A survey of remote sensing image classification based on CNNs

Jia Song\textsuperscript{a,c}, Shaohua Gao\textsuperscript{b}, Yunqiang Zhu\textsuperscript{a,c} and Chenyan Ma\textsuperscript{b}

\textsuperscript{a}State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; \textsuperscript{b}School of Resource and Environmental Sciences, Wuhan University, Wuhan, China; \textsuperscript{c}Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China

ABSTRACT
With the development of earth observation technologies, the acquired remote sensing images are increasing dramatically, and a new era of big data in remote sensing is coming. How to effectively mine these massive volumes of remote sensing data are new challenges. Deep learning provides a new approach for analyzing these remote sensing data. As one of the deep learning models, convolutional neural networks (CNNs) can directly extract features from massive amounts of imagery data and is good at exploiting semantic features of imagery data. CNNs have achieved remarkable success in computer vision. In recent years, quite a few researchers have studied remote sensing image classification using CNNs, and CNNs can be applied to realize rapid, economical and accurate analysis and feature extraction from remote sensing data. This paper aims to provide a survey of the current state-of-the-art application of CNN-based deep learning in remote sensing image classification. We first briefly introduce the principles and characteristics of CNNs. We then survey developments and structural improvements on CNN models that make CNNs more suitable for remote sensing image classification, available datasets for remote sensing image classification, and data augmentation techniques. Then, three typical CNN application cases in remote sensing image classification: scene classification, object detection and object segmentation are presented. We also discuss the problems and challenges of CNN-based remote sensing image classification, and propose corresponding measures and suggestions. We hope that the survey can facilitate the advancement of remote sensing image classification research and help remote-sensing scientists to tackle classification tasks with the state-of-art deep learning algorithms and techniques.

1. Introduction
With the development of earth observation technologies, an integrated space-air-ground global observation has been gradually established. The platform is consisted of satellite constellations, unmanned aerial vehicles (UAVs) and ground sensor networks,
with high-resolution satellite remote sensing systems as the principal parts. As a result, our comprehensive abilities to observe the earth have reached unprecedented levels (Li, Zhang, & Xia, 2014). The acquired remote sensing (RS) images are increasing dramatically in terms of volume, velocity, variety, and value. RS images have the massive characteristics of big data as well as other complex characteristics, such as high dimensionality (high spatial, temporal and spectral resolution), multiple scales (multiple spatial and temporal scales) and non-stationary (RS images acquired at different times vary in their characteristics) (Li, Li, Liu, Xie, & Yang, 2015; Li, Tong, Li, Gong, & Zhang, 2012; Song, Liu, Wang, & Lv, 2014; Zhang, 2017). How to effectively extract useful information from massive RS images remains an issue that needs to be addressed urgently.

The efficient extraction of information from RS data is commonly achieved with image classification techniques. RS image classification is a process in which each pixel or area of an image is classified according to certain rules into a category on the basis of its spectral characteristics, spatial structural characteristics, etc. (Zhao, 2013). RS image classification utilizes visual interpretation, which has relatively high accuracy but also a high manpower cost, and thus, it cannot meet the rapid processing demands for massive volumes of remote sensing data in the big data era. There are also couples of shallow machine learning methods for RS image classification. The methods transform the original data into many features (such as color space, texture and time series composites). However, with the increase of data volume and dimension, the implicit relationship between a large number of data bands is more and more abstract, which makes it difficult to design effective or optimized features for RS image classification. Thus, the classification accuracy is not satisfactory. In order to accommodate remote sensing big data, image classification techniques need to be quantitative, automated, and real-time (Li, 2008). Therefore, computer-based automatic classification techniques are gradually becoming the mainstream in tackling the RS image classification. Computer-based RS image classification methods can be classified into two types, namely parametric and non-parametric classifiers. Parametric classifiers include maximum likelihood and minimum distance algorithms. These classifiers assume that the data follow a normal distribution in the classification process. It is also very difficult to introduce other auxiliary data into these classifiers. The classification accuracy can hardly meet the requirements when facing massive volumes of RS data obtained from the complex earth surface (Jia, Li, Tian, & Wu, 2011). By contrast, non-parametric classifiers, including Support Vector Machine (SVM) (Bai, Zhang, & Zhou, 2014; Lv, Shu, Gong, Guo, & Qu, 2017; Tao, Tan, Cai, & Tian, 2011) and Neural Network (NN) algorithms (Luo, Zhou, & Yang, 2001; Tang, Deng, Huang, & Zhao, 2014), do not require the assumption that the data follow a normal distribution. The classification process involves two stages (Cheng & Han, 2016), namely feature extraction and classifier learning. At the first stage, the features are manually designed based on prior human knowledge. Common feature descriptors include the Local Binary Patterns (LBP) (Ojala, Pietikainen, & Maenpaa, 2002), the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) and the Histograms of Oriented Gradients (HOG) (Dalal & Triggs, 2005). At the second stage, a classifier is trained based on the designed features. Compared to parametric classifiers, non-parametric classifiers have higher classification accuracy for images obtained from complex earth surface. However, non-parametric classifiers also have some shortcomings when used to classify remote sensing big data. First, the manual design of features
relies primarily on prior knowledge and the designed features are often in the shallow layer (e.g., the edges or local textures of a ground object), and they cannot describe the complex changes of the objects in the image. Second, machine learning models (e.g., SVM), which are used in classification, are shallow structure models (Cortes & Vapnik, 1995) and have weak modeling capacity, and they are often unable to sufficiently learn the highly nonlinear relationships.

The emergence of deep learning (Hinton & Salakhutdinov, 2006) provides a new approach to solving these problems. Deep learning has been employed to learning models with multiple hidden layers and to design effective parallel learning algorithms (Chang et al., 2016). Deep learning models have more powerful abilities to express and process data and have shown excellent accuracy and precision rates in applications. In 2012, AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), a deep learning model of convolutional neural network (CNN), achieved remarkable accuracy in the computer vision field and won the ImageNet Challenge, a top-level competition in the image classification field. This CNN model is developed from ordinary neural networks, and directly extracts features from massive amounts of imagery data and abstracts the features layer by layer. It learns the boundary and color features of the objects in an image in the relatively shallow layers. As the number of network layers increases, the information in the neurons of the network is continuously combined. Eventually, the network extracts deep concepts and expresses abstract semantic features. AlexNet reduced the error rate for image classification from 25.8% to 16.4%. After that, networks that competed in the ImageNet Challenge continuously reduced the error rate. In 2015, ResNet (He, Zhang, Ren, & Sun, 2015) even reduced the error rate to 3.6%, whereas the error rate of the human eye was approximately 5.1% in the same experiment. Evidently, computer accuracy in image classification has far surpassed that of humans. Apart from image classification (Lin, Chen, & Yan, 2013), CNNs have also achieved satisfactory accuracy in object detection (Simonyan & Zisserman, 2014) and image segmentation (Tai, Xiao, Zhang, Wang, & Weinan, 2015).

Remote sensing data are essentially digit images, but they record richer and more complex characteristics of the earth surface. Parallel to the enormous success of CNNs in computer vision, geoscientists have discovered that CNNs can be applied in the remote sensing field for rapid, economical, and accurate feature extraction. Some articles have reviewed the current state-of-art of deep learning for remote sensing (Zhu, Tuia, & Mou et al., 2017; Zhang, Zhang, & Du, 2016; Ball, Anderson, & Chan, 2017). However, they tend to cover quite broad issues or topics in remote sensing, and is limited in RS image classification which plays a key role in earth science, such as land cover classification, scene interpretation, monitoring of the earth’s surface, etc. Quite a few researchers have studied RS image classification based on CNN models in recent years. Systematically analyzing and summarizing these studies are desirable, and is significant for advancing deep learning in remote sensing. Thus, this article focuses on surveying CNN-based RS image classification. We hope that our work is helpful for remote-sensing scientists to get involved in CNN-based RS image classification. In the following sections, the principles of CNNs are introduced. Then, based on an extensive literature survey, studies of CNN model improvements and CNN training data for RS image classification are systematically analyzed, and CNN application cases in scene classification, object detection and object segmentation are presented and summarized. Finally, the problems and the
challenges of RS image classification based on CNN are elaborated, and inspiration for addressing the challenges is drawn.

2. Convolutional neural network (CNN)

CNNs, as one type of deep leaning networks, have the following advantages over shallow structure models: (1) CNNs directly apply a convolution operation to the pixels of an image to extract abstract data features. This feature extraction can be applied to various scenarios and has a more powerful generalization ability. (2) CNNs are able to represent image information in a distributed manner and rapidly acquire image information from massive volumes of data. The structure of CNNs can effectively solve complex nonlinear problems (e.g., the rotation and translation of an image). (3) CNNs are characterized by sparse connections, weight-sharing and spatial subsampling, which result in a simpler network structure that is more adaptable to image structures. In order to better understand CNN-based image classification, this section will briefly introduce the structure of CNNs and its training method, followed by several popular CNN models in the computer vision field.

2.1. Structure of CNN

CNNs are multilayer perceptrons that are specially designed to identify two-dimensional (2D) shapes and can be used to establish mapping from the original input to the desired output. In a CNN, each neuron is connected to the neuron in a local area of the previous layer of the network, thereby reducing the number of weights in the network. Similar to ordinary neural networks, a hierarchical connection structure is also used in CNNs. In other words, a CNN consists of components stacked layer by layer. They are convolutional, pooling and fully connected layers and an output layer, as shown in Figure 1. In a typical CNN, the convolutional and pooling layers alternate as the first few layers, followed by the fully connected layer. The final output layer generates classification results.

![Figure 1. Structure of CNN.](image-url)
2.1.1. **Convolutional layer**
A convolutional layer is the basic layer of a CNN. The convolution operation works on a small local area of the image with a certain size of a convolutional kernel. The convolutional kernel is a learnable weight matrix. The output of the convolutional layer goes through an activation function, then a convolved feature map is obtained. The feature map can be the input of the subsequent convolutional layer. Therefore, more sophisticated features can be extracted after several convolutional layers being stacked layer by layer. Moreover, the neurons in each feature map share the weight of a convolutional kernel in a convolution operation, which ensures that the number of parameters in the network will not increase significantly, even if the number of convolutional layers continues to increase, thereby reducing the storage requirement for the model. Consequently, this model can facilitate the establishment of a deeper network structure.

2.1.2. **Pooling layer**
A pooling layer generally comes after a convolutional layer. General pooling layers include the maximum, average and random pooling. The maximum pooling and the average pooling find the maximum and average values of the neighborhood neurons respectively, and the random pooling selects values for the neurons according to a certain probability. There are other forms of pooling layer, which are often improved on the general pooling layers, including overlapping pooling and spatial pyramid pooling. Regardless of which form of pooling layer is used, pooling layers aims to capture features but is insensitive to their precise locations, which ensures that the network can still learn effective features, even if a small amount of input data shift occurs. In addition, a pooling layer does not alter the number of feature maps of the previous layer, but it reduces the spatial dimensionality of feature maps and preserves the important information in the feature maps, thereby further reducing the computation of network training.

2.1.3. **Fully connected layer and output layer**
A fully connected layer is composed of multiple hidden layers. Each hidden layer contains multiple neurons, and each neuron is fully interconnected with the neurons of the subsequent layer. One-dimensional (1D) feature vectors obtained by flattening feature maps after operations in the convolutional and pooling layers are used as the input for a fully connected layer. The objective of a fully connected layer is to map these features into a linearly separable space and coordinate with the output layer in classification. The output layer primarily uses a classification function to output the classification results. Currently, the Softmax function (Liu, Wen, Yu, & Yang, 2016) and SVMs are the common classification functions used in CNNs.

2.1.4. **Activation and loss functions**
In addition, activation and loss functions are also essential modules in a CNN. The activation function is generally nonlinear, which enables the network to be capable of learning on layer-wise nonlinear mapping. Common activation functions include Sigmoid, Rectified Linear Unit (ReLU) (Hara, Saito, & Shouno, 2015) and Maxout functions (Goodfellow, Warde-Farley, Mirza, Courville, & Bengio, 2013). A loss function, which is also referred to as a cost function or an objective function, is used to represent the
extent of previous inconsistencies between the value predicted by the model and the actual value. Furthermore, extra units, such as L1 regularization and L2 regularization, can be added to the loss function to prevent model overfitting. The L1 regularization and L2 regularization can be treated as restrictions in some parameters of the loss function.

### 2.2. Training of CNN

A CNN is trained primarily through backpropagation (Hecht-Nielsen, 1989). First, the input data are forward-calculated by a network structure composed of stacked convolutional layers, pooling layers, fully connected layer, an output layer and activation functions. The errors between the network output and the ground-truth value are calculated by a predefined loss function, then the errors are backpropagated based on the partial derivatives. Under the preset learning rate, each weight and the corresponding error term are adjusted. The above process is performed in an iterative way until the network model is converged. For a CNN, parameters that need to be obtained through training include the weights of the convolutional kernels, the weights of the fully connected layer, and the bias for each layer. Before a network is trained, all the parameters need to be initialized. Good network parameter initialization will make network training more efficient. Common methods of network parameter initialization include Gaussian distributions initialization, uniform distribution initialization and so on.

### 2.3. Typical CNN models

In the early image classification tasks, AlexNet, a CNN with five convolutional layers and two fully connected layers, is widely regarded as one of the most influential CNNs. AlexNet competed in the ImageNet Large Scale Visual Recognition Challenge in 2012 (He et al., 2015). The network achieved a top-5 error of 15.3%. AlexNet was the first time a model performed so well on a historically difficult ImageNet dataset. After AlexNet, Network-In-Network (NIN) (Lin et al., 2013), VGG-Net (Simonyan & Zisserman, 2014), GoogLeNet (Tai et al., 2015), ResNet (He et al., 2015) and DenseNet (Huang, Liu, Maaten, & Weinberger, 2016) have emerged successively. The NIN improves the network structure unit. In the NIN, a multilayer perceptron is added to each convolutional layer to replace the simple linear convolutional layer, thereby increasing the nonlinearity of the convolutional layers of the network. In addition, in the NIN, global average pooling (GAP) is also used to replace fully connected layers, thereby solving the overfitting problem that can be caused easily by an excess number of parameters of fully connected layers. Other CNN-based networks, including VGG-Net, GoogLeNet and ResNet, focus on increasing the model depth to improve the network structure. VGG-Net is a 19-layer network with even smaller (3 × 3) convolutional kernels. GoogLeNet is a 22-layer network with inception modules that can better use computational resources in the network and increase the network depth and width without increasing computation. The residual network used in ResNet further increases the network depth (>1,000 layers). The residual network primarily solves the problem in which deep networks cannot be trained. In DenseNet, a new dense cross-layer connection is used to improve the
network structure. This connection allows each layer of the network to be directly connected to all the previous layers, thereby allowing the features in the network layers to be used repeatedly. In addition, this connection also effectively solves the vanishing gradient problem that occurs during network training and enables faster model convergence.

After several years of rapid development, CNNs have achieved tremendous success in the computer vision field, and it has been being applied in RS image classification. In the following sections, we will survey and analyze the applications in details.

3. CNN model developments for RS image classification

Compared to ordinary images, RS images contain richer spectral information as well as spatial information reflecting their structure, shape and texture. Therefore, quite a few studies focus on CNN model improvements, so as to enable CNN models to better capture the features in RS images. In this section, we survey and analyze CNN model improvements in terms of each main CNN components, and parameter initialization and optimization for CNN training.

3.1. Input layers of CNNs

In the computer vision field, image information in the red, green and blue (RGB) bands is used as the input for a CNN. In remote sensing field, images often contain information in more bands, and they also contain rich multi-scale information and texture information. Moreover, multi-source images from different sensors are often utilized to better analyze geographic features of the earth. Therefore, taking full advantages of multi-source and multi-spectrum information is essential in CNN-base RS image classification. We have found that several studies focus on this point. For example, Zhang et al. (2018) merged two types of high-resolution RS images (GF-2 and World View satellite images) as the input for a CNN to extract roads, and achieved a high accuracy of 99.2%. Similarly, Xu, Mu, Zhao, and Ma (2016) used low- and high-frequency sub-bands that reflected multiscale image features, which was obtained by a contourlet transform on two training datasets (UC Merced land use dataset and the WHU-RS19 dataset), as the input for a CNN and also obtained a scene classification accuracy of above 90%. Furthermore, Xia, Cao, Wang, and Shang (2017) added four types of texture information, namely the mean value, standard deviation, consistency and entropy, to RS images in the RGB bands, and used the resulting data as the input for a CNN to extract roads, vehicles and vegetation based on the CNN and conditional random field methods. This approach greatly improved the classification accuracy. Also, Xu, Wu, Xie, and Chen (2018) used the normalized differential vegetation index (NDVI), the normalized digital surface model (NDSM) and the first component of the principal component analysis (PCA1), together with the original bands (R-G-B-IR), as the input for a CNN to extract buildings and vegetations. The above studies shown that more index, texture and spectrum information in multi-source data has been used in CNN-based RS image classification, and the result of these experiments proved that this approach could improve accuracy of extracting geographic objects in RS images.
3.2. Fully connected layers of CNNs

CNNs usually contain more parameters to be trained than shallow learning models. The more parameters often requires much more training data to ensure faster CNN model convergence. In computer vision, there are abundant and various datasets available for training, such as ImageNet, COCO, PASCAL VOC, etc., but few of them are suitable for training CNNs for RS image classification. Whereas, in the remote sensing field, there has been open datasets suitable for training CNNs for RS image classification, but they are very limited. Therefore, it is necessary to reduce the parameters when the training data are limited so that a CNN model converges smoothly. In a CNN, the parameters of the fully connected layers occupy the majority of CNN parameters. Thus, reducing the parameters of the fully connected layers of CNNs can make a CNN model converge on relatively limited training data. We selected two typical cases to demonstrate this approach. Li, Fu, et al. (2017) added a dropout and a Restricted Boltzmann Machine (RBM) to the fully connected layers, which reduced the number of parameters. They used UC Merced land use dataset to evaluate the performance of the method. Their method achieved the best performance compared to other methods, and reaches an overall accuracy of 95.11%. Zhong, Fei, and Zhang (2016) proposed a global average pooling layer (GAP), which is used to replace the fully connected network as the classifier. It greatly reduces the total parameters in the improved CNN model (large patch convolutional neural network, LPCNN), and make it easier to train the model with limited training data. LPCNN was evaluated on three different HSR remote-sensing datasets: IKONOS land-use dataset, UC Merced land-use dataset, and Google Earth land-use dataset of SIRIWHU. The improved CNN model achieved the best performance on the three datasets. The overall accuracies (OA) are 92.54%, 89.90% and 89.88% respectively.

3.3. Classifiers of CNNs

In the computer vision field, a Softmax or SVM classifier is commonly used in the output layer of a CNN for dividing the feature space and obtain classification results. Whereas, a Softmax classifier is sometimes inadequate for dividing the feature space in the remote sensing field since the training data in remote sensing could have relatively high noise levels or a relatively large difference between features. Therefore, there are a few studies concerning about more powerful classifiers for getting better classification results on RS images. For example, to address the problem of speckle noises in synthetic aperture radar (SAR) images used to detect ships, Bai, Jiang, Pei, Zhang, and Bai (2018) substituted a fuzzy SVM for a conventional classifier, thereby reducing the impact of the noise sample points on the division of the feature space. As a result, they achieved excellent detection accuracy, with a detection accuracy of 98.6% for the ships in the SAR images. In addition, Xu et al. (2016) established a multi-kernel SVM classifier using a linear combination of multiple kernel functions based on the UC Merced land use dataset and WHU-RS19 scene classification dataset. This classifier adaptively selects a kernel function for classification based on the difference in image features, and it consequently has a higher ability to divide complex feature spaces and thus achieves an improved classification accuracy.
3.4. Loss functions of CNNs

Unlike the common images produced using close-range photographic techniques, RS images are generally produced using aerial photographic techniques from above. As a result, the same geographic object is multidirectional in a RS image. This multidirectional leads to unsatisfactory classification accuracy with commonly used loss functions proposed in the computer vision field. In addition, common loss functions are good at separating different classes, but hard to differentiate features of individuals in the same class; there are a few studies focusing on improving the loss functions. For example, Cheng, Zhou, and Han (2016) added an L2 regularization and a regularization constraint term that restricts the rotation variation in objects to the original loss function. They evaluated the performance of this method with NWPU VHR-10 object detection dataset, which includes aircraft, ships, bridges and so on, and achieved a relatively high detection accuracy. Li, Qu, and Peng (2018) designed a loss function that the intraclass compactness and interclass separability are maximized simultaneously for ship detection in SAR images. They designed a CNN model of dense residual network based on ResNet50 (He et al., 2016), and used OpenSARShip (Huang, Liu, et al., 2017) dataset to evaluate ship classification accuracy. The accuracy averaged on classes is 77.2%, which higher than that of the original ResNet50.

3.5. Network structure of CNNs

The number of layers of CNNs has been increased for achieving higher classification accuracy. However, the increase in the number of network layers lead to the increase in the number of model parameters to be trained, which requires much more training data. Whereas, the training data are seriously insufficient in the remote sensing field. Consequently, several studies seek to improve network structure of CNNs. The main idea is to design several independent CNN models that differ in the depth of the convolutional layers or the number of neurons in the fully connected layers, and then combine them through feature fusion or model integration. This idea will not cause a significant increase in the number of network parameters, and the network model can converge easily with limited training data. For example, in order to extract the objects in a built-up area from SAR images, Li, Zhang, and Li (2016) extracted features at three different scales using three independent CNN models that differed in the depth of the convolutional layers. They then imported the extracted features of various scales into the fully connected layers to fuse them, and they finally classified the fused features using a Softmax classifier. This network structure is capable of learning rich features of buildings in a built-up area. These features contain relatively detailed and abstract information on the buildings, and they are helpful to effectively improve the learning ability of CNNs. Using the UC Merced land use scene classification dataset, Li, Fu, et al. (2017) first trained several independent CNN models that differed in the number of neurons in the fully connected layers and then integrated the classification results using the voting strategy during the test stage. On this basis, they obtained the final scene classification results. The experimental results showed that this structure could yield more accurate scene classification results.
3.6. Parameter initialization and optimization for CNN training

Assigning random values to the parameters is a simple parameter initialization method but has low training efficiency when applying in remote sensing images, and it may result in unstable training results. Therefore, parameter initialization is often concerned in the remote sensing field. The surveyed studies show that parameter values are generally obtained from pretrained CNN models in the computer vision field (e.g., AlexNet and VGG-Net). This pretrained network method can rapidly transfer and apply the learned parameter values from visual image features, thereby making the training for RS image classification more efficient and reducing the complexity and cost of training. According to the literature (Zhou, Shao, & Cheng, 2016; Castelluccio, Poggi, Sansone, & Verdoliva, 2015; Nogueira, Penatti, & Santos, 2017), there are currently two approaches to initialize parameters for CNN training in RS image classification. One of the approaches selects several layers of a pretrained network and fine-tunes them based on the remote sensing image dataset, so that the CNN adapts to achieve satisfactory RS image classification accuracy. For example, Zhang, Wang, Liu, Liu, & Wang (2016) compiled a remote sensing image sample set of five types of urban functional land, namely commercial land, residential land, factory land, land for educational purposes and public land, based on Google Earth remote sensing images. They then trained a prediction model by fine-tuning a pre-trained AlexNet CNN using this sample set. Subsequently, they used the prediction model to classify images into the five types of urban land in the cities of Shenyang and Beijing. The results demonstrated that the fine-tuned pre-trained AlexNet could effectively classify urban functional land. The other approach directly uses a pretrained networks as an extractor for extracting remote sensing image features, and the extracted features are then used to train a classifier. For example, Weng, Mao, Lin, and Guo (2017) used the last convolutional layer of a pretrained AlexNet network structure to extract remote sensing image features to train an extreme learning machine classifier. This classifier achieved a classification accuracy of 95.62% on the UC Merced land use dataset. Marmanis, Datcu, Esch, and Stilla (2016) converted 1D remote sensing image features extracted by the fully connected layers of a pre-trained network to a 2D feature matrix. This matrix was used to train a CNN model containing two convolutional layers, two fully connected layers and one Softmax classifier. They used the CNN model to classify the scenes in the UC Merced land use dataset and achieved an overall classification accuracy of 92.4%. Besides, Lu et al. (2015) used the network parameters obtained by training in a linear land elimination task to as the initial parameters to train a proposed CNN model, and then used the eigenvectors from the trained CNN model as the input for an SVM extracting farmland from UAV images.

Parameter optimization is also a key step during the training process of CNNs, which can be achieved with the help of a parameter optimizer. Training optimizers include stochastic gradient descent (SGD), RMSprop, Adam, Adadelta and Adagrad. In (Zhong et al., 2016), SGD is employed as a parameter optimizer, and a momentum technique is also used for SGD which can prevent the model from getting stuck in local minima and approach the global minima quickly.
4. CNN training data for RS image classification

Like other supervised learning algorithms, CNN-based deep learning needs to learn features from a large number of training data to achieve satisfactory model generalization. Training data from remote sensing images are less than data from natural images. As described in Sec. 3, some improvements of CNNs solve the problem of model overfitting which results from insufficient training data. In practice, there are also some studies on data augmentation or moderate use of weakly labeled training data. In this section, we review these studies. Those open datasets available for model training in remote sensing are summarized.

4.1. Open datasets

Although open datasets in remote sensing field are limited, they still play a very important role in training CNN models for RS image classification. Table 1 summarizes the open datasets used for training and validating CNN models for RS image classification. These datasets are categorized into three groups based on three kinds of classification tasks: scene classification, object detection and object segmentation. Table 1 shows that most datasets are from Google earth, thus the bands of most datasets provide only red, green and blue (R-G-B). Some datasets are from satellite or airborne sensors. They provide an extra near infrared band. Table 1 also shows that the datasets have high resolution since they are normally from high resolution sensors. Also, these datasets are different in the number and definition of categories. Most studies use a single dataset as training data, and few studies use multiple datasets for training.

4.2. Data augmentation techniques

As shown above, although there are several datasets available for training CNN-based RS images classification models, the category types and numbers of labels in the datasets are still extremely limited and often fail to meet the data scale requirement of model training. Acquiring samples with manual visual interpretation has very low efficiency and a relatively high cost. Therefore, some studies also focus on data augmentation techniques and the use of weakly labeled samples. For example, when using the RSOD-Dataset to detect oil barrels, Long et al. (2017) employed three operations, specifically, translation, scaling and rotation to augment the training data. The augmented training data were 60 times the original data in terms of volume. After the data augmentation, the detection accuracy for oil barrels reached as high as 96.7%. Their results demonstrate that properly increasing the sample size can effectively improve a CNN model’s performance. Lacking open datasets, Zhou, Shi, and Ding (2017) also augmented a small volume of manually labeled aircraft training data using three processing methods, mirroring, rotation and bit-plane decomposition. The bit-plane method merged eight bit-plane images obtained from the decomposition of each grayscale image at a new ratio. The size of the training set was increased 32-fold, and the test accuracy increased from 72.4% (based on the original training set) to 97.2% (based on the augmented training set). In addition, Zhong et al. (2016) proposed an augmentation method that was applicable to datasets for classifying scenes in RS images. Taking into consideration the random and multi-scale distribution characteristics of spatial objects in a scene, this
Table 1. Open datasets used for RS image classification.

| Dataset                  | Bands          | Pixel size (m) | Data source                                      | Image size (pixels) | Categories                                      | Num.  |
|--------------------------|----------------|----------------|--------------------------------------------------|---------------------|-------------------------------------------------|-------|
| Scene classification     |                |                |                                                  |                     |                                                 |       |
| UC-Merced Land Use dataset (Yang & Newsam, 2010) | R-G-B          | 0.3            | United States Geological Survey National Map    | 256*256             | 21 (forest, beach harbor etc.)                  | 2100  |
| Brazilian Cerrado-Savanna Scenes dataset (Nogueira, Santos, et al., 2017) | R-G-NIR        | 5              | the RapidEye satellite sensors                   | 64*64               | 4 (agriculture, arboreal vegetation etc.)       | 1311  |
| Brazilian Coffee Scenes dataset (Penatti, Nogueira, & Santos, 2015) | R-G-NIR        | 10             | SPOT 5 sensor                                    | 64*64               | 2 (coffee and non-coffee)                       | 2876  |
| WHU-RS19 dataset (Xia et al., 2010) | R-G-B          | 0.5            | Google Earth                                     | 600*600             | 19 (airport, pond etc.)                        | 1005  |
| RSSC7 dataset (Zou, Ni, Zhang, & Wang, 2015) | R-G-B          |                 | Google Earth                                     | 400*400             | 7 (forest, river etc.)                         | 2800  |
| AID dataset (Xia et al., 2016) | R-G-B          |                 | Google Earth                                     | 600*600             | 30 (forest, park etc.)                         | 10000 |
| SIRI-WHU (Zhao, Zhong, Xia, & Zhang, 2016) | R-G-B          |                 | Google Earth                                     | 200*200             | 12 (park, agriculture etc.)                    | 2400  |
| Object detection         |                |                |                                                  |                     |                                                 |       |
| NWPU VHR-10 dataset (Cheng, Han, Zhou, & Guo, 2014) | R-G-B          | 0.5–2          | Google Earth                                     | 10 (airplane, tank etc.) | 715                                             |       |
| SZTAKI-INRIA building dataset (Benedek, Descombes, & Zerubia, 2011) | R-G-NIR        | 0.08           | Vaihingen data                                   |                     | 85                                              |       |
| UCAS-AOD dataset (Zhu et al., 2015) | R-G-B          |                 | Google Earth, IKONOS, QuickBird                  | 2 (building, not building) | 9                                               |       |
| ISPRS 2D Semantic Labeling (11) | R-G-B          |                 | Google Earth                                     | 3 (airplane, car, background) | 910                                             |       |
| RSOD-Dataset (Long, Gong, Xiao, & Liu, 2017) | R-G-B          |                 | Google Earth                                     | 4 (aircraft, playground, overpass, oil-tank) | 2326                                           |       |
| Object segmentation      |                |                |                                                  |                     |                                                 |       |

Dataset websites: 1 http://weegee.vision.ucmerced.edu/datasets/landuse.html; 2 http://www.patreo.dcc.ufmg.br/2017/11/12/brazilian-coffee-scenes-dataset/; 3 http://captain.whu.edu.cn/repository.html; 4 http://weegee.vision.ucmerced.edu/datasets/landuse.html; 5 http://pan.baidu.com/s/1SlSn6V; 6 http://www.lmars.whu.edu.cn/xia/AID-project.html; 7 http://pan.baidu.com/s/1hqwx2XeG; 8 http://web.eee.sztaki.hu/remotesensing/building_benchmark.html; 9 http://www.ucassdl.cn/resource.asp; 10 https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset-; 11 http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html;
method increases the sample size by adjusting the sampling window size and sampling from a scene based on a sliding scheme. This method was evaluated with the IKONOS land use dataset, the UC Merced land use dataset and the SIRI-WHU dataset. It was found to have effectively improved the scene classification accuracy.

### 4.3. Moderate use of weakly labeled samples

Accurate sample labeling requires considerable amounts of labor and time. There are large amounts of application-related and weakly labeled datasets, e.g., coarse-grained and labeled RS image data in object detection tasks and non-accurately labeled map data for object extraction tasks (e.g., OpenStreetMap (OSM)). The moderate use of these weakly labeled samples meets the basic quality requirement of training data for most RS image classification, and it is an effective means to increase the number of training data. For example, aircraft detection needs to distinguish aircraft from complex and diverse backgrounds. Zhang, Du, Zhang, and Xu (2016) first obtained a CNN model by training on aircraft sample data with simple backgrounds based on this model, they obtained sample data that were misclassified as aircrafts from the UC Merced land use dataset, a weakly labeled dataset representing background information. They then treated the misclassified sample data as samples with complex backgrounds and added them to the training sample set. The resulting training sample set was more extensive and more representative. In addition, to address the issue of lacking accurate training data for a building extraction task, Maggiori, Tarabalka, Charpiat, and Alliez (2017) obtained a preliminary prediction model by pre-training on the OSM dataset, a weakly labeled dataset with errors, and then fine-tuned and corrected the prediction model based on a small number of accurately labeled building samples. With the corrected model, they eventually obtained extraction results with higher accuracy. This result demonstrates that the weakly labeled samples can effectively alleviate the problem of insufficient training data in some cases.

## 5. Application cases

Application cases of CNN-based RS image classification are classified into scene classification, object detection and object extraction. Scene classification is the process of determining the type of a remote sensing image based on its content. Object detection is the process of determining the locations and types of the targets to be detected in a remote sensing image and labeling their locations and types with bounding boxes. Object extraction is the process of determining the accurate boundaries of the objects to be extracted in a remote sensing image. In this section, we summarize these application cases.

### 5.1. Scene classification

Scene classification is a mapping process of learning and discovering the semantic content tags of image scenes (Bosch & Zisserman, 2008). Generally, an image scene is a collection of multiple independent geographic objects. These objects have different structures and contain different texture information, and they form different types of
scenes through different combinations and spatial locations. For scene classification studies in the remote sensing field, the UC Merced land use dataset is commonly viewed as the reference dataset. This dataset is used to validate the methods of scene classification. Methods of scene classification were summarized in Table 2. the LPCNN method is characterized with a specific data augmentation technique to enhance CNN training and global average pooling to reduce parameters, and the MS-DCNN method is characterized with multi-source and a multi-kernel SVM classifier. The next two methods use a pretrained CNN model to learn deep and robust features, but an extreme learning machine (ELM) classifier instead of the fully connected layers of CNN is used in the CNN-ELM method to obtain excellent results. The fifth method combines pretrained networks and RBM retrained network as a two stage-training network, which obtains good results. Latest studies on scene classification proposed methods called GCF-LOF CNN, deep-local-global feature fusion framework (DLGFF) and deep random-scale stretched convolutional neural network (SRSCNN). The GCF-LOF CNN is a novel CNN by integrating the global-context features (GCFs) and local-object-level features (LOFs). Similarly, the DLGFF establishes a framework integrating multi-level semantics from the global texture feature-based method, the (bag-of-visual-words) BoVW model, and a pre-trained CNN as well. The SRSCNN proposes random-scale stretching to force the CNN model to learn a feature representation that is robust to object scale variation. As shown in classification accuracy in Table 2, recent studies have achieved very high classification accuracy.

5.2. Object detection

Object detection from remote sensing images detects the locations and the types of objects. The object detection application cases from remote sensing images use the

| Method | Feature | Accuracy (%) |
|--------|---------|--------------|
| LPCNN (Zhong et al., 2016) | An improved CNN model with global average pooling layer, and trained with large patch sampling | 89.9 |
| MS-DCNN (Xu et al., 2016) | Multisource-driven and a multi-kernel SVM classifier | 91.34 |
| CNN with Overfeat pretrained model (Marmanis et al., 2016) | A pretrained network (Sermanet et al., 2013) is directly used as an extractor for remote sensing image features | 92.4 |
| CNN-ELM (Weng et al., 2017) | A pretrained network is directly used as an extractor for remote sensing image features and an extreme learning machine (ELM) classifier | 95.62 |
| Two-stage neural network ensemble model (Li, Fu, et al., 2017) | Combine pretrained networks and RBM retrained network as a two stage-training network and combine the classification results of several networks in testing stages | 95.11 |
| GCF-LOF CNN (Zeng, Chen, Chen, & Li, 2018) | A novel CNN by integrating the global-context features (GCFs) and local-object-level features (LOFs) | 96 |
| Deep-local-global feature fusion framework (DLGFF) (Zhu, Zhong, Liu, Zhang, & Li, 2018) | DLGFF establishes a framework integrating multi-level semantics from the global texture feature–based method, the bag-of-visual-words (BoVW) model, and a pre-trained CNN. | 96.37 |
| Deep random-scale stretched convolutional neural network (SRSCNN) (Liu, Zhong, Fei, Zhu, & Qin, 2018) | SRSCNN forces the CNN to extract features that are robust to the scale variation. | 95.96 |
candidate region-based object detection method. The method involves three steps: the generation of candidate regions, feature extraction by the CNN and classification of candidate regions. Candidate regions are a series of locations in which the objects may appear in the pre-generated image. All of these locations will be used as the input for the CNN for feature extraction and classification.

Based on the candidate region generation method, existing studies can be classified into two categories, namely those that use the sliding window method and those that use a region proposal method. The sliding window method is a type of exhaustion method. In this method, a sliding window is used to extract candidate regions, and whether there are target objects is determined window by window. A region proposal method establishes regions of interest for object detection. Bounding boxes that may contain target objects are first generated, and whether there are target objects is then determined for each bounding box. To determine whether they rely on an external method for candidate region proposals, region proposal CNNs can further be classified into Region-based CNNs (R-CNNs) and Faster R-CNNs (Ren, Girshick, Girshick, & Sun, 2017). The common region proposal methods used in R-CNNs include selective search and edge boxes. In a Faster R-CNN, a region proposal network is used to generate candidate regions and an internal deep network is used to replace candidate region proposals. Table 3 compares the candidate region-based target detection methods in terms of input images for the CNN as well as advantages and disadvantages.

Table 3. Comparison of object detection methods.

| Method | Sliding window methods for detection | Region proposal methods for detection |
|--------|-----------------------------------|-------------------------------------|
|        | Input images for the CNN           | R-CNN                               | Faster R-CNN                        |
|        | Advantages                         | Image area                          | Image area                          |
|        | Relatively high object identification capability | Relatively high object identification capability; relatively few candidate regions (which can be controlled within thousands or hundreds of regions) | Relatively high object identification capability; end-to-end object detection; independent of a region proposal method; very high detection speed |
|        | Disadvantages                      | The method of exhaustion generates a large number of candidate regions, resulting in a significant volume of repeated calculations in the subsequent operations, and therefore has relatively low efficiency when used to process remote sensing images of large scenes. | Relatively long time to calculate candidate regions. Feature extraction will result in repeated calculations. | Relatively complex training process |
|        | Typical literature                 | Buildings (Li, Li, et al., 2017; Zhang, Wang, et al., 2016); Ships (Huang, Jiang, Zhang, & Yao, 2017); Palm trees (Li, Fu, et al., 2016); Golf courses (Ishii, Nakamura, Nakada, Mochizuki, and Ishikawa, 2015); Aircraft (Yao, Wang, Huo, and Fang, 2017) | Multi-class targets (Cheng, Zhou, & Han, 2016); Electrical devices detection in UAV images (Wang, Tian, Liu, Liu, & Li, 2017); Aircraft (Wu, Zhang, Zhang, & Xu, 2015) | Aircraft (Deng et al., 2017; Zhang, Du, et al., 2016) |
5.3. Object segmentation

To extract objects from a remote sensing image, it is necessary to segment the objects of interest in the image and to produce a pixel-level image classification map. Two types of methods are primarily used in the existing CNN-based studies on object segmentation from remote sensing images, namely patch-based CNN methods and end-to-end CNN methods. A patch-based CNN method generally first obtains a prediction model by training a CNN on a training dataset, and then, based on the prediction model, it generates image patches using a sliding window pixel by pixel and predicts the type of each pixel of the image. The fully convolutional networks (FCN) method, a common end-to-end CNN method, substitutes deconvolutional layers for the fully connected layers of a CNN, allowing the network to accept input images of any size and directly generate pixel-level object extraction results. Table 4 compares the two existing types of deep learning methods in terms of input images as well as their advantages and disadvantages.

| Method                  | Patch-based CNN | End-to-end CNN                  |
|-------------------------|-----------------|---------------------------------|
| Input image             | Image patches   | Entire images (no fixed size is required) |
| Advantages              | Relatively good extraction results | Able to perform end-to-end object extraction and directly generate classification maps |
| Disadvantages           | A large volume of redundant calculations; relatively complex processing process; generally requiring post-processing of the extraction results | Insufficiently fine classification results; lack of spatial homogeneity |
| Typical literature      | Buildings (Li, Zhang, et al., 2016; Qu, Ren, Liu, & Li, 2017; Guo et al., 2016); Farmlands (Lu et al., 2015); Land cover (Chen, Fan, Wang, Wu, & Sun, 2017; Zhang, Zheng, Tang, & Zhao, 2017); Water bodies (He, Du, Chen, & Chen, 2017); Roads (Li, Zang, et al., 2016); Roads and buildings (Alshehhi, Marpu, Woon, & Mura, 2017; Saito & Aoki, 2015) | Buildings (Maggiori et al., 2017; Xu et al., 2018; Wu et al., 2018) |

6. Challenges and conclusions

The emergence of deep learning has provided an opportunity for mining and analyzing big remote sensing data. CNNs, a type of deep learning model, play an important role in RS image classification research. In this paper, we surveyed the current state-of-the-art of CNN-based deep learning for RS image classification. Different from images in the computer vision field, extracting features from RS images is difficult because of the complexity of objects in RS images. Thus, there have been many studies addressing CNN-based RS image classification issues. They have achieved certain breakthroughs in CNN model, training data and training methods for RS image classification. However, these studies are just the beginning of CNN-based RS image classification research. RS image classification is still facing unprecedented and significant challenges, and a number of issues are in need to be thought and investigated in depth, which we summarize as follows:
6.1. Insufficient training data in remote sensing

Sufficient training data is the key for training CNN models for RS image classification. According to our investigation, RS training datasets are much less than image datasets in the computer vision field. It is understandable that the preparation of RS training data are much more time-consuming. Remote sensing scientists who devote to deep-learning-based research are still limited, and less of them pay more efforts in RS training dataset production. Some studies are concerned with data augmentation techniques for addressing insufficient training data. With data augmentation techniques, the training sample size and sample diversity are increased. However, they are inadequate when applied in training of complex or large deep learning models. The large increase of training datasets in remote sensing is indispensable. The insufficient training dataset issue requires the attention from world-wide remote sensing communities, and these communities may promote and sponsor some initiatives of developing RS training datasets. The training datasets could come from multi-source remote sensing data, including visible-light, SAR, hyperspectral and multispectral images. Besides, studies of CNN training with non-accurately labeled data, including weakly supervised, semi-supervised and unsupervised data, is expected to further develop. These studies will be the supplement of the studies on developing accurately labeled training datasets.

6.2. RS image-specific CNN models

RS images are different from images in the computer vision field. They involve sensors with multiple imaging methods, including optical imaging, thermal imaging, LiDAR and radar. They can also come from satellite platforms or airborne platforms so they are various in spatial scale. At the same time, unlike objects covering most of a natural image, objects in RS images are generally small and decentralized. Furthermore, the viewing angle from RS images, unlike natural images, is often top-down, which makes it difficult to extract features from RS images. Therefore, CNN models developed in the computer vision field is not adequate for RS image classification. As investigated in this review, existing studies have improved CNN from various perspectives, including input data, fully connected layers, classifiers, loss functions and network structures, for achieving better RS image classification accuracy. These studies achieve enormous success in RS image classification. There is no doubt that RS image-specific CNN models need to be studied further. The future research deserved more attention include: (a). study of CNNs with multiple RS image inputs. The multiple RS image inputs refers to multi sensors with the same or very close spatial scales. The CNNs with multiple RS image input could utilize much more features in spectrum, shape and texture. In addition, studies on dealing with multi RS image inputs in different spatial scale are required. (b) study of the general structure of CNNs specific for remote sensing images. CNNs have a flexible structure but there is a lack of sufficient theory for designing structure of CNNs. Existing studies on the structure of CNNs are conducted on the basis of empirical knowledge. The general structure of CNNs with the help of remote sensing theory is desirable.
6.3. The CNN’s time efficiency

The majority of the CNN-based RS image classification studies focus more on classification accuracy. Very few studies focus on a CNN’s time efficiency during training CNN models. To meet the requirements of big remote sensing images in practical production, high-performance computing devices (e.g., GPUs) can be used to accelerate model training and testing; advanced model training techniques can be used to accelerate model training and testing. Transfer learning, which is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task, is an effective approach to speed up training of CNNs. CNN’s training time efficiency could be addressed in future studies.

6.4. High-level CNN-based applications in RS image classification

Current CNN-based applications in RS image classification resemble the classification tasks in computer vision. They are scene classification, object detection and object segmentation. The first two tasks being the majority. More attention could be drawn in higher level CNN-based applications, e.g., high-accuracy extraction of semantic information on scenes, extraction of more complex objects, super-resolution reconstruction, multi-label remote sensing image retrieval and so on.

CNN-based classification is a state-of-the-art classification approach to extracting geographic features from remote sensing images. This paper reviews the literature on CNN-based remote sensing image classification. We summarized the improvements on CNN models for remote sensing classification. This work is helpful for understanding how CNN can be better applied to remote sensing image classification. Training data is always the key to deep learning methods. Thus, the available open datasets and data augmentation techniques for remote sensing classification are comprehensively surveyed. We also summarized the methods for three typical remote sensing image classification tasks: scene classification, object detection and object segmentation with specific applications of CNN-based models for remote sensing image classification. Finally, the challenges of CNN-based remote sensing image classification research are listed, and corresponding suggestions are proposed. We hope that this paper can facilitate the advancement of remote sensing image classification and help remote-sensing scientists to further explore or discover more remote sensing image classification methods.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was jointly funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA23100103); the 13th Five-year Informatization Plan of Chinese Academy
References

Alshehhi, R., Marpu, P. R., Woon, W. L., & Mura, M. D. (2017). Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130, 139–149.

Bai, X., Zhang, H., & Zhou, J. (2014). VHR object detection based on structural feature extraction and query expansion. *IEEE Transactions on Geoscience and Remote Sensing*, 52, 6508–6520.

Bai, Y., Jiang, D. M., Pei, J. J., Zhang, N., & Bai, Y. (2018). Application of an improved ELU convolution neural network in the SAR image ship detection. *Bulletin of Surveying and Mapping*, 1, 125–128.

Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). A comprehensive survey of deep learning in remote sensing: theories, tools and challenges for the community. *Journal of Applied Remote Sensing*, 11, 4.

Benedek, C., Descombes, X., & Zerubia, J. (2011). Building development monitoring in multi-temporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 34, 33–50.

Bosch, A., & Zisserman, A. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Computer Society*, 30(4), 712–727.

Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. *Acta Ecologica Sinica*, 28(2), 627–635.

Chang, L., Deng, X. M., Zhou, M. Q., Wu, Z. K., Yuan, Y., Yang, S., & Wang, H. A. (2016). Convolutional neural networks in image understanding. *Acta Automatica Sinica*, 42, 1300–1312.

Chen, Y., Fan, R. S., Wang, J. X., Wu, Z. L., & Sun, R. X. (2017). High resolution Image Classification method combining with minimum noise fraction rotation and convolution neural network. *Laser & Optoelectronics Progress*, 54, 434–439.

Cheng, G., & Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117, 11–28.

Cheng, G., Han, J., Zhou, P., & Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *Isprs Journal of Photogrammetry & Remote Sensing*, 98, 119–132.

Cheng, G., Zhou, P., & Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54, 7405–7415.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 886–893), San Diego, CA.

Deng, Z., Lei, L., Sun, H., Zou, H., Zhou, S., & Zhao, J. (2017). An enhanced deep convolutional neural network for densely packed objects detection in remote sensing images. 2017 *International Workshop on Remote Sensing with Intelligent Processing (RSIP)* (pp. 1–4), Shanghai, China.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout networks. arXiv preprint arXiv:1302.4389.

Guo, Z., Shao, X., Xu, Y., Miyazaki, H., Ohira, W., & Shibasaki, R. (2016). Identification of village building via google earth images and supervised machine learning methods. *Remote Sensing*, 8, 271.

Hara, K., Saito, D., & Shouno, H. (2015). *Analysis of function of rectified linear unit used in deep learning*. International Joint Conference on Neural Networks, Killarney, Ireland.
He, H. Q., Du, J., Chen, T., & Chen, X. Y. (2017). Remote sensing image water body extraction combing NDWI with convolutional neural network. Remote Sensing Information, 32, 82–86.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceeding IEEE Conference Computer Vision and Pattern Recognition, Las Vegas, NV.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778), Boston, MA.

Hecht-Nielsen, R. (1989). Theory of the backpropagation neural network. Neural Networks, 1989.

IJCNN. International Joint Conference, Washington, DC.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504–507.

Huang, G., Liu, Z., Maaten, L. V. D., & Weinberger, K. Q. (2016). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708), Las Vegas, NV.

Huang, J., Jiang, Z. G., Zhang, H. P., & Yao, Y. (2017). Ship object detection in remote sensing images using convolutional neural networks. Journal of Beijing University of Aeronautics and Astronautics, 43, 1841–1849.

Huang, L., Liu, B., Li, B., Guo, W., Yu, W., Zhang, Z., & Yu, W. (2017). OpenSARShip: A dataset dedicated to sentinel-1 ship interpretation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11, 1.

Ishii, T., Nakamura, R., Nakada, H., Mochizuki, Y., & Ishikawa, H. (2015). Surface object recognition with cnn and svm in landsat 8 images. 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan.

Jia, K., Li, Q. Z., Tian, Y. C., & Wu, B. F. (2011). A review of classification methods of remote sensing. Spectroscopy and Spectral Analysis, 31, 2618–2623.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems (pp. 1097–1105), Lake Tahoe, NV.

Li, D. R. (2008). Development prospect of photogrammetry and remote sensing. Geomatics and Information Science of Wuhan University, 33, 1211–1215.

Li, D. R., Tong, Q. X., Li, R. X., Gong, J. Y., & Zhang, L. P. (2012). Current issues in high-resolution Earth observation technology. Sci China Earth Sci, 55, 1043–1051.

Li, D. R., Zhang, L. P., & Xia, G. S. (2014). Automatic analysis and mining of remote sensing big data. Acta Geodaetica Et Cartographica Sinica, 43, 1211–1216.

Li, H., Fu, K., Xu, G., Zheng, X., Ren, W., & Sun, X. (2017). Scene classification in remote sensing images using a two-stage neural network ensemble model. Remote Sensing Letters, 8, 557–566.

Li, J., Zhang, R., & Li, Y. (2016). Multiscale convolutional neural network for the detection of built-up areas in high-resolution SAR images. Geoscience and Remote Sensing Symposium (IGARSS) (pp. 910–913), Beijing, China.

Li, J. W., Qu, C. W., & Peng, S. J. (2018). Ship classification for unbalanced SAR dataset based on convolutional neural network. Journal of Applied Remote Sensing, 12(3), 035010.

Li, P., Zang, Y., Wang, C., Li, J., Cheng, M., Luo, L., & Yu, Y. (2016). Road network extraction via deep learning and line integral convolution. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1599–1602), Beijing, China.

Li, W., Fu, H., Yu, L., & Cracknell, A. (2016). Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. Remote Sensing, 9, 22.

Li, Z., Li, Y. S., Wu, X., Liu, G., Lu, H., & Tang, M. (2017). Hollow village building detection method using high resolution remote sensing image based on CNN. Transactions of the Chinese Society of Agricultural Machinery, 48, 160–165.

Li, Z. J., Li, X., J. J., Liu, T., Xie, J. W., & Yang, S. (2015). Remote sensing cloud computing: Current research and prospect. Journal of Equipment Academy, 26, 95–100.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. ICML, 2(3), 7.
Liu, Y., Zhong, Y., Fei, F., Zhu, Q., & Qin, Q. (2018). Scene classification based on a deep random-scale stretched convolutional neural network. *Remote Sensing, 10*, 444.

Long, Y., Gong, Y., Xiao, Z., & Liu, Q. (2017). Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience & Remote Sensing, 55*, 2486–2498.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*, 91–110.

Lu, H., Fu, X., He, Y. N., Li, L. G., Zhuang, W. H., & Liu, T. G. (2015). Cultivated land information extraction from high resolution UAV images based on transfer learning. *Transactions of the Chinese Society of Agricultural Machinery, 46*, 274–279.

Luo, J. C., Zhou, C. H., & Yang, Y. (2001). ANN remote sensing classification model and its integration approach with Geo-knowledge. *Journal of Remote Sensing, 5*, 122–129.

Lv, F. H., Shu, N., Gong, Y., Guo, Q., & Qu, X. G. (2017). Regular building extraction from high resolution image based on multilevel-features. *Geomatics and Information Science of Wuhan University, 42*, 1–5.

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*, 645–657.

Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2016). Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters, 13*, 105–109.

Nogueira, K., Penatti, O. A. B., & Santos, J. A. D. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition, 61*, 539–556.

Nogueira, K., Santos, J. A. D., Fornazari, T., Silva, T. S. F., Morellato, L. P., & Torres, R. D. S. (2017). *In Towards vegetation species discrimination by using data-driven descriptors*. 2016 9th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS) (pp. 1–6), IEEE, Cancun, Mexico.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 971–987.

Penatti, O. A. B., Nogueira, K., & Santos, J. A. D. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains. *Computer Vision and Pattern Recognition Workshops 44–51.*

Qu, C., Ren, Y. H., Liu, Y. L., & Li, Y. (2017). Functional classification of urban buildings in high resolution remote sensing images through POI-assisted analysis. *Journal of Geo-information Science, 19*, 831–837.

Ren, S., Girshick, R., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 39*, 1137–1149.

Saito, S., & Aoki, Y. (2015). *Building and road detection from large aerial imagery*. Image Processing: Machine Vision Applications VIII (Vol. 9405, p. 94050K), International Society for Optics and Photonics, San Diego, CA.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. unpublished paper. [Online]. Retrieved from http://arxiv.org/abs/1312.6229

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Song, W. J., Liu, P., Wang, L. Z., & Lv, K. (2014). Intelligent processing of remote sensing big data: status and challenges. *Journal of Engineering Studies, 6*, 259–265.

Tai, C., Xiao, T., Zhang, Y., Wang, X., & Weinan, E. (2015). Convolutional neural networks with low-rank regularization. arXiv preprint arXiv:1511.06067.

Tang, J., Deng, C., Huang, G. B., & Zhao, B. (2014). Compressed-domain ship detection on space-borne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience & Remote Sensing, 53*, 1174–1185.

Tao, C., Tan, Y., Cai, H., & Tian, J. (2011). Airport detection from large ikonos images using clustered sift keypoints and region information. *IEEE Geoscience and Remote Sensing Letters, 8*, 128–132.
Wang, W. G., Tian, B., Liu, Y., Liu, L., & Li, J. X. (2017). Study on the electrical devices in UAV images based on region based convolutional neural networks. *Journal of Geo-information Science, 19*, 256–263.

Weng, Q., Mao, Z., Lin, J., & Guo, W. (2017). Land-use classification via extreme learning classifier based on deep convolutional features. *IEEE Geoscience and Remote Sensing Letters, 14*, 704–708.

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., … Shibasaki, R. (2018). Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing, 10*, 407.

Wu, H., Zhang, H., Zhang, J., & Xu, F. (2015). Typical target detection in satellite images based on convolutional neural networks. *IEEE International Conference on Systems* (pp. 2956–2961), Las Vegas, NV.

Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., … Lu, X. (2016). Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience & Remote Sensing, 55*(7), 3965–3981.

Xia, G. S., Yang, W., Delon, J., Gousseau, Y., Sun, H., & Maitre, H. (2010). Structural high-resolution satellite image indexing. ISPRS TC VII Symposium - 100 Years ISPRS, Vienna, Austria.

Xia, M., Cao, G., Wang, G. Y., & Shang, Y. F. (2017). Remote sensing image classification based on deep learning and conditional random fields. *Journal of Image and Graphics, 22*, 1289–1301.

Xu, S. H., Mu, X. D., Zhao, P., & Ma, J. (2016). Scene classification of remote sensing image based on multi-scale feature and deep neural network. *Acta Geodaetica Et Cartographica Sinica, 45*, 834–840.

Xu, Y., Wu, L., Xie, Z., & Chen, Z. (2018). Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing, 10*, 144.

Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. *SIGspatial International Conference on Advances in Geographic Information Systems, New York, NY.*

Yao, X. K., Wang, L. H., Huo, H., & Fang, T. (2017). Airplane object detection in high resolution remote sensing imagery based on multi-structure convolutional neural network. *Computer Engineering, 43*, 259–267.

Zeng, D., Chen, S., Chen, B., & Li, S. (2018). Improving remote sensing scene classification by integrating global-context and local-object features. *Remote Sensing, 10*, 734.

Zhang, B. (2017). Current status and future prospects of remote sensing. *Bulletin of Chinese Academy of Sciences, 32*, 774–784.

Zhang, F., Du, B., Zhang, L., & Xu, M. (2016). Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. *IEEE Transactions on Geoscience and Remote Sensing, 54*, 5553–5563.

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 4*(2), 22–40.

Zhang, Q., Wang, Y., Liu, Q., Liu, X., & Wang, W. (2016). CNN based suburban building detection using monocular high resolution google earth images. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 661–664), IEEE, Beijing, China.

Zhang, W., Zheng, K., Tang, P., & Zhao, L. J. (2017). Land cover classification with features extracted by deep convolutional neural network. *Journal of Image and Graphics, 22*, 1144–1153.

Zhang, Y. H., Xia, G. H., Han, X., He, J., Ge, T. T., & Wang, J. G. (2018). Road extraction from multi-source high resolution remote sensing image based on fully convolutional network. 2018 International Conference on Audio, Language and Image Processing (ICALIP) (pp. 201–204), IEEE, Shanghai, China.

Zhang, Z. X., Wang, Y. H., Liu, Q. J., Li, L. L., & Wang, P. (2016). A CNN based functional zone classification method for aerial images. IGARSS, Beijing, China.

Zhao, B., Zhong, Y., Xia, G. S., & Zhang, L. (2016). Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience & Remote Sensing, 54*, 2108–2123.

Zhao, Y. S. (2013). *Principles and methods of analysis of remote sensing applications*. Beijing, China: Science Press.
Zhong, Y., Fei, F., & Zhang, L. (2016). Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing, 10*, 025006.

Zhou, M., Shi, Z. W., & Ding, H. P. (2017). Aircraft classification in remote-sensing images using convolutional neural networks. *Journal of Image and Graphics, 22*, 702–708.

Zhou, W., Shao, Z., & Cheng, Q. (2016). Deep feature representations for high-resolution remote sensing scene classification. 2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA) (pp. 338–342), Guangzhou, China.

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., & Jiao, J. (2015). Orientation robust object detection in aerial images using deep convolutional neural network. IEEE International Conference on Image Processing (pp. 3735–3739), Quebec City, QC.

Zhu, Q., Zhong, Y., Liu, Y., Zhang, L., & Li, D. (2018). A deep-local-global feature fusion framework for high spatial resolution imagery scene classification. *Remote Sensing, 10*, 568.

Zhu, X. X., Tuia, D., Mou, L. C., et al. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*, 8–36.

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience & Remote Sensing Letters, 12*, 2321–2325.