International Journal of Geo-Information

MDPI

Article

Semantic Segmentation of Remote-Sensing Imagery Using Heterogeneous Big Data: International Society for Photogrammetry and Remote Sensing Potsdam and Cityscape Datasets

Ahram Song and Yongil Kim *

Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea; aram200@snu.ac.kr

* Correspondence: yik@snu.ac.kr

Received: 3 September 2020; Accepted: 9 October 2020; Published: 12 October 2020

Abstract: Although semantic segmentation of remote-sensing (RS) images using deep-learning networks has demonstrated its effectiveness recently, compared with natural-image datasets, obtaining RS images under the same conditions to construct data labels is difficult. Indeed, small datasets limit the effective learning of deep-learning networks. To address this problem, we propose a combined U-net model that is trained using a combined weighted loss function and can handle heterogeneous datasets. The network consists of encoder and decoder blocks. The convolutional layers that form the encoder blocks are shared with the heterogeneous datasets, and the decoder blocks are assigned separate training weights. Herein, the International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam and Cityscape datasets are used as the RS and natural-image datasets, respectively. When the layers are shared, only visible bands of the ISPRS Potsdam data are used. Experimental results show that when same-sized heterogeneous datasets are used, the semantic segmentation accuracy of the Potsdam data obtained using our proposed method is lower than that obtained using only the Potsdam data (four bands) with other methods, such as SegNet, DeepLab-V3+, and the simplified version of U-net. However, the segmentation accuracy of the Potsdam images is improved when the larger Cityscape dataset is used. The combined U-net model can effectively train heterogeneous datasets and overcome the insufficient training data problem in the context of RS-image datasets. Furthermore, it is expected that the proposed method can not only be applied to segmentation tasks of aerial images but also to tasks with various purposes of using big heterogeneous datasets.

Keywords: semantic segmentation; deep learning; big dataset; ISPRS Potsdam dataset; Cityscape dataset

1. Introduction

Semantic segmentation involves the allocation of a semantic label to each pixel of an image containing an object, which can deliver high-level structure information [1]. Semantic segmentation is a crucial task in intelligent applications, such as mobile robots and autonomous driving vehicles, because it can provide an accurate understanding of a scene [2]. Recently, significant advances have been achieved in semantic segmentation techniques of natural RGB scenes owing to the development of deep convolutional neural networks (CNNs). Deep-learning models can learn high-level abstract features from raw images with excellent performance; however, these approaches rely on large training samples [3]. To satisfy this requirement, various public datasets have been proposed for scene labeling. For example, PASCAL VOC [4] is a large-scale dataset used for object class recognition and contains 2913 images with pixel-level labeling with 20 classes, such as vehicles and animals. Although similar

to PASCAL, ImageNet contains more than 20,000 classes and 14 million images [5]. The COCO dataset provides more than 328,000 images with 80 classes, and the images are split into different training/validation/test datasets [6]. More recently, the Cityscape dataset has provided a semantic understanding of urban street scenes [7]. It contains 5000 images with dense pixel-level labeling of more than 30 classes of scenes that are commonly encountered during driving, such as vehicles, roads and fences.

Semantic segmentation can be referred to as image classification in the remote-sensing (RS) field, and it has been used in various applications such as land-cover classification, geological and environment surveys, and urban planning. [8–10]. Deep-learning methods have been successfully adopted to solve the problem of satellite and aerial image segmentation, and they outperform original image classifiers [11]. Various deep-learning networks have been established for semantic segmentation, and some methods have achieved good performance for RS images [12]. Fully convolutional networks (FCNs), which were proposed by Long et al. [13], have been used for semantic segmentation of very-high-resolution aerial images [14,15]. In this approach, the fully connected layer is replaced with a convolutional layer, which allows arbitrarily sized input datasets. Based on the concept of FCN, U-net was proposed by Ronneberger et al. [16]; it uses an encoder and decoder architecture. U-net was originally designed to segment medical images, although previous studies have demonstrated that U-net can be successfully used to segment RS images [17]. Similar to U-net, SegNet uses the encoder and decoder architecture. The encoder form is based on convolutional layers from VGG-16, and the decoder performs both upsampling and classifications [18]. Audevert et al. [18] demonstrated the effectiveness of using multiscale SegNet for segmenting International Society for Photogrammetry and Remote Sensing (ISPRS) datasets. Moreover, DeepLab-V3+ uses the encoder and decoder architecture with atrous convolution and fully connected conditional random fields for semantic segmentation. The DeepLab system was successfully applied to RS-images, and it could appropriately handle multiscale objects in high-resolution satellite images [19].

Generally, to compensate for the current lack of large datasets, semantic segmentation of RS-images using deep-learning methods with pretrained networks has been applied to natural RGB-image datasets such as ImageNet and PASCAL VOC [20,21]. However, unlike natural RGB-images, RS-images contain several types of low-resolution objects that are irregularly shaped, which impact subsequent object classifications [12]. Furthermore, as RS-images are acquired from a bird's eye perspective, the objects lie within a flat two-dimensional (2D) plane where only the top of the objects is observed [22]. Additionally, constructing a large-scale RS-image dataset is more difficult than using natural RGB-images, and creating data labels for RS-images obtained from various sensors is time consuming. Indeed, various errors can be introduced due to factors such as relief displacement caused by differences in elevation and shadows in the RS-images. Moreover, it can be difficult to define meaningful classes in a scene in case of numerous surface materials. However, despite these difficulties, large-scale RS public datasets have recently been released. For example, ISPRS provided the benchmark Vaihingen/Germany and Potsdam/Germany datasets containing 33 images with three-band infrared, red, and green (IRRG) and 38 images with four-band infrared, red, green, and blue (IRRGB), respectively, with pixel-level labeling maps comprising five classes and digital elevation models [23]. Moreover, Zurich [24] and Kaggle [25] challenges yielded very-high-resolution satellite images, namely, Quickbird and Worldview-3, respectively, which contain labeling maps with 8–10 classes.

The need for research capable of learning images with different characteristics increases at once with the increase in the big datasets acquired by various sensors. However, when using heterogeneous datasets, it is difficult to account for the differences in the characteristics of input images and class types. For instance, Meletis et al. [26] used the heterogeneous Cityscape and German traffic sign detection benchmark datasets [27] for semantic segmentation of street scenes. The considered datasets were different; however, they contained semantically connected classes. These authors constructed a hierarchy of classifiers using hierarchical training and inference rules using the semantic relationships between the labels of each dataset. Ghassemi et al. [28] designed an encoder and decoder network

for satellite image segmentation of heterogeneous datasets. They used active learning, wherein a trained network was refined using a few sample images in the heterogeneous training and test datasets, and their method improved the network's performance with minimal human intervention. Several studies have been conducted using heterogeneous datasets with similar types of images and labels; however, very few studies have employed natural RGB- and RS-image datasets. For the segmentation of RS-images, there were cases in which pre-trained segmentation networks from natural-image datasets, such as ImageNet, were used as initial values of the segmentation network for RS-images, but the research on directly learning RS and natural RGB-images together is insufficient [3,29].

To address this paucity of research, we analyze the possibility of sharing networks of heterogeneous datasets and identify the impact of learning using the combined weighted loss function between two networks trained on different datasets to overcome the limitations imposed by a lack of training datasets. In this study, aerial and road-view image datasets were used to evaluate the proposed method. Finally, it is expected that the proposed method can be applied to not only segmentation tasks of aerial images, but also the segmentation or object detection of street-view image on web sites and scanned floorplans. The remainder of this paper is organized as follows. The architecture of the proposed method is presented in Section 2, and the datasets and environmental conditions for the experiments are described in Sections 3 and 4. The results and discussion are provided in Sections 5 and 6, respectively, and the conclusions are drawn in Section 7.

2. Methods

The aim of this study is to develop a deep-learning network that can improve the accuracy of semantic segmentation of RS-images using large-scale natural-image datasets with different characteristics. To this end, assuming that the number of RS-images is insufficient, experiments are conducted to determine whether the problem of insufficient training data can be solved when training is performed using both RS-image and natural-image datasets.

2.1. Proposed Combined U-net Model

The proposed method performs semantic segmentation using a technique based on a simplified version of U-net. U-net is a representative deep-learning model used for segmenting RS-images. To reduce the training burden on the proposed model, simplified U-net is used here. Figure 1 shows the architecture of the combined U-net model.

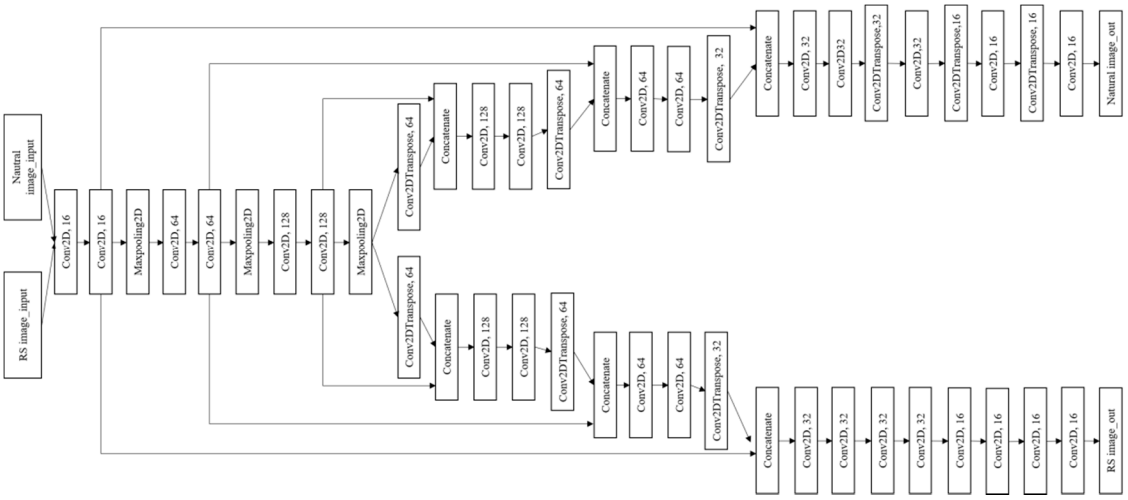

Figure 1. Architecture of the combined U-net model. "Conv2D" denotes the two-dimensional (2D) convolutional layers, and "Conv2DTranspose" denotes a transposed 2D convolutional layer. "Concatenate" denotes a concatenated layer.

The combined U-net model is based on the U-net architecture. The two main components of the U-net architecture are encoder and decoder. For encoding, the combined U-net encoder consists of three main blocks that are shared between the two datasets during the training phase. The model receives input from two different data sources (i.e., the RS-image and natural-image datasets), whose input widths and heights are $n \times n$. The three encoding blocks receive data from the two sources and share the training-parameter weights between them. Every block in the shareable blocks mainly consists of two 2D convolutional layers followed by a max-polling layer for downscaling. Thus, at the end of the encoding phase, the feature map has a size of $\frac{n}{8} \times \frac{n}{8}$. The shared convolutional layers can learn common information that applies to all datasets across all domains [3]. Lee et al. [3] confirmed that shared layers across a domain can be more effective in CNN optimization instead of using only a single dataset because the sharing-layer approach notably improves the classification accuracy compared with the individually trained case. However, in the study by Lee et al. [3], only the middle of the network was shared and only three RS-images were used to train the combined network. In the present study, the initial three encoder blocks are shared during the learning process and the later blocks handle the dataset-specific segmentation tasks. After finishing the encoding phase, the feature maps are decoded separately, implying that every dataset follows a separate decoding path and has separate training weights.

There are two different decoding paths. The first path for the RS-image data consists of three blocks. The initial two blocks primarily consist of one transposed convolutional layer to upscale the feature map by two, followed by two convolutional layers. Furthermore, the last block contains eight convolutional layers. The second path for the natural-image data consists of six blocks. Each block mainly consists of one transposed convolutional layer to upscale the feature maps, followed by two convolutional layers. The decoding blocks end with a feature map of size $n \times n \times 16$, which is scanned using a 1×1 2D convolutional filter (3, 3) to generate an output of size $n \times n \times c$, which matches the shape of the data labels; c is the number of label classes.

In this paper, we use natural-image and RS-image datasets. During the training, the combined U-net handles two inputs in parallel by sharing the three encoder blocks. Moreover, the network is trained with the combined weighted loss, L_c, which is defined as the weighted sum of losses of the two paths. The combined model is then updated using the combined weighted losses (Figure 2). The losses of the first and second paths are denoted by L_{n1} and L_{n2}, respectively, and the spatial cross entropy loss can be defined as shown in Equation (1).

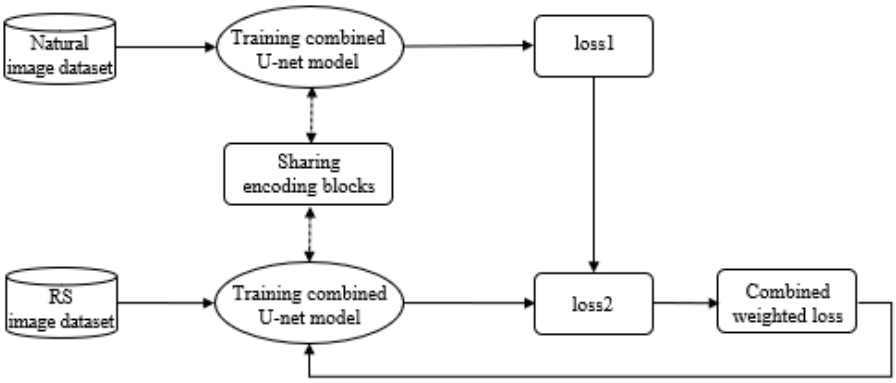

Figure 2. Framework of the proposed method. The combined U-net model shares encoding blocks and is trained using the combined weighted loss.

$$L_n = -\sum_{i=1}^{H \times W} \sum_{k=1}^{c} y_{i,k} \log\left(\hat{y}_{i,k}\right), \tag{1}$$

where $H \times W$ represents the height and width of input x to the network, respectively, c is the number of classes, and $y_{i,k}$ and $\hat{y}_{i,k}$ are the ground-truth and predicted values for the ith pixel x_i and kth class among C different possible classes, respectively [30]. The combined weighted loss, L_c, is defined as follows:

$$L_c = w_1 \cdot L_{n1} + w_2 \cdot L_{n2}, \tag{2}$$

where w_1 and w_2 are the weights of the two networks; these values are empirically determined. Because the ultimate goal of this network is to improve the semantic segmentation accuracy of RS-images, the weight of the network handling RS-image data is set higher than natural-image data.

2.2. Performance Evaluation

There are a number of different ways to evaluate the accuracy of semantic segmentation. In this paper, the overall accuracy (OA), precision, recall, and F1 score are used. OA represents the proportion of correctly classified observations relative to the ground-truth values, and it can be described in terms of true positive (TP), true negative (TN), false negative (FN), and false positive (FP) as

$$\text{OA} = \frac{\text{corrected prediction}}{\text{total prediction}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

OA is a simple and easy approach for evaluating the classification accuracy; however, when the class distributions are dissimilar, OA cannot be appropriately used to show the effectiveness of the results. Instead, the F1 score is a better approach for evaluating the results when there are imbalanced classes, similar to our case. The F1 score (Equation (4)) is the harmonic mean of precision and recall, which in turn measure the correctly identified positive cases from all predicted and actual positive cases, respectively, as given in Equations (5) and (6):

$$\text{F1 score} = \frac{2 \times (\text{recall} \times \text{precision})}{\text{recall} + \text{precision}} \tag{4}$$

$$\text{precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{5}$$

$$\text{recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{6}$$

3. Datasets

Here, two different heterogeneous datasets are used to show the effectiveness of the proposed method. The ISPRS Potsdam and Cityscape datasets are used as the RS and natural-image datasets, respectively. Considering that the aim of this paper is to perform effective semantic segmentation of RS-images using a large-scale natural-image dataset, two datasets with high similarity between the data-label types were selected, even though the properties of the images are different. Both Cityscape and ISPRS Potsdam datasets mainly consist of images of civil areas and have several identical semantic-level features such as roads, cars, and buildings.

3.1. ISPRS Potsdam/Germany Dataset

The ISPRS 2D semantic label Potsdam/Germany dataset is an open benchmark dataset provided online [31] that contains high-resolution airborne images with a spatial resolution of 5 cm and consists of near-infrared (NIR), red, blue, and green orthorectified imagery with corresponding digital surface models. Furthermore, it contains ground-truth images comprising impervious surfaces, buildings, trees, low vegetation, cars, and unknown objects (Figure 3).

The 2D Potsdam dataset contains 38 patches; only 24 images with corresponding labeling images were used for training and validation. Table 1 shows the patch numbers of the labeled data. Twenty-four

large multispectral images of size 6000 × 6000 × 4 pixels in the.tiff format were used as inputs. As the size of the ISPRS images is very large, slices of 256 × 256 × 4 pixels with labels were extracted, separated into batches, and then stored. We trained the network using only a subset of the subimages to consider cases where RS-image data are limited.

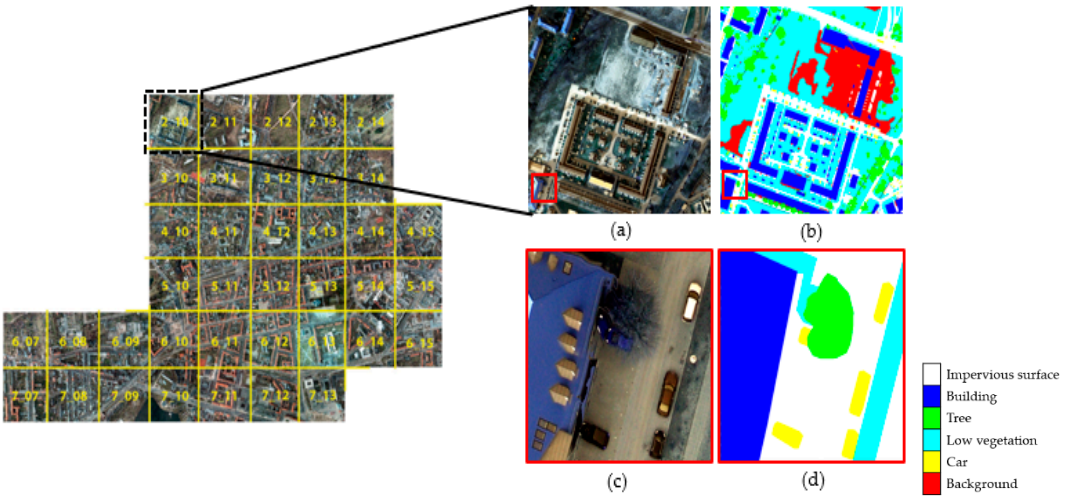

Figure 3. Example of the International Society for Photogrammetry and Remote Sensing (ISPRS) dataset; the patch number is 2–10: (**a**) RGB-image, (**b**) labeling image, (**c**) enlarged RGB-image, and (**d**) data labels.

Table 1. Patch numbers of labeled data.

	Patch Numbers
Potsdam dataset	2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_10, 7_11, 7_12

3.2. Cityscape Dataset

Cityscape provides a large and diverse set of stereo video sequences recoded from streets from 50 different cities with pixel-level ground-truth images [7]. It is a large road scene dataset that can provide semantic understanding of urban street scenes. Cityscape defines 19 semantic labels containing the "void" class for "do-not-care" regions. Figure 4a,d show some examples of the Cityscape images, and Figure 4b,e show their respective label images. Table 2 shows the semantic labels of the Cityscape dataset.

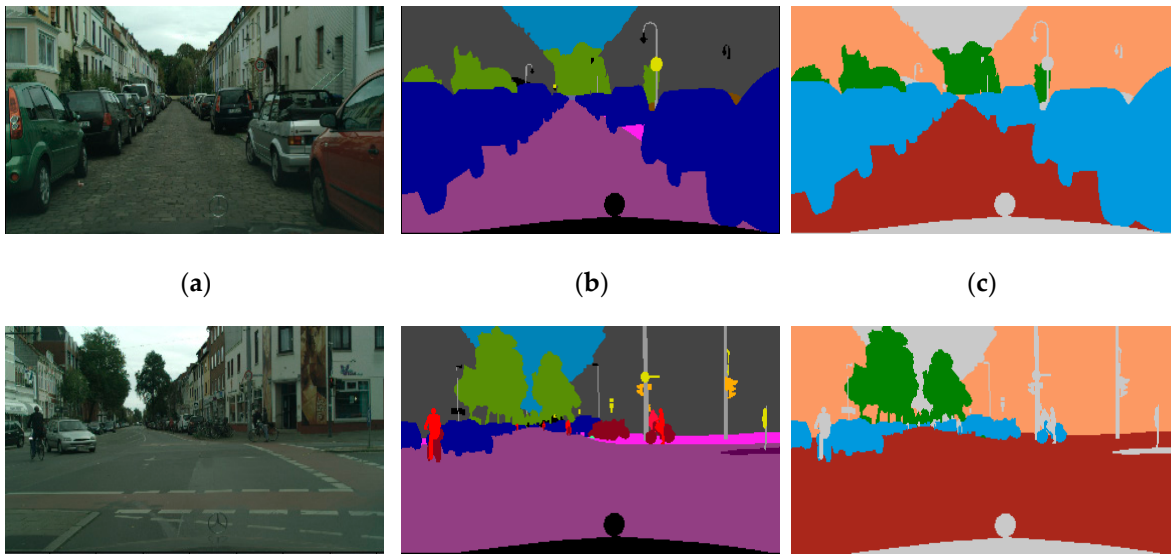

Figure 4. Examples from the Cityscape dataset: (**a**,**d**) RGB-images, (**b**,**e**) original label images, and (**c**,**f**) redefined label images.

Table 2. Original, subclass, and final classes of the Cityscape dataset.

Original Class	Subclass	Final Class
flat areas	road, sidewalk, parking, rail track	flat
human	person, rider	void
vehicle	car, truck, bus, motorcycle, bicycle, caravan, trailer	vehicle
construction	building, wall, fence, guard rail, bridge, tunnel	construction
object	pole, traffic sign, traffic light	void
nature	vegetation, terrain	nature
sky	sky	void
void	ground, dynamic, static	void

To improve the performance efficiency of the proposed method by minimizing the differences between the two datasets, several preprocessing steps were employed. For instance, in the case of the Cityscape dataset, which has deeper semantic-level details than the ISPRS Potsdam dataset, the final classes were adjusted to match those in the Potsdam dataset (Figure 5). Thus, the final classes of the Cityscape dataset included flat areas, construction, nature, vehicles, and voids. Furthermore, all items not included in the above final classes were reclassified as voids. Table 3 shows the relationship between the original and final classes in the Cityscape dataset, and Figure 4c,f show the label maps of the final classes correspondingly shown in Figure 4a,d. Moreover, to share the encoded blocks of the combined U-net model, only the visible bands of the ISPRS Potsdam dataset were used because Cityscape consists of only these bands.

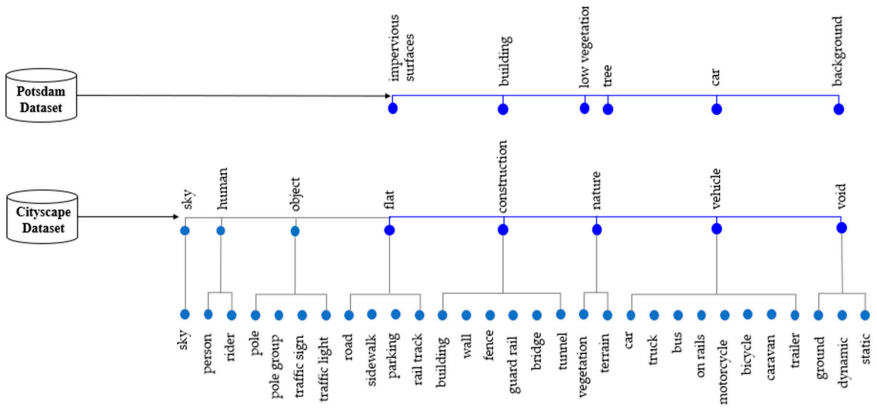

Figure 5. Final label classes of the Potsdam and Cityscape datasets.

Table 3. F1 scores of the five classes and overall accuracy of the test set derived from the Potsdam dataset for the six considered cases.

Model	The Number of Spectral Bands	F1 Score					OA
		Impervious Surface	Building	Low Vegetation	Tree	Car	
SegNet	4bands (RGB+NIR)	0.8615	0.9002	0.7544	0.8111	0.8683	0.8346
DeepLab-V3+		0.8878	0.919	0.8015	0.8343	0.877	0.8605
Simplified U-net		0.8623	0.8535	0.8205	0.8457	0.8812	0.8477
Simplified U-net	3bands (RGB)	0.8417	0.8581	0.7004	0.6848	0.8541	0.7841
Combined U-net-Case 1		0.8789	0.8917	0.7634	0.7757	0.8743	0.8268
Combined U-net-Case 2		0.8924	0.9295	0.7973	0.7982	0.8852	0.8721

4. Experimental Conditions

To demonstrate the effectiveness of the proposed method, several experiments were conducted to compare the segmentation results of the Potsdam images. SegNet, DeepLab-V3+, and simplified U-net, which followed the RS-image path in the combined U-net model, were used to segment the Potsdam dataset with its original four bands. In addition, the Potsdam dataset was trained with only RGB bands using the simplified U-net model to compare with the combined U-net model, which only deals with the RGB bands in the Potsdam images to share convolutional layers. We randomly selected a subset of images from the Potsdam dataset to train the networks. In particular, 1600 images were used as training data, and 400 and 150 images were used as validation and test data, respectively. Finally, the combined U-net model trained the Potsdam dataset with different two conditions. In Case 1, the same number of training datasets of Cityscape dataset was used. Moreover, we varied the number of Cityscape images to confirm the effect of the Cityscape dataset size when training using the combined U-net model. In Case 2, 3000, 550, and 300 images from the Cityscape dataset were used as training, validation, and test data, respectively. Further, to provide larger weights to the Potsdam dataset, w_1 (the weight of loss in RS-image path) and w_2 (the weight of loss in natural-image path) were set as 0.8 and 0.2, respectively.

The combined U-net model was trained on the free Google Colaboratory (Colab) platform [32]. The final epoch was set to 1000 of Adam with a learning rate of 10^{-3}. Considering the available memory on Colab, the batch size was set to 4. The two inputs had the same size ($256 \times 256 \times 3$); hence, we could share the weights in the initial layers. Moreover, $256 \times 256 \times 3$ is a suitable size when using limited training resources such as RAMs and GPUs. Therefore, the original $6000 \times 6000 \times 4$ Potsdam images were divided into smaller samples of $256 \times 256 \times 3$, and the Cityscape images were rescaled to $256 \times 256 \times 3$. In particular, in the second path of the combined U-net, as the six decoding block ends output a feature map of size $1024 \times 2048 \times 16$, this feature map was scanned using a 1×1 2D convolution filter to generate output maps of size $1024 \times 2048 \times 5$ to match the shape of the labels.

We did not rescale the label images to obtain the same size as the input images because rescaling would lead to a loss of class information, which in turn would yield a distorted output. Thus, we decided to match the output of the Cityscape path with the size of the original label size (i.e., 1024 × 2048) to keep the class information untouched.

5. Results

Figure 6 shows the learning graphs of the OA of the training and validation sets for the six aforementioned cases. Since a limited number of images were used for training, there was a difference in the OA between the training and validation sets. When a model has high training and low validation accuracies, this case is probably known as overfitting. Since only a part of the Potsdam dataset was used for training, insufficient training data can sometimes lead to overfitting problems [33]. DeepLab-V3+ showed a higher OA for the validation set in comparison with the SegNet and simplified U-net models (Figure 6b). Also, the difference in the OAs between the validation and training sets was less than when using the SegNet model. In the case of the simplified U-net model, when training was performed by only using the RGB bands of the Potsdam images, the OA of the validation set was lower than when using the original four bands (Figure 6c,d). Also, the learning graph of the simplified U-net model of the training Potsdam images with its four bands showed that the OA of the training set was higher than in the other two cases with the combined U-net models; however, the OA of the validation set was relatively lower than that of the training set.

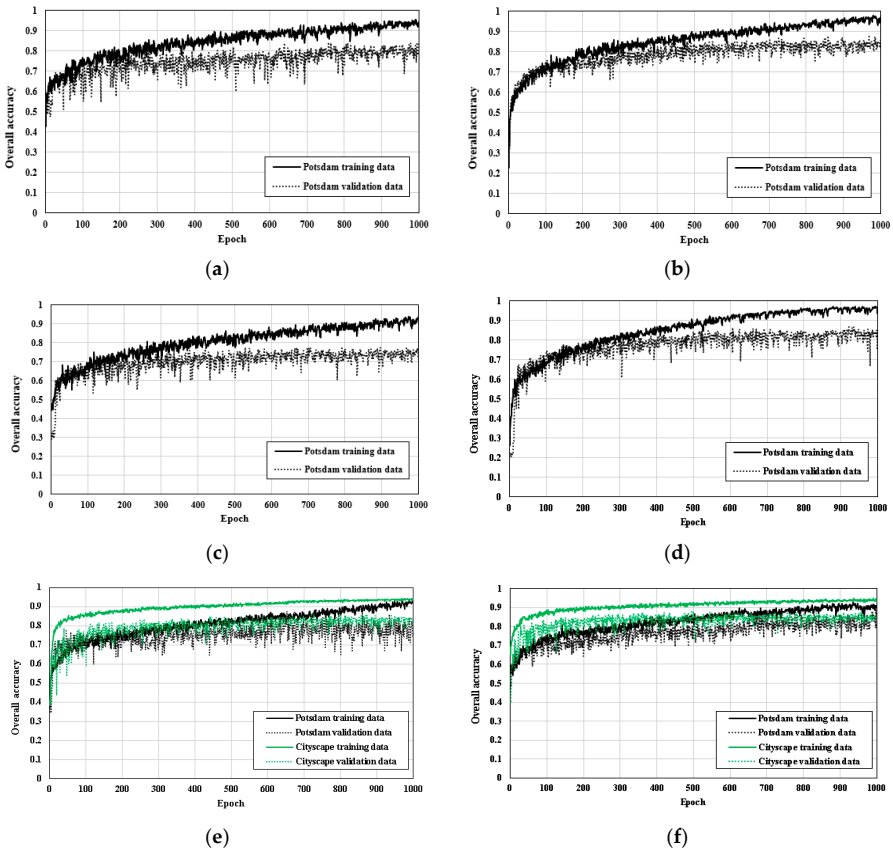

Figure 6. Learning graph of the overall accuracy (OA) for each epoch. (**a**) SegNet with original bands of Potsdam dataset, (**b**) DeepLab-V3+ with original bands of Potsdam dataset, (**c**) simplified U-net with RGB bands of Potsdam dataset, (**d**) simplified U-net with original bands of Potsdam dataset, (**e**) Case 1 wherein training proceeded using both the Potsdam and Cityscape datasets (same sizes) by the combined U-net method. (**f**) Case 2 wherein training proceeded using the Potsdam and Cityscape datasets by the combined U-net method; however, the Cityscape dataset was about twice as large as the Potsdam dataset.

When comparing between the simplified U-net model training with original four bands and Case 1 in the combined model, it can be seen that the OAs of the validation sets of both are similar, but the OAs of the training set in Case 1 were lower than those in the simplified U-net model (Figure 6d,e). Although the combined U-net model used for Case 1 trained the Potsdam and Cityscape datasets together, it used only the RGB-images from the Potsdam data. In contrast, the simplified U-net model also used the NIR band of the Potsdam data rather than simply using only the RGB bands, which aided the meaningful classification of objects such as trees and low vegetation that are particularly prominent at NIR wavelengths.

In Case 2 of the combined U-net model, in which the size of the Cityscape dataset was increased relative to Case 1, the OAs of the training and validation sets further improved relative to Case 1. Although the OA of the training data in Case 2 was lower than that of DeepLab-V3+ and simplified U-net, the gap in training and validation accuracy was decreased. As the amount of data in the Cityscape dataset increased, the overfitting problem was effectively reduced. For the OAs of the Cityscape dataset in Case 1 and Case 2, they showed a similar tendency; furthermore, as the amount of data increased, the validation accuracy of the Cityscape dataset improved.

Table 3 shows the average F1 scores of the five classes and the OA of the test set for the six models. The OAs of the SegNet, DeepLab-V3+, simplified U-net models with the two cases, and combined U-net models with the two cases are 0.8346, 0.8605, 0.8477, 0.7841, 0.8268 and 0.8721, respectively. Among the single models, the OA of the DeepLab-V3+ was highest, and Case 2 of the proposed combined U-net model had a higher OA than that of the DeepLab-V3+. The simplified U-net model trained using only the RGB bands of the Potsdam images had the lowest OA.

In addition, in the simplified U-net trained using original four bands, the F1 scores of the impervious surface, building, low vegetation, tree, and car classes are 0.8623, 0.8535, 0.8205, 0.8457 and 0.8123, respectively. Although Case 1 in the combined U-net model exhibited higher F1 scores for the impervious surface and building classes, it showed lower F1 scores for the vegetation and tree classes. In particular, Case 2 in the combined U-net model exhibited higher F1 scores for the impervious surface, building, and car classes. The road boundaries and car shapes were well predicted in Case 2. In order to visually analyze the segmentation results when using the combined U-net model rather than when using the simplified U-net model, we selected sites that can show the characteristics of three cases as an example, and the semantic segmentation results are shown in Figure 7. In the simplified U-net model trained with original four bands, there were errors in classifying roads and buildings; however, compared with Cases 1 and 2, low vegetation and tree classes were effectively distinguished (Figure 7a–j). In Case 1, materials on the roof were misclassified as car because their shape and colors were similar (Figure 7k–o). Case 2 showed the most efficiency in classifying buildings and roads; however, it could not clearly distinguish between trees and low vegetation (Figure 7k–u).

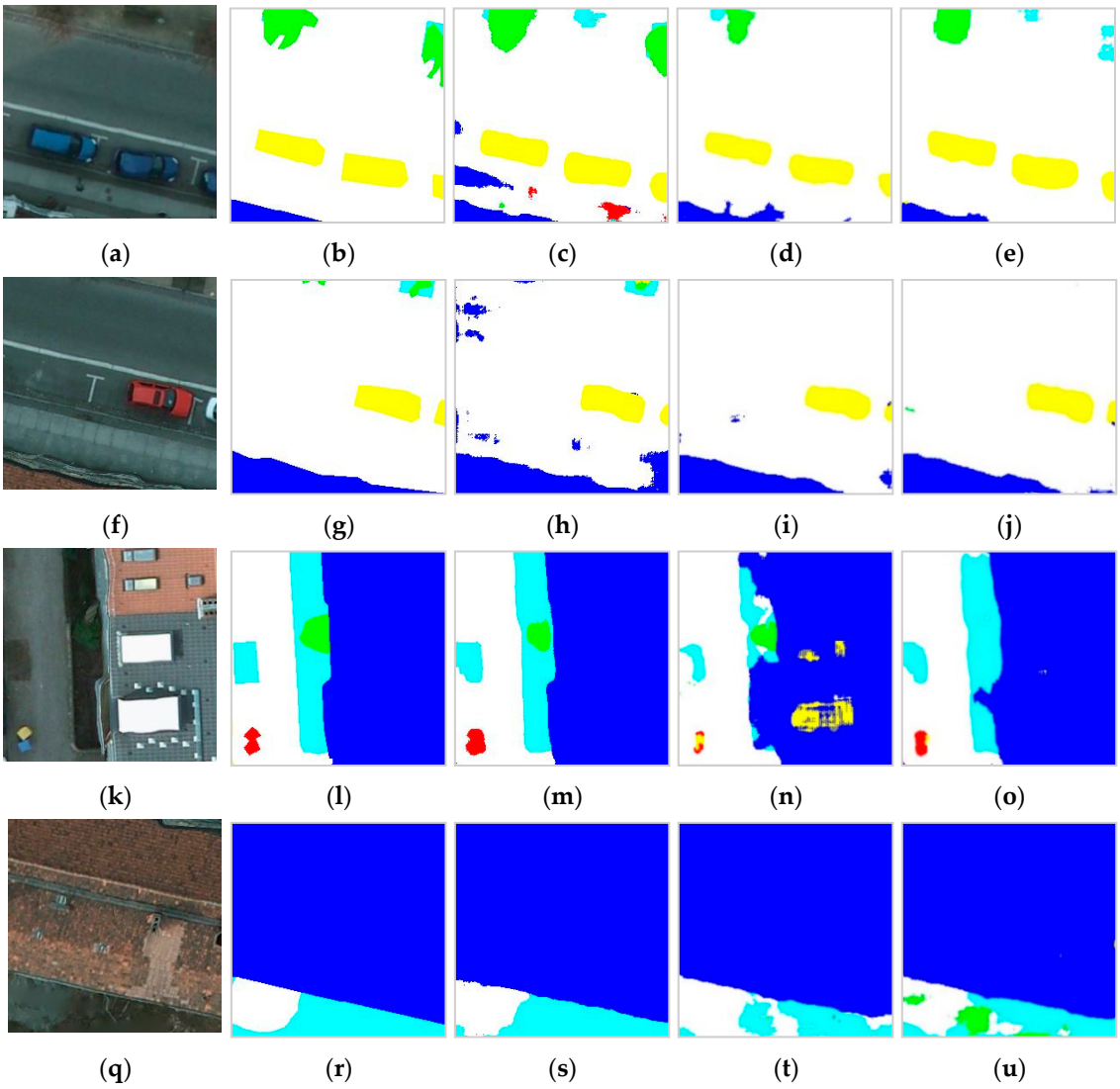

Figure 7. Example of input Potsdam RGB, label, and the resulting semantic segmentation images for the three cases. (**a**,**f**,**k**,**q**) are the input Potsdam images, (**b**,**g**,**l**,**r**) are the label images, (**c**,**h**,**m**,**s**) are the resulting images generated in simplified U-net, (**d**,**i**,**n**,**t**) are the resulting images generated in Case 1, and (**e**,**j**,**o**,**u**) are the resulting images generated in Case 2.

6. Discussion

6.1. Comparison with Other Algorithms

Among the single models, the DeepLab-V3+ model had higher segmentation OAs for the Potsdam images than the SegNet and simplified U-net models. In particular, the simplified U-net model had the lowest OA among the other networks compared with when training with RGB-images only. This is because the segmentation accuracies of the tree and low vegetation classes were reduced when training without the NIR band. In this context, although the Potsdam dataset was trained using an identically sized Cityscape dataset in Case 1, the OA and F1 scores of the vegetation, tree, and car classes showed no improvements compared with DeepLab-V3+ and simplified U-net. The results show that training using the Potsdam data in its original four bands is more effective in classifying vegetation classes than that using just the RGB bands. This is because the NIR band is a key band for classifying trees and low vegetation; in addition, several trees and low vegetation with low reflectance had similar colors to ground and impervious surfaces in the RGB-images. However, the semantic segmentation accuracy was improved by increasing the size of the Cityscape dataset when training using the combined

U-net model. In particular, the F1 scores of impervious surface, buildings, and cars improved in Case 2 relative to other networks. Furthermore, when training Potsdam using the Cityscape dataset, the overfitting problem was decreased. This is because although there are various roads, cars, and buildings in the Cityscape dataset and the shooting angle and shapes are different relative to such objects in the Potsdam images, there may be common points in terms of their color and/or relationships with the surrounding environment.

6.2. Cityscape Dataset Impact

By comparing the results of Case 1 and Case 2, it was confirmed that the combined U-net model was affected by the number of Cityscapes. We varied the number of Cityscape training data and changed the weight values of the Potsdam and Cityscape datasets (w_1 and w_2) (Figure 8). When the same number of Potsdam datasets was used; for example, 1600 images were used as training data and 400 and 150 images were used as validation and test data, respectively, the accuracy of the Potsdam test data improved with the increase in the training data of the Cityscape. In particular, when the number of Cityscape images was about 2500–2900, the OA of the combined U-net model became similar to that of the single models such as SegNet, simplified U-net, and DeepLab-V3+, and the number of Cityscape images was 3000, which is about twice that of the Potsdam dataset, so the OA dramatically increased (Figure 8a).

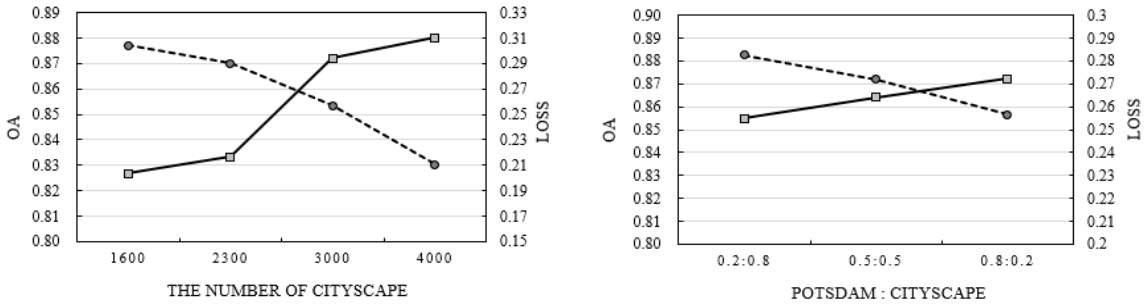

Figure 8. OA and loss of the test set when (**a**) the number of Cityscape datasets was varied and when (**b**) the weight values of the Potsdam and Cityscape were changed

In the experiments, to give larger weights to the Potsdam dataset, w_1 and w_2 were fixed to 0.8 and 0.2, respectively. We examined the weight values effect through an experiment, where w_1 was set to 0.2 and w_2 was set to 0.8, where the weights of the two datasets were the same. In this case, the number of Potsdam and Cityscape datasets was the same as in Case 2 of the combined U-net model. As a result, with the decrease in the weight of the Potsdam dataset, the OA of the test set decreased. This is because the loss due to the Potsdam dataset was less reflected when training the combined U-net model (Figure 8b). In particular, when w_1 and w_2 were set to 0.2 and 0.8, respectively, the OA was less than with the DeepLab-V3+ model.

6.3. Limitations and Future Work

Although the combined U-net model can train heterogeneous datasets, its network structure is relatively complex, and it takes a long time to perform learning processes compared with network learning with a single dataset. For example, when training the combined U-net model in Google Colab, when the batch size was set to be more than 4 (8 or 16), we faced a memory problem. Furthermore, there was a limitation in that the performance of the combined U-net model was changed according to the number of Cityscapes and weight values.

In addition, since RS-images generally have more than four bands, including the NIR band, not using the additional bands for learning with natural RGB-images can lower the segmentation accuracy

of vegetation-related classes. To overcome these limitations, future work is needed to develop a method that can effectively include the NIR band of RS-images while learning with natural RGB-images. Furthermore, to examine the effect of the architecture of the combined U-net model, we plan to see how the OA of the Potsdam dataset changes when the shared phase in the encoder and decoder part of the combined U-net model is changed (e.g., non-shared encoder and shared decoder).

7. Conclusions

In this paper, we proposed the combined U-net model that can train RS-images using a natural-image dataset. The network composed of encoder and decoder blocks, and the encoder blocks were shared with the two different datasets (Potsdam and Cityscape); the network was updated using the combined weighted loss function. The results obtained from the experiments indicated that when training using the identically sized Potsdam data with RBG bands and Cityscape data, the OA was decreased compared with single models training using only the original Potsdam data. However, the accuracy of the Potsdam dataset improved with an increase in the size of the Cityscape dataset. These results show that the use of a large-scale natural-image dataset can improve the semantic segmentation accuracy of the RS-image dataset using the proposed method. The proposed method can solve the problem of insufficiently large RS-image datasets for semantic segmentation. Furthermore, this study confirms the possibility of learning heterogeneous datasets at the same time by sharing the encoder phase and generated weights from two datasets in deep learning networks. It is expected that this approach can not only be applied to segmentation tasks of aerial images but also to tasks with various purposes of using big heterogeneous datasets. For example, when using a relatively limited number of datasets such as newly constructed floorplans and street-view images on web sites for special tasks such as classification and object detection, the big datasets, such as existing scanned floorplans and satellite or unmanned aerial vehicle images which are datasets acquired from different times and sensors but with similar characteristics, can be used to solve the problem of insufficient training data in input dataset by sharing specific part of deep learning networks.

However, the computational burden of the proposed method is relatively high because the combined U-net model trains heterogeneous datasets at the same time. In addition, the segmentation accuracy can be changed according to the number of Cityscape datasets and the weight values between Potsdam and Cityscape. Also, there was a problem in that the provided information by the NIR band could not be used because only the RGB bands were used. For future work, we aim to improve the combined U-net model structure by conducting experiments in regard to including the NIR band information of the Potsdam dataset.

Author Contributions: Conceptualization, Methodology, Software, Formal analysis, Investigation, Ahram Song and Yongil Kim; Resources, Validation, Data curation, Writing (original draft preparation), Funding acquisition, Visualization, Ahram Song; Writing (review and editing), Supervision, Project administration Yongil Kim. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education NRF-2019R1I1A2A01058144 and the BK21 FOUR research program of the National Research Foundation of Korea.

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

References

1. Li, H.; Cai, J.; Nguyen, T.N.A.; Zheng, J. A benchmark for semantic image segmentation. In Proceedings of the IEEE International Conference on Multimedia and Expo, San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
2. Yu, H.S.; Yang, Z.G.; Tan, L.; Wang, Y.N.; Sun, W.; Sun, M.G.; Tang, Y.D. Methods and datasets on semantic segmentation: A review. *Neurocomputing* **2018**, *304*, 82–103. [CrossRef]
3. Li, Y.; Zhang, H.; Xue, X.; Jiang, Y.; Shen, Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1264–e1280. [CrossRef]

4. Everingham, M.; Eslami, S.M.A.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2014**, *111*, 98–136. [CrossRef]
5. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
6. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 1–15.
7. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
8. Sun, W.; Wang, R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
9. Kemker, R.; Salvaggio, C.; Kanan, C.W. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, 60–77. [CrossRef]
10. Rahman, M.T. Detection of land use/land cover changes and urban sprawl in Al-Khobar, Saudi Arabia: An analysis of multi-temporal remote sensing data. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 15. [CrossRef]
11. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [CrossRef]
12. Hu, J.; Li, L.; Lin, Y.; Wu, F.; Zhao, J. A comparison and strategy of semantic segmentation on remote sensing images. In Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Kunming, China, 20–22 July 2019; pp. 21–29.
13. Shelhamer, E.; Long, J.; 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.
14. Jiao, L.; Liang, M.; Chen, H.; Yang, S.; Liu, H.; Cao, X. Deep Fully Convolutional Network-Based Spatial Distribution Prediction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5585–5599. [CrossRef]
15. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Interventions, Munich, Germany, 5–9 October 2015; pp. 234–241.
17. Feng, W.; Sui, H.; Huang, W.; Xu, C.; An, K. Water body extraction from very high-resolution remote sensing imagery using deep u-net and a super pixel -based conditional random field model. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 618–622. [CrossRef]
18. Audebert, N.; Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
19. Liu, Y.; Ren, Q.; Geng, J.; Ding, M.; Li, J. Efficient Patch-Wise Semantic Segmentation for Large-Scale Remote Sensing Images. *Sensors* **2018**, *18*, 3232. [CrossRef] [PubMed]
20. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U. Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 105–109. [CrossRef]
21. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Prague, Czech Republic, 12–19 July 2016.
22. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. In Proceedings of the Computer Vision—ACCV, Taipei, Taiwan, 20–24 November 2016; Springer: Cham, Switzerland, 2016; pp. 180–196.
23. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *1*, 293–298. [CrossRef]

24. Volpi, M.; Ferrari, V. Semantic segmentation of urban scenes by learning local class interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 1–9.
25. Dstl Satellite Imagery Feature Detection. Available online: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/overview/description (accessed on 1 December 2018).
26. Meletis, P.; Dubbelman, G. Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation. In Proceedings of the IEEE Intelligent Vehicles Symposium, Changshu, Suzhou, China, 26–30 June 2018; pp. 1045–1050.
27. Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; Igel, C. Detection of traffic signs in real- world images: The german traffic sign detection benchmark. In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013.
28. Ghassemi, S.; Fiandrotti, A.; Francini, G.; Magli, E. Learning and adapting robust features for satellite image segmentation on heterogeneous datasets. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6517–6529. [CrossRef]
29. Liang, Y.; Monteiro, S.T.; Saber, E.S. Transfer learning for high resolution aerial image classification. *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop* **2016**, *10*, 1–8.
30. Lee, H.; Eum, S.; Kwon, H. Cross-domain CNN for hyperspectral image classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 3627–3630.
31. ISPRS WG III/4. ISPRS 2D Semantic Labeling Contest. Available online: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (accessed on 16 January 2020).
32. Carneiro, T.; Da Nóbrega, R.V.M.; Nepomuceno, T.; Bian, G.B.; De Albuquerque, V.H.C.; Filho, P.P.R. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access* **2018**, *6*, 61677–61685. [CrossRef]
33. Huang, Z.; Pan, Z.; Lei, B. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. *Remote Sens.* **2017**, *9*, 907. [CrossRef]