remote sensing

MDPI

Article

Assessing Deep Convolutional Neural Networks and Assisted Machine Perception for Urban Mapping

Yang Shao [1,*], Austin J. Cooner [1] and Stephen J. Walsh [2]

[1] Department of Geography, Virginia Tech, 238 Wallace Hall, Blacksburg, VA 24060, USA; austincooner@gmail.com
[2] Department of Geography, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3220, USA; swalsh@email.unc.edu
* Correspondence: yshao@vt.edu

Abstract: High-spatial-resolution satellite imagery has been widely applied for detailed urban mapping. Recently, deep convolutional neural networks (DCNNs) have shown promise in certain remote sensing applications, but they are still relatively new techniques for general urban mapping. This study examines the use of two DCNNs (U-Net and VGG16) to provide an automatic schema to support high-resolution mapping of buildings, road/open built-up, and vegetation cover. Using WorldView-2 imagery as input, we first applied an established OBIA method to characterize major urban land cover classes. An OBIA-derived urban map was then divided into a training and testing region to evaluate the DCNNs' performance. For U-Net mapping, we were particularly interested in how sample size or the number of image tiles affect mapping accuracy. U-Net generated cross-validation accuracies ranging from 40.5 to 95.2% for training sample sizes from 32 to 4096 image tiles (each tile was 256 by 256 pixels). A per-pixel accuracy assessment led to 87.8 percent overall accuracy for the testing region, suggesting U-Net's good generalization capabilities. For the VGG16 mapping, we proposed an object-based framing paradigm that retains spatial information and assists machine perception through Gaussian blurring. Gaussian blurring was used as a pre-processing step to enhance the contrast between objects of interest and background (contextual) information. Combined with the pre-trained VGG16 and transfer learning, this analytical approach generated a 77.3 percent overall accuracy for per-object assessment. The mapping accuracy could be further improved given more robust segmentation algorithms and better quantity/quality of training samples. Our study shows significant promise for DCNN implementation for urban mapping and our approach can transfer to a number of other remote sensing applications.

Keywords: deep convolutional neural networks; U-Net; VGG16; urban mapping

Citation: Shao, Y.; Cooner, A.J.; Walsh, S.J. Assessing Deep Convolutional Neural Networks and Assisted Machine Perception for Urban Mapping. *Remote Sens.* **2021**, *13*, 1523. https://doi.org/10.3390/rs13081523

Academic Editor: Garik Gutman

Received: 12 March 2021
Accepted: 13 April 2021
Published: 15 April 2021

1. Introduction

Urban mapping techniques have been rapidly evolving through advances in computer algorithms and integration of a wide range of satellite data. High-resolution urban mapping typically involves characterizing important features such as individual buildings, roads, open built-up, and urban trees and associated vegetation [1,2]. These small-scale urban features can be best mapped using very high-resolution (VHR, <5 m spatial resolution) satellite data such as those from IKONOS, QuickBird, and the WorldView series of sensors [3–5]. For mapping tasks using VHR data, object-based image analysis (OBIA) is preferred over traditional per-pixel classification [6–10], because pixels of a homogeneous land-cover patch often have heterogeneous spectral responses or high information content. Combined with various classification algorithms, OBIA has been routinely used to map detailed urban features with some success [4,11,12]. Previous studies also demonstrated the advantage of data fusion of VHR and LiDAR or synthetic aperture radar (SAR) images for urban-mapping applications [13–15].

Performance of VHR-based urban mapping depends on the choice of image classification algorithms. Currently, the most commonly used algorithms include support vector machines, random forests, feed-forward artificial neural networks, and radial basis function neural networks [16]. While each of these approaches show promise for automatic urban mapping, they often require a diverse set of input data for the given algorithm to properly function. For example, in addition to the original spectral bands of VHR data, a rich set of textural and structural features, such as gray-level co-occurrence matrix and wavelet textures, have been examined to improve urban mapping accuracy [2,16]. Relying on hand-crafted features requires additional processing time and remote sensing expertise and can potentially confuse other landscape features. It is also unclear whether the textural features from each study can be applied with similar results to other study areas.

Deep convolutional neural networks (DCNNs) have recently shown great promise in the field of computer vision after the landmark paper by Krizhevsky et al. [17]. DCNNs work by learning convolutions (or features) that best represent image classes through error minimization via backpropogation. For example, one of the famous DCNNs, the VGG16, was designed by researchers from the Visual Geometry Group Lab of Oxford University and the 16-layer network architecture achieved 92.7 percent test accuracy in ImageNet [18]. The automated feature engineering is particularly appealing compared to traditional, hand-crafted features using domain knowledge [19]. While DCNNs were originally designed for large-scale image recognition [17,18,20], recent work has shown that fully convolutional networks (FCN) and Markov conditional random fields (CRF) can be used for effective semantic segmentation or pixel-wise classification [21,22]. The U-Net, a specific type of convolutional network with creative design of contracting and expanding architecture, shows great potential in biomedical image segmentation or pixel labeling [23].

DCNNs have only recently transitioned into the field of remote sensing. Researchers are increasingly interested in using DCNNs or other deep learning models for scene classification, object detection, and land use and land cover classification [24]. For example, Zou et al. [25] applied transfer learning to VHR remote sensing imagery for classifying 400 $\times$ 400 pixel samples into seven distinct scene types. Sun et al. [26] modified three DCNNs (i.e., AlexNet, VGG16, and ResNet50) to classify tree species, and they found that VGG16 performed best. Full scene classification has also been applied using SVMs on DCNN features [27], DCNN-CRF [28], and DCNN-FCN [29]. More recently, several studies evaluated U-Net for vegetation mapping and obtained very accurate map products [30–32]. The U-Net allows for end-to-end training at the pixel level and it is less demanding on the training sample size [23].

Although DCNNs showed high potential for scene classification and object detection, their overall effectiveness in pixel-wise land use and land cover mapping is still unclear. Several recent articles suggested that the application of DCNNs for pixel-wise land cover mapping remains sparse, and the potential is not fully explored [24,31]. For DCNNs and many other machine learning algorithms, it is also important to investigate machine perception that imitates human perception to improve learning and predictive performance. For a given image/scene, the object of interest and background or contextual information can be rapidly identified by humans. It is thus potentially beneficial to enhance or synthesize observations that may improve computer understanding of remotely sensed images. Within the DCNNs, image tiles are commonly used as input, and the object of interest (e.g., building) is mixed with other background features or integrated pixels. We propose using contextual Gaussian blurring to reduce the impact from background features, an assisted machine perception method that has not been thoroughly examined in the literature, especially within the DCNN framework.

The main objective of this study is to investigate two DCNNs, U-Net and VGG16, for urban land cover mapping using VHR data. To support the training and validation, we developed an urban land cover map product through a commonly used OBIA classification of WorldView-2 imagery. For U-Net mapping, we were particularly interested in how classification performance varies with respect to training sample sizes. For the VGG16

mapping, we designed a new object-based image classification approach that involves image segmentation, framing, and object recognition. For each image segment, we incorporated contextual Gaussian blurring to assist machine perception. The experiments are implemented for urban mapping of San Cristobal Island, one of the Islands of the Galapagos archipelago of Ecuador.

2. Materials and Methods

2.1. Study Area and Data

The Galapagos Islands are a chain of islands known for their natural beauty, history, diverse wildlife, and conservation efforts. Over the last several decades, the growing tourism and human migration have been exerting increasing pressure on the fragile and sensitive island ecosystems [8,33]. San Cristobal is one of the four populated islands in the archipelago and has an area of 558 km^2. Total population for San Cristobal is around 7000. Most residents live in the port city, Puerto Baquerizo Moreno, although the smaller upland town of El Progreso, close to the agricultural zone, accounts for under 1000 residents. The port community is bounded by the Pacific Ocean, and land is managed and controlled by the Galapagos National Park. As such, the community has limited space for urban development, with hard borders occurring with the park that limit peripheral growth; hence, most new development occurs through urban in-filling and through land swaps with the park. The urban structures are relatively small in size, although two-story residential units are common. Generally, residential units are less than 30 square meters in area, although larger structures, primarily hotels and associated commercial buildings, occur, but they are generally close to the water's edge, where most tourist facilities are concentrated. Streets are relatively narrow in the town. Most roads are paved, although dirt roads persist in the town.

The image used for this study was acquired by the WorldView-2 satellite on 1 December 2016. The multispectral product (2.4 m resolution) was converted to TOA reflectance and then the subtractive resolution merge process was applied to create a pan-sharpened image consisting of four spectral bands. Figure 1 shows the study area and WorldView-2 image used for this research.

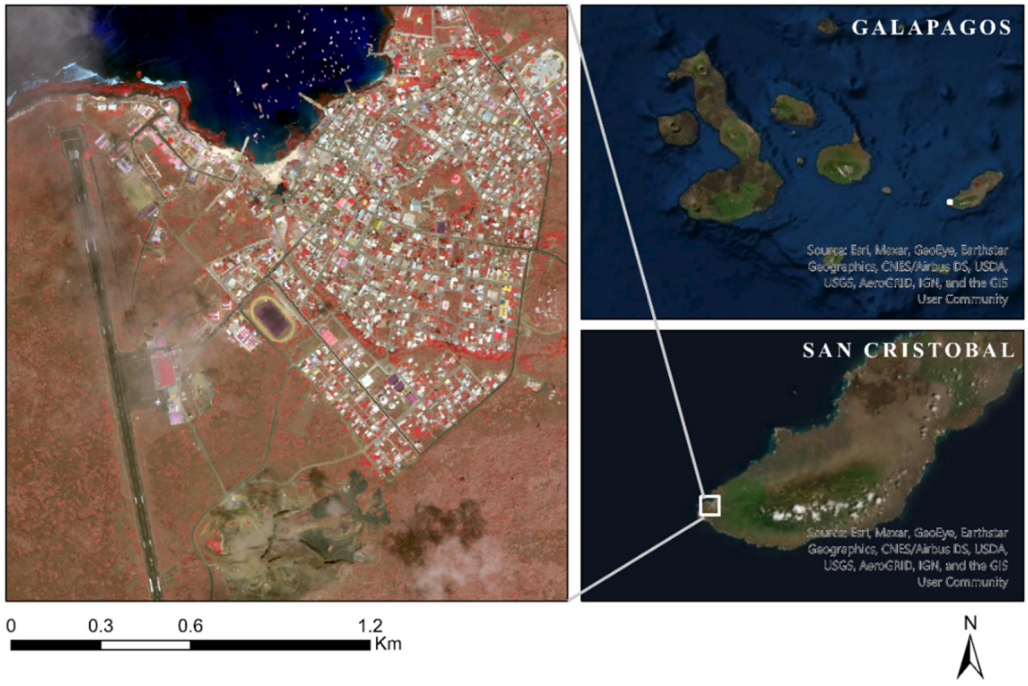

Figure 1. The port city of Puerto Baquerizo Moreno on San Cristobal Island of Galapagos. The VHR WorldView-2 imagery is presented in RGB false color combination.

2.2. OBIA Image Classification

To support the DCNN-based urban mapping, we first implemented OBIA image classification to derive a land use and land cover reference. This OBIA-derived land use and land cover map served as the primary benchmark dataset for DCNN training and validation. Using the WorldView-2 image as input, we applied the multiresolution segmentation algorithm within the eCognition Essentials software package to generate image objects. Among several adjustable parameters (scale, shape, and compactness), the scale factor is the most important parameter for image object size [34,35]. We examined a range of scales (75, 100, and 125) through a trial-and-error approach to determine the optimum scale factor. The weighting between color and shape was set to 0.7/0.3 based on previous studies showing the relative importance of color components [4,6]. The compactness and smoothness were assigned equal weights. After each segmentation, we visually assessed the image objects, using the original WorldView-2 image as a reference. We found that a scale of 75 was adequate to separate buildings and road objects from surrounding areas, while minimizing over-segmentation. A small building was typically represented by one image object, while a large building was divided into several homogenous patches.

Following the image segmentation, we conducted the object-based image classification using a random forest algorithm. The mean value of each spectral band was used to represent each object to support the classification. Six land cover classes were considered: building, road/open built-up, vegetation, beach, volcanic lava/soil, and water. A few locations with obvious cloud/shadow were manually masked out. Minimal post-classification manual editing was conducted to remove the obvious classification errors. For example, some buildings were misclassified as roads and vice versa. Classification accuracy was assessed using 50 randomly selected image objects (polygons) per class for three major urban classes of building, road/open built-up, and vegetation. The other three classes, beach, volcanic lava/soil, and water, were not significant components of urban land cover. We visually interpreted each polygon using the WorldView-2 image and Google Earth's very high-resolution imagery archive as references. Reference polygons contain heterogeneous land cover types, and the dominant land cover was used as the label. Error matrix and accuracy statistics were generated by comparing the OBIA classification result and visually interpreted land cover references.

2.3. Image Classification Using U-Net

The U-Net was developed by Ronneberger et al. [23] for localized pixel classification for biomedical image segmentation. The U-Net architecture (Figure 2) features a contracting path, where convolution and max pooling operations are used to extract image context, and an expanding path, where up-sampling and convolution are used for sequential localization. The localization is enhanced by integrating the extracted features from the contracting path at each spatial scale. This U-Net design allows for end-to-end training at the pixel level and shows very good performance on many image segmentation tasks [23,36]. The U-Net design is particularly appealing for the remote sensing community, because it provides a class label for individual pixels, instead of focusing on scene labeling. The localization is the key for land cover mapping and change detection tasks.

The four-band WorldView-2 image was divided into two parts of northern (50%) and southern (50%) sub-regions. The northern and southern sub-regions were used as the training and testing sets, respectively. These areas were chosen to balance the need to provide the network with sufficient training data, while ensuring that the network can be tested for overall generalization. The classification map derived from the OBIA method (Section 2.2) was used as the labeled image (or target) to support U-Net training and testing. The original classification scheme included six land cover classes of building, road/open built-up, vegetation, beach, volcanic lava/soil, and water. We masked out land cover classes of beach, volcanic lava/soil, and water for U-Net classification, because they are either not significant components of urban land cover or they have very limited spatial coverage (i.e., beach).

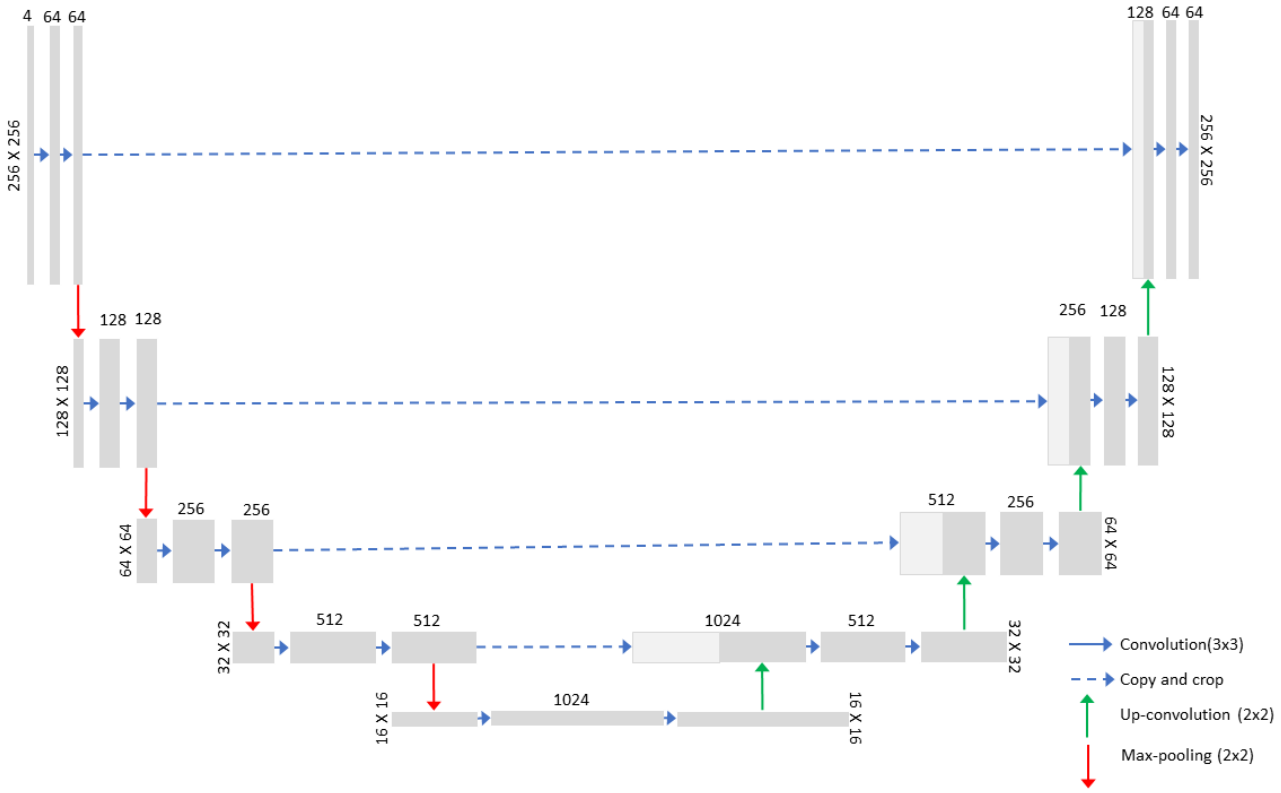

Figure 2. The U-Net architecture features a contracting path and an expanding path. The input data consist of four spectral bands of a WorldView-2 image.

From the training set, the northern sub-region, we randomly extracted image tiles of varying sample sizes from 32 to 4096 to evaluate how sample size affects classification performance. Each tile has four spectral bands with 256 by 256 pixels for each layer. For a given training set, the image tiles were further divided into image batches (batch size = 32) during U-Net training. The final layer of U-Net includes a pixel-wise softmax activation combined with the cross-entropy loss function. The U-Net was trained using the stochastic gradient descent with momentum (SGDM) optimization. The initial learning rate was specified as 0.05 and the gradient clipping threshold was set as 0.05 to improve the stability of network training. The maximum training epoch was set as 10, because the training accuracy typically became saturating after 3–5 epochs. To reduce potential overfitting, we recorded cross-validation (20% hold-out) accuracy for each training epoch. The trained U-Net with the best cross-validation accuracy was then applied to the southern sub-region of the study area to generate pixel labels.

2.4. Image Classification Using VGG16

The data preparation of the VGG16 urban mapping included three basic steps: segmentation, framing, and labeling. The WorldView-2 image was previously segmented using the multiresolution segmentation algorithm within the eCognition Essentials software package (scale factor is 75, Section 2.2). Once the image was segmented, an image database was created by framing an individual scene around each object, regardless of image object size. Frame dimension was determined by considering natural image perception: smaller objects required larger frames to place in context, while larger objects (such as roads, large vegetation patches, etc.) required little to no framing for identification (see Table 1). Object size was considered by averaging width and height in the spatial x and y domain. Objects with size less than 50 pixels were assigned a window size of 75 × 75 pixels, the smallest resolution at which objects can be placed in context and identified by the human eye. Objects with size between 50 and 500 pixels were given a frame dependent on their

size; a scale factor was developed that provides linear interpolation between a 50 pixels object being given a window size of 75 × 75 and a 500 pixels object being given a window size of 500 × 500. Objects larger than 500 pixels were given framing equal to their width and height.

Table 1. Object frame size. W is object width, H is object height, S is scale factor, AvgBox is defined by averaging width and height in the spatial x and y domain.

Object Size (Average of Width and Height, Pixels)	Frame Size
≤50	75 × 75
50–500	$S^*W \times S^*H$ $S = (-0.0011 \times AvgBox) + 1.555$
≥500	W × H

To test our hypothesis that machine perception could be assisted by highlighting the pertinent object, two separate image databases were produced, one where each object's background was blurred using a Gaussian filter (σ = 1, 5 × 5 filter size) and one with no augmentation. Figure 3 shows a comparison of a chipped image example of an object with/without the Gaussian assistance.

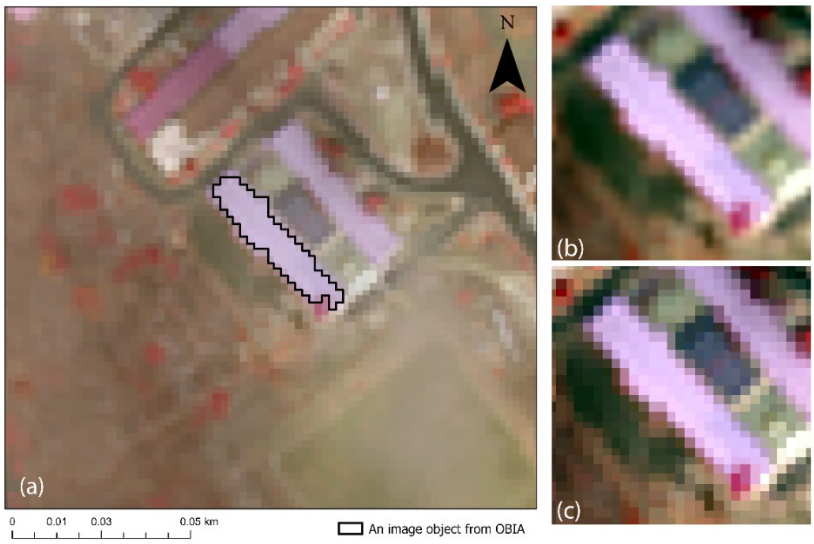

Figure 3. A comparison of chipped image example of an object with/without the Gaussian assistance: (**a**) a selected example of image object from OBIA segmentation, (**b**) the object's background was blurred using a Gaussian filter, (**c**) the object's background was not blurred.

Finally, the image database was labeled according to the class labels from the OBIA classification results.

Training was accomplished on the VGG16 architectures using the deep learning package from Matlab2020a. To capitalize on the millions of images that these networks have already been exposed to, transfer learning was performed on the image databases using the network pre-trained for the ImageNet challenge. The network architectures used were identical to the original networks, except that the networks' last layer's output was reduced from 1000 to 3 to classify three distinct object types (building, road/other built-up, and vegetation). Because the shallower layers contain basic image feature information such as edge or color detection, which are shared between the ImageNet and remote sensing datasets, the learning rates for the earlier network layers were set to zero, thus freezing the weights of these layers during transfer learning [37]. For the last fully connected layer, the network's learning rate was set to 1×10^{-5}. This enabled fine-tuning of the deep network

layers, while more dramatically altering the weights of the last fully connected layer that assigns class probabilities.

The original VGG16 uses 3-band RBG images as input. For our study, we selected near-infrared, red, and green bands as input channels, because certain land cover classes (i.e., vegetation cover) can be best mapped by including the near-infrared band. Similar to the U-Net urban mapping, the northern portion of the WorldView-2 image was used as the training data. Once the network was trained, the VGG16 network predicted the label for the segmented test dataset from the southern portion of the image. The spatial location of each object was retained so that object labels could be mapped directly onto the georeferenced scene.

3. Results

3.1. OBIA

Figure 4 shows the OBIA-derived land cover map and a highlighted area focusing on the detection of buildings. Overall, the well-preserved shapes of buildings and roads suggested good classification performance using OBIA. The OBIA image classification results were assessed using visual interpretation of the WorldView-2 image and the Google Earth's very high-resolution imagery archive as references. For 150 randomly selected OBIA objects for building, road/open built-up, and vegetation classes, the overall accuracy was 86.7 percent (n = 150) with a Kappa coefficient of 0.80 (Table 2). The user's accuracies of building and road/open built-up were 84.0 percent. Certain roofing materials and roads/open built-up parcels may have similar spectral characteristics so the confusion between these two classes was expected. The producer's accuracies for building, road/open built-up, and vegetation were 85.7, 84.0, and 90.2 percent, respectively. We note that per-object accuracy assessment was used here to maintain the consistency with OBIA's analytical unit. The main advantages of using per-object over per-pixel accuracy assessment include reducing positional errors and difficulty in interpreting edge pixels. However, the associated accuracy statistics cannot be directly generalized to areas of agreement and disagreement due to the varying size of OBIA objects.

Table 2. Error matrix for OBIA image classification was generated using reference data derived from visual interpretation of WorldView-2 image and the Google Earth's very high-resolution imagery archive. A total of 150 OBIA-derived polygons were randomly selected for the accuracy assessment. UA and PA denote user's and producer's accuracy, respectively.

	Building	**Road/Open Built-Up**	**Vegetation**	**Total**	**UA**
Building	42	5	3	50	84.0
Road/open built-up	6	42	2	50	84.0
Vegetation	1	3	46	50	92.0
Total	49	50	51		
PA	85.7	84.0	90.2	Overall = 86.7%	Kappa = 0.80

3.2. U-Net Mapping

The training sample size or the number of randomly extracted image tiles had large impacts on U-Net classification performance. Figure 5 compares cross-validation accuracies of U-Net classification across various training sample sizes (32 to 4096 image tiles). At each training sample size, the boxplot shows the variation in overall accuracy across training epochs from one to ten. Using 32 randomly selected image tiles in training, the overall accuracies varied from 40.5 to 83.0 percent. The variation in overall accuracy across training epochs decreased when more image tiles were used in training. With 2048 image tiles, the overall accuracies were above 92 percent across all training epochs. The further increase in image tiles in training did not result in improved cross-validation accuracy. The highest cross-validation accuracy (95.3 percent) was obtained after three training epochs using 2048

image tiles as input. The resultant U-Net was applied to the testing region of the image to obtain the final classification map.

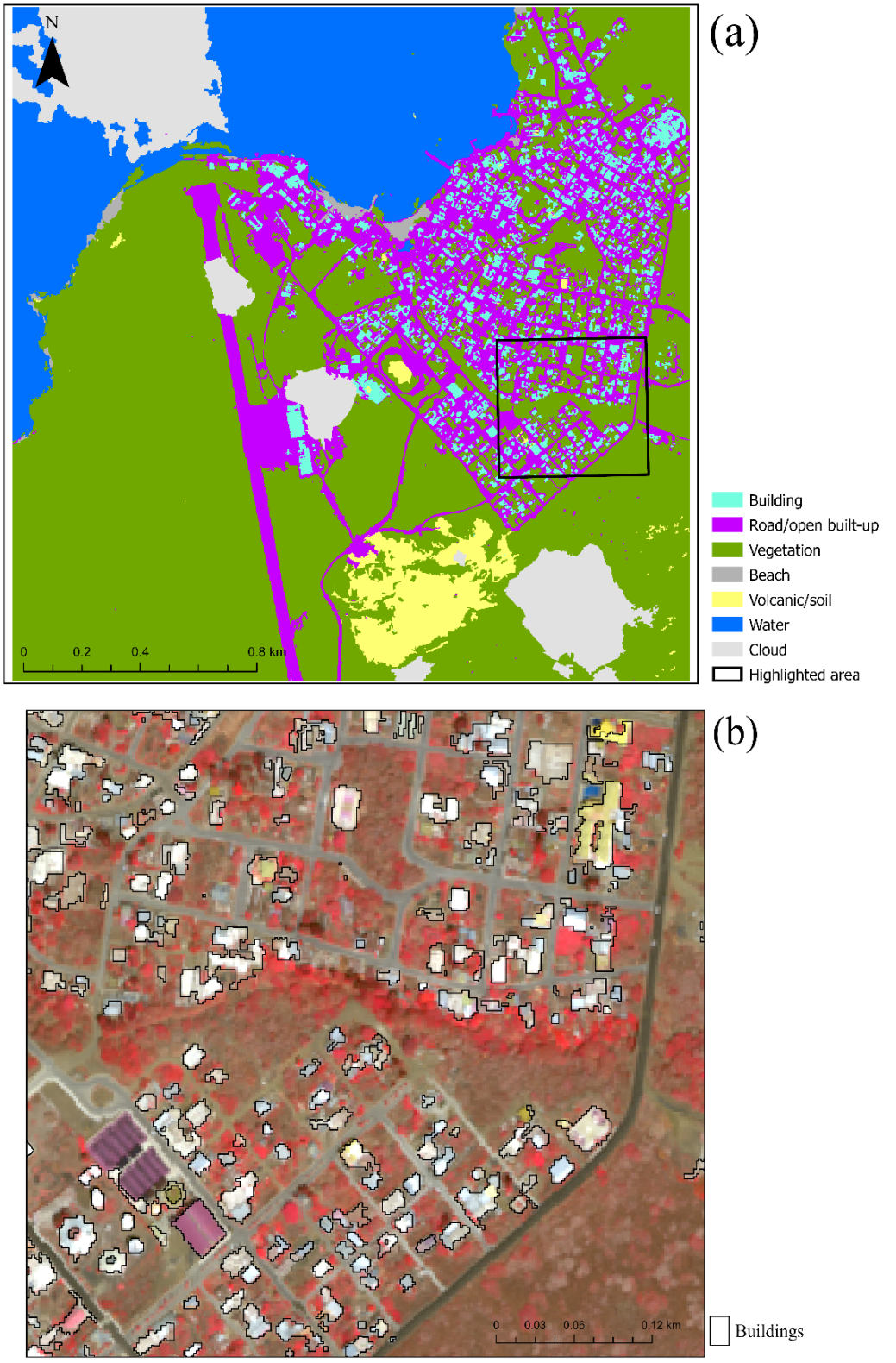

Figure 4. OBIA classification results using four spectral bands of a WorldView-2 image as input: (**a**) Overview of all land cover classes; locations with obvious cloud/shadow were manually masked out. (**b**) Buildings (polygons) for the highlighted area.

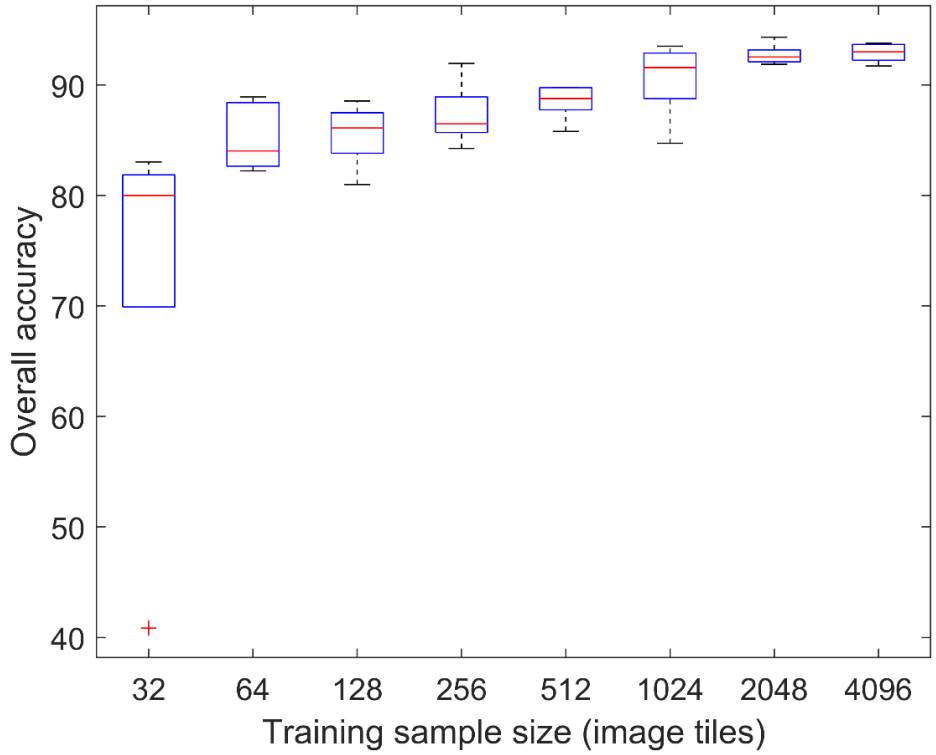

Figure 5. Cross-validation accuracies across varying sample sizes. Image tiles (32 to 4096) were compared to evaluate how sample size affects classification performance.

Figure 6 shows the U-Net classification result for the testing region. Visually, the buildings and road/open built-up areas match well with the visual interpretation of the WorldView-2 image for the test site: objects marked as buildings are scattered around residential areas and objects marked as roads follow the city's transport network. A close visual evaluation indicated a high level of agreement of building outlines. There were apparent classification errors in areas close to the airport and track/soccer complex. Some "salt-and-pepper" noise was observed within the vegetation patches located in the southeast corner of the testing image. Pixel-wise accuracy assessment was conducted for U-Net classification results. The following error matrix (Table 3) presents U-Net classification accuracy statistics using a total of 90 randomly selected points. The overall accuracy was 87.8 percent (Kappa = 0.82). For building class, the user's and producer's accuracies were 86.7 and 92.9 percent, respectively. Road/open built-up class had a lower user's accuracy of 80.0 percent. The error is attributed to confusion between open built-up (with sparse vegetation cover) and vegetation classes.

Table 3. Error matrix for U-Net image classification was generated using reference data derived from visual interpretation of WorldView-2 image and the Google Earth's very high-resolution imagery archive. A total of 90 points were randomly selected for the accuracy assessment. UA and PA denote user's and producer's accuracy, respectively.

	Building	**Road/Open Built-Up**	**Vegetation**	**Sum**	**UA**
Building	26	3	1	30	86.7
Road/open built-up	2	24	4	30	80.0
Vegetation	0	1	29	30	96.7
Sum	28	28	34		
PA	92.9	85.7	85.3	OA = 87.8	Kappa = 0.82

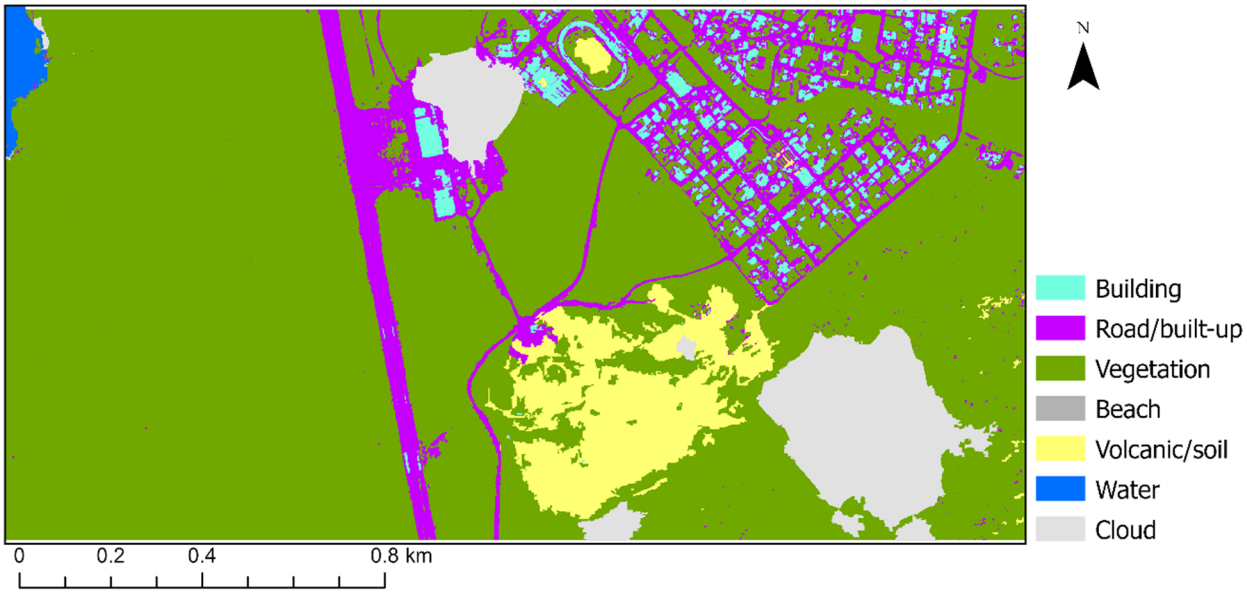

Figure 6. U-Net classification results using four spectral bands of a WorldView-2 image as input. Only testing image (southern portion of the study area) is presented to emphasize the U-Net's generalization capability.

3.3. VGG16 Mapping

Using the image database with Gaussian assistance as input, the highest cross-validation accuracy (85.3 percent) was obtained at five training epochs. Figure 7a shows the classification result for the testing region. Classification errors are shown as image objects, because an object-based classification approach was used. For example, some large vegetation patches were misclassified as road/open built-up classes. A few road/open built-up segments on the runway of the airport were labeled as buildings. Compared to the U-Net mapping results, there was a higher level of confusion between the building and road/open built-up classes. Using 150 randomly selected image objects as a reference, the VGG16 resulted in an overall accuracy of 77.3 percent (Kappa = 0.66). The user's accuracies for building, road/open built-up, and vegetation were 76.0, 74.0, and 82.0 percent, respectively (Table 4).

Table 4. Error matrix for VGG16 mapping (with Gaussian assistance) was generated using reference data derived from the visual interpretation of a WorldView-2 image and the Google Earth's very high-resolution imagery archive. A total of 150 image objects were randomly selected for the accuracy assessment. UA and PA denote user's and producer's accuracy, respectively.

	Building	**Road/Open Built-Up**	**Vegetation**	**Total**	**UA**
Building	38	10	2	50	76.0
Road/open built-up	8	37	5	50	74.0
Vegetation	3	6	41	50	82.0
Total	49	53	48		
PA	77.6	69.8	85.4	Overall = 77.3%	Kappa = 0.66

Using the image database without Gaussian assistance as input, we obtained the highest cross-validation accuracy of 76.5 percent for the training data. The trained VGG16 network was then applied to the testing region to generate the classification map (Figure 7b). Classification errors were apparent for all three classes of building, road/open built-up, and vegetation cover. Overall, the VGG16 mapping without Gaussian assistance performed far worse compared to those with Gaussian assistance. For the same set of validation image objects, the overall accuracy decreased to 67.3 percent.

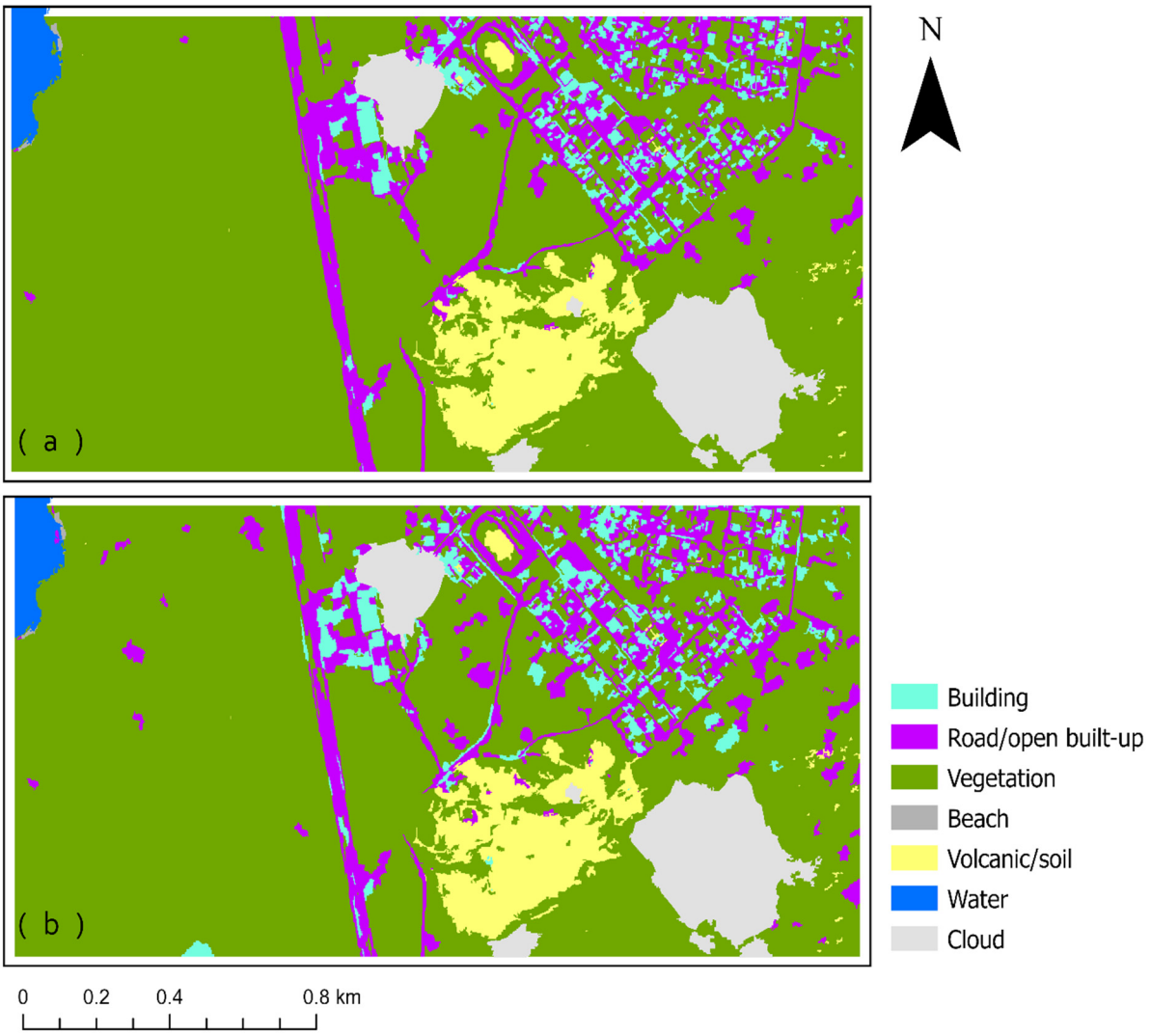

Figure 7. VGG16 classification results: (**a**) with Gaussian assistance and (**b**) without Gaussian assistance. Only testing image (southern portion of the study area) is presented to evaluate generalization capability.

4. Discussion

Currently, the major challenge and opportunity for remote sensing researchers involves developing image analytical approaches that can take full advantage of DCNNs designed for computer vision or other pattern recognition fields. Our study serves as one of these experiments. The U-Net network, combined with an OBIA-derived land cover map as training data, performed very well at high-resolution urban mapping. The buildings, road/open built-up, and vegetation were identified with an overall accuracy of 87.8 percent. The classification result from OBIA showed errors. For example, the class-specific accuracies for building and road/open built-up objects were close to 85 percent. Building boundaries in certain residential areas were not well defined because of combined limiting factors of sensor data, segmentation procedure, and classification algorithms. It is interesting to note that U-Net generated accurate building and road products using the OBIA-derived reference with reasonable accuracy, rather than using reference data acquired by visual interpretation or manual digitizing [31,38]. This points to the utility of traditional, shallow machine learning methods to efficiently generate training data, particularly for the express purpose of evaluating various deep learning schemas.

With an OBIA-derived reference map, image tiles can be easily extracted for U-Net training. With a relatively small training sample size (e.g., 32–256 tiles), U-Net generated a

detailed urban map with acceptable accuracy. However, it may require multiple trials to generate acceptable results, and it is difficult to know which training tiles lead to acceptable generalization performance. For this study, a sample size of 2048 image tiles generated consistent cross-validation performance across various training epochs. This suggests that trial-and-error learning can be significantly reduced by increasing training samples. The actual pixel counts for 2048 image tiles (256 by 256 pixel) were approximately 12.5 times the total pixels within the training image (2300 by 4649 pixels). For computational efficiency, the U-Net training can be best accomplished through the use of GPU. We used GeForce GTX1080 GPUs from the advanced research computing facility of the University of North Carolina at Chapel Hill. With multiple GPUs, the training, validation, and testing could be accomplished within tens of minutes. The availability of GPUs is clearly important if a large amount of U-Net training and testing needs to be implemented. Currently, this is still a limiting factor for incorporating the U-Net as a routine tool for the general remote sensing community.

For VGG16-based urban mapping, an OBIA approach using pre-training segmentation and Gaussian assistance was used as a shortcut for scene classification. Our experiment suggested that the pre-trained VGG16 and transfer learning were not as good as U-Net for detailed urban land cover mapping. For example, this paradigm produced more errors of commission for road/open built-up than both OBIA and U-Net classification. The spatially clustered error of commission was most obvious in a testing area with dominant vegetation cover, where the WorldView-2 image has sufficient resolution to distinguish them. There are several possible explanations. We relied on pre-trained VGG16 and transfer learning. Some image segments in the vegetation class may have very similar VGG16-derived features compared to those of the open built-up class. The fine-tuning appeared to be insufficient in separating those image segments. We conducted additional experiments to re-train the entire VGG16. However, the limited training samples (i.e., several thousand of image objects) did not warrant full network training. It should be noted that the DCNN, such as VGG16, typically requires a large amount of meticulously labeled reference imagery in training. Larger image databases could lead to improved classification accuracy.

With this paradigm of VGG16 mapping, performance is dependent on the objects derived through segmentation, representing real objects in space. For this study, we used the multiresolution segmentation algorithm within the eCognition Essentials software package (scale factor is 75, Section 2.2) to generate image segments. A more representative pre-DCNN segmentation algorithm may result in better accuracies for VGG16 mapping. Apparently, the selection of segmentation algorithm and associated parameters call for future research. The OBIA-derived land cover map has approximately 87 percent of object-level accuracy. The direct use of such a noisy product for VGG16 training may generate high uncertainties in prediction. Therefore, the availability of high quality/quantity of training data remains a major challenge for the application of DCNNs for detailed urban mapping. For our experiment, the VGG16 could not generate usable urban map products without Gaussian blurring. This indicates that Gaussian blurring was essential to DCNN performance as it assisted machine perception in the same way that it guides the human eye; thus, machine perception is limited (for the present) to at least what the human brain can visually distinguish.

5. Conclusions

This paper examined the application of two DCNNs, U-Net and VGG16, for urban land cover mapping, using VHR WorldView-2 imagery as input. The use of traditional OBIA land cover mapping was an important first step to generate a reference map in supporting DCNN training and testing. We evaluated U-Net performance using a range of training sample sizes or image tiles (32–4096). U-Net yielded high performance in pixel-wise classification (overall accuracy 87.8 percent) when more than 2000 image tiles were used as input. The main advantage of U-Net included reducing data requirements and eliminating the need for hand-crafted feature extraction. For VGG16-based urban

mapping, we developed a sequential image processing paradigm that includes image segmentation, framing, and VGG16 transfer learning. Although the VGG16-derived urban map was not as good as maps derived from U-Net, our study demonstrated an alternate solution in linking OBIA and DCNNs designed for computer vision tasks. With Gaussian assistance, the pre-trained VGG16 and transfer learning of VGG16 generated moderately accurate urban maps. Urban mapping accuracy could be further improved with more robust segmentation algorithms and better quality/quality of training samples.

Author Contributions: Y.S. and A.J.C. designed the experiment, coded the software, and wrote the manuscript. S.J.W. provided direction, knowledge of the local Galapagos area and context, and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NASA's Land Cover/Land Use Change Program under grant number NNH16ZDA001N-LCLUC. We are also thankful to VT OASF (Virginia Tech Open Access Subvention Funding) support in our published article.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Manandhar, D.; Shibasaki, R. Auto-extraction of urban features from vehicle-borne laser data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2002**, *34*, 650–655.
2. Huang, X.; Lu, Q.; Zhang, L. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. *ISPRS J. Photogramm. Remote Sens.* **2014**, *90*, 36–48. [CrossRef]
3. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *5*, 161–172. [CrossRef]
4. Pu, R.; Landry, S.; Yu, Q. Object-based urban detailed land cover classification with high spatial resolution IKONOS imagery. *Int. J. Remote Sens.* **2011**, *32*, 3285–3308. [CrossRef]
5. Hamedianfar, A.; Shafri, H.Z.M.; Mansor, S.; Ahmad, N. Improving detailed rule-based feature extraction of urban areas from WorldView-2 image and lidar data. *Int. J. Remote Sens.* **2014**, *35*, 1876–1899. [CrossRef]
6. Yu, Q.; Gong, P.; Clinton, N.; Biging, G.; Kelly, M.; Schirokauer, D. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 799–811. [CrossRef]
7. Blaschke, T.; Lang, S.; Hay, G. (Eds.) *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
8. Walsh, S.J.; McCleary, A.L.; Mena, C.F.; Shao, Y.; Tuttle, J.P.; González, A.; Atkinson, R. QuickBird and Hyperion data analysis of an invasive plant species in the Galapagos Islands of Ecuador: Implications for control and land use management. *Remote Sens. Environ.* **2008**, *112*, 1927–1941. [CrossRef]
9. Shao, Y.; Taff, G.N.; Walsh, S.J. Shadow detection and building-height estimation using IKONOS data. *Int. J. Remote Sens.* **2011**, *32*, 6929–6944. [CrossRef]
10. Pu, R.; Landry, S. A comparative analysis of high spatial resolution IKONOS and WorldView-2 imagery for mapping urban tree species. *Remote Sens. Environ.* **2012**, *124*, 516–533. [CrossRef]
11. Moskal, L.M.; Styers, D.M.; Halabisky, M. Monitoring urban tree cover using object-based image analysis and public domain remotely sensed data. *Remote Sens.* **2011**, *3*, 2243–2262. [CrossRef]
12. Shahi, K.; Shafri, H.Z.M.; Hamedianfar, A. Road condition assessment by OBIA and feature selection techniques using very high-resolution WorldView-2 imagery. *Geocarto Int.* **2017**, *32*, 1389–1406. [CrossRef]
13. Ito, Y.; Hosokawa, M.; Lee, H.; Liu, J.G. Extraction of damaged regions using SAR data and neural networks. *Int. Arch. Photogramm. Remote Sens.* **2000**, *33*, 156–163.
14. Dong, L.; Shan, J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 85–99. [CrossRef]
15. Sohn, G.; Dowman, I. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 43–63. [CrossRef]
16. Cooner, A.J.; Shao, Y.; Campbell, J.B. Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti earthquake. *Remote Sens.* **2016**, *8*, 868. [CrossRef]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

19. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
20. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
22. 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.
23. 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 Interventions, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland; pp. 234–241.
24. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
25. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
26. Sun, Y.; Huang, J.; Ao, Z.; Lao, D.; Xin, Q. Deep Learning Approaches for the Mapping of Tree Species Diversity in a Tropical Wetland Using Airborne LiDAR and High-Spatial-Resolution Remote Sensing Images. *Forests* **2019**, *10*, 1047. [CrossRef]
27. Lagrange, A.; Le Saux, B.; Beaupere, A.; Boulch, A.; Chan-Hon-Tong, A.; Herbin, S.; Randrianarivo, H.; Ferecatu, M. Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4173–4176.
28. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
29. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, V.D. Effective semantic pixel labelling with convolutional networks and conditional random fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 36–43.
30. Flood, N.; Watson, F.; Collett, L. Using a U-net convolutional neural network to map woody vegetation extent from high resolution satellite imagery across Queensland, Australia. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *82*, 101897. [CrossRef]
31. Kattenborn, T.; Eichel, J.; Fassnacht, F.E. Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Sci. Rep.* **2019**, *9*, 1–9. [CrossRef] [PubMed]
32. Wagner, F.H.; Sanchez, A.; Tarabalka, Y.; Lotte, R.G.; Ferreira, M.P.; Aidar, M.P.; Gloor, E.; Phillips, O.L.; Aragao, L.E. Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. *Remote Sens. Ecol. Conserv.* **2019**, *5*, 360–375. [CrossRef]
33. Brewington, L.; Frizzelle, B.G.; Walsh, S.J.; Mena, C.F.; Sampedro, C. Remote sensing of the marine environment: Challenges and opportunities in the Galapagos Islands of Ecuador. In *The Galapagos Marine Reserve*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 109–136.
34. Zheng, B.; Campbell, J.B.; Shao, Y.; Wynne, R.H. Broad-Scale Monitoring of Tillage Practices Using Sequential Landsat Imagery. *Soil Sci. Soc. Am. J.* **2013**, *77*, 1755–1764. [CrossRef]
35. Tonbul, H.; Kavzoglu, T. Semi-Automatic Building Extraction from WorldView-2 Imagery Using Taguchi Optimization. *Photogramm. Eng. Remote Sens.* **2020**, *86*, 547–555. [CrossRef]
36. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Cham, Switzeland, 2016; pp. 424–432.
37. Pires de Lima, R.; Marfurt, K. Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis. *Remote Sens.* **2020**, *12*, 86. [CrossRef]
38. 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]