remote sensing

MDPI

Article

A Deep-Local-Global Feature Fusion Framework for High Spatial Resolution Imagery Scene Classification

Qiqi Zhu, Yanfei Zhong *, Yanfei Liu *, Liangpei Zhang and Deren Li

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; wuxi5477@126.com (Q.Z.); zlp62@whu.edu.cn (L.Z.); drli@whu.edu.cn (D.L.)

* Correspondence: zhongyanfei@whu.edu.cn (Y.Z.); yanfeiliu@whu.edu.cn (Y.L.)

Received: 7 March 2018; Accepted: 4 April 2018; Published: 6 April 2018

Abstract: High spatial resolution (HSR) imagery scene classification has recently attracted increased attention. The bag-of-visual-words (BoVW) model is an effective method for scene classification. However, it can only extract handcrafted features, and it disregards the spatial layout information, whereas deep learning can automatically mine the intrinsic features as well as preserve the spatial location, but it may lose the characteristic information of the HSR images. Although previous methods based on the combination of BoVW and deep learning have achieved comparatively high classification accuracies, they have not explored the combination of handcrafted and deep features, and they just used the BoVW model as a feature coding method to encode the deep features. This means that the intrinsic characteristics of these models were not combined in the previous works. In this paper, to discover more discriminative semantics for HSR imagery, the deep-local-global feature fusion (DLGFF) framework is proposed for HSR imagery scene classification. Differing from the conventional scene classification methods, which utilize only handcrafted features or deep features, DLGFF establishes a framework integrating multi-level semantics from the global texture feature–based method, the BoVW model, and a pre-trained convolutional neural network (CNN). In DLGFF, two different approaches are proposed, i.e., the local and global features fused with the pooling-stretched convolutional features (LGCF) and the local and global features fused with the fully connected features (LGFF), to exploit the multi-level semantics for complex scenes. The experimental results obtained with three HSR image classification datasets confirm the effectiveness of the proposed DLGFF framework. Compared with the published results of the previous scene classification methods, the classification accuracies of the DLGFF framework on the 21-class UC Merced dataset and 12-class Google dataset of SIRI-WHU can reach 99.76%, which is superior to the current state-of-the-art methods. The classification accuracy of the DLGFF framework on the 45-class NWPU-RESISC45 dataset, 96.37 $\pm$ 0.05%, is an increase of about 6% when compared with the current state-of-the-art methods. This indicates that the fusion of the global low-level feature, the local mid-level feature, and the deep high-level feature can provide a representative description for HSR imagery.

Keywords: scene classification; deep feature; global low-level features; local feature; BoVW; high spatial resolution image; fusion

1. Introduction

The technology of satellite sensors has provided us with a large amount of high spatial resolution (HSR) images with abundant spectral and spatial information for precise land-cover/land-use (LULC) investigation. HSR images have great potential for a wide range of applications, such as road extraction, urban functional analysis, and landslide mapping. Due to the low interclass disparity and high intra-class variability in HSR images, object-based methods have been one of the most promising

methods for HSR image information extraction [1–3]. However, diverse object classes, e.g., buildings, trees, and roads, with different spatial distributions can usually be found in HSR images. This makes it a challenging task to obtain the semantic information of the whole image scene, e.g., a residential scene or an industrial scene, and leads to the so-called semantic gap [4]. Scene classification plays an important role in solving the semantic gap problem and has recently been receiving a great deal of attention. Cheriyadat [5] employed sparse coding to learn a set of basis functions from the low-level feature descriptors. Li et al. [6] combined zero-shot learning with a label refinement approach based on sparse learning for HSR imagery scene classification, which is an approach that can suppress the noise in the zero-shot classification results.

The scene classification methods first begin by extracting global low-level features as the visual descriptor, e.g., the color histogram, GIST [7], and local binary patterns (LBP) [8]. These features are extracted from the images based on pre-defined algorithms. Among the various scene classification methods, the well-known bag-of-visual-words (BoVW) model is still an important way to extract the visual descriptors of the scenes [9–12]. By mapping the local features to a visual dictionary, a set of visual words is acquired by the BoVW model. The simple statistics of each visual word occurrence are then computed based on the visual dictionary to represent the images, which narrows the gap between the low-level features and the high-level semantics. Based on BoVW, some classical variants, e.g., the probabilistic topic model and feature coding, have been developed to improve the ability to describe complex HSR images [5,13–18]. These methods are largely dependent on the selection of the handcrafted low-level features and the design of the mid-level feature representation, which requires a lot of prior information, and they have limited transferability between different fields and datasets. In addition, the methods using BoVW simply count the occurrences of the local features, and they do not model the spatial relationships in an image. In this way, an image is represented as an orderless collection of local features, and the information about the spatial layout of the features is disregarded, which severely limits the descriptive ability for HSR imagery.

As a branch of machine learning, deep learning has emerged as a powerful tool to automatically discover the intricate structures hidden in high-dimensional data and has been applied to image classification [19], video analysis [20], facial recognition [21,22], drug discovery [23], objection detection [24], and hyperspectral pixel classification [25]. For HSR imagery scene classification, deep learning has also shown an impressive feature representation ability [26]. Convolutional neural networks (CNNs), as the most popular deep learning–based networks, are generally trained on large amounts of labeled training data, e.g., ImageNet [27], and can directly extract features from raw image pixels without prior knowledge. CNNs are biologically inspired multi-stage architectures, which consist of convolutional (Conv), pooling, and fully connected (FC) layers. The pooling process is able to preserve the spatial location of the HSR image, as well as selecting the most useful features obtained by the Conv layer. In addition, CNNs have shown astounding performances in datasets with different characteristics from which they were trained, which confirms the transferability of deep features from one dataset to another [28]. Compared with the previous state-of-the-art handcrafted feature–based methods, CNNs employ an end-to-end feature learning approach and perform better for HSR imagery scene classification [29,30]. However, the methods based on CNNs need a large amount of data for training, and large amounts of training samples are rare for most remote sensing problems. Based on the generalizability of CNNs [31,32], transfer learning–based scene classification methods have been proposed for HSR imagery scene classification. There are two main transfer learning approaches. The first approach uses a pre-trained CNN to partially initialize the transferred CNN [33,34]. This approach utilizes a fine-tuning strategy to improve the discriminative ability of the new set of HSR imagery, which increases the speed of convergence during the training process. The other approach treats the pre-trained CNN as the feature extractor and is an approach that has no network training process, is effective and simple, and leads to a high speed [35–37]. For instance, Hu et al. [35] encoded the CNN activations from the Conv layer via feature coding methods, e.g., locality-constrained linear coding (LLC), the vector of locally aggregated descriptors (VLAD), and the

improved fisher kernel (IFK), by regarding the pre-trained CNN on a large dataset like ImageNet as a feature extractor, instead of the handcrafted low-level feature extraction operation. However, without the training process on HSR images, the unique and representative characteristics of HSR images may be lost, which weakens the classification performance.

In this paper, considering the respective shortcomings of BoVW and CNNs, a novel deep-local-global feature fusion (DLGFF) framework is established for HSR imagery scene classification. In DLGFF, a low-level visual feature–based method, the BoVW model, and a CNN are combined to generate a more discriminative feature representation for HSR imagery. The BoVW model is utilized to generate the mid-level local features, including the mean and standard deviation (MSD)-based spectral feature and the commonly used scale-invariant feature transform (SIFT) feature. The MSD and SIFT features are then quantized separately to circumvent the hard-assignment characteristics of *k*-means clustering. The global shape-based invariant texture index (SITI) [38] is employed as the low-level texture feature to compensate for the local features. To obtain the deep features, a pre-trained CNN is employed, and the convolutional (Conv) feature and FC feature are separately extracted to be fused with the handcrafted features. Due to the difference between handcrafted and deep features, two different approaches are provided in the DLGFF framework. Finally, the fused semantics by DLGFF are classified by a support vector machine (SVM) classifier with a histogram intersection kernel (HIK).

The major contributions of this paper are as follows:

(1) The DLGFF framework is proposed to discover discriminative information from the HSR imagery. Based on the low-level visual information, three heterogeneous handcrafted features and two deep features are designed to capture the complex geometrical structures and spatial patterns of HSR images. The feature spaces of different types of features are separately mined and effectively fused to circumvent feature interruption in the feature learning process (see also Section 3).

(2) In order to capture representative semantics and spatial structures for the scenes, the low-level based SITI texture, the mid-level based BoVW model, and the high-level based CNN scene classification method are first combined in DLGFF. The spatial information and intrinsic information are acquired from the CNN, whereas the unique characteristics of HSR imagery can be extracted from the handcrafted feature–based methods. The integration of the low-level, mid-level, and deep features provides a multi-level feature description for distinct scenes (see also Sections 3.1 and 3.2).

(3) Two approaches are proposed in the DLGFF framework, i.e., the local and global features fused with the fully connected features (LGFF) and the local and global features fused with the pooling-stretched convolutional features (LGCF). To make full use of the spatial information and intrinsic information in the intermediate Conv layers and FC layer, the DLGFF framework provides an effective strategy for the BoVW features to complement the deep features from different layers (see also Section 3.3).

Comprehensive evaluations on three distinct datasets, i.e., the 21-class UC Merced dataset, the 12-class Google dataset of SIRI-WHU, and the challenging 45-class NWPU-RESISC45 dataset, confirm the effectiveness of the DLGFF framework. In addition, compared with the published results of the previous scene classification methods, LGCF and LGFF obtain state-of-the-art results, further confirming the superiority of the DLGFF framework.

The rest of this paper is organized as follows. In Section 2, scene classification based on the BoVW model and CNNs is described in detail. Section 3 describes the proposed DLGFF framework for HSR imagery scene classification. A description of the datasets and a discussion of the experimental results are presented in Section 4. Finally, the conclusions are drawn in Section 5.

2. Background

2.1. Scene Classification Based on the BoVW Model

The BoVW model was first proposed for text document analysis and has since been widely applied in image interpretation, as a result of the similarity between text analysis and image processing. Due to its simplicity and good performance, the BoVW model has been a popular approach for content-based image classification. This model considers the image as a collection of several independent words, without grammar or word order. The "words" in the scene images are defined as the feature vectors of the image patches, which are obtained by a sampling method. Given a dataset consisting of I images, each image can be described by a set of N visual words w_n from a visual dictionary. The feature vectors are then transformed to 1D frequency vectors, where each element denotes the occurrence number of a visual word in an image. By extracting the local features of the scenes, scene classification based on BoVW maps the local low-level features to the corresponding parameter space to obtain the mid-level features. These mid-level features are called the "bags of visual words". However, the methods based on the BoVW model disregard the spatial relationship of the local features. The procedure of scene classification based on the BoVW model is shown in Figure 1.

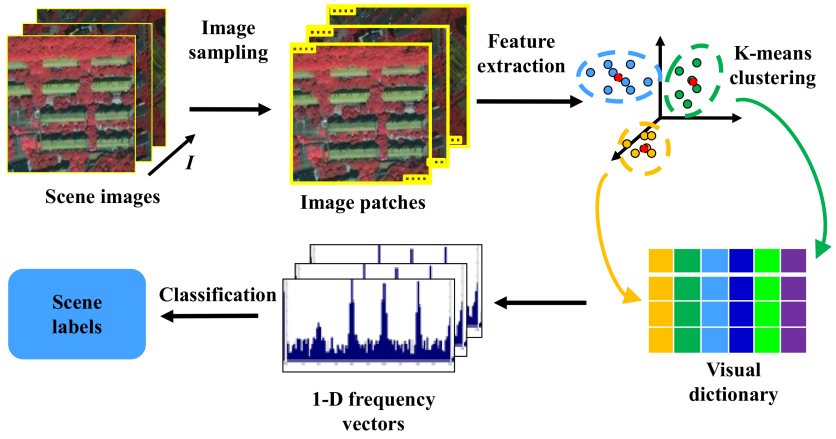


Figure 1. The procedure of scene classification based on the bag-of-visual-words (BoVW) model.

2.2. Scene Classification Based on a CNN

As popular representation learning models, CNNs can directly extract features from raw data without prior knowledge, and have been widely used in many tasks, including HSR imagery scene classification. A typical CNN consists of Conv layers, pooling layers, FC layers, and a Softmax layer, which are stacked layer by layer. The common flowchart of scene classification based on a CNN is shown in Figure 2, which can be divided into three parts: pre-processing, feature extraction, and feature classification.

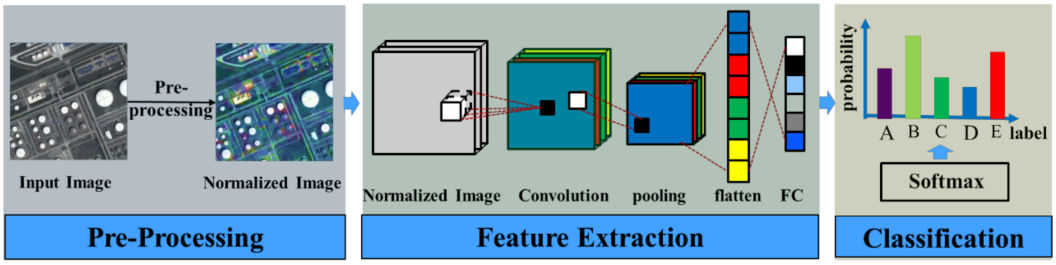


Figure 2. Scene classification based on a convolutional neural network.

In the pre-processing phase, the input image is divided by the maximum pixel value (e.g., 255 in an RGB image) and a normalized image is obtained. After the pre-processing, the CNN is applied to extract the features for the scene classification. We take the CNN shown in Figure 2 to make a brief introduction to this process. Given the normalized image I_n, Conv computation is conducted by sliding a Conv filter on the Conv layer, obtaining feature map C_1. A pooling operation is then applied to compute the local maximum value or mean value of the region pixels in each channel of C_1, and thus features which are robust to noise and clutter and invariant to image transformation are obtained [39]. As shown in Figure 2, the output of the pooling layer is then flattened into a feature vector and fed into the FC layer to further extract features of the whole image for the scene classification. In the classification phase, the feature vectors extracted from the CNN are classified by the Softmax layer. The Softmax layer is a generalization of logistic regression for a multi-class problem, giving the possibility of a sample belonging to each class. Compared with the low-level and mid-level based scene classification methods, scene classification methods based on CNNs are simple, do not require prior knowledge, and perform better. However, the methods based on deep networks need a considerable amount of data and a lot of computational power for training, leading to difficulty in training the CNNs. Hence, it is common to use a pre-trained network, either as a fixed feature extractor for the task of interest or as an initialization for fine-tuning the parameters [40].

3. The DLGFF Framework for HSR Imagery Scene Classification

In this paper, the effective DLGFF framework is carefully designed to improve the performance of scene classification for HSR images. Three tasks need to be addressed for HSR imagery scene classification: (1) local and global feature generation, (2) Conv and FC feature generation by CNN, and (3) feature fusion and classification based on LGCF and LGFF. The flowchart of the proposed DLGFF framework is shown in Figure 3.

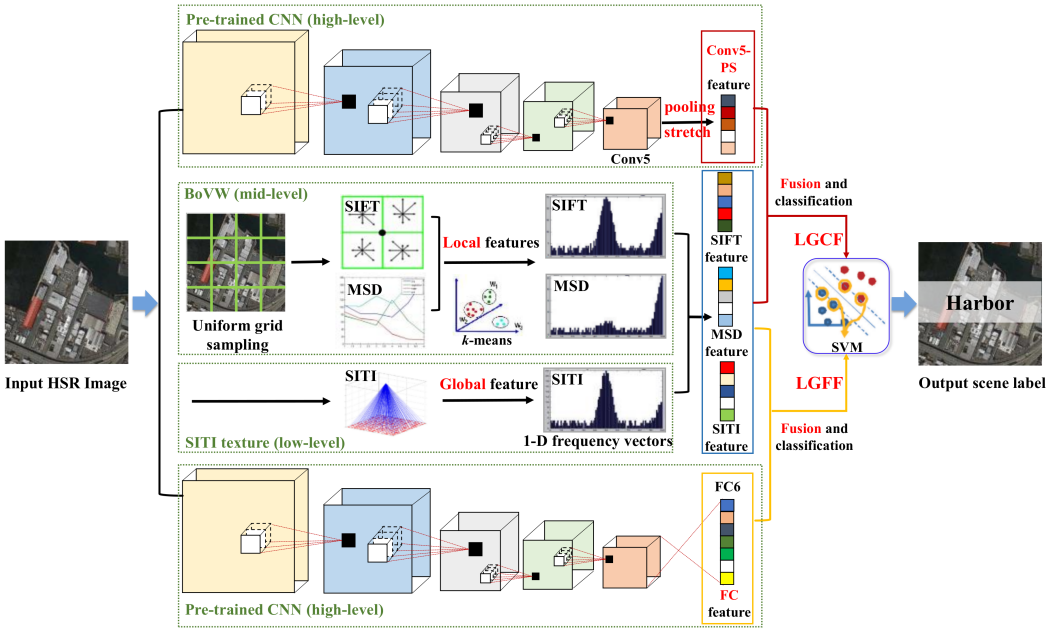


Figure 3. Flowchart of scene classification based on the deep-local-global feature fusion (DLGFF) framework for high spatial resolution (HSR) images, where the descriptions in red indicate the main contributions of the DLGFF framework.

3.1. Local and Global Feature Generation

To acquire the representative handcrafted visual descriptors of the HSR images, the low-level global feature and the BoVW-based local features are combined in DLGFF. SITI, MSD, and SIFT are

employed as the global texture feature, local spectral feature, and local structure feature, respectively. For the local MSD and SIFT features, uniform grid sampling is first employed to obtain uniform image patches from the HSR image. The image patches are then digitized by K types of local feature descriptors. On the other hand, the global feature descriptors are directly computed from the global scale of the image.

SITI was first proposed for texture image retrieval and classification [38]. For HSR images, SITI is acquired according to the complete set of level lines in the images, i.e., the topographic map. Captured from the local keypoints, the commonly used SIFT feature is based on edges, whereas the shapes are the basic elements on which SITI is performed. The shapes are the interiors of the connected components of the level lines. SITI is first computed by extracting the shape descriptors, i.e., the compactness histogram (CpH), the elongation histogram (EH), the contrast histogram (CtH), and the scale ratio histogram (SRH), according to Equations (1) and (2). In Equation (1), λ_1 and λ_2 are the eigenvalues of the inertia matrix. In Equation (2), $\mu_{00}(s)$ is the area of the shape s, $\langle\bullet\rangle_{s'\in\mathbb{N}^M}$ is the mean operator on $\mathbb{N}^M$, and $\mathbb{N}^M$ is the partial neighborhood of order M of s. At each pixel x, $s(x)$ is the smallest shape of the topographic map containing x, and $\mathrm{mean}_{s(x)}(u)$ and $\mathrm{var}_{s(x)}(u)$ are, respectively, the mean and variance of u over $s(x)$. The parameters in our experiments with the SITI feature were set according to the recommendations of the authors [38].

$$Compactness: \quad k = \frac{1}{4\pi\sqrt{\lambda_1\lambda_2}}, \quad Elongation: \quad \in = \lambda_2/\lambda_1 \tag{1}$$

$$Contrast: \quad \gamma(x) = \frac{\mu(x) - \mathrm{mean}_{s(x)}(u)}{\sqrt{\mathrm{var}_{s(x)}(u)}}, \quad Scale\ ratio: \quad \alpha(s) = \frac{\mu_{00}(s)}{\langle\mu_{00}(s')\rangle_{s'\in\mathbb{N}^M}} \tag{2}$$

For HSR images, the spectral feature reflects the attributes that constitute the ground components and structures. The MSD-based spectral feature refers to the first- and second-order statistics of the patches, i.e., the mean and standard deviation values, which are computed in each spectral channel. We let I be the number of pixels in the sampled patch, and b_{ij} denotes the j-th band value of the i-th pixel in a patch. In this way, the mean (M_j) and standard deviation (SD_j) of the spectral vector of the patch are then acquired through Equation (3).

$$M_j = \frac{\sum_{i=1}^{I} b_{ij}}{I}, \qquad SD_j = \sqrt{\frac{\sum_{i=1}^{I} (b_{ij} - M_j)^2}{I}} \tag{3}$$

The SIFT feature [41] is invariant to changes in illumination and affine transformation and has been widely applied in image analysis. Inspired by the work of Fei-Fei and Perona [42], i.e., dense features perform better for scene classification, the DLGFF framework utilizes gray dense SIFT as the patch descriptor. The image patches are divided into 4×4 neighborhood regions, where the gradient orientation histograms of eight directions are counted. We then acquire $4 \times 4 \times 8 = 128$-dimension vectors to describe the keypoint.

After acquiring the local and global low-level features, DLGFF utilizes a visual analog of a word, acquired by vector quantizing the local features. The same visual word in different images may be endowed with different feature values, due to the influence of rotation, illumination, and scale variation. Hence, the local features are quantized by k-means clustering to generate the visual words. In this way, the image patches with similar feature values correspond to the same visual word. By statistical analysis of the frequency of each visual word, the corresponding 1D frequency vector with a size of D_1 and D_2 for the MSD and SIFT features can be obtained, respectively. The global SITI feature can be directly used to represent the 1-D frequency vector of the HSR image scene with a size of D_3. Specifically, given I images, they can each be described by $D_1 + D_2 + D_3$ 1D frequency vectors by fusing the local and global features to obtain the representative handcrafted features.

3.2. Conv and FC Feature Generation by CNN

To obtain the deep features, the DLGFF framework employs a pre-trained CNN, i.e., the commonly used Convolutional Architecture for Fast Feature Embedding (Caffe) [43], as a fixed feature extractor to generate the deep features. Caffe is a fully open-source framework that affords clear, modifiable, and easy implementations for effectively training and deploying general-purpose CNNs and other deep models. The pre-trained CNN provided by Caffe, i.e., CaffeNet, is almost a replication of AlexNet [19], with two important modifications: (1) training is undertaken without data augmentation; and (2) the order of the pooling and normalization layers is exchanged, and pooling is undertaken before normalization in CaffeNet. CaffeNet is obtained using the same dataset as the ILSVRC 2012 competition [44] and basically the same parameters as Krizhevsky's network [19]. CaffeNet allows feature extraction for any layer of the network. In the DLGFF framework, the 4D Conv5 feature from the last Conv layer is employed, in order to make full use of the dense information in the intermediate Conv layers. The first FC layer (FC6) is extracted as the FC feature for the HSR images to capture the high-dimensional information in the CNN. The dense Conv5 feature has a size of $x \times y \times D_4 \times M$, consisting of a feature map measuring $x \times y$ for each image, where each pixel is described by F features. The features of the FC layer mainly describe the spatial layout information, with D_5 dimension features for each image.

3.3. Feature Fusion and Classification Based on LGCF and LGFF

In the task of feature fusion, the global feature, the BoVW-based local features, the Conv5 feature, and the FC feature are denoted as F_g, F_l, F_c, and F_f, respectively. The feature values of the local features acquired by BoVW can reach thousands, whereas the feature values of the Conv5 feature are often much smaller, e.g., 0.02. Hence, in the LGCF approach, an average-pooling operation is first employed. The Conv5 feature is transferred to a 2D feature, with $D_4 \times M$ dimension. In addition, a feature stretching strategy is adopted to solve the problem. The Conv5 feature is stretched to 0–255, which can be denoted as F_{c-PS}, to arrive at a better fusion for the deep feature and the BoVW feature. In this way, the final LGCF representation can be denoted as $LGCF = \left\{F_l, F_g, F_{c-PS}\right\}^{\mathrm{T}}$. As the dimension and structure of the FC features are similar to those of the local and global features, the FC features can be directly fused with the local and global features. Hence, in the LGFF approach, a more discriminative representation of the image can be denoted as $LGCF = \left\{F_l, F_g, F_f\right\}^{\mathrm{T}}$.

In the task of LGCF and LGFF classification, LGCF and LGFF with discriminative semantics are classified by the SVM classifier with a HIK [45]. By measuring the degree of similarity between two histograms, the HIK deals with the scale change, and has been applied to image classification using color histogram features. Assuming $\widetilde{\mathrm{V}} = (\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \ldots, \widetilde{\mathbf{v}}_M)$ and $\widetilde{\mathrm{X}} = (\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots, \widetilde{\mathbf{x}}_M)$ to be the LGCF and LGFF representation vectors of M images, the HIK is calculated as written in Equation (4). Finally, the scene label of each image can be predicted by the two different approaches.

$$K(\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{v}}_j) = \sum_k \min(\widetilde{\mathrm{v}}_{i,k}, \widetilde{\mathrm{v}}_{j,k}) \tag{4}$$

4. Experiments and Analysis

4.1. Experimental Setup

In order to test the performance of the DLGFF framework, the commonly used 21-class UC Merced dataset, the 12-class Google dataset of SIRI-WHU, and the challenging large-scale NWPU-RESISC45 dataset were evaluated in the experiments. All the parameters in the experiments were set based on preliminary experiments. In the experiments with uniform grid-based region sampling of BoVW, the patch size and spacing were set to 8×8 pixels and 4×4 pixels for the spectral features with the three datasets, respectively. The patch size and spacing for the SIFT feature were set to 16×16 pixels and 8×8 pixels, respectively, for the UC Merced dataset and NWPU-RESISC45 dataset, and 8×8 pixels

and 4 × 4 pixels for the Google dataset of SIRI-WHU. The visual dictionary for BoVW with V visual words was constructed by employing Euclidean distance measurement–based k-means clustering over the image patches from the training data. The increase of the dictionary size leads to slight fluctuation of the accuracy and may result in much higher dimensional features for BoVW, which needs more time and space consumption. Hence, in the experiments, the visual word number V was set to 1000 for the spectral and SIFT features and 350 for the SITI feature for the three datasets. The images in the three datasets were all resized to 227 × 227 for CaffeNet, giving consideration to its pre-defined size requirement for the input image. The experiments with the low-level and mid-level methods were run on a personal computer with a single Intel core i3 CPU, an NVIDIA Quadro 600 GPU, and 8 GB of RAM. The operating system was Windows 10, and the implementation environment was under MATLAB 2012a. The computation load of the low-level and mid-level methods equated to 22–30% of the CPU and 500–1000 MB of the memory. The experiments with the high-level methods were performed on a personal computer equipped with dual Intel Xeon E5-2650 v2 processors, a single Tesla K20m GPU, and 64 GB of RAM, running Centos 6.6 with the CUDA 6.5 release. The different methods were implemented 20 times by randomly selecting the training samples, to ensure that convincing results were obtained and the stability of the proposed DLGFF could be tested. The running time of the high-level methods can be ignored, and the running time of the low-level and mid-level methods was about 30 min.

In Tables 1–3, SITI denotes scene classification using the global SITI texture feature, BoVW-MSD denotes BoVW-based scene classification utilizing the MSD feature, and LGFBOVW denotes the combination of BoVW-based scene classification utilizing the MSD and SIFT features with the global SITI feature. CAFFE-CONV5 and CAFFE-FC6 denote pre-trained CaffeNet-based scene classification utilizing the last Conv feature and the first FC feature, respectively, with the SVM classifier for HSR images. LGCF and LGFF denote the two proposed approaches in the DLGFF framework. To further evaluate the performance of DLGFF, the experimental results obtained with the UC Merced dataset, as published in the latest papers by Yang and Newsam [10], Chen and Tian [9], Zhao et al. [46], Castelluccio et al. [33], Zhao et al. [47], Li et al. [37], Nogueira et al. [40], and Zhu et al. [18], are shown for comparison. We also provide the experimental results obtained for the Google dataset of SIRI-WHU by Zhao et al. [46] and Zhu et al. [18] and the results obtained for the NWPU-RESISC45 dataset by Cheng et al. [26] Liu et al. [34], and Cheng et al. [36] for comparison with the proposed LGCF and LGFF approaches.

4.2. Experiment 1: The UC Merced Image Dataset

The UC Merced dataset (http://vision.ucmerced.edu/datasets/landuse.html) was downloaded from the USGS National Map Urban Area Imagery collection [10]. This dataset consists of 21 land-use scenes (Figure 4), namely, 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 courts. Each class contains 100 images, measuring 256 × 256 pixels, with a 1-ft spatial resolution. Following the experimental setup published in Yang and Newsam [10], 80 samples were randomly selected per class from the UC Merced dataset for training, and the rest were kept for testing.

Figure 4. Example images from the UC Merced dataset.

The classification performance of the single feature based SITI, BoVW-MSD, the conventional LGFBOVW, CAFFE-CONV5, CAFFE-FC6, the proposed LGCF and LGFF, and the experimental results of previous methods for the UC Merced dataset are listed in Table 1. As can be seen in Table 1, LGFBOVW outperforms SITI and BoVW-MSD, which indicates that the fusion of the local and global features at the visual word level, is effective. The experimental results of CAFFE-CONV5 and CAFFE-FC6 are comparable to those of LGFBOVW, which confirms the powerful feature learning ability of the deep networks. The classification accuracies for the proposed LGCF and LGFF, 99.52 $\pm$ 1.38% and 99.76 $\pm$ 0.06%, respectively, are the best among all the different methods. This indicates that the combination of the local, global, and deep features is able to provide discriminative image representations for scene classification. LGFF performs slightly better than LGCF, which infers that the FC feature and handcrafted feature are more complementary to describe HSR images. In addition, it can be seen that LGCF and LGFF perform better than the current state-of-the-art methods, i.e., the mid-level feature based methods such as the methods of Yang and Newsam [10], Chen and Tian [9], Zhu et al. [18], Zhao et al. [46], and Zhao et al. [47] and the deep learning based methods, i.e., the methods of Castelluccio et al. [33], Li et al. [27], and Nogueira et al. [40].

Table 1. Overall classification accuracy (%) comparison with the UC Merced dataset.

Method	Classification Accuracy (%)
SITI	81.54 $\pm$ 1.46
BoVW-MSD	85.30 $\pm$ 1.67
LGFBOVW	96.88 $\pm$ 1.32
CAFFE-CONV5	95.34 $\pm$ 0.74
CAFFE-FC6	95.89 $\pm$ 0.74
Yang and Newsam [10]	81.19
Chen and Tian [9]	89.10
Zhao et al. [46]	92.92 $\pm$ 1.23
Zhao et al. [47]	91.67 $\pm$ 1.70
Castelluccio et al. [33]	97.10
Zhu et al. [18]	95.71 $\pm$ 1.01
Li et al. [37]	98.81 $\pm$ 0.38
Nogueira et al. [40]	99.47 $\pm$ 0.50
LGCF	99.52 $\pm$ 0.38
LGFF	99.76 $\pm$ 0.06

An overview of the performance of LGCF and LGFF is shown in the confusion matrices in Figures 5 and 6, respectively. As can be seen in Figures 5 and 6, most of the scene categories can be fully recognized by LGCF. All of the scene categories can be fully recognized by LGFF, except for the tennis court scene. Compared to the confusion matrix of LGCF, the scene categories in the confusion matrix of LGFF obtain a better performance. For example, the baseball diamond and freeway scenes, which are confused in LGCF, are fully recognized by LGFF. The tennis court scene is misclassified by both LGCF and LGFF, which may be due to the high variability of the tennis court scene. For example, the tennis court may be surrounded by residential building, road, vegetation, or idle, and usually occupies a small area in the whole image scene. In addition, the scale of the tennis court varies a lot. These characteristics lead to the misclassification of the tennis court scene.

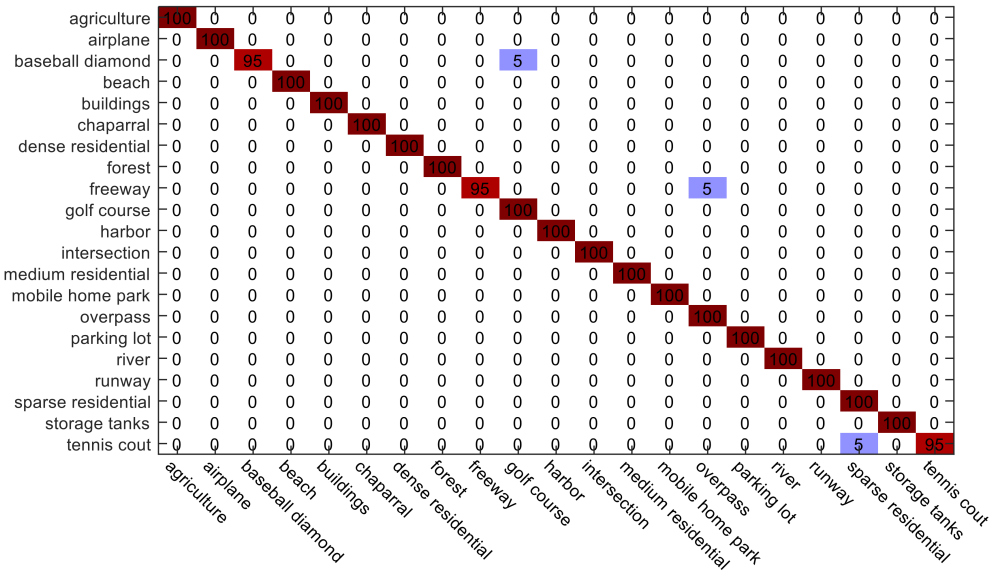


Figure 5. Confusion matrix of local and global features fused with the pooling-stretched convolutional features (LGCF) with the UC Merced dataset.

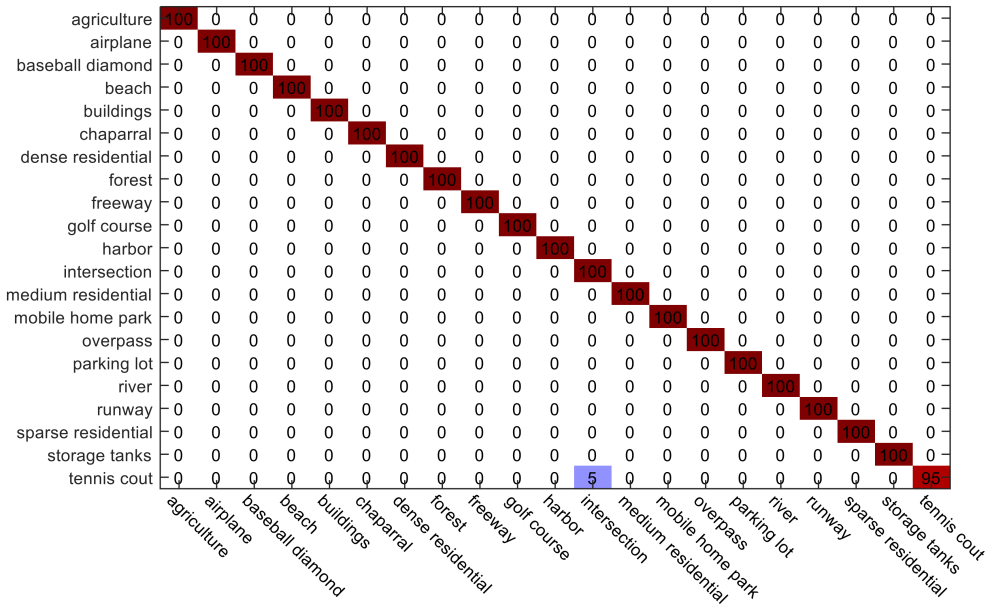


Figure 6. Confusion matrix of local and global features fused with the fully connected features (LGFF) with the UC Merced dataset.

4.3. Experiment 2: The Google Dataset of SIRI-WHU

The Google dataset of SIRI-WHU (http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html) was acquired Mountain View., Mountain View, CA, USA), covering urban areas in China, and the scene image dataset was designed by the Intelligent Data Extraction and Analysis of Remote Sensing (RS_IDEA) Group of Wuhan University (SIRI-WHU) [12,18,46]. The dataset consists of 12 land-use classes, which are labeled as follows: agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water, as shown in Figure 7. Each class separately contains 200 images, which are cropped to 200 × 200 pixels, with a spatial resolution of 2 m. In this experiment, 100 training samples were randomly selected per class from the Google dataset of SIRI-WHU, and the remaining samples were retained for the testing.

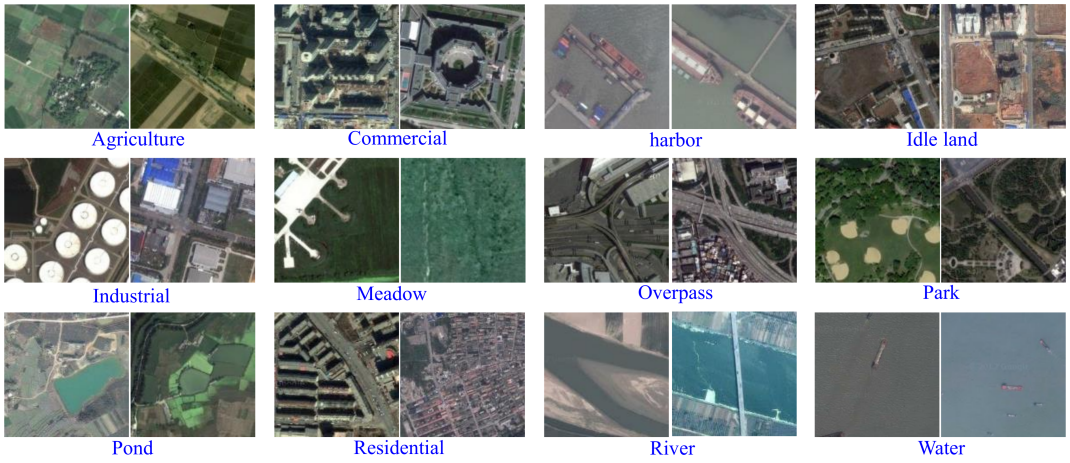


Figure 7. Example images from the Google dataset of SIRI-WHU.

The classification performance of the single feature based SITI, BoVW-MSD, the conventional LGFBOVW, CAFFE-CONV5, CAFFE-FC6, the proposed LGCF and LGFF, and the experimental results of previous methods for the Google dataset of SIRI-WHU are listed in Table 2. As can be seen in Table 2, the classification results of BoVW-MSD are slightly better than those of SITI, which indicates that the mid-level feature performs better than the global low-level feature for HSR imagery scene classification. LGFBOVW also outperforms SITI and BoVW-MSD, which confirms the effectiveness of the local and global feature description for HSR images. The experimental results of CAFFE-CONV5 and CAFFE-FC6 are poorer than those of LGFBOVW, which indicates that the handcrafted features may sometimes be more effective than the deep features for certain HSR image datasets, especially the small datasets. The classification accuracies of the proposed LGCF and LGFF, 99.67 ± 0.22% and 99.75 ± 0.08%, respectively, are better than the results of SITI, BoVW-MSD, LGFBOVW, CAFFE-CONV5, and CAFFE-FC6, which confirms that the DLGFF framework is an effective approach for HSR imagery scene classification. In Table 2, compared to the experimental results reported by Zhao et al. [46], Zhao et al. [47], and Zhu et al. [18], the highest accuracy is acquired by the proposed LGCF and LGFF.

Figures 8 and 9 display the confusion matrices of LGFF and LGCF, respectively, for the Google dataset of SIRI-WHU. On the whole, most of the scene classes achieve good classification performances with LGFF and LGCF. Compared to the confusion matrix of LGCF, the performances of the meadow and river scenes are improved. There is, however, some confusion between certain scenes. For instance, the scenes belonging to the river and pond scenes, respectively, are both classified as the harbor scene. This can be explained by the fact that the three categories are all mainly composed of the same object type, i.e., water.

Table 2. Overall classification accuracy (%) comparison with the Google dataset of SIRI-WHU.

Method	Classification Accuracy (%)
SITI	79.23 ± 1.01
BoVW-MSD	86.51 ± 0.92
LGFBOVW	96.96 ± 0.95
CAFFE-CONV5	93.14 ± 0.82
CAFFE-FC6	91.79 ± 0.75
Zhao et al. [46]	91.52 ± 0.64
Zhao et al. [47]	90.86 ± 0.85
Zhu et al. [18]	97.83 ± 0.93
LGCF	99.67 ± 0.22
LGFF	99.75 ± 0.08

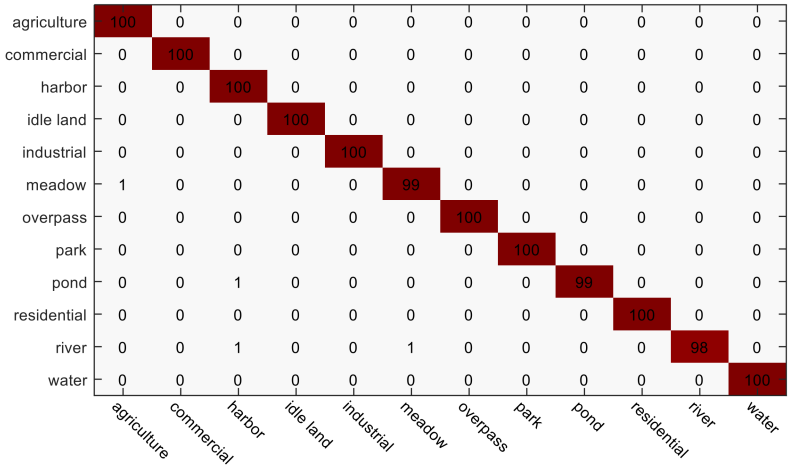


Figure 8. Confusion matrix of LGCF with the Google dataset of SIRI-WHU.

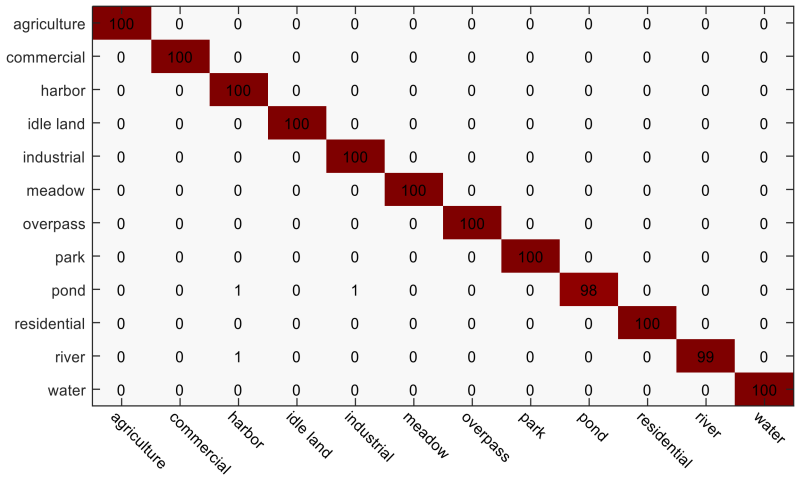


Figure 9. Confusion matrix of LGFF with the Google dataset of SIRI-WHU.

4.4. Experiment 3: The NWPU-RESISC45 Dataset

The NWPU-RESISC45 dataset (http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html) was acquired from Google Earth (Google Inc.), covering more than 100 countries and regions all over the world, including developing, transitional, and highly developed economies [26]. The NWPU-RESISC45 dataset consists of 31,500 remote sensing images divided into 45 scene classes, namely, airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, residential, desert, forest, freeway, golf course, ground track

field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland, as shown in Figure 10. Each class separately contains 700 labeled images, which are cropped to 256 × 256 pixels. The spatial resolutions for most of the scene classes vary from about 30 m to 0.2 m, except for the island, lake, mountain, and snowberg scene classes, which have lower spatial resolutions.

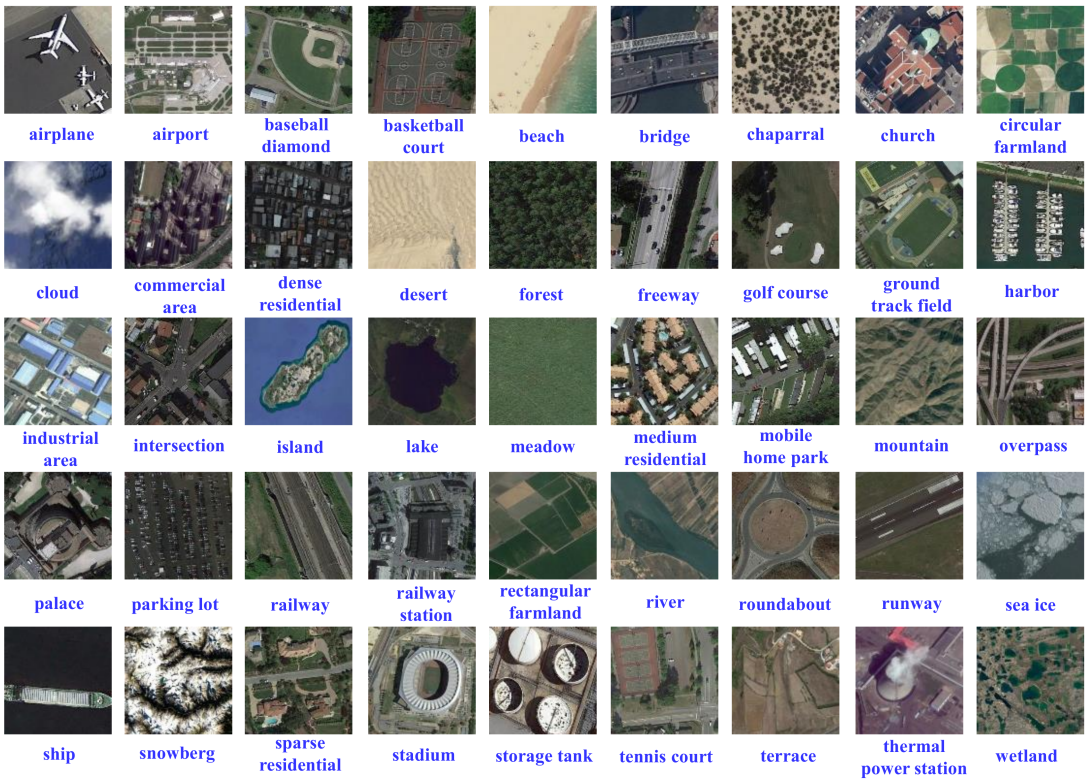


Figure 10. Example images from the NWPU-RESISC45 dataset.

The classification performance of the single feature based SITI, BoVW-MSD, the conventional LGFBOVW, CAFFE-CONV5, CAFFE-FC6, the proposed LGCF and LGFF, and the experimental results of previous methods for the NWPU-RESISC45 dataset are listed in Table 3. As can be seen in Table 3, under the training sample proportions of 10% and 20%, the classification results of the deep feature based methods, i.e., CAFFE-CONV5 and CAFFE-FC6, are better than those of the handcrafted feature based methods, which demonstrates that the deep networks are better able to capture the discriminative representations for a dataset with a large amount of images. The classification accuracies of the proposed LGCF and LGFF, under the training sample proportions of 10% and 20%, respectively, are the best among the different methods, i.e., SITI, BoVW-MSD, LGFBOVW, CAFFE-CONV5, CAFFE-FC6, and the experimental results reported by Cheng et al. [26], Liu and Huang [34], and Cheng et al. [36], which demonstrates the superiority of the DLGFF framework when compared with the previous state-of-the-art methods for HSR imagery scene classification. The classification results of LGFF are about 3% higher than those of LGCF under the training sample proportions of both 10% or 20%. This indicates that the fusion of the FC feature and the local and global features provides a more semantic and robust representation than the fusion of the Conv feature and the local and global features.

Table 3. Overall classification accuracy (%) comparison with the NWPU-RESISC45 dataset.

Method	Training Sample Proportion	
	10%	20%
SITI	53.26 ± 0.45	58.94 ± 1.12
BoVW-MSD	57.90 ± 0.17	63.59 ± 0.18
LGFBOVW	74.78 ± 0.28	81.67 ± 0.22
CAFFE-CONV5	76.90 ± 0.22	80.19 ± 0.21
CAFFE-FC6	78.65 ± 0.36	81.27 ± 0.39
Cheng et al. [26]	87.15 ± 0.45	90.36 ± 0.18
Cheng et al. [36]	82.65 ± 0.31	84.32 ± 0.17
Liu and Huang [34]		92.33 ± 0.20
LGCF	91.07 ± 0.17	94.10 ± 0.09
LGFF	93.61 ± 0.10	96.37 ± 0.05

To prove the contribution of the different modules for the DLGFF framework, the experimental results of the global, local, deep method, and the combinations of these methods for the NWPU-RESISC45 dataset under the ratio of 20% are displayed in Table 4. In Table 4, MSBoVW denotes the local BoVW method using the MSD and SIFT features. Global-local, global-deep, and deep-local denote the fusion of the global and local methods, global and deep methods, and deep and local methods, respectively. As can be seen from Table 4, the experimental results of the two-module based methods, i.e., the global-local, global-deep, and deep-local methods, are better than those of the single-module based methods, i.e., SITI (local), MS (BoVW), and CAFFE-FC6 (deep). In addition, the proposed LGFF fusing three modules performs better than the two-module based methods. This demonstrates that each module in the DLGFF framework does indeed improve the performance of scene classification.

The false-positive rate and false-negative rate of CAFFE-FC6, LGFBOVW, and LGFF are reported in Table 4, to further describe the contribution of the different modules. As can be seen from Table 4, the false-positive rate and false-negative rate of CAFFE-FC6 are slightly higher than those of LGFBOVW, whereas the false-positive rate and false-negative rate of LGFF are much lower than those of CAFFE-FC6 and LGFBOVW. This indicates that the deep module acts as a complement to the global and local modules, and their combination is able to improve the accuracy by reducing both the false-positive rate and false-negative rate.

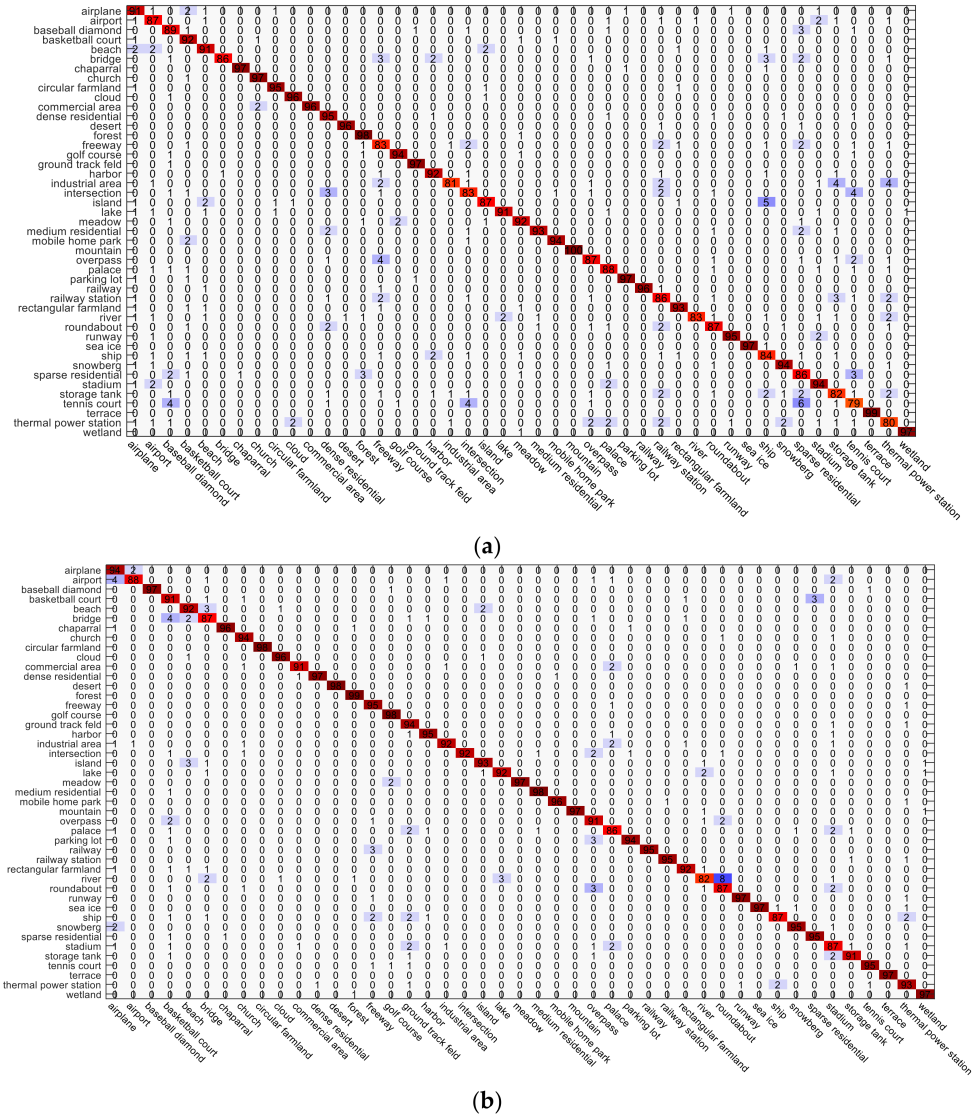
Table 4 can also be used to analyze how much the global, local, and deep features contribute to the DLGFF framework. From Table 4, it can be seen that the classification results of global-local, global-deep, and deep-local are 81.67%, 89.47%, and 94.73%, respectively. Hence, the deep features contribute about 8% more than the local features, the local features contribute about 5% more than the global features, and the deep features contribute about 13% more than the global features. In this way, it can be calculated that for the score of 96.37% for the DLGFF framework, the global, local, and deep features contribute about 26%, 31%, and 39%, respectively.

Table 4. Overall classification accuracy (%) of the different methods for the NWPU-RESISC45 dataset under the training sample proportion of 20%.

Method	Classification Accuracy (%)	False-Positive Rate	False-Negative Rate
SITI (global)	58.94 ± 1.12		
MSBoVW (local)	71.98 ± 0.68		
CAFFE-FC6 (deep)	81.27 ± 0.39	0.0042	0.1852
LGFBOVW (global-local)	81.67 ± 0.27	0.0041	0.1819
Global-deep	89.47 ± 0.31		
Deep-local	94.73 ± 0.23		
LGFF (deep-local-global)	96.37 ± 0.05	0.0008	0.0363

The confusion matrices obtained by LGCF and LGFF for the NWPU-RESISC45 dataset under the training sample proportions of 10% and 20% are shown in Figures 11 and 12, respectively. Compared to the confusion matrix of LGCF, the scene categories in the confusion matrix of LGFF show a better

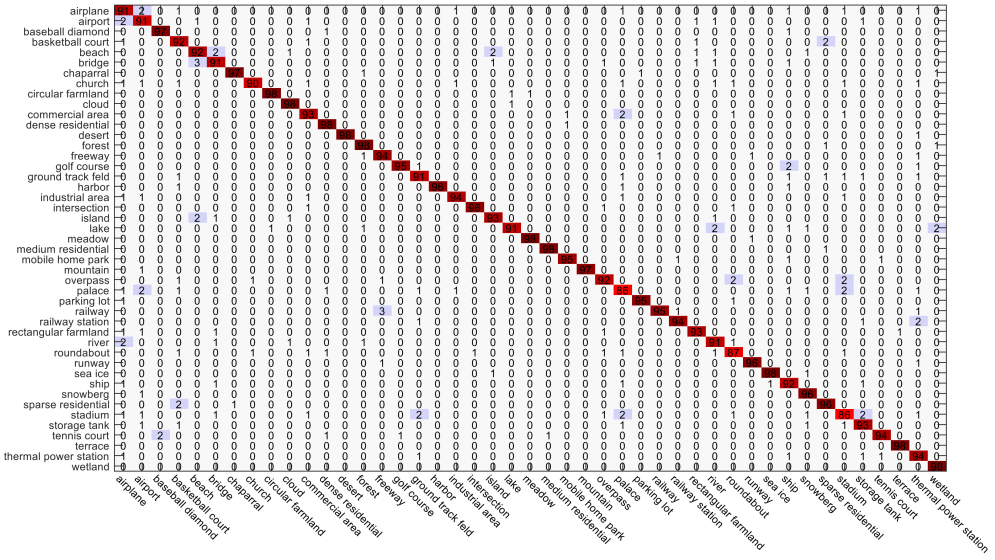
performance. For LGCF, the main confusion happens between the stadium, palace, and ground track field scenes, because they are all composed of building and vegetation, and are similar in the spectral and structural characteristics. Compared with the ground track field scene, some of the stadium scenes even contain the same ground track field. For LGFF under the training sample proportion of 10%, the main confusion happens between the river and roundabout scenes, for the reason that they are both surrounded by vegetation or idle, and some rivers run in a roundabout-type shape. For LGFF under the training sample proportion of 20%, the main confusion also happens between the stadium, palace, and ground track field scenes. On the whole, most of the scenes show a satisfactory classification performance with LGCF and LGFF, which demonstrates the effectiveness of the DLGFF framework.



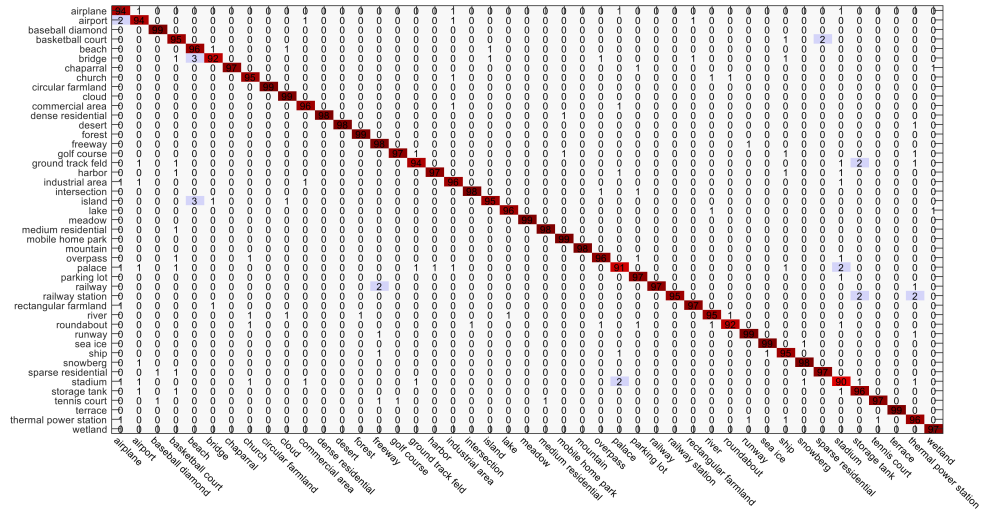
(**a**)

(**b**)

Figure 11. Confusion matrices of LGCF and LGFF under the training sample proportion of 10% with the NWPU-RESISC45 dataset. (**a**) LGCF. (**b**) LGFF.



(**a**)



(**b**)

Figure 12. Confusion matrices of LGCF and LGFF under the training sample proportion of 20% with the NWPU-RESISC45 dataset. (**a**) LGCF. (**b**) LGFF.

5. Discussion

In the DLGFF scene classification framework, the visual word number V is an important parameter, which is discussed in this section (Figure 13). In addition, the effects of the different Conv layers for the three HSR datasets are also analyzed (Figure 14).

(1) The effect of the visual word number V for the LGFBOVW and DLGFF scene classification methods. In the experiments, the visual word number V was varied over the range of [1350,2350,2850,3350] for the UC Merced dataset, the Google dataset of SIRI-WHU, and the NWPU-RESISC45 dataset, under the training sample proportions of 10% and 20%. As shown in Figure 13, with the increase of the visual word number V, the overall accuracy (OA) curves of LGCF and LGFF become higher at the beginning and then fluctuate. For the three datasets, the highest accuracy is acquired when the visual word number is 2350.

Although the classification results of LGFF using the FC feature are better than LGCF using the Conv5 feature, LGCF performs better for the scene images with more complex spatial structures. This is because the Conv5 feature preserves more spatial information than the FC feature. To allow a better visual inspection, some of the classification results of LGCF and LGFF under the training sample proportion of 20% for the NWPU-RESISC45 dataset are shown in Figure 14.

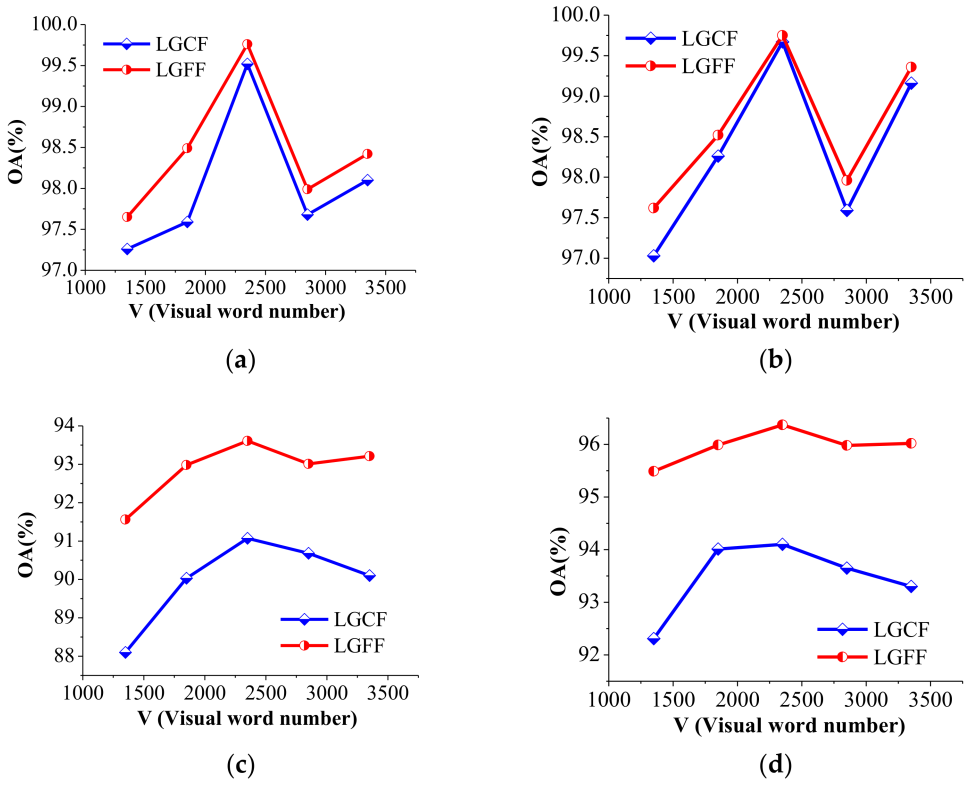
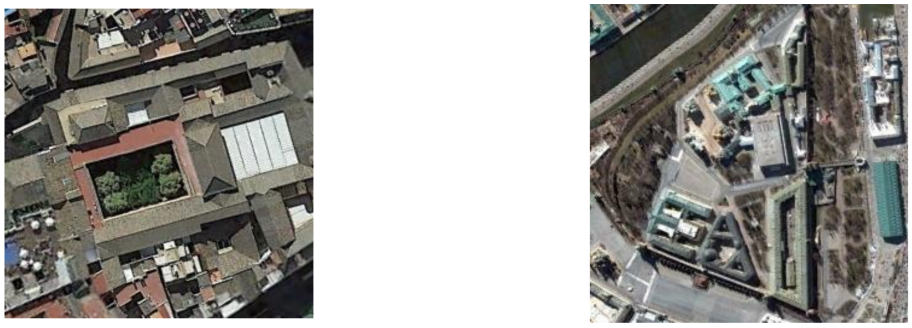


Figure 13. Classification accuracies of the LGCF and LGFF methods with different visual word numbers for the three datasets. (**a**) UC Merced dataset. (**b**) Google dataset of SIRI-WHU. (**c**) NWPU-RESISC45 dataset under the training sample proportion of 10%. (**d**) NWPU-RESISC45 dataset under the training sample proportion of 20%.

Stadium

Figure 14. *Cont.*



Palace

Figure 14. Some of the classification results of LGCF and LGFF under the training sample proportion of 20%, where the images with complex spatial structures are correctly classified by LGCF, but incorrectly classified by LGFF.

(2) The effect of the different Conv layers for the DLGFF framework. In the experiments, the parameter settings of the global SITI feature and the local BoVW-based feature were the same as the settings of the preliminary experiments. The first, third, and fifth Conv layers and the first and second FC layers from the pre-trained CaffeNet were selected for comparison. As shown in Figure 15, the feature from the first FC layer, i.e., Fc6, outperforms the features from the other layers. For the Conv layers, as the depth of the layers increases, the OA also tends to increase. For the FC layers, Fc6 performs better than Fc7. Compared with Fc7, the fifth Conv layer, i.e., Conv5, does not perform as well in distinguishing the images.

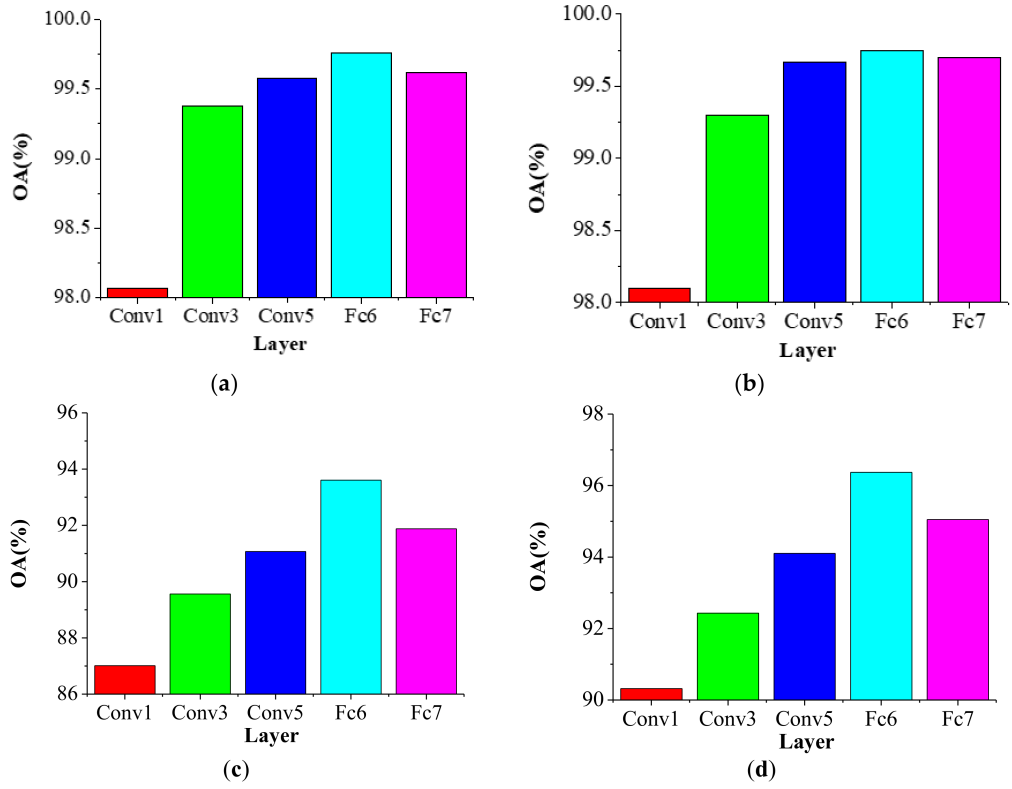


Figure 15. Classification accuracies of the DLGFF framework in relation to the different convolutional layers for the three datasets. (**a**) UC Merced dataset. (**b**) Google dataset of SIRI-WHU. (**c**) NWPU-RESISC45 dataset under the training sample proportion of 10%. (**d**) NWPU-RESISC45 dataset under the training sample proportion of 20%.

6. Conclusions

In this paper, the deep-local-global feature fusion framework (DLGFF) framework has been proposed for high spatial resolution (HSR) remote sensing imagery scene classification. In DLGFF, two effective feature fusion approaches, i.e., the local and global features fused with the pooling-stretched convolutional features (LGCF) and the local and global features fused with the fully connected features (LGFF), are employed for modeling the images. The pooling and stretching operation improves the fusion of the convolutional (Conv) features and the local and global features. The fully connected (FC) feature is similar to the BoVW feature and can be directly fused with the local and global features, and thus a discriminative semantic description is obtained for distinguishing the scenes. The classification experiments undertaken in this study showed that the proposed LGCF and LGFF perform better than the conventional single global feature, local feature, and deep feature–based scene classification methods in discovering high-quality semantics from HSR images, with high robustness. In addition, LGFF achieves a higher accuracy than LGCF, especially for a large dataset, e.g., the NWPU-RESISC45 dataset, which demonstrates that LGFF provides better scene classification results.

End-to-end learning of deep feature architectures is a big challenge, and the findings of this paper can serve as baseline results for future work. In our future research, we plan to use more social media data, e.g., point of interest (POI) data, volunteered geographic information (VGI) data, and OpenStreetMap (OSM) data, to further improve scene classification. To further analyze the scenes, multi-temporal HSR images and images with different resolutions from diverse remote sensing sensors will also be considered. In addition, based on the similarities between HSR images and some non-optical images, we plan to apply our methods to non-optical data, such as synthetic aperture radar (SAR) and Light Detection and Ranging (Lidar) images.

Acknowledgments: The authors would like to thank the editor and the anonymous reviewers for their comments and suggestions. This work was supported by National Natural Science Foundation of China under Grant Nos. 41771385 and 41622107, 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. Qiqi Zhu, Yanfei Zhong, and Yanfei Liu designed the research and analyzed the results. Liangpei Zhang and Deren Li provided advice for the preparation of the paper.

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

Abbreviations

HSR	high spatial resolution
LULC	land-cover/land-use
LBP	local binary patterns
BoVW	bag-of-visual-words
CNN	Convolutional neural network
FC	fully connected
LLC	locality-constrained linear coding
VLAD	vector of locally aggregated descriptors
IFK	improved fisher kernel
DLGFF	deep-local-global feature fusion
MSD	mean and standard deviation
SIFT	scale-invariant feature transform
SITI	shape-based invariant texture index
Conv	convolutional
SVM	support vector machine
HIK	histogram intersection kernel
LGFF	local and global features fused with the fully connected features
LGCF	local and global features fused with the pooling-stretched convolutional features
CpH	compactness histogram

EH	elongation histogram
CtH	contrast histogram
SRH	scale ratio histogram
Caffe	Convolutional Architecture for Fast Feature Embedding
Conv5	the last convolutional layer
FC6	the first fully connected layer
BoVW-MSD	BoVW-based scene classification utilizing the MSD feature
LGFBOVW	the combination of BoVW-based scene classification utilizing the MSD and SIFT features with the global SITI feature
CAFFE-CONV5	pre-trained CaffeNet based scene classification utilizing the last Conv feature
CAFFE-FC6	pre-trained CaffeNet based scene classification utilizing the first FC feature
RS_IDEA	Intelligent Data Extraction and Analysis of Remote Sensing
SIRI-WHU	scene image dataset designed by the RS_IDEA Group of Wuhan University
OA	overall accuracy
POI	point of interest
VGI	volunteered geographic information
OSM	OpenStreetMap

References

1. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Feitosa, R.Q.; van der Meer, F.; van der Werff, H.; van Coillie, F. Geographic object-based image analysis—Towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
2. Hay, G.J.; Blaschke, T.; Marceau, D.J.; Bouchard, A. A comparison of three image-object methods for the multiscale analysis of landscape structure. *ISPRS J. Photogramm. Remote Sens.* **2003**, *57*, 327–345. [CrossRef]
3. Tilton, J.C.; Tarabalka, Y.; Montesano, P.M.; Gofman, E. Best merge region-growing segmentation with integrated nonadjacent region object aggregation. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 4454–4467. [CrossRef]
4. Bratasanu, D.; Nedelcu, I.; Datcu, M. Bridging the semantic gap for satellite image annotation and automatic mapping applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 193–204. [CrossRef]
5. Cheriyadat, A.M. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
6. Li, A.; Lu, Z.; Wang, L.; Xiang, T.; Wen, J.-R. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4157–4167. [CrossRef]
7. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
8. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
9. Chen, S.; Tian, Y. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1947–1957. [CrossRef]
10. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
11. 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]
12. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.-S.; Zhang, L. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. [CrossRef]
13. Zhao, B.; Zhong, Y.; Zhang, L. Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. *Remote Sens. Lett.* **2013**, *4*, 1204–1213. [CrossRef]
14. Fan, J.; Chen, T.; Lu, S. Unsupervised feature learning for land-use scene recognition. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2250–2261. [CrossRef]

15. Wang, Y.; Zhang, L.; Deng, H.; Lu, J.; Huang, H.; Zhang, L.; Liu, J.; Tang, H.; Xing, X. Learning a discriminative distance metric with label consistency for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4427–4440. [CrossRef]
16. Zhong, Y.; Cui, M.; Zhu, Q.; Zhang, L. Scene classification based on multifeature probabilistic latent semantic analysis for high spatial resolution remote sensing images. *J. Appl. Remote Sens.* **2015**, *9*, 0950640. [CrossRef]
17. 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]
18. Zhu, Q.; Zhong, Y.; Zhang, L.; Li, D. Scene classification based on the fully sparse semantic topic model. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5525–5538. [CrossRef]
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, CA, USA, 3–8 December 2012; pp. 1097–1105.
20. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
21. Taigman, Y.; Yang, M.; Ranzato, M.A.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
22. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
23. Wallach, I.; Dzamba, M.; Heifets, A. Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv*, 2015.
24. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3325–3337. [CrossRef]
25. Ma, X.; Wang, H.; Geng, J. Spectral–spatial classification of hyperspectral image based on deep auto-encoder. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 4073–4085. [CrossRef]
26. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
27. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database, Computer Vision and Pattern Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
28. Penatti, O.A.; Nogueira, K.; dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–13 December 2015; pp. 44–51.
29. Liu, Q.; Hang, R.; Song, H.; Zhu, F.; Plaza, J.; Plaza, A. Adaptive deep pyramid matching for remote sensing scene classification. *arXiv*, 2016.
30. Wang, J.; Luo, C.; Huang, H.; Zhao, H.; Wang, S. Transferring pre-trained deep CNNs for remote scene classification with general features learned from linear PCA network. *Remote Sens.* **2017**, *9*, 225. [CrossRef]
31. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1717–1724.
32. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.
33. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land use classification in remote sensing images by convolutional neural networks. *arXiv*, 2015.
34. Liu, Y.; Huang, C. Scene classification via triplet networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *11*, 220–237. [CrossRef]
35. 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]

36. 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]
37. Li, E.; Xia, J.; Du, P.; Lin, C.; Samat, A. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5653–5665. [CrossRef]
38. Xia, G.-S.; Delon, J.; Gousseau, Y. Shape-based invariant texture indexing. *Int. J. Comput. Vis.* **2010**, *88*, 382–403. [CrossRef]
39. Boureau, Y.-L.; Ponce, J.; LeCun, Y. A theoretical analysis of feature pooling in visual recognition. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 111–118.
40. Nogueira, K.; Penatti, O.A.; dos Santos, J.A. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* **2017**, *61*, 539–556. [CrossRef]
41. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
42. Fei-Fei, L.; 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 2005; pp. 524–531.
43. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
44. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
45. Barla, A.; Odone, F.; Verri, A. Histogram intersection kernel for image classification. In Proceedings of the International Conference on Image Processing, Barcelona, Spain, 14–17 September 2003; p. III-513.
46. 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]
47. Zhao, B.; Zhong, Y.; Zhang, L. A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *116*, 73–85. [CrossRef]