applied sciences

MDPI

Review

A Survey on Deep Learning-Driven Remote Sensing Image Scene Understanding: Scene Classification, Scene Retrieval and Scene-Guided Object Detection

Yating Gu, Yantian Wang and Yansheng Li *

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; yatinggu@whu.edu.cn (Y.G.); yantianwang@whu.edu.cn (Y.W.)

* Correspondence: yansheng.li@whu.edu.cn

Received: 27 March 2019; Accepted: 16 May 2019; Published: 23 May 2019

Abstract: As a fundamental and important task in remote sensing, remote sensing image scene understanding (RSISU) has attracted tremendous research interest in recent years. RSISU includes the following sub-tasks: remote sensing image scene classification, remote sensing image scene retrieval, and scene-driven remote sensing image object detection. Although these sub-tasks have different goals, they share some communal hints. Hence, this paper tries to discuss them as a whole. Similar to other domains (e.g., speech recognition and natural image recognition), deep learning has also become the state-of-the-art technique in RSISU. To facilitate the sustainable progress of RSISU, this paper presents a comprehensive review of deep-learning-based RSISU methods, and points out some future research directions and potential applications of RSISU.

Keywords: deep learning; remote sensing image scene understanding (RSISU); remote sensing image scene classification; remote sensing image scene retrieval; remote sensing image object detection

1. Introduction

With the advancement of remote sensing imaging technology, the spatial resolution of remote sensing images has been continuously improved. In low- and medium-resolution remote sensing images, the problem is that different objects may share the same spectral response curve, but the same object may have different spectral response curves, thus making the classification methods at the pixel or object level show many limitations. Using the image block (i.e., the scene) as the basic unit for conducting image interpretation can make effective use of the spatial context information to eliminate the ambiguity of interpretation. As a basic unit of remote sensing image interpretation, the scene is a combination of multiple objects, environments, and semantics. Remote sensing image scene understanding (RSISU) not only needs to recognize each object, but also needs to perceive the topology distribution of multiple objects in a remote sensing image scene. Consequently, RSISU benefits by improving things such as the stability of remote sensing image scene classification, remote sensing image scene retrieval, and scene-driven geospatial object detection. Until now, large amounts of methods have been proposed around RSISU. To pursue the sustainable development of RSISU, this paper mainly focuses on giving a comprehensive review of RSISU, which also discusses the difference and relationship among various sub-tasks in RSISU.

As is well-known, visual abstraction is an important prerequisite for remote sensing image scene understanding (RSISU). As a classic representative of deep learning, the convolutional neural network (CNN) [1] is the milestone technique for abstracting the visual content of remote sensing image scenes. Because of this, CNN has been widely applied in remote sensing image scene classification, remote sensing image retrieval, scene-driven object detection, and so on. Among all of the sub-tasks in RSISU,

deep learning outperforms the traditional technology (e.g., hand-crafted features) by a large margin. With this consideration, this paper mainly reviews and discusses RSISU driven by deep learning.

As the basic tasks of RSISU, remote sensing image scene classification, remote sensing image scene retrieval, and scene-driven geospatial object detection share some communal hints (e.g., scene abstraction) but have key respective techniques, as well as different goals. More specifically, remote sensing image scene classification pursues a high-accuracy classification result by perceiving the objects and their spatial layout, whereas remote sensing image scene retrieval often maps the remote sensing image scene to an efficient representation (e.g., the low-dimensional feature vector) to address the large-scale image retrieval problem. Scene-driven geospatial object detection aims at recognizing the objects by exploiting the scene context. Around each sub-task, large amounts of methods have been proposed and summarized in a specific review. For example, Cheng et al. [2] summarized the remote sensing image scene classification methods, Xia et al. [3] reviewed the remote sensing image scene retrieval methods, and Cheng et al. [4] conducted a survey on the different kinds of remote sensing image object detection methods. Currently, there are no systematic surveys available in regard to the bigger research topic—namely, RSISU.

To achieve a collaborative development of remote sensing image scene classification, remote sensing image scene retrieval, and scene-driven geospatial object detection, this paper aims at summarizing the deep learning driven achievements around RSISU and systematically depicting the relationship among the sub-tasks in RSISU. As a whole, this review includes the following highlights:

1. In contrast to the existing reviews which often review the related techniques on a specific topic, this paper clarifies the relationship and difference among the sub-tasks in RSISU, and gives a systematic review. Hence, this review is of benefit by mining the common problems among the sub-tasks and highlighting the technologies that need to be specially studied in each sub-task.
2. This review not only summarizes the existing achievements, but also points out several promising research directions around RSISU. From this perspective, this review helps the potential readers find the research point, and motivates the engineers to develop the advanced application schema.

The rest of this paper is organized as follows. Section 2 gives a brief review of tasks in remote sensing image scene understanding; Section 3 mainly introduces remote sensing image scene classification; Section 4 mainly introduces remote sensing image scene retrieval; Section 5 shows remote sensing image object detection based on the scene level; Section 6 introduces the future research directions and potential applications; and Section 7 is the conclusion of this paper.

2. A Brief Review of Tasks in Remote Sensing Image Scene Understanding

As mentioned previously, a scene is a combination of multiple objects, environments, and semantics. Scene-based feature representation has been acknowledged to be a more effective way to interpret high-resolution remote sensing images [5,6] which have a spatial resolution of 1.5 m to 4 m, and usually do not have high spectral resolution. Scene classification, scene retrieval, and object detection have all been of benefit to many important applications [7].

Scene classification is a process of learning and discovering image and scene semantic content tags—that is, after feature learning of a large number of samples, the classification results can be achieved after the classifier. Common scene classification methods include things such as scene classification with local semantics, scene classification with middle semantics, and scene classification with a semantic subject model.

Primal remote sensing (RS) image scene retrieval systems, such as those which are text-based, queried merely with the geographical area, and time of acquisition or sensor type, might be very imprecise and inefficient because they largely rely on manually annotated keywords, which are less relevant to the visual content of RS images. The disadvantages of primal RS image scene retrieval based on text or geographic region is imprecision and inefficiency. They mainly rely on manual annotation, which may not be very relevant to the visual content of RS images. Hence, content-based image

scene retrieval (storing and retrieving the scenes based on visual content in the order of decreasing similarity) came into being in the early 1990s, and since then, RS image retrieval efficiency has shown remarkable improvement. Content-based, rather than text-based remote sensing image scene retrieval implementation is what we focus on in this paper, which is achieved by calculating the similarity between visual features. This means that, besides the common feature learning block which also exists in scene classification, there is an independent block of similarity-matching in scene retrieval. Early studies mainly focused on seeking various feature representation methods (e.g., Scale-invariant Feature Transform (SIFT) [8], Bag of Visual Words (BOVW) [9,10], fusion of local and global features [11], deep features, and sparse coded deep features), but a drastic increase of the volume and complexity of RS data caused visual features to be subjective and ambiguous. To solve this problem, researchers have recently proposed the development of more suitable similarity metrics, such as deep features with hash learning, and the end-to-end deep hash learning method. The more satisfactory performances proved that these similarity metrics were more suitable than the distance measures and graphs for large-scale scene retrieval.

By contrast, object detection not only requires learning features and getting classification results, but also knowledge of the direction and location of objects. In previous studies, object detection has often been cast as a classification problem [12]. Features which can characterize the objects are extracted first, and then combined with the predefined classifiers to get the classification result.

The specific functions of each sub-task are shown in Figure 1.

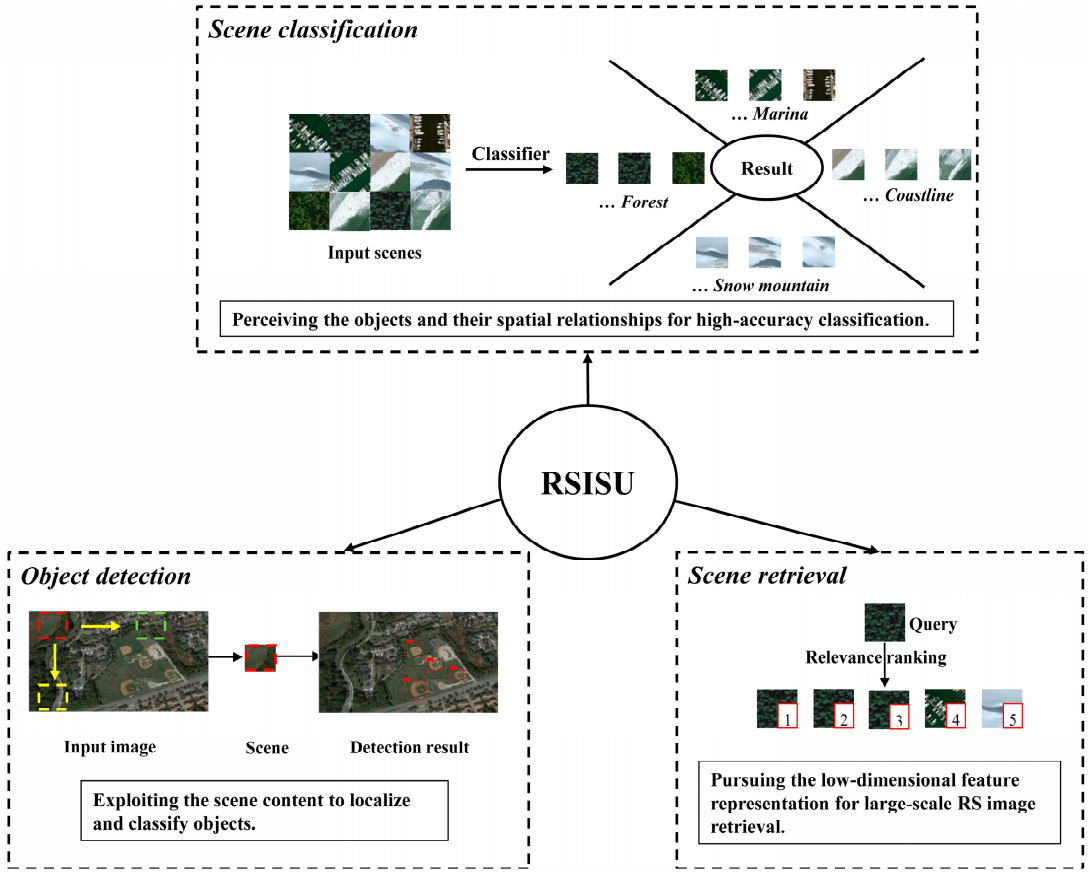

Figure 1. Taxonomy of methods for remote sensing image scene understanding (RSISU). This figure demonstrates each sub-task's function. The common characteristic of three tasks is that they all pursue a scene-level semantic representation of features. Scene classification emphasizes the perceiving of objects and relationships for higher classification accuracy. Object detection excavates the content of the scene in a bid to localize and classify objects, while in scene retrieval, the dimension needs to be reduced for large-scale retrieval.

3. Remote Sensing Image Scene Understanding

Derived from the traditional machine learning field, deep learning has established a multi-layer processing learning model that can hierarchically describe the characteristics of raw data. Consequently, deep learning has made great progress in solving many problems of pattern recognition and the traditional machine learning algorithm by nonlinear processing. One of the reasons why deep learning can achieve high accuracy and efficiency is because a complicated deep learning architecture can be fitted with a large number of training data with diversity and variability to form a better deep network structure. According to the level of scene classification, it can be divided into low-level feature description and middle-level feature description. The middle-level features are aggregation and integration of low-level features, including BoVW [13], SVM [14] based on the structural risk minimization principle, a random forest [15] classifier based on bootstrap resampling [16], and the Markov random field (MRF) [17]. In addition, with the rapid development of artificial intelligence (AI), sparse coding and deep learning algorithms are widely used in scene classification because of their high classification accuracy.

As an important application, scene classification needs a big volume of data, and RS datasets that are frequently used for scene classification are introduced in Section 3.1. Furthermore, methods of supervision based on deep learning have their own strengths and weaknesses, which are identified in Section 3.2.

3.1. Datasets Used for Scene Understanding

With the development of the remote sensing sensor production level, there are many kinds of remote sensing sensors, ranging from the multi-spectral (MS) sensor Landsat-8, to the high-resolution sensor WorldView-III. In addition, we usually need to preprocess the raw data before RSISU. One of the main difficulties in the quantification of remote sensing is the adjustment of the measured values of remote sensing images according to the atmospheric conditions to eliminate the atmospheric impact. The core of this problem is solving the radiative transfer problem in the Earth's atmosphere and estimating the optical parameters of this atmosphere. Among them, the atmospheric correction model based on the radiation transfer model is a method with high accuracy, and the classical models are the 6S model, LOW Resolution TRANsmission (LOWTRAN) model, and MODerate resolution TRANsmission (MODTRAN) model. In addition to the atmospheric correction model based on the radiation transfer model, there are atmospheric correction models based on image characteristics, the typical dark target method, plane field model, internal average method, and so on. At present, many remote sensing scene data sets come from Google Earth, where Google Earth images are an integration of aerial and satellite images. Some of the satellite images are captured through QuickBird and WorldView satellites. Their spatial resolution is less than 1 m, and all sensors contain panchromatic bands, whereas most sensors contain near-infrared bands. The imaging mode is push-broom, rather than swing-broom. Although very high-resolution (VHR) remote sensing images can clearly identify the types of objects and have a high amount of data, the huge amount of data requires high computational efficiency in our classification process. Meanwhile, scene classification in the field of remote sensing is more inclined to semantic segmentation [18]. Thus, for richer semantic information, we need to train the scene-level remote sensing dataset to obtain the semantic label of the scene. The most commonly used scene data sets are shown in Table 1:

Table 1. Introduction of several typical scene databases.

Dataset	Land-Use Classes	Volume	Size	Resolution (m)	Year
UC-Merced [6]	21	2100	256 × 256	0.3	2010
WHU-RS19 [19]	19	1005	600 × 600	Up to 0.5	2010
RSSCN7	7	2800	400 × 400	-	2015
RSC11	11	1232	512 × 512	0.2	2016
SIRI-WHU [20]	12	2400	200 × 200	2	2016
AID [21]	30	10,000	600 × 600	0.5–0.8	2017
NWPU-RESISC45 [2]	45	31,500	256 × 256	0.2–30	2017
PatternNet [22]	38	30,400	256 × 256	0.062–4.693	2017
RSI-CB128 [23]	45	>36,000	128 × 128	0.3–3	2017
RSI-CB256	35	>24,000	256 × 256	0.3–3	2017
AID++ [24]	46	>400,000	512 × 512	-	2018

In addition, as a non-optical remote sensing method, the synthetic aperture radar (SAR) also plays a pivotal role in remote sensing images. Due to its independence on atmosphere and sunlight conditions, SAR images are also obtained to categorize the scene [25] like remote sensing images that rely on spectral information. However, the SAR image is also limited by the existence of speckle noises and the absence of effective feature representation.

It is worth mentioning that through being confronted with higher spatial resolution, scene classification datasets have been greatly developed. However, training large-scale remote sensing datasets from scratch are still subject to many limitations. The database, which only has three bands—R, G, and B—is also different from the real remote sensing spectral band. Like Aerial Image Dataset++ (AID++) and Remote Sensing Image (RSI), we can make more use of the tagging of crowd-sourced geographic data, such as OpenStreetMap (OSM).

3.2. Deep-Learning-Driven Scene Classification

The deep convolution neural network [26] has achieved promising results in the field of natural image scene classification, and it also has many similarities with remote sensing image scene classification. Through the migration of the data set training model, the classification accuracy of the remote sensing image dataset can also be very encouraging. However, there are only three channels of red, green, and blue in natural images, while remote sensing image data with more than three bands have not yet been able to be trained in deep learning training networks, such as ImageNet. This section mainly introduces the pre-trained convolutional neural network model, the unsupervised classification method of scene classification, and fully supervised deep feature networks. The typical classification process is shown in Figure 2. After the input original image passes through the convolution layer, the regularization and pooling layer, the convolution operation, the full connection layer, and a normalization and pooling operation, a classifier is used to get a classification result. The most commonly used classifiers are: the support vector machine (SVM), AdaBoost [27], k-nearest-neighbor [28] (kNN), conditional random field [29] (CRF), sparse representation-based classification [30] (SRC), artificial neural network [31] (ANN), random forest, Markov random field (MRF), and support tensor machine [32] (STM).

Typical deep learning network structures include the deep belief network which needs to vectorize the raw image and can lead to loss of topology information, the stacked autoencoder (SAE) [33], and convolutional neural networks (CNNs). For a large number of labeled sample datasets, convolutional neural networks (CNNs) is the most effective learning model for feature extraction. The basic network structure of feature learning is the convolution layer, normalization, pooling layer, and full connection layer [34]. As mentioned above, deep learning has incomparable advantages over traditional algorithms. Scene classification mainly includes artificial features (Grey-Level Co-occurrence Matrix (GLCM) [35], Local binary patterns (LBP) [36], Histogram of Oriented Gradient (HOG) [37], Gist [38]), data-driven supervised classification features, and data-driven unsupervised classification features.

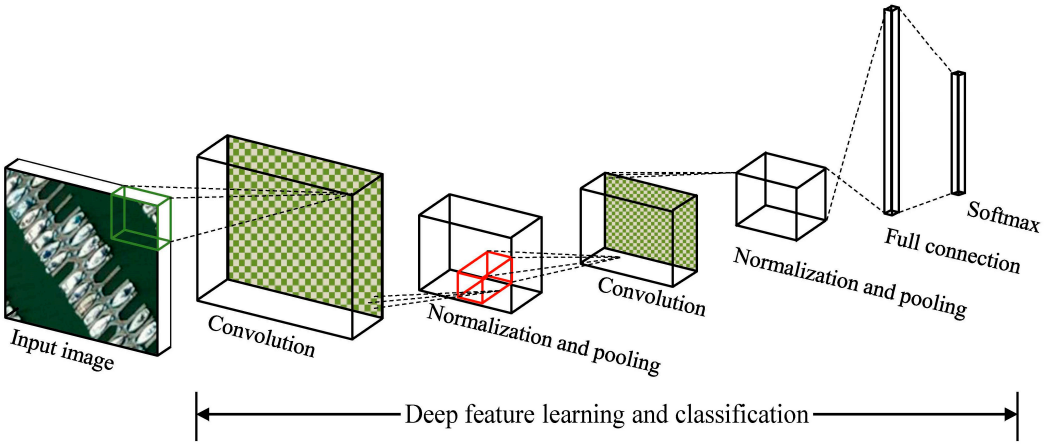

Figure 2. Typical architecture of a convolutional neural network (CNN) for scene classification. The input image gets feature representation through deep networks (pre-trained CNN, unsupervised deep feature learning, and an end-to-end architecture can be included).

3.2.1. Feature from Pre-Trained CNN for Remote Sensing Image Scene Classification

As depicted in [26], the authors utilized the pre-trained CNN network model based on ImageNet, which performs well on scene-level datasets, such as The University of California, Merced (UC Merced) Land-use and Brazilian Coffee Scenes, and they also used the conservative SVM. The paper depicts that in different scene-level datasets, the scene-level aerial image datasets often achieve better results than agricultural datasets. The reason for this is because many objects in the data set of aerial images (e.g., aircrafts, cars) have similar salient edges and boundaries with the datasets pre-trained in the CNN network. Thus, training samples have a great influence on the accuracy of scene classification in supervised classification. The proposition in [39] also used pre-trained CNN models and domain-specifically fine-tuned, pre-trained CNN models on the scene-level dataset, which functioned as universal feature extractors to extract richer high-level representation features. In [39], the writer adopts two strategies for feature extraction. The first is that the parameters of the convolutional neural network model are trained on an ImageNet dataset, and the results are migrated directly to a scene classification task. That is to say, the last classification layer is removed, and the feature output vector of the former layer is used as the input vector of the latter layer. The second strategy is trained on the ImageNet dataset as before, but through fine-tuning in a specific domain. Owing to fine-tuning, the average classification accuracy is further boosted by 2–5% for AlexNet, a 16 layer VGGNet, and GoogleNet, respectively. A Discriminative CNN (D-CNN) model is proposed in [40], which optimizes the discriminant objective function and introduces a metric learning regularization term into the existing CNN model features, in addition to minimizing classification errors. Through testing on the UC Merced, Aerial Image Dataset (AID), NorthWestern Polytechnical University-REmote Sensing Image Scene Classification (NWPU-RESISC45) datasets, it has been found that this method can reduce classification errors caused by the similarity between classes. Even though the training ratio is only 20%, the overall accuracy can reach 90.82 ± 0.16% and 89.22 ± 0.50% on the aforementioned AID and UC Merced dataset, respectively. The paper [41] puts forward a gradient-boosting random convolutional network (GBRCN). This is the first time that a deep integration framework for scene classification is proposed, and it introduces a multiclass soft-max loss function into the framework. The basic classifier is Random CNet (RCNet), which can reuse the weight of each convolution model and reduce the parameters to make the feature extraction process more efficient. Compared with traditional methods, such as Spatial Pyramid Matching Kernel (SPMK) (89.67%), SIFT + Sparse Coding (SSC) (91.33%), Saliency-guided Sparse Auto Encoder (SSAE) (92.20%), the final classification accuracy can reach 98.78%. Unlike most methods which concentrate on deep network architectures, [42] focuses more on feature fusion methods, where Visual Geometry Group (VGG)-Net is still used for feature

extraction, and the full connection layer obtained from the deep learning network model is considered to be a separate feature description method. Unlike the traditional feature fusion methods based on addition and connection, discriminant correlation analysis (DCA) is used as a feature fusion method. This fusion method can describe scene-level images, reduce dimension and data information, and achieve better results for Very High Resolution (VHR) images than traditional images. When the training ratio is 80%, the scene classification performance tested on the UC Merced dataset of DCA fusion can still reach 92.32 ± 1.02% with slight improvement (about 4–5% over AlexNet and CaffeNet).

3.2.2. Unsupervised Deep Feature Learning for Remote Sensing Image Scene Classification

Deep neural networks [8] have been widely used to learn low-dimensional feature representations to reduce the dimensionality of VHR images. Sparse coding [43] is another famous unsupervised feature learning method, which is highly effective for scene classification when compared to the traditional Bag of Visual Words (BoVW)-based approaches [44] and generates a set of basic functions from the unlabeled data. Recently, a method combining scale-invariant-feature-transform (SIFT)-based feature descriptors and sparse coding (Sift + SC) has been put forward. Although unsupervised classification has many advantages, compared with supervised classification, it is still unable to fully automate, and has a long way to go.

In order to overcome the influence of pre-training samples on classification accuracy and the shortcomings of given data with few labels, the unsupervised learning algorithm is also an important part of remote sensing. In [45], an unsupervised feature learning method is proposed. As we all know, one of the fatal drawbacks of unsupervised classification is that the effect of the classifier is very sensitive to the selection of parameters. The author uses a two-level feature extraction algorithm based on K-means clustering. The first layer generates a contour basis, and the second layer generates a corner basis. K-means clustering only needs to adjust one parameter, has strong robustness, and can extract complex structural features (such as corners and junctions). Using the University of California, Merced (UCM)-21 data set based on the scene level, by comparing the original training samples (83.3%) and preventing the rotation of the training samples after fitting (89.1%), the latter can usually achieve a higher level of accuracy. Therefore, in scene level classification, it is very important to use augmentation to prevent over-fitting in deep network training. The paper [46] proposed a new method for training a scene-level UC Merced dataset. Yu proposed a balanced data-driven sparsity (BDDS) method, which designed a feasible sample input for a better-enforced lifetime and population sparsity (EPLS), and hence improved the supervised trained CNN. The proposed deep feature learning algorithm could overcome difficulties by smoothing patches with similar textures, and the same components lacked edge structure information which could be helpful for the classification and failed to present the key evidence for classification. No matter what clustering method (such as k-means and gradient modules) it is combined with, the classification accuracy can reach more than 79%.

Inspired by SAR images unlike the optical remote sensing images, the paper [47] utilized deep unsupervised discriminative feature learning in order to overcome the lack of labeled samples to train the DNN, which may lead to overfitting.

For a good deep learning algorithm, besides improving the algorithm's efficiency, it is also very important to improve the diversity of training samples, which is crucial in order to train a robust deep learning model. This paper [48] proposes a data augmentation method, which can decrease the constraints of remote sensing image scenes limited by spectral and topological relations, as well as enhance visual diversity and the reliability of scene classification. The accuracy of the augmentation method can reach 99.127% and 99.297%, respectively.

3.2.3. Fully Supervised Deep Networks for Remote Sensing Image Scene Classification

Unlike pre-trained and unsupervised deep feature learning, which need classifiers such as a random forest or SVM, metric learning (ML) has gradually developed, which is able to enhance the learned features separability. ML is an end-to-end learning method mainly used to find an appropriate

measure of similarity between pairs of data, which can preserve the desired distance structure very well. Existing metric learning is usually categorized into two streams: contrastive embedding and triplet embedding [49–51].

Contrastive embedding: Contrastive embedding is trained on paired data (x_i, x_j). Concretely, the cost function is defined as,

$$J = \sum_{i,j} \ell_{ij} D^2(x_i, x_j) + \left(1 - \ell_{ij}\right) h\left(a - D\left(x_i, x_j\right)\right)^2 \quad (1)$$

where the label indicator $\ell_{ij} \in \{1, 0\}$ indicates whether the paired data (x_i, x_j) are from the same class or not. $h(x) = \max(0, x)$ denotes the hinge loss function, and $D\left(x_i, x_j\right) = \| f(x_i) - f\left(x_j\right) \|_2$ is the Euclidean distance of paired data (x_i, x_j).

Triplet Embedding: Triplet embedding is trained on triplet data (x_a, x_p, x_n). Concretely, the cost function is defined as:

$$J = \sum_{a,p,n} h\left(D\left(x_a, x_p\right) - D(x_a, x_n) + a\right)^2 \quad (2)$$

Here, (x_a, x_p, x_n) are the triplet data, where (x_a, x_p) have the same class labels and (x_a, x_n) have different class labels.

Compared with traditional measurement learning, which focuses on preserving interclass separability, [44] pays more attention to label consistency (LC). In order to ensure the intraclass compactness of VHR images, a discriminable distance metric learning method (DDML) is proposed in this paper. In the learning process, besides the enforced intraclass compactness and interclass separability, the global and local label consistency (LC) is constrained, which is aimed to jointly optimize the feature manifold, distance metric, and label distribution.

As is well-known, the last layer of the network is structured as FC layers, which are typically used in a bid to better summarize the low-level feature learning for a final strategic decision. Apart from reducing cross entropy loss in FC layers, a metric learning regularization method was proposed in [21] to make the network model more discriminative. In this paper, a new objective function is proposed with three terms, including a cross entropy loss term, a metric learning regularization term, and a weight decay term. When the percentage of labeled images is 15%, the classification accuracy (AC) and standard deviation (STD) can reach 69.33 ± 1.25%.

In order to utilize the structural information between each scene, [52] proposed a special structured loss to measure pairwise distances. Hence, this paper puts forward diversity-promoting deep structural metric learning (D-DSML). Compared with the similar learned metric parameters of a traditional DSML, this method can overcome the overlapping of representations of different factors. The D-DSML method is implemented by imposing a diversity-promoting prior which can prevent the factors in DSML from being correlated, so that the redundancy caused by the similarity between the hidden factors can be decreased. D-DSML-CaffeNet's classification accuracy can reach 96.76 ± 0.36% with a slight boost, compared with CaffeNet (95.48%), GoogLeNet (94.31 ± 0.89%), and VGG-VD16 (95.21 ± 1.20%).

4. Remote Sensing Image Scene Retrieval

Remote sensing image scene retrieval (RSISR) is also an important application of content-based image scene retrieval (CBISR). Remote sensing (RS) image retrieval, by analyzing the visual content, is crucial for mining geological information. Finding the specific region from a large number of RS images needs to be done before producing applications of RS. A fundamental flowchart of the RSISR method includes two indispensable blocks: feature learning or extraction, which is aimed to map the image which is queried or in the dataset into the feature space, and similarity matching (i.e., feature searching, using an appropriate measure to judge the similarities between the query image and images in the dataset). A block named "relevance ranking" is a popular post-processing method for RSISR, which focuses on enhancing the original performance of retrieval. However, it is essentially an

interactive block of similarity-matching using the feedback information of users. As this block is also used in many subsequent articles, we will not repeat it. The overall architecture of scene retrieval is shown in Figure 3.

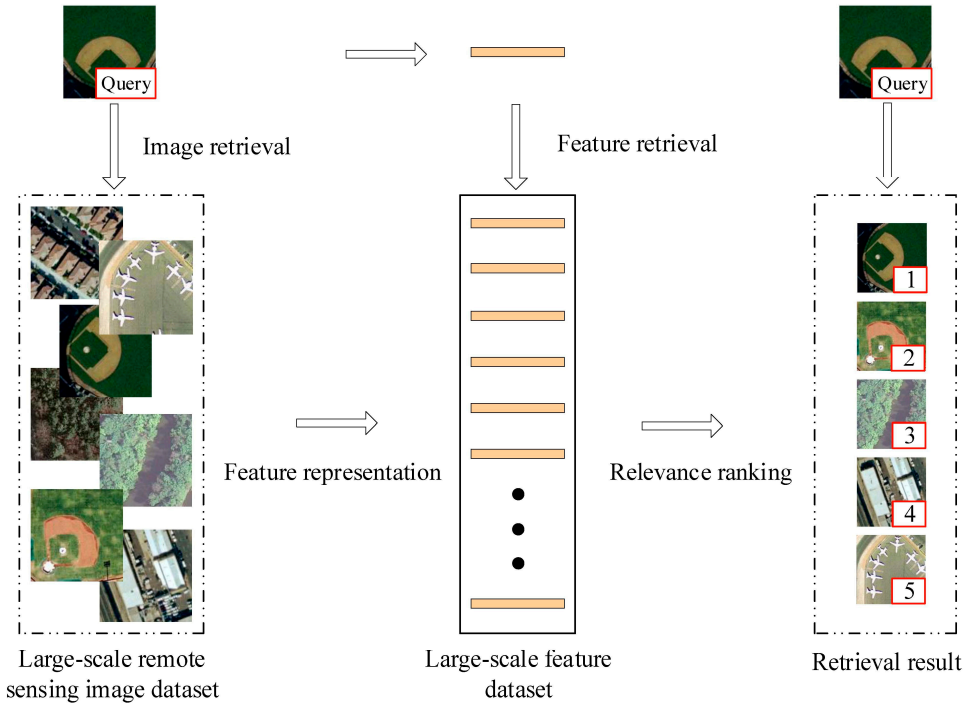

Figure 3. Flowchart of remote sensing image scene retrieval (RSISR). The input remote sensing (RS) scene datasets are trained in deep networks. After feature hashing, which is aimed to reduce dimension, and the first similarity matching from large-scale images, relevance ranking can be further achieved according to the similarity among the query image and retrieved images.

In the block of feature learning, the low-level (e.g., GLCM) and mid-level features (e.g., Bag-of-Words) are always traditional and handcrafted [53]. Meanwhile, since the same type of scenes might emerge at different scales, orientations, and illuminations, relevant images of diverse appearance are difficult to be retrieved. An increasing amount of triumphant applications prove that the features learned by the rapidly-growing deep learning (e.g., CNN) are more effective for narrowing the semantic gap and describing the RS images. Deep learning can excavate intricate structure in huge data sets, and automatically learn feature representations optimally and powerfully. These kinds of features are, namely, deep features [54,55]. The feature fusion and the dimension reduction can be done in this block, and the following RSISR methods are merely about deep features in the block of feature representation. As Figure 3 reveals, we transfer the task of RS image retrieval to the task of feature retrieval.

In the block of similarity matching, features excavated from the images in the dataset are compared to features excavated from the query image [56]. The images in the dataset are ranked according to decreasing similarity. In the methodology section, we will talk about distance measures, the graph models, and hash learning, respectively. The block of similarity matching is what we will focus on.

4.1. Retrieval by Distance Measures

Bao et al. [57] added eight similarity metrics into RSISR, and showed the impact of this. For a particular retrieval task, a suitable similarity metric can get K nearest neighbors, preferably by weighing the similarities between features, which makes the results more desirable. Distance measures in his paper include two categories: general feature vector-based metrics, which includes three distances of Minkowski, and the cosine of the angle; and histogram vector-based measures, which includes

the histogram intersection, center moment, χ^2 statistical distance, and Bhattacharyya distance. This research intuitively demonstrates that similarity metrics for a content-based RS retrieval task are very significant, and various similarity metrics may result in distinctive ranking results, but the distance metrics in his work were not applied here to measure the deep features.

Assume the feature vectors of query image scene Q and image scene D in the data set are $q = \{q_1, \ldots, q_n\}$ and $d = \{d_1, \ldots, d_n\}$. For instance, $L1$, $L2$, $L\infty$ that were developed from a set of distance functions known as the Minkowski r-metric can be defined.

For r = l, it is the city-block distance (i.e., $L1$ or Manhattan distance),

$$d1(q,d) = \sum_{i=1}^{n} \left| q_i - d_i \right| \tag{3}$$

For r = 2, it is the Euclidean distance (i.e., $L2$),

$$d2(q,d) = \sqrt{\sum_{i=1}^{n} (q_i - d_i)^2} \tag{4}$$

And for r = ∞, it is the dominance distance (i.e., $L\infty$),

$$d\infty(q,d) = \max_i \left| q_i - d_i \right| \tag{5}$$

For RSISR, deep learning techniques can be roughly divided into unsupervised feature learning and supervised feature learning methods. Due to the spatial complexity of image scenes, unsupervised feature learning can accurately describe semantic information. Unsupervised feature learning is attractive for RS, since RS has less labeled data than a few image analysis areas. For example, in the ImageNet data set in the field of computer vision, more than 15 million images were tagged. However, few RS data sets contain more than 10,000 annotated images. Zhou et al. [56] first learned features sparser from RS images utilizing the auto-encoder, as well as minimizing the memory of features. They developed an unsupervised feature learning framework (UFLF), and $L1$ and $L2$ were utilized to weigh the similarity for the sparse features produced by their framework. Besides these two measures, he also employed a histogram intersection. However, the auto-encoder network of the framework is very shallow, and only has a single hidden layer, which makes it incapable of producing sufficiently powerful feature representations. Furthermore, the features learned in an unsupervised way may require longer codes to attain satisfactory results of retrieval, which will largely reduce the image retrieval efficiency. Deeper networks are necessary for RSISR.

In high-spatial resolution RS images retrieval, the deep features produced by the Full Connection (FC) layers are usually large, and present computational and storage challenges. This problem has been well-addressed by Yaakoub [58], and the limitations, which merely consider features produced by the last FC layer, are solved. They performed transfer learning for RSISR by taking into consideration features from the FC and convolutional layers from a wider range of CNN. In their framework, scenes were decomposed under the Quin-tree decompositional principle. The Low-Dimensional Convolutional Neural Network model (LDCNN), which is pre-trained, is used to automatically extract deep features from the scenes and predict the labels of them [59]. They adopted $L1$, $L2$, and the Correlation and Cosine similarity as distance measures. Just as Napoletano [60] utilizes CNN networks later, for both LandUse and SceneSat datasets, the descriptors based on CNN are more excellent than other descriptors in RSISR of any class. The retrieval performance of Cosine is near to that of $L2$. Meanwhile, the Average Normalized Modified Retrieval Rank (ANMRR) of χ^2 outperforms that of $L2$ by 0.163 using SatResNet-50, which is slightly smaller (0.045) than that of $L1$ in his work. Deep metric learning CNN was introduced to RSISR by constructing a Triplet network with a metric

learning objective function, which extracts the representative deep features in a semantic space. In such a semantic space, simple distance measures such as $L2$ can be used directly. A method based on the FC layer of the CNN (supervised learning) reduces the dimensionality of the semantic features, which further guarantees that the similarity matching is efficient and that the storage is not too large. Instead of simply taking a distance as the similarity, as many content-based RSISR ways do, a retrieval method [61] based on weighted distance and fundamental features of fine-tuned CNN is demonstrated. The CNN model is used to weigh the weight of image classes one by one, and employ them to calculate the $L2$ between the image which is queried and the images in the data set. As one can see, it is simple but efficient.

For multi-scale information problems, [62] combined fine-tuned CNN features excavated from multi-scale images or from multiple sub-patches. They utilized concatenation from various scales, and multi-patch pooling based on the retrained GoogLeNet CNN, and were able to obtain higher levels of accuracy in that the Average Normalized Modified Retrieval Rank (ANMRR) value was approximately 3.1% less than that attained by the manual Relevance Feedback VGG Medium (RF VGG-M) descriptor. The similarity metrics are $L2$, Cosine, Manhattan, and Chi-square. To make the scene sampling more reasonable, Tang et al. [53] used two schemes. Then, they learned latent features using the deep convolutional auto-encoder (DCAE). To find a proper metric for their deep Bag-Of-Words features, they selected eight common distance metrics: $L1$, histogram intersection, Bhattacharyya, chi-square, Cosine, correlation, $L2$, and inner products. A relatively good retrieval performance could be attained using $L1$ and $L2$ among these metrics. The low-dimensional global descriptors in some of the methods above mean that irrelevant background information may still be encoded. Then, the local convolutional descriptors extracted in an end-to-end manner and based on feature saliency were combined with Vector of Locally Aggregated Descriptors (VLAD) aggregation to achieve compact and descriptive image representations [63]. This effectively enabled fast database searching and made it able to capture deep semantic multi-scale information. The similarity was determined by $L2$ of the query and database vectors.

For multi-label problems, conventional RSISR systems usually perform single-label retrieval, which underestimates the complicacy of RS images, where an image may be labeled by several tags, thus leading to worse retrieval performance. Zhou et al. [64] therefore proposed a multi-label RSISR approach with FC networks, and excavated region convolutional features according to a segmentation map of each image. The region features were processed further to gain a feature vector for similarity matching using $L2$, thus improving the search efficiency and retrieval performance, as well as solving the multi-label problem at the same time.

For multi-band information problems, traditional red, green and blue (RGB)-based image representation has been widely used for RSISR. Whereas true RS images are hyperspectral, the rich spectral information makes them well-suited for RSISU. In [65], deep features were extracted using CNN from the proposed new hyperspectral dataset for the RS community, and $L2$ was used to compute the similarity. From the examples mentioned above, we can conclude that supervised feature learning methods (e.g., CNN) generally outperform unsupervised feature learning methods by a wide margin. For instance, using UC Merced data sets, the precision of the unsupervised method employed in [56] is about 65%, but the precision of the supervised method used by [59] can reach more than 80%. The distance measures allow the retrieval performance conducted on the deep features to be more perfect. RSISR for SAR is not mentioned, in that deep features are scarcely exploited for SAR.

However, the aforementioned methods are all vector-based; they all calculate the similarity between two scenes by $L2$ or other metrics between two feature vectors. The intrinsic relationship between different types of features had not been excavated. They only consider the images themselves, and they become unsuitable if the feature vector is extremely hybrid. Even when multiple types of features are extracted, the $L2$ distance is used to calculate the similarity between two super-feature vectors, which are integrated from multiple features in the Greedy Affinity Metric Fusion (GAMF) method [66]. All of these lead to a final conclusion: the distance measures are so feature-level isolated

that the retrieval performance is relatively poor. Taking [66] as an example, using Local Binary Pattern (LBP) and three other features as the feature combination, 16 of the 21 classes of scenes that are retrieved by Collaborative Affinity Metric Fusion (GAMF) combined with *L*2 have a smaller level of precision than those by the Collaborative Affinity Metric Fusion combined with a graph model.

4.2. Retrieval by Graph Models

Unlike the aforementioned RSISR methods, which often concatenate features of all genres into one vector, the graph models enriched with node and edge attributes are usually the elective data structures for RSISR, in terms of scenes constitution and the relations among them. The high-resolution RS scenes can be precisely retrieved by graph models. Facing the problem caused by multi-type and large-quantity features, the graph models outperform the vector-based methods in that they make multiple complementary features effective.

In this part, we should be aware that many existing region-based graph models have been utilized in RSISR. For example, [67–70] used attributed relational graphs (ARG) to represent the images instead of using deep features, while their similarity metrics are region-based. These graph models are not what we are pursuing. Only graph-based retrieval is discussed below in this section, which means there is retrieval in the data set by graph-based algorithms. Each image in the data set corresponds to a node in a graph, and the edge between two nodes represents the connection between scenes, rather than regions. A weight on one edge represents the similarity. As seen in Figure 4, such a mesh structure shows that auxiliary information from other scenes in the dataset is actually taken into consideration, from vector-based one-to-one feature retrieval to one-to-many feature retrieval.

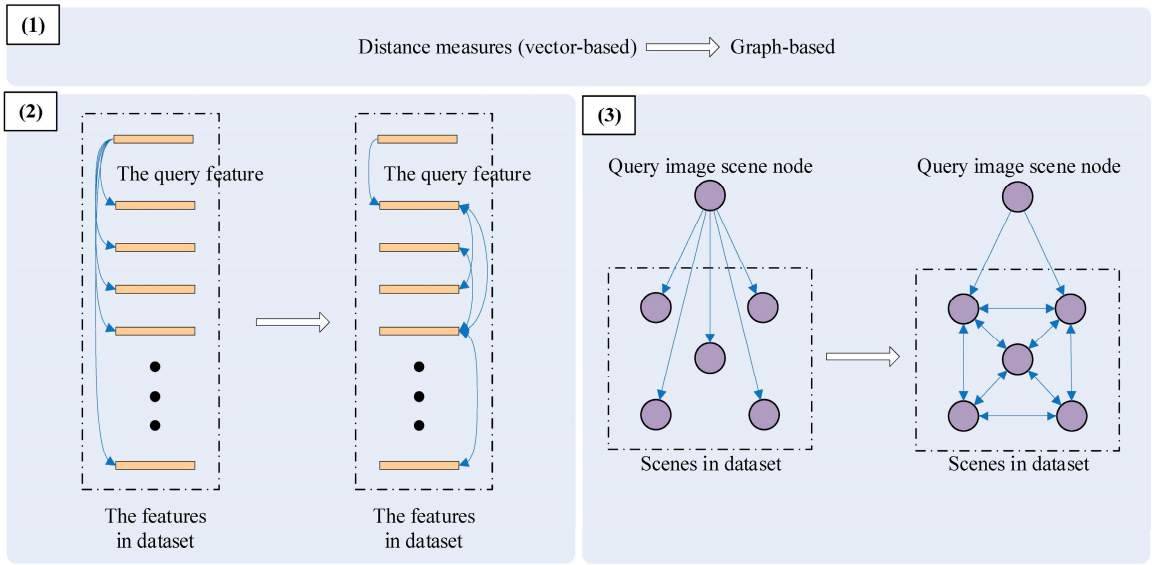

Figure 4. Schematic diagrams of distance measures and graph-based measures. The picture on the left is the principle of feature-level retrieval, and the picture on the right is the principle of scene-level retrieval. A one-way arrow indicates similarity matching between the query scene or corresponding feature and the dataset scene or corresponding feature. Two-way arrows indicate there exists transmission of information between scenes or scenes features, so the scenes in the dataset aren't feature-level isolated.

Owing to the single-image feature vector-based algorithms, to alleviate the incredible difference of the retrieval accuracy among queries, a three-layer graph learning method was applied to holistic and semi-supervised local features [71], which is reasonable and scalable. Here, the graph-based algorithm of retrieval consists of a kernel matrix, which represents the similarity among images using multiple features. The paper in [72] also uses RS images to make a graph which is undirected (fully connected), and then utilizes graph learning for retrieval. Here, original rank scenes are also graph nodes, and the similarity between nodes is weighted by the resemblance between RS scenes under the

chosen measures like *L*2. That is to say, they first use different "distance measures" to attain various affinity graphs, and then re-rank scenes by calculating the similarities between nodes within the affinity graph. To enhance multiple complementary features efficiency in RSISR, Li et al. [66] employed the model that crossed diffusion. In this paper, the collaborative metric fusion is the graph-based retrieval algorithm, which shares information in multiple feature spaces in the fusion process. It builds the relationship among multiple types of deep features, and imports other auxiliary images in the dataset while maintaining the discriminating power. Here, each initial graph is constructed by one corresponding feature. The image retrieval task can be finished immediately just by searching the graph's affinity matrix rather than distance measures, and this method can achieve a dataset-level precision of 76.57%, which is 33.26% higher than the framework which uses *L*2 in [56].

4.3. Retrieval with the Aid of Hash Learning

The approaches in the last part use the graph-based model to measure the similarity, but the graph model cannot easily be applied to large-scale RSISR, because it requires a large memory and the computation of it is very complex. In addition to improving feature retrieval skills, there is also a feasible way to reduce the dimensionality of features. In the former method, there are methods based on trees, such as the K-Dimensional (K-D) tree which divides data into subspaces and stores partition patterns through trees. This method does improve the retrieval speed, but can greatly damage the performance, especially when the high dimension of features makes the tree more complex. Obviously, they are not suitable for RSISR.

There are alternative solutions (i.e., the methods which reduce the dimension of feature) to overcome the shortcomings of methods based on the tree. For instance, [73] employed the Principal Component Analysis (PCA) to map RS features learned by traditional CNN to Deep Compact Codes (DCC). Recently, hash learning methods have been used to address large-scale RSISR, and have been shown to behave promisingly [74–78]. The hash functions change features from high vectors into binary low vectors (i.e., the vectors are made of binary hash codes), and the primitive structure can be preserved. Figure 5 shows how the amount of memory required is largely reduced, whereas the retrieval efficiency is enhanced (i.e., the Hamming distance can be calculated efficiently) in the low-dimensional binary feature space (i.e., Hamming space).

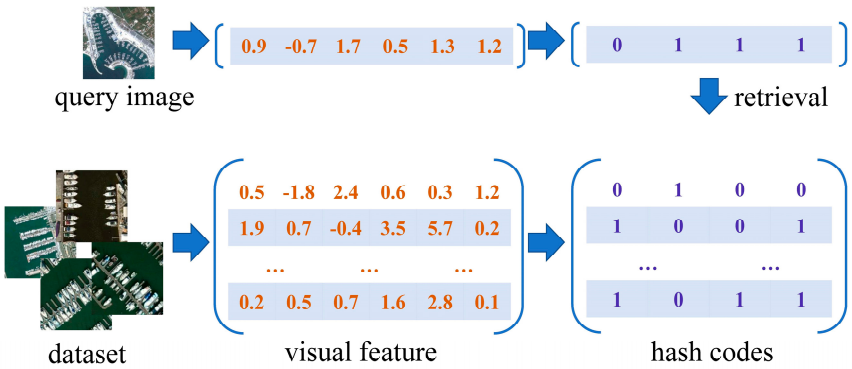

Figure 5. Schematic diagram of hash learning. In large-scale remote sensing image scene retrieval (RSISR), hash learning enormously reduces the storage while maintains the discriminating power of features, as well as transferring visual features to hash codes.

Compared to the graph model, there is neither high computational complexity nor storage complexity of graphs in hash learning. The constraint among the query image and retrieved images are denser. To overcome the problem of how linear search is time-consuming and unsuitable for multi-scale situations, [75] first introduces two nonlinear kernel-based hashing methods. The first method only uses untagged images, and the hash functions are defined in the kernel space. By contrast, the second method uses the semantic similarity excavated from annotated images to describe the

unique hash function in the kernel space. However, the learning time complexity of these methods can be high and cannot achieve a quick response, so [76] proposes a novel data-independent hashing method, named partial randomness hashing (PRH). Later, Li [74] advocated an unsupervised feature learning method. The hash learning module probes the hashing function. Li learned the hybrid feature mapping function on column sampling hashing, where the precision and recalls are even larger than when using the former KSH (kernel-based nonlinear hashing in [75]). The approach has surprisingly better retrieval performance in that the mean average precision (MAP) values exceed those of KSH by about 45%, which can train all of the data that is available. As Zhang et al. [79] brought the hashing method to hyperspectral RS scenes, deep features which are spectral-spatial and produced by the Deep Convolutional Generative Adversarial Networks (DCGAN) model were then Nonlinear Manifold [80] (NM) hashed to reduce their dimensionality. Multi-index hashing was also employed to search hyperspectral images.

There is a trend where deep feature extracting and hash learning need to be jointly trained in one network, so that the retrieval performance could be enhanced in an end-to-end manner. The similarity metrics that are learned can be more appropriate for the deep features. The deep hashing neural networks [81] (i.e., DHNNs) are developed, while both deep feature learning and hash learning neural networks are constructed and linked. The time spent on the feature extraction can be reduced, and the final features generated from the Deep Hashing Neural Networks (DHNNs) in the large-scale RS image dataset can be calculated before and stored as the data set of features without leading to huge costs of memory. However, a high retrieval accuracy can only be achieved if hash codes are long enough and training images annotated are adequate in this method. To solve these problems, this paper [77] advocates a solution that learns a metric space based on semantics, while only using a few annotated training images for fast and accurate RSISR in large archives. Also in the paper [78], the deep semantic hashing (DSH) model was also end-to-end, but only exploited a few tagged images. Meanwhile, DSH abates the accuracy of the hash codes in that there is no relaxation in its learning. The author experimented on a Canadian Institute for Advanced Research (CIFAR)-10 data set and utilized his DSH, as well as a few hashing methods such as Hashing with Mutual Information (MIHASH), which is famous in the computer vision community. Mean Average Precision (MAP), which was chosen to be the accuracy, showed that the DSH was about 12.7% more efficient than MIHASH when the scenes were 32-bits.

Dealing with cross-source (CS) RSISR is also in need. Usually, the scenes retrieved are from the same RS data source. If the same method is applied to retrieve scenes in another data source, then the data shift problem emerges, and the performance decreases drastically. The CS messages can be fused to achieve a higher retrieval performance in the end, and the CS tasks firstly need to be dealt with [82]. For instance, the MAP values of the proposed approach overwhelm those of other famous methods like DCMH for about 0.15 in the computer vision community when they are applied to solve RSISR CS tasks. The CS deep hashing is even more necessary because the volume of CS remote sensing data is continually increasing. Meanwhile, RSISR results vary among datasets, since the CNN models learned are different. Only source-invariant deep hashing methods, such as source-invariant deep hashing convolutional neural networks (SIDHCNNs) can overcome these two shortcomings. In this example, CS messages in multi-spectral images (high spectral resolution) and panchromatic images (high spatial resolution) are fused to achieve a higher retrieval performance in the end.

From distance measures to graph models and hash learning, existing methods can mine complementary messages of multiple features and multiple scales more effectively [74]. The retrieval performance is becoming better (generally, the precision is enhanced from less than 60% to more than 70%). To conclude, the deep feature obtains cracking behaviors in RSISR [53], and the expensive learning process needs to be more unsupervised. Hence, creating more effective and cheaper deep feature extraction methods are of great necessity. Although these approaches could largely accelerate searching speeds, the accuracy levels of retrieval cannot practically apply to applications. On average, the cost of an operation is quite small. By contrast, the individual cost of an operation might be quite

high. Moreover, the modeling and learning of deep hashing neural networks (DHNNs) based on specific RSISR tasks deserve more exploration. In the literature, CS RSISR is rarely discussed [82], but CS RS data is increasing ceaselessly, thus stimulating us to further our work on CS RSISR.

5. Scene-Driven Remote Sensing Image Object Detection

As shown in Figure 6, scene-level object detection has, in recent years, been gradually applied in scientific research due to how can extract segmented image information to improve the accuracy of object detection. The scenes trained for deep networks are gained through sliding scenes. By comparing the labeled scene and unlabeled scenes, the training scenes can be labeled, and the object detection result can be achieved. There are several methods used to extract multi-layer features of images by using deep networks. One is based on region nomination, such as Region-based Convolutional Neural Networks (R-CNN) [83], Spatial Pyramid Pooling Net (SPP-net) [84], Fast R-CNN [85], and Region-based Full Connection Network (R-FCN), and the other is end-to-end, which does not require region nomination, including You Only Look Once (YOLO) and Single Shot MultiBox Detector (SSD) [86]. Furthermore, the feature maps from higher layers of the deep neural network display semantic abstracting properties [87].

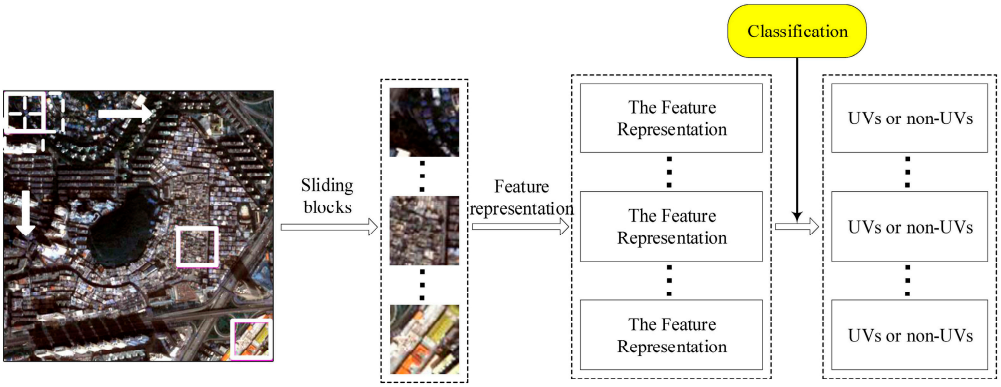

Figure 6. Flowchart of scene-driven object detection: take urban village (UV) detection as an example. The scene trained for deep networks is gained through sliding scenes. By comparing the scene and scenes which have been labeled "UVs", the training scenes can be labeled UVs or non-UVs, and the object detection result can be achieved.

5.1. Taking the Scene as the Primary Unit to Interpret Objects from Remote Sensing Images

For scene-based detection, the primary unit is the image block, and its semantic scene category is determined not only by the contained objects, but also their relative positions. Hence, scene-based classification methods are able to better distinguish complex categories, and can simultaneously consider the relationships between objects.

Li et al. [88] first utilized multi-kernel learning to incorporate multiple features to implement the block-level image interpretation. Then, multi-field integrating was proposed to obtain the block-level result. Subsequently, multi-hypothesis voting was finally utilized for the built-up area detection result through multi-hypothesis super-pixel representation and graph smoothing. Compared with pixel-based Grey-Level Co-occurrence Matrix (GLCM) (89.9%), Li's method can reach 94.3%. In addition, Li [89] presented an unsupervised deep neural network (UDNN), which is composed of an unsupervised deep convolutional neural network (UDCNN) with UDCNN3's accuracy reaching 97.08 ± 1.76% and UDNN3-1-2's accuracy reaching 98.55 ± 0.87%, as well as an unsupervised deep fully connected neural network (UDFNN). The former can hierarchically extract the local features, from simple to complex. UDFNN attaches importance to the merits of fully connected layers by stacking the RBMs, and works to further abstract the feature representation from a global perspective.

In our previous work [90], a double-stream convolutional neural network (DSCNN) model based on Inception V3 was proposed to extract the built-up area automatically. The network consists of two branches: the upper branches manipulate the panchromatic image providing primary information for object classification, and the lower branch attaches more importance to the multispectral image, providing auxiliary information for improvement in accuracy (with the overall accuracy reaching 98.39%).

5.2. Deep Networks under Scene-Level Supervision for Geospatial Object Detection

The intrinsic difference between deep learning methods and traditional visual recognition methods is that the deep learning methods can autonomously learn feature representations from a large amount of data, without much expertise or effort in designing features. The deep learning methods can automatically learn hierarchical feature representations. In the hierarchical feature representations, a simple concept is first learned, and then complex concepts are built successively by being composed of simpler ones. As depicted in Figure 7, this feature abstraction process also accords with the human visual cognition process [89].

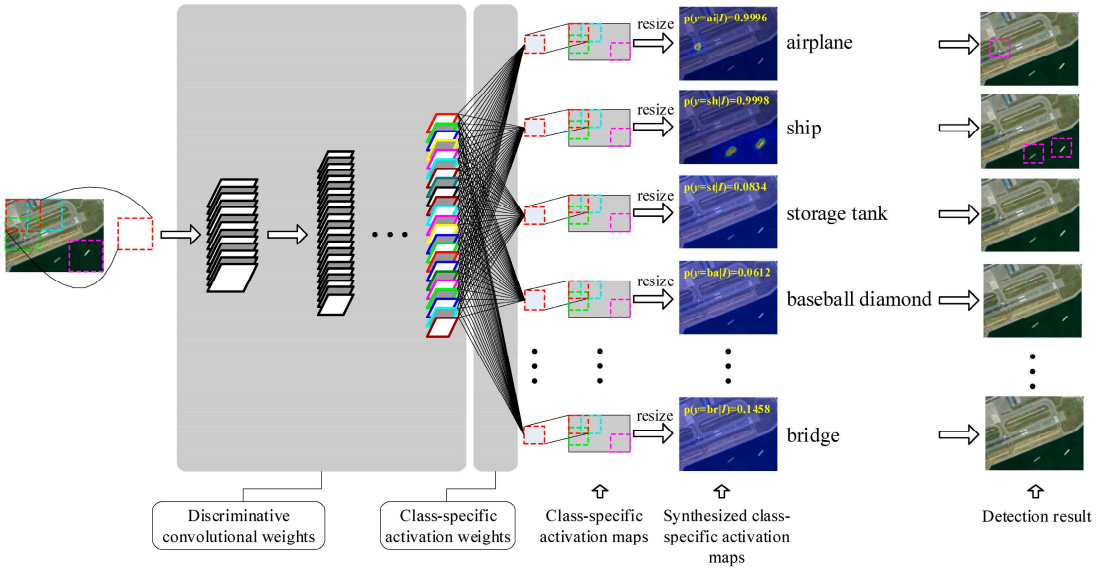

Figure 7. Scene-level supervised object detection based on deep networks.

In [91], weakly supervised learning (WSL) obtains the initial training examples via saliency-based self-adaptive segmentation and negative mining based on weakly labeled remote sensing images (RSIs). On the basis of the trained target detector refined in advance, a candidate-patch-based scheme was adopted so as to detect targets effectively. Analogously, weakly supervised learning (WSL) combined with high-level feature learning [12] was adopted for automatic detection of the optical remote sensing images (RSIs). Furthermore, in a bid to obtained more structural and semantic representation of the image patches, they extracted the low-level and mid-level features to capture the spatial information, and a deep Boltzmann machine (DBM) was utilized to learn the hidden patterns of the mid-level features, even though it can only detect single objects.

Li et al. [92] analysed the drawbacks of weak supervision, which takes scenes as isolated ones and ignores the mutual cues between scene pairs. Hence, an end-to-end multi-class method based on two stages is put forward, which were intended to learn discriminative convolutional weights through exploiting the mutual information between scene pairs, as well as the class-specific activation weights, respectively. Besides, the paper proposes a multi-scale scene-sliding-voting strategy to compute the class-specific activation maps (CAM) and a new set of CAM-oriented segmentation methods

aimed to detect objects. Even under weak supervision, the result of training deep networks has an encouraging result.

To better classify, multi-scale convolutional neural networks (MCNN) [93] can also be used to extract deep features of the scale by considering the context information of the image. MCNN uses a series of images of different scales to perform a pyramid algorithm, and uses the back propagation algorithm [94] to automatically optimize various parameters in a deep network. In addition, because the non-linear feature is often better than the linear feature in representing the characteristics of the object, and in order to better classify, the extracted multi-scale spectral features become non-linear after the action of the non-linear activation function. Compared with classic SVM, whose overall accuracy (OA) is 80.13%, MCNN has an overwhelming advantage in accuracy (96.78%).

6. Future Research Directions

6.1. Unsupervised Model Transfer Cross Different Scene Datasets

In the past, classification methodologies were normally based on pixels or objects. However, with the appearance of the data shift problem (i.e., tagged and untagged images acquired by various sensors of completely different geographical areas), the classification scenarios subjected to it need to be dealt with [34]. The retrieval results from different regions vary dramatically, even if the scenes are born from the same sensor. As revealed, the cross-source RSISU problems should be scene-based and jointly optimized in an unsupervised way. Thus a domain adaptation network was developed in this paper [34]. For the analysis of the tagged and untagged images, the authors employed pre-trained CNN to produce the original feature representation. Then, they threw the resulting features into the other network based on the pre-trained CNN for further learning. In their work, they created a cross-dataset made by aggregating images of the same class from the UC Merced data set and Kingdom of Saudi Arabia (KSA) data set. Obviously, there exists the data shift problem, in that images in the source and target domains are from different places or different sensors. However, the shift is narrowed largely with the aid of the proposed network. At the same time, the overall accuracy of retrieval was enhanced from 73.25% to 91.50% when the source domain data set was KSA and the target domain data set was UC Merced. Besides uni-scene classification, they were able to achieve cross-scene classification.

6.2. Recognition of Unseen Scenes via Knowledge Transfer

Currently, several distinct drawbacks exist in the scene classification of HSR (high spatial resolution) RS images [34]:

- Under no circumstance can a scene class get an appropriate classifier without labeled data using most state-of-art approaches.
- Only messages of HSR RS images can be used. The images of a scene class which are unseen cannot be recognized.
- The names of typical classes are less semantically related than those of the object classes in natural image recognition, which restricts the zero-shot (ZS) learning approaches from being employed in the RS community.

It is normal for humans to be able to identify a new scene class. Since a huge amount of knowledge is contained and delivered to us through texts and other channels in our daily life, the former statement is reasonable. Hence, by combining seen instances and ancillary information (e.g., texts), humans may effortlessly identify new scene classes. We can use ZS learning, which is famous in natural language for transferring knowledge from seen classes to unseen classes based on accessorial information. In this paper, a novel ZS scene classification approach for High Spectral Resolution (HSR) RS images (the word2vec model) changed the names of classes to semantic vectors for the first time [95]. Through sorting all the zero-shot models in the community of RS, only the accuracy of the ZS scene classification

method was able to achieve more than 70%. This indicates that the aforementioned limitations are being dealt with.

6.3. Language-Level Understanding of Remote Sensing Image Scenes

Regarded as a significant but rare task for AI, RS image captioning is more challenging [96], which is actually closer to RSISU itself. The words generated are preliminary comprehensions from the machine, which can be regarded as a bridge linking up computer vision and language.

Instead of simply predicting individual tags, it generates a complete sentence that is comprehensible. This not only means classifying each image scene as scene classification does, but also mining the dependencies between classes. It means adding contextual information between scenes to RSISR. Furthermore, it not only means recognizing objects and analyzing each image scene at a multi-scale, but also finding out the spatial relationship between the objects. Such descriptive sentences are closer to RSISU. By using deep learning and FC networks, an RS image capture framework was constructed in this article. The descriptions produced show wonderful performance. For instance, the lowest accuracy of the scenes in Google Earth images achieved 79%. This inspires us in at least two ways:

- Image Retrieval: Rather than keyword searching, users can further describe their needs and improve the approachability of gathering useful images.
- Military Intelligence Generation: Battlefield images can immediately be converted to text messages by machine. These messages can be delivered to soldiers and help them to fight.

6.4. Greedy Annotation of RS Image Scenes

In addition to the aforementioned zero-shot problem, the low-shot problem (i.e., only a small number of samples are tagged) is more common in RSISU. To target the problem of annotating for remote sensing large data in an unsupervised way, we can solve it by clustering, instead of using knowledge transfer. For example, in the article [97], hierarchical similarity diffusion is used. This method is based on the assumption that scenes in a cluster should have the same label, which is obviously reasonable.

To some degree, it tolerates some wrong annotations. An accuracy level higher than 90% can be attained even when only one-tenth of the scenes are annotated. Therefore, the idea of further solving the low-shot problem in RSISU can diverge from this method.

6.5. Multi-Source Remote Sensing Image Scene Understanding

By coupling multi-source data, RSISU can play a more beneficial role in today's society. True and reliable data sources are scarce in the real developing world, so in order to obtain more accurate socio-economic information [98], we sometimes need to combine existing social survey information with remote sensing image information. In [99], satellite images of luminosity at night ("nightlights" [100]) and country-level economic production statistics were used to train the deep learning model with the method of "transfer learning". Through the method of this article, we were able to properly detect the level of economic development of a certain region, so as to help improve the poverty evaluation.

In addition to the need for testing in poor areas, human residential areas are still largely affected by natural disasters [101]. It is also essential to find the law of the time and location of natural disasters, and to reasonably predict the place and time of possible disasters in the future so that we can assist the relevant departments when making decisions. As depicted in [102], the multi-kernel learning method is combined with 3D point-cloud data (which is known to be a very precise data source that can extract the geometric information of the objects) to detect natural disasters.

Nowadays, remote sensing data sources are still a challenge [103] to the understanding of remote sensing image scenes. Thanks to the further development of WIFI, GPS, and other systems [104], users can upload their location information and corresponding geographic data by Twitter or Instagram. We

should learn how to use "big data" or multi-source data [105], such as OpenStreetMap (OSM) and Google Maps to obtain image information, and apply the deep learning geography theory to these kinds of mass data geosciences. In the long run, learning and transfer learning based on social media and multi-source dataset between different data sources is a promising application.

6.6. Automatic Target Detection under Scene-Tag-Supervision

In many practical applications, it is not a simple task to label the accurate boundary of targets in the infrared imagery [106] or SAR imagery [107]. However, we can easily determine whether one image block (i.e., scene) contains the target of interest or not. Hence, training deep networks under scene-tag-supervision is a promising way to address many challenging target detection tasks (e.g., target detection from infrared imagery or SAR imagery).

7. Conclusions

This comprehensive review may be of benefit to both the industry and academia in the remote sensing domain. In this paper, we introduced the different applications of remote sensing image interpretation based on deep learning. Through sorting out and summarizing the existing research, the latest technologies of scene classification, scene retrieval, and object detection have been more thoroughly analyzed. However, for deep learning, because of the need for a large number of training data sets and speed constraints, even with high operational accuracy, there are still challenges and many areas for improvement.

Author Contributions: Conceptualization, Y.L.; writing—original draft preparation, Y.G., Y.W., and Y.L.; writing—review and editing, Y.G., Y.W., and Y.L.; visualization, Y.G. and Y.W.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L.

Funding: This research was partially supported by the National Natural Science Foundation of China under grant 41601352; the China Postdoctoral Science Foundation under grants 2016M590716 and 2017T100581; the Hubei Provincial Natural Science Foundation of China under grant 2018CFB501.

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

References

1. Liang, M.; Hu, X. Recurrent convolutional neural network for object recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.
2. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
3. Xia, G.S.; Tong, X.Y.; Hu, F.; Zhong, Y.; Datcu, M.; Zhang, L. Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation. *arXiv* **2017**, arXiv:1707.07321.
4. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
5. Chen, S.; Tian, Y.L. Pyramid of Spatial Relatons for Scene-Level Land Use Classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 1947–1957. [CrossRef]
6. 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.
7. Zhang, X.; Du, S. A Linear Dirichlet Mixture Model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sens. Environ.* **2015**, *169*, 37–49. [CrossRef]
8. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]
9. Liu, J.; Shah, M.; Kuipers, B.; Savarese, S. Cross-view action recognition via view knowledge transfer. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3209–3216.

10. Zafar, B.; Ashraf, R.; Ali, N.; Iqbal, M.K.; Sajid, M.; Dar, S.H.; Ratyal, N.I. A Novel Discriminating and Relative Global Spatial Image Representation with Applications in CBIR. *Appl. Sci.* **2018**, *8*, 2242. [CrossRef]
11. Ahmed, K.T.; Irtaza, A.; Iqbal, M.A. Fusion of local and global features for effective image extraction. *Appl. Intell.* **2017**, *47*, 526–543. [CrossRef]
12. 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]
13. Wu, L.; Hoi, S.C.; Yu, N. Semantics-preserving bag-of-words models and applications. *IEEE Trans. Image Process.* **2010**, *19*, 1908–1920. [PubMed]
14. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
15. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random Forests for land cover classification. *Pattern Recognit. Lett.* **2006**, *27*, 294–300. [CrossRef]
16. Kybic, J. Bootstrap resampling for image registration uncertainty estimation without ground truth. *IEEE Trans. Image Process. A Publ. IEEE Signal. Process. Soc.* **2010**, *19*, 64–73. [CrossRef]
17. Boykov, Y.; Veksler, O.; Zabih, R. *Markov Random Fields with Efficient Approximations*; Cornell University: Ithaca, NY, USA, 1997.
18. 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]
19. Xia, G.-S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; Maître, H. Structural High-resolution Satellite Image Indexing. In Proceedings of the ISPRS TC VII Symposium—100 Years ISPRS, Vienna, Austria, 5 July 2010; pp. 298–303.
20. Bei, Z.; 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.
21. Wang, Y.; Zhang, L.; Hao, D.; Lu, J.; Huang, H.; Liang, Z.; Liu, J.; Hong, T.; Xing, X. Learning a Discriminative Distance Metric With Label Consistency for Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4427–4440. [CrossRef]
22. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
23. Li, H.; Tao, C.; Wu, Z.; Chen, J.; Gong, J.; Deng, M. RSI-CB: A Large Scale Remote Sensing Image Classification Benchmark via Crowdsource Data. 2017.
24. Pu, J.; Xia, G.S.; Fan, H.; Lu, Q.; Zhang, L. AID++: An Updated Version of AID on Scene Classification. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018.
25. Gao, F.; Huang, T.; Wang, J.; Sun, J.; Hussain, A.; Yang, E. Dual-Branch Deep Convolution Neural Network for Polarimetric SAR Image Classification. *Appl. Sci.* **2017**, *7*, 447. [CrossRef]
26. Penatti, O.A.B.; Nogueira, K.; Santos, J.A.d. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 44–51.
27. An, T.; Kim, M. A New Diverse AdaBoost Classifier. In Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence, Sanya, China, 23–24 October 2010; pp. 359–363.
28. Keller, J.M.; Gray, M.R.; Givens, J.A. A fuzzy K-nearest neighbor algorithm. *IEEE Trans. Syst. ManCybern.* **1985**, *SMC-15*, 580–585. [CrossRef]
29. Xuming, H.; Zemel, R.S.; Carreira-Perpinan, M.A. Multiscale conditional random fields for image labeling. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; p. II.
30. Yin, J.; Liu, Z.; Jin, Z.; Yang, W. Kernel sparse representation based classification. *Neurocomputing* **2012**, *77*, 120–128. [CrossRef]
31. Rajendra Acharya, U.; Subbanna Bhat, P.; Iyengar, S.S.; Rao, A.; Dua, S. Classification of heart rate data using artificial neural network and fuzzy equivalence relation. *Pattern Recognit.* **2003**, *36*, 61–68. [CrossRef]
32. Hao, Z.; He, L.; Chen, B.; Yang, X. A Linear Support Higher-Order Tensor Machine for Classification. *IEEE Trans. Image Process.* **2013**, *22*, 2911–2920. [CrossRef]

33. Li, W.; Fu, H.; Le, Y.; Peng, G.; Feng, D.; Li, C.; Clinton, N. Stacked Autoencoder-based deep learning for remote-sensing image classification: A case study of African land-cover mapping. *Int. J. Remote Sens.* **2016**, *37*, 5632–5646. [CrossRef]
34. Othman, E.; Bazi, Y.; Melgani, F.; Alhichri, H.; Zuair, M. Domain Adaptation Network for Cross-Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4441–4456. [CrossRef]
35. Lan, Z.; Yang, L. Study on Multi-Scale Window Determination for GLCM Texture Description in High-Resolution Remote Sensing Image Geo-Analysis Supported by GIS and Domain Knowledge. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 175. [CrossRef]
36. Cheng, J.; Li, L.; Luo, B.; Wang, S.; Liu, H. High-resolution remote sensing image segmentation based on improved RIU-LBP and SRM. *Eurasip J. Wirel. Commun. Netw.* **2013**, *2013*, 263. [CrossRef]
37. Cheng, G.; Zhou, P.; Yao, X.; Yao, C.; Zhang, Y.; Han, J. Object detection in VHR optical remote sensing images via learning rotation-invariant HOG feature. In Proceedings of the 2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA), Guangzhou, China, 4–6 July 2016; pp. 433–436.
38. Yin, J.; Hui, L.; Jia, X. Crater Detection Based on Gist Features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 23–29. [CrossRef]
39. Cheng, G.; Ma, C.; Zhou, P.; Yao, X.; Han, J. Scene classification of high resolution remote sensing images using convolutional neural networks. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 767–770.
40. Cheng, G.; Ceyuan, Y.; Xiwen, Y.; Guo, L.; Junwei, H. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
41. 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]
42. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep Feature Fusion for VHR Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4775–4784. [CrossRef]
43. Zang, M.; Wen, D.; Liu, T.; Zou, H.; Liu, C. A Fast Sparse Coding Method for Image Classification. *Appl. Sci.* **2019**, *9*, 505. [CrossRef]
44. Wright, J.; Ma, Y.; Mairal, J.; Sapiro, G.; Huang, T.S.; Yan, S. Sparse Representation for Computer Vision and Pattern Recognition. *Proc. IEEE* **2010**, *98*, 1031–1044. [CrossRef]
45. Li, Y.; Tao, C.; Tan, Y.; Shang, K.; Tian, J. Unsupervised Multilayer Feature Learning for Satellite Image Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 157–161. [CrossRef]
46. Yu, Y.; Zhong, P.; Gong, Z. Balanced data driven sparsity for unsupervised deep feature learning in remote sensing images classification. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 668–671.
47. Ren, Z.; Hou, B.; Wen, Z.; Jiao, L. Patch-Sorted Deep Feature Learning for High Resolution SAR Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3113–3126. [CrossRef]
48. Yu, X.; Wu, X.; Luo, C.; Peng, R. Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework. *Gisci. Remote Sens.* **2017**, *54*, 1–18. [CrossRef]
49. Song, H.O.; Xiang, Y.; Jegelka, S.; Savarese, S. Deep Metric Learning via Lifted Structured Feature Embedding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4004–4012.
50. Cheng, G.; Zhou, P.; Han, J. Duplex Metric Learning for Image Set Classification. *IEEE Trans. Image Process.* **2018**, *27*, 281–292. [CrossRef] [PubMed]
51. Han, J.; Cheng, G.; Li, Z.; Zhang, D. A Unified Metric Learning-Based Framework for Co-Saliency Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2473–2483. [CrossRef]
52. Gong, Z.; Zhong, P.; Yu, Y.; Hu, W. Diversity-Promoting Deep Structural Metric Learning for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 371–390. [CrossRef]
53. Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Unsupervised Deep Feature Learning for Remote Sensing Image Retrieval. *Remote Sens.* **2018**, *10*, 1243. [CrossRef]
54. Zhang, X.; Liang, Y.; Chen, L.; Ning, H.; Jiao, L.; Zhou, H. Recursive Autoencoders-Based Unsupervised Feature Learning for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1928–1932. [CrossRef]

55. Zhou, W.; Li, C. Deep feature representations for high-resolution remote-sensing imagery retrieval. *Remote Sens.* **2016**, *9*, 489. [CrossRef]
56. Qian, B.; Ping, G. Comparative studies on similarity measures for remote sensing image retrieval. In Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), the Hague, the Netherlands, 10–13 October 2004; Volume 1111, pp. 1112–1116.
57. Zhou, W.; Shao, Z.; Diao, C.; Cheng, Q. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. *Remote Sens. Lett.* **2015**, *6*, 775–783. [CrossRef]
58. Boualleg, Y.; Farah, M. Enhanced Interactive Remote Sensing Image Retrieval with Scene Classification Convolutional Neural Networks Model. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 4748–4751.
59. 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]
60. Napoletano, P. Visual descriptors for content-based retrieval of remote-sensing images. *Int. J. Remote Sens.* **2017**, *39*, 1–34. [CrossRef]
61. Ye, F.; Xiao, H.; Zhao, X.; Dong, M.; Luo, W.; Min, W. Remote Sensing Image Retrieval Using Convolutional Neural Network Features and Weighted Distance. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1535–1539. [CrossRef]
62. Hu, F.; Tong, X.; Xia, G.; Zhang, L. Delving into deep representations for remote sensing image retrieval. In Proceedings of the 2016 IEEE 13th International Conference on Signal Processing (ICSP), Chengdu, China, 6–10 November 2016; pp. 198–203.
63. Yandex, A.B.; Lempitsky, V. Aggregating Local Deep Features for Image Retrieval. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 Decmber 2015; pp. 1269–1277.
64. Zhou, W.; Deng, X.; Shao, Z. Region Convolutional Features for Multi-Label Remote Sensing Image Retrieval. *arXiv* **2018**, arXiv:1807.08634.
65. Ben Ahmed, O.; Urruty, T.; Richard, N.; Christine, F.-M. Toward Content-Based Hyperspectral Remote Sensing Image Retrieval (CB-HRSIR): A Preliminary Study Based on Spectral Sensitivity Functions. *Remote Sens.* **2019**, *11*, 600. [CrossRef]
66. 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]
67. Wang, M.; Song, T. Remote Sensing Image Retrieval by Scene Semantic Matching. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2874–2886. [CrossRef]
68. Aksoy, S. Modeling of Remote Sensing Image Content Using Attributed Relational Graphs. In Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition, Berlin, Heidelberg, 17 August 2006; pp. 475–483.
69. Chaudhuri, B.; Demir, B.; Bruzzone, L.; Chaudhuri, S. Region-Based Retrieval of Remote Sensing Images Using an Unsupervised Graph-Theoretic Approach. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 987–991. [CrossRef]
70. Chaudhuri, B.; Demir, B.; Chaudhuri, S.; Bruzzone, L. Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 1144–1158. [CrossRef]
71. Wang, Y.; Zhang, L.; Tong, X.; Zhang, L.; Zhang, Z.; Liu, H.; Xing, X.; Mathiopoulos, P.T. A Three-Layered Graph-Based Learning Approach for Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6020–6034. [CrossRef]
72. Tang, X.; Jiao, L.; Emery, W.J.; Liu, F.; Zhang, D. Two-Stage Reranking for Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5798–5817. [CrossRef]
73. Xiao, Z.; Long, Y.; Li, D.; Wei, C.; Tang, G.; Liu, J. High-Resolution Remote Sensing Image Retrieval Based on CNNs from a Dimensional Perspective. *Remote Sens.* **2017**, *9*, 725. [CrossRef]
74. 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]
75. 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]
76. Li, P.; Ren, P. Partial Randomness Hashing for Large-Scale Remote Sensing Image Retrieval. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 464–468. [CrossRef]

77. Roy, S.; Sangineto, E.; Demir, B.; Sebe, N. Deep Metric and Hash-Code Learning for Content-Based Retrieval of Remote Sensing Images. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 4539–4542.
78. Chen, C.; Zou, H.; Shao, N.; Sun, J.; Qin, X. Deep Semantic Hashing Retrieval of Remotec Sensing Images. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1124–1127.
79. Zhang, J.; Chen, L.; Zhuo, L.; Liang, X.; Li, J. An Efficient Hyperspectral Image Retrieval Method: Deep Spectral-Spatial Feature Extraction with DCGAN and Dimensionality Reduction Using t-SNE-Based NM Hashing. *Remote Sens.* **2018**, *10*, 271. [CrossRef]
80. Chen, Y.; Crawford, M.M.; Ghosh, J. Applying nonlinear manifold learning to hyperspectral data for land cover classification. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005 IGARSS'05, Seoul, Korea, 29–29 July 2005; pp. 4311–4314.
81. Li, Y.; Zhang, Y.; Xin, H.; Hu, Z.; Ma, J. Large-Scale Remote Sensing Image Retrieval by Deep Hashing Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 950–965. [CrossRef]
82. Li, Y.; Zhang, Y.; Huang, X.; Ma, J. Learning Source-Invariant Deep Hashing Convolutional Neural Networks for Cross-Source Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6521–6536. [CrossRef]
83. Cao, Y.; Niu, X.; Dou, Y. Region-based convolutional neural networks for object detection in very high resolution remote sensing images. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, 13–15 August 2016; pp. 548–554.
84. Liu, Q.; Hang, R.; Song, H.; Zhi, L. Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *56*, 117–126. [CrossRef]
85. Yun, R.; Zhu, C.; Xiao, S. Small Object Detection in Optical Remote Sensing Images via Modified Faster R-CNN. *Appl. Sci.* **2018**, *8*, 813.
86. Zhang, W.C.; Chen, Z.P. Research for the SSD-Based Storage Technology of Mass Remote Sensing Image. *Adv. Mater. Res.* **2012**, *532–533*, 1339–1343. [CrossRef]
87. Bei, Z.; 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.
88. Li, Y.; Tan, Y.; Yi, L.; Qi, S.; Tian, J. Built-Up Area Detection From Satellite Images Using Multikernel Learning, Multifield Integrating, and Multihypothesis Voting. *IEEE Geosci. Remote Sens. Lett.* **2017**, *12*, 1190–1194.
89. Li, Y.; Huang, X.; Liu, H. Unsupervised Deep Feature Learning for Urban Village Detection from High-Resolution Remote Sensing Images. *Photogramm. Eng. Remote Sens.* **2017**, *83*, 567–579. [CrossRef]
90. Tan, Y.; Xiong, S.; Li, Y. Automatic Extraction of Built-Up Areas from Panchromatic and Multispectral Remote Sensing Images Using Double-Stream Deep Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3988–4004. [CrossRef]
91. Zhang, D.; Han, J.; Gong, C.; Liu, Z.; Bu, S.; Lei, G. Weakly Supervised Learning for Target Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 701–705. [CrossRef]
92. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 182–196. [CrossRef]
93. Zhao, W.; Du, S. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *113*, 155–165. [CrossRef]
94. Heermann, P.D.; Khazenie, N. Classification of multispectral remote sensing data using a back-propagation neural network. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 81–88. [CrossRef]
95. Xia, G.S.; Hu, J.; Fan, H.; Shi, B.; Xiang, B.; Zhong, Y.; Zhang, L. AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *PP*, 1–17. [CrossRef]
96. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
97. Li, Y.; Ye, D. Greedy Annotation of Remote Sensing Image Scenes Based on Automatic Aggregation via Hierarchical Similarity Diffusion. *IEEE Access* **2018**, *6*, 57376–57388. [CrossRef]
98. Doll, C.N.H.; Muller, J.-P.; Morley, J.G. Mapping regional economic activity from night-time light satellite imagery. *Ecol. Econ.* **2006**, *57*, 75–92. [CrossRef]

99. Jean, N.; Burke, M.; Xie, M.; Davis, W.M.; Lobell, D.B.; Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science* **2016**, *353*, 790. [CrossRef]
100. Zhou, Y.; Smith, S.J.; Elvidge, C.D.; Zhao, K.; Thomson, A.; Imhoff, M. A cluster-based method to map urban area from DMSP/OLS nightlights. *Remote Sens. Environ.* **2014**, *147*, 173–185. [CrossRef]
101. Joyce, K.E.; Belliss, S.E.; Samsonov, S.V.; McNeill, S.J.; Glassey, P.J. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Prog. Phys. Geogr. Earth Environ.* **2009**, *33*, 183–207. [CrossRef]
102. Vetrivel, A.; Gerke, M.; Kerle, N.; Nex, F.; Vosselman, G. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS J. Photogramm. Remote Sens.* **2017**, *140*, 45–59. [CrossRef]
103. Ma, Y.; Wu, H.; Wang, L.; Huang, B.; Ranjan, R.; Zomaya, A.; Jie, W. Remote sensing big data computing: Challenges and opportunities. *Future Gener. Comput. Syst.* **2015**, *51*, 47–60. [CrossRef]
104. Chi, M.; Plaza, A.; Benediktsson, J.A.; Sun, Z.; Shen, J.; Zhu, Y. Big Data for Remote Sensing: Challenges and Opportunities. *Proc. IEEE* **2016**, *104*, 2207–2219. [CrossRef]
105. Zhang, J. Multi-source remote sensing data fusion: Status and trends. *Int. J. Image Data Fusion* **2010**, *1*, 5–24. [CrossRef]
106. Li, Y.; Zhang, Y. Robust infrared small target detection using local steering kernel reconstruction. *Pattern Recognit.* **2018**, *77*, 113–125. [CrossRef]
107. Tan, Y.; Li, Q.; Li, Y.; Tian, J. Aircraft Detection in High-Resolution SAR Images Based on a Gradient Textural Saliency Map. *Sensors* **2015**, *15*, 23071–23094. [CrossRef]