remote sensing

MDPI

Article

Effective Fusion of Multi-Modal Remote Sensing Data in a Fully Convolutional Network for Semantic Labeling

Wenkai Zhang [1,2], Hai Huang [3,*], Matthias Schmitz [3], Xian Sun [1], Hongqi Wang [1] and Helmut Mayer [3]

[1] Key Laboratory of Spatial Information Processing and Application System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China; zhang.wenkai@outlook.com (W.Z.); sunxian@mail.ie.ac.cn (X.S.); wiecas@sina.com (H.W.)

[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Huairou District, Beijing 100049, China

[3] Institute for Applied Computer Science, Bundeswehr University Munich, Werner-Heisenberg-Weg 39, D-85577 Neubiberg, Germany; matthias.schmitz@unibw.de (M.S.); helmut.mayer@unibw.de (H.M.)

* Correspondence: hai.huang@unibw.de; Tel.: +49-89-6004-3447

Received: 9 November 2017; Accepted: 28 December 2017; Published: 29 December 2017

Abstract: In recent years, Fully Convolutional Networks (FCN) have led to a great improvement of semantic labeling for various applications including multi-modal remote sensing data. Although different fusion strategies have been reported for multi-modal data, there is no in-depth study of the reasons of performance limits. For example, it is unclear, why an early fusion of multi-modal data in FCN does not lead to a satisfying result. In this paper, we investigate the contribution of individual layers inside FCN and propose an effective fusion strategy for the semantic labeling of color or infrared imagery together with elevation (e.g., Digital Surface Models). The sensitivity and contribution of layers concerning classes and multi-modal data are quantified by recall and descent rate of recall in a multi-resolution model. The contribution of different modalities to the pixel-wise prediction is analyzed explaining the reason of the poor performance caused by the plain concatenation of different modalities. Finally, based on the analysis an optimized scheme for the fusion of layers with image and elevation information into a single FCN model is derived. Experiments are performed on the ISPRS Vaihingen 2D Semantic Labeling dataset (infrared and RGB imagery as well as elevation) and the Potsdam dataset (RGB imagery and elevation). Comprehensive evaluations demonstrate the potential of the proposed approach.

Keywords: semantic labeling; Fully Convolutional Networks; multi-modal dataset; fusion nets

1. Introduction

Semantic labeling is of great interest for scene understanding and object detection. The introduction of Fully Convolutional Networks (FCN) by Long et al. [1] as a special variant of Convolutional Neural Networks (CNNs) [2,3] has introduced a new means for semantic labeling. Many approaches [1,4–13] based on FCN have been proposed to improve the performance of semantic labeling.

FCNs consist of convolution, ReLU (Rectified Linear Unit), pooling, deconvolution, and classification layers. The convolution layers extract features and encode location as well as semantic information together with deconvolution layers, while the classification layers generate coarse predictions simultaneously. FCN is the first pixel-wise prediction model which can be trained end-to-end and pixel-to-pixel (pixel-wise). This model can take input of arbitrary size and produce correspondingly sized output with efficient inference and learning. In contrast, previous approaches have applied small CNNs without efficient supervised pre-training [14,15]. It is not only used for image data, but also to analyze multi-modal data, i.e., imagery and elevation. Their utilization can improve the quality of semantic labeling. Elevation in the form of a DSM (Digital Surface Model) contains the different heights of objects, which can help to discriminate impervious surfaces from buildings as well as trees from low vegetation. There are two popular ways to utilize multi-modal data: The first is that imagery and elevation are concatenated into a 4-channel (RGB + elevation) vector as the input of the model [1,11]. The other is the ensemble way [10], where imagery and elevation are input into dual streams and trained in separate networks which are merged at the end in the prediction layer. The former way does not produce satisfying prediction results, as the qualitative distinction of different data modalities is ignored. The latter shows an improved performance, but has a complex structure and, thus, requires more computational effort.

In [16,17] methods have been proposed to visualize and understand how deep neural networks for imagery work. Zeiler et al. [16] propose a deconvolution method to visualize that various regions of a layer may have different sensitivities during learning. Zintgraf et al. [17] present a probabilistic methodology which generates a saliency map explaining which parts of the input provide the evidence for the prediction. The two methods present different ways to understand the deep neural networks from the inside. However, they focus only on imagery and ignore the traits of multi-modal data.

In previous work [18], we have proposed a multi-resolution model as a basis to understand the layers' sensitivity to specific classes of multi-modal data. It is quantified by the recall and descent rate of recall. This helps to comprehend what the layers learn and to explain why an early fusion cannot produce a satisfying result. The contributions of different modalities for the pixel-wise prediction have been examined based on the proposed model. Finally, different strategies of layer fusion for imagery and elevation information have been analyzed to find an optimal position (i.e., a specific layer) for an effective data fusion.

This paper extends [18] with improvements and a detailed presentation of the proposed method, further experiments on additional data, and a comprehensive analysis of the results. Additional implementation details of the multi-resolution model and the fusion models are given. The contribution of different modalities for the pixel-wise prediction is visualized and quantified by an evaluation. Experiments are performed on two datasets investigating the effectiveness of fusion models for elevation with both infrared and RGB color data.

The rest of the paper is organized as follows. The related work on FCN for semantic labeling is described in Section 2. Section 3 introduces the proposed multi-resolution model, fusion model and fusion strategy. The experiments and analysis of the sensitivity and contribution of individual layers to specific classes are described in Section 4. The paper ends up with a conclusion in Section 5.

2. Related Work

Our approach is based on an FCN for semantic labeling of imagery and elevation data. An FCN model leads to state-of-the-art performance for pixel-wise prediction for imagery. However, an early fusion of the multi-modal data does not produce satisfying results. In the following we describe the development of FCN for imagery and present existing methods for processing multi-modal remote sensing data.

2.1. FCN for Images

Much research has focused on remote sensing images. The results have been greatly improved since FCN has been proposed by Long et al. [1].

FCN is the first pixel-wise prediction model which can be trained end-to-end and pixel-to-pixel. The basic model named FCN-32s consists of convolution layers extracting features, deconvolution layers, and classification layers generating coarse predictions. Because of the large stride of FCN-32s, it generates dissatisfying coarse results. To overcome this drawback, a skip architecture model, which directly makes use of shallow features and reduces the stride for up-sampled prediction, has been proposed in [1]. This skip model fuses several predictions from shallow layers with deep layers. A model named DeepLab presented in [8] improves the coarse spatial resolution caused by repeated pooling by means of the so called "atrous" convolution. The authors employ multiple parallel "atrous" convolutions with different rates to extract multi-scale features for the prediction. Fully-Connected Conditional Random Fields (CRFs) are applied for smoothing the output score map to generate a more accurate localization. Badrinarayanan et al. [19,20] have proposed a model named SegNet. It is composed of two stacks: The first stack is an encoder that extracts the features of the input images, while the second stack consists of a decoder followed by a prediction layer generating pixel labels. CRF–RNN [7] fully integrates the CRF with Recurrent Neural Networks (RNN) instead of applying the CRF on trained class scores, which makes it possible to train the network end-to-end with back-propagation.

2.2. FCN for Multi-Modal Data

Several researchers have proposed different methods based on FCN to label high-resolution multi-modal remote sensing data. Kampffmeyer et al. [21] have presented an FCN model which removes the fifth convolution layer as well as the fully convolutional layers and keeps the first four convolutions and the up-sampling layers. RGB, DSM and normalized DSM are concatenated into a 5-channel vector used as input. Audebert et al. [22] introduce an improved model based on SegNet, which includes multi-kernel convolution layers for multi-scale prediction. They use a dual-stream SegNet architecture, processing the color and depth images simultaneously. Maggiori et al. [13] present an ensemble dual-stream CNN to combine color with elevation information. Imagery and DSM data are employed in two separate streams. The features derived from color and elevation are only merged at the last high-level convolution layer before the final prediction layer. Sherrah et al. [12] propose a model based on FCN without down-sampling. It preserves the output resolution, but it is time-consuming. It can process not only color imagery, but also the combination of color and elevation data. Color and elevation features are merged before the fully convolutional layer. Concatenating color and depth as 4-channel vector is the most common strategy. It does, however, usually not produce satisfying results. Gupta et al. [23] proposed a new representation of depth, which consists of three different features named HHA consisting of disparity, height of the pixels and the angle between the normal direction and the gravity vector. Long et al. [11] combine RGB and HHA by late fusion averaging the final scores from both networks. Hazirbas et al. [9] explore a network for fusion based on SegNet to improve the semantic labeling for natural scenes.

3. Multi-Resolution Model and Fusion Framework

In this part we describe in detail the model and fusion strategy that we use. First, we introduce the multi-resolution model which is employed to quantify the layers' sensitivity to specific classes of multi-modal data. Then, we describe the fusion strategy. Finally, we propose a fusion model for an effective incorporation of multi-modal data.

3.1. Multi-Resolution Model

FCNs can directly incorporate multi-modal data such as image and elevation information as a n-channel input. However, it is not clear why an early fusion does not lead to a satisfying result [1,12]. Basic FCNs are not suitable to analyze how sensitive the layers are concerning each class of data and different modalities. To this end, we, thus, propose an improved FCN model named "multi-resolution model" (Figure 1). It is similar to FCN-8s proposed by Long et al. [1] which has fifteen convolution layers. We add a shallower layer, i.e., the second layer, as a skip layer to generate predictions. Each skip consists of a convolution layer with kernel size 1×1 and an up-sampling layer (including a subsequent cropping layer to remove the margin caused by pooling and padding) to obtain the same resolution as the original data. The up-sampling layer is implemented using bilinear interpolation, and the up-sampling factors are 4, 8, 16, 32 for layer 2, 3, 4, 5, respectively. The element-wise sum of the up-sampling layers is the input to the classification layer generating the prediction. The nets are trained separately on imagery and elevation data to analyze the contribution of different modalities to each class. They are referred to as image model and elevation model, respectively.

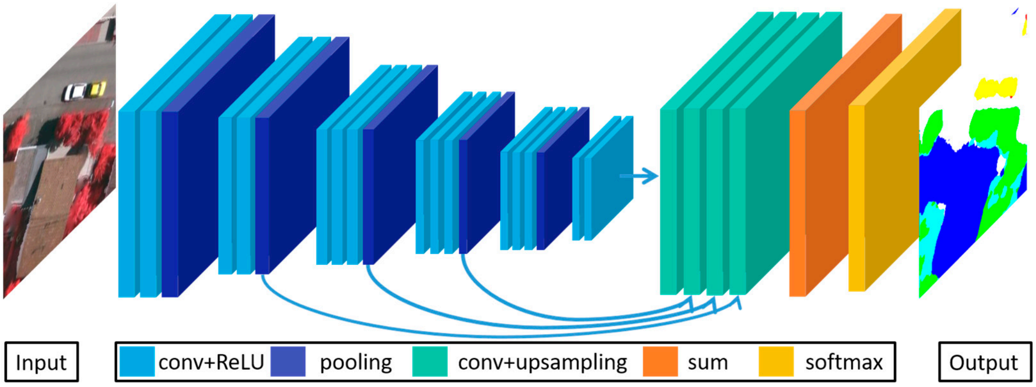


Figure 1. Multi-resolution model for the analysis of the contribution of individual layers to each class. conv: Convolution layer, relu: Rectifying Linear Unit (nonlinearity), sum: Element-wise sum.

3.2. Fusion Strategy

This section explains the fusion strategy, namely the element-wise summation after the ReLU layer. Given, is a training set $\{(X_i, Y_i), for\ i = 1, 2, \ldots, M\}$ with number of samples M, $X_i \in R^{C*H*W}, Y_i \in R^N$, where C is the number of channels (image and elevation), H and W are the height and width of the input data and N is the label set size. For an FCN model with L layers, we represent the weights as $(W, b) = \{W^{(1)}, b^{(1)}, \ldots, W^{(l)}, b^{(l)}, \ldots, W^{(L)}, b^{(L)}\}$, where $W^{(l)}$ is the parameter associated with the connection between layer l, and layer $l + 1$, and $b^{(l)}$ is the bias in layer $l + 1$. Activations in convolution layer l are derived as follows:

$$Z^{(l)} = b^{(l)} + W^{(l-1)} A^{(l-1)} \tag{1}$$

The outputs of ReLU layer l are:

$$A^{(l)} = f(Z^l) \tag{2}$$

where f is the activation function. We employ the ReLU function in this paper with the definition $\max(0, x)$. Let $A^{(1)}$ be equal to $X^{(1)}$.

For data of different modality, we obtain $Z(image)$ and $Z(elevation)$. If we fuse the data before the ReLU layer, the fusion layer's output is

$$Z^{(l)} = Z^{(l)}(image) + Z^{(l)}(elevation) \tag{3}$$

This fusion strategy weakens the outputs in ReLU layer l. Given the two real-valued $Z^{(l)}(image)$ and $Z^{(l)}(elevation)$ the following holds:

$$f(Z^{(l)}(image) + Z^{(l)}(elevation)) \leq f(Z^{(l)}(image)) + f(Z^{(l)}(elevation)) \tag{4}$$

The input of an activation function can be negative, while the output of an activation function is always non-negative. The sum can strengthen the activation while it preserves most features of the multi-modal data. We accordingly adopt the fusion strategy of element-wise summation after the ReLU layer.

3.3. Fusion Model

Figure 2 illustrates the fusion model which incorporates imagery and elevation in one model instead of an ensemble of models. The proposed model consists of two parts: The encoder part extracts features and the decoder part up-samples the (heat-) maps which contain the probability to which class the pixels belong on the original image resolution. The encoder part consists of the left two streams in Figure 2. The upper is the elevation channel and the lower the image channel. The elevation is normalized to the same range of values as the image, i.e., [0,255]. In order to incorporate imagery and elevation in one model, we fuse the feature maps from the elevation stream with those from the image stream. An element-wise summation fusion strategy (cf. Section 3.2) is applied for the fusion layers, shown as red boxes in Figure 2. The fusion layers are inserted before the pooling layers. This fusion strategy helps to preserve the essential information of both streams. Since the fused feature maps preserve more useful information, the network extracts better high-level features, which in turn enhances the final accuracy. We denote the fusion model by "Fusion" followed by the number of the fusion layers used in the FCN (cf. Figure 2). The decoder part consists of the convolution layer, the up-sampling layer, the crop layer and the sum layer. The pooling layer is followed by a 1 × 1 convolution with channel dimension 6 to predict scores for each class. The up-sampling layer is conducted using bilinear interpolation to generate a dense prediction. The up-sampling factors are set to 4, 8, 16 and 32 for the layers 2, 3, 4 and 5, respectively. The cropping layer removes the margin caused by pooling and padding. The offset of the cropped region depends on the re-sampling and padding parameters of all intermediate layers. After the cropping layer, an element-wise sum of the layers is used as the input to the classification layer generating the final prediction. If the fusion is conducted at Layer 1 ("Fusion1"), the elevation is fused to the color stream in layer 1. The up-sampling layers are layers 5, 4, 3, 2 of the color stream. If the fusion occurs at Layer 2 ("Fusion2"), the up-sampling layers are the same as "Fusion1". Yet, when the fusion is accomplished at Layer 3, the up-sampling layers are layers 5, 4, 3, 2 from the color stream and Layer 2 from the elevation stream.

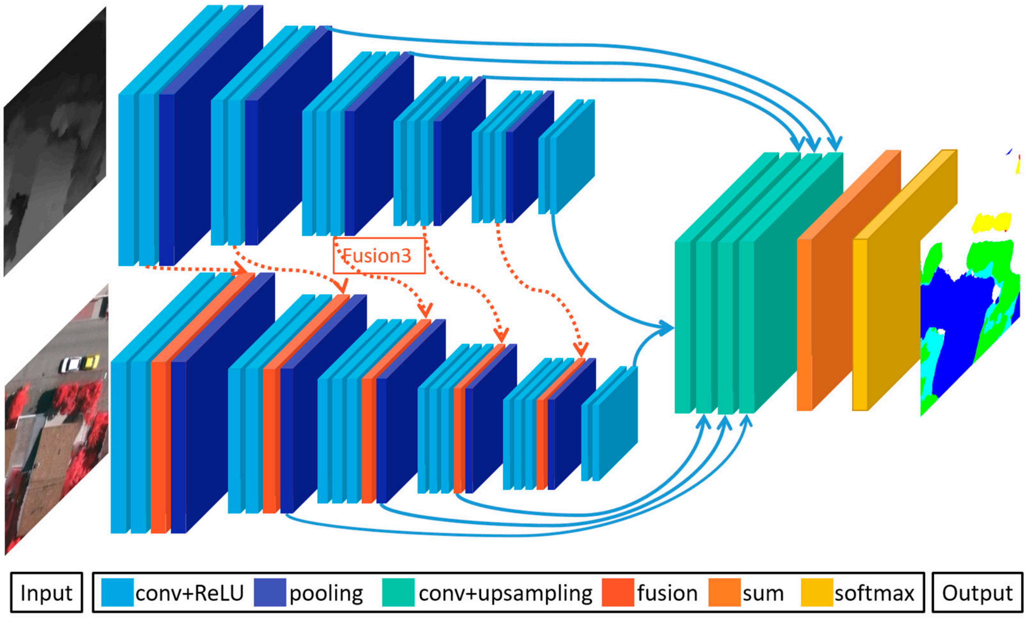


Figure 2. Fusion model incorporating multi-modal remote sensing data: Imagery and elevation. Fusion 3 means that the dual streams (elevation and color) are fused in layer 3, the same holds for Fusion1 up to Fusion5. The fusion strategy works element-wise after the ReLU layer.

4. Experiments

We employ the ISPRS datasets of Vaihingen and Potsdam [24] for experiments. The training of the multi-resolution and the fusion model is described and a quantitative evaluation is given. We especially have conducted a series of experiments concerning the layers' sensitivity to specific classes as well as different modalities and visualize the results. The exploration of the layers' sensitivity indicates the proposed optimal fusion model is able to improve the performance. This fusion model is verified by a comparison of all possible fusion models.

4.1. Datasets

Experiments are performed on the ISPRS 2D semantic labeling datasets showing the urban areas of Vaihingen and Potsdam. The datasets provide high resolution True Orthophotos (TOPs) and corresponding DSMs with a Ground Sampling Distance (GSD) of 9 cm. In the experiments, we use color infrared imagery (infrared-red-green) for the Vaihingen dataset and color imagery (red-green-blue) for the Potsdam dataset to test our approach on different data, even though the Potsdam data contains also infrared imagery. The Vaihingen dataset consists of 32 tiles in total with ground truth released for half of them, designated for training and validation. The Vaihingen dataset employs pixel-wise labeling for six semantic categories (impervious surface, building, low vegetation, tree, car, and clutter/background). We used twelve of the 16 labeled tiles for training and four tiles for validation. For training, we divided the selected tiles into patches with a size of 256 × 256 pixels with 128 pixels' overlap. The patches are rotated (each time by 90 degrees) and flipped (top down, left right) for data augmentation. A set of 16,782 patches is generated. We selected four tiles (13, 15, 23, and 28) as the validation set. The validation set tiles are clipped into patches with a size of 256 × 256 pixels without overlap.

The Potsdam dataset consists of 38 tiles in total with ground truth released for 24. Each contains the infrared, red, blue, green channel and corresponding elevation. The resolution is 0.5 m. In the experiments, patches are divided into small patches with a size of 256 × 256 pixels. Patches smaller than 256 × 256 are discarded. A set of 11,638 patches is generated. 9310 of them are randomly selected as training set, the rest as test set.

4.2. Training

4.2.1. Multi-resolution model training

To investigate the contribution of different modalities for the interpretation of each class, we trained different nets separately on imagery and elevation data. The model for the imagery is only trained on color data and the elevation model is only trained on DSM data. All layers of the multi-resolution model are trained for 60,000 iterations. During training, we utilize a step policy to accelerate the convergence of the networks starting with a reasonably small learning rate and decreasing it based on the step-size. Net model VGG-16 [25] is employed with pre-trained (imagery only) initial parameters for fine tuning of the model. We begin with a learning rate $\text{lr}_{\text{base}} = 10^{-10}$, and reduce it by a factor of 10 every 20,000 iterations. The elevation data have to be learned from scratch. We use a high momentum of 0.99. The gradients of the nets are accumulated over 20 images. Each training process contains a forward pass, which infers the prediction results and compares them with ground truth labels to generate loss, and a backward pass, in which the weights of the nets are updated via stochastic gradient descent.

4.2.2. Fusion model training

The fusion model is trained on both color and elevation data simultaneously. All layers are trained together for 60,000 iterations. To accelerate the learning, we again utilize a step policy with a reasonably initial learning rate of $\text{lr}_{\text{base}} = 10^{-10}$, decreasing it every 20,000 iterations by a factor of 10. The gradients of the nets are accumulated over 20 images. Accumulation accelerates the training of the nets and reduces the memory required. We again employ a high momentum of 0.99. The open source deep learning framework Caffe [26] is used for implementation and experiments.

4.3. Evaluation of Experimental Results

What layers learn about different objects can be represented by the recall for the specific class. The recall is the fraction of correct pixels of a class which are retrieved in the semantic labeling. It is defined together with the precision:

$$recall = \frac{TP}{TP + FN}; precision = \frac{TP}{TP + FP} \tag{5}$$

A higher recall of a layer for a specific class indicates in turn a higher sensitivity to the class. However, the recall for a single layer is not a reasonable measure for the sensitivity towards classes. When the recall for two layers is computed, it is useful to employ the descent rate of recall, defined as difference to the previous recall, to determine which layer has the primary influence. Thus, we evaluate the contribution of layers for each class based on recall and the descent rate of recall (DR) defined as:

$$DR = \begin{cases} \max(\frac{|R_m - R_l|}{R_m}, 0), & if\ R_m \geq 0.5 \\ 0, & if\ R_m \leq 0.5 \end{cases} \tag{6}$$

where R_m is the recall of the model with all layers and R_l is the recall of the model with the specific layer removed. In our experiment, we only compute the DR of a specific class when the recall of the class is larger than 50% and the DR is positive, otherwise the DR is set to 0.

To evaluate the fusion model, we use the $F1$ score and the overall accuracy (OA). The $F1$ score is defined in Equation (7). OA is the percentage of the correctly classified pixels, as defined in Equation (8).

$$F1 = 2\frac{precision \times recall}{precision + recall} \tag{7}$$

$$OA = \sum_i n_{ii} / \sum_i t_i \tag{8}$$

where n_{ij} is the number of pixels of class i predicted to belong to class j. There are n_{cl} different classes and t_i is the total number of pixels of class i.

4.4. Quantitative Results

Imagery is included at the beginning and elevation is integrated at different layers. Table 1 (image) and Table 2 (elevation) list the recalls of single and combined layers: 'layers-all' includes all layers, 'Layer-2' represents the recall of layer 2, 'Layers-345' represents the recall of all layers without layer 2, and so on up to 'Layers-234'.

Table 1. The recall of different layers for different classes in imagery.

Recall	Imp_Surf	Building	Low_Veg	Tree	Car	Clutter
Layers-all	0.86	0.86	0.72	0.84	0.43	0.52
Layer-2	0.19	0.04	0	0.69	0.6	0.91
Layers-345	0.84	0.88	0.84	0.67	0.28	0
Layer-3	0.59	0.03	0.01	0.01	0.86	0.66
Layers-245	0.78	0.88	0.74	0.86	0.05	0.49
Layer-4	0.83	0.83	0.65	0.88	0.01	0
Layers-235	0.26	0.53	0.5	0.39	0.5	0.91
Layer-5	0	0.99	0.23	0	0	0
Layers-234	0.88	0.78	0.45	0.93	0.66	0.47

Table 2. The recall of different layers for different classes in elevation data.

Recall	Imp_Surf	Building	Low_Veg	Tree	Car	Clutter
Layers-all	0.79	0.68	0.27	0.62	0	0
Layer-2	0.92	0	0.07	0.52	0	0.01
Layers-345	0.04	0.84	0.14	0.55	0	0
Layer-3	0.98	0	0	0.24	0	0
Layers-245	0.72	0.8	0.34	0.53	0	0
Layer-4	0.59	0.51	0.11	0.8	0	0
Layers-235	0.85	0.8	0.19	0.2	0	0
Layer-5	0	1	0.04	0	0	0
Layers-234	0.87	0.1	0.21	0.65	0	0

The recalls and *DR*s for imagery are summarized in Table 1 and Figure 3 respectively. One can see that for imagery layer 2 is mostly sensitive to the class car, the *DR* being 34%, and is slightly less sensitive to the class tree with a descent rate of 20.2%. Layer 3 is mostly sensitive to the classes car and impervious surface, the *DR* being 88% and 9.3%, respectively. Impervious surfaces reach the top *DR* value at layer 4, with trees, buildings, and low vegetation next. Layer 5 is only sensitive to buildings. They have the highest recall, although the *DR* is not steeper than for layer 4. Low vegetation, on the other hand, has the steepest *DR* at layer 5.

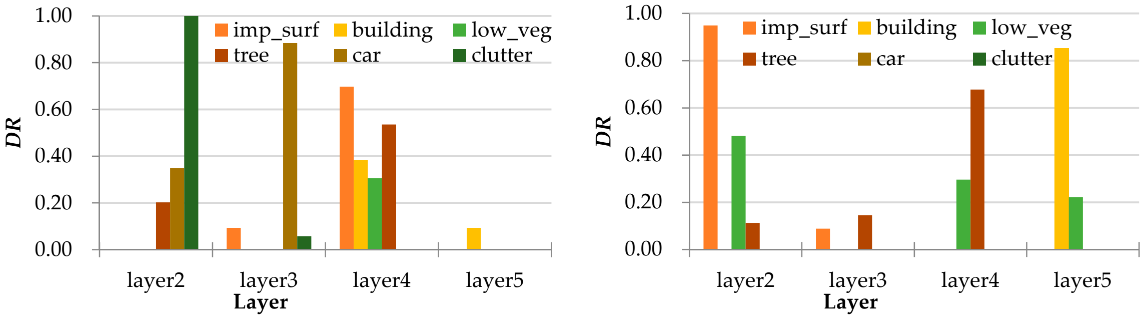
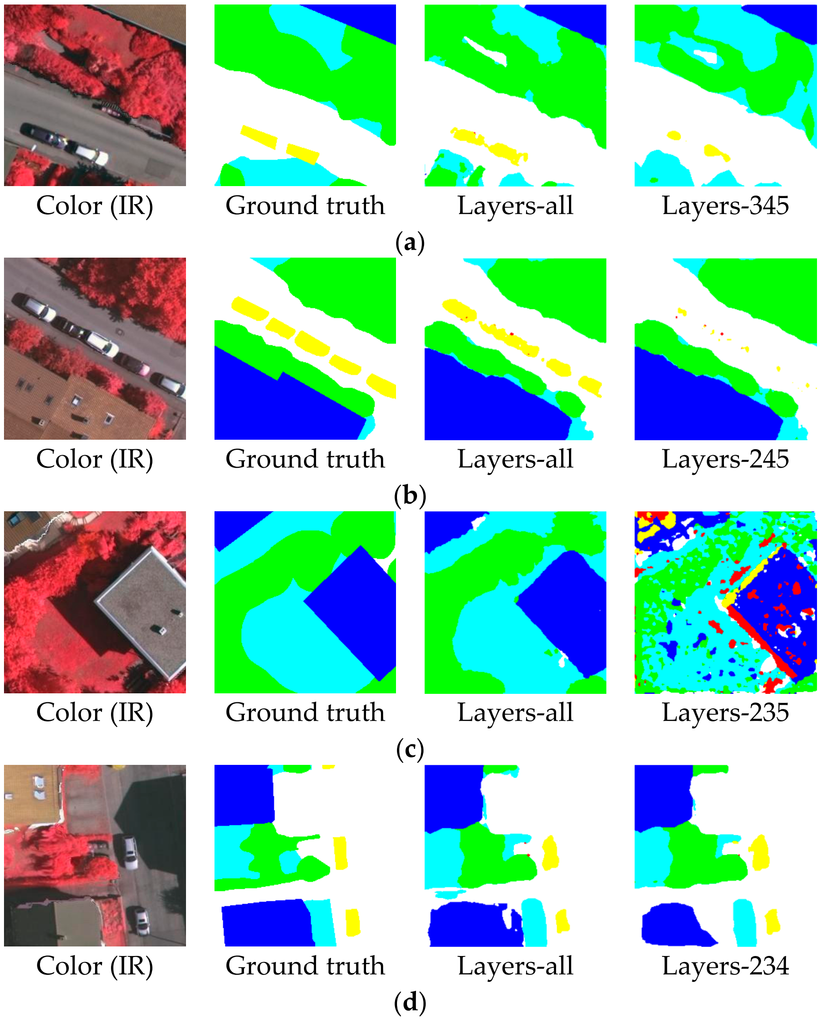


Figure 3. The *DR* (descent rate of recall) of different layers. The left graph is for imagery, the right for elevation data.

For imagery we, thus, conclude that the shallower layers containing the high resolution information, i.e., layers 2 and 3, are sensitive to small objects like cars. As demonstrated in Figure 4a,b, when layer 2 or 3 is removed, cars or trees, respectively, cannot be recognized any more. Deeper layers, i.e., layers 4 and 5, are sensitive to objects which comprise a more complex texture and occupy larger parts of the image, i.e., buildings and trees. Deeper layers also learn the discriminative parts and link them together. As shown in Figure 4c, when layer 4 is removed, the result is noisy, while adding layer 4 eliminates clutter. Without layer 5 parts of buildings are not detected, as shown in Figure 4d.

For the elevation data layer 2 is sensitive to impervious surfaces with a *DR* of 94%, while that for low vegetation is 48% and for tree 11%. Layer 3 reacts stronger to trees and impervious surfaces. The fourth layer is sensitive to trees and low vegetation. Layer 5 is essential for the correct classification of buildings. If the fifth layer is removed, the recall for buildings decreases from 0.68 to 0.1.

For the elevation data we, thus, conclude that the shallower layers show relatively higher sensitivities to impervious surfaces and low vegetation than other classes of objects. As demonstrated in Figure 5a, after removing Layer 2, the impervious surfaces and low vegetation are not segmented any more. Layer 3 reacts strongly to trees. Adding Layer 3 improves the classification of trees, as shown in Figure 5b. Layer 4 is essential for the overall classification accuracy. As shown in Figure 5c, when it is removed, the results are very noisy. When Layer 5 is removed, most of buildings are not detected anymore (Figure 5d).


Color (IR) Ground truth Layers-all Layers-345

(**a**)

Color (IR) Ground truth Layers-all Layers-245

(**b**)

Color (IR) Ground truth Layers-all Layers-235

(**c**)

Color (IR) Ground truth Layers-all Layers-234

(**d**)

Figure 4. The contribution of layers 2, 3, 4 for different classes in images: (**a**) When layer 2 is removed, half of the trees in the scene are classified as low vegetation; (**b**) without layer 3 cars are mostly not detected anymore; (**c**) when layer 4 is removed, the results are cluttered. However, adding layer 4 eliminates clutter; (**d**) without layer 5 parts of buildings are not detected. (White = impervious surface; dark blue = building; green = high vegetation; cyan = low vegetation).

We conclude that the layers for each modality have a different contribution for specific classes. Figure 3 shows that the *DR*s of different layers are not identical. The effectiveness of features and the sensitivities of classes are different for the layers. If the multi-modal data are fused too early, the layers might learn conflicting features, which leads to an unexpected outcome. This explains why an early fusion does not lead to a satisfying result.

4.5. Fusion Model Results

A detailed account of the overall accuracy of the fusion model for the Vaihingen data is given in Table 3. In Figure 6, we demonstrate some visual comparison of the fusion models. Color-D indicates a simple four-channel input. We denote the fusion model by "Fusion" followed by the number of the fusion layer used in the FCN. The results demonstrate that Fusion3 obtains the highest *OA* of 82.69% and all fusion models outperform the simple Color-D model. This verifies that the fusion nets improve urban scene classification compared to the early fusion of color and elevation data. We have learned from Section 4.4 that for imagery Layer 2 is sensitive to trees, cars and clutter/background. It reacts, however, most strongly to impervious surfaces and trees for the elevation data. Layers 3, 4, and 5 have a similar sensitivity for all classes.

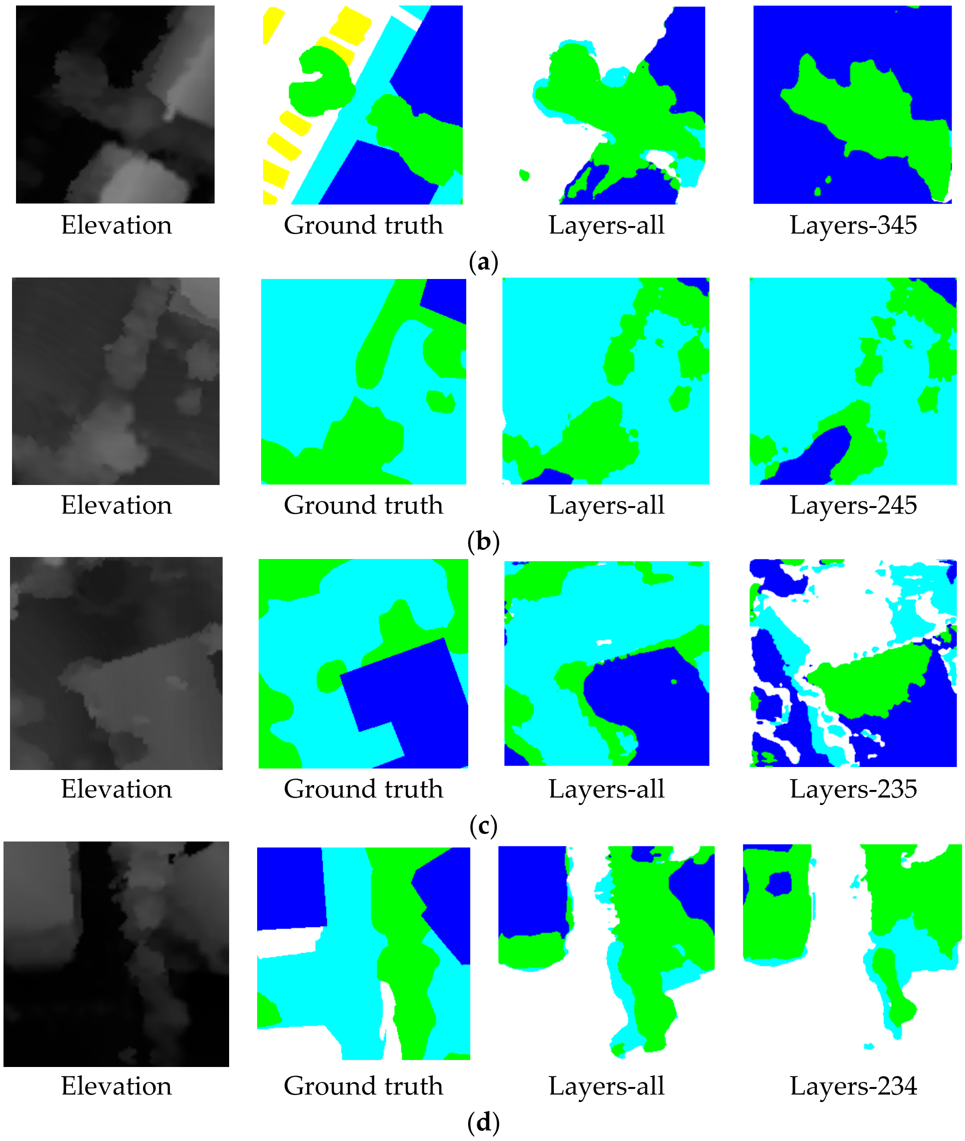
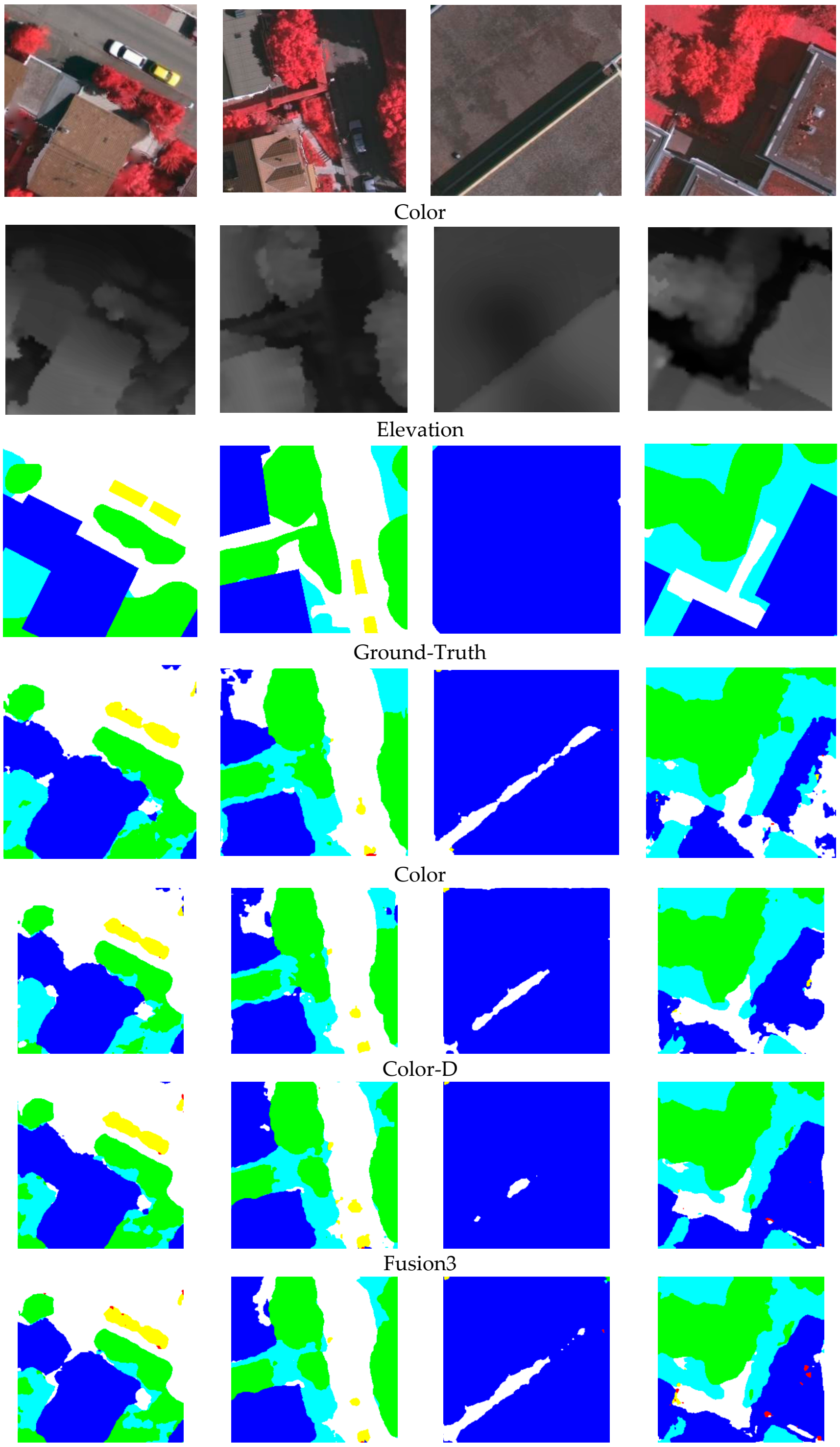


Figure 5. The contribution of layers 2, 3, 4 for different classes in elevation data: (**a**) Layer 2 contributes to impervious surfaces and low vegetation. When layer 2 is removed, they are no longer classified correctly; (**b**) shows the contribution of layer 3 for trees. Without layer 3, trees are mostly not detected anymore; (**c**) when layer 4 is removed, the results are noisy. Trees and low vegetation are not correctly classified; (**d**) without layer 5, large parts of buildings are not detected anymore.

Table 3. Overall Accuracy (*OA*) for dataset Vaihingen.

	Color-D	**Fusion1**	**Fusion2**	**Fusion3**	**Fusion4**	**Fusion5**	**Color**	**Elevation**
OA (%)	79.77	81.79	82.21	82.69	81.75	81.70	81.86	60.53
*F*1	0.80	0.81	0.82	0.83	0.81	0.81	0.81	0.58


Color

Elevation

Ground-Truth

Color

Color-D

Fusion3

Fusion5

Figure 6. Qualitative segmentation results. The first three rows give the input data color, elevation and ground-truth. The following rows present the segmentation results of different models. The last two rows show our fusion model results. The Fusion3 model obtains the best accuracy for the Vaihingen dataset.

In Table 4, we report F1 scores for the individual classes. The fusion models result in very competitive F1 scores. Concerning the six classes, Fusion3 outperforms all other for five of the six classes. When comparing the results for the individual features with the fusion model, we find that avoiding a particular conflicting layer (the layer having different sensitivity to specific classes) can improve the result for those classes. Thus, this investigation helps to effectively incorporate multi-modal data into a single model instead of using ensemble models with higher complexity and computational effort.

Table 4. Class-wise *F*1 score of six classes for dataset Vaihingen.

*F*1	Imp_Surf	Building	Low_Veg	Tree	Car	Clutter
Color-D	0.82	0.86	0.69	0.81	0.56	0.59
Fusion1	0.85	0.88	0.72	0.83	0.5	0.54
Fusion2	0.85	0.89	0.72	0.82	0.47	0.58
Fusion3	0.85	0.9	0.72	0.84	0.56	0.57
Fusion4	0.85	0.89	0.71	0.82	0.51	0.58
Fusion5	0.84	0.89	0.7	0.83	0.55	0.62
Color	0.84	0.87	0.72	0.84	0.54	0.66
Elevation	0.69	0.71	0.34	0.59	0	0

Results for the overall accuracy of the fusion model on the Potsdam data are presented in Table 5. Also these results demonstrate that Fusion models obtain a higher *OA* than other models. The later fusion produces a better *OA*. This additionally verifies that avoiding a particular conflicting layer can improve the result. In Table 6, we report *F*1 scores for the individual classes. Compared to RGB and RGB-D, our fusion model greatly improves the *F*1 score of building and low vegetation. This again demonstrates that the fusion model efficiently utilizes the multi-modal data.

Table 5. Overall Accuracy (*OA*) for dataset Potsdam.

	RGB-D	Fusion1	Fusion2	Fusion3	Fusion4	Fusion5	RGB	Elevation
OA (%)	79.21	80.35	80.99	81.64	81.78	81.80	78.31	63.16
*F*1	0.79	0.78	0.78	0.79	0.80	0.80	0.78	0.48

Table 6. Class-wise *F*1 score of six classes for dataset Potsdam.

*F*1	Imp_Surf	Building	Low_Veg	Tree	Car	Clutter
RGB-D	0.80	0.86	0.74	0.69	0.70	0.50
Fusion1	0.80	0.85	0.74	0.64	0.55	0.29
Fusion2	0.81	0.88	0.75	0.68	0.68	0.45
Fusion3	0.82	0.89	0.76	0.71	0.74	0.52
Fusion4	0.83	0.90	0.77	0.72	0.74	0.51
Fusion5	0.83	0.90	0.77	0.72	0.74	0.51
RGB	0.81	0.85	0.76	0.71	0.77	0.50
Elevation	0.56	0.80	0.16	0.47	0	0

5. Discussion

The main difference between our work and the state of the art is that we have explicitly investigated the impact of each layer of the FCNs quantitatively and visually. Recall and the descent rate of recall are used to quantitatively evaluate the layers' sensitivity concerning specific classes. Through visualization, the different contributions for specific classes are presented in an intuitive way. Additionally, we have proposed an effective layer fusion strategy for a combined exploitation of image and elevation information in a single FCN model which improves the quality of semantic labeling.

The experimental results concerning the layers' sensitivity to specific classes show that the contribution for specific classes depends on the features as shown in Figures 4 and 5. For the color, the shallower layers are more sensitive to small object like cars. Deeper layers, i.e., layers 4 and 5, are sensitive to objects with a more complex texture and occupy larger parts of the image. For the elevation, the shallower layers are more sensitive to impervious surfaces and low vegetation. This different contribution for specific classes explains why an early fusion does not lead to a satisfying result. If the multi-modal data are fused too early, the layers learn conflicting features, which leads to unexpected results. Based on this analysis, we have proposed an effective fusion strategy for the incorporation of multi-modal data for remote sensing semantic segmentation. On the Vaihingen and the Potsdam datasets, our fusion model outperforms multi-modal data which are concatenated into a 4-channel vector as the input of the model in terms of both average F-score and overall accuracy.

6. Conclusions

The main contribution of this paper is a detailed in-depth investigation of the sensitivities of individual layers of FCN to different object classes as well as multi-modal remote sensing data and their contributions to the semantic labeling as basis for an optimal fusion strategy. By exploring different layer fusion strategies, we explain why an early fusion in FCN cannot obtain satisfying results. In-depth understanding of the sensitivity and, thus, functionality of the layers allows an adaptive design of an FCN fusion model for multi-modal data, through which both classification results and efficiency are improved. The major conclusions are:

(i) For multi-modal data, the modes have different contributions to specific classes. The performance can be improved by avoiding conflicts between layers, i.e., different sensitivities to specific classes. By this means, we can, thus, design adaptive models suitable for semantic labeling tasks.

(ii) Summation fusion after the ReLU layer is applied in layer fusion. This preserves more information about features and strengthens the activation. The deeper layers are sensitive to objects which have a more complex texture and occupy a larger parts of the scene.

(iii) The skip architecture in the fusion allows to utilize the shallow as well as the deep features. Combining all these components, the multi-modal fusion model incorporates heterogeneous data precisely and effectively.

Concerning future work, we consider to extend the proposed approach to terrestrial data, i.e., terrestrial images with depth information. The latter can conventionally be obtained by laser scanning or depth estimation from stereo images. Nowadays the data can also be easily acquired by low-cost sensor combinations such as Microsoft Kinect (indoor scenes).

Acknowledgments: This work was supported partly by the National Natural Science Foundation of China under Grant 41501485 and the High Resolution Earth Observation Science Foundation of China under grant GFZX04060103. I am thankful to University of Chinese Academy Sciences for the scholarship.

Author Contributions: Wenkai Zhang, Hai Huang and Xian Sun conceived and designed the experiments; Hong qi Wang and Helmut Mayer supervised and made contribution to the article's organization; Wenkai Zhang, Hai Huang and Matthias Schmitz performed the experiments and analyzed the data; Wenkai Zhang and Hai Huang drafted the manuscript, which was revised by all authors. All authors have read and approved the final manuscript.

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

References

1. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
2. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
4. 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]

5. Längkvist, M.; Kiselev, A.; Alirezaie, M.; Loutfi, A. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sens.* **2016**, *8*, 329. [CrossRef]
6. Zhang, M.; Hu, X.; Zhao, L.; Lv, Y.; Luo, M.; Pang, S. Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images. *Remote Sens.* **2017**, *9*, 500. [CrossRef]
7. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H. Conditional random fields as recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1529–1537.
8. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv* **2016**.
9. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In Proceedings of the Asian Conference on Computer Vision ACCV, Taipei, Taiwan, 20–24 November 2016; Volume 2.
10. 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]
11. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 640–651. [CrossRef] [PubMed]
12. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**.
13. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [CrossRef]
14. Ning, F.; Delhomme, D.; LeCun, Y.; Piano, F.; Bottou, L.; Barbano, P.E. Toward automatic phenotyping of developing embryos from videos. *IEEE Trans. Image Process.* **2005**, *14*, 1360–1371. [CrossRef] [PubMed]
15. Pinheiro, P.H.O.; Collobert, R. Recurrent convolutional neural networks for scene parsing. *arXiv* **2013**.
16. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
17. Zintgraf, L.M.; Cohen, T.S.; Adel, T.; Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv* **2017**.
18. Zhang, W.; Huang, H.; Schmitz, M.; Sun, X.; Wang, H.; Mayer, H. A multi-resolution fusion model incorporating color and elevation for semantic segmentation. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-1/W1*, 513–517. [CrossRef]
19. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**. [CrossRef] [PubMed]
20. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv* **2015**.
21. Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
22. Audebert, N.; Saux, B.L.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *arXiv* **2016**.
23. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In *Computer Vision—ECCV 2014*; Springer: Cham, Switzerland, 2014; pp. 345–360.
24. Gerke, M.; Rottensteiner, F.; Wegner, J.D.; Sohn, G. ISPRS semantic labeling contest. In Proceedings of the Photogrammetric Computer Vision—PCV, Zurich, Switzerland, 5–7 September 2014.
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**.
26. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.