remote sensing

MDPI

Article

Scene Semantic Understanding Based on the Spatial Context Relations of Multiple Objects

Yanfei Zhong [1,2,*], **Siqi Wu** [1,2,*] **and Bei Zhao** [3]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[2] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
[3] Department of Geography &Resource Management, The Chinese University of Hong Kong, Hong Kong, China; zhaoys@whu.edu.cn
* Correspondence: zhongyanfei@whu.edu.cn (Y.Z.); qhywsqnjnu@163.com (S.W.); Tel.: +86-27-68779969 (Y.Z.)

Received: 18 August 2017; Accepted: 28 September 2017; Published: 9 October 2017

Abstract: As a result of the large semantic gap between the low-level features and the high-level semantics, scene understanding is a challenging task for high satellite resolution images. To achieve scene understanding, we need to know the contents of the scene. However, most of the existing scene classification methods, such as the bag-of-visual-words model (BoVW), feature coding, topic models, and neural networks, can only classify the scene while ignoring the components and the semantic and spatial relations between these components. Therefore, in this paper, a bottom-up scene understanding framework based on the multi-object spatial context relationship model (MOSCRF) is proposed to combine the co-occurrence relations and position relations at the object level. In MOSCRF, the co-occurrence relation features are modeled by the fisher kernel coding of objects (oFK), while the position relation features are represented by the multi-object force histogram (MOFH). The MOFH is the evolution of the force histogram between pairwise objects. The MOFH not only has the property of being invariant to rotation and mirroring, but also acquires the spatial distribution of the scene by calculating the acting force between multiple land-cover objects. Due to the utilization of the prior knowledge of the objects' information, MOSCRF can explain the objects and their relations to allow understanding of the scene. The experiments confirm that the proposed MOSCRF can reflect the layout mode of the scene both semantically and spatially, with a higher precision than the traditional methods.

Keywords: scene understanding; object-oriented classification; co-occurrence relations; position relations; multi-object force histogram

1. Introduction

With the development of high resolution satellites, the spatial resolution of remote sensing images has been better than 0.5 m. Compared to mid-low resolution satellite images, high resolution satellite images (HRSIs) can obtain more details of clearer ground objects. To take full advantage of this rich information, scene understanding of higher level is needed. However, it is difficult to cross the chasm between the low-level features and the high-level scene semantics because of the diversity of the objects, the variability of the low-level features, and the complex spatial layouts [1,2].

To bridge this semantic chasm, scene classification methods based on mid-level features have been proposed, including the bag-of-visual-words model (BoVW), feature coding, topic models, and some deep learning models. Among these methods, traditional BoVW [3,4], feature coding [5,6], and topic models [7,8] treat the image as a set of local features called visual words, and then describe the scene according to the coding of the visual dictionary formed by visual words or the distribution of topics. However, visual words or topics are modeled based on pixels, which ignores the information of objects and is not helpful to understand the internal compositions of the scene. As for deep learning, the end-to-end learning method only tells the scene category of the image, instead of what constitute the scene. Though feature expressed by the final layer of the network can express some information, it is too abstract to understand [9,10].

Compared to scene classification, scene understanding based on objects concerns more about recognizing the objects and describing the relations of the objects. The scene categories are obtained based on the relations of objects [11–13]. Object recognition has evolved from pixel-based classification based on single features to object-oriented classification based on multiple features [14,15]. To construct the relations of objects more precisely, the objects should be crisp objects with continuous boundaries. Therefore, object-oriented classification is more suitable. Object-oriented classification usually involves segmenting the images into meaningful homogeneous regions and then classifying these regions into different land-cover types. The relations of objects can be summarized into visual context and semantic context, where visual context is the low-level association and semantic context is the high-level association, fusing the prior knowledge of the objects. Semantic context includes spatial context and scene context [16]. There are two types of spatial context relations: one is the co-occurrence relations, meaning the categories of objects being relevant (e.g., if water is the main part of the scene, buildings will be less likely to appear); and the other is the position relations, meaning that the distribution of the objects follows certain rules, such as trees being on both sides of a road. Co-occurrence relations and position relations are complementary, because position is essential in distinguishing scene categories with objects of similar frequencies but different spatial distributions, and co-occurrence excludes those objects with a similar distribution but totally different number.

Co-occurrence relations can be described by mixture of topics modeling by latent Dirichlet allocation (LDA) [17], concept occurrence vector modeling by the proportion of object patches [18], and object bank representation by the use of a set of filters to calculate the object responses [19]. However, these methods express the co-occurrence relations with clusters or patches of features extracted from pixels, instead of the real geo-objects, or they need a feature library. Position relations include distance relations, direction relations, and topology relations. Those methods using the basic geometric features, such as the ratio of the perimeter, ratio of the area, azimuth, and moment invariants, model the topology relations, the distance relations, and the direction relations of the pairwise objects separately [20,21]. Compared to methods based on geometric features, the histogram of force (*F*-histogram) is sensitive to size, distance, shape, and direction, indirectly uniting the three types of position relations by calculating the force between two objects. The *F*-histogram is also more convenient to obtain because it does not require the calculation of the boundary perimeter of the objects [22–27]. However, the *F*-histogram is designed to acquire the position relations between pairwise objects [28], and it is not suitable for modeling multiple objects in remote sensing images. In image indexing and retrieval, the *F*-histogram has been extended to describe a group of objects located near the image center [24,25]. However, this method needs two reference objects placed outside the circumcircle of the group of objects.

In this paper, to solve the problem of scene understanding, a bottom-up scene understanding framework based on the multi-object spatial context relationship model (MOSCRF) is proposed to bridge the semantic gap between pixels and the high-level semantics of HRSIs scene understanding. MOSCRF abides by a bottom-up sequence of pixels-objects-scenes, making it easy to parse the image hierarchically. In MOSCRF, the scene understanding includes three parts: (1) object-oriented classification; (2) construction of the co-occurrence relations and position relations; and (3) scene sematic category understanding.

The contributions of this paper are as follows:

(1) MOSCRF is used to understand the scene components and their co-occurrence relations and position relations. When determining the scene, MOSCRF takes advantage of the complementary nature of the fisher kernel coding of objects (oFK) and the multi-object force histogram (MOFH) to dissect the scene from two different aspects. The oFK is concerned more about the co-occurrence

relations of different objects, whereas the MOFH pays more attention to the spatial distribution of the scene.

(2) The oFK is used to model the co-occurrence relations by introducing the fisher kernel coding to objects. Compared to the traditional methods, the oFK is a compact, low-dimensional and refined representation of the distribution of objects categories by using a gradient vector.

(3) The MOFH is used to model the position relations for multiple objects. Aimed at dealing with the multiple objects in HRSIs, the MOFH adds a global scan line strategy and a rollback strategy to the traditional *F*-histogram. To keep the invariance to rotation and mirroring, the initial direction is defined as the centroid line between the object with the biggest area and the object with the smallest area. Finally, the feature is the mean and standard deviation of the MOFH curve. The MOFH explains the interaction of the internal objects in different directions.

The remainder of this paper is organized as follows. Section 2 introduces the basic theory of the fisher kernel coding (FK) and *F*-histogram. Section 3 describes the MOSCRF construction procedure. The experimental results are reported in Section 4, followed by the sensitivity analyses in Section 5. Finally, the conclusions are provided in Section 6.

2. Basic Theory

2.1. Fisher Kernel Coding

Fisher kernel (FK) was first proposed to model the generation process of the signal with a gradient vector derived from a probability density function (pdf) and later introduced by Perronnin and Dance to image classification as the extensions of BoVW to overcome the diversity of the low-level features and the complexity of the distribution [6,29]. The main idea is as follows:

Let $X = \{x_i\}_{i=1}^{n}$ be the n local features extracted from the image, which can be described by the gradient vector G_λ^X. $p(X|\lambda)$ is the pdf and λ is the parameters (Equation (1)). To make the classifier more efficient, Fisher information matrix (Equation (2)) is used to normalize the gradient vector (Equation (3)).

$$G_\lambda^X = \nabla_\lambda \log p(X|\lambda) \tag{1}$$

$$F_\lambda = E_X\left[G_\lambda^X \left(G_\lambda^X\right)'\right] \tag{2}$$

$$g_\lambda^X = F_\lambda^{-\frac{1}{2}} G_\lambda^X \tag{3}$$

Finally, the normalized vector is the representation of the image and can be the input of the classifier to recognize the image.

2.2. F-Histogram

The *F*-histogram, developed from the histogram of angles and first proposed by Matsakis in 1999 [28], is an effective way to build direction relationships between a pair of objects [30]. The *F*-histogram treats the image as a set of longitudinal sections instead of points, leading to rapid computation (Figure 1).

In Figure 1, A and B are two objects, and A is to the left of B. It is assumed that it is the scalar resultant of the elementary forces that forms the present relative position. The force between two homogeneous objects can be obtained by the sum of the force generated by all the longitudinal sections in the interior. The force in direction θ is calculated by the integral of the secant lines with Equation (4) (Figure 2). In Equation (4), F_{AB} represents the force between the object pair; w and u are the point positions in the secant line; x, y, z is the secant line; and θ is the direction. According to Equation (4), if there are more points involved in the action, the force will be larger in direction θ. Therefore, the *F*-histogram directly reflects the direction relations, while reflecting the distance, topology, size, and shape information of the objects. When scanning the object pair between 0 and 360 degrees, the force distribution in all directions can be acquired and expressed as the histogram.

$$F_{AB}(\theta) = \int_{y+z}^{x+y+z} \left(\int_0^z \varphi(u-w)dw \right) du \tag{4}$$

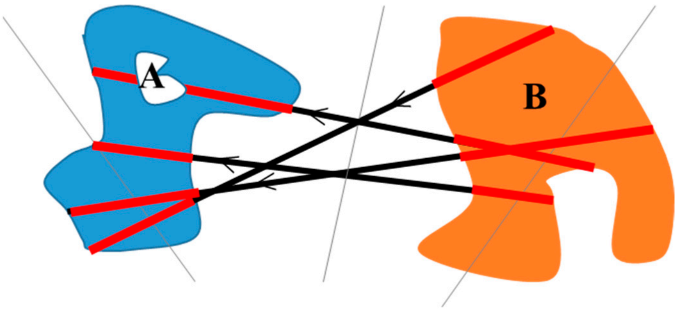


Figure 1. Expression of the histogram of force. The arrow means that position *A* is relative to position *B*, and the red line represents the secant line of the object.

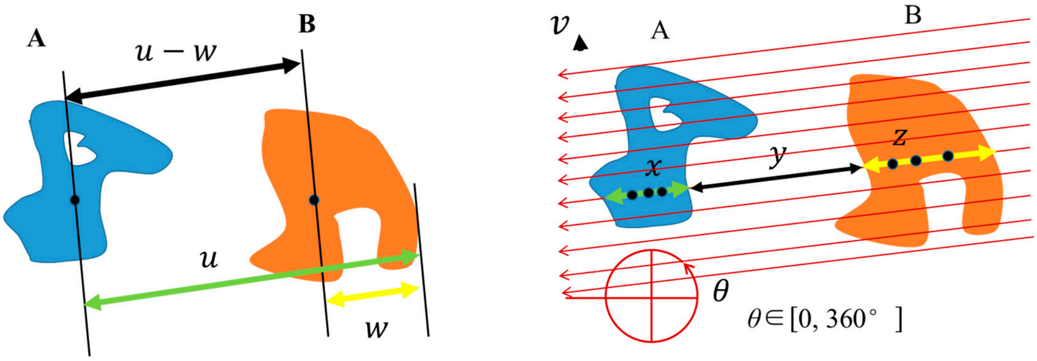


Figure 2. Calculation of the histogram of force. The histogram in one direction is calculated by Equation (1). We then sum the force in all the scan lines with the same direction, and traverse to acquire all the directions.

The *F*-histogram of a pair of objects has the property of invariance to rotation, mirroring, and scale. It can therefore be introduced into remote sensing images as a method of building the spatial relationship between a pair of homogenous objects.

3. Scene Understanding Framework Based on the Multi-Object Spatial Context Relationship Model

To solve the problem of scene understanding, a scene understanding framework based on the multi-object spatial context relationship model is proposed to understand scenes according to the objects and their spatial context relations. MOSCRF is modeled under the following three steps: (1) scene component understanding by object-oriented classification; (2) components' relations understanding by the *atf*-*idf* and the MOFH; and (3) scene sematic category understanding. The flowchart of scene understanding based on MOSCRF is shown in Figure 3, and is described as follows.

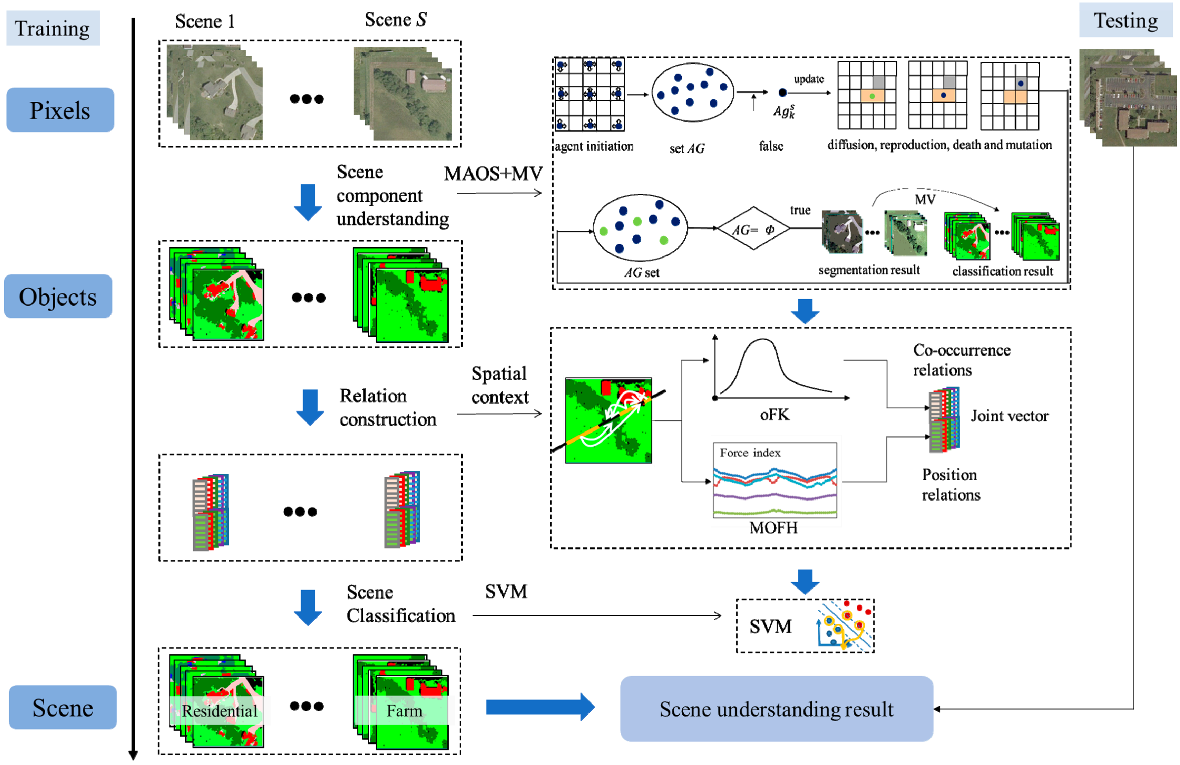


Figure 3. The flowchart of scene understanding based on MOSCRF. Multi-agent object-based segmentation (MAOS) and majority voting (MV) is used to realize the object recognition. *AG* means the agent. oFK and MOFH is used to construct the spatial context of objects. Finally, the scene classification is recognized based on support vector machine (SVM).

3.1. Scene Component Understanding by Object-Oriented Classification

The object-oriented classification accuracy directly influences the authenticity of the scene understanding. Object-oriented classification consists of segmentation and classification. Compared to using a classic segmentation algorithm such as the mean shift algorithm [31], the fractal net evolution approach (FNEA) algorithm [32], or the split-and-merge algorithm [33], and classification algorithms such as KNN, support vector machine (SVM) [14,34], or deep learning methods [35,36], the multi-agent object-based segmentation (MAOS) algorithm can achieve a better result [37,38] by taking advantage of the strong interaction, high flexibility, and parallel global control capability of the multiple agents. The segmentation result and majority voting (MV) are then used to constrain the classification result according to the spectral feature and texture feature. Consequently, the change from the pixel level to the object level is achieved.

3.2. Components' Relations Understanding

3.2.1. Co-Occurrence Relations Based on the oFK

After object-oriented classification, the relations between different objects can be mined. Co-occurrence relations can reduce the conflict caused by ambiguous objects. Compared to traditional fisher kernel coding in scene classification, the biggest differences of oFK are the following two points: the local features are extracted from the object-oriented classification and the local features are the mean and standard deviation of the object categories. Figure 4 shows the process flow of the co-occurrence relations.

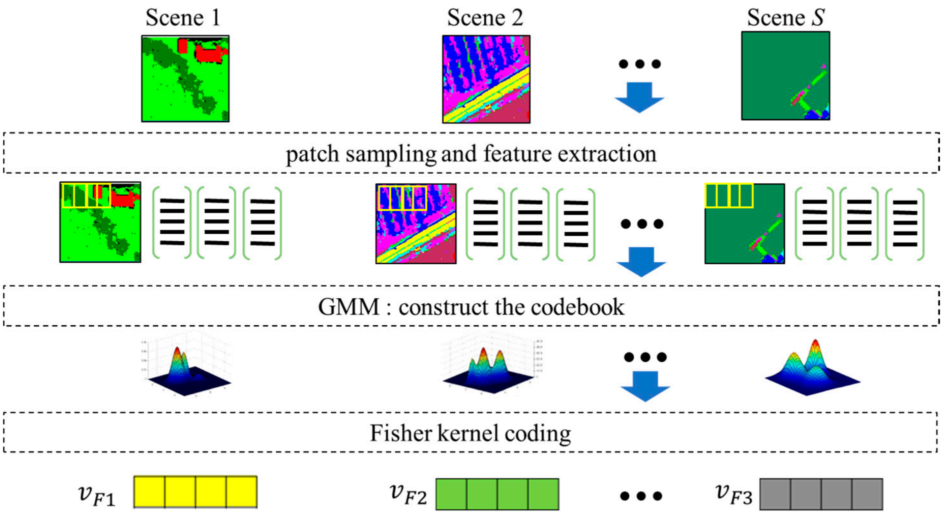

Scene 1
Scene 2
Scene S
patch sampling and feature extraction
GMM : construct the codebook
Fisher kernel coding
v_{F1}
v_{F2}
v_{F3}

Figure 4. The construction of the co-occurrence relations.

The input of the co-occurrence relation construction process is the land-cover classification image. Uniform patch sampling is used to obtain the mean and standard deviation of each patch, and the extracted features are represented as $X = \left\{x_i^m, x_i^{std}\right\}_{i=1}^{n}$, where n is the number of patches, m is the mean and std is the standard deviation. Let Equation (5) be the likelihood function of the pdf of the input image, $p(X|\lambda)$ be the pdf and λ be the parameters. Under the assumption that the feature is independent, $p(x_i|\lambda)$ can be represented by some models generated by Gaussian Mixture Model (GMM). Expectations-Maximum algorithm (EM) is used to learn the parameters and the codebook is then constructed.

$$\varsigma(X|\lambda) = logp(X|\lambda) \tag{5}$$

Finally, the co-occurrence vector v_F is derived based on the codebook according to fisher kernel coding theory.

3.2.2. Position Relations Based on the MOFH

The MOFH is designed to acquire the position relations among multiple objects. Compared to the F-histogram of a pair of objects, there are three main differences. The first is that each scan line crosses several objects, and the MOFH uses a rollback strategy to calculate the force between it and the other objects. The second is that the initial direction is defined as the centroid line between the object with the biggest area and the object with the smallest area. The final difference is that the mean and standard deviation of the F-histogram are used as the position feature. The calculation process of the MOFH is described in Figure 5.

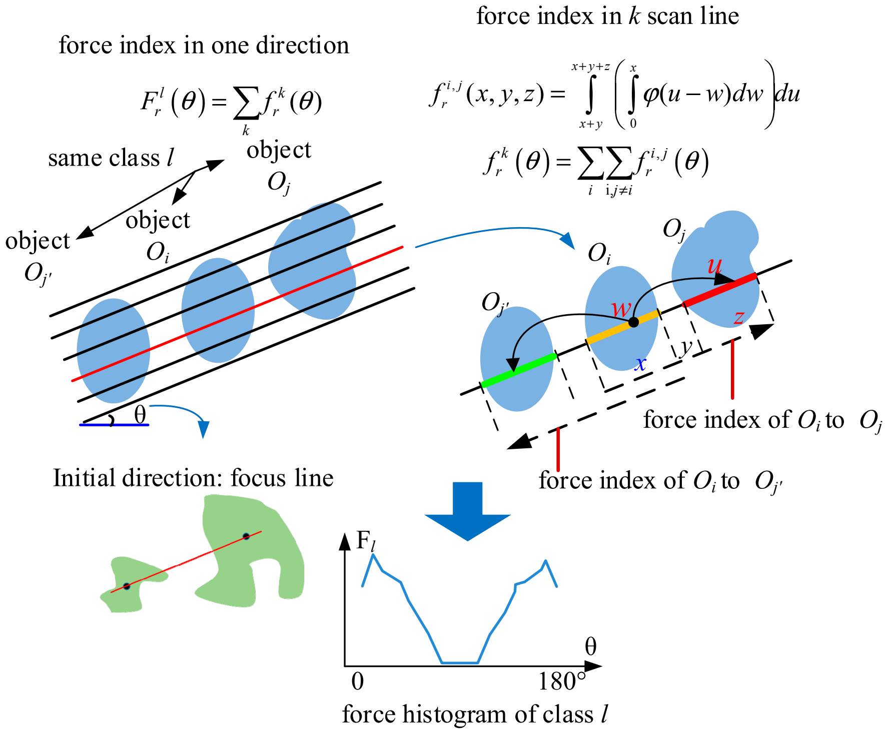
The input image is first translated into a binary image consisting of the target objects $\{O_i\}_i^N$ belonging to category l and the background objects, where N means the number of objects in category l. The task is to calculate the relative position between the target objects and background objects. We let θ be the current direction of the K scan lines, which can be generated by the Bresenham algorithm. The objects O_i and O_j are crossed by the scan line k with secant lengths x and z, respectively, and the distance between x and z is y. The force index $f_r(\theta)$ between O_i and O_j is then calculated by Equation (1), and $\varphi_r(d)$ is the force function (Equation (6)). In Equation (6), the different values of r represent different meanings. When $0 < r < 1$, the F-histogram can deal with any spatial relations. When $1 < r < 2$, the F-histogram reflects the disjoint or tangential objects. When $r \geq 2$, the F-histogram can only describe the disjoint relationships.

$$\varphi_r(d) = \frac{1}{d^r} \tag{6}$$

When using the *F*-histogram to calculate the position relations in a traditional environment, the next step is to use Equation (4) for every pair of objects separately. In contrast, in the MOFH, the scan line crosses the image, i.e., multiple objects are considered. To traverse all the objects, forward and backward actions are used. The force index of scan line *k* is then obtained by Equation (7):

$$f_r^k(\theta) = \sum_i \sum_{i,j \neq i} f_r(\theta) \tag{7}$$

Considering Equations (4), (6) and (7), when the sizes of the objects remain constant, $f_r^k(\theta)$ becomes larger with the increase of the degree of dispersion. When the distance is fixed, $f_r^k(\theta)$ increases as the sizes broaden. Therefore, the force index is proportional to the size, frequency, and compactness of the objects. In other words, $f_r^k(\theta)$ records the scale of the objects and the degree of dispersion of object *l* along this scan line.



force index in one direction
$F_r^l(\theta)=\sum_k f_r^k(\theta)$
force index in k scan line
$f_r^{i,j}(x,y,z)=\int_{x+y}^{x+y+z}\left(\int_0^x \varphi(u-w)dw\right)du$
$f_r^k(\theta)=\sum_i\sum_{i,j\neq i} f_r^{i,j}(\theta)$
same class l
object O_j
object O_i
object $O_{j'}$
θ
Initial direction: focus line
O_j
O_i
$O_{j'}$
u
w
z
x
y
force index of O_i to O_j
force index of O_i to $O_{j'}$
F_l
θ
0
180°
force histogram of class l

Figure 5. The MOFH of class *l*. O_i, O_j, $O_{j'}$ means three different objects from the same class *l*. The force index in direction θ can be calculated by Equation (8). As for the force in each scan line, is calculated by Equation (4). The force histogram is obtained by calculating the force index in direction from 0 to 180 degrees [0, 180°).

Through traversing all the scan lines, the force index of object *l* in direction θ can be described by Equation (8). By varying the direction θ, the force indices of the different directions can be obtained. Here, the direction range is [0, 180°). The increase interval $\Delta\theta$ is usually set to 3°. The smaller $\Delta\theta$ is, the more detail information is retained, but the potential error and computing cost will be larger. After traversing all the directions, the histogram of $F_r^l = \left(F_r^l(\theta_1), F_r^l(\theta_2), \ldots, F_r^l(\theta_p)\right)$ is acquired, where *p* is the number of directions. Once this step is completed, the spatial distribution of class *l* in the scene is determined. The *F*-histograms of all the classes can then be connected to build a position feature vector $F_H = \left(F_r^1, F_r^2, \ldots, F_r^L\right)$.

$$F_r^l(\theta) = \sum_k f_r^k(\theta) \tag{8}$$

To ensure that the MOFH is invariant to rotation and mirroring, the initial direction of the *F*-histogram is defined as the direction from the focus of the largest object in the land-cover type with the highest area proportion to the largest object in the land-cover type with the lowest area proportion. Furthermore, to prevent false classification of scene categories caused by the large difference of F_H in local direction, F_H is normalized and represented by the mean and standard deviation to make the feature more robust. Finally, the *F*-histogram of class *l* is expressed as $v_H^l = (\mu_H^l, \sigma_H^l)$. The final form of the MOFH is $v_H = \left(v_H^1, v_H^2, \ldots, v_H^L\right)^T$.

The pseudo-code of the MOFH is described in Algorithm 1. The input of Algorithm 1 is the object-oriented classification result *I* of the image.

Algorithm 1. Pseudo-code of the MOFH.

```
Procedure $v_H = MOFH(I, \Delta\theta, r)$
  Initial $\theta$
  For $\theta$ = 0° to 180° do
    Generate scan lines to cover all the image with $\theta$.
    For each scan line $k$ do
      For $l$ = 1 to $L$ do
        Compute $f_r^k(\theta)$ for current class $l$ using Equations (4), (6) and (7).
      End For
      Compute the histogram of force $f_r^l(\theta)$ using Equation (8).
    End For
  End for
  Compute $\mu_H^l$, $\sigma_H^l$ of $F_H$
  Normalize $v_H^l$
  Return $v_H^l$
End Procedure
```

3.3. Scene Sematic Category Understanding

After acquiring the spatial context relations of the objects, the two relation feature vectors are connected as a long normalized vector end to end as the new feature vector $v = (v_F, v_H)$ to describe the high-level semantics. Finally, a traditional classifier such as SVM is used to train and classify the images into different scene categories.

4. Experiments and Analysis

To test the performance of the proposed MOSCRF, three datasets with progressive levels of complexity were put into use: a synthetic dataset, a USGS dataset, and the Wuhan IKONOS dataset. The synthetic dataset was used to verify the spatial layout through a visual inspection. The other two datasets were used to test the classification accuracy of MOSCRF.

In the step of object-oriented classification, the features were the spectral feature and the homogeneity of the gray-level co-occurrence matrix (GLCM) feature, and the classifier was SVM with radial basis function (RBF) kernel. The parameters of SVM were obtained by five-fold cross-validation. MAOS and MV were combined to restrict the result. The scale of the segmentation was 20 and the number of initial agents was 2000. In the step of scene sematic category understanding, we compared methods based on visual words of low-level features, such as the BoVW, spatial pyramid match (SPM), LDA and FK, methods based on deep learning features such as CNN, and methods based on objects such as frequency vector and a pair of objects (FH2) with MOSCRF. Besides, we compared the performance of different classifiers like SVM, naive Bayesian (NB), k-Nearest Neighbor (kNN), random forest (RF), and artificial neural network (ANN) acting on MOSCRF. The overall accuracy (OA) was measured by the mean, along with the standard deviation. The consistency test is Wilcoxon test. The details of the experiments are as follows.

4.1. Experiment 1: Synthetic Dataset

The synthetic dataset is defined based on the Google dataset of SIRI-WHU [6] (http://rsidea.whu.edu.cn/resource_sharing.htm), including three types of scenes: residential, commercial, and industrial area (Figure 6). The process of generating the synthetic is as follows. First, select five images randomly from the Google dataset. Second, divide the images into building, road, tree, and other land-cover classes by artificial annotation. Third, rotate each image 90°, 180°, and 270° to generate other images. Fourth, flip all images in the horizontal and vertical directions to acquire the remaining images. Consequently, the synthetic dataset consists of 120 images and each scene contains 40 images.

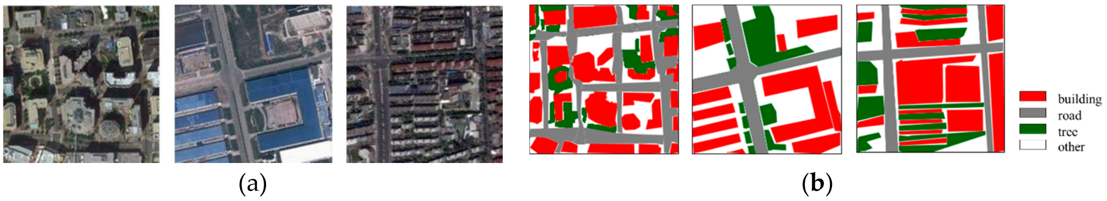


Figure 6. Google dataset and Synthetic dataset: (**a**) Google dataset (from left to right: commercial, industrial, and residential area); and (**b**) synthetic dataset (from left to right: commercial, industrial, and residential area).

Figure 7 shows the visual result. From left to right in Figure 7a are the initial image, its rotation, its mirror. Figure 7b is the MOFH. From top to bottom, the scene categories are commercial, commercial, industrial, industrial, residential, and residential. Since the horizontal three images have the same force histogram, they are represented by a graph. According to the transverse comparison, it is easy to see the invariance to rotation and mirroring of the MOFH. According to the vertical comparison, it is clear that the different scenes have different spatial configurations.

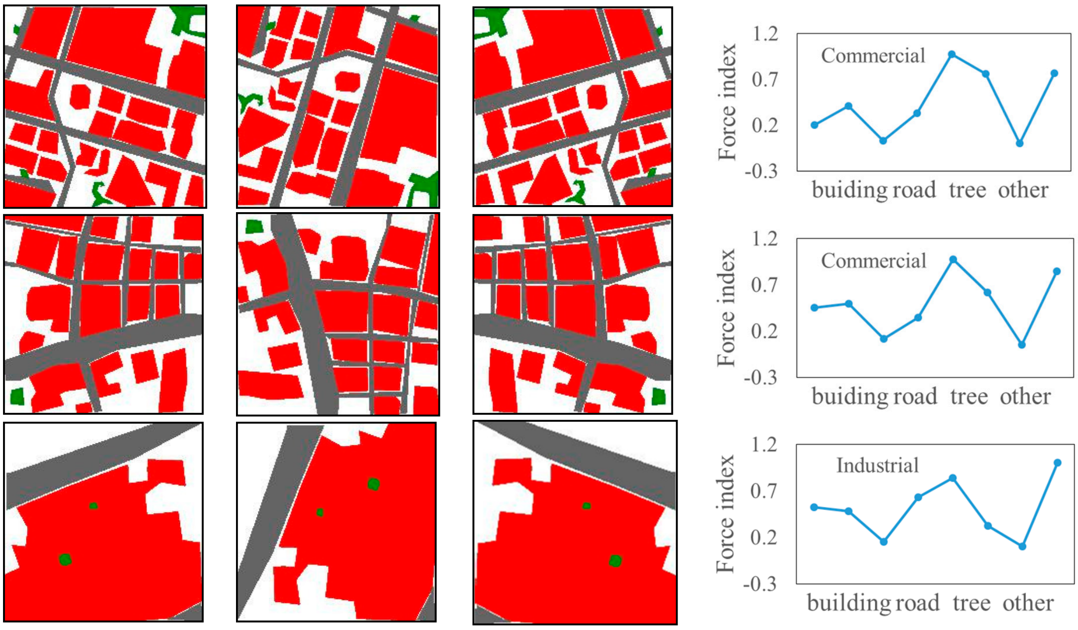


Figure 7. *Cont.*

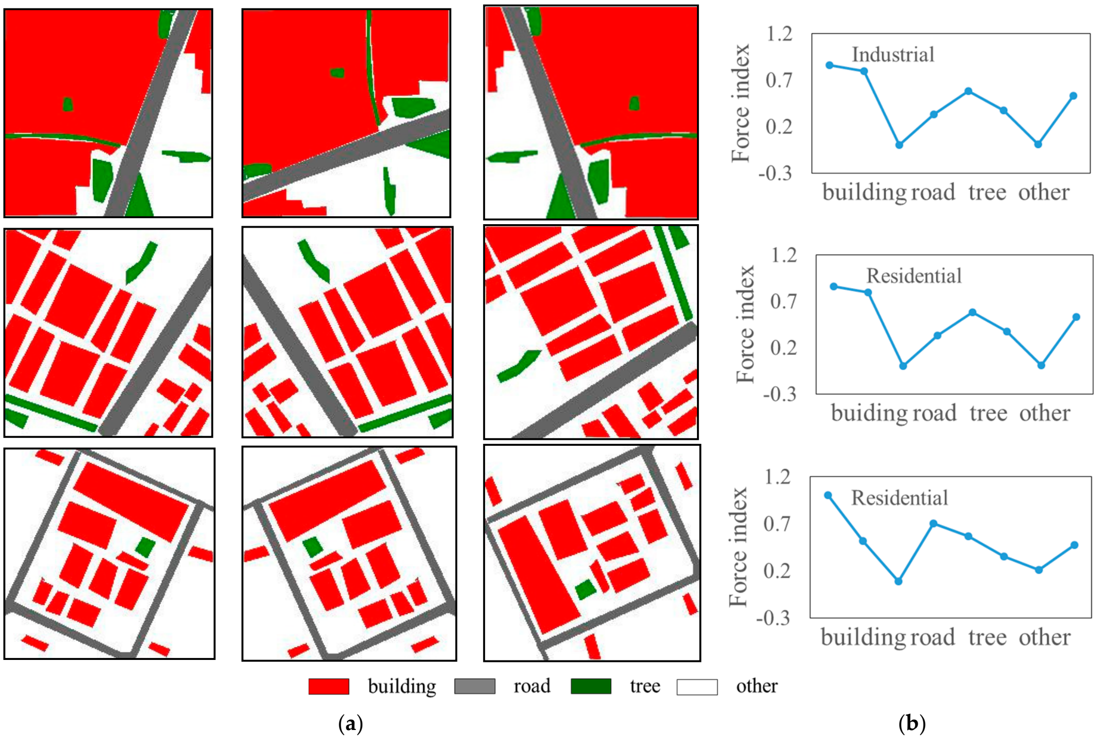


Figure 7. Visualization of the spatial layout of different scenes. (**a**) from left to right are the initial image, its rotation and its mirror. (**b**) is the MOFH. From top to bottom, the scene categories are commercial, commercial, industrial, industrial and residential, and residential.

4.2. Experiment 2: USGS Dataset

The USGS dataset was generated from a USGS database of Montgomery, Ohio, and contains a large image (Figure 8a) with the size of 10,000 × 9000 pixels and four scene classes of residential, farm, forest, and parking lot, with 143, 133, 100, and 139 images, respectively (Figure 9). For all the images, the size was 150 × 150 pixels and the resolution was 0.61 m. In these scenes, the objects were divided into five land-cover classes, namely, water, grass, tree, road, and building, whose numbers of samples were 208,299, 637,054, 594,930, 304,919, and 246,824, respectively.

For the object-oriented classification, the accuracies of the land-cover classification with the different types of features and different numbers of training samples are shown in Figure 10. It can be seen that it is better to use the spectral feature and the GLCM for the USGS dataset, as the OA is generally higher than when just using the spectral feature. The OA increases rapidly at first and then tends to stabilize with the increase of the training samples. The best OA, 92.2%, is obtained with 400 training samples in each class, which is the same as the Wuhan IKONOS dataset. Therefore, for the land-cover classification, the low-level feature was the combination of the spectral feature and the GLCM. The number of training samples was 400 in each class of land cover.

For oFK, the patch size was 8, the grid spacing was 4, and the number of cluster center was 32. For MOFH, $\Delta\theta$ of MOSCRF was set to 3°, and r was chosen as 0.5. The force indices were then from 60 different directions. SVM classifiers with RBF kernel were selected. The penalty factor and the bandwidth coefficient were tested by three-fold cross-validation. In the process of scene sematic category understanding, 50 images in every scene class were randomly selected as training samples, and the process was repeated 100 times. The accuracies of the scene understanding are listed in Table 1. The accuracies of MOSCRF based on different classifiers are listed in Table 2.

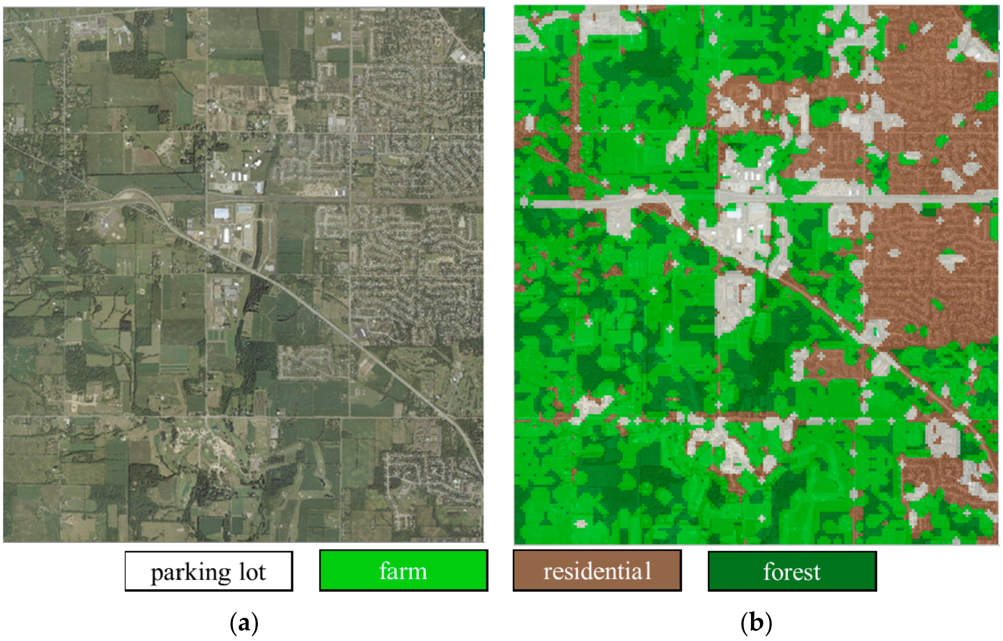


Figure 8. USGS large image and its scene understanding result: (**a**) the original image; and (**b**) the scene understanding result.

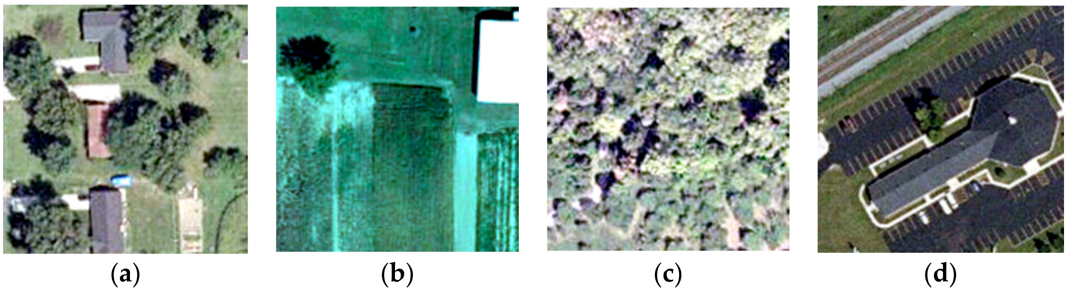


Figure 9. USGS dataset: (**a**) residential; (**b**) farm; (**c**) forest; and (**d**) parking lot.

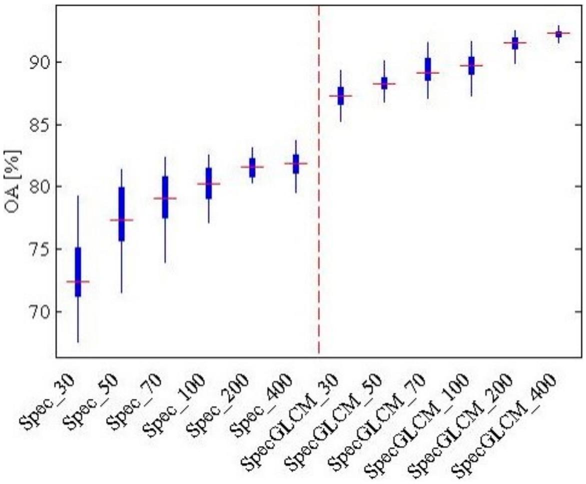


Figure 10. The accuracies of the land-cover classification for the USGS dataset with different features and samples. The underscore connects the feature and the number of samples of each class.

Table 1. Accuracies of the different methods for the USGS dataset.

Method		Accuracy (%)
Features based on visual words of low-level features	BoVW [4]	95.43
	SPM [39]	94.63
	LDA [7]	95.24
	FK [6]	96.48
Deep learning features	CNN [40]	87.08
Features based on objects	Frequency vector [18]	74.80
	FH2 [32]	90.39
	MOSCRF	**92.73**

Table 2. Accuracies of MOSCRF based on different classifiers.

Methods	OA (%)	Wilcoxon Test (α = 0.05)
SVM	**92.73 ± 1.34**	**0.8344**
NB	86.89 ± 2.25	0.5343
kNN	91.85 ± 1.29	0.4483
RF	90.05 ± 1.37	0.9428
ANN	90.93 ± 1.22	0.8179

In Table 1, it can be seen that, compared to traditional methods based on objects, MOSCRF has the highest accuracy of 92.73%; it especially has an improvement of about 17% compared to frequency vector. When the dataset is relatively small, the performance of MOSCRF exceeds the simple CNN about 5%. Although MOSCRF is lower than methods based on visual words of low-level features, it considers the distribution of internal components of the scene and is more in line with people's understanding of the scene. In Table 2, it is obvious SVM performs best, followed by ANN, and NB performs worst. The *p*-values of all classifiers are bigger than 0.05, reflecting the classifier is efficient.

According to the confusion matrix in Figure 11a, MOSCRF performs the best in distinguishing the forest area with almost zero error. Though the number of test images of farm is smaller than residential area and parking lot, the misclassification is larger than other scene categories because the farm area contains cars, buildings and trees, leading to divide into residential and forest. The training ratio and accuracy of parking lot and residential is similar, meaning the recognizing ability of MOSCRF is similar, too.

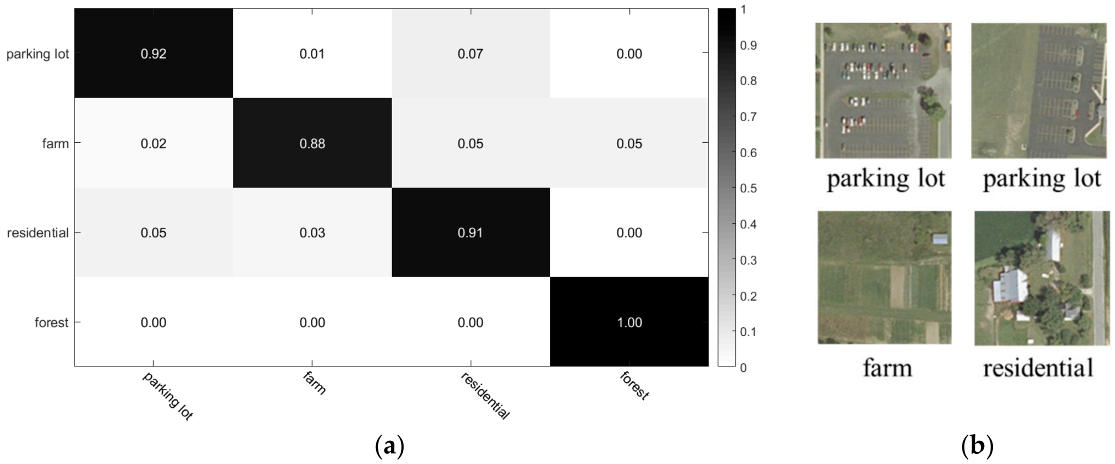


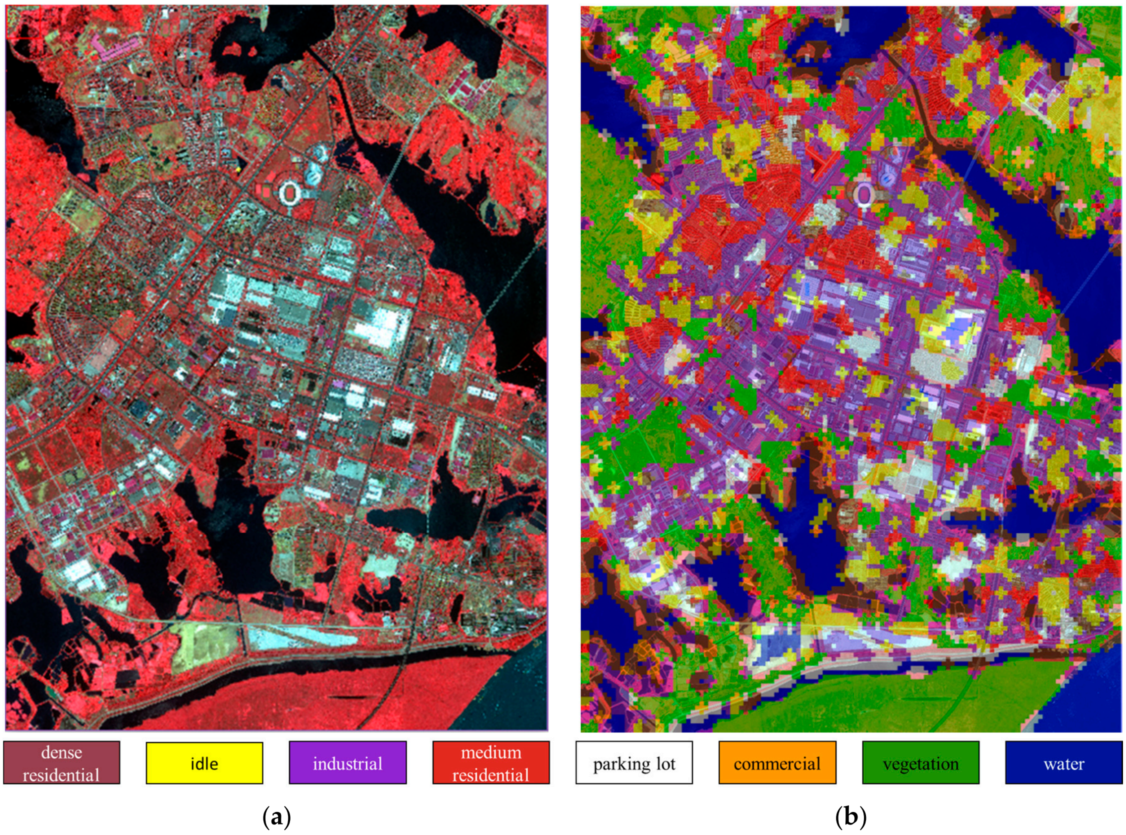
Figure 11. (**a**) The confusion matrix for the USGS dataset based on MOSCRF; (**b**) some of the classification results that MOSCRF (SP) can correctly recognize while frequency vector cannot.

Figure 11b shows some of the images classified correctly by MOSCRF but incorrectly by frequency vector. It can be seen that both methods can classify the pure scenes such as forest, but MOSCRF has a better ability to recognize those scene categories with complex spatial configuration.

According to Figure 8b, we can see that the Montgomery area is covered by water, grass, trees, roads, and buildings, with farm being the commonest scene class. Parking lots are found at the sides of the roads, and the residential area is to the northeast and east. Parking lots are obviously more common in the residential area.

4.3. Experiment 3: Wuhan IKONOS Dataset

The third dataset is the IKONOS images of the Wuhan Hanyang district obtained in 2009, including a large image (Figure 12a) with the size of 6150 × 8250 pixels and eight scene classes: dense residential, idle land, industrial, medium residential, parking lot, commercial district, vegetation, and water (Figure 13). Each scene includes 30 images with the size of 150 × 150 pixels and a 1 m resolution. The objects are divided into nine land-cover classes, including three types of buildings, three types of roads, vegetation, water, and soil, with 23,637, 46,307, 118,710, 118,238, 119,392, 22,996, 89,955, 102,614, and 13,006 samples, respectively.



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

Figure 12. Wuhan IKONOS large image and its scene understanding result: (**a**) the original image; and (**b**) the scene understanding result.

In Figure 14, for the Wuhan IKONOS dataset, it can be seen that it is more valid to use joint features (i.e., the spectral feature and the GLCM) than just the spectral feature when the number of training samples ranges from 30 to 400. The best accuracy, 87.5%, is acquired when there are 400 training samples and the features are joint features. The land-cover recognition result with the highest accuracy was then used in the subsequent scene sematic category understanding.

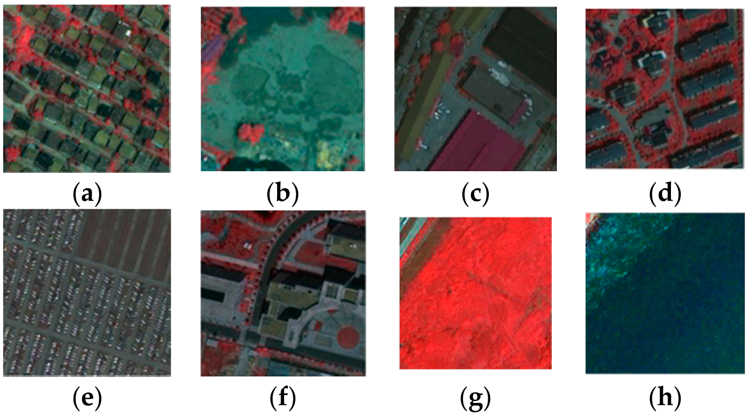


Figure 13. Wuhan IKONOS dataset: (**a**) dense residential; (**b**) idle land; (**c**) industrial; (**d**) medium residential; (**e**) parking lot; (**f**) commercial; (**g**) vegetation; and (**h**) water.

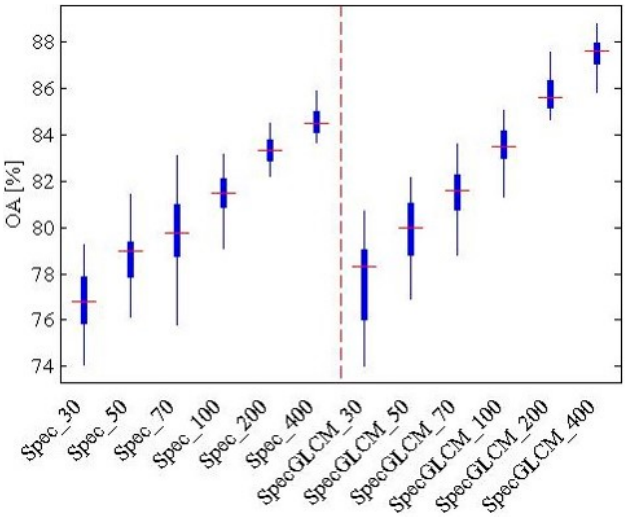


Figure 14. The accuracies of the land-cover classification for the Wuhan IKONOS dataset using different features and samples. The underscore connects the feature and the number of samples of each class.

In this case, 80% of the images in every scene class were randomly selected as training samples. The other parameters of the relation construction and scene sematic category understanding were the same as in the USGS dataset. Table 3 lists the accuracies of the different methods.

Table 3. Accuracies of the different methods for the Wuhan IKONOS dataset.

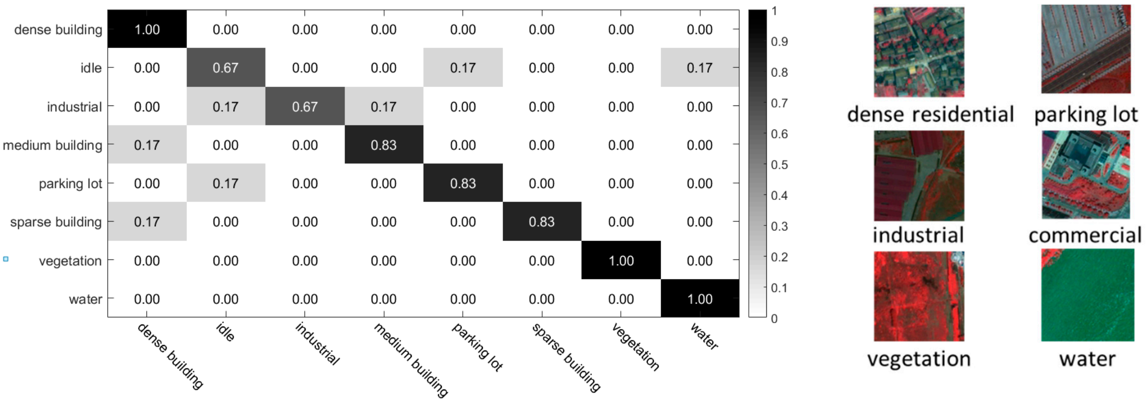
Method		Accuracy (%)
Features based on visual words of low-level features	BoVW [4]	73.85
	SPM [39]	71.69
	LDA [7]	77.34
	FK [6]	77.35
Deep learning features	CNN [40]	74.45
Features based on objects	Frequency vector [18]	73.31
	FH2 [32]	73.27
	MOSCRF	**80.63**

In Table 3, it can be seen that the performance is different from the USGS dataset. Here, MOSCRF acquires the best accuracy, 80.63%, which is at least 3% higher than the other methods. Comparing CNN to MOSCRF explains that MOSCRF is more suitable when lacking training samples, while comparing FK or LDA to MOSCRF reflects that ground object information is useful to scene classification. That is, the relations of objects are helpful in scene understanding. Both co-occurrence relations and positions relations are essential in scene understanding, according to the result that MOSCRF is 7% higher than frequency vector and FH2. In Table 4, all classifiers are usable, while SVM is the best, followed by RF. NB is the worst and is nearly 20% lower than SVM.

Table 4. Accuracies of MOSCRF based on different classifiers.

Methods	OA (%)	Wilcoxon Test (α = 0.05)
SVM	**80.63 ± 4.54**	**0.5971**
NB	60.73 ± 4.70	0.6602
KNN	70.60 ± 6.11	0.5096
RF	72.92 ± 5.70	0.8260
ANN	71.88 ± 5.10	0.8969

According to the confusion matrix in Figure 15a, MOSCRF performs the best in distinguishing dense building, vegetation and water because of their regular distribution and relatively pure objects. Meanwhile, the idle and industrial scene classes are difficult to distinguish because of the vague and similar spatial configurations of their internal components: roads, buildings, and soil. Figure 15b shows some of the images that are classified correctly by MOSCRF but incorrectly by frequency vector. It can be seen that MOSCRF is good at recognizing those objects with similar area proportions but different spatial configurations.



	dense building	idle	industrial	medium building	parking lot	sparse building	vegetation	water
dense building	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
idle	0.00	0.67	0.00	0.00	0.17	0.00	0.00	0.17
industrial	0.00	0.17	0.67	0.17	0.00	0.00	0.00	0.00
medium building	0.17	0.00	0.00	0.83	0.00	0.00	0.00	0.00
parking lot	0.00	0.17	0.00	0.00	0.83	0.00	0.00	0.00
sparse building	0.17	0.00	0.00	0.00	0.00	0.83	0.00	0.00
vegetation	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
water	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

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

Figure 15. (**a**) The confusion matrix for the Wuhan IKONOS dataset based on MOSCRF; and (**b**) some of the classification results that MOSCRF can recognize while frequency vector cannot.

In Figure 12b, we can see that the land-cover types are complex in Hanyang, including three types of buildings, three types of roads, vegetation, water, and soil. The area is still developing and the city planning is incomplete. As a result, the spatial layout of Hanyang is not regular.

5. Sensitivity Analysis

5.1. Sensitivity Analysis of the Interval $\Delta\theta$ of MOFH

$\Delta\theta$ is an important parameter for the MOFH, and it indicates the direction continuity of the MOFH. To test the sensitivity of MOSCRF to $\Delta\theta$, land-cover classification results of the same accuracy and SVM with linear kernel were used. $\Delta\theta$ was set to 2, 3, 4, 6, and 9, and the results are shown in Figure 16. Figure 16a is the result for the USGS dataset, where the effect of $\Delta\theta$ is very small as the range of OA is from 89.74% to 89.79%. Figure 16b is the result for the Wuhan IKONOS dataset, where the effect of $\Delta\theta$ is increased. Furthermore, the USGS dataset obtains the best performance at 6°, while, for the Wuhan IKONOS dataset, it is 2°. The specific spatial layout of the different datasets causes the different direction continuity, leading to the highest accuracy being obtained at different values of $\Delta\theta$. However, the smaller the value of $\Delta\theta$, the higher the calculation cost; thus, 3° was chosen in Experiment 1 and Experiment 2.

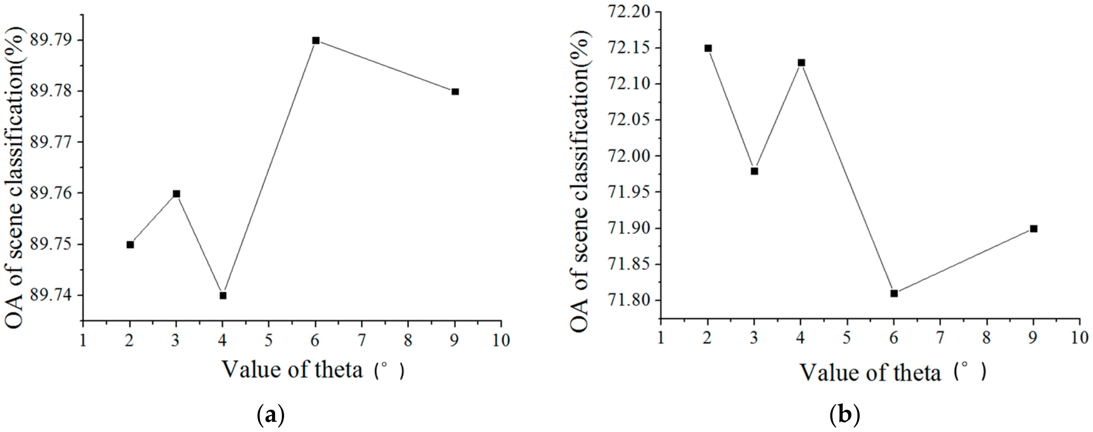


Figure 16. Sensitivity analysis of $\Delta\theta$: (**a**) the scene sematic category understanding result for the USGS dataset; and (**b**) the scene sematic category understanding result for the Wuhan IKONOS dataset.

5.2. Sensitivity Analysis of r in the MOFH

Parameter r is an important parameter to quantify the distance in the MOFH. When r is 0, the forces are constant, while, when r is 2, the form of the force is similar to gravity. Values of $r = 0$, $r = 0.5$, and $r = 2$ were tested. The classifier was SVM with RBF kernel. According to Table 5, it is clear that $r = 2$ is not suitable for MOFH because it can only model disjoint relationships. $r = 0$ and $r = 0.5$ can both be used in MOFH, but $r = 0.5$ performs better. Therefore, the choice of r should follow the actual distribution of the objects.

Table 5. Sensitivity analysis of r in the MOFH.

Value	OA of the USGS Dataset (%)	OA of the Wuhan IKONOS Dataset (%)
0	90.23	78.54
0.5	92.73	80.63
2	16.37	14.23

5.3. Sensitivity Analysis of the Land-Cover Classification Accuracy

The land-cover classification accuracy affects the scene sematic category understanding accuracy. To test how this affects MOSCRF, five groups of different land-cover classification results and SVM with linear kernel were used. The values of $\Delta\theta$ and r were kept the same. In Figure 17a, it is clear that the higher the accuracy of the land-cover classification, the higher the accuracy of the scene sematic category understanding, as the curve is monotonically increasing. This can be explained by the fact that, if the land-cover classification accuracy is high, there will be less loss of land-cover information. However, for the Wuhan IKONOS dataset, the curve is irregular (Figure 17b). The highest accuracy of scene sematic category understanding, 75.58%, is obtained when the land-cover classification accuracy is 78.419%, instead of 84.187%. This may be because, when the land-cover classification result is poor, the spatial configuration is easily affected by small but incorrect regions, leading to the turbulence of the curve. When the land-cover classification accuracy is relatively high, the trend is the same as for the USGS dataset, as seen in the end of the curve in Figure 17b. Therefore, it is important to improve the land-cover recognition result to reduce the noise in constructing MOSCRF.

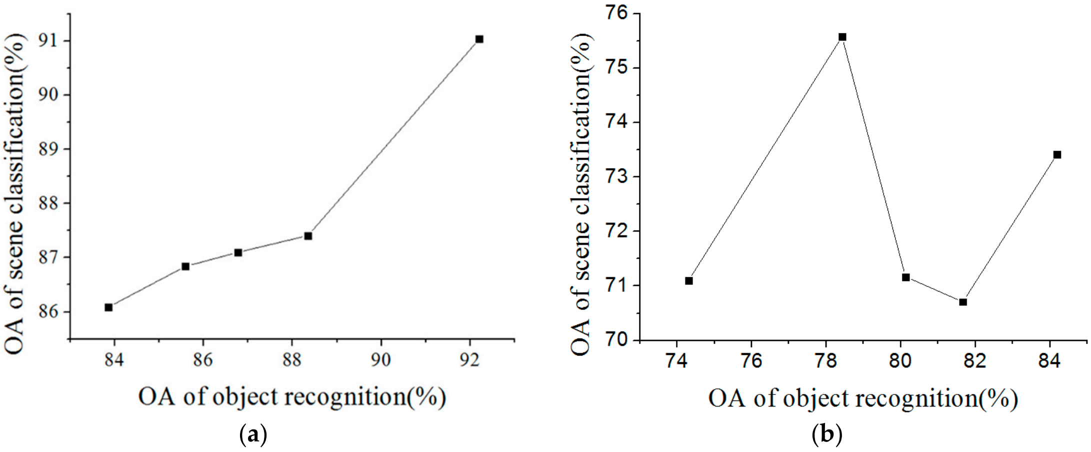


Figure 17. Sensitivity analysis of the land-cover classification accuracy: (**a**) the scene sematic category understanding result for the USGS dataset; and (**b**) the scene sematic category understanding result for the Wuhan IKONOS dataset.

6. Conclusions

Although BoVW, topic models, and deep learning algorithms can acquire a relatively high accuracy in scene classification, they do not take the prior knowledge of the objects into consideration. Therefore, they cannot fully understand the components of the images and their relations in the scene. In this paper, to solve this problem, we have proposed scene understanding based on the spatial context relations of multiple objects. The proposed approach consists of three main steps: (1) object-oriented classification based on MAOS + MV; (2) spatial context relations construction consisting of co-occurrence relations construction by the oFK and position relations construction by the MOFH; and (3) scene sematic category understanding by SVM-RBF. The oFK is the extension of the traditional FK based on low-level features in replacing the low-level features with the category information to justify the distribution of the object categories. MOFH extends the *F*-histogram of pairwise objects into multiple objects to serve the HRSIs and express the spatial layout of the scene. Moreover, the proposed MOFH has the characteristics of invariance to rotation and mirroring. MOSCRF is the framework of scene understanding based on these three steps.

The experimental results not only show that the proposed method performs better with than the traditional methods and classifiers, but it also identifies the internal composition of the scene and the relations of the objects. Therefore, MOSCRF has clear geographical significance in researching the internal patterns of scenes and is very deserving of further study to mine more information and improve the accuracy of scene understanding.

Acknowledgments: This work was supported by National Natural Science Foundation of China under Grant Nos. 41771385, 41622107 and 41371344, National Key Research and Development Program of China under Grant No. 2017YFB0504202, and Natural Science Foundation of Hubei Province in China under Grant No. 2016CFA029.

Author Contributions: All the authors made significant contributions to the work. Yanfei Zhong and Siqi Wu designed the research and analyzed the results. Bei Zhao provided advice for the preparation and revision of the paper.

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

References

1. Dumitru, C.O.; Cui, S.; Schwarz, G.; Datcu, M. Information content of very-high-resolution sar images: Semantics, geospatial context, and ontologies. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *8*, 1635–1650. [CrossRef]
2. Zhong, Y.; Zhu, Q.; Zhang, L. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [CrossRef]
3. Chen, S.; Tian, Y.L. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1947–1957. [CrossRef]
4. Zhao, L.J.; Tang, P.; Huo, L.Z. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4620–4631. [CrossRef]
5. Yang, W.; Yin, X.; Xia, G.S. Learning high-level features for satellite image classification with limited labeled samples. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4472–4482. [CrossRef]
6. Zhao, B.; Zhong, Y.; Zhang, L.; Huang, B. The fisher kernel coding framework for high spatial resolution scene classification. *Remote Sens.* **2016**, *8*, 157. [CrossRef]
7. Luo, W.; Li, H.; Liu, G.; Zeng, L. Semantic annotation of satellite images using author-genre-topic model. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 1356–1368. [CrossRef]
8. Zhao, B.; Zhong, Y.; Xia, G.S.; Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2108–2123. [CrossRef]
9. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 1349–1362. [CrossRef]
10. 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. [CrossRef]
11. Porway, J.; Wang, Q.; Zhu, S.C. A hierarchical and contextual model for aerial image parsing. *Int. J. Comput. Vis.* **2009**, *88*, 254–283. [CrossRef]
12. Li, F.F.; Perona, P. A bayesian hierarchical model for learning natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2015; pp. 524–531.
13. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
14. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
15. 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]
16. Biederman, I.; Mezzanotte, R.J.; Rabinowitz, J.C. Scene perception: Detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **1982**, *14*, 143. [CrossRef]
17. Tang, H.; Shen, L.; Qi, Y.; Chen, Y.; Shu, Y.; Li, J.; Clausi, D.A. A multiscale latent dirichlet allocation model for object-oriented clustering of vhr panchromatic satellite images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 1680–1692. [CrossRef]
18. Vogel, J.; Schiele, B. Semantic modeling of natural scenes for content-based image retrieval. *Int. J. Comput. Vis.* **2006**, *72*, 133–157. [CrossRef]
19. Li, L.-J.; Su, H.; Lim, Y.; Fei-Fei, L. Object bank: An object-level image representation for high-level visual recognition. *Int. J. Comput. Vis.* **2013**, *107*, 20–39. [CrossRef]
20. Aksoy, S.; Koperski, K.; Tusk, C.; Marchisio, G.; Tilton, J.C. Learning bayesian classifiers for scene classification with a visual grammar. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 581–589. [CrossRef]
21. Hu, M. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theory* **1962**, *8*, 179–187.
22. Bondugula, R.; Matsakis, P.; Keller, J.M. Force histograms and neural networks for human-based spatial relationship generalization. In Proceedings of the Iasted International Conference on Neural Networks and Computational Intelligence, Grindelwald, Switzerland, 23–25 February 2004; pp. 185–190.
23. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Urban land use extraction from very high resolution remote sensing imagery using a bayesian network. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 192–205. [CrossRef]

24. Scott, G.; Klaric, M.; Shyu, C.R. Modeling multi-object spatial relationships for satellite image database indexing and retrieval. In Proceedings of the International Conference on Image and Video Retrieval, Singapore, 20–22 July 2005; pp. 247–256.
25. Shyu, C.R.; Klaric, M.; Scott, G.J.; Barb, A.S.; Davis, C.H.; Palaniappan, K. Geoiris: Geospatial information retrieval and indexing system—Content mining, semantics modeling, and complex queries. *Appl. Phys. Lett.* **2013**, *102*, 2564–2567. [CrossRef] [PubMed]
26. Sjahputera, O.; Keller, J.M. Scene matching using f-histogram-based features with possibilistic c-means optimization. *Fuzzy Sets Syst.* **2007**, *158*, 253–269. [CrossRef]
27. Vaduva, C.; Gavat, I.; Datcu, M. Latent dirichlet allocation for spatial analysis of satellite images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2770–2786. [CrossRef]
28. Matsakis, P.; Wendling, L. A new way to represent the relative position between areal objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 634–643. [CrossRef]
29. Perronnin, F.; Dance, C. Fisher Kernels on Visual Vocabularies for Image Categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
30. Matsakis, P. Understanding the Spatial Organization of Image Regions by Means of Force Histograms: A Guided Tour. In *Applying Soft Computing in Defining Spatial Relations*; Matsakis, P., Sztandera, L.M., Eds.; Physica: Heidelberg, Germany, 2002; Volume 106, pp. 99–122.
31. Huang, X.; Zhang, L. An adaptive mean-shift analysis approach for object extraction and classification from urban hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 4173–4185. [CrossRef]
32. Trimble. Available online: http://www.ecognition.com/suite/ecognition-developer (accessed on 1 October 2017).
33. Gonzalez, R.C.; Wintz, P. *Digital Image Processing*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2010; pp. 484–486.
34. Wang, L.; Sousa, W.P.; Gong, P. Integration of object-based and pixel-based classification for mapping mangroves with ikonos imagery. *Int. J. Remote Sens.* **2004**, *25*, 5655–5668. [CrossRef]
35. Wang, Q.; Lin, J.; Yuan, Y. Salient band selection for hyperspectral image classification via manifold ranking. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *27*, 1279–1289. [CrossRef] [PubMed]
36. Wang, Q.; Gao, J.; Yuan, Y. A joint convolutional neural networks and context transfer for street scenes labeling. *IEEE Trans. Intell. Transp. Syst.* **2017**, *99*, 1–14. [CrossRef]
37. Russell, S.J.; Norvig, P. Artificial intelligence: A modern approach. *Appl. Mech. Mater.* **2010**, *263*, 2829–2833.
38. Zhong, Y.; Zhao, B.; Zhang, L. Multiagent object-based classifier for high spatial resolution imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 841–857. [CrossRef]
39. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
40. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *Acta Ecol. Sin.* **2015**, *28*, 627–635.