remote sensing

MDPI

Article

Water Body Extraction from Very High Spatial Resolution Remote Sensing Data Based on Fully Convolutional Networks

Liwei Li [1], Zhi Yan [1,2], Qian Shen [1], Gang Cheng [2], Lianru Gao [1] and Bing Zhang [1,3,*]

[1] The Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, No. 9 Deng Zhuang South Road, Beijing 100094, China; lilw@radi.ac.cn (L.L.); yanzhi@radi.ac.cn (Z.Y.); shenqian@radi.ac.cn (Q.S.); gaolr@radi.ac.cn (L.G.)

[2] School of Surveying and Land Information Engineering, Henan Polytechnic University, No. 2001 Shiji Road, Jiaozuo 454000, China; chenggang@hpu.edu.cn

[3] University of Chinese Academy of Sciences, No. 19 (A) Yuquan Road, Shijingshan District, Beijing 100049, China

* Correspondence: zb@radi.ac.cn

Received: 14 April 2019; Accepted: 13 May 2019; Published: 15 May 2019

Abstract: This paper studies the use of the Fully Convolutional Networks (FCN) model in the extraction of water bodies from Very High spatial Resolution (VHR) optical images in the case of limited training samples. Two different seasonal GaoFen-2 images with a spatial resolution of 0.8 m in the south of the Beijing metropolitan area were used to extensively validate the FCN model. Four key factors including input features, training data, transfer learning, and data augmentation related to the performance of the FCN model were empirically analyzed by using 36 combinations of various parameter settings. Our findings indicate that the FCN-based method can work as a robust and cost-effective tool in the extraction of water bodies from VHR images. The FCN-based method trained on a small amount of labeled L1A data can also significantly outperform the Normalized Difference Water Index (NDWI) based method, the Support Vector Machine (SVM) based method, and the Sparsity Model (SM) based method, even when radiometric normalization and spatial contexts are introduced to preprocess the input data for the latter three methods. The advantages of the FCN-based method are mainly due to its capability to exploit spatial contexts in the image, especially in urban areas with mixed water and shadows. Though the settings of four key factors significantly affect the performance of the FCN based method, choosing a qualified setting for the FCN model is not difficult. Our lessons learned from the successful use of the FCN model for the extraction of water from VHR images can be extended to extract other land covers.

Keywords: water body; extraction; very high spatial resolution; remote sensing; fully convolutional networks

1. Introduction

The rapid development of urbanization in the last decades has greatly changed the spatial distribution and quality of the surface water body in urban areas in China. One noticeable consequence is the deterioration of water quality due to frequent human activities. Thus, timely water body spatial distribution and quality information in urban areas is important in the management of public health and the living environment [1,2]. With the need of high frequency monitoring in large areas, traditional ground surveying cannot meet current demands due to its high labor cost and low efficiency. The development of remote sensing technology in recent years has largely improved the quality and availability of Very High spatial Resolution (VHR) optical remote sensing images (usually <1 m spatial

resolution), which have great potential for monitoring the fine scale surface water body in urban areas. However, developing a fast and effective way to extract surface water bodies from VHR images is still a significant problem.

Over the past decades, various water body mapping methods have been developed for remote sensing data [3–15]. A commonly used approach is based on water spectral indices such as the Normalized Difference Water Index (NDWI) [3,4]. The difficulty of using existing water indices is that the optimal thresholds vary a lot across regions. Although novel methods have been developed to alleviate this difficulty [5–11], most of them concentrate on low or middle spatial resolution optical images such as Sentinel-2 and Landsat. Low spatial resolution limits the capability of the data to detect small or narrow water bodies in urban areas [12]. Additionally, in the case of high spatial resolution optical images, the prevalence of shadows from tall buildings in urban areas makes the problem more complex, especially when the sun's zenith angle is low.

To overcome the limitation of these indices in high spatial resolution images, Huang et al. [13] made use of a group of information indexes and a machine learning technique to differentiate non-water responses such as shadows from water bodies. However, this strategy is sensitive to the quality of the threshold for the indexes and the training data for machine learning techniques. Yao et al. [14] selected the optimal threshold by a Support Vector Machine (SVM) based learning process to derive an initial result, and then refined it by a shadow removal process. However, the shadow removal method has a strong assumption of a rigid geometric relationship between a tall building and its shadow in the VHR image, which does not always hold true due to the existence of artificial targets with irregular shapes in urban areas. Recently Wu et al. [15] proposed a Two-Step Urban Water Index to extract water bodies from high spatial resolution images in a robust way. The method follows a strategy similar to [14] but it models the spectral features of water and shadows differently by combining an Urban Water Index (UWI) and an Urban Shadow Index (USI) both of which are established on the basis of spectral analysis and linear SVM. Although the effectiveness of the method has been extensively validated on high spatial resolution images from various sites, it heavily relies on atmospheric correction of the test image and is also not well validated on data from different seasons.

Different from using traditional expert designed features to capture spatial contexts in images, Convolutional Neural Networks (CNN) are shown to have great advantages in representing spatial contexts in images based on an end-to-end learning framework [16,17]. The deep learning model has gained increasing popularity in the remote sensing community nowadays [18,19]. Various novel deep learning models have been developed for scene classification and change detection on remote sensing data [20,21]. Among the developed deep learning models, the recently developed Fully Convolutional Networks (FCN) [22,23] can perform sematic segmentation and fits well to the pixel-wise classification of remote sensing data. Based on the FCN architecture, Isikdogan et al. [24] proposed a novel method to map water from Landsat data. The method can distinguish water from snow, ice, cloud, and terrain shadows, without requiring a locally varying threshold. However, the model was trained from scratch and was largely based on the freely available Landsat archives and the global inland water mask layer. The situation is quite different for VHR images in urban areas in several aspects such as the abundant availability of free training samples, the stability of spectral reflectance, the characteristics of spatial contexts, and the range of spectral coverage.

As far as we know, no previous work has seriously studied the extraction of water bodies from VHR images based on the FCN model. In this paper, we study the capability of the FCN model for the extraction of water bodies from VHR images, especially in the case of limited training samples. We extensively evaluate the performance of the FCN based method and compare it with the commonly used spectral index based method and supervised learning methods. We also empirically analyze several key factors in the FCN model implementation. Our lessons learned from this study can be extended to extract other land covers from VHR images.

2. Materials and Methods

2.1. Data

The study area selected was in the south of the Beijing metropolitan area. The high intensity of the human activity in this area has brought great attention to the surface water quality. Thus, fine scale and high frequency surface water body monitoring in this area has now become a routine task. Two different seasonal images from Chinese GaoFen-2, namely D302 and D609 acquired on March 2th 2017 and June 9th 2017, respectively, were used to extensively validate the FCN model by visual inspection and quantitative analysis. Each image has four bands (Blue (B), Green (G), Red (R), and Near Infrared (NIR)) and 0.8 m spatial resolution after pan sharpening. As most absolute methods for atmospheric correction of VHR images require the properties of atmosphere at the image acquisition time, which are usually difficult to obtain [25], all images used in the FCN-based method are L1A data products based on systematic radiometric correction without atmospheric correction for the purpose of generalization. For training data preparation, major water bodies in D302 were manually interpreted, and a typical region (namely W1) in D302 with abundant water and shadows was selected. Typical water and shadow areas in the W1 were also selected. For testing data preparation, 10 subsets (1500 × 1500 pixels for each) representing typical water and shadow areas, namely T1-T10, were selected based on visual inspection. Among them, T1-T6 were from D302 and T7-T10 were from D609. Meanwhile, the two full images were also used in the test to further validate the results by visual inspection. The study data and the related subsets are illustrated in Figure 1.

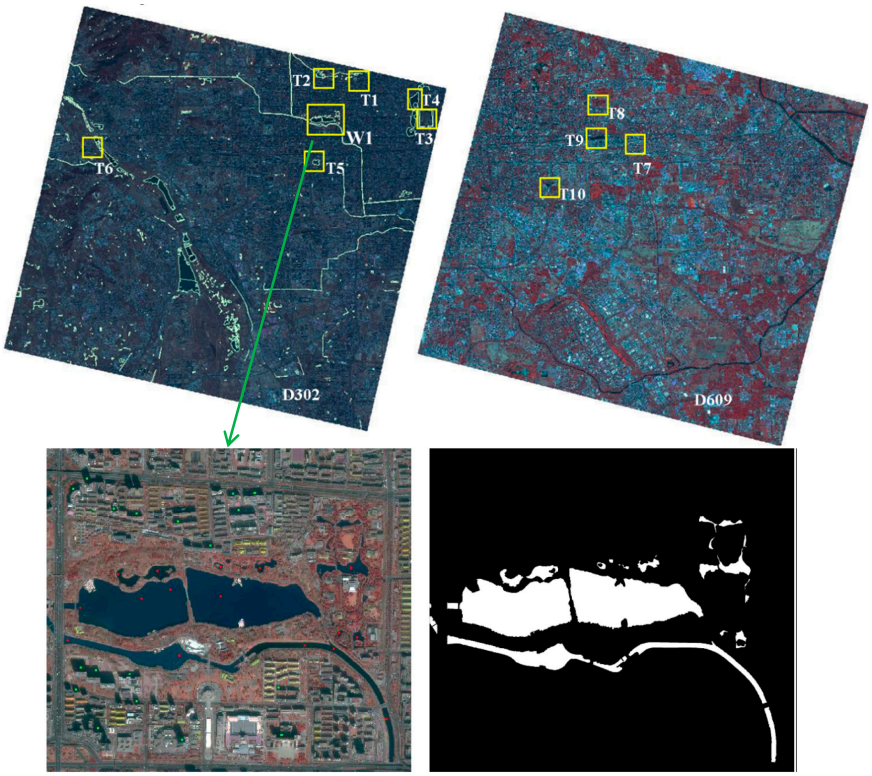

Figure 1. The two upper images are pan-sharpened GaoFen-2 images (namely D302 and D609) and the related training and test data in our experiment. W1 refers to the training subset. T1 to T10 refer to subsets for the test, and the size of each test data is 1500 × 1500 pixels. Water areas enclosed by light green lines in D302 indicate manually interpreted ground truths. The images are aligned based on their true geographical locations and are displayed with R (NIR)-G (R)-B (G) after being linearly stretched to [0,255]. The original spatial resolution of the image is 0.8 m. At the bottom, the W1 is specifically enlarged with its ground truth, and typical water (in red) and shadow samples (in green) selected from the W1 are displayed on the W1.

2.2. FCN Based Method

The flowchart of our FCN-based method is illustrated at the top of Figure 2. An input image is firstly clipped into patches (256×256 in our study). The patches are fed into a trained FCN model as shown at the bottom of Figure 2. The resulting patches of the FCN inference are spatially aligned accordingly. The output of the FCN based method is a map referring to water bodies in the input VHR image. To alleviate the possible boundary effect caused by spatial alignment in the final result, patches prepared for the inference are also clipped with spatial overlaps of 32 pixels, which are dropped in the following results alignment. The strategy is illustrated simply in Figure 3.

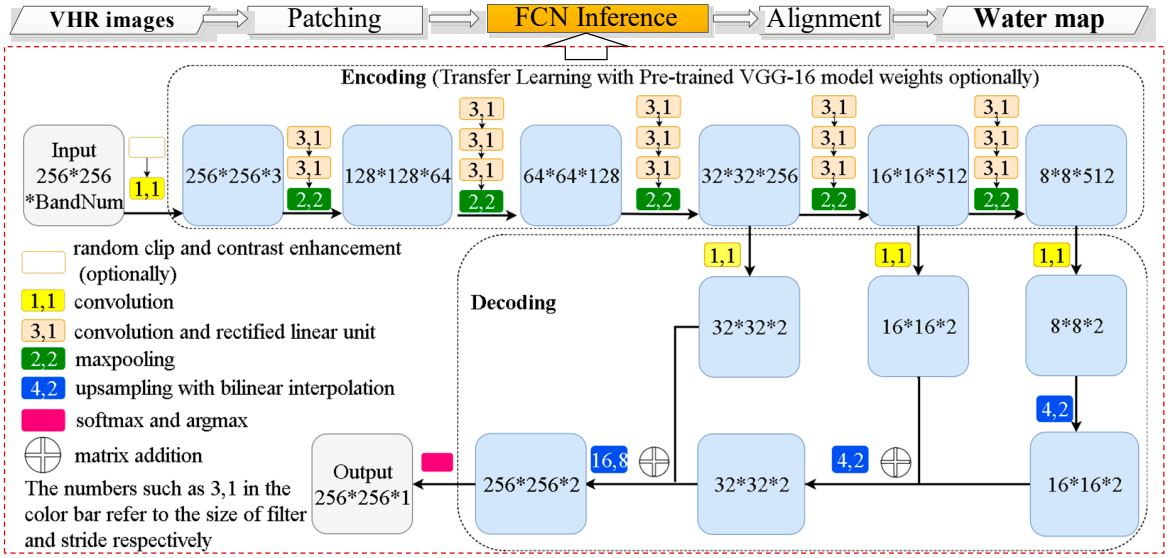

Figure 2. Illustration of the FCN-based method. The top refers to the main procedure; the bottom refers to the FCN model. In the model, the large blue boxes refer to intermediate features whose sizes are shown in each box. Each color bar above the main connector represents a specific module whose function is indicated by its color and is explained at the lower-left corner.

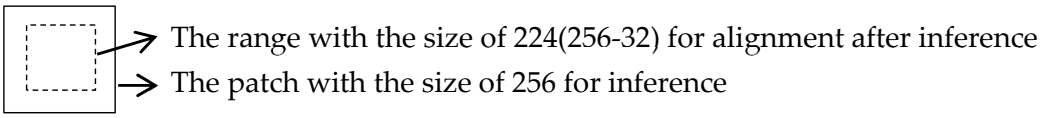

Figure 3. Illustration of the strategy to alleviate the possible boundary effect caused by spatial alignment in the final result.

The FCN model used here can be divided into four parts: a preprocessing layer, an encoder, a decoder, and an output layer. In the preprocessing layer, the input patch is optionally clipped and contrast enhanced in a random manner, and then transformed into a 3-channel patch with the same spatial size of the input patch by a one-to-one convolution. In the encoder, five layers in sequence similar to those in the VGG-16 model [26] are used to gradually encode each 3-channel patch into a group of small but informative features. The detail of each layer is illustrated by a sequence of color bars above each main connector. The two numbers in the color bar refer to the sizes of the filter and stride, respectively. The output of the encoder is fed into the decoder, which contains a sequence of upsampling layers to gradually recover spatial contexts of the label. The number of features in each decoded layer is equal to the number of classes (two in our case, water and non-water). The upsampling is realized by a transposed convolution with a fixed filter defined by the bilinear interpolation. Two skip layers are also introduced to enhance the spatial detail of label recovery. Two encoding features with finer details are convoluted into features with the dimension of the number of classes, and the convoluted features are added to corresponding decoding layers. In the output layer, a normalized

exponential function, namely softmax, transforms the decoded features into probabilities of water and non-water, and then argmax selects the label with the highest probability for each location and gets a pixel-wise map of the water body mask. The one-to-one convolution at the front makes the model flexible for reusing weights of the VGG-16 model trained on the ImageNet. This is useful in situations when the number of training samples is small [27].

2.3. Experimental Setup

Our experiments are designed to validate the capability of the FCN-based method for water body extraction from VHR images and also to analyze key factors affecting its performance. The number of factors affecting the FCN model's behavior is quite large. In this study, we focus on the analysis of four key factors including input features, training data, transfer learning, and input augmentation, while setting others to default values, as listed in the Table 1.

A total of 36 models with various parameters as listed in Table 2 were trained, and the results were analyzed in terms of overall performance and for the four selected factors. To be specific, for each selected factor, each choice of factor is fully studied under different combinations of other parameters. Additionally, to evaluate the stability of training of the FCN based method, we further selected 3 typical groups of parameters to train the FCN model independently 3 times.

To further validate the effectiveness of the FCN-based method, we compared it with three classic methods, including the NDWI based method, the SVM-based method, and the Sparsity Model (SM)-based method. All components of the methods are comparatively described in Table 3. In total, we used 32 combinations for the three methods compared, as listed in Table 4. Finally, we selected the combination with the highest average accuracy on the 10 test data for each method. It should be noted that because the three methods may be sensitive to radiometric changes between training and testing images, we employed the histogram matching based method [28] and the Iterative Reweighted Multivariate Alternate Detection (IR-MAD) based method [25] to normalize all images to the same radiometric level. The three methods also do not exploit spatial context information as the FCN-based method does in nature. For the NDWI and SVM based methods, we used Simple Linear Iterative Clustering (SLIC) [29] to segment the image into objects and then used the mean spectral feature of each object as the input for training and testing. For the SM-based method, we adapted the joint sparsity model by assuming that all pixels within a small neighborhood share the same group of words but have different coefficients [30].

Table 1. Key factors in the model training; we only evaluate factors with multiple choices.

Factors	Choices	Explanations
Input feature	if1	B-G-R-NIR
	if2	G-R-NIR
	if3	B-G-R
Transfer learning	tf1	None
	tf2	Reusing trained weights of VGG-16 on ImageNet as shown in Figure 2.
Training data	td1	Patches covering at least one pixel of water in the W1; 70 in total; all patches are extracted based on a moving step that is the same as the patch size. The following is the same
	td2	All patches in the W1; 340 in total
	td3	Patches covering at least one pixel of water in the image of D302; 1666 in total
Data augmentation	da1	None
	da2	The sequential combination of random clip and contrast enhancement as shown in Figure 2.
Initializer	*Xavier*	*Initialize the weight of the network before training* [31]
Batch size	*1*	*The number of patch used in each round of training*
Patch size	*256*	*The size of input patch*
Training step	*24000*	*We output a trained model at each 3000 steps and select the one with the best performance on the training data*
Loss function	*Cross-entropy*	*Measurement of loss in the optimization*
Optimizer	*Adam*	*Algorithm for updating the weight* [32]
Learning rate	*0.00001*	*Key parameter in the Adam*

Table 2. List of experimental setups for analyzing overall performance of the FCN-based method and four key factors of the model, respectively.

Purposes	Experiment Setup (the Order of Factor Is Irrelevance)
Analysis of overall performance of the FCN-based method with all combinations of selected parameters	if1-tf1-td1-da1, if1-tf2-td1-da1, if1-tf1-td1-da2, if1-tf2-td1-da2, if1-tf1-td2-da1, if1-tf2-td2-da1, if1-tf1-td2-da2, if1-tf2-td2-da2, if1-tf1-td3-da1, if1-tf2-td3-da1, if1-tf1-td3-da2, if1-tf2-td3-da2, if2-tf1-td1-da1, if2-tf2-td1-da1, if2-tf1-td1-da2, if2-tf2-td1-da2, if2-tf1-td2-da1, if2-tf2-td2-da1, if2-tf1-td2-da2, if2-tf2-td2-da2, if2-tf1-td3-da1, if2-tf2-td3-da1, if2-tf1-td3-da2, if2-tf2-td3-da2, if3-tf1-td1-da1, if3-tf2-td1-da1, if3-tf1-td1-da2, if3-tf2-td1-da2, if3-tf1-td2-da1, if3-tf2-td2-da1, if3-tf1-td2-da2, if3-tf2-td2-da2, if3-tf1-td3-da1, if3-tf2-td3-da1, if3-tf1-td3-da2, if3-tf2-td3-da2
Analysis of which type of input feature is more effective	if1-tf1/tf2-td1/td2/td3-da1/da2 if2-tf1/tf2-td1/td2/td3-da1/da2 If3-tf1/tf2-td1/td2/td3-da1/da2
Analysis of whether transfer learning is useful	if1/if2/if3-tf1-td1/td2/td3-da1/da2 if1/if2/if3-tf2-td1/td2/td3-da1/da2
Analysis of which group of training data is more effective	if1/if2/if3-tf1/tf2-td1-da1/da2 if1/if2/if3-tf1/tf2-td2-da1/da2 if1/if2/if3-tf1/tf2-td3-da1/da2
Analysis of whether data augmentation is useful	if1/if2/if3-tf1/tf2-td1/td2/td3-da1 if1/if2/if3-tf1/tf2-td1/td2/td3-da2
Analysis of the stability of the training process	if1-tf1-td1-da1-01, if1-tf1-td1-da1-02, if1-tf1-td1-da1-03 if3-tf2-td3-da2-01, if3-tf2-td3-da2-02, if3-tf2-td3-da2-03 if2-tf2-td2-da2-01, if2-tf2-td2-da2-02, if2-tf2-td2-da2-03 -01 indicates round 1 training with if3-tf2-td3-da2-01

Table 3. List of components of the methods in the comparison.

Component	Description
water	We randomly select 10000 samples from the W1 and then divide them into water and others based on their labels.
water-shadow	We manually collect typical water and shadows samples in the W1 as illustrated in the lower-left of the Figure 1. These samples lie near the boundary of decision function and is useful for discriminative methods such as the NDWI and SVM based methods [13].
norm	norm refers to the IR-MAD based radiometric normalization method. Here we select the atmospherically corrected Sentinel-2 image spatially coded as T50TMK and acquired on 9 March 2017 as reference. All VHR images are normalized to the reference.
hm	hm refers to histogram matching based radiometric normalization. D302 works as reference and D609 is transformed in our experiment.
grid-svm	grid-svm refers to the linear SVM model with an optimized penalty C. the linear SVM model is employed here for its efficiency and effectiveness compared with the RBF based SVM model after rigorous comparisons on training samples. The C is set by a grid search.
ndwi	ndwi refers to the NDWI based method which uses the mean index value of the selected water and other samples in the training data as the threshold. The index is calculated according to Equation (1).
slic	slic refers to the SLIC method and is used to segment an input image into small objects based on which spatial contexts can be exploited in the water extraction. SLIC on a large image is computational expensive. Here we cut a large input image into patches with a size of 500*500 pixels, and then segment each patch into approximately 10000 regions, finally all segmented patched are combined into a single image. To mitigate the boundary effect, we overlap 200 pixels vertically and horizontally in the patch cutting.
best-sm	best-sm refers to the joint sparsity model. Since it does not scale well with the size of dictionary, we randomly select 250 samples for water and others, respectively from the training data. To keep the uncertainty brought by training samples to the minimum level, we run the SM based method for 10 times and select the one with the best performance.

Table 4. List of experimental setups for analyzing overall performance of the three methods in comparison. The combination in the table is composed by components listed in Table 3. For example, norm-hm-water-shadow-best-sm indicates a customized SM-based method. The method is trained on water-shadow samples from the normalized D302 image and is tested on the normalized D302 image and the preprocessed D609 image. In the preprocessing, the D609 image is firstly normalized to the Sentinel-2 image, and then has its histogram matched to the normalized D302 image.

The SM Based Method	The SVM Based Method	The NDWI Based Method
	water-grid-svm	water-ndwi
	water-shadow-grid-svm	water-shadow-ndwi
water-best-sm	norm-water-grid-svm	norm-water-ndwi
water-shadow-best-sm	norm-water-shadow-grid-svm	norm-water-shadow-ndwi
norm-water-best-sm	norm-hm-water-grid-svm	norm-hm-water-ndwi
norm-water-shadow-best-sm	norm-hm-water-shadow-grid-svm	norm-hm-water-shadow-ndwi
norm-hm-water-best-sm	hm-water-grid-svm	hm-water-ndwi
norm-hm-water-shadow-best-sm	hm-water-shadow-grid-svm	hm-water-shadow-ndwi
hm-water-best-sm	water-slic-grid-svm	water-slic-ndwi
hm-water-shadow-best-sm	water-shadow-slic-grid-svm	water-shadow-slic-ndwi
	norm-water-slic-grid-svm	norm-water-slic-ndwi
	norm-water-shadow-slic-grid-svm	norm-water-shadow-slic-ndwi

$$\text{NDWI} = (\text{G} - \text{NIR})/(\text{G} + \text{NIR}) \tag{1}$$

In all experiments, we used the F1 score to assess the performance of each method. The F1 score is the harmonic average of the precision and recall, as indicated in Equation (2) where an F1 score reaches its best value at 1 and worst at 0. It is more objective than overall accuracy in our binary classification case because a water body mostly covers a small portion of the image under evaluation.

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

where precision is the number of correct positive pixels divided by the number of all positive pixels returned by the method, and recall is the number of correct positive pixels divided by the number of all relevant pixels.

3. Results and Analysis

3.1. Analysis of the Performance of the FCN Based Method

3.1.1. Overall Performance Analysis of the Trained FCN Models

Figure 4 illustrates the mean and variance of F1 scores on all 10 test data based on the 36 FCN models with different parameter settings, as listed in Table 2. The mean F1 score has a large dynamic range from 0 to ~0.92. There are 15 trained models with a mean F1 score above 0.8, and there are 7 trained models with a mean F1 score below 0.5. Meanwhile, the variance of the F1 score varies from 0 to ~0.2 with the models. It can be inferred from Figure 4 that models with large mean values have low variances. Figure 5 illustrates the mean and variance of F1 scores on the 36 FCN models for each of the 10 test data with different parameter settings as indicated in Table 2. The mean F1 score fluctuates from ~0.5 to ~0.8, and the variance remains relatively stable around 0.1. T5 and T7 are the most and the least accurately processed test data, respectively. In general, results of the test data (T1-T6) from D302 where training data were collected are better than those from D609 (T7-T10).

The results imply that the performance of the FCN based method is significantly affected by the parameter setup, though the choice of a qualified parameter setting for the FCN model is not hard to find. The bad average performance on T7 may be due to the ratio of the area of water bodies to shadows in this test data being small compared to that of others. Thus few omissions or false alarms may lead to a large change of the F1 score. The situations for T1, T4, and T5 are largely the opposite. To further validate the effectiveness of the FCN-based method, we applied a few well-trained FCN models to the two full images. Consistent results were obtained based on visual inspection of the resultant water maps.

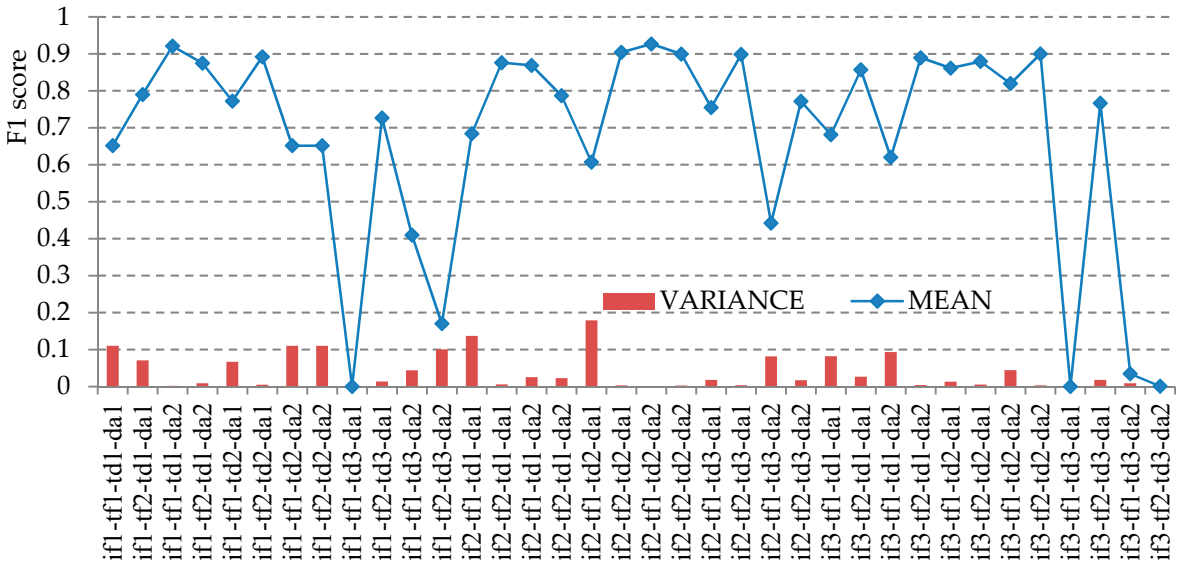

Figure 4. Illustration of the mean and variance of F1 scores on all 10 test data based on the 36 FCN models with different parameter settings, as indicated in Table 2.

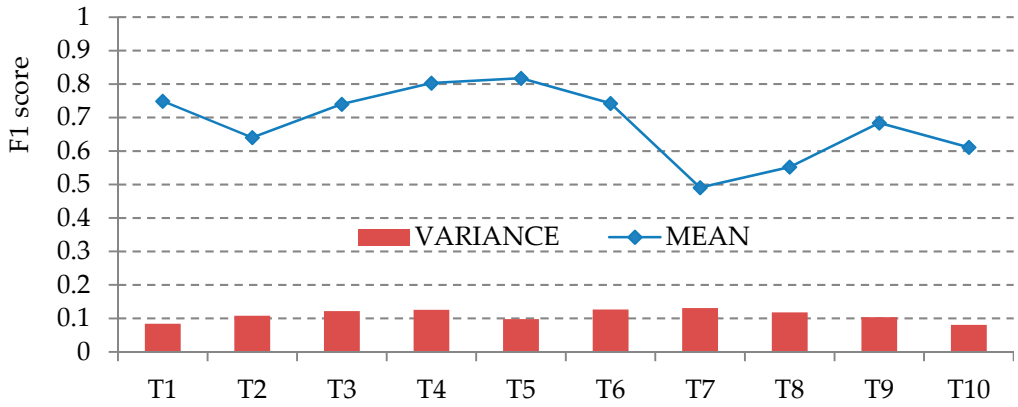

Figure 5. Illustration of the mean and variance of F1 scores on the 36 FCN models for each of the 10 test data with different parameter settings, as indicated in Table 2.

3.1.2. Comparing the FCN-Based Method with Classic Methods

Figure 6 shows the mean accuracies on all 10 test data and on test data T7-T10 of all combinations as listed in Table 4. Generally, the classic methods performed better on test data T1-T6 than on test data T7-T10. The boundary training samples, namely water-shadow, did not necessarily improve the accuracy of the methods, nor did the SLIC and radiometric normalization. Nevertheless, the best combinations for each of the three methods are norm-water-slic-ndwi, hm-water-grid-svm, and water-best-smc, respectively, and the mean F1 scores on all 10 test data of the three best combinations are about 0.85, 0.88 and 0.89. This implies that the strategies employed in our experiments such as radiometric normalization and segmentation are helpful for improving the three methods on the extraction of water from VHR. However, we did not find a single best strategy for all three methods. Here, we focus on the best combination from each of the three methods. Deeper analysis of the classic methods is beyond the scope of this paper.

Figure 7 shows the comparison of the best FCN-based method with the best SM-based method (best-SM), the best SVM based method (best-SVM), and the best NDWI (best-NDWI) based method in terms of per test data F1 score. Figure 8 illustrates the extracted water body maps from 10 test data by all 4 methods. From Figure 7, best-FCN outperforms the other three methods, especially on the test data T2, T7, and T8. Best-SM achieves slightly better results than best-SVM, while best-NDWI is comparatively the worst in terms of overall accuracy. According to Figure 8, there are plenty of small false alarms in the results except that of best-FCN. However, the distribution of false alarms for each of the three methods differs a lot on specific test data. Overall best-FCN is obviously more effective than the other three methods, as a result of a prominent advantage for dealing with areas mixed with water and shadow. This advantage of best-FCN comes from its ability to exploit the spatial context in the image so as to discriminate the shadows of high-rising buildings from water bodies and becoming largely invariant to the radiometric change. The latter merit is quite distinct as the input to the FCN-based method is in L1A level.

Figure 9 illustrates the mean sample spectrum of water in W1 and four typical test data from D609. The radiometric differences between W1 and other test data are quite clear in L1A data. hm and norm as described in Table 3 do reduce radiometric changes between images acquired in different conditions, especially for hm. Their effects will be more distinct if the whole images are taken into consideration. However, due to hm being a nonlinear point operator concentrating on matching the global histogram between different images, it may not be helpful to discriminate between water and shadow, which are located at the low end of the radiometric range and share similar spectral features that vary a lot from space and time. The situation is the same for norm, which also works on the whole image. The effect of slic on radiometric normalization is not as distinct as that of the other two strategies, but it may help in reducing spatial noises in the image.

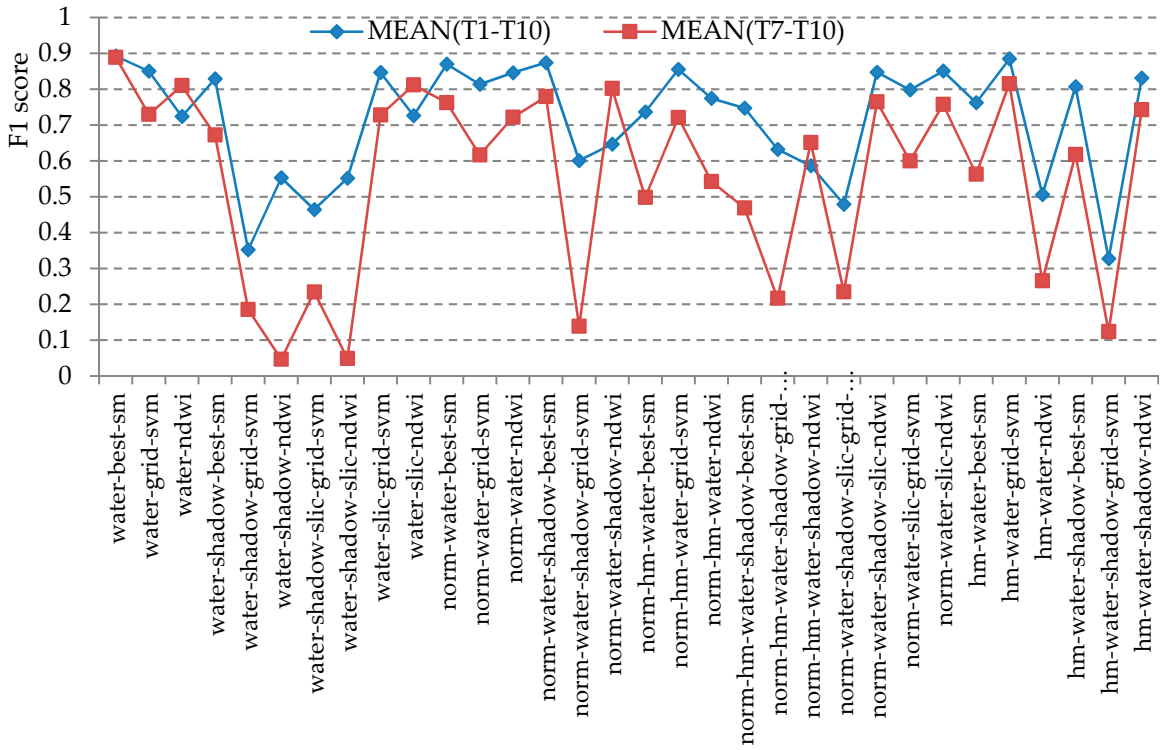

Figure 6. Illustration of the mean and variance of the F1 scores on all 10 test data based on the 32 combinations from the three classic methods, as indicated in Table 4.

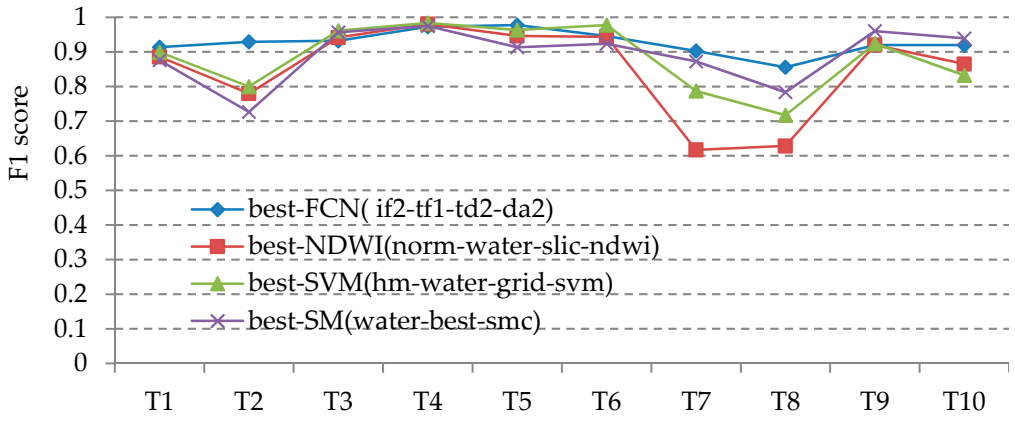

Figure 7. Comparison of best-FCN with best-SM, best-SVM, and best-NDWI in terms of per test data F1 score.

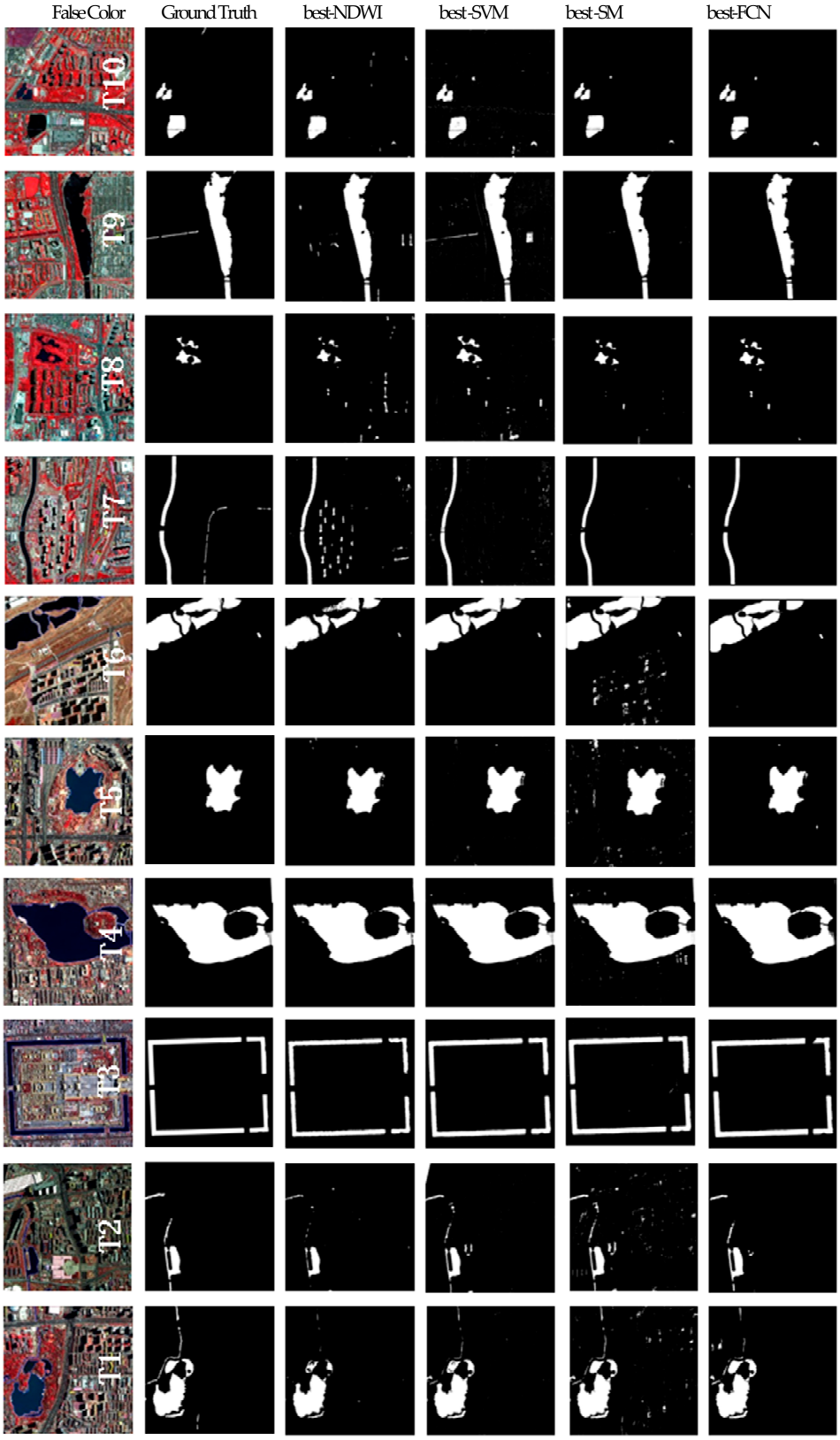

Figure 8. Illustration of false color images (column 1), ground truth (column 2) and results of T1-T10 (rows 1-10) based on best-NDWI, best-SVM, best-SM and best-FCN (columns 3, 4, 5, 6).

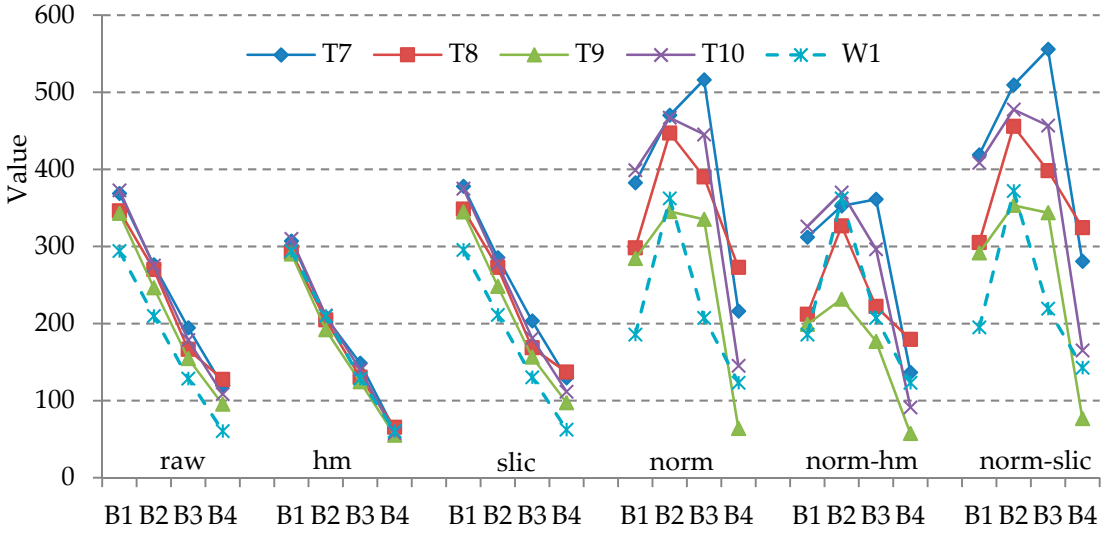

Figure 9. Each group of spectra refers to the mean spectrum of water samples from W1 and T7-T10 (D609). From left to right, they are from L1A data, hm enhanced image, slic enhanced image, norm enhanced image, norm-hm enhanced image, and norm-slic enhanced image. Refer to Table 3 for a detailed explanation of those terms. Values on the y-axis refer to the digital number for the first three groups and reflectance with a scale factor of 10000 for the last three groups.

3.2. Analysis of Key Factors of the FCN-Based Method

3.2.1. Analysis of the Input Feature

Figure 10 shows the average F1 scores of 10 test areas for 12 groups of models with different input features, as explained in Table 2. Figure 11 shows the average F1 scores of 12 models for 10 test data with different input features. For each group of models, the model parameters used in the training are the same except for the input feature, as indicated in Figure 10. Overall, input feature 2 clearly gives better performance than that of input future 1 and input feature 3. Input feature 1 is also slightly better than input feature 3. The results indicate that the NIR band may contribute most to the water body extraction, while the blue band may be least helpful or even harmful to the water body extraction. This phenomenon can be explained based on the fact that the blue band is more sensitive to atmospheric scattering as well as suspended matters on the surface of the water when compared with the NIR band.

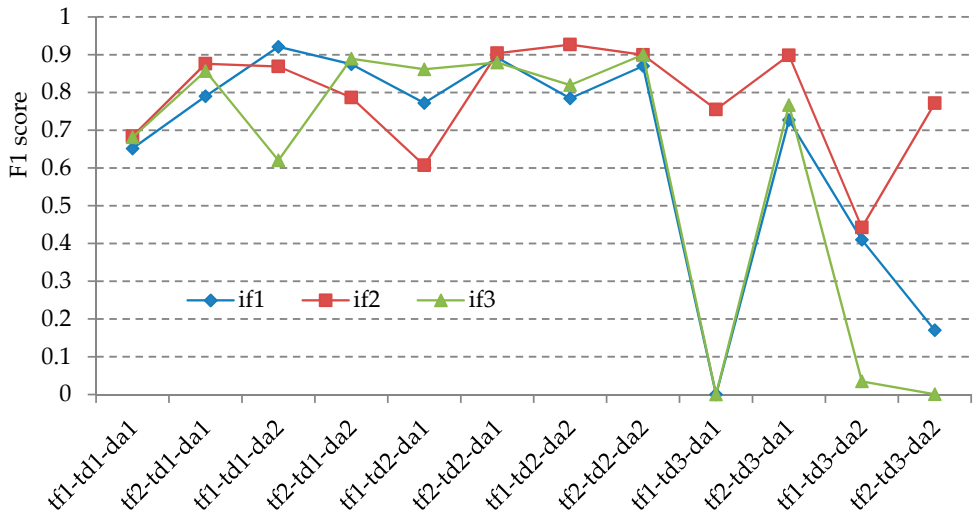

Figure 10. Average F1 scores of 10 test areas for 12 groups of models with different input features, as explained in Table 2.

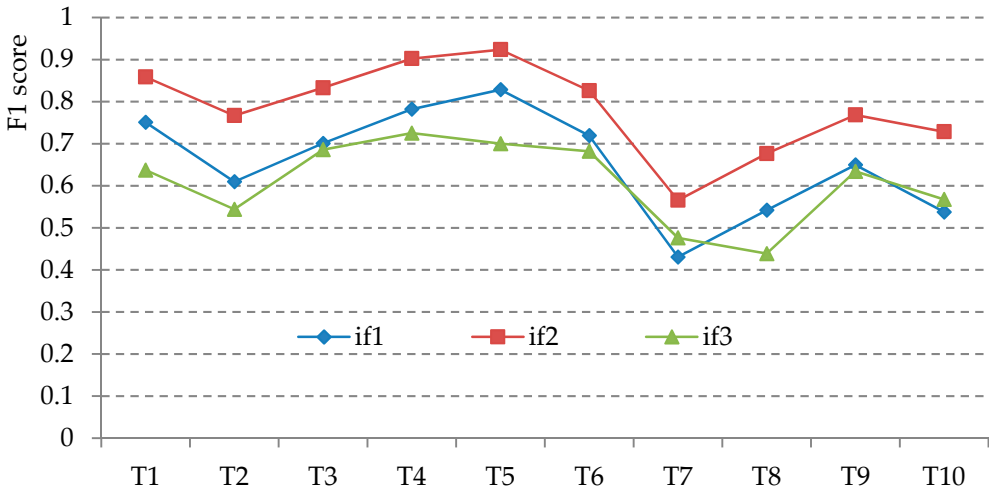

Figure 11. Average F1 scores of 12 models for 10 test data with different input features.

3.2.2. Analysis of the Training Data

Figure 12 shows the average F1 scores of 10 test areas for 12 groups of models with different training data, as explained in Table 2. Figure 13 shows the average F1 scores of 12 models for 10 test data with different training data. For each group of models, the model parameters used in the training are the same, except for training data, as indicated in Figure 12. Overall, training data 2 and training data 1 clearly show better performance than that of training data 3. Training data 2 is also slightly better than training data 1. The results largely indicate that representativeness is more important than the amount of size in terms of the quality of training data in our experiment. The high quality of training data 2 may be due to it containing samples covering water body areas as well as typical high-rising building areas with plenty of shadows in a balanced manner. The shadow is quite spectrally similar with the water body but has a different spatial context. Although the size of the samples in the training data 3 is larger than that of the samples in training data 2, the bad performance of training data 3 may be due to its lack of balance in the amount of samples covering water body areas and other low reflectance targets such as shadows. This should be taken into consideration in the application of deep learning models in remote sensing data when the total amount of training samples is relatively small.

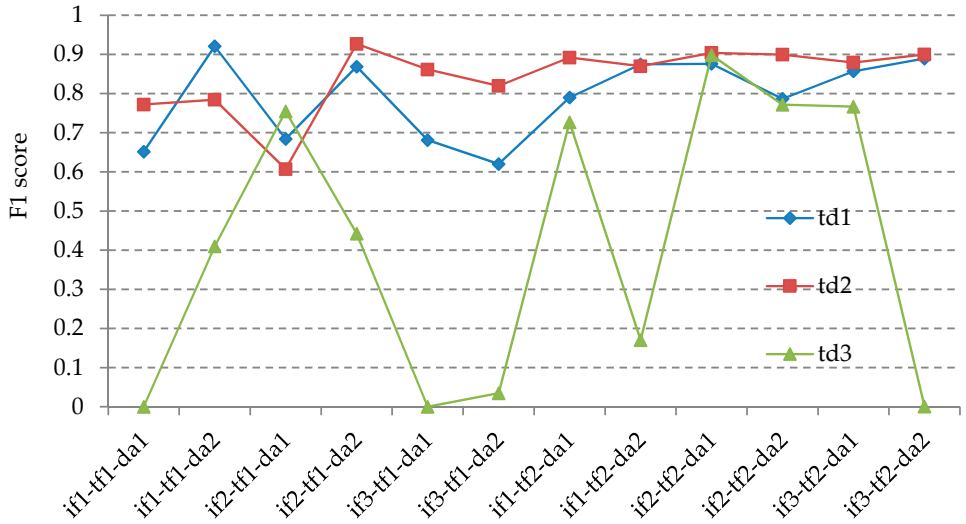

Figure 12. Average F1 scores of 10 test areas for 12 groups of models with different training data as explained in Table 2.

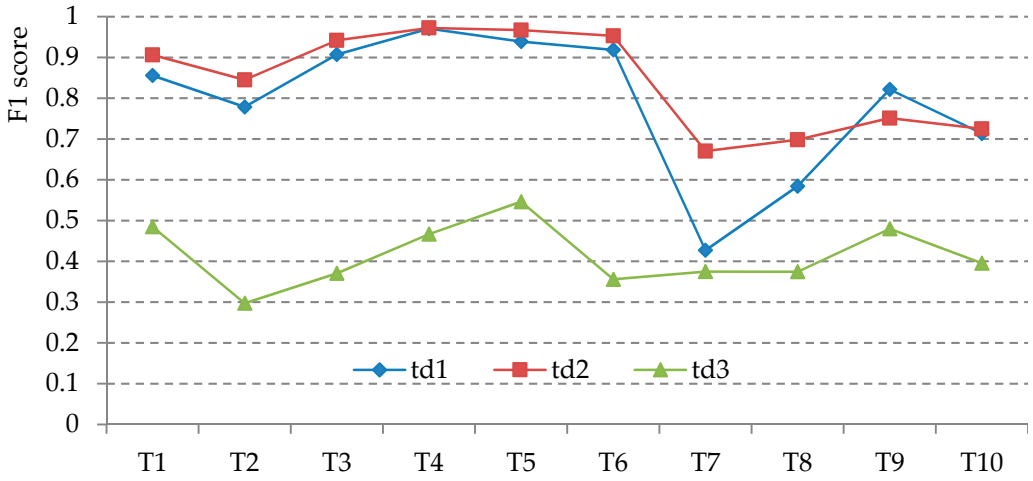

Figure 13. Average F1 scores of 12 models for 10 test areas with different training data.

3.2.3. Analysis of the Transfer Learning

Figure 14 shows the average F1 scores of 10 test data for 18 groups of models with and without transfer learning. Figure 15 shows the average F1 scores of 18 models for 10 test data with and without transfer learning. For each group of models, the model parameter used in the training is the same, except for transfer learning, as indicated in Figure 14. In nearly all cases, a model with transfer learning is better than the one without it in terms of the performance measured by the F1 score. Introduction of the transfer learning in the training process increases the F1 score by about 10% for the FCN-based method, as indicated in Figure 15. We consider that the merit of the transfer learning is based on the similarity of the spatial context between VHR images and natural images. For model training without transfer learning, the model initialization plays an important role. Xavier [24], used in our experiment, initializes weights with a properly scaled uniform distribution and has been proven to be substantially helpful for increasing the training convergence. That may be the reason that, in a few cases, trained models without transfer learning can still achieve very good results. As shown in Figure 14, the trained model with the best mean F1 score on all test data was actually achieved without transfer learning. Overall, taking the stability of the training into consideration, the inclusion of transfer learning in the FCN model training for processing VHR images is strongly suggested in practical use.

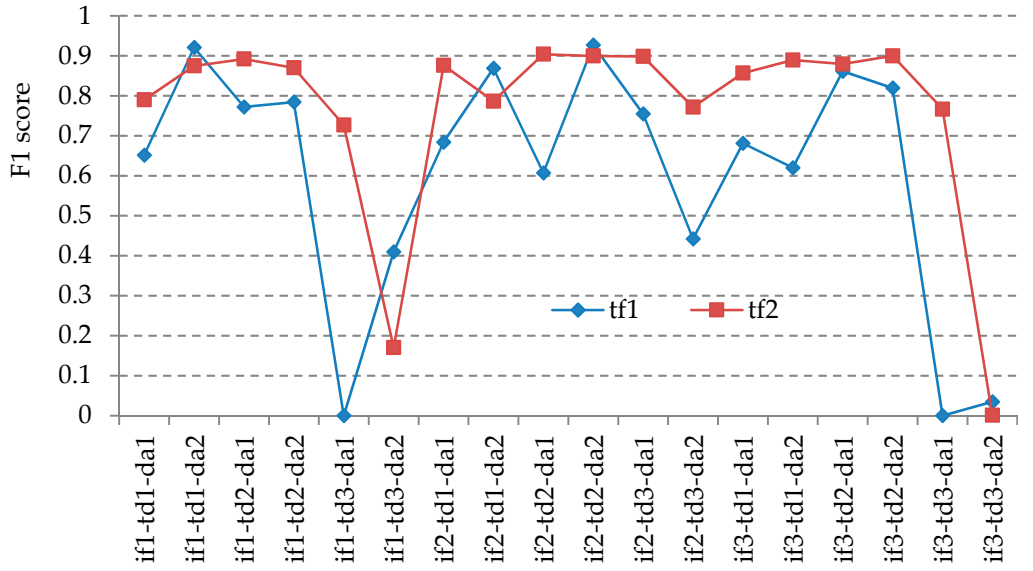

Figure 14. Average F1 scores of 10 test data for 18 groups of models with and without transfer learning.

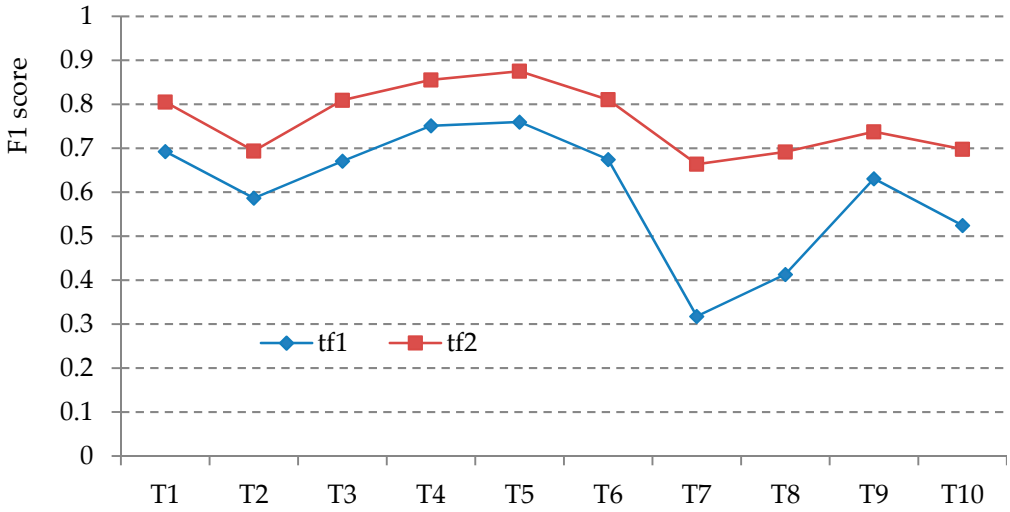

Figure 15. Average F1 scores of 18 models for 10 test data with and without transfer learning.

3.2.4. Analysis of the Data Augmentation

Figure 16 shows the average F1 scores on 10 test data for 18 groups of models with and without data augmentation. Figure 17 shows the average F1 scores on 18 models for 10 test data with and without data augmentation. For each group of models, the model parameter used in the training is the same, except data augmentation, as indicated in Figure 16. Overall, data augmentation does not show clear advantages over the original input data in our experiment. However, trained models without data augmentation perform slightly better for test data for images where training data is collected, but the situation is vice versa for test data in another image. The result shows weak support for the effectiveness of data augmentation in the training. However, the typical contrast enhancement does not work in a distinct way as it works for natural images. This may due to the remote sensing image being relatively physical calibrated in its quantity compared with natural images and also the self-similarity of geometric structures being represented in the remote sensing image prevailing on the earth surface. The latter would confuse the trained model by targets that show similar textures of the water body but have different spectral properties from the water body such as soil and large roofs.

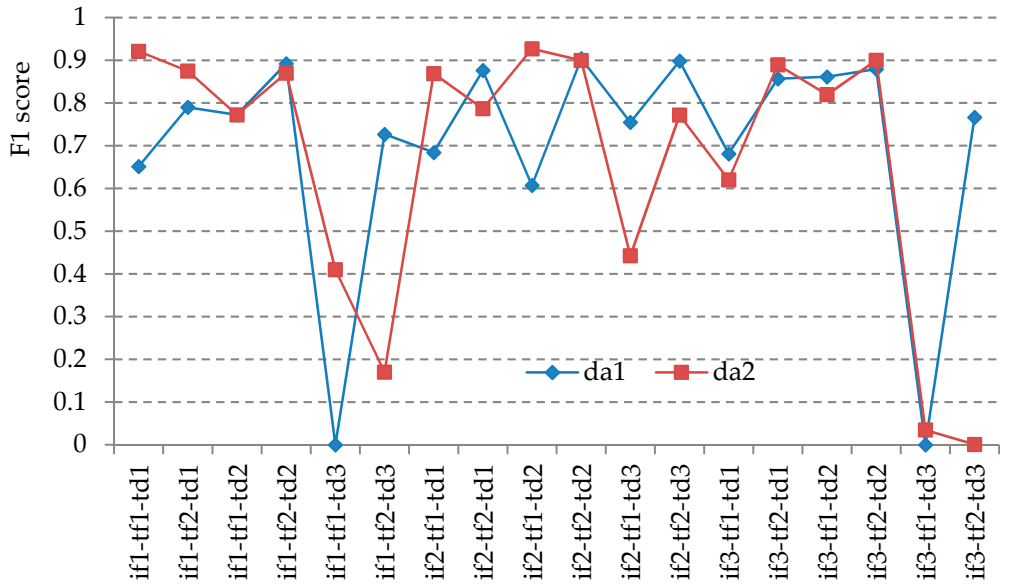

Figure 16. Average F1 scores of 10 test areas for 18 groups of models with and without data augmentation as explained in Table 2.

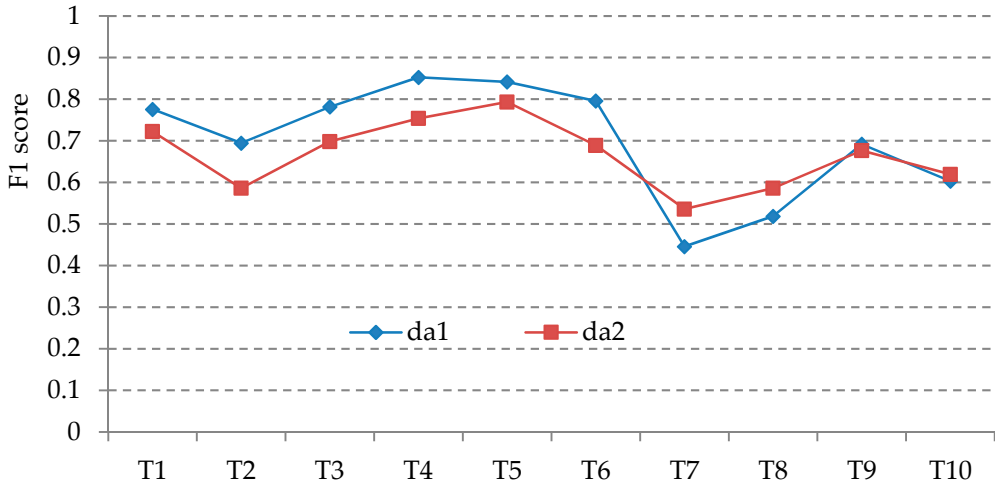

Figure 17. Average F1 scores of 18 models for 10 test areas with and without data augmentation.

3.2.5. Analysis of the Stability in the Training of FCN-Based Method

Figure 18 shows the average F1 scores on 10 test data for 9 trained models as indicated in the last row of Table 2. It can be seen that the stability of the model training is relatively weak. This is especially true for models without transfer learning. With a better input feature such as input feature 2, the stability of the training can also be largely enhanced. The randomness of the results is mainly brought by the stochastic method exploited in the model training. The results encourage the inclusion of transfer learning and good input features to weaken the side-effect of randomness in the FCN-based method, meanwhile, any parameter setting should be tried at least twice before rejection.

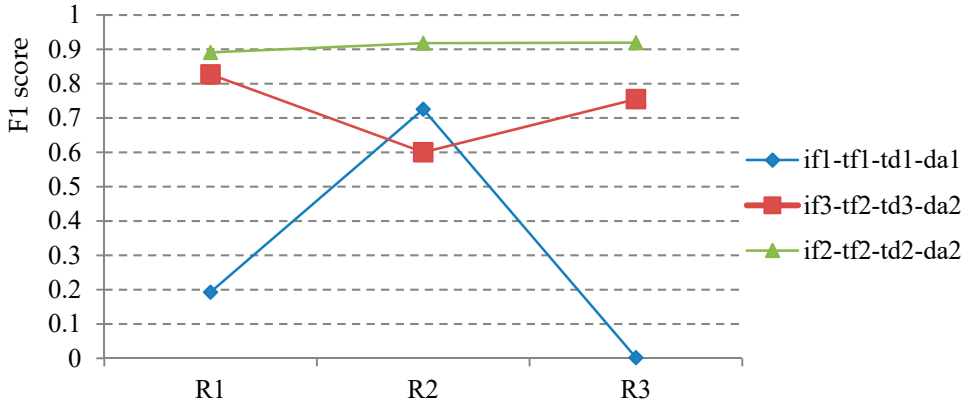

Figure 18. Average F1 scores on 10 test data for 9 trained models in the last row of Table 2.

4. Discussion and Conclusions

In this paper, we study the use of the FCN model to extract water bodies from VHR images. To better adapt to the property of remote sensing data, we introduce a flexible one-to-one convolution layer in the typical FCN model. Two seasonal representative Chinese GaoFen-2 images with a spatial resolution of 0.8 m experimentally validated the FCN based method in the extraction of water bodies from VHR images. Meanwhile, we selected and analyzed four key factors using 36 FCN models with different parameter settings with the purpose of understanding the contribution of each key factor and how to make good choices of them in practical use.

The FCN-based method can work as a robust and effective tool in the extraction of water bodies from VHR images. If properly trained with a small number of labeled samples, the FCN based method can also significantly outperform the SM-based method, the SVM-based method, and the NDWI-based method on either capability or transferability, especially for urban areas with mixed water and shadows.

The advantage of the FCN based method remains even when radiometric normalization and spatial context information are introduced to preprocess the input data for the other three methods. This is largely due to the capability of the FCN model to exploit the spatial context in VHR images well. This capability also makes the FCN-based method withstand radiometric changes, which is a quite challenging task for typical feature extraction methods [13–15]. It should be noted that the high accuracy of the FCN-based method in our experiments is at the expense of efficiency compared with the NDWI based method. We still consider that the NDWI is the best choice in appropriate sceneries when efficiency is more important than precision.

Our study supports the prevalent existence of the qualified FCN models with various parameter settings. Although the randomness in the training is unavoidable due to the nature of the optimization method for the FCN, and the performance of the FCN-based method is easily affected by model parameter settings, a well-trained FCN model for water body extraction from VHR images is not hard to find and the choice of a qualified parameter setting for the FCN model varies. This eases the difficulty of the selection of the optimal parameter setting in the use of the FCN model. Empirical results encourage the selection of the input feature, the inclusion of transfer learning, and the use of training data with balanced positive and negative samples to improve the stability and accuracy of the FCN based method. These findings help to build a stable and accurate FCN-based method in real applications. Finally, due to the holistic nature of remote sensing data, the extraction of a specific type of target from VHR images covering a large area is commonly seen in real remote sensing applications. Though the appearance of the type of target in the image may diverge a lot depending on applications, our lessons learned from the successful use of the FCN model in the water body extraction from VHR images can be extended to extract other land covers especially in cases with limited training samples.

We obtained FCN models with satisfactory results based on relatively small training data. However, our experimental images were acquired with clear sky and flat terrain. These criteria may not always be fulfilled in practical use. The remote sensing image may be contaminated by vapor and haze. Shadows of mountains may be different from that of buildings. One research direction is to enrich the training data by accounting for more situations. As the manual interpretation of water body from VHR images from scratch is quite tedious work, we may collect the training data in an iterative way by using extracted results. However, the representativeness of the samples should be kept in mind in the preparation of the training data.

Our experiments were carried out on 0.8 m GaoFen-2 images. The feasibility of the proposed FCN-based method on more images from other VHR sensors with different specifications should be validated further because short-term temporal monitoring of water bodies needs collaborative data from multiple sensors in practice, but the accuracy of water body extraction is sensitive to minor spatial resolution changes [33] of the data. Very narrow rivers also are not extracted well by the FCN-based method in our experiments. We are considering improving this by introducing an advanced FCN model [34]. This will be studied in the future. Last but not least, treating water bodies as a single class like in our experiments is not enough in practical use. Our next step is to work on extracting more subtle types of water bodies with different depths and turbidity.

Author Contributions: B.Z. and L.L. came up with the original idea for the study, and L.L. and Q.S. carried out the design. L.L. was responsible for recruitment and follow-up of study participants. Z.Y. and G.C. were responsible for programming and data processing. L.L. and L.G. conceived the experiments and carried out the analysis with assistance from Z.Y., L.L. structured and drafted the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China from MOST (Grant Number 2016YFB0501501), National Natural Science Foundation of China (Grant Number 91638201) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Number XDA19080304).

Acknowledgments: The authors appreciate the anonymous referees for their valuable suggestions and questions. The authors thank Mengyun Ren and Qinglan Zhang for their hard work on manual interpretation of the water body from the D302 image and Jinming Zhu for arrangement of F1 scores data. All of them are graduate students from the School of Surveying and Land Information Engineering, Henan Polytechnic University.

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

References

1. Niemczynowicz, J. Urban hydrology and water management-present and future challenges. *Urban Water* **1999**, *1*, 1–14. [CrossRef]
2. Varis, O.; Vakkilainen, P. China's 8 challenges to water resources management in the first quarter of the 21st century. *Geomorphology* **2001**, *41*, 93–104. [CrossRef]
3. Gao, B.C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
4. McFeeters, S.K. The use of the normalized difference water index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
5. Feyisa, G.L.; Meilby, H.; Fensholt, R.; Proud, S.R. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sens. Environ.* **2014**, *140*, 23–35. [CrossRef]
6. Xie, H.; Luo, X.; Xu, X.; Pan, H.Y.; Tong, X.H. Automated Subpixel Surface Water Mapping from Heterogeneous Urban Environments Using Landsat 8 OLI Imagery. *Remote Sens.* **2016**, *8*, 584. [CrossRef]
7. Hollstein, A.; Segl, K.; Guanter, L.; Brell, M.; Enesco, M. Ready-to-Use Methods for the Detection of Clouds, Cirrus, Snow, Shadow, Water and Clear Sky Pixels in Sentinel-2 MSI Images. *Remote Sens.* **2016**, *8*, 666. [CrossRef]
8. Fisher, A.; Danaher, T. A Water Index for SPOT5 HRG Satellite Imagery, New South Wales, Australia, Determined by Linear Discriminant Analysis. *Remote Sens.* **2013**, *5*, 5907–5925. [CrossRef]
9. Yang, X.C.; Zhao, S.S.; Qin, X.B.; Zhao, N.; Liang, L.G. Mapping of Urban Surface Water Bodies from Sentinel-2 MSI Imagery at 10 m Resolution via NDWI-Based Image Sharpening. *Remote Sens.* **2017**, *9*, 596. [CrossRef]
10. Yang, Y.H.; Liu, Y.X.; Zhou, M.X.; Zhang, S.Y.; Zhan, W.F.; Sun, C.; Duan, Y.W. Landsat 8 OLI image based terrestrial water extraction from heterogeneous backgrounds using a reflectance homogenization approach. *Remote Sens. Environ.* **2015**, *171*, 14–32. [CrossRef]
11. Jia, K.; Jiang, W.G.; Li, J.; Tang, Z.H. Spectral matching based on discrete particle swarm optimization: A new method for terrestrial water body extraction using multi-temporal Landsat 8 images. *Remote Sens. Environ.* **2018**, *209*, 1–18. [CrossRef]
12. Sun, X.X.; Li, L.W.; Zhang, B.; Chen, D.M.; Gao, L.R. Soft urban water cover extraction using mixed training samples and Support Vector Machines. *Int. J. Remote Sens.* **2015**, *36*, 3331–3344. [CrossRef]
13. Huang, X.; Xie, C.; Fang, X.; Zhang, L.P. Combining Pixel- and Object-Based Machine Learning for Identification of Water-Body Types from Urban High-Resolution Remote-Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2097–2110. [CrossRef]
14. Yao, F.F.; Wang, C.; Dong, D.; Luo, J.C.; Shen, Z.F.; Yang, K.H. High-Resolution Mapping of Urban Surface Water Using ZY-3 Multi-Spectral Imagery. *Remote Sens.* **2015**, *7*, 12336–12355. [CrossRef]
15. Wu, W.; Li, Q.; Zhang, Y.; Du, X.; Wang, H. Two-Step Urban Water Index (TSUWI): A New Technique for High-Resolution Mapping of Urban Surface Water. *Remote Sens.* **2018**, *10*, 1704. [CrossRef]
16. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [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. Zhang, L.P.; Zhang, L.F.; Du, B. Deep Learning for Remote Sensing Data A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
19. Zhu, X.X.; Tuia, D.; Mou, L.C.; Xia, G.S.; Zhang, L.P.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
20. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [CrossRef]
21. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene Classification with Recurrent Attention of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1155–1167. [CrossRef]
22. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
23. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]

24. Isikdogan, F.; Bovik, A.C.; Passalacqua, P. Surface Water Mapping by Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4909–4918. [CrossRef]
25. Canty, M.J.; Nielsen, A.A.; Schmidt, M. Automatic radiometric normalization of multitemporal satellite imagery. *Remote Sens. Environ.* **2004**, *91*, 441–451. [CrossRef]
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Teichmann, M.; Weber, M.; Zoellner, M.; Cipolla, R.; Urtasun, R. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv* **2016**, arXiv:1612.07695.
28. Vivone, G.; Alparone, L.; Chanussot, J.; Mura, M.D.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586. [CrossRef]
29. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Susstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2281. [CrossRef]
30. Dao, M.; Kwan, C.; Koperski, K.; Marchisio, G. A joint sparsity approach to tunnel activity monitoring using high resolution satellite images. In Proceedings of the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), New York, NY, USA, 19–21 October 2017; pp. 322–328.
31. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural network. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
32. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
33. Enwright, N.M.; Jones, W.R.; Garber, A.L.; Keller, M.J. Analysis of the impact of spatial resolution on land/water classifications using high-resolution aerial imagery. *Int. J. Remote Sens.* **2014**, *35*, 5280–5288. [CrossRef]
34. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.