 doi:10.20944/preprints201808.0112.v2

Article

Fully Connected Conditional Random Fields for High-Resolution Remote Sensing Land Use/Land Cover Classification with Convolutional Neural Networks

Bin Zhang [1,*], CunPeng Wang, Yonglin Shen and Yueyan Liu

School of Geography and Information Engineering, China University of Geoscience, Wuhan 430074; Wangcunpeng@cug.edu.cn (C.W.); shenyl@cug.edu.cn (Y.S.); yueyanliu@cug.edu.cn (Y.L.)
* Correspondence: zhangbin@cug.edu.cn; Tel.: +86-134-3717-2339

Abstract: The interpretation of land use and land cover (LULC) is an important issue in the fields of high-resolution remote sensing (RS) image processing and land resource management. Fully training a new or existing convolutional neural network (CNN) architecture for LULC classification requires a large amount of remote sensing images. Thus, fine-tuning a pre-trained CNN for LULC detection is required. To improve the classification accuracy for high resolution remote sensing images, it is necessary to use another feature descriptor and to adopt a classifier for post-processing. A fully connected conditional random fields (FC-CRF), to use the fine-tuned CNN layers, spectral features, and fully connected pairwise potentials, is proposed for image classification of high-resolution remote sensing images. First, an existing CNN model is adopted, and the parameters of CNN are fine-tuned by training datasets. Then, the probabilities of image pixels belong to each class type are calculated. Second, we consider the spectral features and digital surface model (DSM) and combined with a support vector machine (SVM) classifier, the probabilities belong to each LULC class type are determined. Combined with the probabilities achieved by the fine-tuned CNN, new feature descriptors are built. Finally, FC-CRF are introduced to produce the classification results, whereas the unary potentials are achieved by the new feature descriptors and SVM classifier, and the pairwise potentials are achieved by the three-band RS imagery and DSM. Experimental results show that the proposed classification scheme achieves good performance when the total accuracy is about 85%.

Keywords: remote sensing; image classification; fully connected conditional random fields (FC-CRF); convolutional neural networks (CNN)

1. Introduction

Remote sensing has become an important means of obtaining information about the earth's surface. With the continuous improvements in sensor technology, high-resolution multi-spectral images can be obtained, such as IKONOS, QuickBird, WorldView-2, ZY-3C, and GF-1, as well as high-resolution remote sensing data with spatial resolution close to one meter or even the sub-meter level. In addition, with Unmanned Aerial Vehicle (UAV) aerial technology, a large number of decimeter-scale ultra-high resolution remote sensing images can be acquired. At present, the interpretation of high-resolution remote sensing images, especially high spatial resolution images, is of paramount importance in many practical areas, such as urban environments, precision agriculture, infrastructure, forestry survey, military target identification, and disaster assessment. Recent

 doi:10.20944/preprints201808.0112.v2

technological developments have significantly increased the amount of available remote sensing imagery.

Image classification is an important step in many remote sensing applications and refers to the task of identifying the category of every pixel in an image. Due to the introduction of deep learning methods, especially convolutional neural networks (CNN), and the availability of large-scale annotated datasets, remarkable progress has been achieved in image classification [1–3]. Currently, three main CNN strategies have been successfully adopted for remote sensing image classification: (1) full trained CNN, (2) fine-tuned CNN, and (3) pre-trained CNN used as feature extractors [1,2].

Fully training a new or existing CNN gives full control of the architecture and parameters, which tends to yield a more robust network. However, this strategy not only requires a large amount of computational resources to train CNN's parameters, but also needs a large amount of annotated remote sensing data. Although this amount of annotated remote sensing data is unusual, there are many works, usually using reasonable datasets (more than 2000 images), that achieved promising results by proposing the full training of new CNN [4–6]. Nogueira et al. [4] fully trained a new CNN architecture to classify images from aerial and multispectral images. Yue et al. [5] proposed a hybrid method combining principal component analysis, CNN, and logistic regression to classify hyperspectral image using both spectral and spatial features. Maggiori et al. [6] proposed a CNN architecture that is fully convolutional that only involves a series of convolution and deconvolution operations to produce the output classification maps. Makantasis et al. [7] exploited a CNN to encode pixels' spectral and spatial information and a multi-layer perceptron to conduct the classification task. Volpi [8] presented a CNN-based system relying on a down-sample-then-up-sample architecture for semantic labeling of sub-decimeter resolution images. However, some drawbacks exist in this strategy [8]. For example, Nogueira et al. [1] and Castelluccio et al. [9] trained the networks by only using the existing satellite image dataset, which had lower classification accuracy compared with using the pre-trained networks as global feature extractors or fine-tuning the pre-trained networks. The reason for this may be because the large-scale networks usually contain millions of parameters to be learned. Thus, training them using small-scale satellite image datasets causes overfitting and local minimum problems. Consequently, some constructed a new smaller network and trained it from scratch using satellite images to better fit the satellite images [1].

Another strategy to exploit CNN is to fine-tune its parameters using new data because the first-layer features resemble either Gabor filters, edge detectors, or color blobs. Specifically, fine-tuning adjusts the parameters of a pre-trained network by resuming the training of the network from a current parameter setting but considers a new dataset. In Girshick et al. [10], the authors showed that fine-tuning a pre-trained CNN on the target data significantly improves the performance. They fine-tuned AlexNet [11] and outperformed results for semantic segmentation. Zhao et al. [12] fine-tuned a couple of networks, outperforming state-of-the-art results in classification of traditional datasets. Several works [5,13] in the remote sensing community also exploited the benefits of fine-tuning pre-trained CNN. Xie et al. [13] evaluated a fully-trained CNN against a fine-tuned one to detect poverty using remote sensing data. Yue et al. [5] used the fine-tuning method to classify hyperspectral images.

Based on the aforementioned characteristics, CNN can also be exploited as a feature extractor. These features, usually called deep features, are obtained by removing the last classification layer and considering the output of previous layer (or layers). In some prior studies, CNN were shown to perform well even for datasets with different characteristics from the ones on which they were trained, without performing any adjustment, using them as feature extractors only and using the features according to the application (e.g., classification, retrieval, etc.). In remote sensing domains, Penatti et al. [14] evaluated the use of different CNN as feature extractors, achieving state-of-the-art results in two remote sensing datasets, outperforming several well-known visual descriptors. Hu et al. [15] extracted features of several pre-trained CNN to classify high-resolution remote sensing imagery.

Though these three strategies have their own advantages and disadvantages, to make features more discriminative, choosing the second strategy, which involves fine-tuning CNN's parameters with a new training dataset, is most suitable. Remote sensing datasets for image classification to train

 doi:10.20944/preprints201808.0112.v2

CNN are available, such as the ISPRS 2D semantic labeling datasets, and the Zurich Summer Dataset. Although CNNs are commonly considered a well-performing and promising solution for image classification [1], the problem of segmentation can have multiple solutions, and therefore defines the exact architecture of the CNN. For instance, previous per-pixel approaches that classify each pixel in remote sensing data have used the spectral information in a single pixel from hyperspectral (or multi-spectral) imagery that consists of different channels with narrow frequency bands. This pixel-based classification method alone is known to produce salt-and-pepper effects due to misclassified pixels [16] and has had difficulties in dealing with the rich information from very high-resolution data [17,18]. Works that include more spatial information in the neighborhood of the pixel to be classified have been published [19,20]. Moreover, conditional random fields (CRF) have been broadly used in semantic segmentation to combine class scores computed by multi-way classifiers with low level information captured by the local interactions of pixels and edges [21,22] or superpixels [23]. Even though increasingly sophisticated methods have been proposed to model the hierarchical dependency [24–26] and/or high order dependencies of segments [27,28], we used the fully connected pairwise CRF proposed by Chen et al. [20] for its efficient computation and ability to capture fine edge details while also managing long range dependencies.

Thus, we propose a new classification architecture that combines fine-tuned CNN, spectral features, and digital surface model (DSM) with support vector machine (SVM), called Fully Connected CRF (FC-CRF). First, inspired by the down-sample and up-sample of CNN architecture [8], we fine-tune the parameters of CNN to form probability features. Then, the spectral features and DSM are used to form another probability feature using a SVM classifier. Combining these two types of probabilities, the feature descriptor of each pixel in entire image is achieved. Finally, a SVM classifier is adopted to form the unary potentials of FC-CRF, and pairwise potentials are formed by spectral features and DSM.

The main contributions of this paper are summarized as follows: We use true UAV multispectral imagery with a spatial resolution of 0.4 m along with a DSM of the area for LULC classification. We develop a novel approach for high-resolution remote sensing imagery per-pixel classification of four classes (homestead, impervious surface, tree, and background) using fine-tuned CNN, spectral features, and FC-CRF, which are effective for LULC classification, achieving a classification accuracy of 85%. We show how the proposed method can improve the classification and reduce the limitations of using per-pixel approaches, effectively removing salt-and-pepper effects.

The remainder of the paper is organized as follows. The method is presented in Section 2. The experimental results and analysis are shown in Section 3. Finally, a conclusion and future work are given in Section 4. Some important items and corresponding abbreviations are listed in Table 1.

Table 1. List of some important items and corresponding abbreviations.

LULC	Land Use and Land Cover
DSM	Digital Surface Model
UAV	Unmanned Aerial Vehicle
NDSM	Normalized Digital Surface Model
dCNN	Convolutional Neural Networks
SVM	Support Vector Machine
FC-CRF	Fully connected Conditional Random Fields
Overall Accuracy	OA

2. CNN and Fully Connected CRF

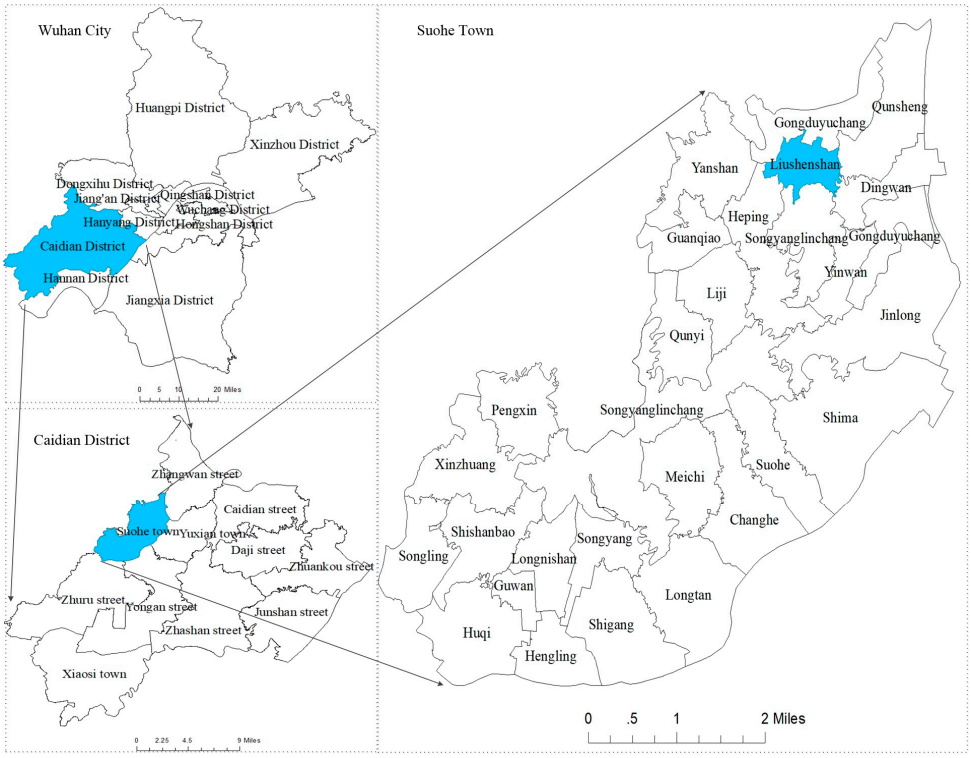
The data and the process of manually labeling the data are outlined in Section 2.1. An introduction to CNN is provided in Section 2.2., and the process of concatenating probabilities for feature descriptors is described in Section 2.3. An introduction to the FC-CRF model and the work flow of the proposed method are provided in Sections 2.4. and 2.5., respectively.

2.1. Data and Manual Labeling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

The data that were used in this work included a rural area map whose location is in the rural area of Wuhan City, China, and consisted of several north-oriented multispectral true orthography bands, and a digital surface model (DSM) that was generated from the UAV imagery using stereo vision. True ortho is a composition of many images to an accurate, seamless two-dimensional (2D) image mosaic that represents a true nadir rendering. The true ortho imagery was exactly matched to the used DSM. The use of a DSM increases classification accuracy by providing height information that can help distinguish between similar looking LULC categories, for example buildings and dark-colored ground. A DSM is invariant to lighting and color variations and can produce a better geometry estimation and background separation [29]. However, the measurements in a DSM need to be normalized because they are measured from a global reference point and not from the local surrounding ground level. In remote sensing, the definition and acquisition of reference data are often critical problems [30]. Most datasets that use classification at a pixel-level only use a few hundred reference points [29,31].

We used a rural area around Wuhan City as the study area (Figure 1a), which has different LULC types, such as homestead, water, farmland, road, trees, and low vegetation. The Vaihingen datasets for the 2D semantic labeling contest was organized by the International Society for Photogrammetry and Remote Sensing (ISPRS) Commission III.4. There are six class types in the Vaihingen datasets: buildings, cars, low vegetation, trees, impervious surfaces, and background. Compared with the Vaihingen datasets, there are four class types in the LULC classification dataset in rural areas of Wuhan city: homestead, impervious surfaces, trees, and background, which contains farmland, low vegetation, and water. Each map from Figure 1b–d contains 8801 × 9007 pixels. In order to validate the classifier using a much larger pool of validation points and for supervised learning and fine-tuning, each pixel in the full map was manually labeled. Figure 1b,c show the high-resolution remote sensing image and DSM, respectively, which were used to manually classify the map, and Figure 1d shows the finished labeled map.



(**a**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

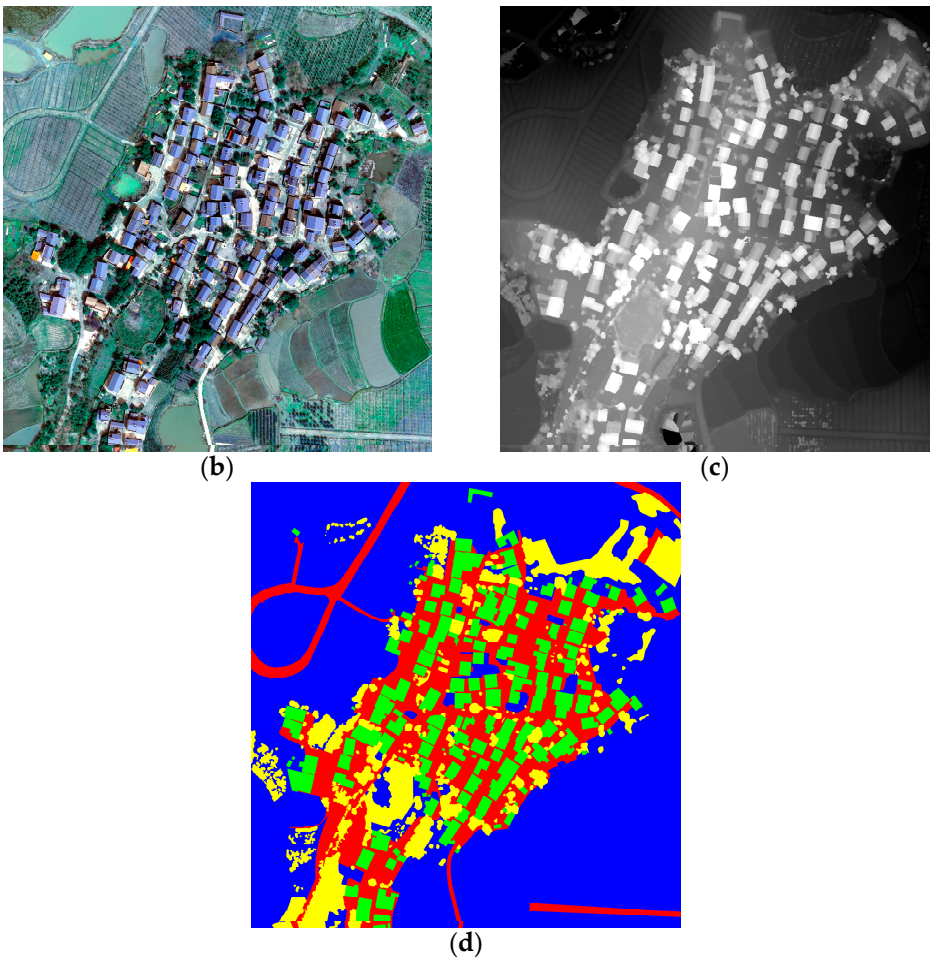


Figure 1. The LULC classification datasets in rural areas of Wuhan city. (**a**) Research location in Suohe Town, Caidian District, Wuhan City, Hubei Province, China; (**b**) high-resolution remote sensing image; (**c**) NDSM; and (**d**) ground truth (C1 Red: impervious surfaces; C2 Green: homestead; C3 Blue: background; C4 Yellow: trees).

2.2. CNN Architecture

The CNNs are composed by a sequential hierarchy of processing layers displayed in Figure 2. From the input to the final classification layer, the data pass through a series of trainable units. The architecture contained down-sampling, up-sampling, and prediction layers, which is described in detail in Volpi et al. [8]. The topmost layer, which is a prediction layer, is composed by a classifier. We used the commonly employed multinomial logistic regression, whose scores (class-conditional probabilities) are given by the SoftMax function:

$$y_i = \frac{\exp(x_i)}{\sum_{i=1}^{C} \exp(x_i)} \quad (1)$$

For C classes, x_i is a C-dimensional vector representing unnormalized scores for the location i, as given by the penultimate layer (the one before the loss function), and y_i is the class-conditional probabilities.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

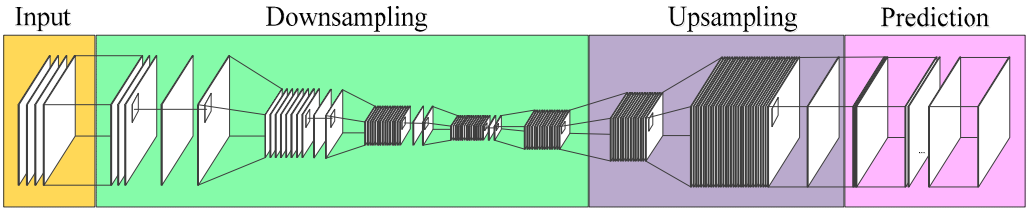


Figure 2. The structure of a dense CNN.

2.3. Feature Descriptors

According to Equation (1), the probability belong to each LULC type can be calculated as:

$$P_{FCNN}=[p_1^{FCNN},\cdots,p_C^{FCNN}]^T \quad (2)$$

where *C* is the number of LULC class type, *T* is transpose operation, and P_{FCNN} is a C-dimensional vector representing normalized scores (class-conditional probabilities), which can be calculated using Equation (1).

The spectral features and normalized DSM combined with training samples were used to estimate the SVM classifier parameters and the probability of vectors belonging to each LULC class, expressed as:

$$P_{CMP}=[p_1^{CMP},\cdots,p_C^{CMP}]^T \quad (3)$$

The feature descriptors for the final classification are given as follows:

$$P_{MF}=[P_{FCNN}{}^T, P_{CMP}{}^T]^T \quad (4)$$

The probabilities of Equations (2) and (3) are concatenated to form a new feature descriptor. The feature descriptor P_{MF} is a 2C-dimensional vector that was adopted to form unary potentials for the FC-CRF model.

2.4. Fully Connected CRF

CRFs have been often employed to smooth noisy segmentation maps [25,33]. Typically these models couple neighboring nodes, favoring same-label assignments to spatially proximal pixels. Qualitatively, the primary function of these short-range CRF is to clean up the spurious predictions of weak classifiers built on top of local hand-engineered features.

Using contrast sensitive potentials [24] in conjunction with local-range CRF can potentially improve localization but thin structures are often still missed, typically requiring solving a computationally expensive discrete optimization problem. To overcome these limitations of short-range CRF, we integrated the FC-CRF model of Lucchi et al. [23] into our system. The model employs the energy function:

$$E(x)=\sum_i \theta_i(x_i)+\sum_{ij}\theta_{ij}(x_i,x_j) \quad (5)$$

where x is the label assignment for pixels. We used $\theta_i(x_i)=-\log(p(x_i))$ as the unary potential, where $p(x_i)$ is the label assignment probability at pixel *i* as computed by the SVM classifier. Equation (4) was adopted to produce the feature descriptors for the SVM classifier.

In the FC-CRF model, each node of the graph is assumed to be linked to every other pixel in the image. Using these higher-order potentials, the method is able to consider not only neighboring information but also long-range interactions between pixels. The pairwise potential has a form that allows for efficient inference while using a fully-connected graph, i.e., when connecting all pairs of image pixels, *i,j*. In particular, as in Rother et al. [21], we used the following expression:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

$$\theta_{ij}(x_i,x_j)=\omega u(x_i,x_j)[\exp(-\frac{||p_i-p_j||^2}{2\delta_\alpha^2}-\frac{||I_i-I_j||^2}{2\delta_\beta^2}) + \exp(-\frac{||p_i-p_j||^2}{2\delta_\gamma^2})] \quad (6)$$

where $u(x_i,x_j)=1$, if $x_i \neq x_j$, and zero otherwise, which, as in the Potts model, means that only nodes with distinct labels are penalized. The remaining expression uses two Gaussian kernels in different feature spaces: the first bilateral kernel depends on both pixel positions (denoted as i, j) and feature vector (denoted as I), and the second kernel only depends on pixel positions. The hyper parameters δ_α, δ_β, and δ_γ control the scale of Gaussian kernels. The first kernel forces pixels with similar color and position to have similar labels, while the second kernel only considers spatial proximity when enforcing smoothness. An efficient inference approach was introduced under the restriction that the pairwise potentials are a linear combinations of Gaussian kernels over an Euclidean feature space [33,34]. This approach, which is based on taking a mean field approximation of the original CRF, is able to produce accurate segmentations in a few seconds. We solved Equation (9) efficiently using the mean field approximation approach proposed in Orlando et al. [33].

2.5. Work Flow of Proposed Method

An overview of the method used can be seen in Figure 3. The pre-trained CNN parameters, which were pre-trained by the Vaihingen dataset combined with the LULC classification dataset in rural areas of Wuhan city, were fine-tuned. The remote sensing images and normalized DSM with training samples were adopted to achieve SVM parameters. Then, the probabilities of the testing image in Equations (2) and (3) were achieved. The feature descriptors in Equation (4) were extracted by concatenating these probabilities. Finally, by introducing the FC-CRF model, the classification results were obtained. A method that concatenates multiple probabilities for feature descriptor, and adopts FC-CRF model for post-processing (CMP-FCCRF) is proposed for LULC classification in this study.

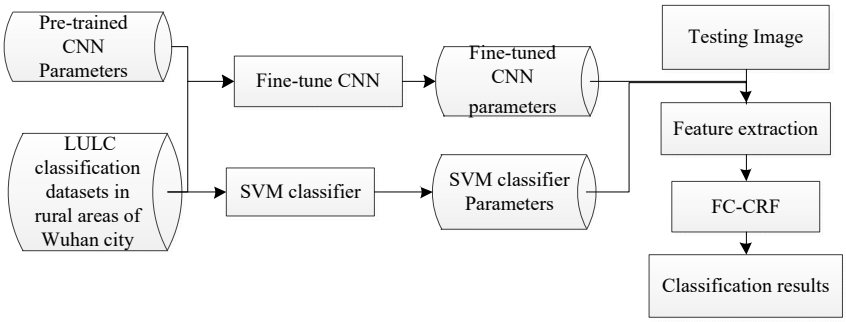


Figure 3. The flow chart of proposed method.

We conducted experiments using high-resolution aerial images to evaluate the effectiveness of the proposed CMP-FCCRF framework for LULC classification. Based on Nogueira's study [1], comparative experiments were conducted by combining feature descriptors and classification methods. We compared the different methods using a confusion matrix and overall accuracy. The fine-tuned CNN features, hand-crafted features, and classifiers associated with FCNN-SVM, SD-SVM, CMP-SVM, FCNN-FCCRF, SD-FCCRF, and CMP-FCCRF are reported in Table 2. The details of these methods are described as follows.

Table 2. Information of different classification schemes.

Method	Feature Descriptors	Classifier
FCNN-SVM	Fine-tuned CNN–Equation (2)	SVM
SD-SVM	Spectral feature and NDSM combined with SVM–Equation (3)	SVM
CMP-SVM	Concatenating probability vector–Equation (4)	SVM
FCNN-FCCRF	Fine-tuned CNN–Equation (2)	FC-CRF
SD-FCCRF	Spectral feature and NDSM combined with SVM–Equation (3)	FC-CRF
CMP-FCCRF	Concatenating probability vector–Equation (4)	FC-CRF

FCNN-SVM uses only the fine-tuned CNN probabilities as a feature. After the CNN's parameters fine-tuned, the SVM method is adapted to achieve classification results [34]. SD-SVM is similar to FCNN-SVM, but they differ in the selection of feature descriptors. Spectral features and normalized DSM are considered feature descriptors in this technique. In CMP-SVM, the concatenating feature of fine-tuned CNN probabilities and SD-SVM probabilities are considered as the feature descriptors, and SVM classifier is used for LULC classification. In FCNN-FCCRF, fine-tuned CNN probabilities are used for the feature vector, and FC-CRF is adopted for LULC classification. In SD-FCCRF, spectral features and NDSM are considered as the feature descriptors in this method, which is combined with FC-CRF to achieve the classification results. Finally, in CMP-FCCRF, the concatenating feature of fine-tuned CNN probabilities and SD-SVM probabilities are considered as the feature descriptors, and the FC-CRF classifier is used for LULC classification.

3. Results

3.1. Experimental Results and Discussion

The proposed method was evaluated on the LULC classification datasets in rural areas of Wuhan City, as shown in Figure 1. Six images were cut from Figure 1b,c, where each image contained about 2000×2000 pixels, to fine-tune the CNN parameters. The CNN parameters were pre-trained by Vaihingen datasets [6]. When we fine-tuned the CNN parameters, the learning rates were initialized at 10−5 and kept uniform for 20 epochs. The batch size of the LULC classification datasets in rural areas of Wuhan city was 128. The codes ran on a computer with Intel® Core™ i7-6500U CPU @ 2.50 GHz 2.50 GHz, NVIDIA Quadro M500M (2 GB), 16 GB RAM, 512 GB SSD and MATLAB 2016a. The gradient was computed via batch gradient descent, which was not computed by GPU.

The testing image shown in Figure 4a contains 2361 × 2406 pixels. We used 500 training samples for the SVM classifier for each method. In Equation (6), the parameter ω for FCNN-FCCRF, CMP-FCCRF, and SD-FCCRF methods, which was adopted to balance the unary potentials and pairwise potentials, was five. The effects of classification results using different methods can be seen in Figure 4. In Figure 4d–f, there are some salt-and-pepper misclassifications, impervious surfaces being classified as homestead, trees being classified as homestead, and patches of impervious ground in the middle of the homestead. The shadows from the homestead caused impervious ground to be misclassified as homestead. Figure 4g–i show the result after FC-CRF smoothing, and t the salt-and-pepper effects were removed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

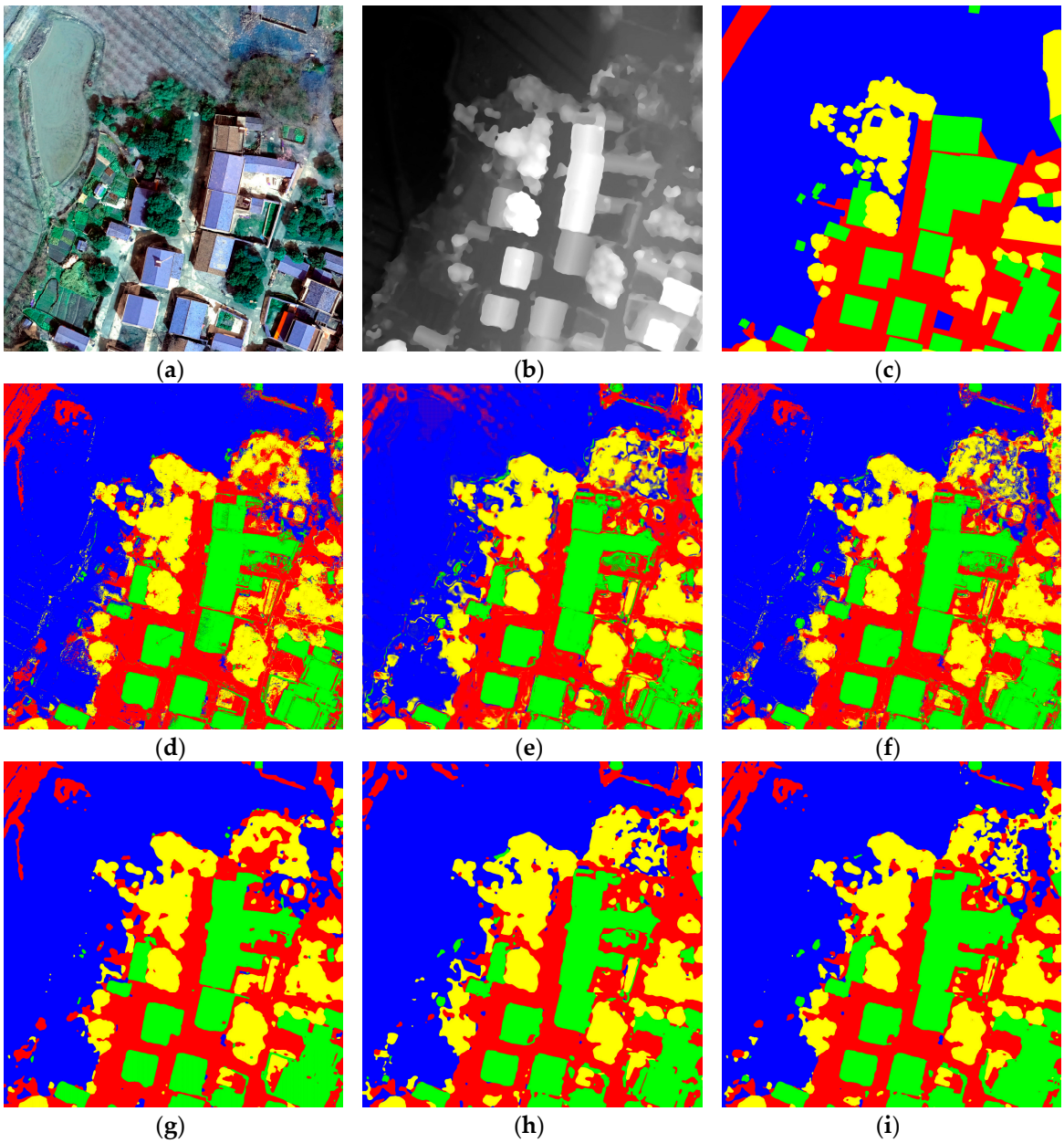


Figure 4. Classification results of LULC classification dataset in rural areas of Wuhan city. (**a**) remote sensing image; (**b**) NDSM; (**c**) ground truth; (**d**) SD-SVM classification; (**e**) FCNN-SVM classification; (**f**) CMP -SVM classification; (**g**) SD-FCCRF classification; (**h**) FCNN-FCCRF classification; (**i**) CMP-FCCRF classification (C1 Red: impervious surfaces; C2 Green: homestead; C3 Blue: background; C4 Yellow: trees).

Table 3 shows the confusion matrix for the different classification methods when combining different feature descriptors and classifiers. It can be seen that confusion often occurred between homesteads that were classified as impervious surfaces. We think this was caused by the large amount of shadows that existed on many of the roads from the surrounding forest and buildings, which is known to complicate the interpretation of areas nearby such objects [27]. There was also confusion between background and impervious surfaces. This is because these two class types may have similar spectral features. We think this occurred when buildings were located close to dense forest areas. The results show that the algorithms in which spatial contextual information is considered significantly outperformed the SVM classification in classification accuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

Table 3. Confusion matrix and classification accuracy of different classifiers.

Classification Method	Confusion Matrix (%)					OA (%)
		C1	C2	C3	C4	
SD-SVM	C1	82.28	6.69	3.81	7.23	81.07 ± 0.6050
	C2	11.78	85.11	2.60	0.51	
	C3	13.91	0.99	77.78	7.32	
	C4	9.85	1.30	7.10	81.74	
FCNN-SVM	C1	76.13	8.84	7.09	7.95	78.89 ± 0.3141
	C2	11.38	86.54	0.86	1.22	
	C3	12.71	1.25	75.77	10.28	
	C4	7.11	0.81	3.68	88.40	
CMP-SVM	C1	80.34	9.04	3.69	6.93	83.29 ± 0.1192
	C2	7.67	90.13	1.17	1.02	
	C3	10.32	1.05	80.02	8.62	
	C4	6.66	0.80	5.02	87.51	
SD-FCCRF	C1	86.09	6.05	1.83	6.03	83.03 ± 1.1030
	C2	10.19	87.47	2.00	0.35	
	C3	13.07	0.42	79.50	7.02	
	C4	7.59	1.24	6.39	84.78	
FCNN-FCCRF	C1	79.42	8.62	5.34	6.62	80.06 ± 0.4239
	C2	9.31	89.19	0.70	0.81	
	C3	11.79	0.62	77.71	9.88	
	C4	6.18	0.61	2.63	90.58	
CMP -FCCRF	C1	83.34	9.04	1.75	5.87	84.73 ± 0.1566
	C2	6.04	92.51	0.80	0.65	
	C3	9.72	0.67	81.23	8.38	
	C4	4.58	0.72	4.48	90.22	

The accuracy of CMP-FCCRF was higher than the two other FC-CRF-based classification methods (SD-FCCRF and FCNN-FCCRF), indicating that the CMP-FCCRF can adaptively incorporate different feature descriptors. In the LULC classification dataset in rural areas of Wuhan (Table 3), the reported quantitative performance of CMP-FCCRF exhibited an improvement in OA. Additionally, the 1.44% higher accuracy (from 83.29% to 84.73%) of CMP-FCCRF compared with CMP-SVM shows that CMP-FCCRF focuses more on spatial contextual information. Compared with FCNN-FCCRF and SD-FCCRF classification methods, the classification accuracy of CMP-FCCRF was better by 4.67% and 1.70%, respectively. Thus, spatial contextual information and other feature descriptors should be considered. Finally, the CMP-FCCRF obtained the highest accuracy.

Even though a rigorous comparison of the results using different datasets and training/test sets could not be performed, the results obtained by applying the proposed classification scheme on a relatively large size validation image, which contained 2361 × 2406 pixels, attest to the effectiveness of the proposed framework.

3.2. Parameter Sensitivity Analysis

The performance of the proposed CMP-FCCRF method was further evaluated using different numbers of training samples. As shown in Figure 3, there are two training steps for the CMP-FCCRF method: fine-tuning of CNN and parameter estimation of SVM. The first step, is described in Section 3.1. Figure 4 was also selected as the validation image. Compared with other methods, different sizes ranging from 100 to 500 to train the parameters of SVM were tested for each LULC class.

As shown in Figure 5, the classification accuracy of CMP-FCCRF initially increased for the data sets with the gradual increase in the number of training samples per class (from 78.9% to 84.73%). The classification accuracy of CMP-FCCRF was slightly higher than CMP-SVM (78.09% and 83.29%,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2018 doi:10.20944/preprints201808.0112.v2

respectively) classification approaches with LULC classification dataset in rural areas of Wuhan city. The classification accuracy of the proposed method remained higher than the other five methods for each training number. The training samples were randomly selected from the overall ground truth, and the remaining samples were used to evaluate the classification accuracies. The experiments showed that the classification accuracies of the methods incorporating spatial contextual information (i.e., SD-FCCRF, FCNN-FCCRF, and the proposed CMP-FCCRF) were all better than SVM-based classification methods. Moreover, the CMP-FCCRF method was more robust than the other classification methods for different training samples.

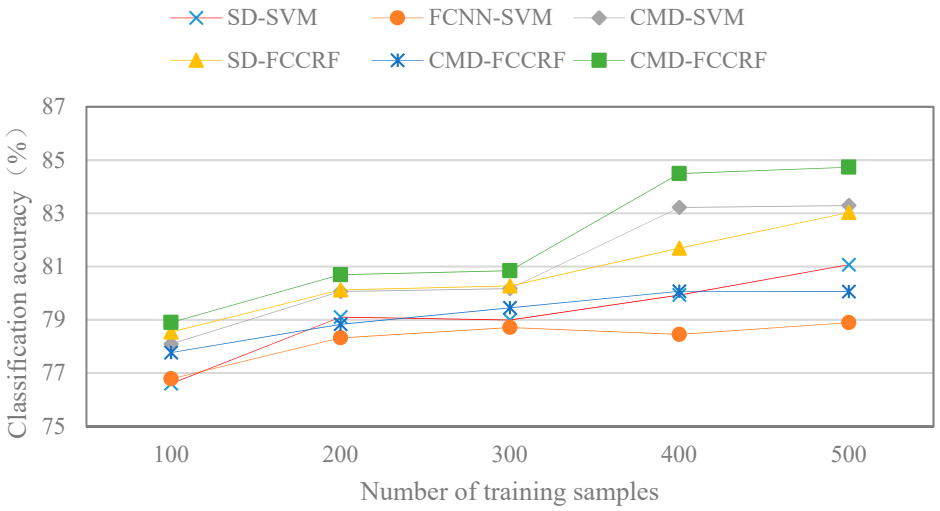


Figure 5. Effect of training data size on the classification results.

3.3. Discussion

As was highlighted in Hu et al. [15] for scene classification, this work demonstrates that the features that are concatenated from the CNN layer and spectral features and NDSM by SVM classifier can effectively perform LULC classification using high spatial resolution remote sensing imagery. We found that simple spectral and NDSM features achieved high accuracy. This is not surprising because the input data were designed to discriminate the target classes: the DSM highlights homestead and trees, and infrared highlights impervious surfaces and background. In order to improve the classification accuracy, it was necessary to incorporate the information produced by DSM.

The FC-CRF model effectively removed the salt-and-pepper effects, which was employed in this work for post-processing. Compared with SVM-based classification methods, FC-CRF methods achieved higher accuracy. Gerke et al. [35] found that CRF smoothing had a negative effect on accuracy, whereas, in our work, the accuracy improved. In contrast to the CRF model, this most likely occurred because our FC-CRF was defined at fully-connected cliques or higher cliques than the CRF model in Gerke et al. [35]. Along with Gerke et al. [35], we conclude that FC-CRF improves the labelling visually by removing speckle from classifier output labels. The FC-CRF-based methods are easy to operate when the unary potentials are calculated by the SVM classifier with some training samples and the pairwise potentials are automatically achieved by taking a mean field approximation of the original CRF with pairwise features.

4. Conclusions

In this paper, to improve LULC classification of high-resolution remote sensing images, a classification framework based on FC-CRF and fine-tuned CNN that takes full advantage of both spectral and spatial information contained within high-resolution remote sensing images was proposed. With the fine-tuned CNN parameters, spectral features, and NDSM processed by a SVM

classifier, new feature descriptors were formed for the FC-CRF models. We demonstrated that CMP-FCCRF is effective for LULC classification and removes the salt-and-pepper effects.

In terms of future research, we plan to investigate potentially more effective classification techniques based on deep learning for LULC classification, which can exploit unlabeled samples. In the high-resolution imagery domain, unlabeled samples are much easier to access than labeled samples. Supervised classification methods based on CNN fail to make full use of these unlabeled samples. Generative adversarial networks (GAN), which consist of both a generative model (G) and a discriminative model (D), could effectively learn image representation in an unsupervised way [36,37]. In the future, integration of unsupervised and supervised learning methods based on CNN for image classification would be desirable to better address this issue.

Author Contributions: All the authors made significant contributions to this work. B.Z. devised the approach and analyzed the data; Y.S. helped design the remote sensing classification experiments and provided advice for the preparation and revision of the work; C.W. performed the experiments; and Y.L. helped to pre-process the image data, including DSM map processing and producing ground truth map.

Acknowledgments: This work was supported by the National Natural Science Foundations of China (Grant No. 41601480).

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

References

1. Nogueira, K.; Penatti ,O.A.B.; Santos, J.A.D. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* **2016**, *61*, 539–556.
2. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Review [PDF]. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36.
3. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. **2014**, *7*, 2094–2107.
4. Nogueira, K.; Miranda, W.O.; Santos, J.A.D. Improving Spatial Feature Representation from Aerial Scenes by Using Convolutional Network. In Proceedings of the 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Bahia, Brazil, 26–29 August 2015; IEEE Computer Society: Washington DC, USA, 2015; 289–296.
5. Yue, J.; Zhao, W.; Mao, S.; Liu, H. Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. Remote Sens. Lett. 2015, 6, 468–477.
6. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens*. **2016**, *55*, doi:10.1109/TGRS.2016.2612821
7. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep Supervised Learning for Hyperspectral data Classification Through Convolutional Neural Networks. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; IEEE Computer Society: Washington DC, USA, 2015; pp. 4959–4962.
8. Volpi, M.; Tuia, D. Dense Semantic Labeling of Subdecimeter Resolution Images with Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 881–893.
9. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *Acta Ecol. Sinica* **2015**, *28*, 627–635.
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2012**, *60*, 1097–1105.
12. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *Proceedings of the British Machine Vision Conference 2014,* Nottingham, UK, September 1-5, 2014; BMVA Press 2014.

13. Xie, M.; Jean, N.; Burke, M.; Ermon, S. Transfer learning from deep features for remote sensing and poverty mapping. Thirtieth AAAI In Proceedings of the Conference on Artificial Intelligence, Phoenix, Arizona, USA, 12–17 February 2016; AAAI Press: Paolo Alto, California, USA, 2016; 3929–3935.
14. Penatti, O.A.B.; Nogueira, K.; Santos, J.A.D. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington DC, USA 2015, 44–51.
15. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens*. **2015**, *7*, 14680–14707, doi:10.3390/rs71114680
16. Li, X.; Shao, G.; Object-Based Land-Cover Mapping with High Resolution Aerial Photography at a County Scale in Midwestern USA. *Remote Sens*. **2014**, *6*, 11372–11390.
17. Lucieer, V.; Object-oriented classification of side-scan sonar data for mapping benthic marine habitats. *Int. J. Remote Sens*. **2008**, *29*, 905–921.
18. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens*. **2010**, *65*, 2–16.
19. Qian, Y.; Ye, M. Hyperspectral imagery restoration using nonlocal spectral-spatial structured sparse representation with noise estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. **2013**, *6*, 499–515.
20. Chen, Y.; Zhao, X.; Jia, X. Spectral-Spatial classification of hyperspectral data based on deep belief network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. **2015**, *8*, 2381–2392.
21. Rother, C.; Kolmogorov, V.; Blake, A. GrabCut: Interactive foreground extraction using iterated graph cuts. In Proccedings of the 31st International conference on computer graphic and interactive techniques SIGGRAPH '04 ACM SIGGRAPH, Los Angeles, CA, USA, 8–12 August 2004; ACM Trans Graphics: New York, NY, USA 2004; Volume 23, pp. 309–314.
22. Shotton, J.; Winn, J.; Rother, C.; Criminisi, A. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vis.* **2009**, *81*, 2–23.
23. Lucchi, A.; Li, Y.; Boix, X.; Smith, K.; Fua, P. Are spatial and global constraints really necessary for segmentation? In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
24. He, X.; Zemel, R.S.; Carreira-Perpindn, M. Multiscale conditional random fields for image labeling. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, Washington, DC, USA, 27 June–2 July 2004.
25. Ladicky, L.; Russell, C.; Kohli, P.; Torr, P.H. Associative hierarchical crfs for object class image segmentation. In Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009.
26. Lempitsky, V.; Vedaldi, A.; Zisserman A. A Pylon Model for Semantic Segmentation. Advances in Neural Information Processing Systems.2011, 1485-1493.
27. Delong, A.; Osokin, A.; Isack, H.N.; Boykov, Y. Fast approximate energy minimization with label costs. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
28. Gonfaus, J.M.; Boix, X.; Van de Weijer, J.; Bagdanov, A.D.; Serrat, J.; Gonzàlez, J. Harmony potentials for joint classification and segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
29. Matikainen, L.; Karila, K. Segment-Based Land Cover Mapping of a Suburban Area—Comparison of High-Resolution Remotely Sensed Datasets Using Classification Trees and Test Field Points. *Remote Sens*. **2011**, *3*, 1777–1804, doi: 10.3390/rs3081777.
30. Chi, M.; Feng, R.; Bruzzone, L. Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. *Adv. Space Res*. **2008**, *41*,1793–1799,doi:10.1016/j.asr.2008.02.012.
31. Huang, M.J.; Shyue, S.W.; Lee, L.H.; Kao, C.C. A Knowledge-based Approach to Urban Feature Classification Using Aerial Imagery with Lidar Data. *Photogramm. Eng. Remote Sens*. **2008**, *74*,1473–1485.
32. Salem, S.A.; Salem, N.M.; Nandi, A.K. Segmentation of retinal blood vessels using a novel clustering algorithm (RACAL) with a partial supervision strategy. *Med. Biol. Eng. Comput*. **2007**, *45*,261–273.

33. Orlando, J.I.; Prokofyeva, E.; Blaschko, M.B. A Discriminatively Trained Fully Connected Conditional Random Field Model for Blood Vessel Segmentation in Fundus Images. *IEEE Trans. Biomed. Eng*. **2016**, *64*, 16–27.
34. Wagner, S.A. SAR ATR by a combination of convolutional neural network and support vector machines. *IEEE Trans. Aerosp. Electron. Syst*. **2018**, 52, 2861–2872.
35. Gerke, M. *Use of the Stair Vision Library Within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*. Technical Report, University of Twente; Researche Gate: Berlin, Germany, 2015.
36. Lin, D.; Fu, K.; Wang, Y.; Xu, G.; Sun, X. Marta gans: Unsupervised representation learning for remote sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 2092–2096.
37. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedinsg of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico,2–4 May 2016.