Cong Jin [1,2], Lijuan Yang [5] and Hui He [6,7,*]

1 College of Information and Communication Engineering, Communication University of China, Beijing 100024, China; lijianguang@cuc.edu.cn (J.L.); jincong0623@cuc.edu.cn (C.J.)
2 State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
3 Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China; wen.li@siat.ac.cn
4 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China
5 Ocean College, Minjiang University, Fuzhou 350108, China; 2611@mju.edu.cn
6 College of Information Technology, Beijing Normal University, Zhuhai 519087, China
7 Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing 100875, China
* Correspondence: hehui@bnuz.edu.cn
† These authors contributed equally to this work.

Received: 14 October 2019; Accepted: 18 December 2019; Published: 24 December 2019

Abstract: The segmentation of buildings in remote-sensing (RS) images plays an important role in monitoring landscape changes. Quantification of these changes can be used to balance economic and environmental benefits and most importantly, to support the sustainable urban development. Deep learning has been upgrading the techniques for RS image analysis. However, it requires a large-scale data set for hyper-parameter optimization. To address this issue, the concept of "one view per city" is proposed and it explores the use of one RS image for parameter settings with the purpose of handling the rest images of the same city by the trained model. The proposal of this concept comes from the observation that buildings of a same city in single-source RS images demonstrate similar intensity distributions. To verify the feasibility, a proof-of-concept study is conducted and five fully convolutional networks are evaluated on five cities in the Inria Aerial Image Labeling database. Experimental results suggest that the concept can be explored to decrease the number of images for model training and it enables us to achieve competitive performance in buildings segmentation with decreased time consumption. Based on model optimization and universal image representation, it is full of potential to improve the segmentation performance, to enhance the generalization capacity, and to extend the application of the concept in RS image analysis.

Keywords: remote-sensing; one view per city; buildings segmentation; fully convolutional network

1. Introduction

Buildings segmentation using remote-sensing (RS) images plays an important role in monitoring and modeling the process of urban landscape changes. Quantification of these changes can deliver useful products to individual users and public administrations and most importantly, it can be used to support the sustainable urban development and to balance both the economic and environmental benefits [1–5]. Via analyzing the data source of Landsat TM/ETM+ in 1990s, 2000s and 2010s, a study

estimated China's urban expansion from the urban built-up area, the area of croplands converted into urban as well as the speed of urbanization in different cities [6]. It gave clues on the relationship among urbanization, land use efficiency of urban expansion and population growth. Moreover, these factors are highly related to carbon emissions, climate change and urban environmental development that facilitate urban planning and management [7].

A large number of computational methods have been developed for RS image segmentation [8], since manual annotation of high-resolution RS images is not only time-consuming and labor-intensive but also error-prone, and therefore, to develop high-performance methods is urgent for RS image analysis. Techniques for image segmentation can be broadly grouped into semi- and full-automatic methods. Semi-automatic methods require user assistance and graph cuts is one of the most notable methods [9–12]. The method takes intensity, textures and edges of an image into consideration and after some pixels are manually localized in background, foreground or unknown regions, it addresses the problem of binary segmentation by using Gaussian mixture models [13]. Finally, one shot of object segmentation is achieved from iterative energy minimization. Wang et al. [14] integrated a graph cuts model into spectral-spatial classification of hyper-spectral images and in each smoothed probability map, the model extracted the object to a certain information class. Peng et al. [15] took advantage of a visual attention model and a graph cuts model to extract the rare-earth ore mining area information using high-resolution RS images. Notably, semi-automatic methods enable a user to incorporate prior knowledge, to validate results and to correct errors in the process of iterative image segmentation.

It is imperative to develop full-automated methods for RS image analysis, particularly when the spatial and temporal resolution of RS imaging has been considerably and continuously increased. The approaches for full-automated segmentation of RS images can be divided into conventional methods and deep learning (DL) methods in general. The former is developed based on the analysis of pixels, edges, textures and regions [8,16,17]. Hu et al. [18] designed an approach that consisted of algorithms for determination of region-growing criteria, edge-guided image object detection and assessment of edges. The approach detected image edges with embedded confidence and the edges were stored in an R-tree structure. After that, initial objects were coarsely segmented and then organized in a region adjacency graph. In the end, multi-scale segmentation was incorporated and the curve of edge completeness was analyzed. Interestingly, some methods incorporate machine learning principles and recast RS image segmentation as a pixel- or region-level classification problem [19,20]. However, parameters in most approaches are set empirically or adjusted toward high performance and thus, the generalization capacities might be restricted.

Recently, DL has revolutionized image representation [21], visual understanding [22], numerical regression [23] and cancer diagnosis [24]. Many novel methods have been developed for RS image segmentation [25–29]. Volpi and Tuia [30] presented a fully convolutional neural network (FCN) and it achieved high geometric accuracy of land-cover prediction. Kampffmeyer et al. [31] incorporated median frequency balance and uncertainty estimation which aimed to address class imbalance in semantic segmentation of small objects in urban images. Langkvist et al. [32] compared various design choices of a deep network for land use classification and the land areas were labeled with vegetation, ground, roads, buildings and water. Wu et al. [33] explored an ensemble of convolutional networks for better generalization and less over-fitting and furthermore, an alignment framework was designed to balance the similarity and variety in multi-label land-cover segmentation. Alshehhi et al. [34] proposed a convolutional network model for the extraction of roads and buildings and to improve the segmentation performance, low-level features of roads and buildings were integrated with deep features for post-processing. Vakalopoulou et al. [35] used deep features to represent image patches and support vector machine was employed to differentiate buildings from the background regions. Gao et al. [36] designed a deep residual network and it consisted of a residual connected unit and a dilated perception unit and in the post-processing stage, a morphologic operation and a tensor voting algorithm were employed. Yuan [37] proposed a deep network with a final stage that integrated activations from multiple preceding stages for pixel-wise prediction. The network introduced the

signed distance function of building boundaries as the output representation and the segmentation performance was improved.

It has achieved promising results by using DL methods for automated objects segmentation in RS images. However, DL requires considerable data for hyper-parameter optimization [38]. In the field of RS imaging, to collect sufficient images with accurately annotated labels is challenging, since a lot of objects of interest are buried in a complex background and a large area mapping [39]. To address this issue, a concept of "one view per city" (OVPC) is proposed. It explores to make the most of one RS image for parameter settings in the stage of model training, with the hope of handling the rest images of the same city by the trained model. As such, challenges could be relieved to some extent. It requires only one image per city and thus, time and labor can be reduced in the labeling of ground truth for specific purposes as well as the model training. Moreover, an algorithm is trained and tested on images from the same city and thus, intrinsic similarity between the foreground and background regions could be well explored. In fact, the concept comes from the observation that buildings of a same city in singe-source RS images illustrate similar intensity distributions. To verify its feasibility, a prove-of-concept study is conducted and five FCN models are evaluated in the segmentation of buildings in RS images. In addition, five cities in the Inria Aerial Image Labeling (IAIL) database [40] are analyzed.

The rest of this paper is organized as follows. Section 2 shows the observation that the buildings of a same city acquired by a same sensor show similar intensity distributions. Section 3 describes the involved FCN models and then introduces the data collection, experiment design, performance metrics and algorithm implementation. Section 4 demonstrates experimental results and Section 5 discusses the findings. This proof-of-concept study is concluded in Section 6.

2. One View Per City

The proposal of the concept "one view per city" (OVPC) comes from the observation that most of buildings from a same city acquired by a same sensor demonstrate a similar appearance in RS images. To show the observation, the IAIL database [40] is analyzed. Specially, the appearance of buildings is quantified with the distribution of pixel intensities in the annotated regions in RS images. As shown in Figure 1, each row stands for a city (Austin, Chicago, Kitsap County, Western Tyrol and Vienna), each column indicates the red, green or blue channel of images, and each plot shows the intensity distributions of all 36 images. Moreover, in each plot, the horizontal axis shows the intensity range ([1, 255]), and the vertical axis shows the number of pixels to each intensity value.

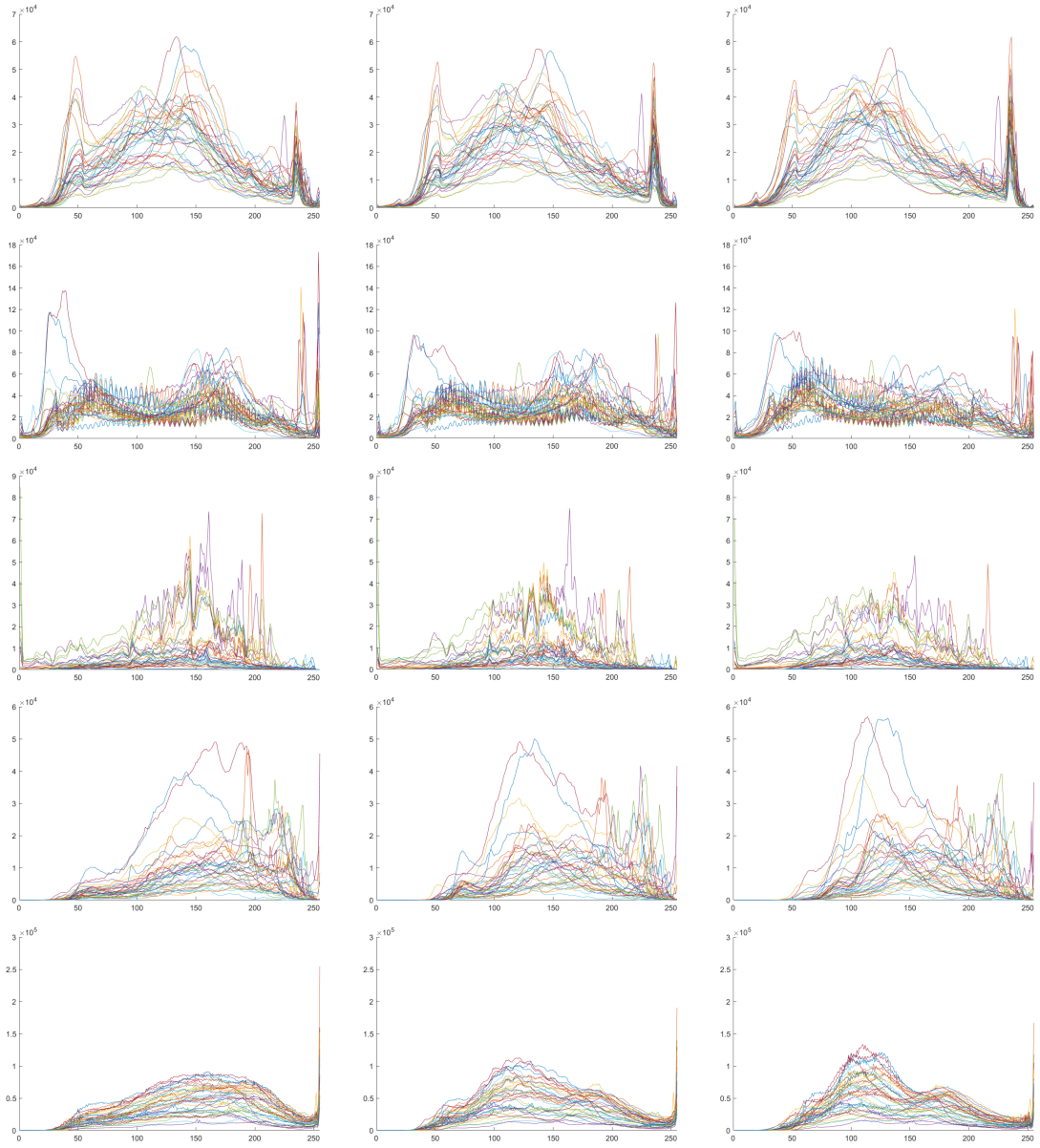

Figure 1. The intensity distribution of each city from Austin, Chicago, Kitsap County, Western Tyrol to Vienna (plots in each row) and each channel from R, G to B (plots in each column). In each plot, the horizontal axis shows the intensity range ([1, 255]) and the vertical axis shows the number of pixels to each intensity value.

Pair-wise linear correlation coefficients (LCCs) of distributions of pixel intensities are calculated. In Figure 2, each row stands for a city, each column indicates the red, green and blue channel of images and each plot shows a LCC matrix. Note that in each plot, both the horizontal and the vertical axis shows the image index. It is observed that when 0.5 is defined as the threshold of LCC values, 98.61% RS image pairs from the city Austin shows a higher correlation, followed by Vienna (97.69%), Western Tyrol (82.72%), Kitsap County (75.15%) and Chicago (57.87%).

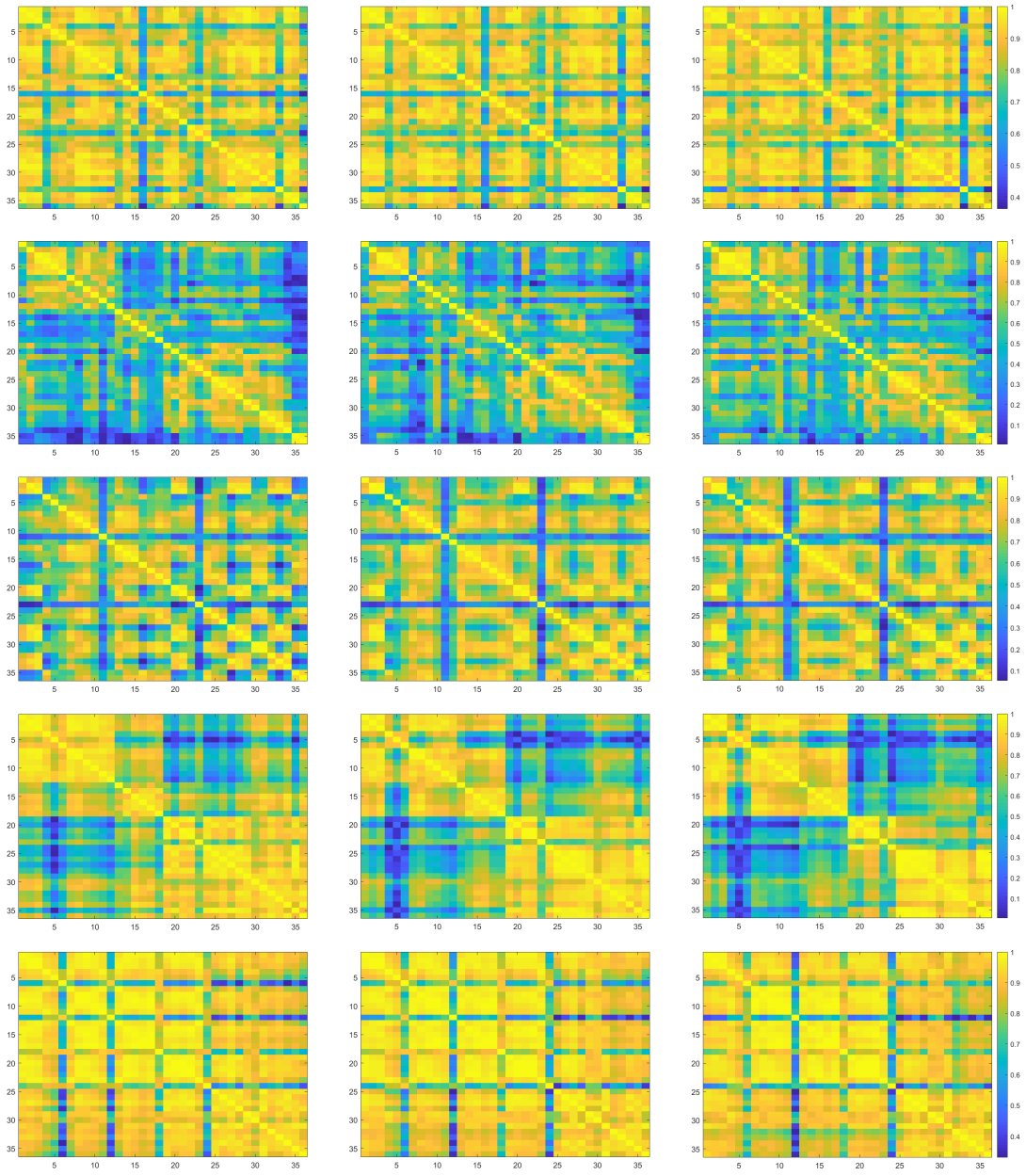

Figure 2. The pair-wise linear correlation coefficient matrix of each city from Austin, Chicago, Kitsap County, Western Tyrol to Vienna (plots in each row) and each channel from R, G to B (plots in each column). In each plot, both the horizontal and the vertical axis shows the RS image index.

The observation can be visually perceived. Figure 3 shows RS images of the 6th and 21st (**A**ustin, noted as A06 and A21), the 14th and 26th (**C**hicago, noted as C14 and C26), the 33rd and 36th (**K**itsap County, noted as K33 and K36), the 13th and 22nd (Western **T**yrol, noted as T13 and T22), and the 5th and 33rd (**V**ienna, noted as V05 and V33). It is found that these cities can be visually distinguished from each other by comparing the major appearances of buildings.

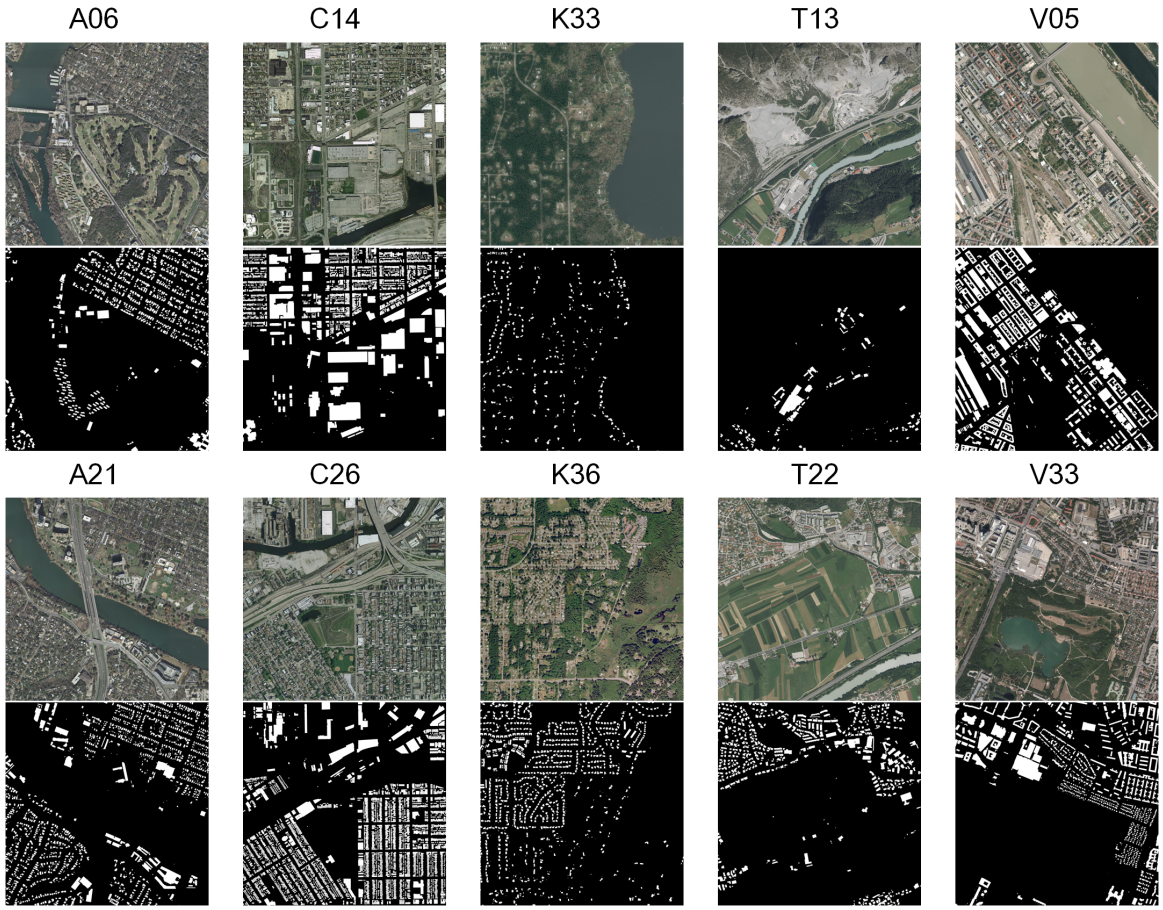

Figure 3. Visual comparison of remote-sensing images from different cities (Austin, Chicago, Kitsap County, Western Tyrol and Vienna). The binary images under each remote-sensing images correspond to the annotated labels of buildings regions.

Both the quantitative comparison (Figure 2) and the visual observation (Figure 3) suggest that the buildings of a same city illustrate a similar appearance in single-source RS images. Intuitively, this kind of information redundancy can be utilized to address the issue of limited data in DL based RS image analysis. Therefore, the proposal of OVPC might benefit pixel-wise segmentation of buildings in RS images.

3. A Proof-of-Concept Study

To verify the feasibility, a proof-of-concept study is conducted. The task is to segment *buildings* in RS images using OVPC based FCN models. In short, five FCN models are evaluated on RS images of five cities in the IAIL database [40].

3.1. Fully Convolutional Neural Network

In general, FCN architectures consist of an encoder and a decoder network symmetrically and Figure 4 illustrates a classic network, U-Net [41]. The main blocks are convolutional layers, pooling layers, up-convolutional layers and in particular, the concatenate parts propagate information from the encoder to the decoder which keeps the visual information fidelity during image restoration.

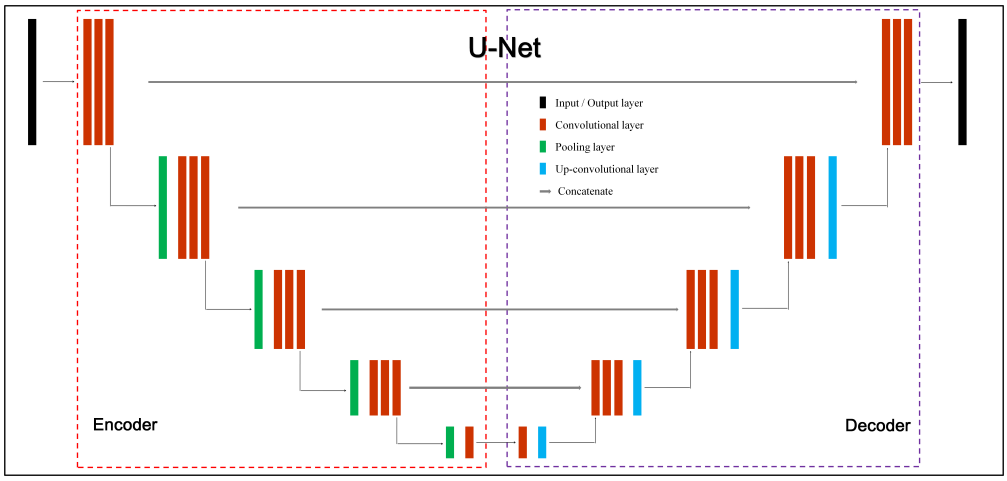

Figure 4. The architecture of U-Net. It consists of an encoder and a decoder network symmetrically. The main blocks of the network include convolutional layers, pooling layers, up-convolutional layers and concatenate parts. The concatenate parts aim to propagate original information for accurate image restoration.

The models FCN8 and FCN32 [42] modify pre-trained neural networks for pixel-wise prediction. In particular, a skip architecture is added which embodies not only deep, coarse, sematic information, but also shallow, fine, appearance messages. It is the prototype of encoder-decoder architectures for pixel-wise image segmentation. In addition, both FCN8 and FCN32 are based on the 16-layer VGG net [43] and the difference comes from the restoration position of the skip architecture.

The model SegNet [44] is made up of an encoder network, a decoder network and a pixel-wise classification layer. Topologically, the encoder network is identical to the 13-layered VGG net [43]. Its decoder makes use of pooling indices in the max-pooling step of the corresponding encoder to perform non-linear up-sampling. Because up-sampled maps are sparse, the SegNet further employs trainable convolutional filters and produces dense feature maps for pixel-wise labeling.

The model TernausNet [45] is also an encoder-decoder architecture. It employs the VGG network [43] and contains 11 sequential layers as its encoder. Comparing three weight initialization schemes, experimental results suggest that the VGG network pre-trained on ImageNet [46] achieves relatively better performance in image segmentation.

The U-Net [41] takes advantage of a contracting path for context capturing and a symmetric expanding path for precise localization. It allows for the propagation of context information to higher resolution layers by using these previously extracted feature channels. In particular, a weighted loss function is additionally used to separate the background regions between touching objects. It has been widely used in RS image analysis, such as buildings segmentation [47], damage mapping [48] and crop mapping [49].

Among the five encoder-decoder architectures, four (FCN8, FCN32, SegNet and TernausNet) utilize the VGG net [43] with different number of successive layers. The 16-layer VGG net is shown in Figure 5 and it consists of 13 convolutional layers, 5 pooling layers, 3 full-connection layers and 1 softmax layer.

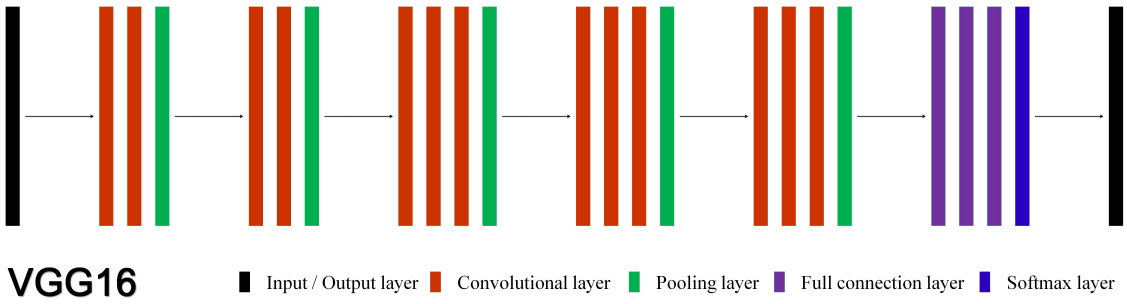

Figure 5. The 16-layer VGG net. Different colors stand for different layers. Besides the input and the output layer, the net consists of 13 convolutional layers, 5 pooling layers, 3 full-connection layers and 1 softmax layer.

3.2. Data Collection

The IAIL database [40] is analyzed. It contains a total of 360 RS images with regard to 10 cities and each city is with 36 images (https://project.inria.fr/aerialimagelabeling/). Images are formatted with GeoTIFF, the spatial resolution of aerial orthorectified color images is 0.3 m and the image matrix size is [5000, 5000]. Furthermore, manual labels of five cities are provided and images are annotated into *building* and *not building* regions. Therefore, in this study, the images with ground truth are used. The involved satellite images cover 5 cities and 405 km^2. These cities are Austin (America), Chicago (America), Kitsap County (America), Western Tyrol (Austria) and Vienna (Austria). The RS images demonstrate different density of urban settlements, various urban landscapes and illumination [50]. For instance, buildings in Kitsap County are sparsely scattered, while buildings in Chicago are densely distributed (Figure 3).

The area of buildings is further concerned. To make it clear, a parameter is defined which aims to describe the coverage of buildings regions in each image. The parameter, buildings region ratio (*brr*), is computed as the pixel number of buildings (B) over that of each image (I) as shown in Equation (1),

$$brr = \frac{||B||}{||I||}, \tag{1}$$

where $||\cdot||$ denotes the number of pixels.

Figure 6 shows the distributions of *brr* values where the x-axis indicates cities and the y-axis is *brr* values. In violin plots, solid lines denote median values and dashed lines stand for the 1st and the 3rd quartiles. The mean of *brr* values is 0.1504 ± 0.0477 (Austin), 0.2433 ± 0.0599 (Chicago), 0.0485 ± 0.0420 (Kitsap County), 0.0584 ± 0.0378 (Western Tyrol) and 0.2884 ± 0.1153 (Vienna). It indicates that the city Vienna has the highest buildings density, followed by Chicago and Austin, while RS images of Western Tyrol and Kitsap County are with buildings thinly scattered.

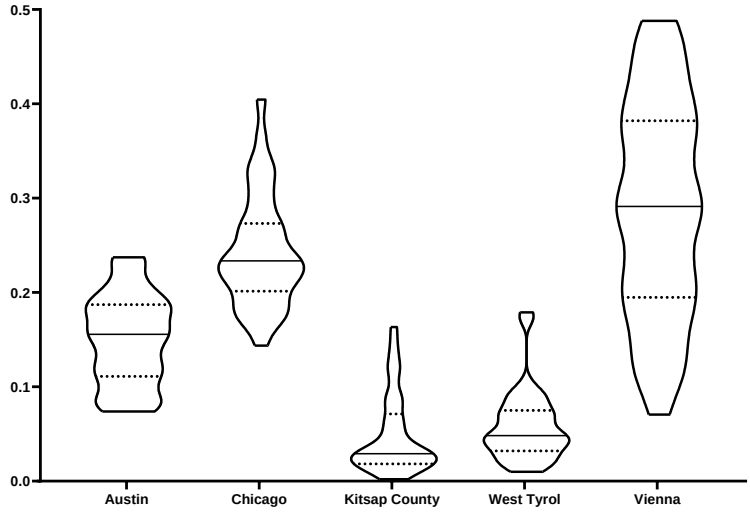

Figure 6. Distribution of *brr* values. The *brr* value reflects the coverage of buildings region in each remote-sensing image. The x-axis indicates cities and the y-axis shows *brr* values. In violin plots, solid lines denote median values and dashed lines stand for quartiles. It indicates that the city Vienna has the highest buildings density, followed by Chicago and Austin, while remote-sensing images of Western Tyrol and Kitsap County are with buildings thinly scattered.

3.3. Experiment Design

Based on the concept of OVPC, one image is used for parameter tuning and the rest images of the same city are used for testing (Table 1). Specifically, to each city, 5 images are randomly selected and each plays as the input for model training. After the model is trained, the rest 35 images are tested (intra-city test). Moreover, images from other cities (a total of 144 images) are also tested using this trained model (inter-city test).

Table 1. One view per city (OVPC) based experiment design for buildings segmentation in RS images.

Training	Image No.	Intra-City Test	Image No.	Inter-City Test	Image No.
Austin	1	Austin	35	Other cities	36 × 4
Chicago	1	Chicago	35	Other cities	36 × 4
Kitsap County	1	Kitsap County	35	Other cities	36 × 4
West Tyrol	1	West Tyrol	35	Other cities	36 × 4
Vienna	1	Vienna	35	Other cities	36 × 4

3.4. Performance Metrics

Two metrics are used to evaluate the segmentation performance. One is segmentation accuracy (ACC) and it measures the percentage of correctly classified pixels (Equation (2)). The other metric is intersection over union (IoU) which is the ratio of the number of pixels labeled as buildings in both the prediction and the reference divided by the number of pixels labeled as pixel in the prediction or the reference (Equation (3)). Given the prediction result P and the reference S, the metrics of ACC and IoU are respectively defined as

$$ACC = \frac{|P \cap S|}{|S|}, \tag{2}$$

and

$$IoU = \frac{|P \cap S|}{|P \cup S|}, \tag{3}$$

where $|\cdot|$ denotes the number of foreground pixels in the binary images.

3.5. Algorithm Implementation

Algorithms were run with Linux system (Ubuntu 16.04.10) on 3 workstations. The workstations are all embedded with 16 Intel(R) Xeon(R) CPU (3.00 GHz), 64 GB DDR4 RAM and one GPU card (TITAN X (Pascal), 12 GB). FCN models are available online (https://github.com/divamgupta/image-segmentation-keras). Deep networks are implemented with Keras (https://keras.io/) (Python 2.7.6) and the backend is Tensorflow (https://www.tensorflow.org/).

In detail, images are cropped to $[2^{12}, 2^{12}]$ and the boundary of the original images is discarded. A total of 15,625 patches are extracted from each cropped image. Note that 2^5 pixels are overlapped between successive patches. To deep models, the size of input patches is defined as $[2^7, 2^7]$ and the pixel intensities in each patch are linearly scaled to [0, 1]. Parameters are set as follows. The binary cross entropy is set as the loss function, Adam as the optimizer, the learning rate is 10^{-4}, the batch size is 2^6 and the number of epochs is 10^2. Other parameters are set as default. In addition, neither fine-tuning nor data augmentation are used.

4. Results

4.1. Intra-City Test

Table 2 shows the mean and standard deviation of ACC values (mean $\pm$ std, %) and the highest mean values in each intra-city test are bold-faced. In general, it shows that major building regions are correctly predicted (>80%) except the SegNet on Chicago (79.41 $\pm$ 7.18%). Based on the performance analysis of five FCN models with regard to different cities, it is found that the buildings in Kitsap County might be the easiest to be segmented (>94%), followed by Austin (≈90%), West Tyrol and Vienna, and the last one is Chicago. Notably, Unet achieves the overall best segmentation results.

Table 2. Segmentation accuracy (ACC) values of OVPC based buildings segmentation in RS images (mean $\pm$ std, %). The highest mean values in each intra-city test are in bold.

	Austin	**Chicago**	**Kitsap County**	**West Tyrol**	**Vienna**
FC8	89.89 $\pm$ 3.87	83.20 $\pm$ 5.77	94.56 $\pm$ 6.06	86.17 $\pm$ 12.12	82.45 $\pm$ 6.77
FC32	89.45 $\pm$ 3.72	82.83 $\pm$ 4.76	**95.91 $\pm$ 3.84**	90.43 $\pm$ 7.63	83.78 $\pm$ 5.47
SegNet	90.08 $\pm$ 3.66	79.41 $\pm$ 7.18	94.04 $\pm$ 6.68	90.25 $\pm$ 6.85	84.07 $\pm$ 5.78
TernausNet	**91.54 $\pm$ 3.57**	81.16 $\pm$ 7.78	95.17 $\pm$ 5.75	88.67 $\pm$ 10.45	85.91 $\pm$ 4.24
Unet	91.07 $\pm$ 3.78	**84.41 $\pm$ 5.41**	95.62 $\pm$ 4.70	**95.70 $\pm$ 3.11**	**87.05 $\pm$ 4.20**

Table 3 shows the IoU values (mean $\pm$ std, %) and the highest mean values in each intra-city test are in bold. It is found that the performance of buildings segmentation is moderate. The worst result is from FC32 on West Tyrol (54.22 $\pm$ 9.75%) and the best is from Unet on Vienna (72.11 $\pm$ 6.43%). Based on the performance analysis with regard to different cities, it is observed that the buildings in Vienna and Austin might be much easier to be outlined (>64%). It is worth mentioning that Unet gets best results on three cities (Chicago, West Tyrol and Vienna) and in general, superior performance over other FCN models (>63%).

Figure 7 shows the values of performance metrics with regard to different models and cities. The rows from top to bottom denote FCN models of FCN8, FCN32, SegNet, TernausNet and U-Net, and the columns from left to right correspond to metrics of ACC and IoU, respectively. In each box and whisker plot shown with the minimum and maximum values, the x-axis indicates the cities, the y-axis denotes the metric values (%).

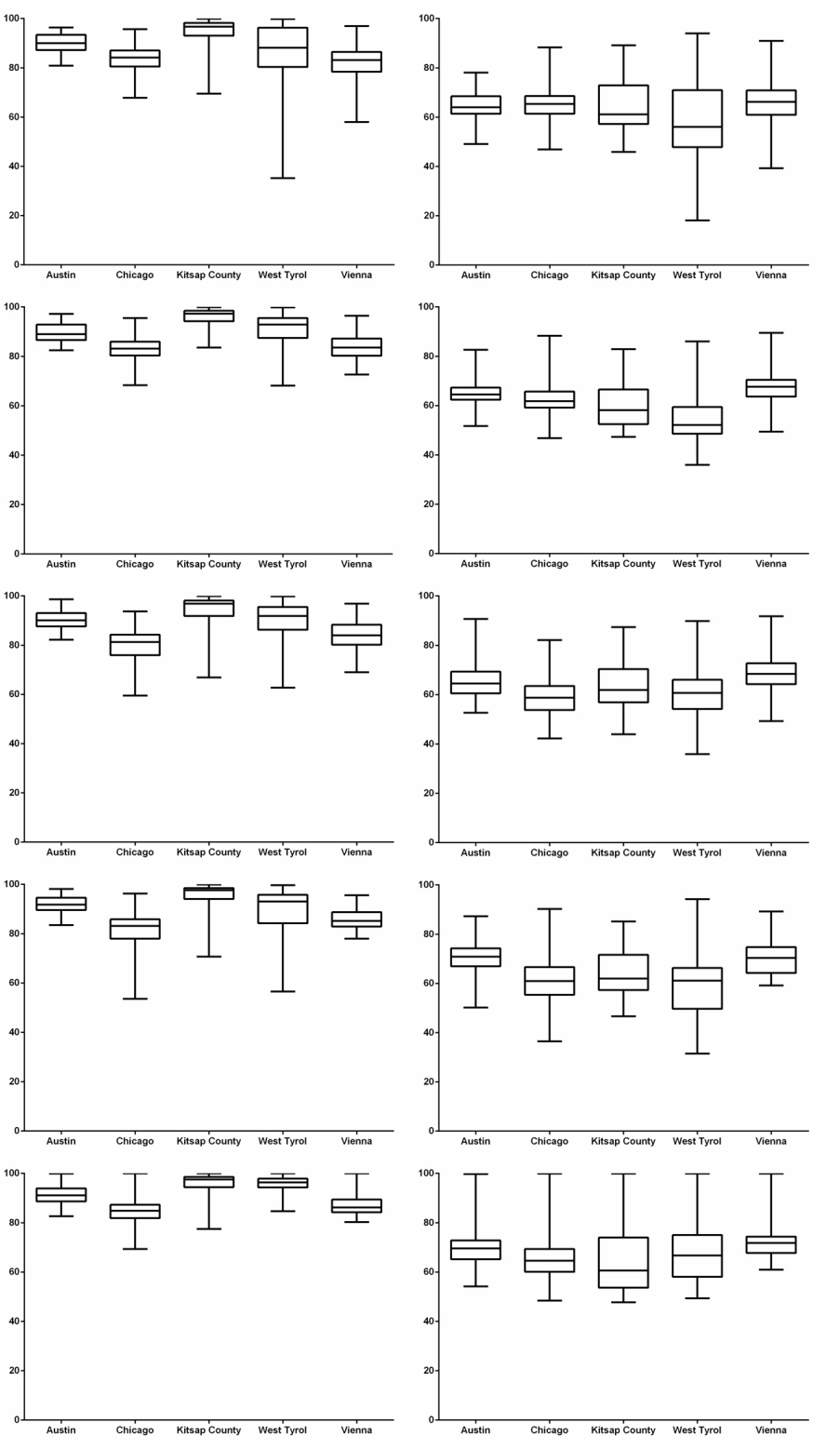

Figure 7. Comparison of fully convolutional neural network (FCN) models on OVPC based buildings segmentation using RS images (intra-city test). Each row indicates results from a FCN model from FCN8, FCN32, SegNet, TernausNet to U-Net, and each column shows a performance metric from ACC to intersection over union (IoU). In each box and whisker plot, the x-axis denotes the cities and the y-axis stands for metric values (%).

Table 3. IoU values of OVPC based buildings segmentation in RS images (mean ± std, %). The highest mean values in each intra-city test are in bold.

	Austin	Chicago	Kitsap County	West Tyrol	Vienna
FC8	64.53 ± 5.69	64.86 ± 7.45	**64.07 ± 9.27**	58.69 ± 15.53	65.67 ± 8.13
FC32	64.86 ± 5.01	62.19 ± 5.83	59.47 ± 7.96	54.22 ± 9.75	66.90 ± 6.92
SegNet	64.96 ± 6.24	58.47 ± 7.29	63.14 ± 8.68	61.41 ± 10.68	68.34 ± 7.73
TernausNet	**70.08 ± 6.47**	60.47 ± 8.94	63.69 ± 8.34	58.55 ± 11.86	69.97 ± 6.30
Unet	69.60 ± 7.01	**65.36 ± 8.58**	63.31 ± 11.66	**67.63 ± 10.99**	**72.11 ± 6.43**

At present, tens of studies are conducted on the IAIL database [50–58]. Several studies [50–54] follow a common practice as [40] that the first 5 images of each city are used for testing. In other words, there are 31 images for training and the rest 5 images for testing regarding each city. Current outcomes on the IAIL database are summarized in Table 4. The highest metric values are in bold. It is observed that the generative adversarial network [59] with spatial and channel attention mechanisms (GAN-SCA) [54] achieves the best overall segmentation results.

Table 4. Recent outcomes on the Inria Aerial Image Labeling (IAIL) database based on intra-city test (mean value, %).

		Austin	Chicago	Kitsap County	West Tyrol	Vienna
MLP in [40] (baseline)	ACC	94.20	90.43	98.92	96.66	91.87
	IoU	61.20	61.30	51.50	57.95	72.13
FCN in [40]	ACC	92.22	88.59	98.58	95.83	88.72
	IoU	47.66	53.62	33.70	46.86	60.60
Mask R-CNN in [50]	ACC	94.09	85.56	97.32	98.14	87.40
	IoU	65.63	48.07	54.38	70.84	64.40
Two-level U-Net in [51]	ACC	96.69	92.40	99.25	98.11	93.79
	IoU	77.29	68.52	72.84	75.38	78.72
Multi-task SegNet in [52]	ACC	93.21	**99.25**	97.84	91.71	**96.61**
	IoU	76.76	67.06	**73.30**	66.91	76.68
MSMT in [53]	ACC	95.99	92.02	99.24	97.78	92.49
	IoU	75.39	67.93	66.35	74.07	77.12
GAN-SCA in [54]	ACC	**97.26**	93.32	**99.30**	**98.32**	94.84
	IoU	**81.01**	**71.73**	68.54	**78.62**	**81.62**
Unet [a]	ACC	94.34	88.72	98.67	97.52	92.48
	IoU	72.48	68.14	67.50	72.16	74.35
Unet (OVPC)	ACC	91.07	84.41	95.62	95.70	87.05
	IoU	69.60	65.36	63.31	67.63	72.11

[a] The Unet is with the same network architecture and parameters as the OVPC based Unet, while it is tested with the first 5 images of each city and trained with the other 31 images.

Table 4 reports the results of OVPC based Unet. To RS images of one city, the model uses 1 image for training and the rest 35 for testing. Comparing the results of OVPC based Unet with the recent best outcome (GAN-SCA [54]), the ACC drop is between 2.62% (West Tyrol) and 4.60% (Chicago) and IoU decrease is between 5.32% (Kitsap County) and 11.41% (Austin). On the other hand, the Unet achieves competitive or superior performance over the baseline of MLP network [40]. It is found that

ACC values slightly reduce from 0.96% (West Tyrol) to 6.02% (Chicago), while IoU values obviously increase for the cities of Austin (8.40%), Kitsap County (11.81%) and West Tyrol (9.68%).

Table 4 also lists the performance of Unet that takes 31 images of each city for training. Note that the architecture and parameter settings of implemented Unet models are the same. The comparison indicates that using more images of a city for training leads to slight increase (2% to 5%) on both ACC and IoU values.

4.2. Inter-City Test

The generalization capabilities of deep networks on IAIL database are concerned [40,60]. The results of inter-city test using OVPC based Unet are shown in Tables 5 and 6. The ACC values indicate that the segmentation performance decreases when using the model trained on one city to predict another city. Specifically, the model trained on West Tyrol gets the minimum decrease when it tests on other cities ($\leq$5.28%), followed by the model trained on Chicago ($\leq$5.72), Kitsap County ($\leq$5.80), Vienna ($\leq$7.42) and Austin ($\leq$8.60).

Table 5. ACC values of OVPC based Unet on buildings segmentation in RS images (mean $\pm$ std, %). The inter-city test results are in bold.

	Austin	**Chicago**	**Kitsap County**	**West Tryol**	**Vienna**
Austin	**91.07 $\pm$ 3.78**	88.72 $\pm$ 6.72	82.47 $\pm$ 8.58	84.13 $\pm$ 7.32	90.04 $\pm$ 4.33
Chicago	82.68 $\pm$ 6.82	**84.41 $\pm$ 5.41**	78.69 $\pm$ 8.36	80.25 $\pm$ 6.53	82.65 $\pm$ 4.75
Kitsap County	92.52 $\pm$ 6.43	91.75 $\pm$ 7.32	**95.62 $\pm$ 4.70**	93.48 $\pm$ 3.79	89.82 $\pm$ 8.09
West Tryol	93.21 $\pm$ 8.52	90.42 $\pm$ 7.12	94.85 $\pm$ 5.84	**95.70 $\pm$ 3.11**	91.46 $\pm$ 8.42
Vienna	82.1 $\pm$ 5.47	82.78 $\pm$ 7.38	78.49 $\pm$ 6.12	80.54 $\pm$ 6.79	**85.91 $\pm$ 4.24**

Comparing the IoU values in Table 6 demonstrates that the performance decreases when using the model trained on one city to predict another city. Specifically, the model trained on Chicago gets a relatively small decrease ($\leq$7.82%), followed by the model trained on Kitsap County ($\leq$10.48), Austin ($\leq$14.87), West Tyrol ($\leq$15.53) and Vienna ($\leq$16.88).

Table 6. IoU values of OVPC based Unet on buildings segmentation in RS images (mean $\pm$ std, %). The inter-city test results are in bold.

	Austin	**Chicago**	**Kitsap County**	**West Tryol**	**Vienna**
Austin	**69.60 $\pm$ 7.01**	62.58 $\pm$ 10.23	58.62 $\pm$ 12.35	60.27 $\pm$ 14.23	54.73 $\pm$ 15.29
Chicago	60.18 $\pm$ 10.32	**65.36 $\pm$ 8.58**	57.54 $\pm$ 14.46	59.73 $\pm$ 10.89	63.08 $\pm$ 13.65
Kitsap County	59.73 $\pm$ 10.54	56.26 $\pm$ 16.46	**63.31 $\pm$ 11.66**	61.35 $\pm$ 12.36	52.83 $\pm$ 15.23
West Tryol	61.42 $\pm$ 12.49	58.76 $\pm$ 13.72	64.45 $\pm$ 14.36	**67.63 $\pm$ 10.99**	52.10 $\pm$ 16.47
Vienna	64.48 $\pm$ 8.25	66.29 $\pm$ 9.21	55.23 $\pm$ 10.23	57.58 $\pm$ 8.82	**72.11 $\pm$ 6.43**

4.3. Time Consumption

Different implementations lead to various time consumption. In this study, one epoch takes about 3.2 min to the OVPC based Unet and it costs 5.4 h to complete the whole model training. While using the 31 RS images as the input for training (Unet ^{a}), the time per epoch increases to 21.0 min and the complete training takes about 35.0 h. In addition, to predict a whole RS image takes $\approx$6.2 min, including thresholding and small patch merging.

5. Discussion

This study proposes the concept of OVPC. It explores to relieve the requirement of a large-scale data set to some degree in DL based RS image analysis. The concept proposal is inspired by the observation that buildings of a same city share similar appearance in single-source RS images. This study illustrates the observation qualitatively (Figure 2) and quantitatively (Figure 3). Furthermore, the concept is verified on buildings segmentation in RS images via DL methods and five FCN models are evaluated on RS images of five cities in the IAIL database. At last, its pros and cons are analyzed through intra-city test, inter-city test and time consumption.

The building regions from a same city shares similar intensity distribution in RS images. The quantitative analysis (pair-wise LCCs) of intensity histograms indicates that building regions between the RS images from Austin correlates strongly, followed by the images from Vienna (Figure 2). The similar appearance can also be observed from visual comparison (Figure 3). Therefore, the proposal of the concept is reasonable and it is possible to use the similarity in intensity distributions in RS image analysis. And further, it might be able to reduce time and labor in data annotation, algorithm design and parameter optimization.

In this proof-of-concept study, intra-city test shows that the OVPC-based Unet achieves superior performance over other networks (Tables 2 and 3), and its generalization capacity is inadequate as shown in inter-city test (Tables 5 and 6). The result of this study is similar to the findings in [60] which suggests the Unet architecture is well suited for image dense labeling, while outcome of the cross-city test is not satisfactory yet (ACC $\approx$ 95% and IoU $\approx$ 73%). In particular, in this study, the intra-city test shows that FCN models achieve moderate to excellent performance. The metric ACC indicates that RS images are correctly portioned into *buildings* and *not buildings* (>80%), in particular in the images from Kitsap County and Austin (Table 2), while the metric IoU shows that background regions are misclassified into *buildings* regions and several values are less than 60%, such as SegNet on Chicago (Table 3). Unsurprisingly, the inter-city test finds out that it is challenging to accurately and precisely isolate *buildings* regions of one city by using an OVPC based Unet model trained on another city (Tables 5 and 6). In detail, ACC values show slight decrease (Table 5), while IoU values reveal around 10% drop in buildings segmentation (Table 6).

To OVPC-based deep models, the reason of promising results in intra-city test mainly comes from the high-performance image representation of deep networks and these networks can represent complex patterns with hierarchical features. Moreover, these models have demonstrated capacities of pixel-wise semantic segmentation in various fields, such as computer vision [46] and biomedical imaging [42,44,45]. In particular, OVPC utilizes a large number of patches (i.e., 15,625) from one image as the input of deep networks and then, the information redundancy of building appearance is further used for the segmentation of *buildings* in other images from the same city. On the other hand, reasons for moderate IoU values are manifold. First, OVPC decreases the capacity of image representation of DL models due to limited training samples. Furthermore, the area of *buildings* regions over the RS image (i.e., *brr*) is tiny, such as 0.0485 ± 0.0420 of Kitsap County (Figure 6), that dramatically imposes difficulties on deep networks to learn effective representation. Second, some key parameters, such as loss function, should be fine-tuned or carefully designed [60]. From the technical point of view, data augmentation, batch normalization and transfer learning can be further integrated to improve the segmentation performance. In addition, RS image segmentation is a long-standing problem. Due to the unique cultures of western countries, buildings in RS images are distributed with different sizes and shapes. For instance, the buildings in Kitsap County are scattered, while buildings in Chicago are densely distributed and most buildings are small in size [54]. Furthermore, boundaries between buildings are ambiguous that makes accurate segmentation challenging.

This study suggests that OVPC is beneficial to RS image analysis. It requires one RS image for model training and thereby, the time and labor in manual annotation can be reduced. To annotate a large scale of images, in particular high-resolution RS images, is always an expensive task, and cross checking should be carried out to minimize the risk of false annotation [61]. To address the

challenge, few-shot learning becomes a hot topic in RS image classification [62–64]. Impressively, Song and Xu explored zero-shot learning for automatic target recognition in synthetic aperture radar (SAR) images [65]. Moreover, using one single RS image for model training might save computing resources and decrease time cost in model training. Under the context of a GPU card with 12 GB memory, the experimental design, model implementation and time cost should be fully considered when a large number of samples are as input for training. Based on the IAIL database, when using the entire data set as the input, one epoch would last more than 2.5 h which is inconceivable [50]. In this study, one epoch takes about 2 to 5 min dependant on the FCN model and subsequently, a total of 3.5 to 8.5 h to complete the whole model training. Based on the Unet with same architecture and parameter settings, the time cost is further compared when using different numbers of images as the input for model training. It finds out that one epoch takes about 3.2 min to the OVPC based Unet and ≈21.0 min to the Unet with 31 images as its input. In other words, the proposed OVPC based Unet achieves competitive performance with dramatically decreased time consumption. More importantly, the performance of OVPC based RS image analysis could be further improved when advanced networks are used, which can be observed by comparing the results in Tables 2 and 3 with the recent outcomes in Table 4. This study uses original networks, such as U-Net [41], and advanced networks [50–54] can improve the segmentation results (Table 4). In detail, multi-task SegNet [52] embeds a multi-task loss that can leverage multiple output representation of the segmentation mask and meanwhile bias the network to focus more on pixels near boundaries; MSMT [53] is a multi-task multi-stage network that can handle both semantic segmentation and geolocalization using different loss functions in a unified framework; and GAN-SCA [54] integrated spatial and channel attention mechanisms into a generative adversarial network [59]. At last, the proposed concept can be further extent to different types of RS images and applications. RS image segmentation is indispensable to measure urban metrics [66,67], to monitor landscape changes [68] and to model the pattern and extent of urban sprawl [69]. It is also important to define urban typologies [5], to classify land use [4], to manage urban environment [2] and to support sustainable urban development [6,7]. While for accurate decision making, diverse techniques should be involved [70–74], such as LiDAR and aerial imagery.

On further improving the performance of OVPC based buildings segmentation in RS images, additional techniques could be considered. Above all, the concept requires images should be acquired from a same imaging sensor. To enhance its generalization capacity, universal image representation is indispensable which aims to transform the source and the target images into a same space. For instance, Zhang et al. [75] improved the Kalman filter and harmonized multi-source RS images for summer corn growth monitoring. Notably, generative adversarial network [59] has been used to align both panchromatic and multi-spectral images for data fusion [76]. These methods provide insights on how to generalize the proposed concept into multi-source RS image analysis. Moreover, data augmentation is helpful to enhance representation capacity of deep networks (https://github.com/aleju/imgaug). Data transformation, shape deformation and other various distortions can be used to represent buildings characteristics. Attention can also be paid to architecture design, batch normalization and parameter optimization. Besides, transfer learning is suggested to enhance network performance [77] and it requires domain adaption to balance the data distributions of the source and the target domain [78]. In addition, except for the appearance, it is potential to model buildings from shape and texture and to enrich our understanding of urban buildings in RS images. Last but not the least, post-processing strategies can be employed and prior knowledge and empirical experiences become helpful.

This study has some limitations. At first, the pros and cons of the concept OVPC are not explicitly revealed. It is better to use each of the 180 RS images (36 images per city × 5 cities) for OVOC based buildings segmentation, while that would cost more than 1000 h for model training (≈6 h per experiment × 180 experiments) to one FCN model. Secondly, it is also interesting to compare the OVPC based approaches with the multi-view based approaches and definitely, time consumption would be dramatically increased. Fortunately, results of several multi-view based approaches [50–54] are available for comparison as shown in Table 4. In addition, this study focuses on one database and

the five cities show unique characteristics in urban environment among western cities (America and Austria), while large databases with global cities [61,79] would be more general.

6. Conclusions

This paper proposes the concept of "one view per city" and conducts a proof-of-concept study on the segmentation of buildings in remote-sensing images. Five deep networks are evaluated on images of five cities in the Inria Aerial Image Labeling database. Experimental results suggest that the concept can be explored to decrease the number of images for model training and it enables us to achieve competitive performance in buildings segmentation with decreased time cost. In addition, several techniques to improve and to extend the concept for remote-sensing image analysis are suggested. The proposed concept can relieve the challenge of large-scale data annotation in deep learning based remote-sensing image segmentation. It might pave the way for multi-source remote-sensing image analysis.

Author Contributions: J.L. and W.L. designed and performed the experiments. C.J. and L.Y. participated in data collection. H.H. conceived of the idea and drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Hui He thanks the grant support from Characteristic Innovation (NATURAL SCIENCE) project of Department of Education of Guangdong Province (no. 2017KTSCX207) and Lijuan Yang thanks the grant support from Training Program of National Research Project of Minjiang University (no. MYK19029).

Acknowledgments: The authors would like to thank the editor and reviewers for their valuable advices that have helped to improve the paper quality. Thanks are also given to those people who have participated or are participating in open data sharing program and people who have shared codes, deep models or pre-trained deep models online in the scientific research community.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RS	Remote-Sensing
FCN	Fully Convolutional Network
DL	deep learning
OVPC	One View Per City
IAIL	Inria Aerial Image Labeling
LCC	Linear Correlation Coefficients
brr	buildings region ratio
ACC	Accuracy
IoU	Intersections over Union
SAR	Synthetic Aperture Radar

References

1. Camps-Valls, G.; Tuia, D.; Gomez-Chova, L.; Jimenez, S.; Jimenez, J. Remote sensing image processing. *Synth. Lect. Image Video Multimedia Process.* **2011**, *5*, 1–192. [CrossRef]
2. Musse, M.A.; Barona, D.A.; Rodriguez, L.M.S. Urban environmental quality assessment using remote sensing and census data. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *71*, 95–108. [CrossRef]
3. Pathiranage, I.S.S.; Kantakumar, L.N.; Sundaramoorthy, S. Remote sensing data and SLEUTH urban growth model: As decision support tools for urban planning. *Chin. Geogr. Sci.* **2018**, *28*, 274–286. [CrossRef]
4. Wellmann, T.; Haase, D.; Knapp, S.; Salbach, C.; Selsam, P.; Lausch, A. Urban land use intensity assessment: The potential of spatio-temporal spectral traits with remote sensing. *Ecol. Indic.* **2018**, *85*, 190–203. [CrossRef]
5. Valero Medina, J.A.; Lizarazo Salcedo, I.A.; Elsner, P. Topological challenges in multispectral image segmentation. *Tecnura* **2014**, *18*, 136–149.

6. Wang, L.; Li, C.; Ying Q.; Cheng, X.; Wang, X.; Li, X.; Hu, L.; Liang, L.; Yu, L.; Huang, H.; et al. China's urban expansion from 1990 to 2010 determined with satellite remote sensing. *Chin. Sci. Bull.* **2012**, *57*, 2802–2812. [CrossRef]
7. Zhang, Z.; Liu, F.; Zhao, X.; Wang, X.; Shi, L.; Xu, J.; Yu, S.; Wen, Q.; Zuo, L.; Li, Y.; et al. Urban expansion in China based on remote sensing technology: A review. *Chin. Geogr. Sci.* **2018**, *28*, 727–743. [CrossRef]
8. Vivek, D.; Zhang, Y.; Zhong, M. A review on image segmentation techniques with remote sensing perspective. In Proceedings of the International Society for Photogrammetry and Remote Sensing, Vienna, Austria, 5–7 July 2010; Vol. XXXVIII, Part 7A, pp. 31–42.
9. Boykov, Y.; Funka-Lea, G. Graph cuts and efficient ND image segmentation. *Int. J. Comput. Vis.* **2006**, *70*, 109–131. [CrossRef]
10. Yu, S.; Wu, S.; Zhuang, L.; Wei, X.; Mak, S.; Neb, D.; Hu, J.; Xie, Y. Efficient segmentation of a breast in B-mode ultrasound tomography using three-dimensional GrabCut (GC3D). *Sensors* **2017**, *17*, 1827. [CrossRef]
11. Tarabalka, Y.; Rana, A. Graph-cut-based model for spectral-spatial classification of hyperspectral images. In Proceedings of the IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 3418–3421.
12. Rother, C.; Kolmogorov, V.; Blake, A. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314. [CrossRef]
13. Reynolds, D.A.; Rose, R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 72–83. [CrossRef]
14. Wang, Y.; Song, H.; Zhang, Y. Spectral-spatial classification of hyperspectral images using joint bilateral filter and graph cut based model. *Remote Sens.* **2016**, *8*, 748. [CrossRef]
15. Peng, Y.; Zhang, Z.; He, G.; Wei, M. An improved GrabCut method based on a visual attention model for rare-earth ore mining area recognition with high-resolution remote sensing images. *Remote Sens.* **2019**, *11*, 987. [CrossRef]
16. Huang, G.; Song, Z.; Zhang, S.; Zhu, J. A fast marine sewage detection method for remote-sensing image. *Comput. Appl. Math.* **2018**, *37*, 4544–4553. [CrossRef]
17. Wang, Z.; Jensen, J.R.; Im, J. An automatic region-based image segmentation algorithm for remote sensing applications. *Environ. Model. Softw.* **2010**, *25*, 1149–1165. [CrossRef]
18. Hu, Y.; Chen, J.; Pan, D.; Hao, Z. Edge-guided image object detection in multiscale segmentation for high-resolution remotely sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4702–4711. [CrossRef]
19. Csillik, O. Fast segmentation and classification of very high resolution remote sensing data using SLIC superpixels. *Remote Sens.* **2017**, *9*, 243 . [CrossRef]
20. Zanotta, D.C.; Zortea, M.; Ferreira, M.P. A supervised approach for simultaneous segmentation and classification of remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 162–173. [CrossRef]
21. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
22. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
23. Yu, S.; Wu, S.; Wang, L.; Jiang, F.; Xie, Y.; Li, L. A shallow convolutional neural network for blind image sharpness assessment. *PLoS ONE* **2017**, *12*, 0176632. [CrossRef]
24. Zou, L.; Yu, S.; Meng, T.; Zhang, Z.; Liang, X.; Xie, Y. A technical review of convolutional neural network-based mammographic breast cancer diagnosis. *Comput. Math. Methods Med.* **2019**, *2019*, 6509357. [CrossRef]
25. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
26. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016; pp. 1835–1838.
27. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
28. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
29. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]

30. Volpi, M.; Tuia, D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 881–893. [CrossRef]
31. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
32. Langkvist, M.; Kiselev, A.; Alirezaie, M.; Loutfi, A. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sens.* **2016**, *8*, 329. [CrossRef]
33. Wu, G.; Guo, Y.; Song, X.; Guo, Z.; Zhang, H.; Shi, X.; Shibasaki, R.; Shao, X. A stacked fully convolutional networks with feature alignment framework for multi-label land-cover segmentation. *Remote Sens.* **2019**, *11*, 1051. [CrossRef]
34. Alshehhi, R.; Marpu, P. R.; Woon, W. L.; Dalla Mura, M. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
35. Vakalopoulou, M.; Karantzalos, K.; Komodakis, N.; Paragios, N. Building detection in very high resolution multispectral data with deep learning features. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 1873–1876.
36. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
37. Yuan, J. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2793–2798. [CrossRef]
38. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *J. Appl. Remote Sens.* **2017**, *11*, 042609. [CrossRef]
39. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
40. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing, Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
41. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Germany, 2015; pp. 234–241.
42. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
43. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
44. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv* **2015**, arXiv:1511.00561.
45. Iglovikov, V.; Shvets, A. TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation. *arXiv* **2018**, arXiv:1801.05746.
46. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
47. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
48. Bai, Y.; Mas, E.; Koshimura, S. Towards operational satellite-based damage-mapping using U-net convolutional network: A case study of 2011 Tohoku Earthquake-Tsunami. *Remote Sens.* **2018**, *10*, 1626. [CrossRef]
49. Wei, S.; Zhang, H.; Wang, C.; Wang, Y.; Xu, L. Multi-temporal SAR data large-scale crop mapping based on U-Net model. *Remote Sens.* **2019**, *11*, 68. [CrossRef]
50. Ohleyer, S. *Building Segmentation on Satellite Images*; ENS: Paris, France, 2018.
51. Khalel, A.; El-Saban, M. Automatic pixelwise object labeling for aerial imagery using stacked U-Nets. *arXiv* **2018**, arXiv:1803.04953.

52. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1480–1484.
53. Marcu, A.; Costea, D.; Slusanschi, E.; Leordeanu, M. A multi-stage multi-task neural network for aerial scene interpretation and geolocalization. *arXiv* **2018**, arXiv:1804.01322.
54. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. *Remote Sens.* **2019**, *11*, 917. [CrossRef]
55. Li, X.; Yao, X.; Fang, Y. Building-A-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3680–3687. [CrossRef]
56. Liu, Y.; Zhang, Z.; Zhong, R.; Chen, D.; Ke, Y.; Peethambaran, J.; Chen, C.; Sun, L. Multilevel building detection framework in remote sensing images based on convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3688–3700. [CrossRef]
57. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
58. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
59. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; The MIT Press: Montreal, QC, Canada, 2014; Volume 11, pp. 2672–2680.
60. Huang, B.; Lu, K.; Audeberr, N.; Khalel, A.; Tarabalka, Y.; Malof, J.; Boulch, A.; Le Saux, B.; Collins, L.; Bradbury, K.; et al. Large-scale semantic classification: Outcome of the first year of inria aerial image labeling benchmark. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 6947–6950.
61. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
62. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; Wang, R. Deep few-shot learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2290–2304. [CrossRef]
63. Zhang, T.; Sun, X.; Zhang, Y.; Yan, M.; Wang, Y.; Wang, Z.; Fu, K. A training-free, one-shot detection framework for geospatial objects in remote sensing images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1414–1417.
64. Rostami, M.; Kolouri, S.; Eaton, E.; Kim, K. Deep transfer learning for few-shot SAR image classification. *Remote Sens.* **2019**, *11*, 1374. [CrossRef]
65. Song, Q.; Xu, F. Zero-shot learning of SAR target feature space with deep generative neural networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2245–2249. [CrossRef]
66. Zhang, Y.; Li, Q.; Huang, H.; Wu, W.; Du, X.; Wang, H. The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: A case study in Beijing, China. *Remote Sens.* **2017**, *9*, 865. [CrossRef]
67. Tu, W.; Hu, Z.; Li, L.; Cao, J.; Jiang, J.; Li, Q.; Li, Q. Portraying urban functional zones by coupling remote sensing imagery and human sensing data. *Remote Sens.* **2018**, *10*, 141. [CrossRef]
68. Liu, T.; Yang, X. Monitoring land changes in an urban area using satellite imagery, GIS and landscape metrics. *Appl. Geogr.* **2015**, *56*, 42–54. [CrossRef]
69. Sudhira, H.S.; Ramachandra, T.V.; Jagadish, K.S. Urban sprawl: Metrics, dynamics and modelling using GIS. *Int. J. Appl. Earth Obs. Geoinf.* **2004**, *5*, 29–39. [CrossRef]
70. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. [CrossRef]
71. Sun, Y.; Zhang, X.; Zhao, X.; Xin, Q. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. *Remote Sens.* **2018**, *10*, 1459. [CrossRef]

72. Varol, B.; Yilmaz, E.O.; Maktav, D.; Bayburt, S.; Gurdal, S. Detection of illegal constructions in urban cities: Comparing LIDAR data and stereo KOMPSAT-3 images with development plans. *Eur. J. Remote Sens.* **2019**, *52*, 335–344. [CrossRef]
73. Agouris, P.; Mountrakis, G.; Stefanidis, A. Automated spatiotemporal change detection in digital aerial imagery. In *Automated Geo-Spatial Image and Data Exploitation*; International Society for Optics and Photonics: Orlando, FL, USA, 2000; Volume 4054, pp. 2–12.
74. Al-Dail, M.A. Change Detection in Urban Areas using Satellite Data. *J. King Saud Univ.-Eng. Sci.* **1998**, *10*, 217–227. [CrossRef]
75. Zhang, M.; Zhu, D.; Su, W.; Huang, J.; Zhang, X.; Liu, Z. Harmonizing Multi-Source Remote Sensing Images for Summer Corn Growth Monitoring. *Remote Sens.* **2019**, *11*, 1266. [CrossRef]
76. Lin, D.Y.; Wang, Y.; Xu, G.L.; Fu, K. Synthesizing remote sensing images by conditional adversarial networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 48–50.
77. Yu, S.; Liu, L.; Wang, Z.; Dai, G.; Xie, Y. Transferring deep neural networks for the differentiation of mammographic breast lesions. *Sci. China Technol. Sci.* **2019**, *62*, 441–447. [CrossRef]
78. Gong, R.; Li, W.; Chen, Y.; Gool, L.V. DLOW: Domain flow for adaptation and generalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2477–2486.
79. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datch, M.; Pelillo, M; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.