remote sensing

MDPI

Article

Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification

Xiaobing Han [1,2], Yanfei Zhong [1,2,*], Liqin Cao [3,*] and Liangpei Zhang [1,2]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; whu_hxb@163.com (X.H.); zlp62@whu.edu.cn (L.Z.)
[2] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
[3] School of Printing and Packaging, Wuhan University, Wuhan 430079, China
[*] Correspondence: zhongyanfei@whu.edu.cn (Y.Z.); clq@whu.edu.cn (L.C.); Tel.: +86-27-68779969

Received: 11 July 2017; Accepted: 7 August 2017; Published: 16 August 2017

Abstract: The rapid development of high spatial resolution (HSR) remote sensing imagery techniques not only provide a considerable amount of datasets for scene classification tasks but also request an appropriate scene classification choice when facing with finite labeled samples. AlexNet, as a relatively simple convolutional neural network (CNN) architecture, has obtained great success in scene classification tasks and has been proven to be an excellent foundational hierarchical and automatic scene classification technique. However, current HSR remote sensing imagery scene classification datasets always have the characteristics of small quantities and simple categories, where the limited annotated labeling samples easily cause non-convergence. For HSR remote sensing imagery, multi-scale information of the same scenes can represent the scene semantics to a certain extent but lacks an efficient fusion expression manner. Meanwhile, the current pre-trained AlexNet architecture lacks a kind of appropriate supervision for enhancing the performance of this model, which easily causes overfitting. In this paper, an improved pre-trained AlexNet architecture named pre-trained AlexNet-SPP-SS has been proposed, which incorporates the scale pooling—spatial pyramid pooling (SPP) and side supervision (SS) to improve the above two situations. Extensive experimental results conducted on the UC Merced dataset and the Google Image dataset of SIRI-WHU have demonstrated that the proposed pre-trained AlexNet-SPP-SS model is superior to the original AlexNet architecture as well as the traditional scene classification methods.

Keywords: scene classification; convolutional neural network; pre-trained AlexNet; spatial pyramid pooling; side supervision; high spatial resolution remote sensing imagery

1. Introduction

With the recent launch of remote sensing satellites around the world, a large volume of multi-level, multi-angle, and multi-resolution HSR remote sensing images can now be obtained, where the remote sensing big data brings new understandings for the traditional definition of big data [1–3]. These multi-source remote sensing images allow the ground object observation from multiple perspectives. The rapid development of HSR remote sensing imaging sensors has provided us with a large number of HSR remote sensing images with abundant detail and structural information, and a higher spatial resolution [1]. In addition, these multi-source HSR remote sensing images also provide a huge amount of data without corresponding labels, which may consume a large amount of human labor for labeling. Traditional HSR remote sensing imagery understanding is based on recognizing pixel-based or object-based ground elements, but this cannot describe the whole content of the scene images and cannot well bridge the "semantic gap" between the low-level features and the high-level semantics [4].

Scene classification for HSR remote sensing imagery is aimed at obtaining the semantic category information of the scene images, where the core idea of HSR remote sensing imagery scene classification is to bridge the semantic gap and to explore the high-level semantic category information contained within the scenes [5–8].

To adequately bridge the semantic gap between the low-level features and the high-level semantics, various scene classification methods have been proposed in recent years. The traditional HSR remote sensing imagery scene classification methods include bag of visual words (BOVW) [9–12], spatial pyramid matching (SPM) [13], latent Dirichlet allocation (LDA) [14–16], and probabilistic latent semantic allocation (PLSA) [17,18]. These methods adopt manual feature extraction techniques, namely the spectral features, textural features, and structural features (e.g., scale invariant feature transformation, SIFT [19]), to realize scene semantic recognition. However, all these approaches adopt manually designed feature descriptors for the predefined algorithms [4–10], which require expert engineering experience.

Recently, with the development of deep learning [20–27], much effort has been dedicated to developing automatic and discriminative feature extraction and representation frameworks for HSR remote sensing imagery scene classification [26–34]. Previous works have proven that CNN are excellent deep learning model for HSR remote sensing imagery scene semantic recognition [26,28–30,32,35–38] or image classification [39,40], and they can efficiently and automatically extract features derived from the data. A CNN is a hierarchical feature representation framework consisting of multiple alternate convolutional and pooling layers, with a back-propagation mechanism to tune the whole network to obtain the final classification result [35]. However, according to the current research status, research into CNN models for HSR remote sensing imagery scene classification can be summarized into two research trends.

The first research trend of CNN models is focused on carefully designing an effective and accurate network architecture to obtain a satisfactory HSR remote sensing scene classification result. For instance, to improve HSR remote sensing imagery scene classification performance, Zhang et al. [33,35] proposed an improved gradient boosting CNN ensemble framework to reuse the weights in each random convolutional network. Compared with the large and complicated natural imagery scene datasets, as introduced in [20], the current HSR remote sensing imagery scene datasets have the characteristics of small quantities, simple categories, simple content, multi-scale objects et al., which results in the manually designed CNN models being faced with many critical challenges when the labeled samples are limited. To better deal with the above situations, another effective and meaningful research trend of HSR remote sensing imagery scene classification with CNN models is the introduction of the pre-training mechanism. A pre-trained CNN architecture involves first training the existing CNN model upon a large natural imagery dataset, and then a transfer mechanism is used to convey the network parameters from the natural imagery dataset to the HSR remote sensing imagery dataset [32], considering some of the specific similarities between HSR remote sensing imagery scene dataset and natural imagery scene dataset. Marco et al. [28] were the first to prove that the transfer of a pre-trained CNN can achieve a promising classification performance. Hu et al. [32] further explored that the transferability of the natural image features from the pre-trained CNN applicable to the limited amount of HSR remote sensing scene datasets with the feature coding methods. The advantage of the second research trend of pre-trained CNN models is their effective extensible properties for dealing with the HSR remote sensing imagery scenes with limited labeling. However, the pre-trained CNN models seldom consider fusing the multi-scale information of the last convolved feature maps.

Although the pre-training mechanism can help CNN models achieve satisfactory classification performances for HSR remote sensing imagery scenes with limited labeled samples, the choice of a proper network architecture for making strong and correct assumptions about the nature of the input data is a big challenge for the current HSR remote sensing imagery scene classification techniques. Thus, research into a simple network architecture with a powerful modelling capability is urgently needed. AlexNet, as a simple, typical, foundational, and one of the state-of-the-art CNN architecture, was first proposed by Hinton and was successfully utilized in the 2012 ImageNet Competition [21]. Compared with the other structure-complex and deep CNN architectures (e.g., GoogLeNet [22], VGG et al. [23]), AlexNet is a structure-simple CNN architecture, which is easy to train and optimize. When fine-tuned with HSR remote sensing imagery datasets, a fast and satisfactory classification result can be obtained. However, considering the multi-scale characteristic in some specific semantic scenes with key objects, the properties of the current pre-trained AlexNet architecture are limited. In order to further improve the classification performance and adequately consider the multi-scale information with the AlexNet architecture, an improved pre-trained AlexNet architecture is needed.

In order to better deal with the multi-scale information of the convolved feature maps of the HSR remote sensing scene images and fuse this information, a multi-scale pooling strategy, named spatial pyramid pooling (SPP) [13,41–43], is incorporated into the pre-trained AlexNet classification architecture. SPP is a pooling strategy proposed by He et al. [44] for object detection tasks, which was developed from the SPM model proposed by Lazebnik et al. in [13], and extended research done in [41–44]. The SPP strategy operates on the multi-scale convolved feature maps, and concatenates the different-scale convolved feature maps, which adequately takes the multi-scale spatial information of the same scenes into consideration and can narrow the semantic differences for the scenes with multi-scale information.

Although the pre-trained AlexNet architecture can handle scenes containing multi-scale information with the SPP strategy, the relatively simple pre-trained AlexNet architecture still lacks an efficient side supervision (SS) technique to prevent overfitting of the AlexNet architecture. To further improve the performance of the pre-trained AlexNet architecture, an effective improvement is needed to be incorporated into the pre-trained AlexNet architecture. The SS strategy firstly derived from [45] is an effective companion operation which incorporates deep supervision into both the hidden layers of the deep CNN and the final output layer to propagate this supervision to the previous layers, simultaneously minimizing the classification error. In addition, SS can also reduce the gradient vanishing phenomenon and prevent overfitting of the CNN architecture. By introducing the SPP and SS strategies into the pre-trained AlexNet architecture, the multi-scale operation and the simultaneous minimization operation can be handled at the same time for the pre-trained AlexNet architecture, which enables the pre-trained AlexNet architecture with better properties to better deal with HSR remote sensing imagery scene classification.

The main contributions of this paper can be summarized as follows.

(a) The end-to-end AlexNet classification architecture. Differing from the complicated and stepwise operation of the AlexNet classification architecture, the proposed pre-trained AlexNet-SPP-SS model is an end-to-end operation. Pre-trained AlexNet-SPP-SS deals with the label-limited HSR remote sensing imagery scene classification task with fast and effective one-step heterologous parameter transferring and pre-training operations, enabling the whole procedure to be more convenient, reducing the complicated intermediate operations, and reducing the resource consumption.

(b) The effective multi-scale pyramid pooling scene interpretation capability. The SPP strategy is incorporated into the end-to-end pre-trained AlexNet architecture, and solves the multi-scale scene interpretation task by fusing the different-scale convolved feature maps, which adequately considers the spatial information in different scales and increases the scene interpretation ability.

(c) The simultaneous supervision processing framework. To make the end-to-end pre-trained AlexNet architecture more transparent in dealing with the heterologous parameter transferring in quantity-limited HSR remote sensing imagery scene classification, the SS strategy is incorporated by introducing intermediate supervision to the layers of the pre-trained AlexNet architecture, to reduce the gradient vanishing phenomenon and prevent overfitting of the whole architecture.

To test the performance of the proposed pre-trained AlexNet-SPP-SS model, extensive experiments were conducted on HSR remote sensing datasets—the UC Merced dataset and the Google image dataset of SIRI-WHU—with the pre-trained network parameters transferred from natural image datasets, demonstrating that the proposed pre-trained AlexNet-SPP-SS model can perform better than the pre-trained AlexNet architecture, the AlexNet-SPP architecture, and the AlexNet-SS architecture, as well as the traditional handcrafted feature based HSR remote sensing imagery scene classification approaches.

The rest of this paper is organized as follows. In Section 2, the typical AlexNet architecture is introduced. In Section 3, the SPP strategy, the SS strategy, and the proposed AlexNet-SPP-SS model are described in detail. The experimental datasets, the experimental results, and an analysis are given in Sections 4 and 5. Section 6 presents a discussion. Section 7 draws our conclusions.

2. The AlexNet Architecture

AlexNet, which was first proposed by Alex Krizhevsky et al. in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012) [21], is a fundamental, simple, and effective CNN architecture, which is mainly composed of cascaded stages, namely, convolution layers, pooling layers, rectified linear unit (ReLU) layers and fully connected layers. Specifically, AlexNet is composed of five convolutional layers, the first layer, the second layer, the third layer and the fourth layer followed by the pooling layer, and the fifth layer followed by three fully-connected layers. For the AlexNet architecture, the convolutional kernels are extracted during the back-propagation optimization procedure by optimizing the whole cost function with the stochastic gradient descent (SGD) algorithm. Generally, the convolutional layers act upon the input feature maps with the sliding convolutional kernels to generate the convolved feature maps, and the pooling layers operate on the convolved feature maps to aggregate the information within the given neighborhood window with a max pooling operation or average pooling operation. The reason why AlexNet is successful can be attributed to some of the practical strategies, for instance, the ReLU non-linearity layer and the dropout regularization technique. The ReLU, as shown in Equation (1), is a half-wave rectifier function, which can significantly accelerate the training phase and prevent overfitting. The dropout technique can be regarded as a kind of regularization by stochastically setting a number of the input neurons or hidden neurons to be zero to reduce the co-adaptations of the neurons, which is usually utilized in the fully connected layers in the AlexNet architecture.

$$f(x) = \max(x, 0) \tag{1}$$

The transfer mechanism and the pre-training mechanism allow the CNN network parameters to be transferred from natural imagery datasets to HSR remote sensing imagery datasets. The reason why this can succeed can be explained, to some extent, by the similarities between natural imagery datasets and remote sensing imagery datasets, and the category compatibility. It can also be easily understood that the large and complicated ImageNet datasets can help to obtain a well-trained AlexNet architecture, and well-trained network parameters are important for initializing the subsequent classification framework. Therefore, the pre-training mechanism helps the AlexNet architecture to perform the HSR remote sensing imagery scene classification task. Based on the introduction of the convenient and comprehensive representation ability of the pre-trained AlexNet architecture in dealing with HSR remote sensing imagery scene classification, the pre-training mechanism also makes the AlexNet architecture an end-to-end classification pipeline. The pre-trained AlexNet network architecture is shown in Figure 1.

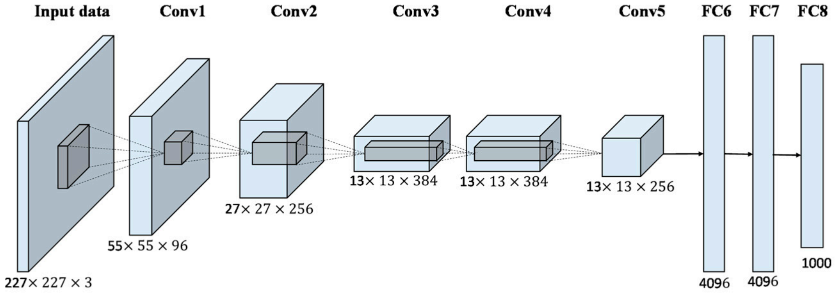


Figure 1. The AlexNet architecture.

3. The Proposed AlexNet-SPP-SS Architecture for High Spatial Resolution Remote Sensing Imagery Scene Classification

It is noted that the simplicity and convenience of the pre-training mechanism in the AlexNet architecture make the pre-trained AlexNet architecture a good choice in dealing with HSR remote sensing imagery scene classification. In order to further mine the properties of the pre-trained AlexNet for the HSR remote sensing imagery scene classification tasks, and to adequately consider the multi-scale properties of the ground objects as well as the simultaneous processing capacity, an improved pre-trained CNN architecture named the pre-trained AlexNet-SPP-SS architecture is proposed in this paper. The proposed pre-trained AlexNet-SPP-SS model is introduced below.

3.1. The Effective Multi-Scale Pyramid Pooling Ground Objects Scene Interpretation Strategy—Spatial Pyramid Pooling (SPP)

SPP developed from the SPM model [13] for object recognition and scene classification [41–43], to improve the performance of the CNN architecture [44], SPP deals with the multi-scale convolved feature maps to generate a fixed-length pooling representation, regardless of the image size, and concatenates the pooled feature maps into a long single vector. As the multi-scale convolved feature maps contain abundant complementary spatial information, especially the scenes containing key ground objects, the incorporation of the SPP strategy can enhance the scene interpretation capability. The SPP strategy also has the outstanding advantage of generating a fixed-length pooling feature representation, regardless of image size, and is thus able to deal with the images of arbitrary scales.

The advantages of incorporating the SPP strategy into the pre-trained AlexNet architecture can be summarized from three aspects. The first advantage of SPP is that it computes the convolved feature maps only once from the entire image, and it pools the convolved features in arbitrary-scale regions to generate a fixed-length representation. The second advantage of SPP is that it can utilize multi-scale spatial bins, which is an approach that has been shown to be robust to object deformation, while the sliding window pooling only uses a single window size. The third advantage of SPP is that it can pool features extracted at different scales. These advantages enable the pre-trained AlexNet architecture to interpret HSR remote sensing imagery scenes with multi-scale ground objects. The multi-scale processing procedure of the SPP strategy is shown in Figure 2.

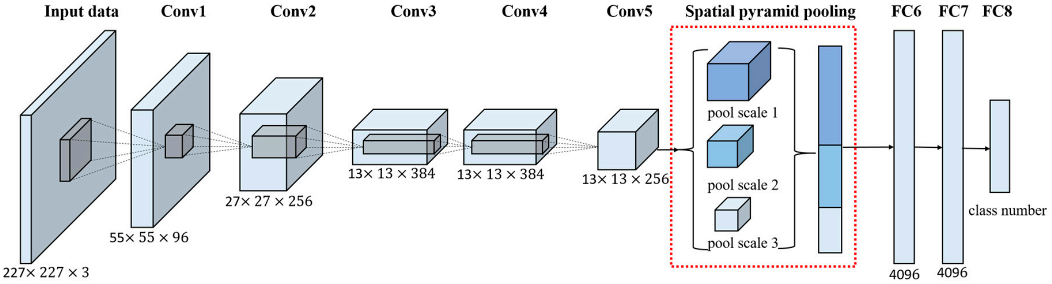


Figure 2. The AlexNet architecture with spatial pyramid pooling.

3.2. The Simultaneous Supervision Processing Framework: the Side Supervision (SS) Strategy

The pre-trained AlexNet architecture is an effective end-to-end HSR remote sensing imagery scene classification framework, but it only deals with the classification task with the final supervision term. It is noted that the goal of the pre-trained AlexNet architecture is to learn layers of filters and weights for the minimization of the classification error at the final output layer. However, the single supervision term limits the ability of the pre-trained AlexNet architecture to deal with the simultaneous and transparent classification error minimization. To alleviate the phenomenon of non-simultaneous and non-transparent processing in the pre-trained AlexNet architecture, a supervision [45] strategy is incorporated into the pre-trained AlexNet architecture. SS is a strong convex strategy, which enforces the feature robustness and discriminative ability through both final-layer supervision and intermediate-layer supervision. Specifically, the core idea of SS is aimed at providing integrated direct supervision to the hidden layers, which is in contrast to the standard approach of providing supervision only at the output layer and propagating this supervision back to earlier layers. The SS is added by the companion objective function for each hidden layer, and can be regarded as an additional constraint within the learning process.

For the pre-trained AlexNet architecture, there are three obvious problems with the current architecture. The first problem is the non-transparency in the intermediate layers during the overall classification procedure, which makes the training process difficult to observe. The second problem refers to the robustness and discriminative ability of the learned features, especially in the latter layers of the network, which can significantly influence the performance. The third problem is the low training effectiveness in the face of "exploding" and "vanishing" gradients. In order to better deal with the problems existing in the current pre-trained AlexNet architecture, there are two significant advantages of introducing the SS companion objective functions into the pre-trained AlexNet architecture. The first advantage is that the SS functions are strong convex regularization functions for both the large training data in deeper networks and the small training data in relatively shallower networks, which can increase the robustness and discriminative ability of the learned features in the pre-trained AlexNet architecture. The second advantage is that SS can make the intermediate layers transparent during the training process.

In order to allow a better understanding of the pre-trained AlexNet architecture with the SPP and SS strategies, an illustration is given below. Suppose that the input sample $\mathrm{X_i} \in \mathrm{R^n}$ denotes the raw input data and $\mathrm{y_i} \in \{1, \ldots, \mathrm{K}\}$ denotes the corresponding ground truth label for sample $\mathrm{X_i}$. Suppose that there are M layers in total in the pre-trained AlexNet architecture, the weight combinations for the pre-trained AlexNet architecture are $\mathrm{W} = (\mathrm{W}^{(1)}, \ldots, \mathrm{W}^{(\mathrm{M})})$. Meanwhile, for each classifier in each hidden layer of the pre-trained AlexNet architecture, the corresponding weights are $\mathrm{w} = (\mathrm{w}^{(1)}, \ldots, \mathrm{w}^{(\mathrm{M}-1)})$. In the pre-trained AlexNet architecture, the relationships between the weight parameters and the filters are respectively shown in Equations (2) and (3):

$$\mathrm{Z}^{(\mathrm{m})} = \mathrm{f}(\mathrm{Q}^{(\mathrm{m})}) \text{ and } \mathrm{Z}^{(0)} = \mathrm{X} \tag{2}$$

$$\mathrm{Q}^{(\mathrm{m})} = \mathrm{W}^{(\mathrm{m})} * \mathrm{Z}^{(\mathrm{m}-1)} \tag{3}$$

In Equations (2) and (3), M denotes the total layer number of the pre-trained AlexNet architecture; m refers to the specific layer of the pre-trained AlexNet architecture; $\mathrm{W}^{(\mathrm{m})}, \mathrm{m} = 1 \ldots \mathrm{M}$ are the network weights to be learned; $\mathrm{Q}^{(\mathrm{m})}$ refers to the convolved responses on the previous feature map; and $\mathrm{f}()$ is the pooling function on Q. The total objective function for the pre-trained AlexNet architecture is shown in Equation (4).

$$\mathrm{F}(\mathrm{W}) = \mathrm{P}(\mathrm{W}) + \mathrm{Q}(\mathrm{W}) \tag{4}$$

where $\mathrm{P(W)}$ and $\mathrm{Q(W)}$ refer to the output objective and the summed companion objectives, which are defined in Equations (5) and (6), respectively.

$$\mathrm{P(W)} \equiv \|\mathrm{w}^{(\mathrm{out})}\|^2 + \mathrm{L}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})}) \tag{5}$$

$$\mathrm{Q(W)} \equiv \sum_{\mathrm{m}=1}^{\mathrm{M}-1} \left[\|\mathrm{w}^{(\mathrm{M})}\|^2 + \mathrm{l}(\mathrm{W}, \mathrm{w}^{(\mathrm{m})}) - \mathrm{r} \right] \tag{6}$$

where $\mathrm{w}^{(\mathrm{out})}$ refers to the classifier weight of the output layer. The final combined objective function of the pre-trained AlexNet architecture is defined in Equation (7).

$$\|\mathrm{w}^{(\mathrm{out})}\|^2 + \mathrm{L}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})}) + \sum_{\mathrm{m}=1}^{\mathrm{M}-1} \mathrm{a_m} \left[\|\mathrm{w}^{(\mathrm{m})}\|^2 + \mathrm{l}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})}) - \mathrm{r} \right] \tag{7}$$

where $\|\mathrm{w}^{(\mathrm{out})}\|^2$ and $\mathrm{L}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})})$ are respectively the margin and squared hinge loss of the support vector machine (SVM) classifier. $\|\mathrm{w}^{(\mathrm{m})}\|^2$ and $\mathrm{l}(\mathrm{W}, \mathrm{w}^{(\mathrm{m})})$ are respectively the margin and squared hinge loss of the SVM classifier at each hidden layer. The overall loss of the output layer $\mathrm{L}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})})$ is as shown in Equation (8):

$$\mathrm{L}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})}) = \sum_{\mathrm{y_k{}^1y}} \left[1- < \mathrm{w}^{(\mathrm{out})}, \mathrm{f}(\mathrm{Z}^{(\mathrm{M})}, \mathrm{y}) - \mathrm{f}(\mathrm{Z}^{(\mathrm{M})}, \mathrm{y_k}) \right] \tag{8}$$

In Equation (7), $\mathrm{l}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})})$ as the companion loss of the intermediate layers is as shown in Equation (9).

$$\mathrm{l}(\mathrm{W}, \mathrm{w}^{(\mathrm{out})}) = \sum_{\mathrm{y_k{}^1y}} \left[1- < \mathrm{w}^{(\mathrm{m})}, \mathrm{f}(\mathrm{Z}^{(\mathrm{m})}, \mathrm{y}) - \mathrm{f}(\mathrm{Z}^{(\mathrm{m})}, \mathrm{y_k}) > \right]_+^2 \tag{9}$$

For Equations (8) and (9), they are both squared hinge losses of the prediction errors. From the above formulations, it can be understood intuitively that, in Equations (8) and (9), the pre-trained AlexNet architecture not only learns the convolutional kernels $W^\star$, but enforces a constraint at each hidden layer to directly make a good label prediction and give a strong push for having discriminative and sensible features at each individual layer. It is noted that for each $\mathrm{l}(\mathrm{W}, \mathrm{w}^{(\mathrm{m})})$, the $\mathrm{w}^{(\mathrm{m})}$ directly depends on $\mathrm{Z}^{(\mathrm{m})}$, which is dependent on $\mathrm{W}^1, \ldots, \mathrm{W}^\mathrm{m}$ up to the mth layer. The second term often goes to zero during the course of training. In this way, the overall goal of producing a good classification result at the output layer is not altered and the companion objective just acts as a proxy or regularization. To achieve this goal, the threshold γ is usually set in the second term of Equation (6). The working mechanism of this companion function is that the hinge losses of the overall function and the companion objective function vanish and no longer play a role in the learning process when the overall value of the hidden layer reaches or is below γ. α_m balances the importance of the error in the output objective and the companion objective.

To summarize, the working mechanism of the pre-trained AlexNet architecture with SS strategy is that the output performance of the entire network is achieved with a "satisfactory" level of performance on the part of the hidden layer classifiers. For the pre-trained AlexNet architecture with SS strategy, the optimization procedure is conducted using the SGD algorithm and the gradient functions in a similar way to the original AlexNet architecture. To better demonstrate the working details and the processing manner of the pre-trained AlexNet architecture, a flowchart is provided in Figure 3.

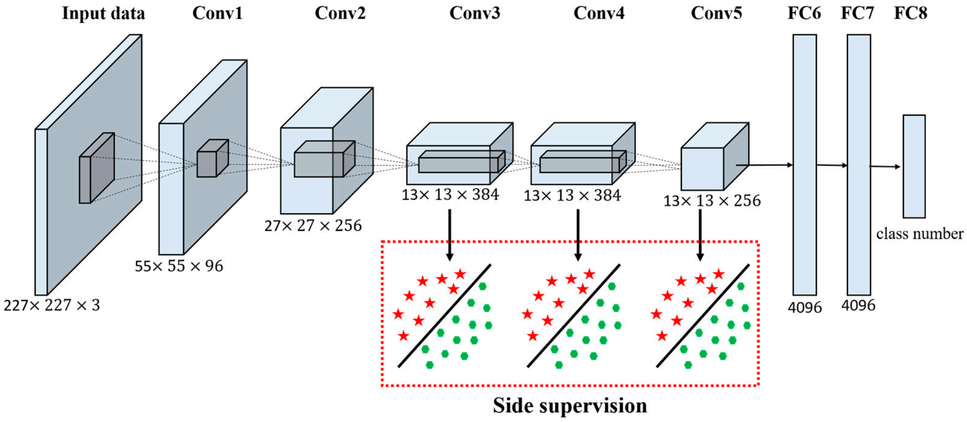


Figure 3. The AlexNet architecture with side supervision.

3.3. The Proposed Pre-Trained AlexNet-SPP-SS Model for High Spatial Resolution Remote Sensing Imagery Scene Classification

As a simple and effective HSR remote sensing imagery scene classification model, the AlexNet architecture has some disadvantages in dealing with both the non-transparency phenomenon of the intermediate layers and scene classification tasks with the multi-scale thematic scenes. To quickly make the objective function converge to an optimal value, the network weight parameters transferred from the natural images are retrained in the HSR remote sensing imagery scene classification tasks, where the pre-trained AlexNet architecture is derived from pre-training AlexNet architecture on large-scale natural imagery datasets. The similar semantic scene information helps the pre-trained AlexNet architecture to obtain a fast and satisfactory result for the HSR remote sensing scene images. In order to deal with the multi-scale phenomenon of the specific multi-scale semantic scenes, SPP, as a kind of effective multi-scale pooling operation is added into the pre-trained AlexNet architecture. To better represent the intermediate layer information of the pre-trained AlexNet architecture, SS, as a kind of useful intermediate supervision incorporation strategy, can help the pre-trained AlexNet architecture improve the classification performance not only from the aspect of the robustness of the network weights but also from the aspect of the transparency of the intermediate layers. Based on the advantages of the SPP and the SS strategies, and for the purpose of further improving the HSR remote sensing imagery scene classification performance with the pre-trained AlexNet architecture, the pre-trained AlexNet-SPP-SS model is proposed to first incorporate the SPP and SS strategies into the pre-trained AlexNet architecture.

The pre-trained AlexNet-SPP-SS architecture is a combinatorial CNN network architecture, which incorporates the supervision layers as the intermediate layers of the pre-trained AlexNet architecture and also combines the SPP layers into the AlexNet architecture to allow the pre-trained AlexNet architecture to have the ability to both deal with the multi-scale information of the pre-trained AlexNet architecture and simultaneously process the SS information. In an overall view, the pre-trained AlexNet-SPP-SS model, as shown in Figure 4, endows the HSR remote sensing imagery with limited samples to obtain an improved scene classification performance.

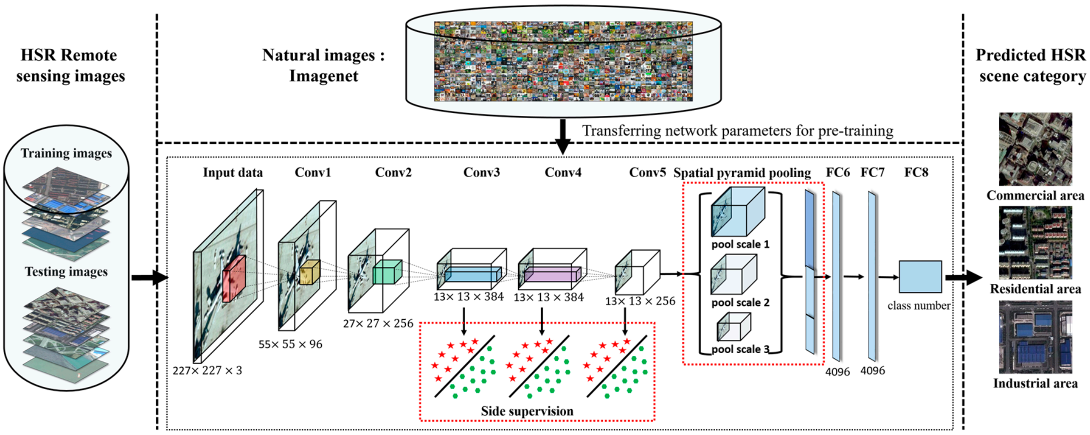


Figure 4. The pre-trained AlexNet architecture with spatial pyramid pooling and side supervision.

4. Datasets and Experiment Scheme

In order to test the performance of the proposed pre-trained AlexNet-SPP-SS model, three datasets, namely the widely utilized UC Merced dataset, the Google image dataset of SIRI-WHU, and WHU-RS dataset were utilized to conduct the experiment.

4.1. Dataset Description

The first dataset utilized for evaluating the performance of the proposed pre-trained AlexNet-SPP-SS model is the UC Merced dataset, which was collected from the USGS National Map Urban Area Imagery collection [46]. This dataset is composed of 21 classes, with 100 samples per class. The image size of the UC Merced dataset is 256 × 256 with a 1-ft spatial resolution. The classes for the UC Merced dataset are: agriculture, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts, as shown in Figure 5. For the UC Merced dataset, 80 samples of each class were stochastically selected as the training samples, and the rest were selected as the testing samples.

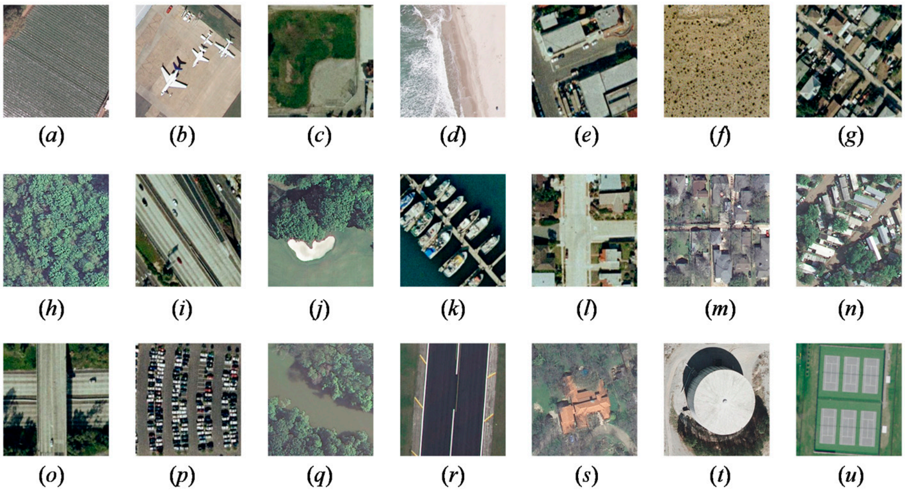


Figure 5. Representative images of the 21 land-use categories in the UC Merced dataset: (**a**) agriculture; (**b**) airplane; (**c**) baseball diamond; (**d**) beach; (**e**) buildings; (**f**) chaparral; (**g**) dense residential; (**h**) forest; (**i**) freeway; (**j**) golf course; (**k**) harbor; (**l**) intersection; (**m**) medium residential; (**n**) mobile home park; (**o**) overpass; (**p**) parking lot; (**q**) river; (**r**) runway; (**s**) sparse residential; (**t**) storage tanks; (**u**) tennis court.

The second dataset utilized for evaluating the performance of the proposed AlexNet-SPP-SS model is the Google image dataset of SIRI-WHU. The Google image dataset of SIRI-WHU covering urban areas in China was collected by the RSIDEA (Intelligent Data Extraction, Analysis and Applications of Remote Sensing) group, LIESMARS, Wuhan University [10,14,47]. The Google image dataset of SIRI-WHU consists of 12 land-use classes, and each class has 200 samples, for which the image size is 200 × 200 and the spatial resolution is 2 m. The class names of the Google image dataset of SIRI-WHU are meadow, pond, harbor, industrial, park, river, residential, overpass, agriculture, water, commercial, and idle land, demonstrate as shown in Figure 6. In this experiment, 160 samples of each class were stochastically selected per class as the training samples, and the rest were retained as the test samples.

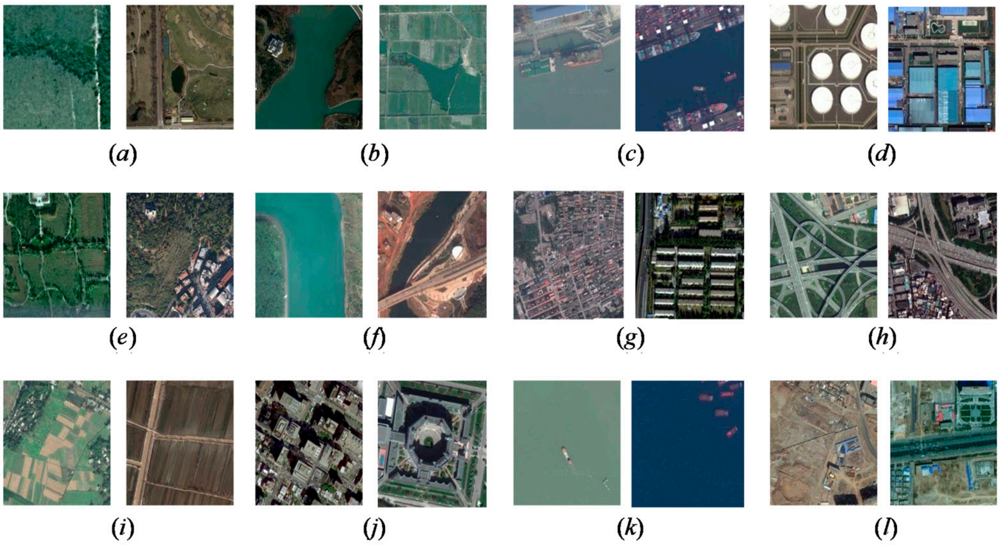


Figure 6. Representative images of the Google Image dataset of SIRI-WHU: (**a**) meadow; (**b**) pond; (**c**) harbor; (**d**) industrial; (**e**) park; (**f**) river; (**g**) residential; (**h**) overpass; (**i**) agriculture; (**j**) commercial; (**k**) water; (**l**) idle land.

The third dataset utilized for evaluating the performance of the proposed AlexNet-SPP-SS model is WHU-RS dataset. The WHU-RS dataset [32], collected from Google Earth (Google Inc., Mountain View, CA, USA), is a new publicly available dataset, which consists of 950 images with a size of 600 × 600 pixels uniformly distributed in 19 scene classes. Some example images are shown in Figure 7. In this experiment, 25 samples of each class were stochastically selected per class as the training samples, and the rest were retained as the test samples.

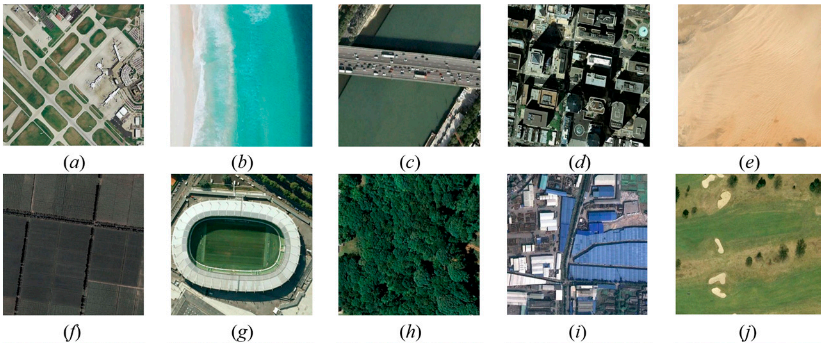


Figure 7. *Cont.*

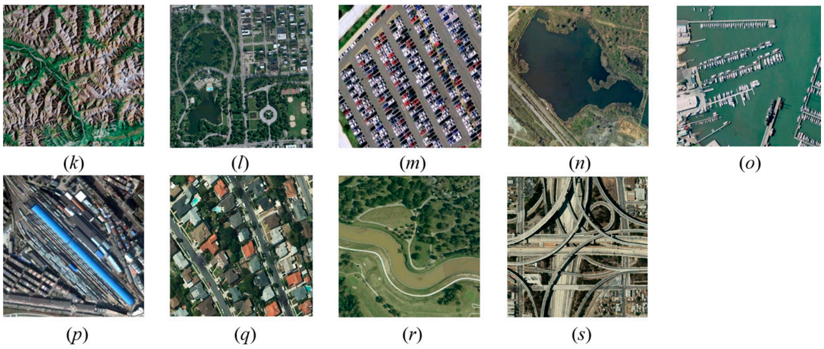


Figure 7. Representative images of the WHU-RS dataset: (**a**) airport; (**b**) beach; (**c**) bridge; (**d**) commercial; (**e**) desert; (**f**) farmland; (**g**) football field; (**h**) forest; (**i**) industrial; (**j**) meadow; (**k**) mountain; (**l**) park; (**m**) parking; (**n**) pond; (**o**) port; (**p**) railway station; (**q**) residential; (**r**) river; (**s**) viaduct.

4.2. Experiment Scheme

To gradually and explicitly demonstrate the advantages of the proposed pre-trained AlexNet-SPP-SS model, the performances of the original AlexNet architecture, the pre-trained AlexNet-SPP architecture, and the pre-trained AlexNet-SS architecture were respectively compared. As introduced in the previous sections of this paper, the pre-trained AlexNet is the AlexNet architecture with weights transferred from the natural images, the pre-trained AlexNet-SPP architecture is the pre-trained AlexNet architecture with SPP strategy, and the pre-trained AlexNet-SS architecture is the pre-trained AlexNet architecture with SS strategy. To demonstrate the advantages of the proposed pre-trained AlexNet-SPP-SS model, the previously proposed pre-trained AlexNet-related architectures, for instance, the AlexNet-BOVW, the AlexNet-FV, the AlexNet-VLAD [32] were also compared. Furthermore, some of the traditional HSR remote sensing imagery scene classification methods—BOW, SPM, LDA, and a non-pre-trained CNN architecture, the gradient boosting random convolutional network (GBRCN) [35] were also compared.

The images were resized to 227 × 227 on the Caffe platform [48]. To increase the diversity of the HSR remote sensing image datasets, a data augmentation strategy was adopted by incorporating five types of cropping with 0° and 180° flipping. The initial learning rates for the three datasets were sequentially set as 0.0001, 0.001, and 0.0001, and the momentum and the weight decay for the three datasets were set as 0.9 and 0.0005 respectively. To test the stability of the proposed pre-trained AlexNet-SPP-SS model, the experiments were executed 10 times to obtain convincing results for the three datasets. The mean value and standard deviation were adopted as the evaluation indicators.

5. Results

In order to evaluate the performance of the proposed pre-trained AlexNet-SPP-SS model and compare with the performances of the traditional classification methods on the UC Merced dataset, the scene classification results are listed in Table 1.

Table 1. Scene classification results for the UC Merced dataset.

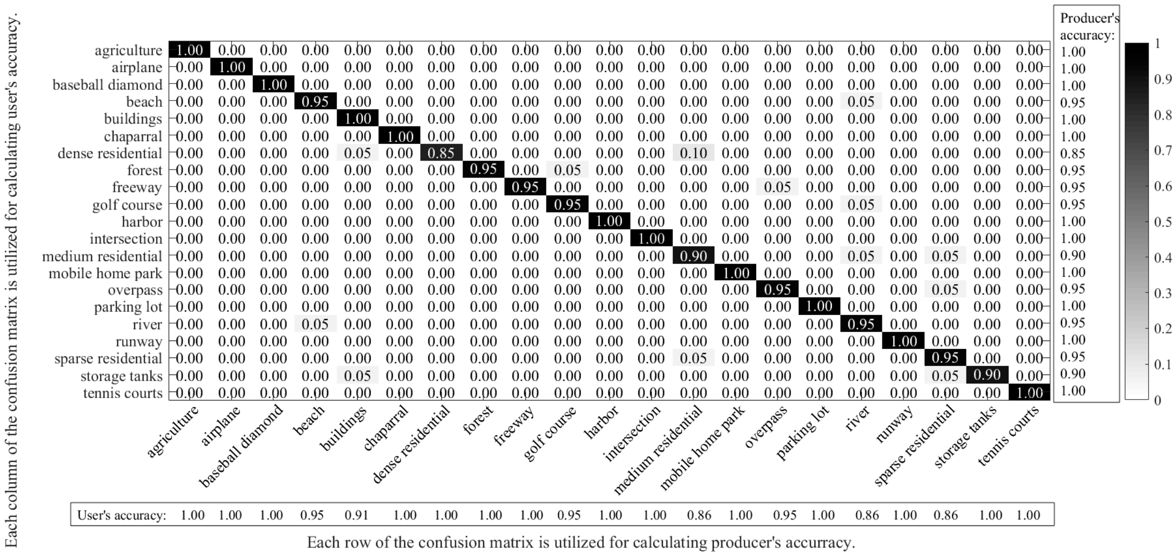
Scene Classification Method	Classification Accuracy (%)
BoW	72.05 ± 1.41
SPM	82.30 ± 1.48
LDA	81.92 ± 1.12
UFL+Saliency [34]	82.72 ± 1.18
GBRCN [35]	94.53
TF-CNN [36]	89.90

Table 1. *Cont.*

Scene Classification Method	Classification Accuracy (%)
Pre-trained AlexNet 1st-FC [32]	95.08
Multiview deep learning [31]	93.48 ± 0.82
CNN with OverFeat [40]	95.48
AlexNet	90.21 ± 1.17
Pre-trained-AlexNet	95.00 ± 0.72
Pre-trained-AlexNet-SPP	95.95 ± 1.01
Pre-trained-AlexNet-SS	95.71 ± 1.21
Pre-trained-AlexNet-SPP-SS	96.67 ± 0.94

From Table 1, it can be seen that the pre-trained-AlexNet-SPP and the pre-trained-AlexNet-SS obtain better scene classification performances than the pre-trained-AlexNet architecture, which proves that the incorporation of either SPP or SS can improve the scene classification performance. The pre-trained-AlexNet-SPP-SS achieves the best scene classification result of 96.67 ± 0.94%. Compared with the traditional scene classification methods, the pre-trained AlexNet-SPP architecture and the pre-trained AlexNet-SPP-SS architecture obtain accuracy of 95.95 ± 1.01% and 95.71 ± 1.21%. This proves that the incorporation of the combination of SPP and SS further improves the pre-trained-AlexNet classification performance in the HSR remote sensing imagery.

In order to better demonstrate the performance of the proposed pre-trained AlexNet-SPP-SS model for the UC Merced dataset, a confusion matrix is shown in Figure 8.



	agriculture	airplane	baseball diamond	beach	buildings	chaparral	dense residential	forest	freeway	golf course	harbor	intersection	medium residential	mobile home park	overpass	parking lot	river	runway	sparse residential	storage tanks	tennis courts	Producer's accuracy:
agriculture	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
airplane	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
baseball diamond	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
beach	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.95
buildings	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
chaparral	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
dense residential	0.00	0.00	0.00	0.00	0.05	0.00	0.85	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85
forest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95
freeway	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.95
golf course	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.95
harbor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
intersection	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
medium residential	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.00	0.90
mobile home park	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
overpass	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.05	0.00	0.00	0.95
parking lot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
river	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.95
runway	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
sparse residential	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.95
storage tanks	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.90	0.00	0.90
tennis courts	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
User's accuracy:	1.00	1.00	1.00	0.95	0.91	1.00	1.00	1.00	1.00	0.95	1.00	1.00	0.86	1.00	0.95	1.00	0.86	1.00	0.86	1.00	1.00	

Figure 8. Confusion matrix for the pre-trained AlexNet-SPP-SS model with the UC Merced dataset.

Figure 8 demonstrates the confusion matrix of the pre-trained AlexNet-SPP-SS model, where the accuracy of the row represents the producer's accuracy and the column represents the user's accuracy. From Figure 8, it can be seen that most of the classes obtain a satisfactory classification result over 90%, but the dense residential class shows a severe misclassification. By analyzing the confusion matrix of the pre-trained AlexNet-SPP-SS model, it can be seen that the samples of the dense residential classes are mainly misclassified as the building and medium residential classes. For the UC Merced dataset, the pre-trained AlexNet-SPP-SS model easily misclassifies the dense residential, building, and medium residential classes, as a result of their similar ground object distributions.

To further demonstrate the performances of the proposed pre-trained AlexNet-SPP-SS model on the UC Merced dataset, the scene classification accuracies for each thematic category are compared with the AlexNet, the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS in Figure 9.

Figure 8 demonstrates the confusion matrix of the pre-trained AlexNet-SPP-SS model, where the accuracy of the row represents the producer's accuracy and the column represents the user's accuracy. From Figure 8, it can be seen that the pre-trained AlexNet-SPP-SS model obtains a better classification accuracy than the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS in an overall view. However, the classes of dense residential and sparse residential obtain a worse classification accuracy as they contain confusing scene images that are difficult to classify. Taking a more in-depth and detailed analysis of the classification accuracy for certain classes, for example, all the classes except for the forest and tennis court classes, it can be seen that a better classification performances is obtained when adopting the SPP strategy. This is mainly due to the performance promotion of the SPP strategy considering the multi-scale information of the HSR remote sensing scene images with key ground objects. However, the classes of forest and tennis court show less improvement on the pre-trained AlexNet and the pre-trained AlexNet-SPP-SS models, because the scene images possesses heterogenous ground object distributions covering the main parts of the images. From Figure 9, it can be seen that the pre-trained AlexNet-SS performs better than the pre-trained AlexNet for most classes, except for the dense residential, forest, freeway, and sparse residential classes.

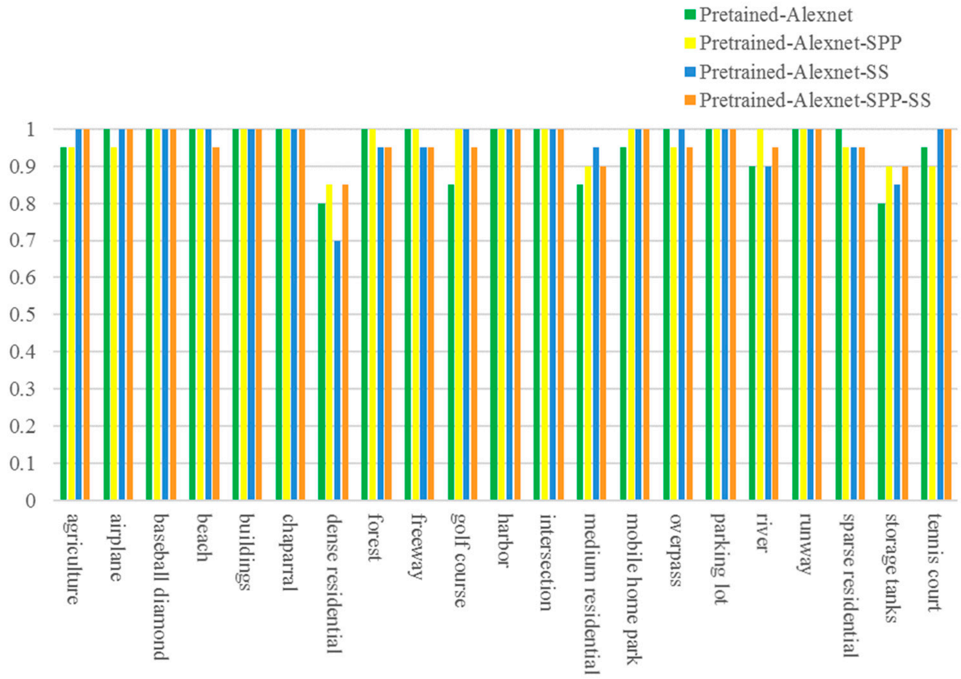


Figure 9. Classification accuracies of each category for the different algorithms with the UC Merced dataset.

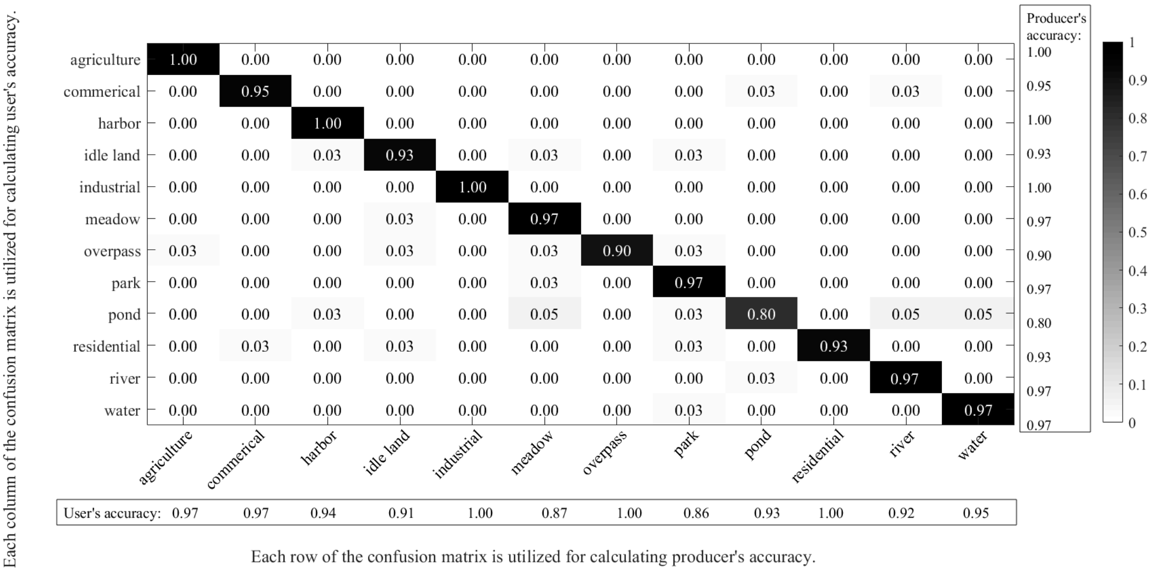
From Table 2, for the Google image dataset of SIRI-WHU, it can be seen that the pre-trained-AlexNet-SPP-SS model achieves the best scene classification result of 95.07 ± 1.09%. The reason why the pre-trained-AlexNet-SPP-SS method obtains a better scene classification result can be attributed to the multi-scale spatial information consideration and the side-supervision incorporation in the relatively simple pre-trained AlexNet architecture.

Table 2. Scene classification results for the Google image dataset of SIRI-WHU.

Scene Classification Method	Classification Accuracy (%)
BoW	73.93 ± 1.41
SPM	80.26 ± 1.86
LDA	66.85 ± 2.12
TF-CNN [36]	82.81
AlexNet	90.42 ± 1.11
Pre-trained-AlexNet	93.64 ± 0.98
Pre-trained-AlexNet-SPP	94.21 ± 1.18
Pre-trained-AlexNet-SS	94.58 ± 0.98
Pre-trained-AlexNet-SPP-SS	95.07 ± 1.09

In order to better demonstrate the specific performances of the proposed pre-trained AlexNet-SPP-SS model for the Google Image dataset of SIRI-WHU, a confusion matrix is shown in Figure 9.

From Figure 10, it can be seen that the classes of agriculture, harbor, and industrial obtain satisfactory classification results, but the pond class shows a severe misclassification. By analyzing the confusion matrix of the pre-trained AlexNet-SPP-SS model, it can be seen that the pre-trained AlexNet-SPP-SS model easily misclassifies the pond, meadow, idle land, and agriculture classes, as a result of their similar ground object distributions.



	agriculture	commerical	harbor	idle land	industrial	meadow	overpass	park	pond	residential	river	water	Producer's accuracy:
agriculture	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
commerical	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.95
harbor	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
idle land	0.00	0.00	0.03	0.93	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.93
industrial	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
meadow	0.00	0.00	0.00	0.03	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.97
overpass	0.03	0.00	0.00	0.03	0.00	0.03	0.90	0.03	0.00	0.00	0.00	0.00	0.90
park	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.97	0.00	0.00	0.00	0.00	0.97
pond	0.00	0.00	0.03	0.00	0.00	0.05	0.00	0.03	0.80	0.00	0.05	0.05	0.80
residential	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.03	0.00	0.93	0.00	0.00	0.93
river	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.97	0.00	0.97
water	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.97	0.97
User's accuracy:	0.97	0.97	0.94	0.91	1.00	0.87	1.00	0.86	0.93	1.00	0.92	0.95	

Figure 10. Confusion matrix for the pre-trained AlexNet-SPP-SS model with the Google image dataset of SIRI-WHU.

In Figure 11, to further demonstrate the performances of the proposed pre-trained AlexNet-SPP-SS model on the Google image dataset of SIRI-WHU, the scene classification accuracies for each thematic category are compared with the results of AlexNet, the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS.

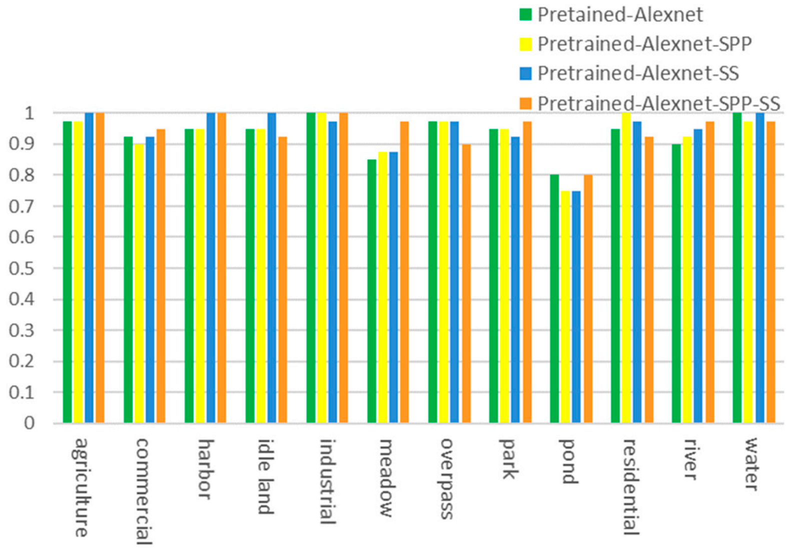


Figure 11. Classification accuracies of each category for different algorithms with the Google image dataset of SIRI-WHU.

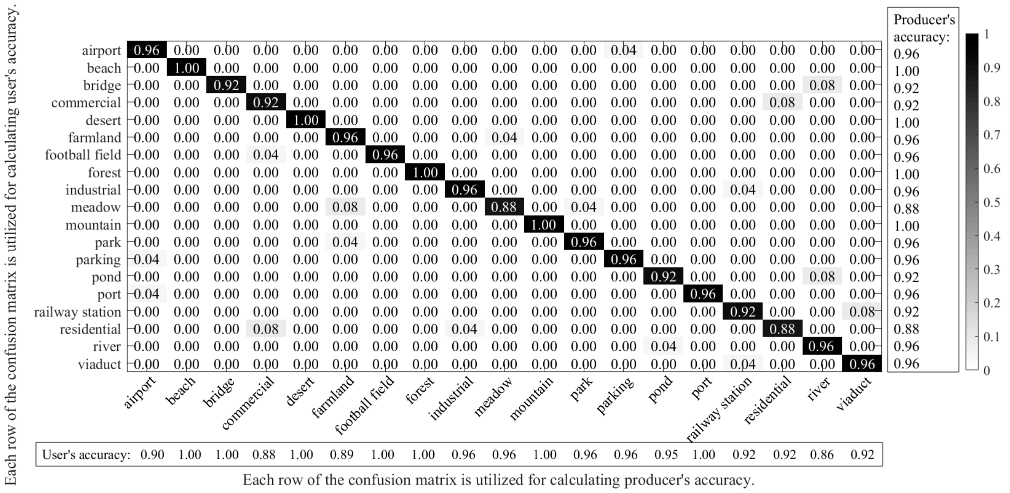
From Figure 11, it can be seen that the pre-trained AlexNet-SPP-SS model obtains a better classification accuracy than the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS, in an overall view. However, the classes of meadow and pond obtain a worse classification accuracy as they contain confusing scene images that are similar to agriculture and river, respectively. Taking a more in-depth and detailed analysis of the classification accuracy for certain classes, for example, all the classes except for the harbor and park classes, it can be seen that a better classification performances when adopting the SPP strategy. This is mainly due to the performance promotion of the SPP strategy considering multi-scale information of the HSR remote sensing scene images with key ground objects. However, for the classes of overpass and water, SPP shows less improvement in the pre-trained AlexNet and the pre-trained AlexNet-SPP-SS, because the scene images possesses heterogeneous ground object distributions covering the main part of the images. For the Google image dataset of SIRI-WHU, the pre-trained AlexNet-SS performs better than the pre-trained AlexNet for most of the classes, except for the pond and industrial classes.

From Table 3, for the WHU-RS dataset, it can be seen that the pre-trained-AlexNet-SPP-SS model achieves the best scene classification result of 95.00 ± 1.12%. The pre-trained AlexNet-SPP and pre-trained AlexNet-SS architectures also obtain superior scene classification performances, which can be attributed to the multi-scale spatial information consideration and the side-supervision incorporation in the relatively simple pre-trained AlexNet architecture.

Table 3. Scene classification results for the WHU-RS dataset.

Scene Classification Method	Classification Accuracy (%)
BoW	69.06 ± 2.26
SPM	85.67 ± 2.13
LDA	75.46 ± 2.50
AlexNet	86.32 ± 1.86
Pre-trained-AlexNet	94.32 ± 1.54
Pre-trained-AlexNet-SPP	94.73 ± 1.09
Pre-trained-AlexNet-SS	94.28 ± 2.10
Pre-trained-AlexNet-SPP-SS	95.00 ± 1.12

In order to better demonstrate the specific performances of the proposed pre-trained AlexNet-SPP-SS model for the WHU-RS dataset, a confusion matrix is shown in Figure 12.



	airport	beach	bridge	commercial	desert	farmland	football field	forest	industrial	meadow	mountain	park	parking	pond	port	railway station	residential	river	viaduct	Producer's accuracy:
airport	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.96
beach	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
bridge	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.92
commercial	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.92
desert	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
farmland	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96
football field	0.00	0.00	0.00	0.04	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96
forest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
industrial	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.96
meadow	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.88	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88
mountain	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
park	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96
parking	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.96
pond	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.08	0.00	0.92
port	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.96
railway station	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.08	0.92
residential	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.88
river	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.96	0.00	0.96
viaduct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.96	0.96
User's accuracy:	0.90	1.00	1.00	0.88	1.00	0.89	1.00	1.00	0.96	0.96	1.00	0.96	0.96	0.95	1.00	0.92	0.92	0.86	0.92	

Figure 12. Confusion matrix for the pre-trained AlexNet-SPP-SS model with the WHU-RS dataset.

From Figure 12, it can be seen that the classes of beach, desert, forest, and mountain obtain satisfactory classification results, but the classes of meadow and residential show severe misclassifications. By analyzing the confusion matrix of the pre-trained AlexNet-SPP-SS model, it can be seen that the pre-trained AlexNet-SPP-SS model easily misclassifies the farmland, meadow, commercial, residential, railway station, and viaduct classes, as a result of their similar ground object distributions.

In Figure 13, to further demonstrate the performances of the proposed pre-trained AlexNet-SPP-SS model on the WHU-RS dataset, the scene classification accuracies for each thematic category are compared with the results of AlexNet, the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS.

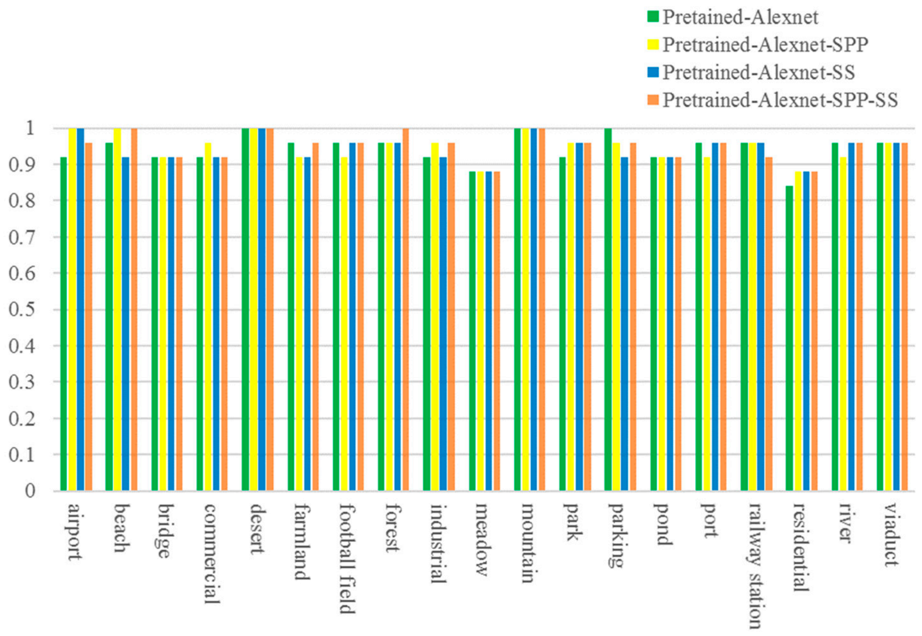


Figure 13. Classification accuracies of each category for different algorithms with the WHU-RS dataset.

From Figure 13, it can be seen that the pre-trained AlexNet-SPP-SS model obtains a better classification accuracy than the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS, in an overall view. However, the classes of bridge, commercial, meadow, pond, and, residential obtain a worse classification accuracy as they contain confusing scene images that are similar to agriculture, residential, commercial, and river, respectively. Taking a more in-depth and detailed analysis of the classification accuracy for certain classes, for example, all the classes except for the farmland, football field, and river classes, it can be seen that a better classification performances when adopting the SPP strategy. This is mainly due to the performance promotion of the SPP strategy considering multi-scale information of the HSR remote sensing scene images with key ground objects. However, for the classes of bridge, forest, pond, railway station, and viaduct, SPP shows less improvement in the pre-trained AlexNet and the pre-trained AlexNet-SPP-SS, because the scene images possesses heterogenous ground object distributions covering the main part of the images. For the WHU-RS dataset, the pre-trained AlexNet-SS performs better than the pre-trained AlexNet for most of the classes, except for the beach, farmland, and parking classes.

6. Discussion

From the above, it is known that the SPP strategy can improve the performance of the pre-trained AlexNet-SPP-SS model. To study the effect of the SPP layer number of the proposed pre-trained AlexNet-SPP-SS model for the UC Merced dataset, the Google image dataset of SIRI-WHU dataset, and the WHU-RS dataset, the other parameters generated by the pre-trained AlexNet and SS strategy were kept the same. The number of SPP layers was then varied over the range of [1–4] for the proposed pre-trained AlexNet-SPP-SS model.

From Figure 14, it can be seen that when the spatial pyramid layer number is equal to 4, the pre-trained AlexNet-SPP-SS model obtains the best classification performance with the UC Merced dataset. In addition, this experiment also indicates that the pre-trained AlexNet-SPP-SS model can better deal with the multi-scale convolutional feature information, as a result of the information fusion ability of the SPP strategy.

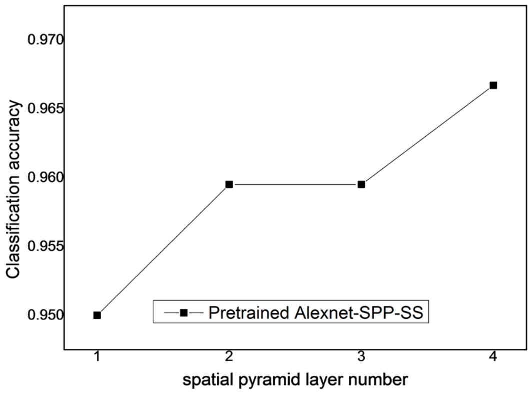


Figure 14. The influence of the spatial pyramid layer number for the pre-trained AlexNet-SPP-SS model with the UC Merced dataset.

From Figure 15, it can be seen that when the spatial pyramid layer number is equal to 3 or 4, the pre-trained AlexNet-SPP-SS model obtains the best classification performance for the Google image dataset of SIRI-WHU. Furthermore, the experimental results of the pre-trained AlexNet-SPP-SS model demonstrate that the SPP strategy has the ability to fuse the information of the multi-scale convolved feature maps and promotes the classification performance.

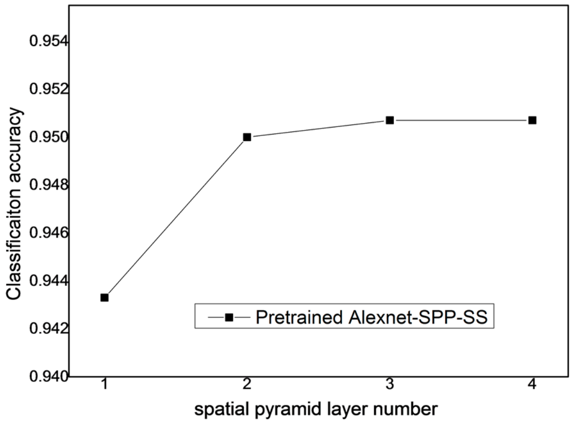


Figure 15. The influence of the spatial pyramid number for the pre-trained AlexNet-SPP-SS model with the Google Image dataset of SIRI-WHU.

From Figure 16, it can be seen that when the spatial pyramid layer number is equal to 4, the pre-trained AlexNet-SPP-SS model obtains the best classification performance for the WHU-RS dataset. Furthermore, the experimental results of the pre-trained AlexNet-SPP-SS model demonstrate that the SPP strategy has the ability to fuse the information of the multi-scale convolved feature maps and promotes the classification performance.

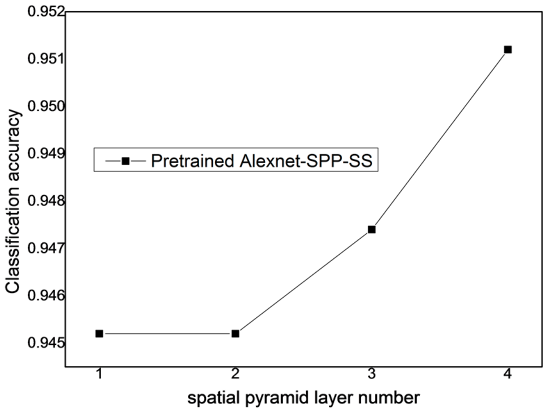


Figure 16. The influence of the spatial pyramid number for the pre-trained AlexNet-SPP-SS model with the WHU-RS dataset.

Although the performance of the pre-trained AlexNet-SPP-SS model was analyzed with the regard to the spatial pyramid layer number, further research into the classification performance of the pre-trained AlexNet-SPP-SS model with different training sample ratios is needed. For a further comparison of the proposed pre-trained AlexNet architecture with the other AlexNet architecture related models, the classification performances with the varying numbers of the training samples are reported in Figures 17–19, for the UC Merced dataset, the Google Image dataset of SIRI-WHU, and WHU-RS dataset, respectively.

From Figure 17, it can be seen that the pre-trained AlexNet-SPP-SS model performs better over the training sample ratios of [10, 20, 30, 40, 50, 60, 70, 80] than the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS for the UC Merced dataset. This figure also demonstrates that, in most of the training sample ratios, the pre-trained AlexNet-SPP-SS model performs better than the other models. In addition, the pre-trained AlexNet-SS model performs slightly better than the pre-trained AlexNet model in most of the training sample ratios.

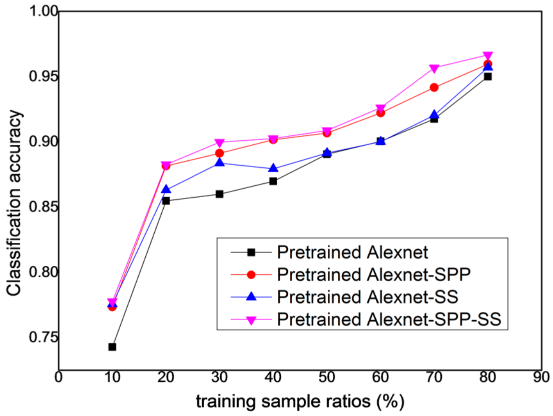


Figure 17. The influence of the training sample ratios with the different algorithms for the UC Merced dataset.

From Figure 18, it can be seen that the pre-trained AlexNet-SPP-SS model performs better over the training sample ratios of [10, 20, 30, 40, 50, 60] than the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS models for the Google image dataset of SIRI-WHU. In this figure, for the Google Image dataset of SIRI-WHU, the performance of the pre-trained AlexNet-SS model is slightly better than the pre-trained AlexNet-SPP model, which can be attributed to the introduction of the side supervision.

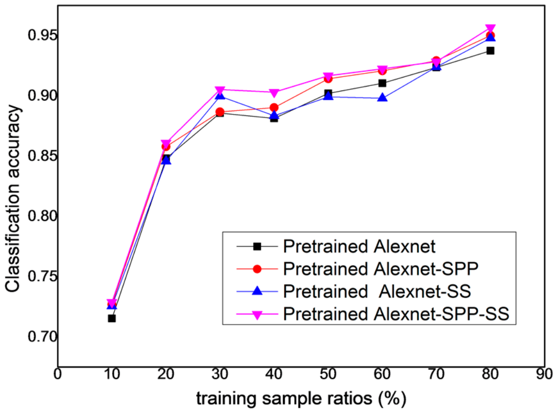


Figure 18. The influence of the training sample ratios with the different algorithms for the Google image dataset of SIRI-WHU.

From Figure 19, it can be seen that the pre-trained AlexNet-SPP-SS model performs better over the training sample ratios of [10, 20, 30, 40, 50] than the pre-trained AlexNet, the pre-trained AlexNet-SPP, and the pre-trained AlexNet-SS models for the Google image dataset of SIRI-WHU. In this figure, for the WHU-RS dataset, the performance of the pre-trained AlexNet-SPP model is slightly better than the pre-trained AlexNet-SS model. When the training sample ratio is small, the pre-trained AlexNet-SPP model and the pre-trained AlexNet-SS model perform much better than the pre-trained AlexNet model.

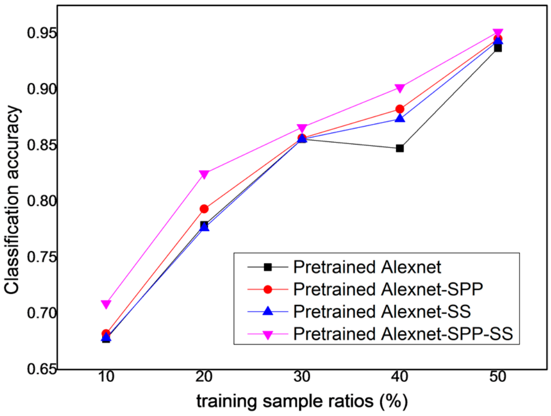


Figure 19. The influence of the training sample ratios with the different algorithms for the Google image dataset of SIRI-WHU.

7. Conclusions

In this paper, an improved pre-trained CNN architecture named the pre-trained AlexNet-SPP-SS model has been proposed for HSR remote sensing imagery scene classification. By fully utilizing both SPP and SS to further improve the performance of the pre-trained AlexNet architecture, the pre-trained AlexNet-SPP-SS demonstrates robust feature description ability for HSR remote sensing imagery. The incorporation of SPP adequately takes the multi-scale spatial information into consideration and helps to maintain theI ha fixed-length representation of the multi-scale convolved information. The incorporation of SS strategy can, to some extent, alleviate the over-fitting problem for HSR remote sensing imagery scene classification. Through the experiments, it was found that the proposed pre-trained AlexNet-SPP-SS model outperforms the current pre-trained AlexNet models, and the handcrafted feature based HSR remote sensing scene classification models on the UC Merced dataset, the Google image dataset of SIRI-WHU, and the WHU-RS dataset. In our future work, the multi-scale SPP strategy will be explored in more CNN architectures, and more automatic and adaptive multi-scale spatial pyramid information will be considered. In the future, an interesting phenomenon that the scene classification performances of images containing different scales of objects not present in the training/testing images will be studied.

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

Author Contributions: All the authors made significant contributions to the work. Xiaobing Han and Yanfei Zhong designed the research and analyzed the results. Liqin Cao and Liangpei Zhang provided advice for the preparation and revision of the paper.

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

References

1. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and efficient midlevel visual elements oriented land-use classification using VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
2. Zhao, J.; Zhong, Y.; Shu, H.; Zhang, L. High-Resolution Image Classification Integrating Spectral-Spatial-Location Cues by Conditional Random Fields. *IEEE Trans. Image Process.* **2016**, *25*, 4033–4045. [CrossRef] [PubMed]
3. Zhong, Y.; Wang, X.; Zhao, L.; Feng, R.; Zhang, L.; Xu, Y. Blind spectral unmixing based on sparse component analysis for hyperspectral remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 49–63. [CrossRef]

4. 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]
5. Cheriyadat, A. Unsupervised Feature Learning for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
6. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [CrossRef]
7. Zhang, X.; Du, S. A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sens. Environ.* **2015**, *169*, 37–49. [CrossRef]
8. Zhang, X.; Du, S. Semantic classification of heterogeneous urban scenes using intra-scene feature similarity and inter-scene semantic dependency. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2005–2014. [CrossRef]
9. Hu, F.; Xia, G.; Wang, Z.; Zhang, L.; Huang, X.; 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]
10. Zhao, B.; Zhong, Y.; Zhang, L. A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *116*, 73–85. [CrossRef]
11. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic, 10–16 May 2004; pp. 1–2.
12. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
13. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Patter Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2169–2178.
14. 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**, *5*, 2108–2123. [CrossRef]
15. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
16. Zhao, B.; Zhong, Y.; Zhang, L. Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. *Remote Sens. Lett.* **2013**, *4*, 1204–1213. [CrossRef]
17. Zhong, Y.; Cui, M.; Zhu, Q.; Zhang, L. Scene Classification Based on Multi-Feature PLSA for High Spatial Resolution Remote Sensing Imagery. *J. Appl. Remote Sens.* **2015**, *9*, 095064-1–095064-14. [CrossRef]
18. Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* **2001**, *42*, 177–196. [CrossRef]
19. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
20. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst. (NIPS)* **2012**, *25*, 1097–1105. [CrossRef]
22. Szegedy, C.; Liu, W.; Jia, Y.Q.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–12 June 2015; pp. 1–9.
23. Simonyan, K.; Andrew, Z. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. Lecun, Y.; Bengio, Y.; Hinton, G.E. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
25. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
26. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 1–46. [CrossRef]

27. Larochelle, H.; Bengio, Y.; Louradour, J.; Lamblin, P. Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.* **2009**, *10*, 1–40.
28. Marco, C.; Giovanni, P.; Carlo, S.; Luisa, V. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *arXiv* **2015**, arXiv:1508.00092.
29. Penatti, O.A.; Nogueira, K.; dos Santos, J.A. Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 12 June 2015; pp. 44–51.
30. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U.; Member, S. Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geosci. Remote Sens.* **2015**, *13*, 105–109. [CrossRef]
31. Luus, F.P.S.; 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]
32. Hu, F.; Xia, G.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
33. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
34. Zhang, F.; Du, B.; Zhang, L. Saliency-Guided Unsupervised Feature Learning for Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2175–2184. [CrossRef]
35. Zhang, F.; Du, B.; Zhang, L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 1–10. [CrossRef]
36. 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*. [CrossRef]
37. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412. [CrossRef]
38. Han, X.; Zhong, Y.; 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]
39. Han, X.; Zhong, Y.; Zhang, L. Spatial-Spectral Unsupervised Convolutional Sparse Auto-Encoder Classifier for Hyperspectral Imagery. *Photogramm. Eng. Remote Sens.* **2017**, *83*, 195–206. [CrossRef]
40. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 105–109. [CrossRef]
41. Chen, S.; Tian, Y. Pyramid of Spatial Relations for Scene-Level Land Use Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1947–1957. [CrossRef]
42. He, J.; Chang, S.H.; Xie, L. Fast kernel learning for spatial pyramid matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
43. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1465–1472.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Learn.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
45. Lee, C.; Xie, S.; Gallagher, P; Zhang, Z.; Tu, Z. Deep-Supervised Nets. *arXiv* **2014**, arXiv:1409.5185.
46. UC Merced Dataset. Available online: http://vision.ucmerced.edu/datasets/landuse.html (accessed on 8 August 2017).
47. SIRI-WHU Dataset. Available online: http://rsidea.whu.edu.cn/e-code.html (accessed on 8 August 2017).
48. Jia, Y.Q.; 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 ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.