remote sensing

MDPI

Article

Mapping Impervious Surfaces in Town–Rural Transition Belts Using China's GF-2 Imagery and Object-Based Deep CNNs

Yongyong Fu [1], Kunkun Liu [1], Zhangquan Shen [1], Jinsong Deng [1,*], Muye Gan [1], Xinguo Liu [2], Dongming Lu [2] and Ke Wang [1]

[1] College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China; yyong_fu@zju.edu.cn (Y.F.); 21714119@zju.edu.cn (K.L.); zhqshen@zju.edu.cn (Z.S.); ganmuye@zju.edu.cn (M.G.); kwang@zju.edu.cn (K.W.)

[2] Department of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China; xgliu@cad.zju.edu.cn (X.L.); ldm@zju.edu.cn (D.L.)

* Correspondence: jsong_deng@zju.edu.cn; Tel.: +86-571-8898-2623

Received: 25 December 2018; Accepted: 29 January 2019; Published: 31 January 2019

Abstract: Impervious surfaces play an important role in urban planning and sustainable environmental management. High-spatial-resolution (HSR) images containing pure pixels have significant potential for the detailed delineation of land surfaces. However, due to high intraclass variability and low interclass distance, the mapping and monitoring of impervious surfaces in complex town–rural areas using HSR images remains a challenge. The fully convolutional network (FCN) model, a variant of convolution neural networks (CNNs), recently achieved state-of-the-art performance in HSR image classification applications. However, due to the inherent nature of FCN processing, it is challenging for an FCN to precisely capture the detailed information of classification targets. To solve this problem, we propose an object-based deep CNN framework that integrates object-based image analysis (OBIA) with deep CNNs to accurately extract and estimate impervious surfaces. Specifically, we also adopted two widely used transfer learning technologies to expedite the training of deep CNNs. Finally, we compare our approach with conventional OBIA classification and state-of-the-art FCN-based methods, such as FCN-8s and the U-Net methods. Both of these FCN-based methods are well designed for pixel-wise classification applications and have achieved great success. Our results show that the proposed approach effectively identified impervious surfaces, with 93.9% overall accuracy. Compared with the existing methods, i.e., OBIA, FCN-8s and U-Net methods, it shows that our method achieves obviously improvement in accuracy. Our findings also suggest that the classification performance of our proposed method is related to training strategy, indicating that significantly higher accuracy can be achieved through transfer learning by fine-tuning rather than feature extraction. Our approach for the automatic extraction and mapping of impervious surfaces also lays a solid foundation for intelligent monitoring and the management of land use and land cover.

Keywords: transfer learning; remote sensing; deep learning; object-based image analysis (OBIA)

1. Introduction

Urban development, which has significantly changed land use and land cover (LULC) patterns over the past 30 years, typically involves the removal of natural surface cover and an increase in impervious surfaces [1]. Impervious surfaces mainly include artificial structures that eliminate water infiltration and soil moisture evaporation; these surfaces include rooftops, roads covered with asphalt and concrete, and parking lots. In recent years, impervious surfaces have been seen as an important indicator of urbanization and play an important role in natural environment assessment [2–4]. A high

impervious surface ratio can cause heavy flooding and "urban heat island" effects, and may also adversely affect ecological environments. Thus, the accurate monitoring and estimation of impervious surfaces is critical for urban planning and sustainable environmental management.

With the increase in the availability of high-spatial-resolution (HSR) imagery, mapping impervious surfaces from HSR images has attracted increasing attention [5–10]. To reduce the high intraclass variability and low interclass distance in HSR imagery, object-based image analysis (OBIA) is a new and evolving paradigm [11] that has achieved significantly high accuracy on information extraction from HSR images [12–14]. Object-based image classification approaches include two main steps: first, an image is divided into homogeneous and continues segments; second, classification is performed based on the attributes of the segments. However, since nearly all the features used are based on the statistical features of pixels or segments, which may exclude the intrinsic qualities of the land cover type from HSR pixels, it is impossible for these features to allow for high discrimination while maintaining robustness [15].

Over the past few years, deep convolutional neural networks (CNNs), which attempt to learn high-level feature representations in a hierarchical manner, have achieved state-of-the-art performance in computer vision, significantly outperforming other methods [16–18]. Thus, CNNs have been applied to RS classification applications [15,19–21]. As deep CNNs require multidimensional inputs, a very simple method has been developed to predict each pixel in an image based on overlapping patches using a sliding-window search method [21–23]. However, this patch-wise procedure has a limited receptive field of a predefined size, and so objects that are obviously larger or smaller than the fixed size may be easily fragmented or misclassified as background. Although some studies try to improve performance using patches centered on the superpixel segmentation as input [24,25], this is not a fundamental solution because of the existence of the inherent tradeoff between the fixed size of the receptive field and the varying size of meaningful sematic image objects in HSR imagery. A new trend in recent research is to employ fully convolutional network (FCN)-based approaches [26], which replace fully connected layers in standard CNNs with convolutional layers for dense class map generation [27–29]. FCN consists of an encoder structure and a decoder structure. The image is first converted to a low-resolution feature representation by using the encoder structure and is then converted to pixel-wise predictions using the decoder structure. However, it is still challenging to restore the identical detailed resolution of the input image during upsampling via learning, which may result in the loss of the detailed information available in HSR imagery [30]. This situation can be worse in complex rural environments because the impervious surface area usually covers a smaller area and is much more sparsely distributed than pervious surfaces. In particular, the capture of valuable edge information remains a challenge.

To solve this problem, it is necessary to combine the edge information provided by image segmentation with the feature learning capability of deep CNNs for information extraction from HSR images. Inspired by this idea, we have developed a framework and applied it to impervious surface extraction. In addition, we investigate two commonly used transfer learning techniques, which are used to expedite the training of deep CNNs. Finally, we compare our approach with the conventional OBIA classification, and FCN-based approaches, such as FCN-8s approach proposed in [26],and the U-Net approach proposed in [31], which are commonly used approaches and achieved success in image classification for remote sensing or nature images.

2. Study Area

We selected the Chongfu subdistrict, which is located in Tongxiang County, the northeastern part of Zhejiang Province, China (120°26′11″E, 30°32′48″N, see Figure 1), as our study area. The Chongfu subdistrict has an area of 110.44 km^2 and is representative of a typical town–rural pattern on the Hang-Jia-Hu plain. It is located in the subtropical monsoon climate zone and has abundant precipitation, distinct seasons, and an annual average temperature of 16.5 degrees centigrade. Due to its favorable climate and physiognomy, the study area provides an ideal environment for agricultural

development. In addition, due to its premium geographical location, essentially the geographical center of the Shanghai-Hangzhou-Suzhou (Huhangsu) triangle within the Huhangsu one-hour economic circle, and its comprehensive transportation network, the area has experienced a rapid development period, during which various impervious surfaces were established due to construction. Thus, its complex town–rural environment renders Tongxiang a suitable area for developing a robust method using HSR images to monitor impervious surfaces.

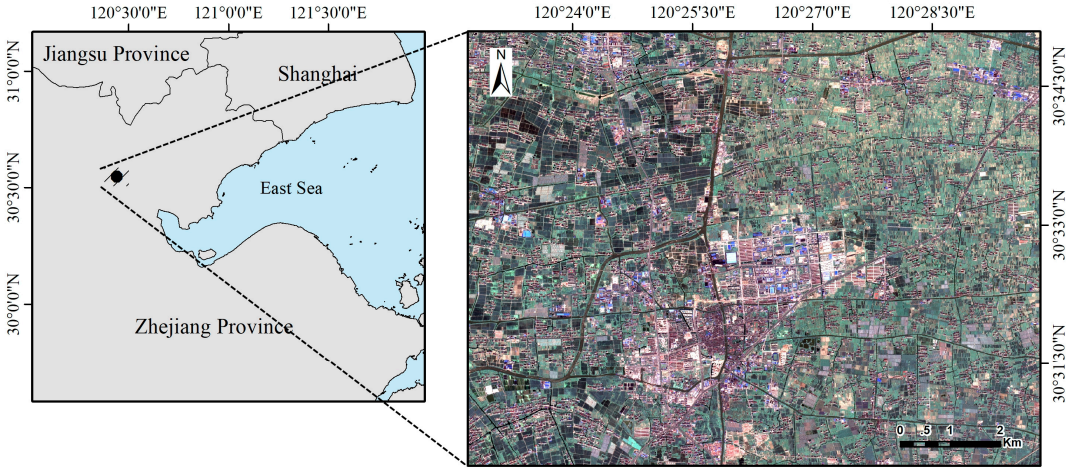


Figure 1. The study area is a typical town–rural region in Zhejiang province. The Gaofen 2 (GF-2) image used in this study is presented in true color.

As shown in Figure 2, visual inspection combining images of the entire area reveals that the impervious surfaces mainly comprise three categories: roads (including asphalt and concrete roads), rooftops (town buildings, industrial warehouses, and rural settlements), and other exposed impervious surfaces (squares, parking lots, and grain-basking fields). The other types of land cover in the study area represent pervious surfaces, such as water bodies (including rivers and ponds), vegetation (crops, shrubs, and trees), and bare land. To effectively extract impervious surfaces, we regard impervious and pervious surfaces as individual classes rather than detailed categories.

(**a**) (**b**)

Figure 2. Image examples on the ground for different types of the impervious and previous surfaces in our study area, including typical examples in rural (**a**), and town areas (**b**).

3. Materials and Methods.

Our overall framework is shown in Figure 3. Following preprocessing and pan sharpening, we firstly acquire the semantically meaningful image objects by applying a segmentation algorithm on the imagery. Next, to follow the data format requirement in transfer learning and avoid abnormal

gradients, standardization and normalization are conducted on every single image object (a set of pixels). Subsequently, the image objects are randomly separated into three individual datasets that are used for training, validation, and testing. Pre-trained inception-resnet v2 is employed for transfer learning with the training set for 30 epochs. We saved and estimated the model after each epoch and selected the best model based on validation set performance. The model was then used to produce the final map of impervious surfaces. To verify whether our method effectively discriminates impervious surfaces, we compared our method with conventional object-based nearest neighbor classification (NNC), FCN-8s, and U-Net.

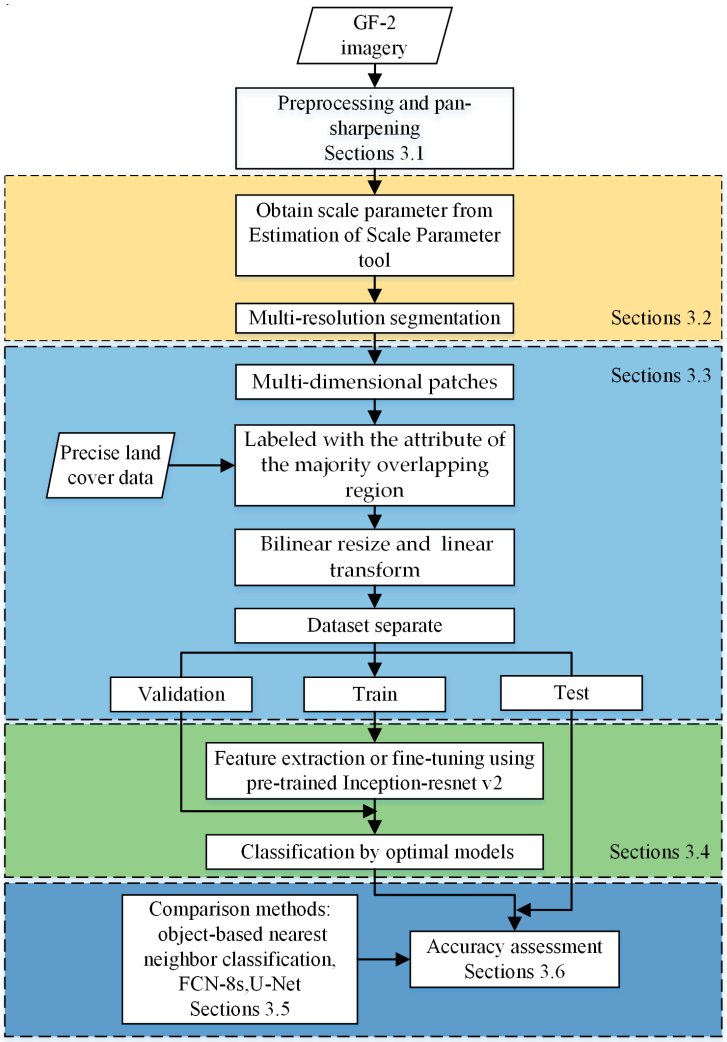


Figure 3. Outline of the overall framework presented in this paper, including image preprocessing, classification, and comparison methods.

3.1. Datasets and Preprocessing

We acquired imagery from the PMS sensor of Gaofen 2 (GF-2), comprising four multispectral bands (MSS) with a spatial resolution of 3.2 m and a panchromatic band (PAN) with a resolution of 0.8 m. The entire study area imagery was acquired on July 22, 2016, under cloud-free atmospheric conditions, thus, atmospheric correction was unnecessary for our LULC classification purposes during preprocessing [32,33]. The acquired MSS image and PAN image were orthorectified into the universal transverse Mercator (UTM) projection system and fused using the Gram–Schmidt (GS) pan-sharpening method in ENVI (version 5.1, Exelis Visual Information Solutions, Boulder, CO, USA, 2014). We then calculated the normalized difference vegetation index (NDVI) and incorporated it into the fused

images by using the layer stacking tool in ENVI. The NDVI values are calculated based on the surface reflectance, which used the Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) atmospheric model implemented in ENVI for atmospheric correction.

We also collected land-use change surveying maps from 2015 (provided by the Bureau of Land and Resources, Tongxiang) covering the entire study area as ancillary data for labeling the image for training and validation. Specifically, we carefully revised the map attributes by visual interpretation and ground surveying.

3.2. Multi-Resolution Segmentation

Following preprocessing and pan-sharpening, image segmentation was performed to produce semantically meaningful image objects. We adopted multiresolution segmentation (MRS), which is a widely used algorithm integrated into eCognition (version 9.0, Trimble Germany GmbH, Munich, Germany, 2014). MRS is a bottom-up region-merging method that iteratively merges small image objects into a larger image object until a heterogeneity threshold is reached [34]. MRS is controlled by three key criteria: scale parameter (SP), shape, and compactness. Choosing an optimal segmentation parameter is typically a subjective trial-and-error process; however, we adopted an objective method based on local variance (LV) to select the optimal SP based on an estimate provided by the Estimation of Scale Parameter 2 (ESP 2) tool [35]. The ESP 2 tool was used to iteratively perform segmentation in fixed step sizes and calculate the LV for each SP. Finally, we plotted the rate of change of the LV (ROC-LV) against the corresponding SP, with the peaks in the ROC-LV curve indicating the optimal SP [36,37]. Because impervious surfaces tend to have a regular shape and compact features, the weight of the shape and compactness criteria was set to 0.8. Meanwhile, we used four bands of GF-2 imagery and NDVI as input raster layers for the MRS algorithm, and assigned them the same weight of 1.

3.3. Standardization and Normalization

CNNs require a multidimensional array as input rather than a single feature vector. Therefore, we extracted multidimensional patches as samples as opposed to single pixels. To accomplish this goal, the image objects were first extracted as patches. These were padded with zero values and labeled with the attribute of the majority overlapping region with the reference data. These patches are original and semantically meaningful image pixels from the HSR imagery; that is, they contain the most useful and essential information for each semantic category. The optimal patch size is expected to cover major semantically meaningful objects; thus, it can vary in accordance with one's interests and the resolution of the images. In this study, we adopted a size of 299 pixels (due to the limitations of transfer learning) and resized the image using a bilinear method. In addition, we eliminated objects with areas of less than 40 m^2 to avoid the influence of defective pixels in ArcGIS (version 10.2, ESRI Inc., Redlands, CA, USA).

In general, the digital values in RS imagery are integers that can have a dynamic range of greater than 8 bits. Thus, during the training phase of CNNs, these values are typically transformed into an approximately normal distribution to avoid abnormal gradients. However, as the transfer learning strategy was adopted in our study, to reduce loss of information and shorten the distance between RS data and the original dataset, we used only three visible bands (the red, green, and blue bands) of the GF-2 imagery in our proposed method, and used a linear transformation method for each band:

$$DN_{\text{norm}} = \frac{DN_{\text{orig}}}{DN_{\text{max}}} \tag{1}$$

where DN_{norm} is the normalized value and DN_{orig} and DN_{max} are the original and maximum pixel values in the image, respectively.

Following normalization, we randomly divided all the image objects into three independent subsets, i.e., the training, validation, and testing sets, which account for 70%, 10%, and 20% of the whole dataset, respectively.

3.4. Transfer Learning Based on a Pre-Trained Inception-Resnet V2 Model

Rather than training a complex deep CNN from scratch, we adopted a transfer learning strategy to reduce the time and amount of labeled data required for training. Transfer learning assumes that knowledge learned from one task can be helpful in improving performance when applied in another task or domain [38]. This strategy has been successfully applied in many image classification applications by employing a CNN from a set of pre-trained weights [39–41]. Typically, pre-trained CNNs include both the model structure and weights, which are fully trained with sufficiently labeled images collected from other applications.

Generally, there are two strategies for employing pre-trained CNNs. As shown in Figure 4, the first one, "feature extraction," regards pre-trained CNNs as feature extractors, as it only reconstructs and fine-tunes the final logits (classifier) layer, while all the rest remain frozen. In our approach, we adopt multinomial logistic regression (also known as softmax regression) as our classifier because it is efficient and straightforward [42]. The second strategy for employing pre-trained CNNs, "fine-tuning," continues training all the layers to keep the output as close to the new task as possible [43]. Both of these commonly used strategies are compared in our study.

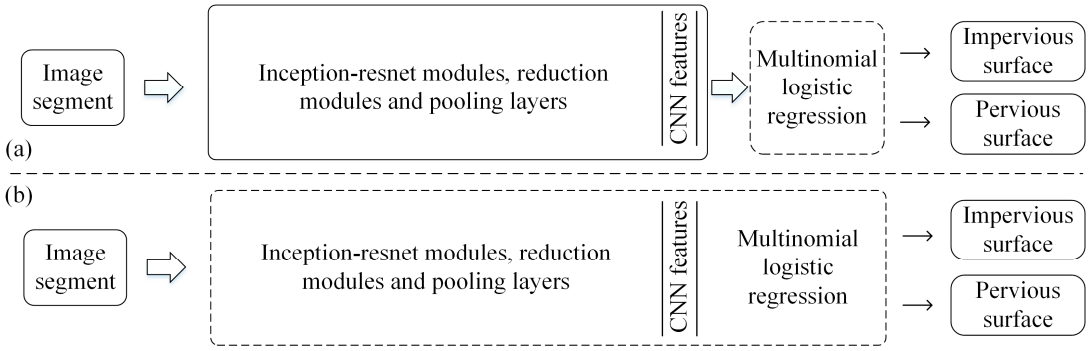


Figure 4. The transfer learning strategies of this study: (**a**) feature extraction and (**b**) fine-tuning. The strategies differ in the trainable part of the pre-training, which is shown in the dotted box.

Considering the performance and accessibility of the deep CNN architecture and our limited amount of labeled RS data, we decided to transfer the pre-trained inception-resnet V2 model [44], which has been fully trained on the ILSVRC-2012-CLS image classification dataset [39]. The inception-resnet V2 model combines the inherent computational efficiency of inception architectures with the accelerative training benefits conferred by residual connections. The inception structure was introduced as a fundamental part in GoogLeNet [40] and has been optimized and refined over a series of iterations [41,44,45]. It was expected that most of the structure used in this model has been optimized with features suitable for image classification purposes.

To implement the pre-trained inception-resnet V2 model, we used the TensorFlow-Slim (TF-Slim, version 1.4) image classification model library, which contains a high-level application programming interface (API) for defining, training, and evaluating complex models. All of the programming code is based on python 3.5.2. Four Tesla P40 graphics processing units (GPUs) on the deep learning service (DLS) of Meituan Open Services (MOS) were used for acceleration. We trained all the models using a batch size of 64, a learning rate of 0.01 decayed exponentially by 0.94 every 2 epochs, and RMSProp optimization with a momentum of 0.9 and decay of 0.9. CNN weights were recorded after every epoch; after all the epochs were completed, we selected the model with the highest accuracy on our validation set as our optimal model.

3.5. Comparison Methods

3.5.1. Object-Based Nearest Neighbor Classification

To provide a comparison with our proposed method, the object-based NNC method was applied to distinguish between impervious and pervious surfaces. NNC is straightforward to implement and

does not require hyperparameter definitions. Moreover, due to its inherent mechanism, NNC can also achieve satisfactory results without detailed classification categories. Thus, NNC is an appropriate method for comparison with our proposed method. Following an identical selection of samples, we used feature space optimization (FSO), a tool available in eCognition, to select the optimal feature combination. Based on selected samples, FSO calculates the Euclidean distance in feature space between classes and chooses the best combination of features, resulting in the largest minimum distances between the least separable classes [46].

3.5.2. Fully Convolutional Neural Networks

FCN-based approaches were also selected for comparison with our proposed method due to their high performance in recent RS applications. We opted to use the widely used FCN-8s and U-Net FCN models, because of their successful performance in image classification for remote sensing or nature images. Both FCN-8s and U-Net are expected to produce accurate and detailed segmentations because they combine semantic information from deep, coarser layers and surface information from shallow, finer layers. FCN-8s, a VGG-16 network with a skip-layer structure, combines its final prediction layer with lower layers (the pool3 pool4 layers). U-Net has a u-shaped architecture consisting of a contracting path (left) and an expensive path (right). Every step in the expensive path combines information in a corresponding lower layer. Detailed information on these model architectures can be found in [26] and [31].

We trained FCN-8s and U-Net on the GF-2 true color images for impervious surface extraction. During training, we modified the number of outputs to be 2, fine-tuned FCN-8s based on an ImageNet pre-trained model, and trained the U-Net from scratch. Additionally, we set the learning rate to 0.0001 to avoid abnormal gradients and batch size to 10 due to the limits of the GPU's memory. The other training parameters for both FCN models are identical to those used in our proposed method. The percent of three independent subsets, i.e., the training, validation and testing sets, are the same with our proposed methods.

3.6. Accuracy Assessment and Comparison

In this paper, we compared the proposed object-based deep CNNs approach with the object-based NNC, FCN-8s, and U-Net methods. We conducted accuracy assessment on the final classification maps, with a total of 44970 randomly selected segments to construct the error matrix. We confirmed whether the segments were correctly identified by visual interpretation. During the visual interpretation, a 2015 land-use change surveying map was used as ancillary data. This map includes accurate spatial distribution information for the detailed categories of impervious and pervious surfaces, thus offering detailed locations of the impervious and pervious surfaces. Finally, the accurate spatial distribution map for these two different types of surfaces was obtained, by visual interpretation, for conducting accuracy assessment. Based on the error matrix, we calculated several commonly used accuracy statistics, including user accuracy (UA), producer accuracy (PA), and overall accuracy (OA).

To compare the accuracies of the classification results between different methods, we employed three commonly used evaluation metrics for impervious surfaces: precision, recall, and F-measure [47,48], which are calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

Another method widely used in the accuracy assessment is the Kappa statistical analysis, which is a discrete multivariate technique to statistically analyze the difference between a classified map and reference map [49]. In this study, the Kappa statistic of every error matrix was calculated. Furthermore, Kappa statistics of two different error matrices were compared in Z-test to measure whether there was significant difference between the two classification methods [49]. The Z-statistic is calculated as follows:

$$Z = \frac{K_1 - K_2}{\sqrt{Var(K_1) + Var(K_2)}} \quad (5)$$

where K_1 and K_2 are the two Kappa statistics, $Var(K_1)$ and $Var(K_2)$ are their estimated variances. The hypothesis that two Kappa statistics are equal is rejected if $|Z|$ is greater than a certain amount (1.96 for a 95% confidence level test).

In this study, we regard the discrimination of impervious and pervious surfaces as binary classification. Therefore, TP and TN are defined as the area of correctly labeled impervious and pervious surfaces. Precision and recall values of different methods were calculated at the pixel level based on the test dataset. Eventually, we calculated five accuracy values for each accuracy statistic using Equations (2), (3), and (4), and a total of 10 Z-statistics between different methods using Equation (5).

4. Results

4.1. Segmentation Results with Optimal Scale Parameter

ROC-LV and LV values are plotted against corresponding SPs in Figure 5. Based on this diagram, the peaks of the ROC-LV curve indicate optimal SPs for segmentation. The graph shows that the scale of 41 represents the first break after a continuous and abrupt decay. As a result, we set 41 as the optimal SP, generating 224,889 segments. As shown in Figure 6, at the scale of 41, rooftops, water bodies, and grain-basking fields can be identified. The segmentation results have a border consistent with the major target in our study and are fully consistent with the expected results in the overall framework shown in Figure 3.

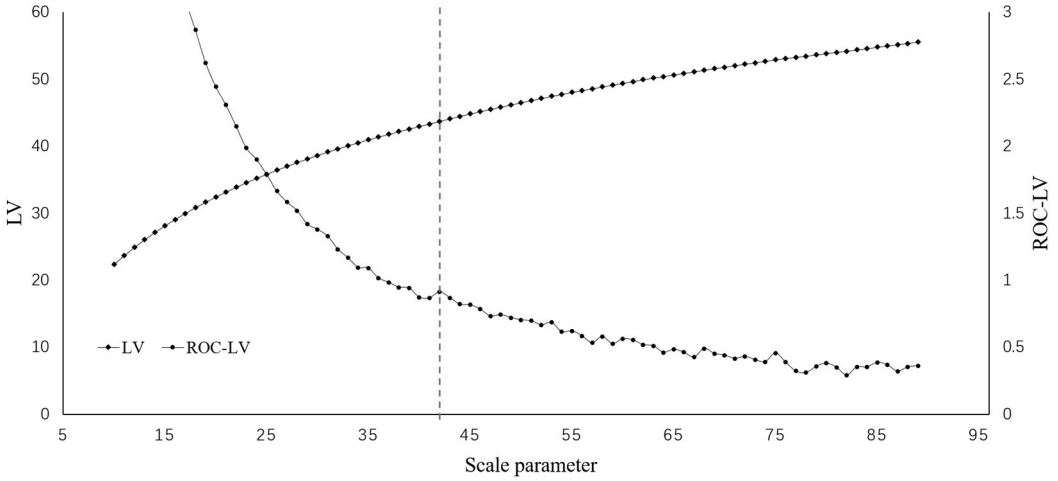
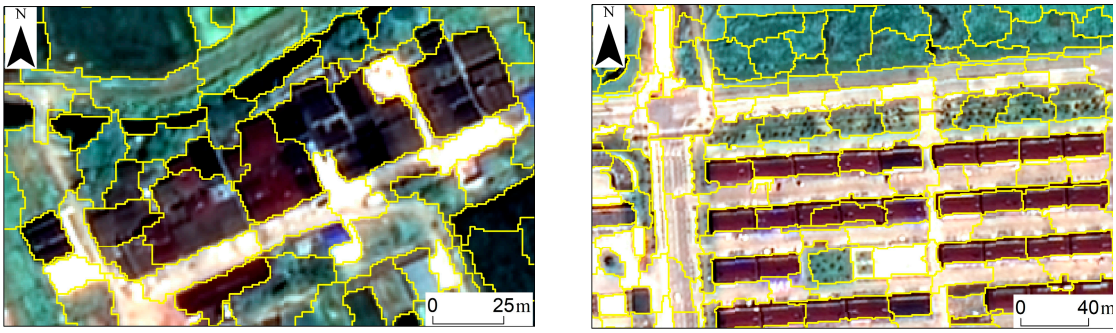


Figure 5. Local variance (LV) and rates of change of LV (ROC-LV) values against corresponding scale parameters (SPs) produced by the Estimation of Scale Parameter 2 (ESP 2) tool. The gray vertical dotted line indicates the optimal SP.



(**a**) (**b**)

Figure 6. Subset examples of the segmentation results in rural (**a**), and town areas (**b**). The yellow lines represent the segmentation results at the scale of 41. Images are presented in true color.

4.2. Optimal Model Selection Results

Figure 7 shows training and validation accuracy values with respect to the number of epochs completed. Clearly, it can be found that the fine-tuning method achieves higher accuracy values than the feature extraction method during the training phase. With an increase in epochs, both of these transfer learning techniques show a more stable trend in accuracy values. The highest validation values achieved by fine-tuning and feature extraction methods are 93.90% and 85.05%, respectively. Based on these highest validation accuracy values, we selected the best model for fine-tuning after training for 17 epochs, and the best model for feature extraction after training for 30 epochs. Vertical dotted lines indicate the optimal models of fine-tuning and feature extraction in Figure 7.

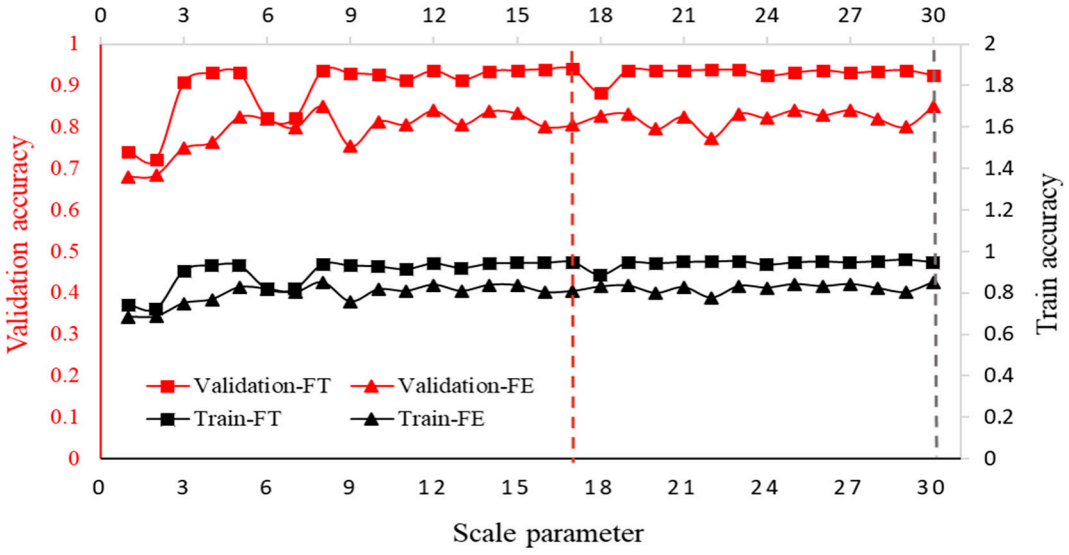


Figure 7. The training and validation accuracy values with respect to the number of epochs. Red and gray vertical dotted lines highlight the highest overall accuracy on validation set using fine-tuning and feature extraction methods, respectively. Validation-FT: The validation accuracy values using our proposed fine-tuning method. Validation-FE: The validation accuracy values using our proposed feature extraction method. Train-FT: The training accuracy values using our proposed fine-tuning method; Train-FE: The training accuracy values using our proposed feature extraction method.

4.3. Final Map and Accuracy Assessment

Some detailed subset examples from the classification results using our proposed method are shown in Figure 8. After visual inspection of the final maps, semantically meaningful objects are accurately delineated, and most of the impervious and pervious surfaces can be classified successfully.

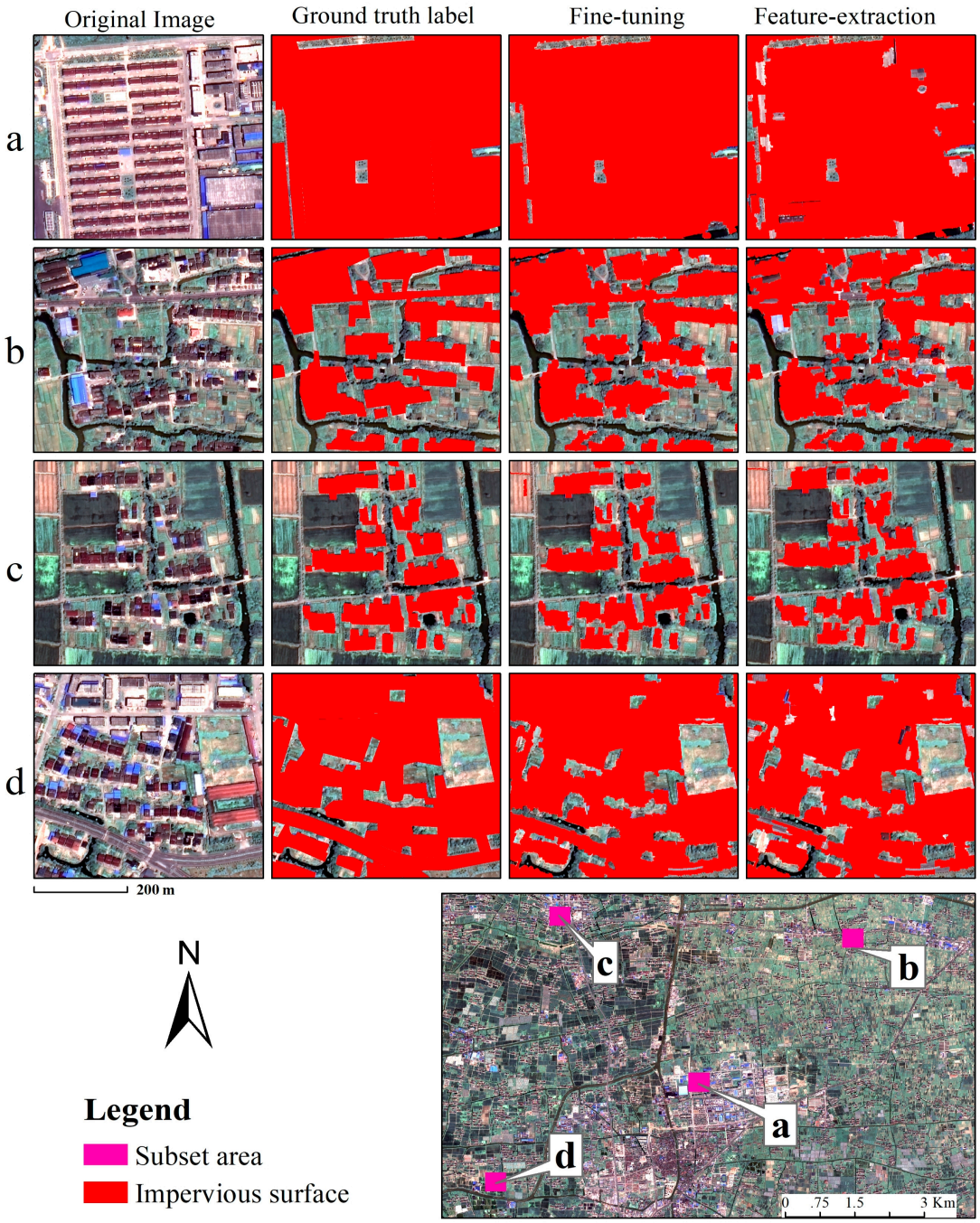


Figure 8. Subset examples from classification results using our proposed object-based deep convolution neural networks (CNNs).

To further quantitatively assess the classification results, we randomly selected over 44,970 segments for the accuracy assessment, which account for approximately 20% of whole image objects. Tables 1 and 2 show the confusion matrices of the classification results. We find that the pervious surfaces produced by the fine-tuning method achieves the highest UA value of 95.85%, indicating that over 95% of the identified pervious surface in the classification results are truly pervious surfaces. The impervious and pervious surfaces have relatively high PA values of 92.13% and 94.82%, respectively, which are produced by the fine-tuning method.

Table 1. Error matrix for the final map using our proposed feature extraction method. PA: producer accuracy, UA: user accuracy, IS: impervious surface, PS: pervious surface.

Predicted Class	**Reference Class**			
	PS	**IS**	**Sum**	**UA**
PS	26,684	3846	30,530	87.40%
IS	2874	11,575	14,449	80.11%
Sum	29,558	15,421	44,979	
PA	90.28%	75.06%		
Overall accuracy	85.06%			

Table 2. Error matrix for the final map using our proposed fine-tuning method.

Predicted Class	**Reference Class**			
	PS	**IS**	**Sum**	**UA**
PS	28,028	1214	29,242	95.85%
IS	1530	14,207	15,737	90.28%
Sum	29,558	15,421	44,979	
PA	94.82%	92.13%		
Overall accuracy	93.90%			

We can also find that the fine-tuning method performs better than the feature extraction method with higher accuracy values, which is consistent with findings in the previous optimal model selection phase. Both the impervious and pervious surfaces classified by feature extraction method have UAs greater than 80%. And all the UAs and PAs of the impervious and pervious surfaces produced by fine-tuning method are greater than 90%. Thus, impervious and pervious surfaces are classified successfully, with all OAs greater than 85%.

4.4. Accuracy Comparison

In this study, we used the object-based NNC and the state-of-the-art FCN-based methods for comparison. Table 3 shows the experiment setup and computational complexity of different classification schemes in this study. In terms of computational complexity, objected-based NNC method is more time intensive relative to the others. With the acceleration of GPU, our proposed methods and the FCN-based methods take less time. Furthermore, our proposed feature extraction method requires the least time. This is because in the other methods that perform training with the fine-tuning strategy, the networks continue training from pre-trained models, where all the weights and biases in the models need to be updated, and more time and resources are required.

Table 3. Experiments setup and computational complexity between classification schemes. OB-NNC: object-based nearest neighbor classification (NNC) method. Ours-FE: our feature extraction method. Ours-FT: our fine-tuning method. GPU: Tesla P40 graphics processing units. CPU: inter i7 7700k with 16 Gb memory.

Methods	**FCN-8s**	**U-Net**	**OB-NNC**	**Ours-FE**	**Ours-FT**
Learning rate	0.0001	0.0001	-	0.01	0.01
strategy	fine-tuning	fine-tuning	-	feature-extraction	fine-tuning
platform	GPU	GPU	CPU	GPU	GPU
time (hours)	4.36	4.91	32.61	4.24	5.18

To quantitatively assess the performance of our proposed method and other methods, precision, recall, and F-measure were calculated at the pixel level (Table 4), and our fine-tuned object-based deep CNN is found to obtain the best performance. The approaches using FCN-8 and U-Net achieve

similar accuracy levels and achieve a F-measure value of approximately 80.0%, and both of them have lower precision and recall values than the fine-tuned object-based deep CNN. We find that the fine-tuned object-based deep CNN achieves the highest F-measure value for impervious surfaces at 88.9%, compared with 80.9% for object-based NNC method, and the highest precision value for impervious surfaces at 89.7%, which is 7.5% higher than that of object-based NNC method.

Table 4. Quantitative comparison between methods using conventional object-based image analysis (OBIA), FCN-8s, U-Net, and our methods at the pixel level.

Methods		FCN-8s	U-Net	OB-NNC	Ours-FE	Ours-FT
Evaluation Criteria	Precision	81.1%	81.7%	82.2%	78.4%	89.7%
	Recall	79.0%	74.3%	79.6%	69.7%	88.1%
	F-measure	80.0%	77.9%	80.9%	73.8%	88.9%
	Kappa coefficient	0.737	0.704	0.751	0.714	0.852

Table 5 summarizes the Kappa analysis results between the five classification methods used in this study. According to the Z-test of the Kappa values, our proposed fine-tuning method was significantly different from all the other methods. Furthermore, since the Kappa value of the fine-tuning method results is greater than all the other methods (Table 4), and $Z \geq 1.96$, the classification results of our proposed fine-tuning method are statistically significantly better than the other five classification results at 5% significance level. However, there was no significant difference among the other four methods.

Table 5. Z-test for a 95% confidence level between methods using conventional OBIA, FCN-8s, U-Net, and our methods.

Methods	FCN-8s	U-Net	OB-NNC	Ours-FE	Ours-FT
FCN-8s		Not significant	Not significant	Not significant	Significant
U-Net	0.57		Not significant	Not significant	Significant
OB-NNC	0.24	0.81		Not significant	Significant
Ours-FE	0.40	0.17	0.64		Significant
Ours-FT	2.30	2.85	2.06	2.66	

It is evident from the evaluation criteria that (1) the fine-tuned object-based deep CNNs achieve the highest precision, recall, and accuracy values, (2) the fine-tuned object-based deep CNNs outperform the feature extraction method, and (3) the fine-tuned object-based deep CNNs produced significantly higher accuracies than all the other methods.

5. Discussion

5.1. Object-Based NNC vs. Our Approach

Our proposed method was first compared with the widely used object-based NNC approach. The major difference between the two methods involves the feature design procedure. First, deep CNNs can automatically learn features, which are present as the weights of each layer [50]. In contrast, the feature space of the object-based NNC method is based on manually designed features. Second, deep CNNs attempt to learn high-level feature representations in a hierarchical manner [51], whereas manually designed feature based methods mainly cover spectrum, shape, and texture features. The majority of these hand-crafted features are based on statistical results, resulting in a lack of generalizability [15]. Thus, the use of deep CNNs improved our classification performance relative to the object-based NNC approach, with an 8% improvement in terms of F-measure at the pixel level.

5.2. Pixel-Wise FCN-Based Methods vs. Our Approach

Compared with state-of-the-art pixel-wise FCN-based methods, our approach takes segments as its basic classification units and extracts accurate semantic information from deep CNNs. Unlike

the standard CNN model, FCN is a pixel-wise classification method composed of downsampling and upsampling processes. The downsampling path is usually a combination of convolutional and max-pooling layers, which are commonly used in the CNNs for extract and interpret the context in image classification tasks. The upsampling path is generally composed of deconvolutional layers, which are used to upsample the feature maps and output the final dense classification results. However, during the downsampling process, FCN's pooling operations can lose a great deal of detailed information. Although the FCN models can be enhanced by combining the feature map of a previous low-level pooling layer, as with FCN-8s and U-Net, it is still challenging to restore highly nonlinear object boundaries during upsampling by learning. Thus, such methods tend to contain salt and pepper noises [52–54] and the detailed boundaries of an object are often lost or smoothed [30].

In our proposed method, we avoid the loss of detailed information by using image objects as basic classification units. First, we employed the MRS algorithm to overcome local spectral variance and provide precise boundary information. With the segments as basic processing units, the detailed and complete image information of land cover in the real word can then be observed by deep CNNs. Additionally, the fixed-size receptive field in FCN-based models becomes a semantically meaningful receptive field before classification; thus, the shape information is also considered in our approach. Besides, compared with the downsampling path in FCN-based models, the pre-trained inception-resnet V2 model with deeper and well-designed structure can provide more accurate semantic information. Therefore, our method can achieve high accuracy with accurate boundary information and avoid the salt and pepper noises. As shown in Figure 9, our approach outperforms FCN-8s and U-Net in terms of detailed information.

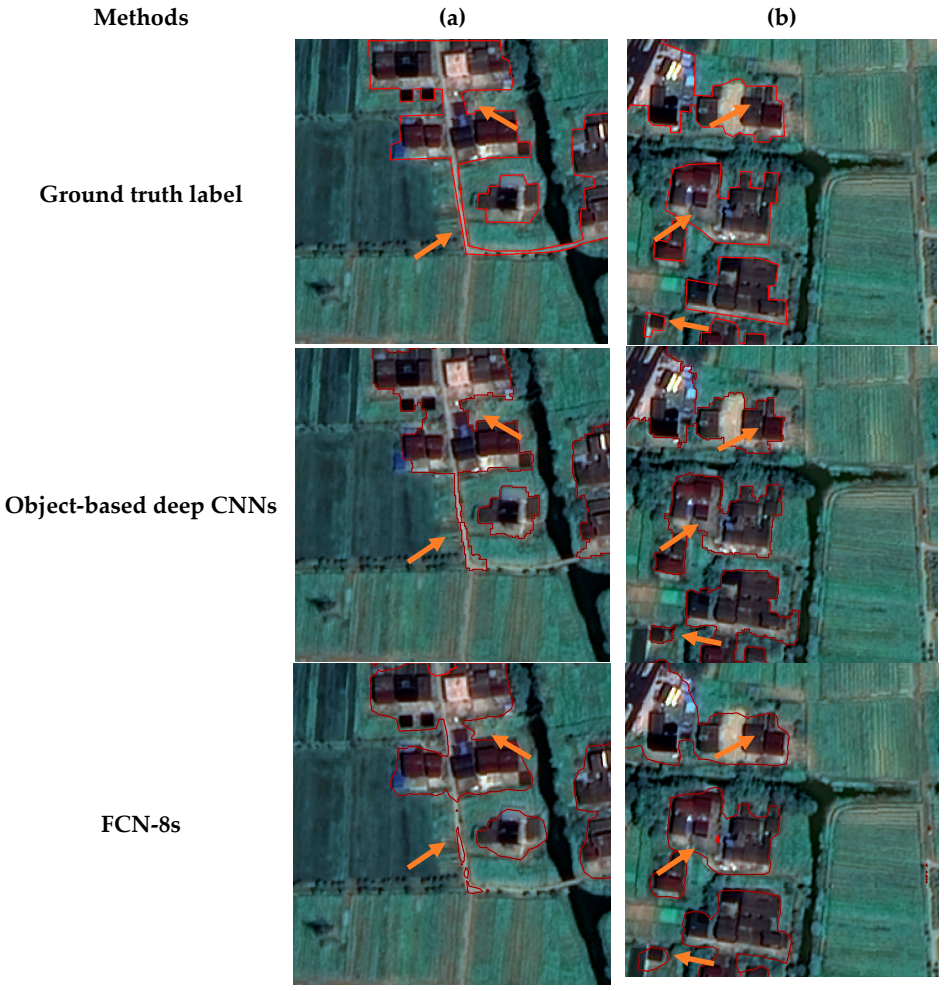


Figure 9. *Cont.*

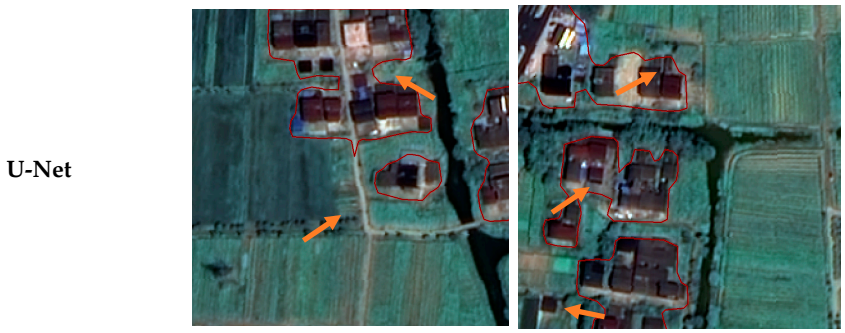


Figure 9. Detailed comparison between our proposed method and FCN based models at subset areas (**a**) and (**b**). The arrows indicate that some impervious surfaces can be extracted accurately by utilizing object-based deep CNNs, while by using the FCN-based models, the salt and pepper noises still persist and detailed boundaries are often smoothed.

5.3. Training Strategies and Scale Effects

Two types of widely used strategies of transfer learning were adopted in this study: feature extraction and fine-tuning. The difference between these options highlights the general applicability of early pre-trained layers. The first one utilizes the early pre-trained layers to produce features that would subsequently be used to extract impervious surfaces, which can achieve satisfactory overall accuracy (lager than 85%), since these layers can extract the low-level and more generic features and can be utilized for other image classification tasks [55]. However, as fine-tuning continues training from a pre-trained model, that is, as it adjusts the weights in each layer to acquire output as close to the new labels as possible, it shows improvements over feature extraction (8.84% in overall accuracy). However, as only the pre-trained inception-resnet V2 model and the multinomial logistic regression as classifier were employed in this study, we emphasize that numerous classifiers and optimization methods should be subjected to further investigation.

In our approach, it is crucial to generate a set of semantically meaningful segments because they are regarded as the basic units for classification. These segments preserve the detailed and complete image information for landcover and provided additional geometry information for the deep CNNs. We chose the MRS algorithm for segmentation since it follows the region-merging principle and can generate satisfactory segmentation results with our imagery. One of the most important and sensitive parameters for the MRS algorithm is SP, which is defined as the maximum threshold for the heterogeneity in an image object. By adjusting SP, image objects at specific-level scales can be generated. Rather than using trial-and-error, we employed an objective method, the ESP 2 tool, to identify optimal SP. However, segmentation errors, as so-called under- and over-segmentation, still persist and have not been completely solved. Therefore, segmentation methods that can directly and accurately delineate the boundaries of different land cover classes are still required.

6. Conclusions

This paper presents an object-based deep CNN framework for impervious surface extraction from VHSR imagery. Compared with the conventional OBIA method and other two commonly used FCN-based methods, the classification accuracy has been obviously improved.

Our proposed method, which is based on a combination of the MRS algorithm and deep CNNs, can effectively map impervious surfaces while retaining detailed information in HSR images. The MRS algorithm, with optimal SP selected by ESP 2 tool, can provide semantically meaningful image objects for our study, and the pre-trained deep CNNs allow us to effectively extract and interpret the context. Besides, our performance comparison using different training strategies indicates that significantly higher accuracy can be achieved through transfer learning by fine-tuning rather than feature extraction.

As future research, we might focus on testing our method for the mapping of other land covers and on images with higher spatial resolution. Additionally, it is necessary to improve and investigate

additional segmentation algorithms to produce a robust and reasonable boundary at different scales. Additionally, other effective models of deep CNNs should be investigated.

Author Contributions: Funding acquisition, J.D. and M.G.; methodology, Y.F.; supervision, J.D. and K.W.; visualization, K.L.; writing—original draft, Y.F.; writing—review & editing, Z.S., J.D., M.G., X.L., D.L., and K.W.

Funding: This research was funded by Zhejiang Provincial Natural Science Foundation of China, grant number LY18G030006; the Basic Public Welfare Research Program of Zhejiang Province, grant number LGN18D010001; National Natural Science Foundation of China, grant number 41701171.

Acknowledgments: We would like to express our appreciation to the Meituan Open Services for freely providing cloud computing services.

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

References

1. Xian, G.; Homer, C. Updating the 2001 National Land Cover Database Impervious Surface Products to 2006 using Landsat Imagery Change Detection Methods. *Remote Sens. Environ.* **2010**, *114*, 1676–1686. [CrossRef]
2. Xu, H.; Shi, T.; Wang, M.; Fang, C.; Lin, Z. Predicting effect of forthcoming population growth–induced impervious surface increase on regional thermal environment: Xiong'an New Area, North China. *Build. Environ.* **2018**, *136*, 98–106. [CrossRef]
3. Lan, Y.; Zhan, Q. How do urban buildings impact summer air temperature? The effects of building configurations in space and time. *Build. Environ.* **2017**, *125*, 88–98. [CrossRef]
4. Mahmoud, S.H.; Gan, T.Y. Long-term impact of rapid urbanization on urban climate and human thermal comfort in hot-arid environment. *Build. Environ.* **2018**, *142*, 83–100. [CrossRef]
5. Cablk, M.E.; Minor, T.B. Detecting and discriminating impervious cover with high-resolution IKONOS data using principal component analysis and morphological operators. *Int. J. Remote Sens.* **2003**, *24*, 4627–4645. [CrossRef]
6. Lu, D.; Weng, Q. Extraction of urban impervious surfaces from an IKONOS image. *Int. J. Remote Sens.* **2009**, *30*, 1297–1311. [CrossRef]
7. Goetz, S.J.; Wright, R.K.; Smith, A.J.; Zinecker, E.; Schaub, E. IKONOS imagery for resource management: Tree cover, impervious surfaces, and riparian buffer analyses in the mid-Atlantic region. *Remote Sens. Environ.* **2003**, *88*, 195–208. [CrossRef]
8. Hu, X.; Weng, Q. Impervious surface area extraction from IKONOS imagery using an object-based fuzzy method. *Geocarto Int.* **2011**, *26*, 3–20. [CrossRef]
9. Zhang, H.; Li, J.; Wang, T.; Lin, H.; Zheng, Z.; Li, Y.; Lu, Y. A manifold learning approach to urban land cover classification with optical and radar data. *Landsc. Urban Plan.* **2018**, *172*, 11–24. [CrossRef]
10. Zhang, H.; Lin, H.; Wang, Y. A new scheme for urban impervious surface classification from SAR images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *139*, 103–118. [CrossRef]
11. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
12. Luo, K.; Li, B.; Moiwo, J.P. Monitoring Land-Use/Land-Cover Changes at a Provincial Large Scale Using an Object-Oriented Technique and Medium-Resolution Remote-Sensing Images. *Remote Sens.* **2018**, *10*, 2012. [CrossRef]
13. Ventura, D.; Bonifazi, A.; Gravina, M.F.; Belluscio, A.; Ardizzone, G. Mapping and classification of ecologically sensitive marine habitats using unmanned aerial vehicle (UAV) imagery and Object-Based Image Analysis (OBIA). *Remote Sens.* **2018**, *10*, 1331. [CrossRef]
14. Zhang, X.; Feng, X. Detecting urban vegetation from IKONOS data using an object-oriented approach. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'05), Seoul, Korea, 29 July 2005.
15. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]

16. Krizhevsky, A.; Sutskever, I.; Geoffrey, E.H. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Neural Information Processing Systems Conference, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25, pp. 1–9.
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
18. Gong, F.Y.; Zeng, Z.C.; Zhang, F.; Li, X.; Ng, E.; Norford, L.K. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Build. Environ.* **2018**, *134*, 155–167. [CrossRef]
19. Romero, A.; Gatta, C.; Camps-valls, G.; Member, S. Unsupervised Deep Feature Extraction for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1349–1362. [CrossRef]
20. Weng, Q.; Mao, Z.; Lin, J.; Guo, W. Land-Use Classification via Extreme Learning Classifier Based on Deep Convolutional Features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 704–708. [CrossRef]
21. Sharma, A.; Liu, X.; Yang, X.; Shi, D. A patch-based convolutional neural network for remote sensing image classification. *Neural Netw.* **2017**, *95*, 19–28. [CrossRef]
22. Santara, A.; Mani, K.; Hatwar, P.; Singh, A.; Garg, A.; Padia, K.; Mitra, P. Bass net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5293–5301. [CrossRef]
23. Lagrange, A.; Le Saux, B.; Beaupere, A.; Boulch, A.; Chan-Hon-Tong, A.; Herbin, S.; Randrianarivo, H.; Ferecatu, M. Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; Volume 15588863, pp. 4173–4176.
24. Audebert, N.; Saux, B.L.; Lefèvre, S. How Useful is Region-based Classification of Remote Sensing Images in a Deep Learning Framework? In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016; pp. 5091–5094. [CrossRef]
25. Xie, F.; Shi, M.; Shi, Z.; Yin, J.; Zhao, D. Multilevel cloud detection in remote sensing images based on deep learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3631–3640. [CrossRef]
26. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 3431–3440.
27. 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]
28. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9. [CrossRef]
29. Volpi, M.; Tuia, D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 881–893. [CrossRef]
30. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
31. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
32. Lin, C.; Wu, C.C.; Tsogt, K.; Ouyang, Y.C.; Chang, C.I. Effects of atmospheric correction and pansharpening on LULC classification accuracy using WorldView-2 imagery. *Inf. Process. Agric.* **2015**, *2*, 25–36. [CrossRef]
33. Song, C.; Woodcock, C.E.; Seto, K.C.; Lenney, M.P.; Macomber, S.A. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects? *Remote Sens. Environ.* **2001**, *75*, 230–244. [CrossRef]
34. Benz, U.C.; Hofmann, P.; Willhauck, G.; Lingenfelder, I.; Heynen, M. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS J. Photogramm. Remote Sens.* **2004**, *58*, 239–258. [CrossRef]
35. Drăguţ, L.; Csillik, O.; Eisank, C.; Tiede, D. Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 119–127. [CrossRef] [PubMed]

36. Drăguţ, L.; Tiede, D.; Levick, S.R. ESP: A tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 859–871. [CrossRef]
37. Belgiu, M.; Drăguţ, L. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS J. Photogramm. Remote Sens.* **2014**, *96*, 67–75. [CrossRef]
38. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
39. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
40. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1–9.
41. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
42. Lei, Y.; Jia, F.; Lin, J.; Xing, S.; Ding, S.X. An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. *IEEE Trans. Ind. Electron.* **2016**, *63*, 3137–3147. [CrossRef]
43. Xu, G.; Zhu, X.; Fu, D.; Dong, J.; Xiao, X. Automatic land cover classification of geo-tagged field photos by deep learning. *Environ. Model. Softw.* **2017**, *91*, 127–134. [CrossRef]
44. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
45. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
46. eCognition Developer. *Trimble eCognition Developer 9.0 User Guide*; Trimble Germany GmbH: Munich, Germany, 2014.
47. Witharana, C.; Lynch, H. An Object-Based Image Analysis Approach for Detecting Penguin Guano in very High Spatial Resolution Satellite Images. *Remote Sens.* **2016**, *8*, 375. [CrossRef]
48. POWERS, D.M.W. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
49. Congalton, R.G.; Green, K. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2008; ISBN 9781420055122.
50. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
51. Luus, F.P.S.; Salmon, B.P.; Van Den Bergh, F.; Maharaj, B.T.J. Multiview Deep Learning for Land-Use Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2448–2452. [CrossRef]
52. Guo, Z.; Shengoku, H.; Wu, G.; Chen, Q.; Yuan, W.; Shi, X.; Shao, X.; Xu, Y.; Shibasaki, R. Semantic Segmentation for Urban Planning Maps Based on U-Net. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 6187–6190.
53. Kellenberger, B.; Volpi, M.; Tuia, D. Learning class- and location-specific priors for urban semantic labeling with CNNs. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, UAE, 6–8 March 2017; p. 16868016.
54. Sun, Y.; Zhang, X.; Zhao, X.; Xin, Q. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. *Remote Sens.* **2018**, *10*, 1459. [CrossRef]
55. Özbulak, G.; Aytar, Y.; Ekenel, H.K. How transferable are CNN-based features for age and gender classification. In *Lecture Notes in Informatics (LNI), Proceedings—Series of the Gesellschaft fur Informatik (GI)*; IEEE: Darmstadt, Germany, 2016; pp. 1–6.