Abstract: The interpretation of land use/land cover (LULC) is a hotspot and difficult issue in the field of high resolution remote sensing image processing as well as land resource management. Training a new (or existing) Convolutional Neural Networks (CNNs) architecture fully for LULC classification needs a large amount of remote sensing images. Thus, fine-tuning a pre-trained CNNs for LULC is acceptable. To improve the classification accuracy for high resolution remote sensing images, it is necessary to utilize some hand-crafted features and adopt a classifier for post-processing. A Fully Connected Conditional Random Fields (FC-CRFs), to utilize the fine-tuned CNNs layers, hand-crafted features and fully connected pairwise potentials, is proposed for image classification of high resolution remote sensing images. First, an existing CNNs model is adopted, and the parameters of CNNs are fine-tuned by training datasets. Then, the probabilities of image pixels belong to each class type also could be calculated. Second, we consider the hand-crafted features, combined with Support Vector Machine (SVM) classifier, the probabilities belong to each LULC class type are achieved. Combined with the probabilities achieved by fine-tuned CNNs, new feature descriptors are built. Finally, FC-CRFs are introduced to get the classification results, while the unary potentials are achieved by the new feature descriptors and SVM classifier, and the pairwise potentials are achieved by the hand-crafted features. Experimental results show that the proposed classification scheme can achieve impressive performance when the total accuracy reached to about 85%.

Keywords: remote sensing; image classification; Fully connected Conditional Random Fields (FC-CRFs); Convolutional Neural Networks (CNNs)

1. Introduction

Remote sensing has become an important means of obtaining Earth surface information. With the continuous improvement of sensor technology, high-resolution multi-spectral images can be obtained, such as IKONOS, QuickBird, WorldView-2, ZY-3C and GF-1, high resolution remote sensing data, when the spatial resolution can be close to 1m or even sub-meter level. In addition, with the Unmanned Aerial Vehicle (UAV) aerial technology promotion, 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 applications, such as urban environments, precision agriculture, transportation and road facilities, forestry survey, military target identification and disaster assessment. Recent 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 of an image. Due to deep learning methods, especially convolutional neural networks (CNNs), and the availability of large-scale annotated datasets, remarkable progress has been achieved in image classification task. Currently, three major strategies, which concern CNNs, have been successfully adopted for remote sensing image classification: (1) full trained CNNs, (2) fine-tuned CNNs, and (3) pre-trained CNNs used as feature extractors [1, 2].

(1) Training a (new or existing) CNNs fully gives full control of the architecture and parameters, which tends to yield a more robust network. However, this strategy not only needs a lot of computational resources to train CNNs’ parameters, but also needs huge annotated remote sensing data. Although huge annotated remote sensing data are unusual, there are many works, usually using reasonable datasets (more than 2,000 images), that achieved promising results by proposing the full training of new CNNs [3, 4, 5]. Nogueira et al. [3] proposed and fully trained a new CNNs architecture to classify images from aerial and multispectral images. In [4], the authors proposed a hybrid method combining principal component analysis, CNNs and logistic regression to classify hyperspectral image using both spectral and spatial features. Maggiori et al proposed a CNNs architecture that is fully convolutional that only involves a series of convolution and deconvolution operations to produce the output classification maps [5]. Volpi present a CNN-based system relying on a downsample-then-upsample architecture for semantic labeling of subdecimeter resolution images [6]. However, some drawbacks exist in this strategy [6]. For example, [1] and [7] train the networks by only using the existing satellite image dataset, which suffers a drop in classification accuracy compared with using the pre-trained networks as global feature extractors or fine-tuning the pre-trained networks. The reason may lie in the fact that the large-scale networks usually contain millions of parameters to be learned. Thus, training them using the small-scale satellite image datasets will easily cause overfitting and local minimum problems. Consequently, some construct a new smaller network and train it from scratch using satellite images to better fit the satellite images.

(2) Another strategy to exploit CNNs is to fine-tune its parameters using the new data, because the first-layer features that resemble either Gabor filters, edge detectors or color blobs. Specifically, fine-tuning realizes adjustment in the parameters of a pre-trained network by resuming the training of the network from a current setting of parameters but considering a new dataset. In [8], the authors showed that fine-tuning a pre-trained CNNs on the target data can significantly improve the performance. Specifically, they fine-tuned AlexNet [9] and outperformed results for semantic segmentation. Zhao et al. [10] fine-tuned a couple of networks outperforming state-of-the-art results in classification of traditional datasets. Several works [11, 12], in the remote sensing community, also exploit the benefits of fine-tuning pre-trained CNNs. In [12], the authors evaluated a full-trained CNNs against a fine-tuned one to detect poverty using remote sensing data. Yue et al. [11] used fine-tuning method to classify hyperspectral images.

(3) Based on aforementioned characteristics, CNNs can also be exploited as a feature extractor. Specifically, 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 recent studies, ConvNets have shown to perform well even in datasets with different characteristics from the ones they were trained with, 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. [13] evaluated the use of different CNNs as feature extractors, achieving state-of-the-art results in two remote sensing datasets, outperforming several well-known visual descriptors. Hu et al. [14] extracted features of several pre-trained CNNs to perform classification of high-resolution remote sensing imagery.

Though three strategies have their own advantages and disadvantages, to make features more discriminative, choosing the second strategy, which is fine-training a existing CNNs, is more suitable. Remote sensing datasets for image classification to train CNNs are available, such as the ISPRS 2D semantic labeling datasets, the Zurich Summer Dataset, et.al. While CNNs are commonly considered as a well-performing and promising solution for image classification, the problem of segmentation can have multiple solutions and thus 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 of misclassified pixels \[15\] and has had difficulty with dealing with the rich information from very high-resolution data \[16, 17\]. Works that include more spatial information in the neighborhood of the pixel to be classified have been done \[18, 19\]. Moreover, Conditional Random Fields (CRFs) have been broadly used in semantic segmentation to combine class scores computed by multi-way classifiers with the low level information captured by the local interactions of pixels and edges \[20], \[21\] or superpixels \[22\]. Even though works of increased sophistication have been proposed to model the hierarchical dependency \[23], \[24], \[25\] and/or high order dependencies of segments \[26], \[27\], we use the fully connected pairwise CRF proposed by \[19\] for its efficient computation, and ability to capture fine edge details while also catering for long range dependencies.

Thus, we propose a new classification architecture which combines fine-tunes CNNs, handcrafted features with SVM, and fully connected CRF (FCNMF-FCCRFs). First, inspired by the down-sample and up-sample of CNNs architecture \[6\], we fine-tune CNNs parameters to form probability features. Then the hand-crafted features are used to form another probability feature by SVM classifier. Cascading these two type probabilities, the feature descriptor of each pixel in entire image is achieved. Finally, SVM classifier is adopted to form the unary potentials of FC-CRFs, and pairwise potentials are formed by the hand-crafted features.

The main contributions of this paper are summarized as follows:
- We used true UAV multispectral imagery with a spatial resolution of 0.4 m along with a digital surface model (DSM) of the area for LULC classification.
- We developed a novel approach for high-resolution remote sensing imagery per-pixel classification of four classes (building, ground, tree, farmland/water) using fine-tuned CNNs, hand-crafted features and FC-CRFs, that outperform 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, that is 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.

<table>
<thead>
<tr>
<th>Table 1. List of Some Important Items and Corresponding Abbreviations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LULC</td>
</tr>
<tr>
<td>DSM</td>
</tr>
<tr>
<td>UAV</td>
</tr>
<tr>
<td>NDSM</td>
</tr>
<tr>
<td>CNNs</td>
</tr>
<tr>
<td>SVM</td>
</tr>
<tr>
<td>FC-CRFs</td>
</tr>
<tr>
<td>Overall Accuracy</td>
</tr>
</tbody>
</table>

2. CNNs and Fully connected CRFs

The data and the process of manual labeling of the data is described in are described in Section 2.1. An introduction to CNNs is given in Section 2.2, and the process of cascading probabilities for feature descriptors is given in Section 2.3. An introduction to FC-CRFs model and the flow chart of proposed method is given in Section 2.4 and 2.5, respectively.

2.1 Data and Manual Labeling

The data that are used in this work is a rural area map whose location is in the rural area of Wuhan city, China and consist of several north-oriented multispectral true orthography bands,
several synthetic image bands, two near-infrared 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 2D image mosaic that represents a true nadir rendering. Moreover, the true
ortho imagery is 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 building and dark-colored ground. Furthermore, a DSM is invariant to lighting and color
variations and can give a better geometry estimation and background separation [28]. 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 [29]. Most datasets that use classification on
a pixel-level only use a few hundred reference points [28,30].

The Vaihingen datasets for the 2-D semantic labeling contest is organized by the International
Society for Photogrammetry and Remote Sensing (ISPRS) Commission III.4. There are six class types
in the Vaihingen datasets, which are named buildings, cars, low vegetation, trees, imperious surfaces,
and background. Compared with the Vaihingen datasets, there are four class types in the LULC
classification dataset in rural areas of Wuhan city, which are named homestead, imperious surfaces,
background which contains farmland, low vegetation and water, trees. Each map in Figure 1 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 is manually labeled. Figure 1a
shows the high-resolution remote sensing image that is used to manually classify the map, and Figure
1c shows the finished labeled map.

![Figure 1](image1.png)

Figure 1 The LULC classification datasets in rural areas of Wuhan city (a) High-resolution remote
sensing image (b) NDSM (c) ground truth (C1 Red: imperious surfaces; C2 Green: homestead; C3
Blue: background; C4 Yellow: trees)
2.2 CNNs architecture

The CNNs are composed by a sequential hierarchy of processing layers in Figure 2. From the input to the final classification layer, data go through a series of trainable units. CNNs are structured in a similar way, but neurons are learnable convolutions shared at each image location. The architecture contains down-sampling, up-sampling and prediction layers, which is described in detail in [6]. Furthermore, the topmost layer, which belongs to prediction layer, is composed by a classifier. We use the commonly employed multinomial logistic regression, whose scores (class-conditional probabilities) are given by the softmax function:

$$ p(y_i | x_i) = \frac{\exp(x_{i,C})}{\sum_{c=1}^{C} \exp(x_{i,c})} \quad (1) $$

for C classes. Inputs $x_i$ are a C-dimensional vector representing unnormalized scores for the location $i$, as given by the penultimate layer (the one before the loss function).

2.3 Feature descriptors

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

$$ \mathbf{P}^{\text{FCNN}} = [p_1^{\text{FCNN}}, \ldots, p_N^{\text{FCNN}}]^T \quad (2) $$

$N$ is the number of LULC class type.

The hand-crafted features, which are spectral features and normalized DSM, combined with training samples are used to estimate the SVM classifier parameters and the probability of vectors belonging to each LULC class, could be expressed as:

$$ \mathbf{P}^{\text{HC}} = [p_1^{\text{HC}}, \ldots, p_N^{\text{HC}}]^T \quad (3) $$

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

$$ \mathbf{P}^{\text{MF}} = [\mathbf{P}^{\text{FCNN}}]^T, \mathbf{P}^{\text{HC}}]^T \quad (4) $$

The probabilities of Equation (2) and Equation (3) are cascaded to form a new feature descriptor.

2.4 Fully connected CRFs

CRFs have been often employed to smooth noisy segmentation maps [23], [31]. Typically these models couple neighboring nodes, favoring same-label assignments to spatially proximal pixels.

Qualitatively, the primary function of these short-range CRFs is to clean up the spurious predictions of weak classifiers built on top of local hand-engineered features.

Using contrast sensitive potentials [23] in conjunction to local-range CRFs can potentially improve localization but still miss thin structures and typically requires solving an expensive discrete optimization problem. To overcome these limitations of short-range CRFs, we integrate into our system the fully connected CRF (FC-CRFs) model of [22]. To overcome these limitations of short-range CRF, we integrate into our system the FC-CRFs model of [22]. 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 use as unary potential $\theta_i(x_i) = -\log(p(x_i))$, where $p(x_i)$ is the label assignment probability at pixel $i$ as computed by a CNNs.

In a FC-CRFs model, each node of the graph is assumed to be linked to every other pixel of the image. Using these higher order potentials, the method is able to take into account not only neigh-

---

**Figure 2 The structure of dense CNNs**
boring 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 [22], we use the following expression:

\[
\theta_{\theta}(x_i, x_j) = \omega u(x_i, x_j) \left[ \exp \left( -\frac{||p_i - p_j||^2}{2\delta^2} - \frac{||I_i - I_j||^2}{2\delta^2} \right) + \exp \left( -\frac{||p_i - p_j||^2}{2\delta^2} \right) \right]
\]

(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 \( p \)) and RGB color (denoted as \( I \)), and the second kernel only depends on pixel positions. The hyper parameters \( \delta^2 \) and \( \delta^2 \) control the scale of Gaussian kernels. The first kernel forces pixels with similar color and position to have similar labels, while these second kernel only considers spatial proximity when enforcing smoothness. Recently, an efficient inference approach is introduced under the restriction that the pairwise potentials are a linear combination of Gaussian kernels over an Euclidean feature space [32,33]. 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.

2.5 The flow chart of proposed method

An overview of the method used can be seen in Figure 3. The pre-trained CNNs parameters, which are pre-trained by the Vaihingen dataset, combined with the LULC classification dataset in rural areas of Wuhan city, are fine-tuned. The remote sensing images and normalized DSM with training samples are adopted to achieved SVM parameters. Then, the probabilities of testing image in Equation (2) and Equation (3) could be achieved. The feature descriptors in Equation (4) are extracted by cascading these probabilities. Finally, by introducing FC-CRFs models, the classification results are got.

![Figure 3 The flow chart of proposed method](image)

3. Results

We conduct experiments using the high-resolution aerial images to evaluate the effectiveness of the proposed FCNNMF-FCRFs framework for LULC classification. Based on the study of Nogueira’s work [1], comparative experiments are conducted by combining feature descriptors and classification methods. We compared the different methods using confusion matrix and overall accuracy. The Fine-tuned CNNs feature, hand-crafted feature and classifier associated with FCNN-SVM, HC-SVM, FCNNMF-SVM, FCNN-FCRFs, HC-FCRFs and FCNNMF-FCRFs, are reported in Table 2. The details are described as follows.

(1) FCNN-SVM. This method uses only the fine-tuned CNNs probabilities as feature. After CNNS parameters fine-tuned, SVM method is adapted to achieve classification results [33].
(2). HC-SVM. This method 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.

(3). FCNNMF-SVM. The cascading feature of fine-tuned CNNs probabilities and HC-SVM probabilities are considered as the feature descriptors in this method, and SVM classifier is used for LULC classification.

(4). FCNN-FCCRFs. Fine-tuned CNNs probabilities is used for the feature vector in this method, and FC-CRFs is adopted for LULC classification.

(5). HC-FCCRFs. Spectral features and NDSM are considered as the feature descriptors in this method, which is combined with FC-CRFs to achieve the classification results.

(6). FCNNMF-FCCRFs. The cascading feature of fine-tuned CNNs probabilities and HC-SVM probabilities are considered as the feature descriptors in this method, and FC-CRFs classifier is used for LULC classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature descriptors</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCNN-SVM</td>
<td>Fine-tuned CNNs (Equation (2))</td>
<td>SVM</td>
</tr>
<tr>
<td>HC-SVM</td>
<td>Spectral feature and NDSM combined with SVM (Equation (3))</td>
<td>SVM</td>
</tr>
<tr>
<td>FCNNMF-SVM</td>
<td>Cascading probability vector (Equation (4))</td>
<td>SVM</td>
</tr>
<tr>
<td>FCNN-FCCRFs</td>
<td>Fine-tuned CNNs (Equation (2))</td>
<td>FC-CRFs</td>
</tr>
<tr>
<td>HC-FCCRFs</td>
<td>Spectral feature and NDSM combined with SVM (Equation (3))</td>
<td>FC-CRFs</td>
</tr>
<tr>
<td>FCNNMF-FCCRFs</td>
<td>Cascading probability vector (Equation (4))</td>
<td>FC-CRFs</td>
</tr>
</tbody>
</table>

3.1 Experimental Results and Discussion

The proposed method is evaluated on the LULC classification datasets in rural areas of Wuhan city, as shown in Figure 1. We subset six images to fine-tune the CNNs parameters. The CNNs parameters are pre-trained by Vaihingen datasets [6]. When we fine-tune the CNNs parameters, the learning rates are initialized at 10^{-5} and kept uniform for 20 epochs. The batch size on the LULC classification datasets in rural areas of Wuhan city is 128. The codes are running on the computer with Intel(R) Core(TM) i7-6500U CPU @ 2.50 GHz 2.50GHz, NVIDIA Quadro M500M (2GB), 16 GB RAM, 512GB SSD and Matlab 2016a. The gradient is computed via batch gradient descent, which is not computed by GPU.

The testing image as shown in Figure 4a, contains 2361×2406 pixels. The training samples of SVM classifier for each method is 500. In Equation (6), the parameter \( \omega \) for FCNN-FCCRFs, FCNNMF-FCCRFs, and HC-FCCRFs methods, which is adopted to balance the unary potentials and pairwise potentials, is 5. The effects of classification results using different methods can be seen in Figure 4. In Figure 4d, 4e and 4f, it can be seen that there are some salt-and-pepper misclassifications, imperious surfaces being classified as homestead, trees being classified as homestead and patches of imperious ground in the middle of the homestead. Furthermore, the shadows from the homestead cause the imperious ground to be misclassified as homestead. In Figure 4g, 4h and 4i, it shows the result after FC-CRFs smoothing, and it can be seen that the salt-and-pepper effects have been removed.

Table 3 shows the confusion matrix for the different classification methods due to combination of different feature descriptors and classifier. It can be seen that the large confusion is between homestead that was classified as imperious surfaces. We believe this is caused by the large amount of shadows that exist on many of the roads from the surrounding forest and buildings, which is known to complicate the interpretation of areas nearby such objects [28]. There is also a large
confusion between background and imperious surfaces. This is because these two class types may have similar spectral feature. We believe this occurs when buildings are located close to dense forest areas. The results show that the algorithms in which spatial contextual information are considered significantly outperformed the SVM classification in classification accuracy. Moreover, the accuracy of FCNNMF-FCCRFs is higher than the three other FC-CRFs-based classification methods (i.e., HC-FCCRFs, and FCNN-FCCRFs), indicating that the FCNNMF-FCCRFs can adaptively incorporate different feature descriptors. In the LULC classification dataset in rural areas of Wuhan city (Table 4), the reported quantitative performance of FCNNMF-FCCRFs exhibits the improvement in OA. Additionally, the 1.44% higher accuracy (from 83.29% to 84.73%) of FCNNMF-FCCRFs compared with FCNNMF-SVM shows that FCNNMF-FCCRFs focuses more on spatial contextual information. Furthermore, compared with FCNN-FCCRFs and HC-FCCRFs classification methods, the classification accuracy of FCNNMF-FCCRFs improved by 4.67% and 1.70%. Thus, spatial contextual information and other hand-crafted feature descriptors should be considered. Finally, the FCNNMF-FCCRFs obtains the highest accuracy.

Figure 4 Classification results of LULC classification dataset in rural areas of Wuhan city (C1 Red: imperious surfaces; C2 Green: homestead; C3 Blue: background; C4 Yellow: trees)
The performance of the proposed FCNNMF-FCCRFs method is further evaluated using different numbers of training samples. Different sizes ranging from 100 to 500 are tested for each LULC class. As shown in Figure 5, the classification accuracy of FCNNMF-FCCRFs initially increases for the data sets with gradual increase in the number of training samples per class (from 78.9% to 84.73%). The classification accuracy of FCNNMF-FCCRFs is slightly higher than FCNNMF-SVM (from 78.09% to 83.29%) classification approaches with LULC classification dataset in rural areas of Wuhan city. The classification accuracy of the proposed method remains higher than the other five methods at each training number. The training samples are randomly selected from the overall ground truth, and the remaining samples are used to evaluate the classification accuracies. The experiments show that the classification accuracies of the methods incorporating spatial contextual information (i.e., HC-FCCRFs, FCNN-FCCRFs and the proposed FCNNMF-FCCRFs) are all better than SVM-based classification methods. Moreover, FCNNMF-FCCRFs method is more robust than the other classification methods with different training samples.
3.3. Discussion

According to the experimental results presented above, we can draw the following conclusions.

First, FC-CNNs-based methods obtained better classification results than SVM-based methods. Although FCNNMF-SVM has achieved higher classification accuracy than two other SVM-based method, the salt-and-pepper effects phenomenon still exists. Furthermore, the FC-CNNs-based methods are easy to operate when the unary potentials are calculated by 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.

Second, the feature descriptors of FCNNMF-FCCRFs method which cascades the fine-tuned CNNs probabilities and the probabilities calculated by hand-crafted features with SVM, are more discriminable. Compared with FCNN-FCCRFs and HC- FCCRFs, FCNNMF- FCCRFs method has achieved better classification results. Moreover, the hand-crafted features, especially the NDSM, which is effect for homestead (or building) extraction, and fine-tuned CNNs features are as important when the training datasets is not large enough. In order to improve the classification accuracy, it is necessary to make use of these two types feature descriptors.

Finally, compared with HC-SVM, FCNN-SVM, FCNNMF-SVM, HC- FCCRFs and FCNN- FCCRFs, the FCNNMF-FCCRFs model achieved a better performance on LULC classification dataset in rural areas of Wuhan city. The six different classification methods have all adopted SVM classifier, when the FC-CRFs methods use SVM classifier to achieve unary potentials. The classification accuracy of FCNNMF-FCCRFs is always higher than five other classification methods with different training samples. Thus, the results show that FCNNMF-FCCRFs method is more robust than the other classification methods.

4. Conclusions

In this paper, to improve LULC classification of high-resolution remote sensing images, a classification framework based on FC-CRFs and fine-tuned CNNs has been proposed that takes full advantage of both spectral and spatial information contained within high-resolution remote sensing images. With the fine-tuned CNNs parameters, and hand-crafted features processed by SVM classifier, new feature descriptors are formed for the FC-CRFs models. It has been shown that FCNNMF-FCCRFs method is effective for LULC classification and removes the salt-and-pepper effects.

In terms of future research, we plan to investigate more effective CNNs-based LULC classification techniques with transfer learning method. Usually, training CNNs architecture fully for remote sensing application task needs a huge labeled remote sensing data. We will develop a
CNNs architecture based on transfer learning method to enhance the classification accuracy with less training samples.

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

Author Contributions: All the authors made significant contributions to this work. Bin Zhang devised the approach and analyzed the data; Yonglin Shen helped design the remote sensing classification experiments and provided advice for the preparation and revision of the work; Cunpeng Wang performed the experiments; Yueyan Liu helped to pre-process the image data, including DSM map processing and producing ground truth map.

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

References


