applied sciences

MDPI

Review

Geographic Scene Understanding of High-Spatial-Resolution Remote Sensing Images: Methodological Trends and Current Challenges

Peng Ye [1,2,3,4], Guowei Liu [2,*] and Yi Huang [5,6]

[1] Urban Planning and Development Institute, Yangzhou University, Yangzhou 225127, China; 007839@yzu.edu.cn
[2] College of Architectural Science and Engineering, Yangzhou University, Yangzhou 225127, China
[3] Key Lab of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China
[4] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[5] Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; huangyi@njupt.edu.cn
[6] School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
* Correspondence: arguoweiliu@yzu.edu.cn; Tel.: +86-155-2142-2026

Abstract: As one of the primary means of Earth observation, high-spatial-resolution remote sensing images can describe the geometry, texture and structure of objects in detail. It has become a research hotspot to recognize the semantic information of objects, analyze the semantic relationship between objects and then understand the more abstract geographic scenes in high-spatial-resolution remote sensing images. Based on the basic connotation of geographic scene understanding of high-spatial-resolution remote sensing images, this paper firstly summarizes the keystones in geographic scene understanding, such as various semantic hierarchies, complex spatial structures and limited labeled samples. Then, the achievements in the processing strategies and techniques of geographic scene understanding in recent years are reviewed from three layers: visual semantics, object semantics and concept semantics. On this basis, the new challenges in the research of geographic scene understanding of high-spatial-resolution remote sensing images are analyzed, and future research prospects have been proposed.

Keywords: geographic scene; high-spatial-resolution remote sensing image; scene understanding; semantic hierarchy of geographic scene; remote sensing image processing

Citation: Ye, P.; Liu, G.; Huang, Y. Geographic Scene Understanding of High-Spatial-Resolution Remote Sensing Images: Methodological Trends and Current Challenges. *Appl. Sci.* **2022**, *12*, 6000. https://doi.org/10.3390/app12126000

Academic Editor: Amerigo Capria

Received: 5 May 2022
Accepted: 8 June 2022
Published: 13 June 2022

1. Introduction

Remote sensing, as a comprehensive modern surveying and mapping technology, plays an important role in Earth observation. In recent years, as a result of the rapid development of sensor technology, aerospace platform technology and data communication technology, as well as the vigorous promotion of relevant international organizations, the global observation capability of the space–air–ground integration has been greatly enhanced [1]. At present, a large number of high-spatial-resolution (HSR) remote sensing images with meters, or even sub-meters, can be obtained. In HSR remote sensing images, various realistic geographic scenes are clearly presented: for instance, artificial construction scenes such as urban residential areas, ports and airports; disaster scenes such as landslides, mudslides and earthquakes; and natural scenes such as forests and beaches [2]. This small-scale observation means that HSR remote sensing images can provide more complex surface structure information and more sophisticated texture and size information. Consequently,

it has been applied to urban planning, disaster management, environmental monitoring, military activities and many other fields [3,4].

As a collection of multiple objects and their surroundings in the real world, understanding the semantics of scenes is an important task in remote sensing image interpretation. Scene understanding is based on the perception of remote sensing image data, combined with visual analysis, image processing, pattern recognition and other technical means, to mine the characteristics and patterns in the image from different levels such as computational statistics, behavioral cognition and semantics, so as to realize the effective analysis, cognition and representation of the scene. However, due to the limitations of space imaging technology, HSR remote sensing images, although of higher spatial resolution, are relatively deficient in spectral information [5]. In HSR remote sensing images, the spectral heterogeneity of the same type of ground objects is enhanced, and the spectral diversity of different ground objects is reduced, which leads to the decline in statistical separability of different types of ground objects in the spectral domain [6]. Therefore, understanding the geographic scenes in the HSR remote sensing images includes the identification of both objects and the relationships between objects, as well as the analysis of themes categories with richer concepts and content implied in the geographic scene. Because of the complexity and intersection of these tasks, the research on geographic scene understanding of HSR remote sensing images still faces many challenges, mainly including the following three aspects:

(1) In terms of the basic principles of geographic scene understanding, the machine will identify the objects or targets contained in the scene according to the similarity of image data. In contrast, humans analyze the semantic information of scene content through the category and spatial distribution of ground objects, and form high-level features through abstract concepts [7]. There is a semantic gap between the conceptual similarity of human understanding and the digital storage form similarity of machine identify. This makes it impossible to relate low-level visual features (such as color, shape, texture, etc.) to high-level semantic information directly.
(2) In terms of the data characteristics of HSR remote sensing images, the improved spatial resolution makes the ground objects in the images have more fine texture features, more obvious geometric structure and clearer location layout. Correspondingly, it also aggravates the difficulty of data processing in intelligent image interpretation. In high resolution images, the spectral heterogeneity of similar objects is enhanced, and the spectral difference of different objects is reduced. This leads to a decrease in the statistical separability of different ground objects in the spectral domain [8]. A high resolution does not necessarily promote an improvement in interpretation accuracy.
(3) In terms of the sophistication of geographic scenes, the structure and composition of the geographic scenes in the HSR remote sensing images are complex, highly variable and even messy. The types of geographic scenes with the same ground objects may be different. However, different types of ground objects also appear in similar geographic scenes [9]. Consequently, understanding the semantic information of the geographic scene and constructing the corresponding semantic feature description is crucial.

As an extension of remote sensing image interpretation, the complexity and comprehensiveness of geographic scene understanding based on HSR images is beyond the general processing task of remote sensing. Although significant progress has been made in the research of feature extraction, target detection, scene classification and other sub-tasks, these sub-tasks lack a unified framework to cross the "semantic gap" to understand the high-level semantics of the geographic scenes. Thus, it is necessary to integrate these sub-tasks according to the human cognitive model in understanding the geographic scenes of HSR remote sensing images. In recent years, many researchers who are engaged in computer vision have realized the importance of a "holistic understanding" of geographic scenes and put forward the research approaches of task integration and feature integration. However, there is no systematic research on the geographic scene of HSR remote sensing images as a comprehensive and complete field of intelligent information processing. This paper is focused on answering the following research questions:

(1) What are the objectives of geographic scene understanding?
(2) How are remote sensing approaches being used for geographic scene understanding?
(3) What are the current gaps in HSR remote-sensing-based geographic scene understanding?

The rest of the paper is organized as follows: Section 2 describes the basic ideas of geographic scene understanding based on HSR remote sensing images; Section 3 presents the semantic understanding approaches of the visual layer; Section 4 presents the semantic understanding approaches of the object layer; Section 5 presents the semantic understanding approaches of the concept layer; and Section 6 discusses the open problems and challenges in the future. The paper closes with a conclusion in Section 7.

2. Basic Ideas

The concept of geographic scene understanding of remote sensing images includes two aspects, namely, "remote sensing image understanding" and the "geographic scene". Remote sensing image understanding is a cognitive process to realize the objective things and their laws reflected by remote sensing images through observing, distinguishing, identifying and reasoning remote sensing images and interpreting the content of remote sensing images semantically. The geographic scene is a regional complex with a specific structure and function, which is composed of various natural and human factors in a certain region [10]. In remote sensing images, the geographic scene is a closed region composed of different ground objects. The geographic scene generally involves three aspects: (1) the constituent elements of the scene structure, (2) the relationship between these elements and (3) the function of the set of these elements. Therefore, the research object of geographic scene understanding of remote sensing images is the regional complex composed of ground objects with certain spatial distribution patterns. The research objective is to interpret the research object as a series of meaningful and understandable semantic information.

HSR remote sensing images can reflect the more detailed composition and spatial distribution of the ground objects, which is a microcosm of the real geographic scene. Owing to increased spatial resolution and unique imaging methods, geographic scenes in HSR remote sensing images have the following characteristics:

(1) The categories of ground objects in the geographic scene are diverse. The same category of geographic scene can contain different ground objects, and different geographic scenes can also contain the same ground objects. Different objects also have different characteristics in terms of spectrum, texture and structure [11].
(2) The categories of ground objects in the geographic scene have variability. The change in the categories of some ground objects in geographic scenes does not necessarily lead to a change in the whole semantic information of geographic scenes [12].
(3) The spatial relationship between ground objects in geographic scenes is complex. Different distribution forms between ground objects lead to different semantic information of geographic scenes. Other relevant characteristics are shown in Figure 1.

The characteristics of HSR remote sensing images also make the following special features in understanding geographic scenes:

(1) The semantic information of geographic scenes in HSR remote sensing images is hierarchical. The content description of HSR remote sensing images has the hierarchical inclusion relation of "Pixel-Region-Target-Scene". Different levels of image content reflect the semantic information with different levels of abstraction, which can be divided into the visual layer, object layer and concept layer (Figure 2). The visual layer is the description of pixel-level image content, including color, texture, shape and other original visual characteristics. The semantic information of the visual layer can be obtained directly from image processing without any external knowledge and experience [13]. The object layer is the description of region-level and target-level image content, including the individual features of objects and the local features of spatial relations among objects [14]. The semantic information of the object layer needs to be obtained through simple reasoning, and it is necessary to use external knowledge and experience to assist this reasoning. The concept layer is a description of the scene-level image content, including the abstract attributes of

the image. The semantic information of scene level involves the semantic features of scene representation or higher-level behavior or emotion analysis, and it needs to link image content with abstract concepts through complex reasoning and subjective judgment [15].

Figure 1. Image characteristics of geographic scenes in HSR remote sensing: (**a**) reflects the diversity of the ground objects in the geographic scene; (**b**) reflects the diversity of imaging conditions of HSR remote sensing images; (**c**) reflects the differences in the types of ground objects in the same category of geographic scenes; (**d**) reflects the similarity in the types of ground objects in different categories of geographic scenes.

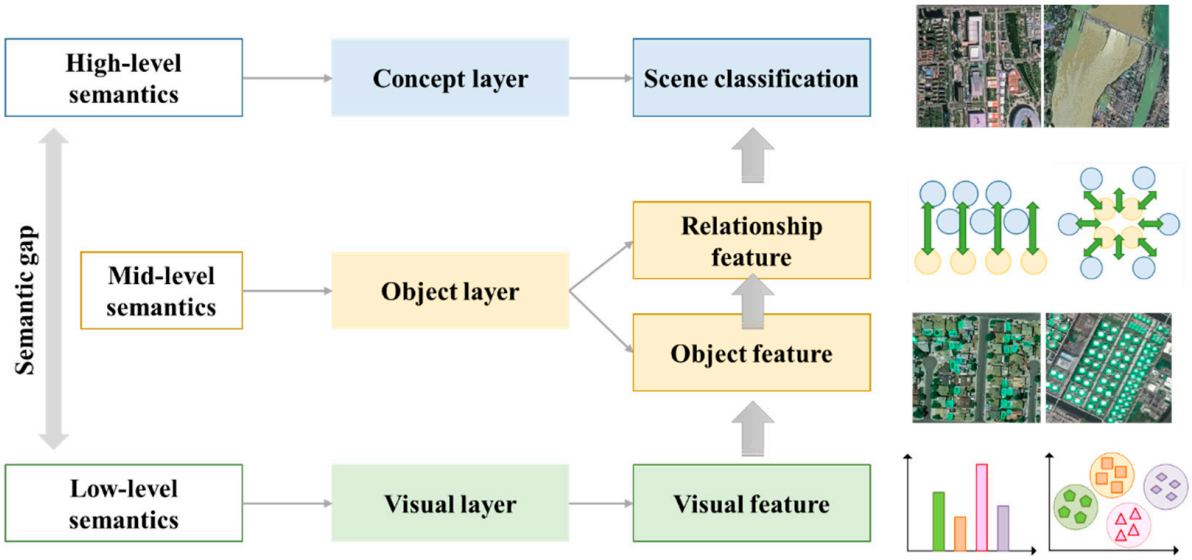

Figure 2. Semantic hierarchy model of three layers structure.

There is a dialectical relationship between the tasks of geographic scenes understanding at different semantic layers. The input of geographic scene understanding is the original HSR remote sensing images, and the output is semantic information of the geographic scene. In the tasks of geographic scene understanding, it is necessary to combine the semantic processing of the visual layer and object layer (feature extraction and target recognition) with the semantic reasoning of the concept layer (scene description and classification), the different tasks are interdependent on each other [16]. The cognition of visual and object layer semantics can form the inference of concept layer semantics, and the cognition of concept layer semantics can be used as knowledge to guide the extraction of visual and object layer semantics.

(2) Spatial structure characteristics play an important role in the geographic scene understanding of HSR remote sensing images. Because of the global and polysemy of the geographic scene, the geographic scene understanding is not a simple stacking of some local semantics [17]. In the same category of the geographic scene, the objects of the same type have similar individual characteristics and spatial distribution patterns. However, there are different structural features among different categories of geographic scenes. In Figure 3, these two geographic scenes of "residential area" and "industrial area" contain similar visual features and object types, which are composed of buildings, roads and vegetation. However, there are great differences in spatial structures between objects, which is a critical factor to distinguish the categories of geographic scenes. Therefore, the spatial structure characteristics of geographic scenes are relatively stable, and making full use of spatial information such as geometry, texture and context of HSR remote sensing images is an effective way to improve the understanding of geographic scenes [18,19].

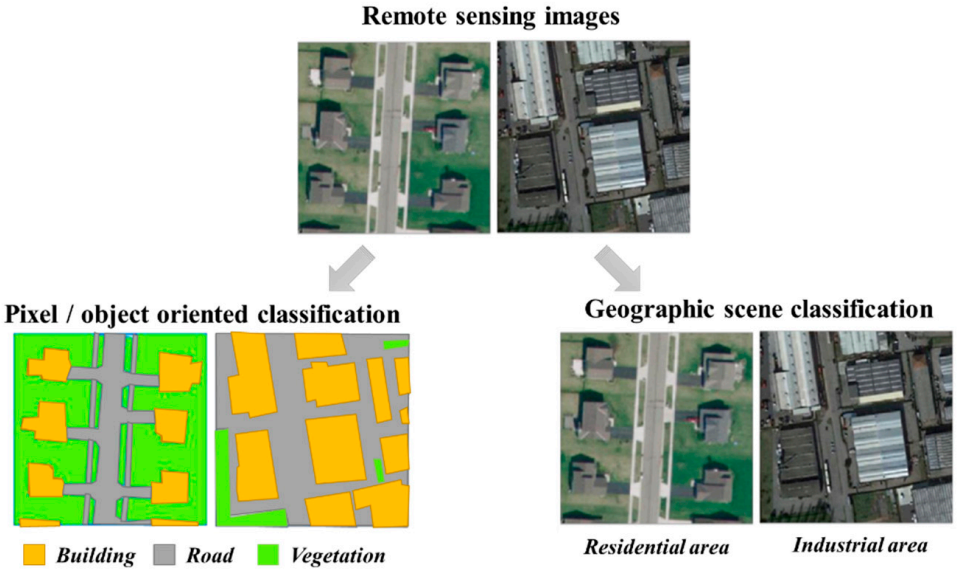

Figure 3. Understanding differences of geographic scenes at different semantic layers.

(3) The data characteristics of HSR remote sensing images have both opportunities and challenges for the geographic scene understanding. The amount of HSR remote sensing images increases significantly. As the spatial resolution increases, the area of the ground covered by each pixel decreases significantly. This makes the ground object details and spatial distribution of HSR remote sensing images clearer. Compared with medium–low-resolution remote sensing images, HSR remote sensing images can be interpreted at the scene level, where semantic information is more abstract. However, compared with natural images, the HSR remote sensing images used for geographic scene understanding are less accessible in terms of data availability, except for the differences in shooting distance, shooting angle and imaging sensors. Natural images can be easily and quickly obtained from the Internet, and a large amount of data has been given the relevant label information when uploaded to the Internet [20]. For example, the ImageNet dataset [21,22] contains more than 14 million labeled samples in 1000 categories. HSR remote sensing images are not freely available for political, military and security reasons. HSR remote sensing images often rely on professional interpretation or even field research to obtain the correct label, the available sample size is limited. For instance, the UC-Merced dataset contains only 21 categories, with a total of 2100 labeled samples. Currently, the HSR remote sensing image datasets commonly used for geographic scene understanding are shown in Table 1.

Table 1. Comparison of existing public datasets.

Datasets	Spatial Resolution (m)	Image Size	Number of Categories	Number of Samples per Category	Total Number of Samples	Year of Publication
UC-Merced [23]	0.3	256 × 256	21	100	2100	2010
WHU-RS19 [24]	0.5	600 × 600	12	50	950	2010
RSSCN7 [25]	-	400 × 400	7	400	2800	2015
RSC11 [26]	0.2	512 × 512	11	About 100	1232	2016
SIRI-WHU [27]	2	200 × 200	12	200	2400	2016
NWPU-RESISC45 [28]	0.2–30	256 × 256	45	700	31,500	2017
PatternNet [29]	0.062–4.693	256 × 256	38	800	-	2017
AID [30]	-	-	30	-	10,000	2017
EuroSAT [31]	-	64 × 64	10	2000–3000	27,000	2019

3. Semantic Understanding of Visual Layer

The semantic understanding of the visual layer of the geographic scene is the extraction of basic characteristics from remote sensing image data. The essence of geographic scene understanding is to establish the mapping relationship between low-level visual features and high-level scene semantics. Thus, extracting the visual features of HSR remote sensing images is the basis of the content description of geographic scenes, which includes local features and global features.

The global feature is the feature that can represent the whole image, and it has good invariance, simple calculation and intuitive representation. Common global features include color features, texture features and shape features. Among them, color features such as color sets, color moments, color correlation diagrams, color histograms [32] and color aggregation vectors [33–35] are insensitive to size and orientation and have good stability [36]. Texture features such as gray level co-occurrence matrix (GLCM) [37], the grayscale difference [38], autocorrelation function [39], gray-level run-length [40], local binary pattern (LBP) [41], etc., are characterized by local irregularity, but macroscopic regularity. Visual features cannot only describe the basic attributes of the image such as color, texture and shape, but also reflect the deep structure information of the image. In 2001, GIST was proposed by Aude Oliva et al. simulating human vision to roughly extract the image and its context information [42]. GIST can extract spectral information from the image globally as its representation without segmenting the image or detecting the target in advance. GIST is simple and easy to use. However, with the increasing complexity of image content and structure, such as the analysis granularity being too coarse to ignore the details of the objects in the scene, the result of image processing is far from the correct result. In general, the global features are sensitive to the actual imaging conditions, and the robustness and generalization ability are relatively poor.

The local feature can effectively resist various affine transformations and have some invariance. David Lowe has come up with a landmark local feature descriptor, the scale invariant feature transform (SIFT), which has good scale invariance and rotational invariance [43]. Thus, SIFT is one of the most widely used features in image processing. Bay et al. present an accelerated robust feature descriptor (SURF) inspired by SIFT [44]. While SURF is inferior to SIFT in scale scaling and rotational invariance, it is superior in blur and illumination variation and is several times faster than SIFT. With the advancement of research, the instability of color, light and gradient features in the process of recognition became an obstacle to image classification. The histograms of oriented gradients (HOG) feature proposed in 2005 continue the high recognition accuracy characteristic of the local feature; the gradient histogram method is used to effectively solve the problem of the low recognition rate of local scene contours due to the sensitivity of light and gradient features [45]. However, HOG features have high dimensions, low computational efficiency and great redundancy and do not consider the effect of scale transformation on classification results. The CENTRIST feature proposed by Wu et al. in 2010 solves this problem well [46]. Through the census transformation of the acquired pixels, these pixels are transformed into statistical histograms to form the CENTRIST feature to extract the object's local shape

structure. After the CENTRIST transformation, the image still retains the global and local structure information. Therefore, it can simulate the human visual system and describe the shape and texture of objects accurately.

Global features and local features have their own advantages and disadvantages. Different visual features are suitable for different tasks of geographic scene understanding. In HSR remote sensing image description, visual feature extraction should not only keep the invariance of features but should also fuse the spatial structure information of the features.

4. Semantic Understanding of Object Layer

Object layer semantics mainly describe the logical concepts of scenes in images, usually based on a large number of visual layer descriptors. Compared with the visual layer, the object layer is closer to the human understanding of the geographic scene. For instance, in the process of geographic scene understanding, we rely more on the houses and roads in the images, rather than recognizing that there are small dense and regular highlighted areas, narrow and long gray-banded areas and so on in the image. Houses, roads, sky and grass, which conform to human cognition, constitute the object layer semantics of geographic scenes. In addition, object layer semantics can also be abstract local areas, such as visual words generated by feature detection algorithms. There is also a certain context structure between different objects, forming a corresponding spatial relationship. Thus, a geographic scene is a combination of a set of specific objects. According to the different semantic forms of objects, there are three types: target object semantics, local area semantics and spatial structure semantics.

4.1. Target Object Semantics

The object layer semantics of the geographic scene are usually concentrated on the basic level of human cognition, which can be represented by many target objects (Figure 4). For the semantic understanding of the target object, it is necessary to use the target detection algorithm to clarify the types of each object. In the fields of computer vision and pattern recognition, many target detection algorithms have been developed, for instance: the threshold-based detection method [47], the template-based detection method [48], target detection based on Hough transform or Hough forest [49], target detection based on classifiers [50], etc. For the target detection of HSR remote sensing images, the method of target detection in the computer vision field is usually used for reference, and the research is carried out around the object of special interest, in particular, the artificial structures closely related to human activities, such as buildings [51], ports [52], airport runways [53], roads [54], warships [55] and so on. For artificial structures with obvious shape features, it is generally possible to directly use their unique shape features for detection, for instance, extracting straight lines to detect linear targets in images [56]. For complex targets, the corresponding models can be constructed; for instance, the "Building" target model can be constructed by texture, shape and SIFT features, and the "Port" target model can be constructed by combining the information of coastline, wharf and embankment [57].

For the target detection of HSR remote sensing images, it is more challenging to detect objects with large image sizes and various details. Target detection of HSR remote sensing images is studied from different perspectives. A multi-layer SVM classifier is used to exclude non-target regions to improve the speed of target detection in high-resolution remote sensing images [58]. The large remote sensing image is divided into smaller blocks, the salient and synopsis features of each block are extracted, and the target detection is realized by classification [59]. Target detection is also accomplished by first segmenting HSR remote sensing images and then merging regions related to the target based on knowledge [60]. In addition, the successful application of visual selective attention to the target location in large-format remote sensing images, and the results show that the visual attention mechanism can quickly focus on the place where the object to be detected appears in the complex large image. These methods are all beneficial explorations in the target detection of HSR remote sensing images [61].

Figure 4. Target detection schematic. (**a**) is the original remote sensing image; (**b**) reflects that the candidate locations of the targets are found in the image; (**c**) reflects the results of target identifications.

The existing target detection methods have strong pertinence and lack universal and robust target detection models and algorithms. The motion characteristics of the target to be detected (such as ship wake, submarine track), the use of the shadow of the target in the image, the removal of the visible cloud cover and so on need to improve the target detection model with pertinence. To realize the practicality of target detection, it is necessary to establish a target detection model and a fast algorithm for multi-source data fusion.

4.2. Local Area Semantics

Remote sensing images can also be divided according to specific rules. By extracting the local image descriptor of each sub-block, the correspondence between the local descriptor and the local semantic concept is established, and the object layer semantics are extracted. Due to the differences of descriptors, feature extraction methods of the local area can be divided into three categories: visual dictionary, feature mapping and topic model.

4.2.1. Visual Dictionary

The visual dictionary, also known as the visual codebook, maps feature data onto individual codewords to generate feature vectors with codebook length [62]. The construction of a visual dictionary is essentially a cluster problem, and the visual codewords correspond to the cluster center. In the task of geographic scene understanding, the visual dictionary connects the image visual features with the scene semantics.

Whether the design of a visual dictionary is effective mainly includes three aspects: resolution, compactness and universality. (1) The resolution of the visual dictionary is reflected in the similarity between visual words. The lower the similarity, the higher the resolution. (2) The compactness is reflected in the choice of codebook length, which corresponds to different classification accuracies. A high recognition rate can be achieved by selecting a suitable visual dictionary. (3) The universality mainly refers to whether the visual dictionary needs to be relearned if the data of new categories are added. Existing dictionary learning includes generative (unsupervised) and discriminative (supervised) approaches. Perronnin et al. design a universal visual dictionary and a category visual dictionary to compete for the description of image content. The universal visual dictionary is used to describe all image scene classes, and the category visual dictionary for a certain scene class can be obtained by adaptive learning from the universal visual dictionary [63]. If an image belongs to a given class, a category visual dictionary is more suitable for describing the image than a general visual dictionary. On the contrary, the general visual dictionary is more suitable to describe the image than the category visual dictionary. However, traditional visual dictionaries are prone to a lack of clear meaning or polysemy. To solve the above problems, Su et al. use semantic attributes to clarify semantic meaning and integrate semantic attributes into the visual dictionary to remove the ambiguity of visual words [64].

4.2.2. Feature Mapping

After constructing the visual dictionary, it is necessary to encode and map the local features of the image, and to represent the semantic information of the image by transforming the local features into some organized form of visual words. In addition, there are some problems such as the low efficiency of dictionary generation, serious quantization errors and the lack of spatial information of visual words. Furthermore, the image semantic representation based on the visual dictionary is a linear representation, which only performs well in the case of the classifier with the nonlinear kernel, such as support vector machine (SVM). This will undoubtedly reduce its usefulness, making it difficult to apply to large-scale data set classifications. In recent years, a semantic representation based on feature mapping has attracted more attention. Feature mapping is used to quantize and code the visual features according to the visual words and generate the representation of the visual features in the visual dictionary.

Vector quantization (VQ) is simple and convenient, but its constraint conditions are too strict, resulting in the lack of information after visual feature quantization. To overcome this shortcoming, the sparse regularization approach can be used to loosen the constraints in the VQ, which translates into a sparse coding [65,66]. Sparse coding (SC) uses a sparse regularization method to reduce quantization errors and improve the uniqueness of feature coding. However, sparse coding is only a shallow learning model with a single hidden layer. The visual dictionary acquired by shallow learning lacks the selectivity of features, which will reduce the semantic resolution of image content. On the basis of SC, the local-constrained linear coding (LLC) is proposed [67]. The sparsity of feature coding cannot guarantee its locality, while the locality of feature coding can guarantee its sparsity. As a result, LLC is more efficient and has a better refactoring effect and local smooth sparsity [68].

4.2.3. Probabilistic Topic Model

In order to improve the performance of image semantic expression, a visual language model is proposed [69], which is inspired by the probabilistic topic model (PTM) of natural language understanding. Based on the visual language model, an image can be divided into many blocks as visual words according to certain rules, and these visual words have certain grammatical rules and spatial dependencies, also called visual grammar. The semantic information is represented by the co-occurrence frequency and spatial dependence of local features in the image. Common PTMs include probabilistic Latent Semantic Analysis (pLSA) [70] and Latent Dirichlet Allocation (LDA) [71].

To ameliorate the robustness of the visual language model to the change of target scale, Wu et al. extended the original model to multi-scale, and proposed the scale-invariant visual language model (m-VLM) [72]. Jing et al. use LDA to realize the scene classification of optical remote sensing images and compare it with the bag of visual words (BOVW) model [73]. The results show that LDA can provide more concise and abundant semantic information for image representation. In the parameter training stage, the probability of a visual language model is estimated by counting the frequency of a visual word or visual word combinations in the image. This approach equates the visual words in the target area of the image with the visual words in the background area, thus ignoring the negative impact of background noise on the target semantic representation [74]. Therefore, if we can distinguish the visual words in the background and assign the weight according to their contribution to the target, we can enhance the resolution of the visual language model to image semantic representation.

4.3. Spatial Structure Semantics

The different arrangements of the objects that comprise the geographic scene will make the geographic scene have a different spatial structure. Spatial structure information in HSR images is contained in spectral features and prior knowledge. For the understanding of spatial structure semantics, it is necessary to describe, model and extract them and obtain

a vector model representing the structural features of processing units (pixels, primitives and targets).

4.3.1. Pixel-Neighborhood-Window-Based Method

Taking a pixel as the basic processing unit, a window is defined for each pixel (also called the central pixel) in the image, which describes the spatial distribution pattern of the pixel values in the window area. The spatial structure features of the pixels in the window area are used as the spatial structure features of the central pixel [75]. This method describes the spatial structure of pixel neighborhoods. It can make up for the lack of spectral information of the central pixel by using the information of neighboring pixels, but it is important and uncertain for the reasonable selection of window size [76].

Among the existing methods, two kinds of neighborhood structure patterns are common. One is the interactive mode between the central pixel and its neighbor pixels, which is a "one-to-multiple" relationship. This relationship is represented using methods such as random fields, local spatial autocorrelation statistics and data fields [77,78]. The other is the spatial structure relationship of multiple pixels in the neighborhood window. The method equates the center pixel with its neighbor pixel, and is a "multiple-to-multiple" relationship, which is represented using methods such as the gray level co-occurrence matrix, global spatial autocorrelation statistic and the spatial semi-variogram function [79,80].

4.3.2. Object-Oriented Method

The basic unit of object-oriented processing is homogeneous objects (image blocks, homogeneous areas or patches) with certain semantic information in images [81]. The method needs to segment the image to obtain the objects to further describe the spatial structure of the objects in the images [82]. The advantage of this method is that it has more abundant spatial relationships for the objects themselves and is convenient for extracting spatial features [83,84]. The deficiency of this method lies in its serious dependence on the quality of image segmentation. In fact, inaccurate image segmentation results in error accumulation when understanding spatial structure semantics [85].

4.3.3. Rule-Partition-Based Method

The rule-partition-based method is similar to grid division. Firstly, the image is divided into regular (generally square) image blocks. Then, each image block is used as the processing unit to describe the spatial structure features of each image block [86]. This method is especially suitable for the detection of the spatial structure semantics of complex objects such as residential areas and aircraft. It does not focus on the detailed structure of objects in the image block but only on the statistical properties of the overall structure [87,88]. The deficiency of this method lies in how to determine the suitable partition of image blocks, especially when it cannot locate the object boundary accurately [89].

4.3.4. Global Organization Method Based on Local Structure

In this method, firstly, the local structural features, such as feature points, feature lines and feature surfaces, are obtained. Then, according to the spatial structure of the objects, the global structure model of the objects is constructed by using certain organization rules and mathematical models [90]. The process is mainly based on the geometric structure of the object itself, spatial relationship information and prior knowledge of the object structure [91]. For instance, when extracting building targets in HSR remote sensing images, we can make full use of the feature that the building roof is a rectangular structure. Firstly, local structure features such as corners, lines and ridges are extracted. Then, the method of perceptual organization is used to organize it into a complete roof contour of the building [92,93]. This method accords with the cognition rule of people to things, but it has a higher request for the construction of mathematical models and the realization of calculation methods [94].

5. Semantic Understanding of Concept Layer

The concept layer belongs to high-level abstract semantics. The concept layer semantics of the geographic scene is the comprehensive judgment and representation of concepts such as function and pattern. The main application of the remote sensing image processing method is scene classification. In general, high-level semantic information can be acquired based on low-level information analysis, and low-level information can be transferred to higher-level by modeling. Through layer-by-layer refinement, the final representation of the concept layer semantics is closer to the abstract thinking of human beings, and then the geographic scene semantics in the HSR remote sensing images have more practical significance. Therefore, the concept layer semantics of geographic scenes are derived from the visual layer semantics or the object layer semantics.

5.1. Visual Features Based Method

The concept layer semantics of the geographic scene can be directly described by low-level visual feature attributes. The scene classification algorithm based on visual features extracts the low-level visual features (such as color, shape and texture), then describes the features and designs the classifier to infer the semantic information of the geographic scene. According to the different sources of low-level feature extraction, scene classification based on low-level features includes two categories: global-feature-based methods and local-feature-based methods. The extraction methods of global features and local features in visual features are detailed in Section 3 (Table 2). Common classifiers used for visual features include maximum likelihood [95], minimum distance [96] and K-means clustering [97].

Table 2. Comparison of main methods in visual features.

Name	Type	Output	Advantage	Disadvantage	Applicable
GIST	Global	Spectral information	Low computational complexity and easy to use	Poor performance in complex scenes with dense targets	Simple natural scenes
SIFT	Local	Neighborhood histogram	Suitable for translation, rotation, scale transformation	Poor performance in complex scenes with overall layout	Natural scenes
HOG		Vector	Representation of contours and edges	Poor performance in scenes with unstable shape structure	Scenes with global structural stability
CENTRIST		Census transformed value	Highlight local characteristics and reflect position information	Poor performance in complex and volatile scenes	Scenes with clear layout and sparse target distribution

A single low-level visual feature is not suitable for the complicated task of geographic scene classification, and more methods of multi-feature fusion are applied. Feature fusion combines color, texture and other features into high-dimensional feature descriptors, and then uses a neural network to achieve feature dimension reduction [98]. In addition, on the basis of local features, the image is divided into local blocks, and the low-level visual features of each block are taken to establish the multi-feature fusion descriptors [99,100]. Nevertheless, the method based on local or global visual features and their fusion of visual features is not effective. The core problem is that the concept layer semantics need to infer from the low-level features to obtain the high-level semantic representation, while the visual-features-based method just lacks this semantic representation.

5.2. Object Semantics Based Method

In order to fully describe the complex characteristics of the geographic scene, the extraction method of concept layer semantics based on object semantics is widely used in geographic scene understanding of HSR remote sensing images. By extracting the local features in the geographic scene, the local features are mapped to the visual dictionary or parameter space to obtain more distinguishable object layer features. Then, these features

are input into the classifier to obtain the comprehensive description features of the whole geographic scene.

5.2.1. Target-Recognition-Based Method

Geographic scenes involve the interaction of many objects in complex semantic patterns. According to the experience of human visual perception, images containing similar objects may represent the same geographic scene. When defining the category of the geographic scene, different objects have different importances in the scene. This prior knowledge provides ideas for the classification of geographic scenes.

The method based on object recognition will identify the semantics of each object in the geographic scene and train the classifier for concept semantic understanding based on the semantic information of each object. Typical approaches include Object Bank [101], Latent Pyramidal Regions [102], Bag of Parts [103] and Latent Semantic Analysis [104] (Table 3). These approaches assume that a scene consists of a series of targets, and that by identifying and recognizing those targets with significant discrimination, the category of the scene can be inferred from the semantics of those targets [105]. In these approaches, the problem of semantic understanding of the concept layer is first transformed into the problem of target recognition, and then the geographic scene is represented by image blocks containing multiple targets. However, the errors caused by target recognition will further result in "error propagation", which will affect the semantic understanding of the geographic scene.

Table 3. Comparison of main methods in target recognition.

Name	Advantage	Disadvantage	Applicable
Object Bank	Identifiable targets and natural scenes	High computational complexity and high feature dimension	Natural scenes with landmark targets
Latent Pyramidal Regions	Good performance for regions with specific structures	Focus on the shape structure of the scene, lack of deep semantic understanding	Scenes with complex background and crowded targets
Bag of Parts	Good performance for areas with boundaries or corners		
Latent Semantic Analysis	The synonym is characterized by dimensionality reduction, and the redundant data are used	Polysemous words have low discrimination and high computational complexity	Scenes with heterogeneous information and clear boundaries

5.2.2. Local Semantics Based Method

To avoid the process of object detection and recognition, the HSR remote sensing image can be divided according to rules and the local image descriptors of each sub-block can be extracted. The correspondence between local descriptors and local semantic concepts is established, and the scene classification is completed by using the probability distribution of local semantic concepts. There are two main algorithms based on local semantic concepts: the probabilistic topic-model-based method and the bag-of-visual-words-model-based method. Because the feature of spatial structure expresses the relationship between objects, it does not exist independently. Therefore, this feature is often used in conjunction with the bag of visual words model or the probabilistic topic model to enhance semantics.

(1) Bag of visual words model

The visual codebook is defined in advance, and the image content is described by the probability distribution of the appearance of the visual codewords. Then, the geographic scenes are classified according to the probability distribution. In the process of constructing the bag of visual words model, feature extraction, visual dictionary learning, feature mapping and whether to add spatial context information all have an impact on the classification results [106].

In the aspect of feature extraction, we consider the construction of multi-feature scenes in low-dimensional space under different perspectives and use feature complementarity to carry out feature fusion to solve the problem of dimension reduction from a multi-

perspective [107]. Existing visual dictionary learning includes generation (unsupervised) and discriminant (supervised) methods [108,109]. After visual dictionary learning, the image local descriptor is mapped to the visual dictionary. The modification of the mapping method can improve the representation of local semantics [110]. In addition, feature description can be incorporated into feature space partitioning to capture the high-order structure inherent in the scene [111]. This makes the local semantics have both local gradient information, local structure information [112] and global spatial information [113–115].

(2) Probabilistic topic model

The scene semantic content is first modeled by probability distribution based on the codewords. Then, the latent semantic topics in the images are learned by using the probability distribution model. In addition, the geographic scenes are classified according to the probability distribution of latent semantic topics.

Semantic topic modeling includes the generative probabilistic model and the discriminative probabilistic model. The generative probabilistic model of the geographic scene is constructed according to the joint probability distribution of the scene category in the feature space, using pLSA [116], LDA [117] and improved LDA (ts-LDA, css-LDA) to mine the latent semantic information of visual words [118]. Because various scenes contain different space-level structures, spatial information can undergo weighted fusion based on local semantic content. The discriminative probabilistic model is based on the conditional probability distribution of the category of the geographic scene in feature space, and its core task is to design kernel function. Wu demonstrates that a support vector machine based on a histogram intersection kernel (HIK) is more efficient than a radial basis function kernel for histogram-based data [119].

The generative probabilistic model and the discriminative probabilistic model have their respective advantages and complementary characteristics. The contradiction between computational complexity and model complexity is the biggest problem in the generative probabilistic model, but it is not a problem in the discriminative probabilistic model. The discriminative probabilistic model does not consider the connection between geographic scenes when modeling different categories of the scene, which belongs to independent modeling. In [120], these two probabilistic models are combined to complete the task of scene classification, and the classification effect is better than the single probabilistic model.

5.3. Feature-Learning-Based Method

Both the method based on visual features and object semantics rely mainly on artificial design in feature extraction, which is not only subjective, and it is not enough for more complex HSR remote sensing images. In recent years, feature learning, especially deep learning, has been introduced into the field of remote sensing for semantic understanding of geographic scenes due to its excellent performance in image classification [121]. The general flow of geographic scene classification based on feature learning is shown in Figure 5.

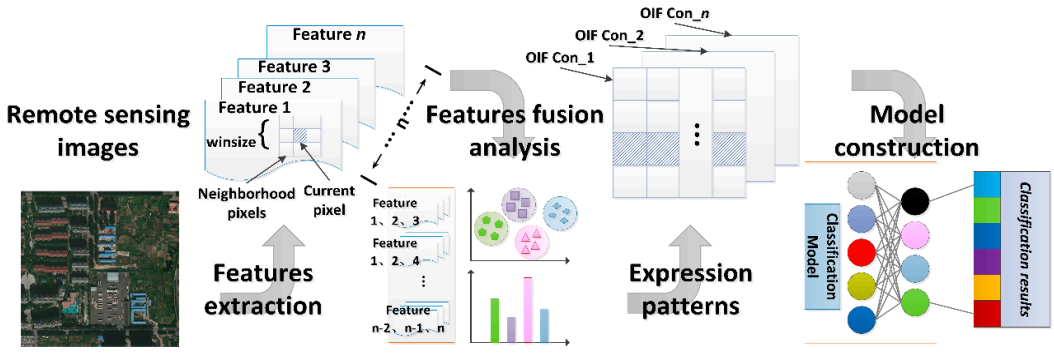

Figure 5. Geographic scene classification diagram.

Semantic understanding of the concept layer based on feature learning refers to the process of learning a potential scene classification feature through a series of mapping and transformation by using HSR remote sensing images as input of the model in machine learning tasks. Machine learning models can autonomously express and extract features from image data, abandoning the previous pattern of extracting features based on pre-designed rules [122,123]. Therefore, in the face of a complex surface environment, better classification results of geographic scenes can be obtained. At present, the commonly used machine learning models include sparse coding [124], neural network [125], support vector machine [126] and deep learning [127] (Figure 6). As a new intelligent method of pattern recognition in recent years, a deep learning network composed of multilaminate nonlinear mapping layers has become an especially important development direction in the field of remote sensing image processing [128]. Deep Learning is a deep structure neural network, which can extract the features of remote sensing images better than shallow structure models such as artificial neural networks and support vector machine models. Moreover, deep learning models can learn more abstract and distinguishable semantic features autonomously. The deep learning approach converts the semantic understanding of the concept layer into an end-to-end problem. On the one hand, the pre-trained deep learning network structure can be directly used to learn the global features in the visual layer of images to understand the semantics of the concept layer [129]. On the other hand, the deep learning network can also be used as a local feature extraction operator to jointly complete the semantic understanding of the concept layer with the help of feature code technology. Common deep learning models include convolutional neural networks (CNN) [130–132], deep belief network (DBN) [133], recurrent neural network (RNN) [134], automatic encoders [135], graph convolutional networks (GCN) [136], generative adversarial networks (GANs) [137,138] and so on. The deep learning method can be divided into three categories according to the supervision mode: (1) full supervision, (2) semi-supervised and (3) weak supervision.

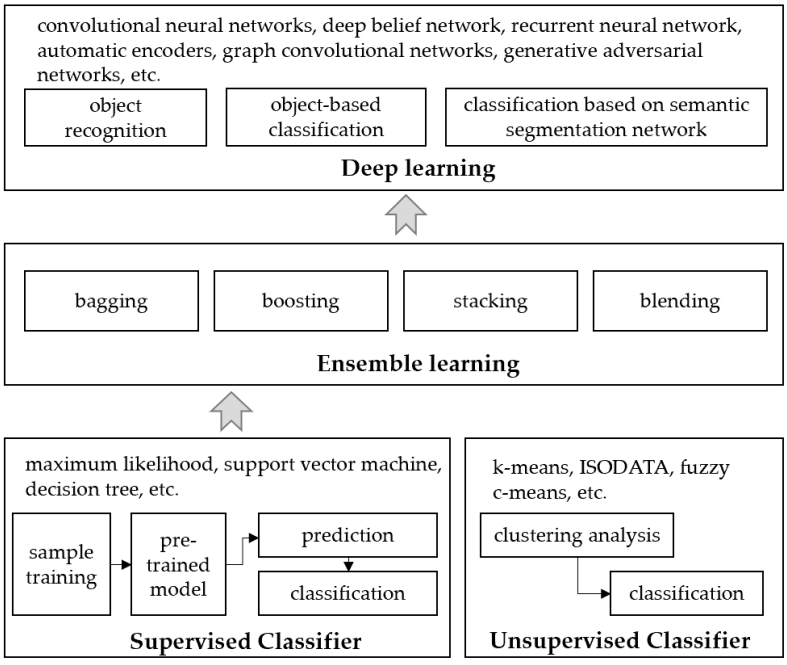

Figure 6. Development of feature learning methods.

5.3.1. The Method Based on Fully Supervised Deep Learning

Nowadays, most geographic scene classifications of HSR remote sensing images based on deep learning can be classified as full supervision. The integration of multiple learning models is one of the ways to improve the learning effect. Zhu et al. [139] proposed an adaptive deep sparse semantic modeling (ADSSM), which combines the topic model with

CNN and effectively integrates sparse topic features and deep features at the semantic level. Cheng et al. [140] proposed a new loss function to train fused deep neural networks by combining deep learning with metric learning. Zhang et al. [141] combined CNN and CapsNet for scene classification. This approach combines the advantages of both networks while leveraging the powerful feature extraction capabilities of CNN and the excellent feature fusion and classification capabilities of CapsNet. He et al. [142] proposed a new skip-connected covariance network (SCCov) for remote sensing image scene classification. Sumbel et al. [143] presented the BigEarthNet, which is a new large-scale, multi-label Sentinel-2 benchmark archive. The experimental results obtained in the framework of scene classification problems show that a shallow CNN architecture trained on the BigEarthNet provides much higher accuracy compared to a state-of-the-art CNN model pre-trained on the ImageNet. Thus, the BigEarthNet opens up promising directions to advance operational remote sensing applications and research in massive Sentinel-2 image archives.

5.3.2. The Method Based on Semi-Supervised Deep Learning

Semi-supervised learning can make use of a large number of unlabeled samples, reducing the need for labeled samples, which, to some extent, solves the problem of insufficient labeled samples in the field of deep learning [144]. Han et al. [145] proposed a generic framework based on semi-supervised deep features from the perspective of expanding the scale of labeled samples. In this framework, multiple support vector machine (SVM) models are applied to the label recognition of easily confused category samples, which improves the label precision and the number of labeled samples, thus improving the generalization ability and classification precision of the network.

It is also an effective semi-supervised deep learning to construct a feature extraction model based on unsupervised learning in the feature learning stage, then train the classifier with labeled samples. Soto et al. [146] used a combination of labeled and unlabeled samples to train generative adversarial networks (GAN) and then used the trained classifiers for scene classification. At this point, the classifier has a large number of unlabeled samples of information, which is helpful to improve the final classification effect. Fan et al. [147] used the representative salient regions extracted from the image as unlabeled samples to train the feature extractor. Then, the extractor is used to extract the features of the samples to be classified. Finally, SVM is used to classify the extracted features.

5.3.3. The Method Based on Weak Supervised Deep Learning

In HSR remote sensing image scene classification tasks, weak supervision usually uses labeled samples similar to target samples to train scene classification models. This method divides the dataset into the source domain and target domain. The former is different from the latter but similar. The latter can obtain labels through various transfer learning and further be used for training scene classification models. Othman et al. [148] took the features extracted from labeled images as the source domain, and the features extracted from unlabeled images as the target domain. Then, apply them to network training and optimize the specified loss function to classify labeled and unlabeled data. Gong et al. [149] further improved deep structural metric learning (DSML) by proposing Diversity-Promoting-DSML (D-DSML), which reduces the parameter redundancy produced by DSML and improves the feature representation ability.

Some existing deep learning classification tools include OverFeat [150], DeCAF [151], Caffe [152], AlexNet and so on (Figure 7). However, in these models, learning millions of network parameters also requires millions of training data as input. In order to reduce the over-fitting problem, a smaller network structure can be constructed. However, the generalization ability of the network model trained by this method is limited, such as gradient enhancement convolutional neural network [153] and multi-perspective convolutional neural network [154]. Therefore, the unsupervised feature-learning method directly uses the network model trained on the data set of images as the feature extractor to extract the deep features of the image directly, or after the feature transformation is input into the clas-

sifier for classification, higher classification results could be obtained [155]. Furthermore, the stacked covariance pooling method transforms the extracted multi-layer convolution layer features to obtain the global deep features of the image, which can effectively fuse the multi-layer deep features [156].

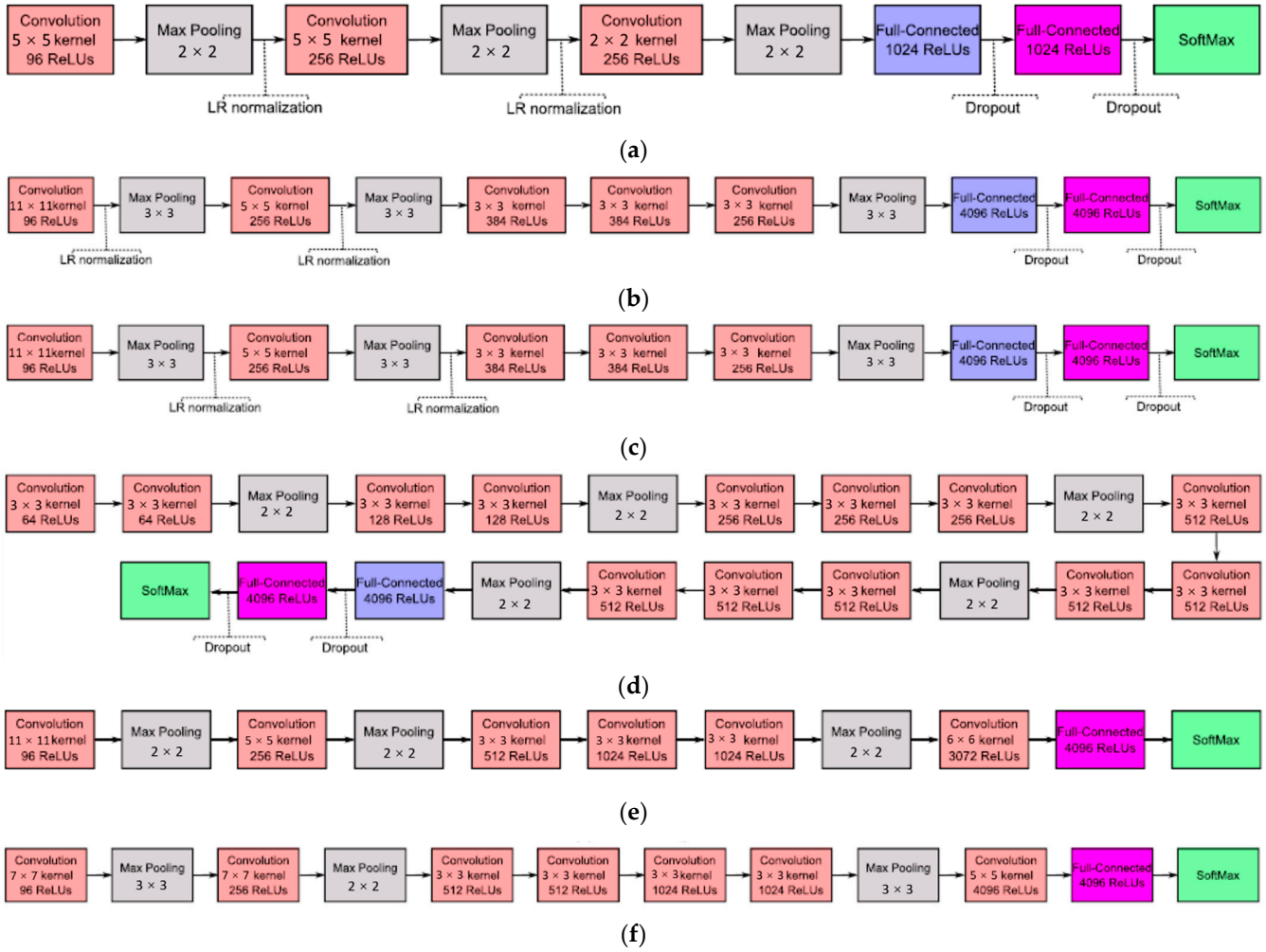

Figure 7. Architectures of different ConvNets evaluated in [157]. Purple boxes indicate the layers from where features were extracted in the case of using the ConvNets as feature extractors. (**a**) PatreoNet; (**b**) AlexNet; (**c**) CaffeNet; (**d**) VGG_{16}; (**e**) $OverFeat_S$; (**f**) $OverFeat_L$.

The method of deep feature fusion can further improve the accuracy of geographic scene classification. The most direct way is to cascade the features of the fully connected layers extracted from different network models [158]. In addition, discriminant correlation analysis (DCA) can be used to fuse the features of different fully connected layers [159]. Alternatively, the classification results of multiple models are fused based on Choquet fuzzy integral [160]. The UC-Merced dataset is one of the classic open resource sets for classification tasks of geographic scenes. Using the UC-Merced dataset as experimental data, the performance of existing geographic scene classification models is summarized (Table 4).

Table 4. Comparison of main methods in geographic scene classification.

Method		Accuracy (%)	Other Indicators
Visual features-based method	Gabor texture [161]	76.91	-
	Color-HLS [161]	81.19	-
	NN-STSIM [162]	86	-
	Quaternion orientation difference [163]	85.48 ± 1.02	-
	MS-CLBP [164]	90.6 ± 1.4	-
Object-semantics-based method	BoVW [161]	76.81	-
	BoVW + SCK [161]	77.71	-
	SPM [161]	75.29	-
	SPCK ++ [165]	77.38	-
	HMFF [166]	92.38 ± 0.62	-
	CCM-BoVW [167]	86.64 ± 0.81	-
	Wavelet BoVW [168]	87.38 ± 1.27	-
	UFL [169]	81.67 ± 1.23	-
	COPD [170]	91.33 ± 1.11	-
	FV [171]	93.8	-
	VLAT [171]	94.3	-
	SG-UFL [172]	82.72 ± 1.18	-
	PSR [173]	89.1	-
	UFL-SC [174]	90.26 ± 1.51	-
	SAL-PLSA [175]	87.62	-
	SAL-LDA [175]	88.33	-
Feature-learning-based method	CaffeNet finetune [176]	95.48	-
	GoogleNet finetune [176]	97.1	-
	Multiview DL [177]	93.48 ± 0.82	84.35 (Sensitivity), 91.72 (Specificity)
	GBRCN [178]	94.53	-
	ADPM [179]	94.86	-
	HCSAE [180]	97.14 ± 1.19	-
	MARTA GANs [181]	94.86 ± 0.80	-
	Fusion by addition [182]	97.42 ± 1.79	-
	salM^{3}LBP-CLM [183]	95.75 ± 0.80	-
	TEX-Nets [184]	97.72	-
	CCP-Net [185]	97.52 ± 0.97	-
	CNN (LOFs+GCFs) [186]	99.00 ± 0.35	-
	ARCNet-VGG16 [187]	99.12 ± 0.40	-
	D-CNN with VGG16 [188]	98.93 ± 0.10	-
	SAL-TS-Net [189]	98.90 ± 0.95	-
	Two-stream deep fusion [190]	98.02 ± 1.03	-
	PMS [191]	98.81	8.32×10^6 (Number of neurons)
	SSF-AlexNet [192]	92.43 ± 0.46	-
	VGG16+MSCP+MRA [193]	98.40 ± 0.34	-
	MCNN [194]	96.66 ± 0.90	-
	Bidirectional adaptive feature fusion [195]	95.48	-

Although great progress has been made in geographic scene classification using deep learning algorithms, compared with the shallow algorithm, the classification effect has been improved obviously. However, the application of deep learning still faces many problems, such as the following:

(1) In terms of training data, the success of a deep neural network is that it can fit large-scale samples without sacrificing generalization ability. In the field of geographic

scene understanding, it is difficult to construct a large-scale, high-quality and complete HSR remote sensing image dataset for training. Firstly, from the perspective of time, a training sample can only represent the sampling of a time section. However, the interpretations of objects are dynamic in different periods. This time heterogeneity puts forward higher requirements for the quality, scale and completeness of sample annotation [196]. Secondly, from the perspective of space, due to the differences in climate and light conditions, the distribution of ground objects in different geographic scenes has natural heterogeneity [197]. This spatial heterogeneity leads to the imbalance of sample categories in the supervised learning process, whether within the training set or between the training set and the test set, which leads to "over-fitting" or "under-fitting" problems.

(2) In terms of learning mechanisms, supervised learning mainly relies on semantic support provided by manual annotation as the only learning signal for model training. If human labeling is regarded as prior knowledge, the machine has been limited in knowledge in the process of labeling [198]. However, for the huge amount of image data, the intrinsic information should be much more abundant than the semantic information provided by sparse labels. Therefore, over-reliance on a manual annotation will cause the risk of "inductive bias" in the trained model. Moreover, the computational cost is high, especially for small samples. Most of the deep learning models are trained on the established network structure and are then fine-tuned to obtain better network parameters. This training pattern is not suitable for ever-expanding datasets.

Although deep learning has a strong learning ability, compared with real artificial intelligence, it still lacks the ability of abstract knowledge representation, reasoning causality and logical relationship. Therefore, there is a long distance to understand the geographic scene automatically through feature learning.

6. Open Problems and Challenges

(1) Integrated system engineering for geographic scene understanding

The research of HSR remote sensing images is often only used the visual information and a little semantic information, such as target detection, image classification, image segmentation, scene classification and so on. These researchers can often only detect a certain target contained in the image, or obtain the category labels of each pixel or the whole image, but they do not make full use of the features of the image. Thus, it is difficult to mine the attributes, characteristics and relationships among the objects in the image in detail. In this way, images are not fully understood at the semantic level and HSR remote sensing data are not fully utilized. The organic integration of single subtask or feature information of HSR remote sensing image processing can enhance the performance of understanding, and it is more suitable for people's understanding mode of the geographic scene. The multi-class feature information and multi-subtasks are not completely independent, and the mutual influence and restriction factors should be considered comprehensively (Figure 8). Therefore, the construction of system engineering for geographic scene understanding can follow the following Formula (1):

$$Y = \{(feature_1 \oplus feature_2 \oplus \ldots \oplus feature_n) \otimes (task_1 \oplus task_2 \oplus \ldots \oplus task_n)\} \quad (1)$$

In the formula, Y is the system engineering for geographic scene understanding, $\otimes$ and $\oplus$ represent the different combinations of features and tasks.

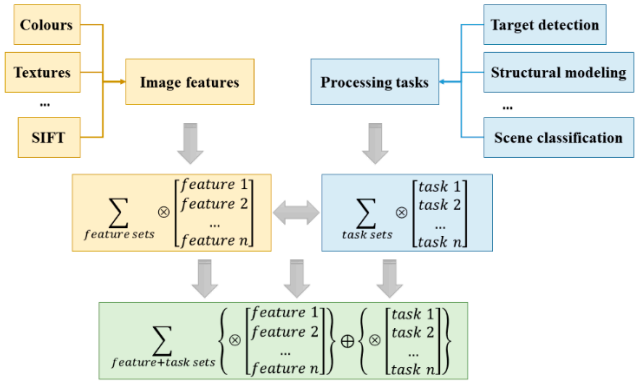

Figure 8. System engineering framework for geographic scene understanding based on HSR remote sensing image.

(2) Comprehensive semantic representation of the geographic scene in HSR remote sensing image

The purpose of geographic scene understanding based on HSR remote sensing image is to semantically explain the content at all layers in the geographic scene. It is necessary to construct a comprehensive semantic representation model of geographic scenes, and to standardize and integrate the various layers and types of semantic information obtained from the understanding of geographic scenes. The semantic parsing tree of a geographic scene can be constructed, and the tree structure of and/or a graph can be used to represent the semantic content of understanding. The semantic parsing tree of the geographic scene follows a unified semantic specification, which is generally divided into four levels "scene-object-part-pixel" (Figure 9). The "And" node represents the decomposition, such as "scene→object", "object→part" and so on, which is followed the syntactic rules of "A→BCD". Any geographic scene semantics can be represented by this parsing tree structure, and the semantic hierarchy of geographic scene is clearly divided, which has both semantic attributes and semantic relations between different levels.

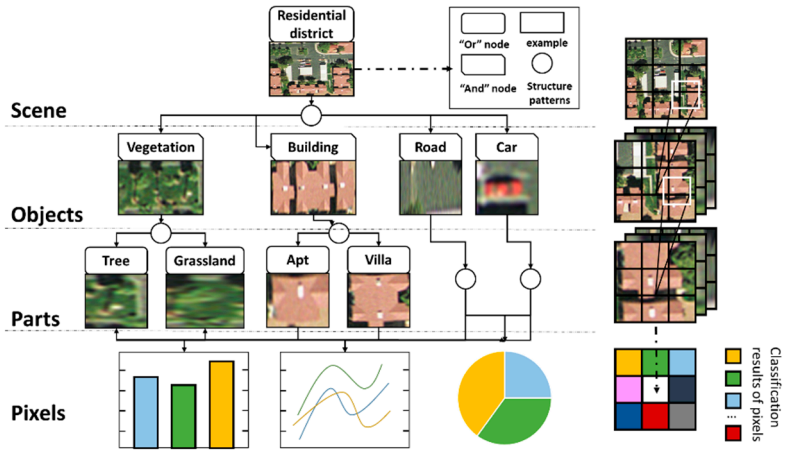

Figure 9. Hierarchical representation structure of the semantic of geographic scene.

(3) Adaptability for large-scale complex geographic scenes

With the massive growth of image data and the continuous subdivision of scene categories, the problem of geographic scene understanding is faced with unprecedented challenges both in image quantity and scene category. The understanding of real geographic scenes requires higher complexity and depth than scene classification, and the solution to this problem will have a profound impact on artificial intelligence technology. For the current semantic understanding approaches of geographic scenes, there are still the

following problems to be solved: (1) The semantic extraction ability of the visual layer is insufficient. Even considering multiple visual features, most of them are the simple superposition of different features. (2) The semantic modeling of object layer is redundant and lacks homogeneity in the description, which makes it difficult to take into account the computational efficiency and effect. (3) The semantic understanding of concept layer ignores spatial location information and globality. Whether it is visual-feature-based or object-sematic-based method, it is difficult to obtain an accurate global description of spatial location relationships and geographic scene characteristics at the same time. Thus, it is not conducive to an accurate understanding of the geographic scene.

(4) Fusion application of abundant multi-source data

In the big data era, the accessibility of various types of data has been broken through, creating conditions for the geographic scene understanding. On the one hand, the promotion of big data technology in the field of remote sensing has promoted the arrival of the era of remote sensing big data. Multi-source remote sensing data collaboration can integrate the advantages of various remote sensing observation methods and make up for the insufficient single sensor, which is one of the important research directions for the breakthrough of remote sensing image processing. On the other hand, the remote sensing image data record the natural environment of the surface, but the perception of changes in the social environment is scarce. The fusion of multi-source data is not only limited to HSR remote sensing data itself, but also needs to combine different types of data resources to make up for the deficiency of remote sensing monitoring. Social media data represented by Twitter, Facebook, Sina Weibo and Internet maps represented by GeoNames, GNIS, OpenStreetMap, etc., have become important sources of data for depicting the scenes of humanities and society. The fusion of HSR remote sensing images and multi-source data provides a new idea and method for geographic scene understanding.

7. Conclusions

Geographic scene understanding is one of the core tasks for middle and high-level cognition in remote sensing image processing tasks. Its complexity and comprehensiveness make it difficult to accurately understand the semantic information of geographic scenes. Based on the analysis of the basic concepts and core connotations of geographic scene understanding, this paper reviews the research status of geographic scene understanding from the tasks of different semantic layers in HSR remote sensing images. Geographic scene understanding decomposes the information of HSR remote sensing images into three semantic layers: based on the visual features of remote sensing images, the local objects, spatial structure and scene functions of the geographic scene are analyzed in a consistent cognitive system. This not only conforms to the logic and order of human cognition but also has significant interpretability of various semantic information. In terms of target detection, efficient and accurate feature representation and fusion of appropriate attention mechanisms are the core of extracting object category semantics. In terms of spatial structure description, pixel neighborhood window, object-oriented, rule partition and local structure are the main methods for extracting spatial structure semantics. In terms of scene classification, according to the semantic abstraction degree of extracted features, it mainly includes the visual feature classification method, object semantic classification method and feature learning classification method.

In future research, it is necessary to deeply study the intrinsic objective laws of various objects, textures, spaces and other information in geographic scene understanding, in order to reveal the relationship and influence mechanism between various features of images and different subtasks of image processing. The system engineering of geographic scene understanding is constructed from the global perspective, and the deep mechanism of human understanding of the geographic scene is explored. This is not only conducive to improving the adaptability of large-scale complex geographic scenes, but it also provides a universal cognitive structure for other HSR remote sensing images processing tasks such as image analysis and landscape investigation.

Author Contributions: Conceptualization, P.Y. and Y.H.; methodology, P.Y. and G.L.; validation, G.L.; formal analysis, P.Y.; investigation, P.Y. and G.L.; writing—original draft preparation, P.Y.; writing—review and editing, P.Y. and Y.H.; visualization, G.L.; supervision, G.L.; project administration, P.Y.; funding acquisition, P.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Open Foundation of Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), the Ministry of Education (grant nos. 2021VGE01, and 2022VGE01), the Humanities and Social Sciences Foundation of Yangzhou University (grant no. xjj2021-08), the Open Foundation of Research Institute of Central Jiangsu Development, Yangzhou University (grant no. szfz202114) and the Open Foundation of Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province (grant no. SHEL221 002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Xueying Zhang and Chunju Zhang for their critical reviews and constructive comments.

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

References

1. Li, D.; Zhang, L.; Xia, G. Automatic Analysis and Mining of Remote Sensing Big Data. *Acta Geod. Cartogr. Sin.* **2014**, *43*, 1211–1216.
2. Dumitru, C.O.; Cui, S.; Schwarz, G.; Datcu, M. Information content of very-high-resolution sar images: Semantics, geospatial context, and ontologies. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *8*, 1635–1650. [CrossRef]
3. Zhang, Y.; Zheng, X.; Liu, G.; Sun, X.; Wang, H.; Fu, K. Semi-Supervised Manifold Learning Based Multigraph Fusion for High-Resolution Remote Sensing Image Classification. *IEEE Geosci. Remote Sens.* **2014**, *11*, 464–468. [CrossRef]
4. Liu, Q.; Hang, R.; Song, H.; Li, Z. Learning multiscale deep features for high-resolution satellite image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 117–126. [CrossRef]
5. Gong, Z.; Zhong, P.; Yu, Y.; Hu, W. Diversified deep structural metric learning for land use classification in remote sensing images. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017.
6. Zhu, Q.; Zhong, Y.; Zhang, L. Scene classfication based on the semantic-feature fusion fully sparse topic model for high spatial resolution remote sensing imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *XLI-B7*, 451–457. [CrossRef]
7. Biederman, I. Human image understanding: Recent research and theory. *Comput. Vis. Graph. Image Process.* **1985**, *31*, 400–401. [CrossRef]
8. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.; Emery, W.J. Active learning methods for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218–2232. [CrossRef]
9. Eakins, J. Automatic image content retrieval—Are we getting anywhere? In Proceedings of the Third International Conference on Electronic Library and Visual Information Research (ELVIRA3), Milton Keynes, UK, 10–12 May 1996.
10. Lv, G.; Chen, M.; Yuan, L.; Zhou, L.; Wen, Y.; Wu, M.; Hu, B.; Yu, Z.; Yue, S.; Sheng, Y. Geographic scenario: A possible foundation for further development of virtual geographic environments. *Int. J. Digit. Earth* **2018**, *11*, 356–368.
11. Zhong, Y.; Fei, F.; Zhang, L. Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *J. Appl. Remote Sens.* **2016**, *10*, 025006. [CrossRef]
12. Lin, B.; Liu, Q.; Li, C.; Ye, Z.; Hui, M.; Jia, X. Using Bag of Visual Words and Spatial Pyramid Matching for Object Classification Along with Applications for RIS. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14.
13. Vyas, K.; Vora, Y.; Vastani, R. Bag-of-visual-words and spatial extensions for land-use classification. *Procedia Comput. Sci.* **2016**, *89*, 457–464. [CrossRef]
14. Kasper, A.; Jäkel, R.; Dillmann, R. Using spatial relations of objects in real world scenes for scene structuring and scene understanding. In Proceedings of the 2011 15th International Conference on Advanced Robotics (ICAR), Tallinn, Estonia, 20–23 June 2011.
15. 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]
16. Gu, Y.; Wang, Y.; Li, Y. A Survey on Deep Learning-Driven Remote Sensing Image Scene Understanding: Scene Classification, Scene Retrieval and Scene-Guided Object Detection. *Appl. Sci.* **2019**, *9*, 2110. [CrossRef]
17. Zhong, Y.; Wu, S.; Zhao, B. Scene Semantic Understanding Based on the Spatial Context Relations of Multiple Objects. *Remote Sens.* **2017**, *9*, 1030. [CrossRef]
18. Qin, K.; Chen, Y.; Gan, S.; Feng, X.; Ren, W. Review on methods of spatial structural feature modeling of high resolution remote sensing images. *J. Image Graph.* **2013**, *18*, 1055–1064.

19. Hu, J. Multi-Level Feature Representation for Scene Classification with High Spatial Resolution Remote Sensing Images. Ph.D. Thesis, Wuhan University, Wuhan, China, 2019.
20. Long, Y.; Xia, G.; Li, S.; Yang, W.; Yang, M.; Zhu, X.; Zhang, L.; Li, D. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID. *IEEE J-STARS.* **2021**, *14*, 4205–4230. [CrossRef]
21. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
22. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
23. Shahriari, M.; Bergevin, R. Land-use scene classification: A comparative study on bag of visual word framework. *Multimed. Tools Appl.* **2017**, *76*, 23059–23075. [CrossRef]
24. Xia, G.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; Maitre, H. Structural high-resolution satellite image indexing. In Proceedings of the ISPRS TC VII Symposium—100 Years ISPRS, Vienna, Austria, 5–7 July 2010; pp. 298–303.
25. Nilakshi, D.; Bhogeswar, B. A novel mutual information-based feature selection approach forefficient transfer learning in aerial scene classification. *Int. J. Remote sens.* **2021**, 2321–2325. [CrossRef]
26. Zhao, L.; Ping, T.; Huo, L. Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *J. Appl. Remote Sens.* **2016**, *10*, 035004. [CrossRef]
27. Zhao, B.; Zhong, Y.; Xia, G.; Zhang, L. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2108–2123. [CrossRef]
28. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
29. 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.* **2017**, *145*, 197–209. [CrossRef]
30. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark data Set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
31. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [CrossRef]
32. Swain, M.J.; Ballard, D.H. Color indexing. *Int. J. Comput Vis.* **1991**, *7*, 11–32. [CrossRef]
33. Forssén, P.E. Maximally Stable Colour Regions for Recognition and Matching. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition., Minneapolis, MN, USA, 17–22 June 2007.
34. Sande, K.; Gevers, T.; Snoek, C. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1582–1596. [CrossRef]
35. Tao, D.; Jin, L.; Zhao, Y.; Li, X. Rank Preserving Sparse Learning for Kinect Based Scene Classification. *IEEE Trans. Cybern.* **2013**, *43*, 1406–1417. [CrossRef]
36. Banerji, S.; Sinha, A.; Liu, C. New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. *Neurocomputing* **2013**, *117*, 173–185. [CrossRef]
37. Iqbal, N.; Mumtaz, R.; Shafi, U.; Zaidi, S.M.H. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms. *PeerJ Comput. Sci.* **2021**, *7*, e536. [CrossRef]
38. Lv, H.; Liu, Y.; Xue, X.; Ma, T. Methods and Experiments of Background Subtraction and Grayscale Stretch for Remote Sensing Images. *Chin. J. Liq. Cryst. Disp.* **2012**, *27*, 235.
39. Li, Y.; Zhang, J.; Zhou, Y.; Niu, J.; Wang, L.; Meng, N.; Zheng, J. ISAR Imaging of Nonuniformly Rotating Targets with Low SNR Based on Coherently Integrated Nonuniform Trilinear Autocorrelation Function. *IEEE Geosci. Remote Sens. Lett.* **2020**, *99*, 1074–1078. [CrossRef]
40. Ru, C.; Li, Z.; Tang, R. A Hyperspectral Imaging Approach for Classifying Geographical Origins of Rhizoma Atractylodis Macrocephalae Using the Fusion of Spectrum-Image in VNIR and SWIR Ranges (VNIR-SWIR-FuSI). *Sensors* **2019**, *19*, 2045. [CrossRef] [PubMed]
41. Huang, L.; Chen, C.; Li, W.; Du, Q. Remote Sensing Image Scene Classification Using Multi-Scale Completed Local Binary Patterns and Fisher Vectors. *Remote Sens.* **2016**, *8*, 483. [CrossRef]
42. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2011**, *42*, 145–175. [CrossRef]
43. Lowe, D. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
44. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded Up Robust Features. In *Computer Vision—ECCV 2006; Lecture Notes in Computer Science; Proceedings of the 9th European Conference on Computer Vision (ECCV 2006), Graz, Austria, 7–13 May 2006*; Leonardis, A., Bischof, H., Pinz, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3951. [CrossRef]
45. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
46. Wu, J.; Rehg, J. CENTRIST: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1489–1501.
47. Zou, C.; Lei, Z.; Lv, S. Remote Sensing Image Dam Detection Based on Dual Threshold Network. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, 23 August 2020.

48. Horhan, M.; Eidenberger, H. An Efficient DCT template-based Object Detection Method using Phase Correlation. In Proceedings of the 2016 50th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 6–9 November 2017.
49. Wu, Z.; Wan, Q.; Liang, J.; Zhou, Z. Line Detection in Remote Sensing Images Using Hough Transform Based on Granular Computing. *Geomat. Inf. Sci. Wuhan Univ.* **2007**, *32*, 860–863.
50. Zhang, L.; Zhang, L.; Tao, D.; Xin, H.; Bo, D. Hyperspectral Remote Sensing Image Subpixel Target Detection Based on Supervised Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4955–4965. [CrossRef]
51. Hermosilla, T.; Ruiz, L.A.; Recio, J.A.; Estornell, J. Evaluation of Automatic Building Detection Approaches Combining High Resolution Images and LiDAR Data. *Remote Sens.* **2011**, *3*, 1188–1210. [CrossRef]
52. Li, X.; Xu, H.; An, S. Monitoring and assessment of intensive utilization of port area based on high spatial resolution remote sensing image with case study of five typical ports in the Bohai Sea. *J. Appl. Oceanogr.* **2019**, *38*, 126–134.
53. Ai, S.; Yan, J.; Li, D.; Xu, J.; Shen, J. An Algorithm for Detecting the Airport Runway in Remote Sensing Image. *Electron. Opt. Control* **2017**, *24*, 43–46.
54. Li, X.; Zhang, Z.; Lv, S.; Pan, M.; Yu, H. Road Extraction from High Spatial Resolution Remote Sensing Image Based on Multi-Task Key Point Constraints. *IEEE Access* **2021**, *9*, 95896–95910. [CrossRef]
55. Wei, S.; Chen, H.; Zhu, X.; Zhang, H. Ship Detection in Remote Sensing Image based on Faster R-CNN with Dilated Convolution. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020.
56. Wang, X.; Luo, G.; Ke, Q.; Chen, A.; Tian, L. A Fast Target Locating Method for Remote Sensing Images Based on Line Features. *Int. J. Signal Process. Image Process. Pattern Recogn.* **2017**, *10*, 61–72. [CrossRef]
57. Zhang, Q.; Lin, Q.; Ming, G.; Li, J. Remote Sensing Image Analysis on Circulation Induced by the Breakwaters in the Huanghua Port. In Proceedings of the International Conference on Estuaries and Coasts, Hangzhou, China, 9–11 November 2003.
58. Song, J.; Hu, W. Experimental Results of Maritime Target Detection Based on SVM Classifier. In Proceedings of the 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China, 12–15 September 2020.
59. Huang, S.; Huang, W.; Zhang, T. A New SAR Image Segmentation Algorithm for the Detection of Target and Shadow Regions. *Sci. Rep.* **2016**, *6*, 38596. [CrossRef] [PubMed]
60. Chaudhuri, D.; Agrawal, A. Split-and-merge Procedure for Image Segmentation using Bimodality Detection Approach. *Def. Sci. J.* **2010**, *60*, 290–301. [CrossRef]
61. Sun, Y.J.; Lei, W.H.; Ren, X.D. Remote sensing image ship target detection method based on visual attention model. In *Proceedings of the Lidar Imaging Detection and Target Recognition 2017*; Lv, D., Lv, Y., Bao, W., Eds.; SPIE-Int. Soc. Optical Engineering: Bellingham, WA, USA, 2017; Volume 10605.
62. Wu, J.; Rehg, J.M. Beyond the Euclidean distance: Creating effective visual codebooks using the Histogram Intersection Kernel. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009.
63. Perronnin, F. Universal and Adapted Vocabularies for Generic Visual Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1243–1256. [CrossRef]
64. Su, Y.; Allan, M.; Jurie, F. Improving Image Classification Using Semantic Attributes. *Int. J. Comput. Vis.* **2012**, *100*, 59–77. [CrossRef]
65. Yang, J.; Kai, Y.; Gong, Y.; Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009.
66. Lee, H.; Battle, A.; Raina, R.; Ng, A.Y. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006*; MIT Press: Cambridge, MA, USA, 2006.
67. Yu, K.; Zhang, T. Improved Local Coordinate Coding using Local Tangents. In Proceedings of the International Conference on International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010.
68. Wang, J.; Yang, J.; Kai, Y.; Lv, F.; Huang, T.; Gong, Y. Locality-constrained Linear Coding for image classification. In Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010.
69. Pham, T.T.; Maisonnasse, L.; Mulhem, P.; Gaussier, E. Visual Language Model for Scene Recognition. In Proceedings of the Singaporean-French Ipal Symposium 2009, Singapore, 18–20 February 2009.
70. Hofmann, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach. Learn.* **2001**, *42*, 177–196. [CrossRef]
71. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
72. Wu, L.; Hu, Y.; Li, M.; Yu, N.; Hua, X. Scale-Invariant Visual Language Modeling for Object Categorization. *IEEE Trans. Multimed.* **2009**, *11*, 286–294. [CrossRef]
73. Jing, H.; Wei, H. Latent Dirichlet Allocation Based Image Retrieval. In *Information Retrieval*; Wen, J., Nie, J., Ruan, T., Liu, Y., Qian, T., Eds.; CCIR 2017, Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10390. [CrossRef]
74. Kato, H.; Harada, T. Visual Language Modeling on CNN Image Representations. *arXiv* **2015**, arXiv:1511.02872.
75. Zhao, H.; Wang, Q.; Wang, Q.; Wu, W.; Yuan, N. SAR image despeckling based on adaptive neighborhood window and rotationally invariant block matching. In Proceedings of the 2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Guilin, China, 5–8 August 2014.
76. Aytekin, Ö.; Koc, M.; Ulusoy, İ. Local Primitive Pattern for the Classification of SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2431–2441. [CrossRef]

77. Hudak, A.T.; Strand, E.K.; Vierling, L.A.; Byrne, J.C.; Eitel, J.; Martinuzzi, S.; Falkowski, M. Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. *Remote Sens. Environ.* **2012**, *123*, 25–40. [CrossRef]
78. Li, S.; Zhang, B.; Li, A.; Jia, X.; Gao, L.; Peng, M. Hyperspectral Imagery Clustering with Neighborhood Constraints. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 588–592. [CrossRef]
79. Rahman, M.H.; Islam, H.; Neema, N. Compactness of Neighborhood Spatial Structure: A Case Study of Selected Neighborhoods of DNCC and DSCC Area. In Proceedings of the International Conference on Sustainability in Natural and Built Environment (iCSNBE 2019), Dhaka, Bangladesh, 19–22 January 2019.
80. Guan, X.; Huang, C.; Yang, J.; Li, A. Remote Sensing Image Classification with a Graph-Based Pre-Trained Neighborhood Spatial Relationship. *Sensors* **2021**, *21*, 5602. [CrossRef]
81. Sha, Z.; Bian, F. Object-Oriented Spatial Knowledge Representation and Its Application. *J. Remote Sens.* **2004**, *19*, 165–171.
82. Wei, C.; Zheng, Z.; Zhou, Q.; Huang, J.; Yuan, Y. Application of a parallel spectral-spatial convolution neural network in object-oriented remote sensing land use classification. *Remote Sens. Lett.* **2018**, *9*, 334–342.
83. Wang, Y.; Bao, W.; Yang, C.; Zhang, Y. A study on the automatic classification method on the basis of high resolution remote sensing image. In Proceedings of the 6th International Digital Earth Conference, Beijing, China, 9–12 September 2009.
84. Liu, X. Object Oriented Information Classification of Remote Sensing Image Based on Segmentation and Merging. *Appl. Mech. Mater.* **2014**, *568–570*, 734–739. [CrossRef]
85. Tan, Y.; Huai, J.; Tang, Z. An Object-Oriented Remote Sensing Image Segmentation Approach Based on Edge Detection. *Spectrosc. Spect. Anal.* **2010**, *30*, 1624–1627.
86. Tong, X.; Jin, B.; Ying, W. A new effective Hexagonal Discrete Global Grid System: Hexagonal quad balanced structure. In Proceedings of the 8th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010.
87. Khromyk, V.; Khromykh, O. Analysis of Spatial Structure and Dynamics of Tom Valley Landscapes based on GIS, Digital Elevation Model and Remote Sensing. *Procedia Soc. Behav. Sci.* **2014**, *120*, 811–815. [CrossRef]
88. Ding, Y.; Pan, S.; Chong, Y. Robust Spatial–Spectral Block-Diagonal Structure Representation with Fuzzy Class Probability for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1747–1762. [CrossRef]
89. Gao, Y.; Zhang, Y.; Alsulaiman, H. Spatial structure system of land use along urban rail transit based on GIS spatial clustering. *Eur. J. Remote Sens.* **2021**, *54*, 438–445. [CrossRef]
90. Wurm, M.; Taubenbck, H.; Dech, S. Quantification of urban structure on building block level utilizing multisensoral remote sensing data. In Proceedings of the Earth Resources and Environmental Remote Sensing/GIS Applications 2010, Toulouse, France, 25 October 2010.
91. Chen, J.; Chen, S.; Chen, X.; Yang, Y.; Xing, L.; Fan, X.; Rao, Y. LSV-ANet: Deep Learning on Local Structure Visualization for Feature Matching. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
92. Du, Z.; Li, X.; Lu, X. Local structure learning in high resolution remote sensing image retrieval. *Neurocomputing* **2016**, *207*, 813–822. [CrossRef]
93. Lei, S.; Shi, Z.; Zou, Z. Super-Resolution for Remote Sensing Images via Local-Global Combined Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1243–1247. [CrossRef]
94. Chen, J.; Fan, X.; Chen, S.; Yang, Y.; Bai, H. Robust Feature Matching via Hierarchical Local Structure Visualization. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
95. Bruzzone, L.; Prieto, D.F. Unsupervised Retraining of a Maximum Likelihood Classifier for the Analysis of Multitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 456–460. [CrossRef]
96. Zeh, A.; Bezzateev, S. A New Bound on the Minimum Distance of Cyclic Codes Using Small-Minimum-Distance Cyclic Codes. *Design. Code. Cryptogr.* **2014**, *71*, 229–246. [CrossRef]
97. Yuan, Y.; Meng, Q. Polyp classification based on Bag of Features and saliency in wireless capsule endoscopy. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014.
98. Guo, Y.; Ji, J.; Shi, D.; Ye, Q.; Xie, H. Multi-view feature learning for VHR remote sensing image classification. *Multimed. Tools Appl.* **2021**, *80*, 23009–23021. [CrossRef]
99. Hu, J.; Li, M.; Xia, G.; Zhang, L. Mining the spatial distribution of visual words for scene classification. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016.
100. 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]
101. Li, L.; Su, H.; Xing, E.; Li, F. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In Proceedings of the 23rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 6–9 December 2010.
102. Sadeghi, F.; Tappen, M.F. Latent Pyramidal Regions for Recognizing Scenes. In *Computer Vision—ECCV 2012, Proceedings of the 12th European Conference on Computer Vision (ECCV 2012), Florence, Italy, 7–13 October 2012*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012.
103. Juneja, M.; Vedaldi, A.; Jawahar, C.; Zisserman, A. Blocks that shout: Distinctive parts for scene classification. In Proceedings of the Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
104. Wang, J.; Sun, X.; Nahavandi, S.; Kouzani, A.; Wu, Y.; She, M. Multichannel biomedical time series clustering via hierarchical probabilistic latent semantic analysis. *Comput. Meth. Prog. Biol.* **2014**, *117*, 238–246. [CrossRef] [PubMed]

105. Gong, C.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote Sensing Image Scene Classification Using Bag of Convolutional Features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1735–1739.
106. Hu, J.; Xia, G.-S.; Hu, F.; Zhang, L. A Comparative Study of Sampling Analysis in the Scene Classification of Optical High-Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14988–15013. [CrossRef]
107. Yu, J.; Tao, D.; Rui, Y.; Cheng, J. Pairwise constraints based multiview features fusion for scene classification. *Pattern Recogn.* **2013**, *46*, 483–496. [CrossRef]
108. Wang, X.; Wang, B.; Bai, X.; Liu, W.; Tu, Z. Max-margin multiple-instance dictionary learning. In Proceedings of the 30th International Conference on International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
109. Shen, L.; Wang, S.; Sun, G.; Jiang, S.; Huang, Q. Multi-level discriminative dictionary learning towards hierarchical visual categorization. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
110. Oliveira, G.; Nascimento, E.; Vieira, A.; Campos, M. Sparse spatial coding: A novel approach for efficient and accurate object recognition. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012.
111. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.; Zhang, L. Bag-of-Visual-Words Scene Classifier with Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *13*, 747–751. [CrossRef]
112. Hu, F.; Xia, G.-S.; Hu, J.; Zhong, Y.; Xu, K. Fast Binary Coding for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2016**, *8*, 555. [CrossRef]
113. Kwitt, R.; Vasconcelos, N.; Rasiwasia, N. Scene recognition on the semantic manifold. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012.
114. Wang, Z.; Feng, J.; Yan, S.; Xi, H. Linear distance coding for image classification. *IEEE Trans. Image Process.* **2013**, *22*, 537–548. [CrossRef]
115. Xie, L.; Wang, J.; Guo, B.; Zhang, B.; Tian, Q. Orientational pyramid matching for recognizing indoor scenes. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
116. Singh, A.; Parmanand; Saurabh. Survey on pLSA based scene classification techniques. In Proceedings of the 2014 5th International Conference—Confluence the Next Generation Information Technology Summit (Confluence), Noida, India, 25–26 September 2014.
117. Veeranjaneyulu, N.; Raghunath, A.; Devi, B.J.; Mandhala, V.N. Scene classification using support vector machines with LDA. *J. Theor. Appl. Inf. Technol.* **2014**, *63*, 741–747.
118. Bosch, A.; Zisserman, A.; Muñoz, X. Scene classification via PLSA. In Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006.
119. Wu, J. A fast dual method for HIK SVM learning. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010.
120. Bosch, A.; Zisserman, A.; Muoz, X. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 712–727. [CrossRef]
121. Gu, Y.; Liu, H.; Wang, T.; Li, S.; Gao, G. Deep feature extraction and motion representation for satellite video scene classification. *Sci. China Inf. Sci.* **2020**, *63*, 140307. [CrossRef]
122. Tuia, D.; Marcos, D.; Schindler, K.; Saux, B.L. Deep Learning-based Semantic Segmentation in Remote Sensing. In *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*; Camps-Valls, G., Tuia, D., Zhu, X., Reichstein, M., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2021; Volume 5, pp. 46–66. [CrossRef]
123. Lin, D. MARTA GANs: Deep Unsupervised Representation Learning for Remote Sensing Images. *arXiv* **2016**, arXiv:1612.08879.
124. Qi, K.; Zhang, X.; Wu, B.; Wu, H. Sparse coding-based correlation model for land-use scene classification in high-resolution remote-sensing images. *J. Appl. Remote Sens.* **2016**, *10*, 042005.
125. Xie, J.; He, N.; Fang, L.; Plaza, A. Scale-Free Convolutional Neural Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6916–6928. [CrossRef]
126. Du, P.; Tan, K.; Xing, X. A novel binary tree support vector machine for hyperspectral remote sensing image classification. *Opt. Commun.* **2012**, *285*, 3054–3060. [CrossRef]
127. Zhao, Z.; Luo, Z.; Li, J.; Chen, C.; Piao, Y. When Self-Supervised Learning Meets Scene Classification: Remote Sensing Scene Classification Based on a Multitask Learning Framework. *Remote Sens.* **2020**, *12*, 3276. [CrossRef]
128. Ma, A.; Wan, Y.; Zhong, Y.; Wang, J.; Zhang, L. SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. *ISPRS J. Photogramm. Remote Sens.* **2021**, *172*, 171–188. [CrossRef]
129. Risojevi, V.; Stojni, V. The Role of Pre-Training in High-Resolution Remote Sensing Scene Classification. *arXiv* **2021**, arXiv:2111.03690.
130. Boualleg, Y.; Farah, M.; Farah, I.R. Remote Sensing Scene Classification Using Convolutional Features and Deep Forest Classifier. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1944–1948. [CrossRef]
131. Li, E.; Samat, A.; Du, P.; Liu, W.; Hu, J. Improved Bilinear CNN Model for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
132. Xu, K.; Huang, H.; Deng, P.; Shi, G. Two-stream Feature Aggregation Deep Neural Network for Scene Classification of Remote Sensing Images. *Inform. Sci.* **2020**, *539*, 250–268. [CrossRef]

133. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1–5. [CrossRef]
134. Liang, L.; Wang, G. Efficient recurrent attention network for remote sensing scene classification. *IET Image Process.* **2021**, *15*, 1712–1721. [CrossRef]
135. Cheng, G.; Zhou, P.; Han, J.; Han, J.; Guo, L.; Han, J. Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images. *IET Comput. Vis.* **2015**, *9*, 639–647. [CrossRef]
136. Liang, J.; Deng, Y.; Zeng, D. A Deep Neural Network Combined CNN and GCN for Remote Sensing Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4325–4338. [CrossRef]
137. Duan, Y.; Tao, X.; Xu, M.; Han, C.; Lu, J. GAN-NL: Unsupervised Representation Learning for Remote Sensing Image Classification. In Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, 26–29 November 2018.
138. Yu, Y.; Li, X.; Liu, F. Attention GANs: Unsupervised Deep Feature Learning for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 519–531. [CrossRef]
139. Zhu, Q.; Zhong, Y.; Zhang, L.; Li, D. Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6180–6195. [CrossRef]
140. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. 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]
141. Zhang, W.; Tang, P.; Zhao, L. Remote Sensing Image Scene Classification Using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
142. He, N.; Fang, L.; Li, S.; Plaza, J.; Plaza, A. Skip-connected covariance network for remote sensing scene classification. *IEEE T. Neur. Net. Lear.* **2020**, *31*, 1461–1474. [CrossRef]
143. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. *arXiv* **2019**, arXiv:1902.06148.
144. Qian, X.; Li, E.; Zhang, J.; Zhao, S.; Wu, Q.; Zhang, H.; Wang, W.; Wu, Y. Hardness recognition of robotic forearm based on semi-supervised generative adversarial networks. *Front. Neurorobot.* **2019**, *13*, 73. [CrossRef] [PubMed]
145. Han, W.; Feng, R.; Wang, L.; Cheng, Y. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 23–43. [CrossRef]
146. Soto, P.J.; Bermudez, J.D.; Happ, P.N.; Feitosa, R. A comparative analysis of unsupervised and semi- supervised representation learning for remote sensing image categorization. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *IV-2/W7*, 167–173. [CrossRef]
147. Fan, J.; Tan, H.; Lu, S. Multipath sparse coding for scene classification in very high resolution satellite imagery. In Proceedings of the SPIE 9643, Image and Signal Processing for Remote Sensing XXI, Toulouse, France, 15 October 2015.
148. Othman, E.; Bazi, Y.; Melgani, F.; Alhichri, H.; Alajlan, N.; Zuair, M. Domain adaptation network for cross-scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4441–4456. [CrossRef]
149. 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]
150. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv* **2014**, arXiv:1312.6229.
151. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *arXiv* **2013**, arXiv:1310.1531.
152. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM international conference on Multimedia, New York, NY, USA, 3–7 November 2014.
153. Chung, A.; Shafiee, M.; Wong, L. Random feature maps via a Layered Random Projection (LARP) framework for object classification. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
154. Luus, F.P.; Salmon, B.P.; Van Den Bergh, F.; Maharaj, B.T.J. Multiview deep learning for land-use classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2448–2452. [CrossRef]
155. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
156. Nogueira, K.; Penatti, O.; Santos, J. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recogn.* **2017**, *61*, 539–556. [CrossRef]
157. He, N.; Fang, L.; Li, S.; Plaza, A.; Plaza, J. Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6899–6910. [CrossRef]
158. Penatti, O.; Nogueira, K.; Dos Santos, J.A. 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.
159. Li, B.; Su, W.; Wu, H.; Li, R.; Zhang, W.; Qin, W.; Zhang, S. Aggregated Deep Fisher Feature for VHR Remote Sensing Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3508–3523. [CrossRef]

160. Scott, G.; Hagan, K.; Marcum, R.; Hurt, J.; Anderson, D.; Davis, C. Enhanced Fusion of Deep Neural Networks for Classification of Benchmark High-Resolution Image Data Sets. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1451–1455. [CrossRef]
161. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL Internationa Conference on Advances in Geographic Information Systems (ACM 2010), San Jose, CA, USA, 3–5 November 2010; pp. 270–279.
162. Risojević, V.; Babić, Z. Aerial image classification using structural texture similarity. In Proceedings of the 2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, Spain, 14–17 December 2011.
163. Risojević, V.; Babić, Z. Orientation difference descriptor for aerial image classification. In Proceedings of the 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), Vienna, Austria, 11–13 April 2012.
164. Chen, C.; Zhang, B.; Su, H.; Li, W.; Wang, L. Land-use scene classification using multi-scale completed local binary patterns. *Signal Image Video Processing* **2016**, *10*, 745–752. [CrossRef]
165. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
166. Shao, W.; Yang, W.; Xia, G.; Liu, G. A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization. In Proceedings of the 9th International Conference, ICVS 2013, Saint Petersburg, Russia, 16–18 July 2013.
167. Zhao, L.; Tang, P.; Huo, L. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4620–4631. [CrossRef]
168. Zhao, L.; Tang, P.; Huo, L. A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification. *Int. J. Remote Sens.* **2014**, *35*, 2296–2310. [CrossRef]
169. Cheriyadat, A.M. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 439–451. [CrossRef]
170. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
171. Negrel, R.; Picard, D.; Gosselin, P. Evaluation of second-order visual features for land-use classification. In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), Klagenfurt, Austria, 18–20 June 2014.
172. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2175–2184. [CrossRef]
173. Chen, S.; Tian, Y. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1947–1957. [CrossRef]
174. Hu, F.; Xia, G.; Wang, Z.; Huang, X.; Zhang, L.; Sun, H. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2015–2030. [CrossRef]
175. Zhong, Y.; Zhu, Q.; Zhang, L. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [CrossRef]
176. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land use classification in remote sensing images by convolutional neural networks. *arXiv* **2015**, arXiv:1508.00092.
177. Luo, J.; Kitamura, G.; Arefan, D.; Doganay, E.; Panigrahy, A.; Wu, S. Knowledge-Guided Multiview Deep Curriculum Learning for Elbow Fracture Classification. In Proceedings of the 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, 27 September 2021.
178. 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]
179. Liu, Q.; Hang, R.; Song, H.; Zhu, H.; Plaza, J.; Plaza, A. Adaptive deep pyramid matching for remote sensing scene classification. *arXiv* **2016**, arXiv:1611.03589.
180. Han, X.; Zhong, Y.; Zhao, B.; Zhang, L. Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery. *Int. J. Remote Sens.* **2017**, *38*, 514–536. [CrossRef]
181. Lin, D.; Fu, K.; Wang, Y.; Xu, G.; Sun, X. MARTA GANs: Unsupervised representation learning for remote sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2092–2096. [CrossRef]
182. Shawky, O.A.; Hagag, A.; El-Dahshan, E.S.A.; Ismail, M. A very high-resolution scene classification model using transfer deep CNNs based on saliency features. *Signal Image Video Processing* **2021**, *15*, 817–825. [CrossRef]
183. Bian, X.; Chen, C.; Tian, L.; Du, Q. Fusing local and global features for high-resolution scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 2889–2901. [CrossRef]
184. Anwer, R.M.; Khan, F.S.; Van De Weijer, J.; Molinier, M.; Laaksonen, J. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 74–85. [CrossRef]
185. Qi, K.; Guan, Q.; Yang, C.; Peng, F.; Shen, S.; Wu, H. Concentric Circle Pooling in Deep Convolutional Networks for Remote Sensing Scene Classification. *Remote Sens.* **2018**, *10*, 934. [CrossRef]
186. Zeng, D.; Chen, S.; Chen, B.; Li, S. Improving Remote Sensing Scene Classification by Integrating Global-Context and Local-Object Features. *Remote Sens.* **2018**, *10*, 734. [CrossRef]
187. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1155–1167. [CrossRef]

188. Wang, W.; Du, L.; Gao, Y.; Su, Y.; Wang, F.; Cheng, J. A Discriminative Learned CNN Embedding for Remote Sensing Image Scene Classification. *arXiv* **2019**, arXiv:1911.12517.
189. Yu, Y.; Liu, F. Dense Connectivity Based Two-Stream Deep Feature Fusion Framework for Aerial Scene Classification. *Remote Sens.* **2018**, *10*, 1158. [CrossRef]
190. Yu, Y.; Liu, F. A two-stream deep fusion framework for high-resolution aerial scene classification. *Comput. Intel. Neurosc.* **2018**, *2018*, 8639367. [CrossRef]
191. Ye, L.; Wang, L.; Sun, Y.; Zhao, L.; Wei, Y. Parallel multi-stage features fusion of deep convolutional neural networks for aerial scene classification. *Remote Sens. Lett.* **2018**, *9*, 294–303. [CrossRef]
192. Chen, J.; Wang, C.; Ma, Z.; Chen, J.; He, D.; Ackland, S. Remote Sensing Scene Classification Based on Convolutional Neural Networks Pre-Trained Using Attention-Guided Sparse Filters. *Remote Sens.* **2018**, *10*, 290. [CrossRef]
193. Akodad, S.; Vilfroy, S.; Bombrun, L.; Cavalcante, C.C.; Germain, C.; Berthoumieu, Y. An ensemble learning approach for the classification of remote sensing scenes based on covariance pooling of CNN features. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019.
194. Liu, Y.; Zhong, Y.; Qin, Q. Scene Classification Based on Multiscale Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7109–7121.
195. Lu, X.; Ji, W.; Liu, W.; Zheng, X. Bidirectional adaptive feature fusion for remote sensing scene classification. *Neurocomputing* **2019**, *328*, 135–146. [CrossRef]
196. He, H.; Garcia, E. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
197. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
198. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies* **2021**, *9*, 2. [CrossRef]