*remote sensing*

MDPI

*Article*

# Region-Wise Deep Feature Representation for Remote Sensing Images

**Peng Li [1,2], Peng Ren [1,2], Xiaoyu Zhang [3,*], Qian Wang [4], Xiaobin Zhu [4,*] and Lei Wang [5]**

1 College of Information and Control Engineering, China University of Petroleum (East China), Qingdao 266580, China; lipeng@upc.edu.cn (P.L.); pengren@upc.edu.cn (P.R.)
2 State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China
3 Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
4 College of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China; wangqianjsj@163.com
5 Academy of Broadcasting Science, SARFT, Beijing 100045, China; wanglei@abs.ac.cn
* Correspondence: zhangxiaoyu@iie.ac.cn (X.Z.); brucezhucas@gmail.com (X.Z.); Tel.: +86-10-8254-6701 (X.Z.); +86-10-6898-5285 (X.Z.)

Received: 8 May 2018; Accepted: 1 June 2018; Published: 5 June 2018


**Abstract:** Effective feature representations play an important role in remote sensing image analysis tasks. With the rapid progress of deep learning techniques, deep features have been widely applied to remote sensing image understanding in recent years and shown powerful ability in image representation. The existing deep feature extraction approaches are usually carried out on the whole image directly. However, such deep feature representation strategies may not effectively capture the local geometric invariance of target regions in remote sensing images. In this paper, we propose a novel region-wise deep feature extraction framework for remote sensing images. First, regions that may contain the target information are extracted from one whole image. Then, these regions are fed into a pre-trained convolutional neural network (CNN) model to extract regional deep features. Finally, the regional deep features are encoded by an improved Vector of Locally Aggregated Descriptors (VLAD) algorithm to generate the feature representation for the image. We conducted extensive experiments on remote sensing image classification and retrieval tasks based on the proposed region-wise deep feature extraction framework. The comparison results show that the proposed approach is superior to the existing CNN feature extraction methods.

**Keywords:** convolutional neural networks (CNNs); feature representation; scene classification; image retrieval

## 1. Introduction

With the developments of satellite imaging techniques, it is much easier to acquire a large collection of remote sensing images. In recent years, automatic remote sensing image analysis [1–4] has become a hot topic due to its wide applications in many fields such as military reconnaissance, agriculture, and environmental monitoring. Feature extraction and representation is the foundation of many remote sensing image processing tasks [5–9]. Developing powerful image feature representation methods helps us understand the image information more accurately.

During the past decades, a variety of feature learning methods for remote sensing images have been proposed. In earlier years, remote sensing image analysis was mainly based on the hand-crafted features which include both global features and local features. Global features [10–12] include color, shape and textual information, which are the primary characteristic of a remote sensing image. The global features are extracted based on the whole image, and they are not able to reflect the local information of interested area. Among the local features [13,14], bag-of-words (BoW) and its

variations [15–17] are one of the most popular types in recent decades, which have comprised the state of the art for several years in the remote sensing community because of their simplicity, efficiency, and invariance to viewpoint changes. In addition to the hand-crafted features, data-driven features are also developed via unsupervised feature learning in terms of content-based remote sensing image retrieval and classification tasks [18–22]. For example, a saliency-guided unsupervised feature learning approach was proposed in [19] for remote sensing scene classification. A multiple feature-based remote sensing image retrieval approach was proposed in [21] by combining hand-crafted features and data-driven features via unsupervised feature learning. Wang et al. [22] proposed a multilayered graph model for hierarchically refining retrieval results from coarse to fine. However, as the remote sensing image understanding task becomes more challenging, the description capabilities of the above low-level features are limited and may not be effective to capture the high-level semantics.

More recently, various deep learning algorithms [23–26], especially convolutional neural networks (CNNs), have shown their much stronger feature representation power in many fields such as traffic scene analysis [27,28] and bush-fire frequency forecasting [29,30]. Convolutional neural networks (CNNs) learn high-level semantic features automatically rather than requiring hand-crafted features and have achieved great success in many remote sensing image analysis applications [31–39]. For example, a low dimensional convolutional neural network was learned in [34] for high-resolution remote sensing image retrieval while an unsupervised convolutional feature fusion network was developed for deep representation of remote sensing images in the scene classification task [36]. In these CNN-based remote sensing image feature learning methods, the whole image is usually directly fed into a pre-trained or fine-tuned network to obtain the deep representation. However, there exists one problem that is seldom exploited in the existing CNN feature extraction methods. Compared with other images, remote sensing images have several special characteristics. For example, even in the same category, the targets in different images may have varied sizes, colors, and angles. More importantly, other materials and the background around the target area may cause high intraclass variance and low interclass variance. Therefore, if we directly extract the CNN features from the whole image in the traditional manner, the image representations in the feature space may not accurately reflect their true category information (as demonstrated in Figure 1a).

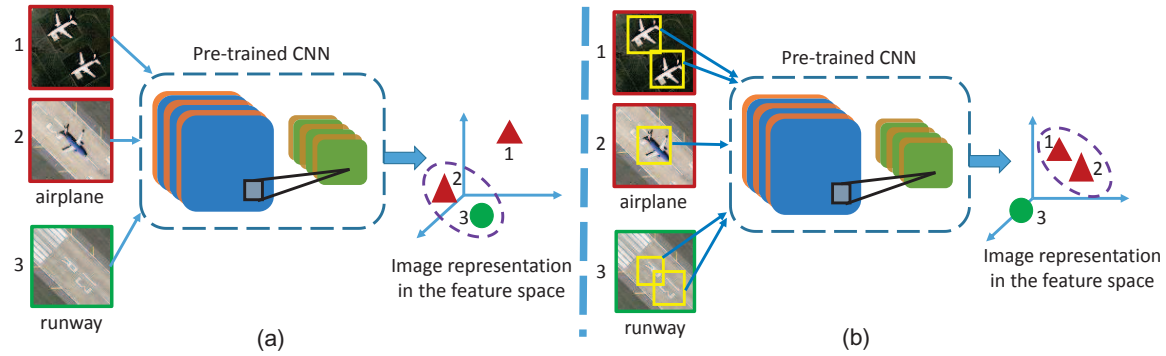



**Figure 1.** Remote sensing image representations by different deep CNN feature extraction frameworks: (**a**) existing deep CNN feature extraction from the whole image; (**b**) region-wise deep CNN feature extraction.

In order to address the above problem, we propose a novel region-wise deep CNN feature representation method for remote sensing image analysis, which extracts the CNN features from regions containing the targets instead of the whole image (see Figure 1b). The proposed feature extraction approach includes the following steps: First, regions that may contain the targets are generated from one whole image. Then, these regions are fed into a pre-trained CNN model to extract the regional deep features. Finally, the regional deep features are encoded by an improved Vector of Locally Aggregated Descriptors (VLAD) algorithm to generate the feature representation for the

image. The image features extracted by our proposed approach have more powerful and effective representation ability to capture the local target information and geometric invariance. The flowchart of the proposed region-wise deep CNN feature learning method is illustrated in Figure 2.

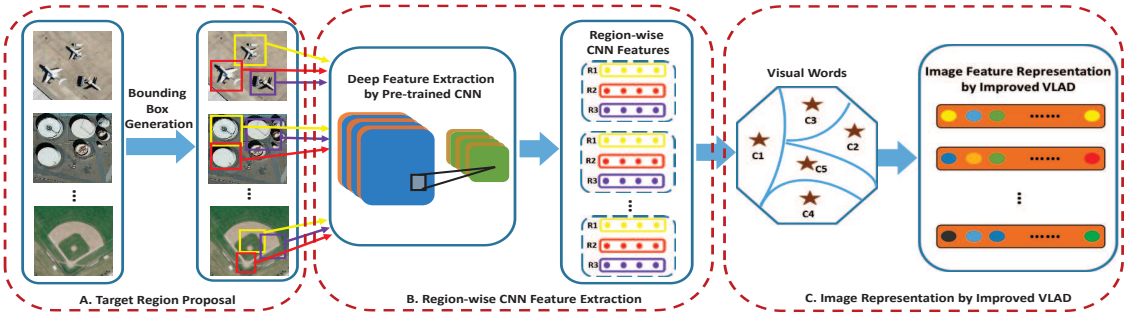



**Figure 2.** The framework of the proposed region-wise deep CNN feature learning approach.

## 2. The Proposed Approach

### *2.1. Target Region Proposal*

As we have introduced in Section 1, our proposed feature extraction method is based on the target regions instead of the whole image. Thus, we have to generate the regions that may contain the targets firstly. The regions are expected to reflect the objects from the same category rather than the varied background, such that the features extracted from these regions have more discriminative power between different classes. Region proposal algorithm can generate a set of bounding boxes, which may contain the interested targets. In this paper, we apply edge-boxes algorithm [40] to generate the object-wise bounding boxes. Edge-boxes method is able to cover most objects in one image with a set of bounding boxes as well as their corresponding confidential scores. It generates the target bounding boxes directly from edges information and each box is scored based on the number of contours wholly enclosed in it. The score is calculated as follows: First, neighboring edge pixels of similar orientation are clustered together to form groups. Then, affinities between edge groups are computed based on their relative positions and orientations. Finally, each bounding box is scored by summing the edge strength of all edge groups within the box, and subtracting the edge magnitudes in the center of the box. Specifically, the score $h_x$ of bounding box $x$ in one image is expressed as follows:

$$h_x = \frac{\sum_i w_x(s_i) m_i - \sum_{p \in x^{in}} m_p}{2(x_w + x_h)^b} \tag{1}$$

where $s_i$ denotes the edge group generated from the neighboring edge pixels, and $w_x(s_i)$ denotes the continuous value which indicates whether $s_i$ is wholly contained in the bounding box $x$. $m_i$ is the sum of the magnitudes of all the edge pixels in edge group $s_i$. $x_w$ and $x_h$ are the width and height of $x$. $p \in x^{in}$ denotes the set of edge groups $p$ located in the center of $x$, which are considered having no contribution to the score and $b$ is the bias. According to the computed scores, we select the first $K$ bounding boxes with the highest scores for each image as the candidate target regions. More details of bounding box generation for target regions can be found in [40].

### *2.2. Region-Wise CNN Feature Extraction*

Based on the edge-boxes algorithm, we obtain a set of candidate target regions and their corresponding scores for each image. In the following step, we will extract the deep features from the target regions because deep features learn better high-level semantic information than hand-crafted low-level features. Among various deep learning algorithms, CNN is one of the most commonly used deep learning architectures for image feature extraction. A typical CNN model is generally structured as a series of layers including convolutional layers, pooling layers, and fully

connected layers. Many deep CNN models have been developed for image analysis in the past few years, such as AlexNet [23], VGG-Net [24], and GoogleNet [25]. Without loss of generality, we choose AlexNet as the CNN feature extraction model in this paper and the candidate regions are fed into AlexNet for deep feature extraction. The AlexNet model has five convolutional layers and three fully-connected layers. We directly copy the model parameters for convolutional layers conv1-conv5 and fully-connected layers fc6-fc7, which are pre-trained on ImageNet dataset [23]. The output 4096-dimensional vector of the fully-connected layer fc7 is extracted as the deep feature for each target region. Therefore, for each image $I$ with $K$ candidate target regions, we can obtain $K$ region-wise CNN features as $I = \{r_1, r_2, \ldots, r_K\}$ where $r_i \in R^D (i = 1, ..., K)$ denote the D-dimensional fc7 CNN feature vector for the $i$-th region.

### 2.3. Image Representation by Improved VLAD

After we have obtained the region-wise CNN features for each image, these massive regional deep features need to be encoded into a single vector for image representation. In this subsection, we propose an improved Vector of Locally Aggregated Descriptors (VLAD, [41]) method to encode the regional feature vectors into a single long vector for each image. Before encoding, we have to generate a set of $M$ visual words $C = \{c_1, c_2, \ldots, c_M\}$ where each visual word $c_i \in R^D (i = 1, ..., M)$ is a D-dimensional vector. This can be simply done by running $k$-means clustering algorithm on all the regional CNN features of the whole image database and each cluster center can be regarded as one visual word. The traditional VLAD representation [41] for each image $I$ is the concatenation of $M$ $D$-dimensional vectors as $V = [v_1, v_2, \ldots, v_M] \in R^{DM}$ where $v_i (i = 1, ..., M)$ is computed as follows:

$$v_i = \sum_{r_k : NN(r_k) = c_i} (r_k - c_i) \tag{2}$$

where $NN(r_k) = c_i$ denotes that the nearest visual word (cluster center) of regional feature vector $r_k$ is $c_i$. Thus, $v_i$ is the aggregation of differences between each visual word and its assigned regional features.

From Equation (2), we can find that the traditional VLAD approach only calculates the differences between the assigned regional features and their nearest neighbor visual word. However, it is often possible that some regional features have similar or even identical distances between two or more visual words. Assigning the regional vectors only to one nearest visual word may not be appropriate and sometimes loses important information. More importantly, each bounding box in one image has a corresponding score obtained in the region proposal stage, which reflects the confidence that the target is contained in the region. If we apply the VLAD encoding method directly, the score weight information is also neglected. To overcome the above shortcomings, we propose a weighted multi-neighbor assignment strategy for the regional CNN features to improve the representation ability of the traditional VLAD method. Specifically, we propose to calculate the new $v_i^{new}$ in VLAD by the following equation:

$$v_i^{new} = \sum_{r_k : NN(r_k) \supseteq c_i} h_{r_k} \cdot \beta_{ki} \cdot (r_k - c_i) \tag{3}$$

where $h_{r_k}$ is the score of region $r_k$ computed by Equation (1). $\beta_{ki}$ is the weight of difference between regional feature $r_k$ and visual word $c_i$, which is simply computed through Gaussian function $\beta_{ki} = \exp -(\|r_k - c_i\|^2 / \sigma)$. $NN(r_k) \supseteq c_i$ denotes that visual word $c_i$ is included in the set of visual words that regional feature vector $r_k$ have been assigned to. By taking the region score and visual word assignment weight into consideration, the VLAD image representation through our method is more accurate than the traditional VLAD approach. The comparison of the improved VLAD with the original VLAD method is demonstrated in Figure 3.

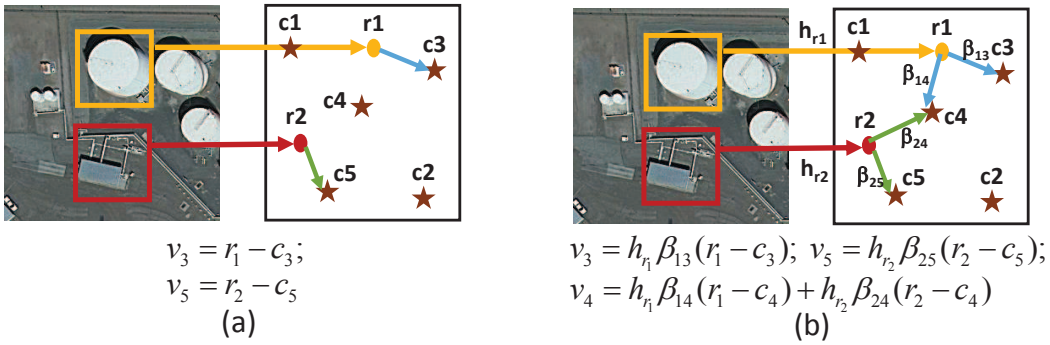



**Figure 3.** The comparison of original VLAD (Vector of Locally Aggregated Descriptors) and the improved VLAD descriptors generation: (**a**) original VLAD with nearest neighbor assignment; (**b**) the improved VLAD with weighted multi-neighbor assignment.

By performing the improved VLAD encoding algorithm, each image is encoded into a single $MD$-dimensional long vector as $V^{new} = [v_1^{new}, v_2^{new}, \ldots, v_M^{new}] \in R^{DM}$. Because $DM$ may be very large which leads to a quite long $V^{new}$, we apply principle component analysis (PCA) on $V^{new}$ for dimensionality reduction and get the final feature representation for each image.

The proposed region-wise deep CNN feature representation framework can be applied to many kinds of remote sensing image analysis tasks such as scene classification, image retrieval and so on. We will show the superiority of our proposed image representation method to the existing CNN feature extraction approaches for remote sensing images in the following section.

## 3. Experiments

In this section, we will conduct extensive experiments to evaluate the performance of the proposed region-wise deep feature representation method on different remote sensing image analysis tasks, i.e., remote sensing scene classification and large-scale remote sensing image retrieval.

### *3.1. Datasets and Settings*

Two public available remote sensing image datesets are used in the experiments: UC-Merced Landuse dataset [16] and Aerial Image Dataset (AID) [42]. The images in the UC-Merced Landuse dataset were manually extracted from large images from the USGS (United States Geological Survey) National Map Urban Area Imagery collection for various urban areas around the country. The pixel resolution of this public domain imagery is 1 foot. The UC-Merced dataset contains 2100 images in total and each image measures 256 × 256 pixels. There are 100 images for each of the following 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. The AID dataset is a new benchmark dataset for performance evaluation of aerial scene analysis, which was released in 2017 and is much larger than the UC-Merced dataset. AID is a new large-scale aerial image dataset, by collecting sample images from Google Earth imagery and the new dataset is made up of the following 30 aerial scene types: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks and viaduct. In all, the AID dataset has a number of 10,000 images within 30 classes and each class contains about 200 to 400 samples of size 600 × 600 pixels. Some sample images from the two datasets are shown in Figure 4.

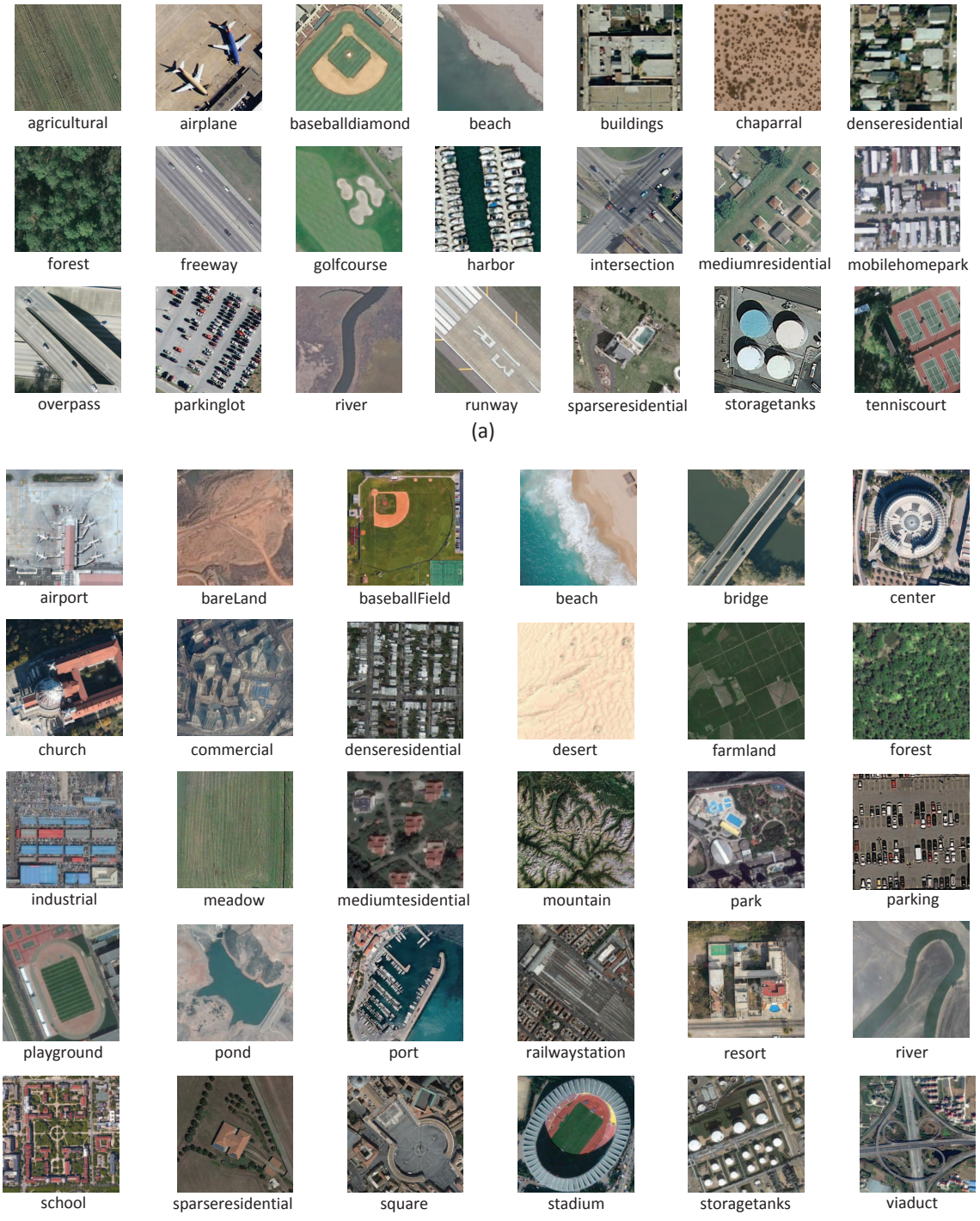



**Figure 4.** Some sample images from the two datasets: (**a**) UC-Merced Landuse dataset and (**b**) AID dataset.

For target region proposal in the proposed approach, we directly use the code of edge-boxes algorithm [40] provided by the authors. The number $K$ of candidate regions for each image is uniformity set to 500 in the experiments. AlexNet model is adopted to extract the deep CNN features for all the candidate regions and the model parameters are directly downloaded from the web, which are pre-trained on the ImageNet dataset. The output D-dimensional (D = 4096) vector of the layer fc7 is extracted as the regional deep CNN features. In the improved VLAD step, $k$-means clustering algorithm is firstly ran on all the regional CNN features to obtain $M = 64$ cluster centers as the visual words. Then, all the images can be encoded into a $MD$ dimensional vector based on the improved VLAD

algorithm through Equation (3) and the number of nearest neighbor visual words for each region is set to 5 in the multi-neighbor assignment stage. Finally, PCA is performed for dimensionality reduction and a 1024-dimensional feature vector as the representation for each image is obtained. The proposed regional deep CNN feature representation method is denoted by CNN-R in the experiments.

For comparison, we also implement the traditional CNN feature representation, where the whole image is directly fed into the pre-trained AlexNet model and a 4096-dimensional vector of the layer fc7 is extracted for each image. We denote the whole image based CNN feature by CNN-W in the experiments. PCA is also performed to reduce the dimensionality to 1024 for fair comparison.

### *3.2. Results and Discussion*

#### 3.2.1. Results for Remote Sensing Scene Classification

We first evaluate the proposed region-wise deep feature representation method CNN-R for remote sensing scene classification task on the UC-Merced dataset. CNN-W as well as state-of-the-art remote sensing image classification methods are used as benchmarks in the experiments. SVM is used as the classifier for CNN-R and CNN-W. Similar to the previous works in the literature [36,37], we randomly select 80% images from each class to train the SVM model and the remained 20% images are used for testing. According to [43], overall accuracy and confusion matrix are usually adopted as the metrics for accuracy assessment. Other related works such as [44] also report measures that derive from the confusion matrix, in which the Bradley-Terry Model was used to quantify association in remotely sensed images. In order for fair comparison with the reported results in previous remote sensing image classification works [19,32,35–37], we adopt the same accuracy assessment measures used in the above literature.

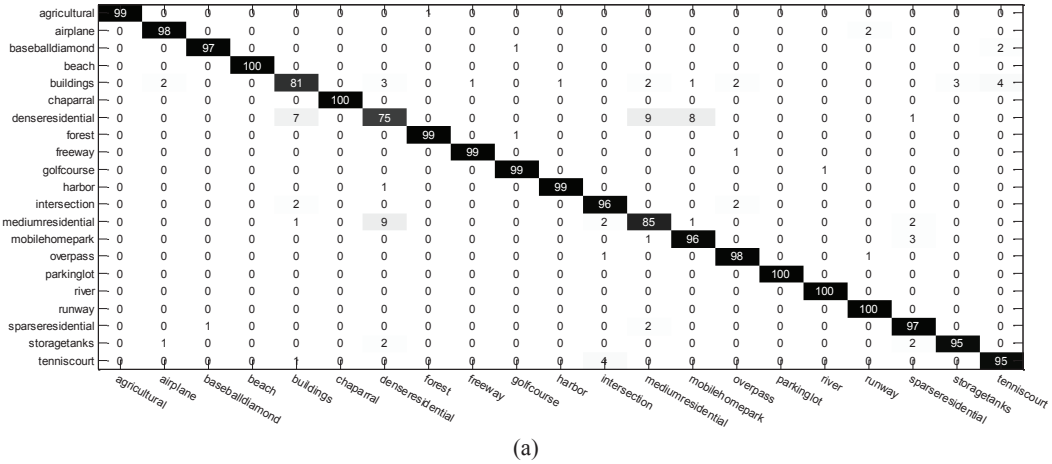
Table 1 shows the overall accuracy of classification results for different remote sensing image feature learning methods. By comparing the CNN-W with CNN-R, we can find that the proposed region-wise deep feature representation method CNN-R achieved better classification results than CNN-W. This can be attributed to the candidate regions capturing more effective local geometric information of the target areas and thus CNN features extracted from these target regions have more powerful discrimination ability and are less influenced by the background. By comparing our CNN-R with state-of-the-art remote sensing scene classification results, we can observe that the performance of CNN-R is still among the top ones, which further validates the effectiveness of the proposed approach.

**Table 1.** Classification accuracy of compared methods on the UC-Merced dataset.

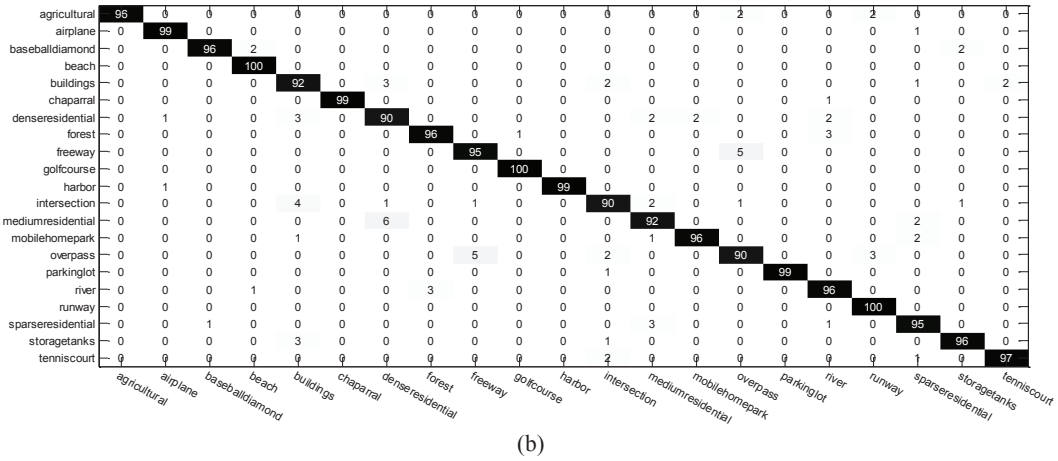
| Method | Year | Accuracy |
|---|---|---|
| SPMK [15] | 2006 | 74% |
| LDA-SVM [2] | 2013 | 80.33% |
| SIFT + SC [18] | 2014 | 81.67% |
| Saliency + SC [19] | 2015 | 82.72% |
| DCGANs [35] (without augmentation) | 2017 | 85.36% |
| MAGANs [35] (without augmentation) | 2017 | 87.69% |
| CaffeNet [26] (without fine-tuning) | 2015 | 93.42% |
| CaffeNet + VLAD [32] | 2015 | 95.39% |
| UCFFN [36] | 2018 | 87.83% |
| WDM [37] | 2017 | 95.71% |
| CNN-W (AlexNet) with SVM | | 95.61% |
| CNN-R (AlexNet) with SVM | | 95.85% |

Figure 5 shows the confusion matrix of the feature representation method CNN-W and our proposed CNN-R on the UC-Merced dataset. From the figure we observe that accuracies above 90% are obtained for all the 21 classes with our proposed CNN-R approach. By comparing CNN-R with CNN-W, significant improvements have been obtained upon the classes “building”, “denseresidential”

and "mediumresidential", where the accuracies have been elevated from 81%, 75%, 85% to 92%, 90%, 92% respectively. The reasons may be that the images in the above three classes have high intraclass variance and low interclass variance while directly extracting deep features from the whole image may not accurately reflect their true category information. In contrast, our proposed method employs the region-wise deep features for image representation, which effectively captures the local geometric invariance and has more discriminative power for remote sensing scene classification.




(a)




(b)

**Figure 5.** The confusion matrix of the classification results based on (**a**) CNN-W and (**b**) CNN-R on the UC-Merced dataset.

### 3.2.2. Results for Large-Scale Remote Sensing Image Retrieval

Hashing-based methods have attracted much attention in handling the large-scale remote sensing image retrieval problem recently [39,45–47]. The hashing methods map the input images from the feature space to a low dimensional code space, i.e., hamming space, where each image is represented by a binary code. The goal of hashing approaches is to generate binary codes for each sample in a database such that similar samples have close codes. One advantage of the binary code representation is significantly reducing the amount of memory required for storing the images' content. In addition, it is extremely fast to perform similarity search over such binary codes for large-scale applications because the hamming distance between binary codes can be efficiently calculated with XOR operation in modern CPU.

In this subsection, we will evaluate the performance of proposed region-wise deep features for hashing-based large-scale remote sensing image retrieval on the AID dataset. We select three state-of-the-art hashing models: kernel supervised hashing (KSH) [48], supervised discrete hashing (SDH) [49], and column sampling based discrete supervised hashing (COSDISH) [50] in the

experiments. CNN-W feature and our proposed CNN-R feature are used as input of the above models respectively to learn binary hash codes. Finally, image retrieval is carried out by comparing hamming distance of the learned codes. The retrieval performance is measured with four widely used metrics: mean average precision (MAP), precision of the top $K$ retrieved images (Precision@K), recall of the top $K$ retrieved images (Recall@K) and precision-recall (P-R) curves. More specifically, the precision and recall are defined as follows:

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{4}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{5}$$

The MAP score is calculated by

$$\text{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{n_i} \sum_{k=1}^{n_i} \text{precision}(R_{ik}) \tag{6}$$

where $q_i \in Q$ is a query, $n_i$ is the number of images relevant to $q_i$ in the database. Suppose the relevant images are ordered as $\{r_i, r_2, \cdots, r_{n_i}\}$, then $R_{ik}$ is the set of ranked retrieval results from the top result until you get to point $r_k$.

Table 2 shows the MAP of different hashing methods for fast image retrieval on the AID dataset based on varied input deep features and hash code length. We observe that nearly all the hashing methods have obtained performance improvements when using our proposed region-wise deep feature representation method for hash code learning. As is illustrated, the KSH, SDH and COSDISH methods have achieved 4%, 4.75% and 7.75% improvement on average when using CNN-R as input feature instead of CNN-W. These results also indicate that learning CNN feature representation from regions is more effective to capture the target information and can generate powerful hash codes for large-scale image retrieval.

**Table 2.** The mean average precision (MAP) of different hashing methods with varied CNN features and code length on AID dataset.

| Method | 8-Bits | 12-Bits | 16-Bits | 24-Bits |
|---|---|---|---|---|
| KSH + CNN-W | 0.35 | 0.45 | 0.48 | 0.55 |
| SDH + CNN-W | 0.52 | 0.63 | 0.67 | 0.46 |
| COSDISH + CNN-W | 0.65 | 0.75 | 0.82 | 0.86 |
| KSH + CNN-R | 0.42 | 0.46 | 0.54 | 0.59 |
| SDH + CNN-R | 0.54 | 0.62 | 0.67 | 0.64 |
| COSDISH + CNN-R | 0.74 | 0.86 | 0.88 | 0.91 |

The precision of the top $K$ retrieved images, recall of the top $K$ retrieved images and precision-recall curves for the compared methods are shown in Figure 6. From the Precision@K curves (Figure 6a,d,g,j) we can find that our CNN-R based hashing methods obtain better results than CNN-W based methods when the retrieved images grow in most cases. The Recall@K scores of different approaches over varied hash bits are shown in Figure 6b,e,h,k, which have shown similar results to Precision@K curves. The above observed results have demonstrated that the hash codes learned by our CNN-R features are more effective than traditional CNN-W features for the large-scale image retrieval task. The P-R curves, which reflect the overall image retrieval performance of different methods, are shown in Figure 6c,f,i,l. By comparing the hashing methods using CNN-R features as input with that using CNN-W, we also find that the proposed CNN-R feature representation still consistently outperforms CNN-W for hashing-based large-scale image retrieval in most cases. This may be because our proposed CNN-R

feature learning scheme is able to generate more informative feature representation for remote sensing images and thus the learned hash codes with CNN-R features are also more accurate to capture the image contents. In fact, MAP score is the area under the precision-recall curve, thus, these detailed results in Figure 6 are consistent with the trends that we observe in the above experiments, which validates the superiority of our CNN-R feature representation strategy.

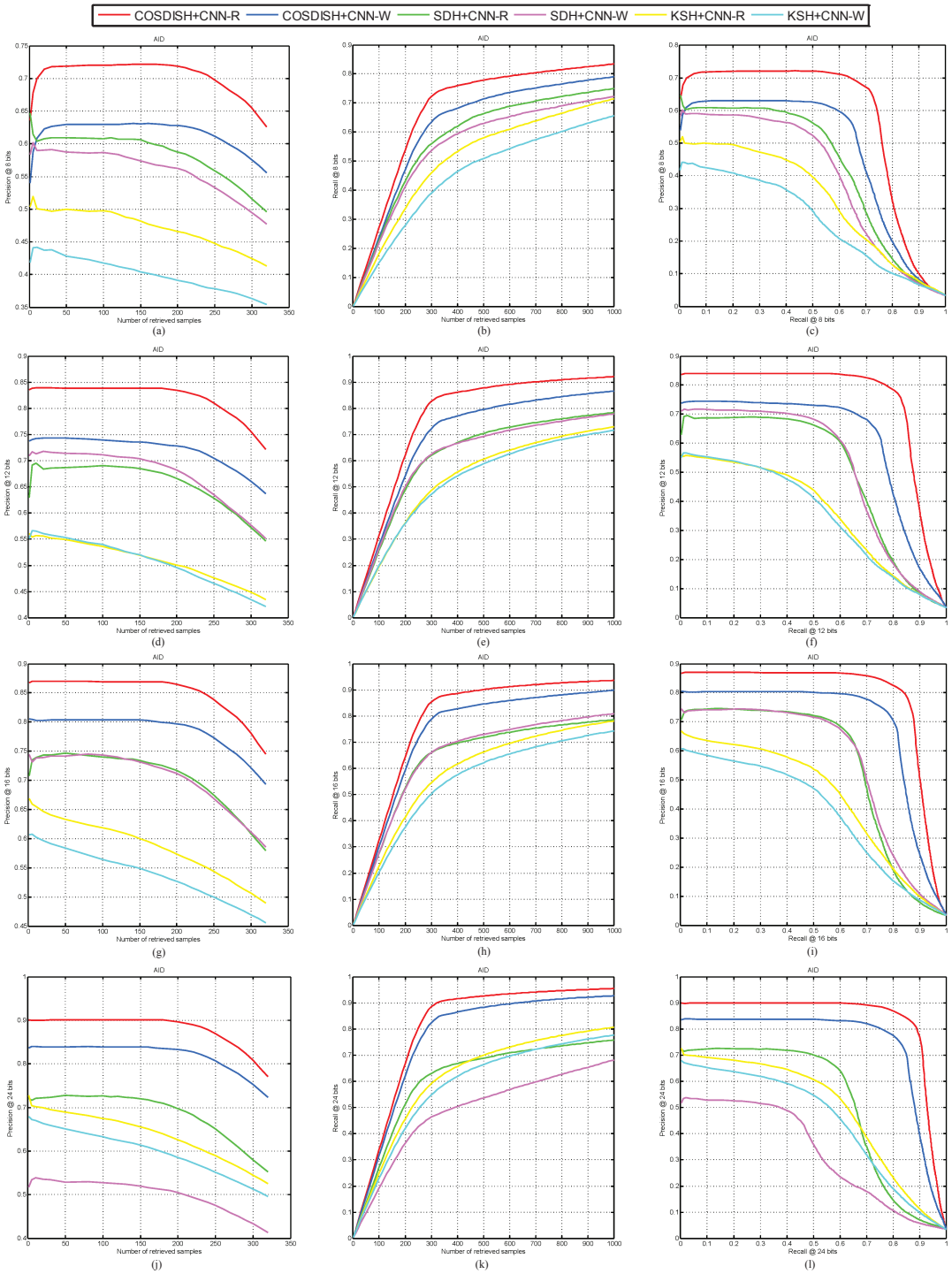



**Figure 6.** The comparison of precision@K, recall@K and precision-recall curves for different hashing methods with varied CNN features and code length on AID dataset ((**a**–**c**) for 8-bits; (**d**–**f**) for 12-bits; (**g**–**i**) for 16-bits; (**j**–**l**) for 24-bits.)

## 4. Future Work

From the above experimental results, we have demonstrated the effectiveness of our proposed region-wise CNN feature representation for remote sensing images. Compared with the traditional CNN features extracted from the whole image, our region-wise CNN feature can keep much more useful information in the final feature representations and thus achieve better performance in remote sensing image classification and retrieval tasks. However, there are also some open issues that remain for future research. For example, the first step of our proposed approach is to locate the target proposals, in which the existing edge-boxes algorithm is directly adopted. The edge-boxes algorithm is a general object proposal method for natural images, which may not be completely suitable for remote sensing targets. Therefore, how to improve the original object proposal algorithm specifically for remote sensing images can be one research direction. Moreover, our proposed feature extraction approach is made up of three individual steps and how to design an end-to-end region-wise deep feature representation for remote sensing images will be another direction for future research.

## 5. Conclusions

In this paper, we have proposed a novel region-wise deep feature representation framework for remote sensing images. In our proposed approach, the target-related bounding boxes are first computed for the candidate regions and a deep CNN model is applied to extract the regional deep features for each image. Then, the regional deep features are encoded into a single feature vector for each image by an improved VLAD algorithm, where a weighted multi-neighbor assignment strategy is proposed to calculate the VLAD representation. The main advantages of the our proposed approach are: (1) representing the images with region-wise deep features is able to capture the local geometric invariance of target information more accurately and retain more specific content information in the final image features. (2) the improved VLAD algorithm takes the region score and visual word assignment weight into consideration when encoding the local regional features and thus can generate more effective unique feature vectors for final image representations. Extensive experiments on two different remote sensing image analysis tasks have demonstrated the superiority of our approach over the traditional feature representing methods.

**Author Contributions:** P.L. and P.R. conceived and designed the experiments; Q.W. and X.Z. (Xiaoyu Zhang) performed the experiments; X.Z. (Xiaobin Zhu) and L.W. analyzed the data; P.L. wrote the paper; All authors read and approved the final manuscript.

**Acknowledgments:** This work was supported in part by the National Natural Science Foundation of China (Grant 61602517, 61501475), in part by the National Key R&D Program of China (Grant 2017YFB1401000), in part by Qingdao Applied Fundamental Research (Grant 16-5-1-11-jch), in part by the Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing (Grant 2017A05), in part by the Open Project Program of National Laboratory of Pattern Recognition (Grant 201800018), and in part by the Fundamental Research Funds for Central Universities (Grant 18CX02110A).

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

## References

1. Du, R.; Chen, Y.; Tang, H.; Fang, T. Study on content-based remote sensing image retrieval. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seoul, Korea, 25–29 July 2005; pp. 707–710. [CrossRef]
2. 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]
3. Rosu, R.; Donias, M.; Bombrun, L.; Said, S.; Regniers, O.; Da Costa, J.-P. Structure tensor Riemannian statistical models for CBIR and classification of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 248–260. [CrossRef]

4. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
5. Ozkan, S.; Ates, T.; Tola, E.; Soysal, M.; Esen, E. Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1996–2000. [CrossRef]
6. Yang, J.; Liu, J.; Dai, Q. An improved Bag-of-Words framework for remote sensing image retrieval in large-scale image databases. *Int. J. Digit. Earth* **2015**, *8*, 273–292. [CrossRef]
7. Dos Santos, J.; Penatti, O.; Da Silva Torres, R. Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In Proceedings of the Fifth International Conference on Computer Vision Theory and Applications (VISAPP), Angers, France, 17–21 May 2010; pp. 203–208.
8. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1349–1362. [CrossRef]
9. Wang, Q.; Wan, J.; Yuan, Y. Locality constraint distance metric learning for traffic congestion detection. *Pattern Recogn.* **2018**, *75*, 272–281. [CrossRef]
10. Aptoula, E. Remote sensing image retrieval with global morphological texture descriptors. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3023–3034. [CrossRef]
11. Newsam, S.; Wang, L.; Bhagavathy, S.; Manjunath, B.S. Using texture to analyze and manage large collections of remote sensed image and video data. *Appl. Opt.* **2004**, *43*, 210–217. [CrossRef] [PubMed]
12. Luo, B.; Aujol, J.F.; Gousseau, Y.; Ladjal, S. Indexing of satellite images with different resolutions by wavelet features. *IEEE Trans. Image Process.* **2008**, *17*, 1465–1472. [CrossRef] [PubMed]
13. Yang, Y.; Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 818–832. [CrossRef]
14. Du, Z.; Li, X.; Lu, X. Local structure learning in high resolution remote sensing image retrieval. *Neurocomputing* **2016**, *207*, 813–822. [CrossRef]
15. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; pp. 2169–2178. [CrossRef]
16. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems (GIS), San Jose, CA, USA, 3–5 November 2010; pp. 270–279. [CrossRef]
17. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1465–1472. [CrossRef]
18. Cheriyadat, A. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
19. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2175–2184. [CrossRef]
20. Zhou, W.; Shao, Z.; Diao, C.; Cheng, Q. High-resolution remotesensing imagery retrieval using sparse features by auto-encoder. *Remote Sens. Lett.* **2015**, *6*, 775–783. [CrossRef]
21. Li, Y.; Zhang, Y.; Tao, C.; Zhu, H. Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion. *Remote Sens.* **2016**, *8*, 709. [CrossRef]
22. Wang, Y.; Zhang, L.; Tong, X.; Zhang, L.; Zhang, Z.; Liu, H.; Xing, X.; Takis Mathiopoulos, P. A three-layered graph-based learning approach for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6020–6034. [CrossRef]
23. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–13.

25. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
26. Penatti, O.; Nogueira, K.; Dos Santos, J. 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 (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 44–51. [CrossRef]
27. Wang, Q.; Wan, J.; Yuan, Y. Deep metric learning for crowdedness regression. *IEEE Trans. Circ. Syst. Video* **2017**. [CrossRef]
28. Wang, Q.; Gao, J.; Yuan, Y. A joint convolutional neural networks and context transfer for street scenes labeling. *IEEE Trans. Intell. Transp.* **2018**, *19*, 1457–1470. [CrossRef]
29. Dutta, R.; Aryal, J.; Das, A.; Kirkpatrick, J.B. Deep cognitive imaging systems enable estimation of continental-scale fire incidence from climate data. *Sci. Rep.* **2013**, *3*, 3188. [CrossRef] [PubMed]
30. Dutta, R.; Das, A.; Aryal, J. Big data integration shows Australian bush-fire frequency is increasing significantly. *R. Soc. Open Sci.* **2016**, *3*, 150241. [CrossRef] [PubMed]
31. Zhang, F.; Du, B.; Zhang, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [CrossRef]
32. Hu, F.; Xia, G.; 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]
33. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
34. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sens.* **2017**, *9*, 489. [CrossRef]
35. 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.* **2017**, *14*, 2092–2096. [CrossRef]
36. Yu, Y.; Gong, Z.; Wang, C.; Zhong, P. An unsupervised convolutional feature fusion network for deep representation of remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 23–27. [CrossRef]
37. Lu, X.; Zheng, X.; Yuan, Y. Remote sensing scene classification by unsupervised representation learning. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5148–5157. [CrossRef]
38. Cheng, G.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1735–1739. [CrossRef]
39. Li, Y.; Zhang, Y.; Huang, X.; Zhu, H.; Ma, J. Large-scale remote sensing image retrieval by deep hashing neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 950–965. [CrossRef]
40. Zitnick, C.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 391–405. [CrossRef]
41. Jégou, H.; Perronnin, F.; Douze, M.; Sanchez, J.; Perez, P.; Schmid, C. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1704–1716. [CrossRef]
42. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
43. Ye, S.; Pontius, R.G.; Rakshit, R. A review of accuracy assessment for object-based image analysis: From per-pixel to per-polygon approaches. *ISPRS J. Photogramm.* **2018**, *141*, 137–147. [CrossRef]
44. Stein, A.; Aryal, J.; Gort, G. Use of the Bradley-Terry model to quantify association in remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 852–856. [CrossRef]
45. Demir, B.; Bruzzone, L. Hashing-based scalable remote sensing image search and retrieval in large archives. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 892–904. [CrossRef]
46. Li, P.; Ren, P. Partial randomness hashing for large-scale remote sensing image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 464–468. [CrossRef]
47. Ye, D.; Li, Y.; Tao, C.; Xie, X.; Wang, X. Multiple feature hashing learning for large-scale remote sensing image retrieval. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 364. [CrossRef]
48. Liu, W.; Wang, J.; Ji, R.; Jiang, Y.; Chang, S. Supervised hashing with kernels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2074–2081. [CrossRef]

49. Shen, F.; Shen, C.; Liu, W.; Shen, H. Supervised discrete hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 37–45. [CrossRef]
50. Kang, W.; Li, W.; Zhou, Z. Column sampling based discrete supervised hashing. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI), Phoenix, AZ, USA, 12–17 February 2016; pp. 1230–1236.