Remote Sensing of Environment 237 (2020) 111322

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Land-cover classification with high-resolution remote sensing images using transferable deep models

Xin-Yi Tong[a], Gui-Song Xia[a,b,*], Qikai Lu[c], Huanfeng Shen[d], Shengyang Li[e], Shucheng You[f], Liangpei Zhang[a]

[a] State Key Laboratory LIESMARS, Wuhan University, China
[b] School of Computer Science, Wuhan University, China
[c] Electronic Information School, Wuhan University, China
[d] School of Resource and Environmental Sciences, Wuhan University, China
[e] Key Laboratory of Space Utilization, Tech. & Eng. Center for Space Utilization, Chinese Academy of Sciences, China
[f] Remote Sensing Department, China Land Survey and Planning Institute, China

ARTICLE INFO

Keywords:
land-cover classification
High-resolution remote sensing
Deep learning
Gaofen-2 satellite images

ABSTRACT

In recent years, large amount of high spatial-resolution remote sensing (HRRS) images are available for land-cover mapping. However, due to the complex information brought by the increased spatial resolution and the data disturbances caused by different conditions of image acquisition, it is often difficult to find an efficient method for achieving accurate land-cover classification with high-resolution and heterogeneous remote sensing images. In this paper, we propose a scheme to apply deep model obtained from labeled land-cover dataset to classify unlabeled HRRS images. The main idea is to rely on deep neural networks for presenting the contextual information contained in different types of land-covers and propose a pseudo-labeling and sample selection scheme for improving the transferability of deep models. More precisely, a deep Convolutional Neural Networks (CNNs) is first pre-trained with a well-annotated land-cover dataset, referred to as the *source data*. Then, given a *target image* with no labels, the pre-trained CNN model is utilized to classify the image in a patch-wise manner. The patches with high confidence are assigned with pseudo-labels and employed as the queries to retrieve related samples from the source data. The pseudo-labels confirmed with the retrieved results are regarded as supervised information for fine-tuning the pre-trained deep model. To obtain a pixel-wise land-cover classification with the target image, we rely on the fine-tuned CNN and develop a hybrid classification by combining patch-wise classification and hierarchical segmentation. In addition, we create a large-scale land-cover dataset containing 150 Gaofen-2 satellite images for CNN pre-training. Experiments on multi-source HRRS images, including Gaofen-2, Gaofen-1, Jilin-1, Ziyuan-3, Sentinel-2A, and Google Earth platform data, show encouraging results and demonstrate the applicability of the proposed scheme to land-cover classification with multi-source HRRS images.

1. Introduction

Land-cover classification with remote sensing (RS) images plays an important role in many applications such as land resource management, urban planning, precision agriculture, and environmental protection (Mathieu et al., 2007; Shi et al., 2015; Ozdarici-Ok et al., 2015; Zhang and Kovacs, 2012; Ardila et al., 2011; Fauvel et al., 2013). In recent years, high-resolution remote sensing (HRRS) images are increasingly available. Meanwhile, multi-source and multi-temporal RS images can be obtained over different geographical areas (Moser et al., 2013). Such large amount of heterogeneous HRRS images provide detailed information of the land surface, and therefore open new avenues for large-coverage and multi-temporal land-cover mapping. However, the rich details of objects emerging in HRRS images, such as the geometrical shape and structural content of objects, bring more challenges to land-cover classification (Bruzzone and Carlin, 2006). Furthermore, diverse imaging conditions usually lead to photographic distortions, variations in scale and changes of illumination in RS images, which often seriously reduces the separability among different classes (Tuia et al., 2016). Due to these influences, optimal classification models learned from certain annotated images always quickly lose their effectiveness on new images captured by different sensors or by the same

* Corresponding author.
E-mail address: guisong.xia@whu.edu.cn (G.-S. Xia).

https://doi.org/10.1016/j.rse.2019.111322
Received 23 July 2018; Received in revised form 14 June 2019; Accepted 14 July 2019
Available online 28 November 2019
0034-4257/

sensor but from different geo-locations. Therefore, it is intractable to find an efficient and accurate land-cover classification method for HRRS images with large diversities.

To characterize the image content of different land-cover categories, many methods investigated the use of spectral and spectral-spatial features to interpret RS images (Jensen and Lulla, 1986; Gong et al., 1992; Casals-Carrasco et al., 2000; Giada et al., 2003; Tarabalka et al., 2010a, b; Zhong et al., 2014; Ma et al., 2017a). However, due to the detailed and structural information brought by the gradually increased spatial resolution, the spectral and spectral-spatial features have difficulty in describing the contextual information contained in the images (Zhao et al., 2016; Zhong et al., 2017; Hu et al., 2016; Yu et al., 2016), which are often essential in depicting land-cover categories in HRRS images. Recently, it has been reported that effective characterization of contextual information in HRRS images can largely improve the classification performance (Shao et al., 2013; Hu et al., 2017; Yang et al., 2015). Among them, deep Convolutional Neural Networks (CNNs) have been drawn much attention in the understanding of HRRS images (Hu et al., 2015a; Zhu et al., 2017), mainly because of their strong capability to depict high-level and semantic aspects of images (Krizhevsky et al., 2012; Zeiler and Fergus, 2014). Currently, various deep models have been adopted to cope with challenging issues in RS image understanding, including *e.g.* scene classification (Hu et al., 2015a; Xia et al., 2017c), object detection (Xia et al., 2018), image retrieval (Napoletano, 2018; Jiang et al., 2017; Xia et al., 2017b), as well as land-cover classification (Zhao and Du, 2016; Zhao et al., 2015; Zhang et al., 2018a; Maggiori et al., 2017b; Kussul et al., 2017; Volpi and Tuia, 2017).

Nevertheless, there are two main problems in applying deep model to land-cover classification with multi-source HRRS images, which are listed below.

- *The inadequate transferability of deep learning models*: Due to the diverse distributions of objects and spectral shifts caused by the different acquisition conditions of images, deep models trained on a certain set of annotated RS images may not be effective when dealing with images acquired by different sensors or from different geo-locations (Othman et al., 2017). To obtain satisfactory land-cover classification on a RS image of interest, referred as the *target image*, new specific annotated samples closely related to it are often necessary for model fine-tuning (Maggiori et al., 2017b). Nevertheless, considering that manual annotation requires high labor intensity and is often time-consuming, it is infeasible to label sufficient samples for continuously accumulated multi-source RS images (Lu et al., 2017; Hu et al., 2015b).
- *The lack of well-annotated large-scale land-cover dataset*: The identification capability of CNN models relies heavily on the quality and quantity of the training data (Chakraborty et al., 2015). Up to now, several land-cover datasets have been proposed in the community, and have advanced a lot deep-learning-based land-cover classification approaches (Gerke et al., 2014; Maggiori et al., 2017a; Mattyus et al., 2015). However, the geographic areas covered by most of existing land-cover datasets (Ma et al., 2017b; Gerke et al., 2014; Mattyus et al., 2015) do not exceed $10 km^2$ and somewhat similar in geographic distributions (Mnih, 2013). The lack of variations in geographic distributions of annotated HRRS images may cause overfitting in model training and limit the generalization ability of learned models. Overall, the insufficient or unqualified training data restrict the availability of deep models for HRRS images.

In this paper, we propose a scheme to adapt deep models to land-cover classification with multi-source HRRS images, which don't have any labeling information. Considering that the textures and structures of the objects are not affected by the spectral shifts, we use contextual information extracted by CNN to automatically mine samples for deep model fine-tuning. Concretely, unlabeled samples in the target image are identified by a CNN model pre-trained on an annotated HRRS dataset, which is referred to as the *source data*. A subset of them with high confidence are assigned with pseudo-labels and employed to retrieve similar samples from the source data. Finally, the returned results are used to determine whether the pseudo-labels are reliable. In our classification process, a patch-wise classification is initially conducted on the image relying on the multi-scale contextual information extracted by CNN. Then, a hierarchical segmentation is used for obtaining the object boundary information, which is integrated into the patch-wise classification map for accurate results. Specifically, for pre-training CNN models, we annotate 150 Gaofen-2 satellite images to construct a land-cover classification dataset, which is named after *Gaofen Image Dataset* (GID).

In summary, the contributions of this paper are as follows:

- We propose a scheme to train transferable deep models, which enables one to achieve land-cover classification by using unlabeled multi-source RS images with high spatial resolution. In addition, we develop a hybrid land-cover classification that can simultaneously extract accurate category and boundary information of HRRS images. Experiments conducted on multi-source HRRS images, including Gaofen-2, Gaofen-1, Jilin-1, Ziyuan-3, Sentinel-2A, and Google Earth platform data obtain promising results and demonstrate the effectiveness of the proposed scheme.
- We present a large-scale land-cover classification dataset, namely GID, which is consist of 150 high-resolution Gaofen-2 images and covers areas more than 50,000 km^2 in China. To our knowledge, GID is the first and largest well-annotated land-cover classification dataset with high-resolution remote sensing images up to 4 m. It can provide the research community a high-quality dataset to advance land-cover classification with HRRS images, *like* Gaofen-2 imagery.

A preliminary version of this work was presented in (Tong et al., 2018).

The remainder of the paper is organized as follows: In Section 2, we introduce the related works. In Section 3, the introduction of our land-cover classification algorithm is presented. In Section 4, the details and properties of GID coupled with other examined images are described. We present the results of experiments and sensitivity analysis in Section 5 and Section 6, and give the discussion in Section 7. Finally, we conclude our work in Section 8.

2. Related work

Land-cover Classification: Land-cover classification with RS images aims to associating each pixel in a RS image with a pre-defined land-cover category. To this end, classification approaches using spectral information have been intensively studied. These methods can interpret RS images using the spectral features of individual pixels (Jensen and Lulla, 1986; Gong et al., 1992; Casals-Carrasco et al., 2000), but their performance is often heavily affected by intra-class spectral variations and noises (Blaschke, 2001; Burnett and Blaschke, 2003; Benz et al., 2004). More recently, the spatial information has been taken into consideration for land-cover classification (Giada et al., 2003; Tarabalka et al., 2010a, b; Zhong et al., 2014; Ma et al., 2017a). Spectral-spatial classification incorporates spatial information, such as texture (Pacifici et al., 2009; Xia et al., 2010a, 2017a), shape (Zhang et al., 2006), and structure features (Tuia et al., 2010; Xia et al., 2010b), to improve the separability of different categories in the feature space. It has been reported that spectral-spatial approaches can effectively boost the categorization accuracy compared with the methods using spectral features (Blaschke, 2010; Yan et al., 2006; Myint et al., 2011; Duro et al., 2012). However, with the improvement of spatial-resolution of RS images, rather than discriminating in spectral or spectral-spatial information of local pixels, land-cover types are more categorized in contextual information and spatial relationship of ground objects (Shao et al., 2013; Yang et al., 2015).

Recently, deep neural network models have been intensively studied to address the task of land-cover classification and reported impressive performance, see *e.g.* (Zhao and Du, 2016; Zhao et al., 2015; Zhang et al., 2018a; Paisitkriangkrai et al., 2015, 2016; Audebert et al., 2016). In contrast with conventional methods that employ spectral or spectral-spatial features for land-cover description, a significant advantage of deep learning approaches is that they are able to adaptively learn discriminative features from images (Zeiler and Fergus, 2014). Land-cover classification approaches that utilize deep features treat CNN models as feature extractors and employ conventional classifiers (Zhao and Du, 2016; Zhao et al., 2015; Paisitkriangkrai et al., 2015, 2016; Audebert et al., 2016), such as support vector machine (SVM) and logistic regression, for classification. As an alternative, end-to-end CNN models are adopted to interpret RS images (Maggiori et al., 2016, 2017b; Kussul et al., 2017; Volpi and Tuia, 2017; Liu et al., 2017). End-to-end CNN models, such as Fully Convolutional Networks (FCN) (Maggiori et al., 2017b), conduct dense land-cover labeling for RS images without using additional classifiers or post-processing. Although, compared with utilizing deep features, end-to-end CNN models are more efficiency for classification, the receptive field in CNNs leads to the loss of fine resolution detail (Liu et al., 2017). To address this problem, an effective solution is to replace down-sampling process with the structure that preserves spatial information (Sherrah, 2016; Persello and Stein, 2017), such as the max pooling indices applied in SegNet (Badrinarayanan et al., 2017), and the dilated convolutions employed in DeepLab (Chen et al., 2018). Moreover, the detailed spatial information can be obtained through complementary classification frameworks. The ensemble MLP-CNN classifier (Zhang et al., 2018a) fuses results acquired from the CNN based on deep spatial feature representation and from the multi-layer perceptron (MLP) based on spectral discrimination, which compensates the uncertainty in object boundary partition. And object-based convolutional neural network (OCNN) (Zhang et al., 2018b) incorporates CNN into the framework of object-based image analysis for more precise object boundaries and achieves encouraging performance in high-resolution urban scenes.

Transfer Learning: In practical land-cover classification applications, the available ground-truth samples are usually not sufficient in number and not adequate in quality for training a high-performance classifier. Thus, to improve the classification accuracy, transfer learning has been adopted as a promising solution (Tuia et al., 2016). Transfer learning aims to adapt models trained to solve a specific task to a new, yet related, task (Li et al., 2017). The existing task is usually referred to as *source domain* and the new task is *target domain*. Two major categories of transfer learning approaches have been studied in the RS community: supervised learning methods and semi-supervised learning methods. Supervised learning approaches assume that the training set is available for both the source and target domain. They are commonly based on the selection of invariant features (Izquierdo-Verdiguier et al., 2013; Bruzzone and Persello, 2009), the adaptation of data distributions (Tuia and Camps-Valls, 2016; Yang and Crawford, 2016; Jun and Ghosh, 2011), and active learning (Persello and Bruzzone, 2012; Demir et al., 2012, 2014). By contrast, the approaches are defined as semi-supervised if they use only unlabeled samples of the target domain. Semi-supervised learning methods exploit the structural information of unlabeled samples in the feature space to better model the distribution of classes (Persello and Bruzzone, 2014). They are effective in a wide range of situations and do not require a strict match between the source and target domains (Bruzzone et al., 2006; Gómez-Chova et al., 2008; Matasci et al., 2015). However, their performance is extremely dependent on the classifier's ability to learn structural information of the target domain.

On the other hand, deep neural networks are widely used for transfer learning in recent years due to their ability to model high-level abstractions of data (LeCun et al., 2015). CNN trained on large-scale natural image dataset have been transferred to interpret RS images, either by directly using the pre-trained network as a feature extractor (Hu et al., 2015a; Marmanis et al., 2016; Zhao et al., 2017), or fine-tuning the network with large-scale RS dataset (Xia et al., 2017c). However, due to the huge number of model parameters, deep models require large amount of supervised information of the target domain to avoid the over-fitting problem (Ge and Yu, 2017). To reduce the required number of training samples, semi-transfer deep convolutional neural networks (STDCNN) is proposed for land-use mapping (Huang et al., 2018). STDCNN fuses a small CNN designed to analyse RS images and a deep CNN used for domain adaptation. It achieves promising performance with a small amount of labeled target samples.

3. Methodology

To efficiently conduct land-cover classification with multi-source HRRS images, we propose a scheme to train transferable deep models, which is pre-trained on labeled land-cover dataset and can be applied to unlabeled HRRS images. Assume that there is a well-annotated large-scale dataset and a newly acquired image without labeling information. We define two domains, called *source domain* $\boldsymbol{D_S}$ and *target domain* $\boldsymbol{D_T}$ that are separately associated with the labeled and unlabeled images. Our aim is to exploit the information learned from the source domain $\boldsymbol{D_S}$ to conduct classification in the target domain $\boldsymbol{D_T}$.

Firstly, we use $\boldsymbol{D_S}$ to pre-train a deep model specific to RS domain, which is presented in Section 3.1. Given a target image χ_T belonging to $\boldsymbol{D_T}$, we divide it into patches $\mathbf{U} = \{\mathbf{x}_i\}_{i=1}^{I}$ with non-overlapping grid partition. Our method automatically searches reliable training samples from $\mathbf{U}$ to learn transferable deep model for χ_T, as introduced in Section 3.2. Subsequently, we utilize the fine-tuned deep model to classify $\mathbf{x}_i$ for all $i \in \{1, ..., I\}$. Our classification scenario combines patch-wise categorization and object-based voting, which is described in Section 3.3.

3.1. Learning deep model for land-cover classification

CNN models are deep hierarchical architectures which commonly consist of three main types of layers: convolutional layers, pooling layers, and fully-connected layers. Convolutional layers perform as hierarchical feature extractors, pooling layers conduct spatial down-sampling of feature maps, while fully-connected layer serve as the classifier to generate the predictive classification probabilities of the input data. In addition to these main layers, Residual Networks (ResNet) (He et al., 2016) adopt residual connections to improve the model performance. The structure of residual connection can greatly reduce the optimization difficulty, as well as enable the training of much deeper networks.

ResNet models have 5 versions, separately with 18, 34, 50, 101, and 152 layers. Compared to the models with shallow architecture, i.e. ResNet-18 and ResNet-34, ResNet-50 can achieve better classification performance. Compared to the models with very deep architecture, i.e. ResNet-101 and ResNet-152, ResNet-50 has fewer parameters and higher computational efficiency. Therefore, for a trad-off consideration between simplicity and computational efficiency, we employ ResNet-50 as the classifier in our work.

ResNet-50 consists of 16 residual blocks, each of which has 3 convolutional layers that constitute a shortcut connection. The first convolutional layer of the overall model is followed by a max pooling layer. An average pooling layer, a full-connected layer, and a softmax layer are subsequent to the last convolutional layer. Fig. 1 shows the detailed structure of ResNet-50 we employed. Note that the default input size of ResNet-50 is $224 \times 224 \times 3$. To transfer deep model to classify images with only R, G, B bands (Jilin-1 satellite images, Google Earth platform data), we utilize ResNet-50 with 3 input channels. To classify images have R, G, B, NIR bands (Gaofen-1, Sentinel-2A, Ziyuan-3 satellite images), we adjust the input size of the conversional ResNet-50 to 4 channels. The difference between 3-channel and 4-channel models is that the kernel sizes in conv1 are $7 \times 7 \times 4$ and $7 \times 7 \times 3$, respectively.

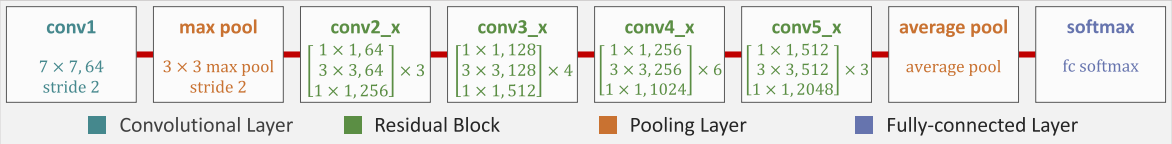

Fig. 1. The structure of ResNet-50. Different structures are represented by different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The rest of structures in the models are the same.

To pre-train a CNN model with strong discrimination ability for HRRS images, we construct a large-scale land-cover dataset, Gaofen Image Dataset (GID), which contains 150 well-annotated Gaofen-2 images and will be represented in Section 4. The training images of GID are referred to as $\boldsymbol{D_S}$ and are used to pre-train the ResNet-50 models.

3.2. Learning transferable model for multi-source images

Although CNNs have a certain degree of generalization ability, they are unable to achieve satisfactory classification results on multi-source RS images because of dramatic changes in acquisition conditions. To transfer CNN models for classifying RS images acquired under different conditions, we extract available information from the unlabeled target data to find a more accurate classification rule than using only labeled source data. Inspired by pseudo-label assignment (Wu and Yap, 2006; Lee, 2013) and joint fine-tuning (Xue et al., 2007; Ge and Yu, 2017) methods, we propose a semi-supervised transfer learning scheme to classify multi-source RS images. The main idea of pseudo-label assignment is to select valuable samples from the target domain based on the predicted classification confidence (Wu and Yap, 2006; Lee, 2013), however, the pseudo-labels may be unreliable. Joint fine-tuning optimizes the classification model by adding promising target samples into the source domain training set (Xue et al., 2007; Ge and Yu, 2017), however, it requires a small amount of labeled target samples. Our scheme combines the advantages of the aforementioned methods to acquire reliable training information from the unlabeled target domain for model optimization.

As shown in Fig. 2, the proposed scheme is divided into two stages: pseudo-label assignment and relevant sample retrieval, which are presented in Section 3.2.1 and Section 3.2.2, respectively. Category predictions and deep features generated by CNN are used to search target samples that possess similar characteristics to the source domain. These relevant samples and their corresponding category predictions, which are referred to as pseudo-labels, are used for CNN model fine-tuning.

3.2.1. Pseudo-label assignment

Given the patch set $\mathbf{U} = \{\mathbf{x}_i\}_{i=1}^{I}$ of a target image χ_T, we input each patch $\mathbf{x}_i$ to ResNet-50 that has been pre-trained on $\boldsymbol{D_S}$. The output vectors of softmax layer form a set $\mathbf{F} = \{\mathbf{p}_i\}_{i=1}^{I}$, where:

$$\mathbf{p}_i = \{p_{i,1}, p_{i,2}, ..., p_{i,K}\},\ \mathbf{p}_i \in \mathbb{R}^K \tag{1}$$

$p_{i,k}$ represents the probability that patch $\mathbf{x}_i$ belongs to class k, $k \in \{1, ..., K\}$, and K is the total number of classes.

$\mathbf{p}_i$ is the predicted classification probability vector, of which the highest probability value is $h = \max_{k\in\{1,...,K\}} p_{i,k}$. Since CNN model has strong discriminating ability, we use the probability value to determine whether a sample is associated with a label. If the value of h is greater than or equal to a threshold σ, the patch $\mathbf{x}_i$ is reserved and assigned with a predicted category l_i. Otherwise, $\mathbf{x}_i$ is removed from $\mathbf{U}$. l_i corresponds to the category represented by h and is referred to as a pseudo-label. After removing all patches with low classification confidence from $\mathbf{U}$, the remaining samples form a new set $\mathbf{U}_1$.

3.2.2. Relevant sample retrieval

Considering that the pseudo-labels may be inaccurate, we search source samples that are similar to the selected target samples, and use the true-labels of the retrieved source samples to determine the reliability of the pseudo-labels. Assume that there are J candidates remaining in the set $\mathbf{U}_1$, $\mathbf{U}_1 = \{\mathbf{x}_j\}_{j=1}^{J}$. For the patch $\mathbf{x}_j$, the information entropy E_j is calculated to measure its classification uncertainty:

$$E_j = -\sum_{k=1}^{K} p_{j,k} \cdot log(p_{j,k}) \tag{2}$$

where $p_{j,k}$ represents the probability that patch $\mathbf{x}_j$ belongs to class k.

The patches with higher information entropy are considered as valuable training samples, hence we treat them as preferred candidates. The patches in $\mathbf{U}_1$ are then sorted according to the descending order of E_j value, forming a sample set $\mathbf{U}_2 = \{\hat{\mathbf{x}}_j\}_{j=1}^{J}$. Considering that data with low information entropy provides insufficient information, we only use the top μ candidates of each category in the set $\mathbf{U}_2$ to perform retrieval as follows.

Given a patch $\hat{\mathbf{x}}_j$ that possesses the pseudo-label $\hat{l}_j$, we take it as a query image and retrieve its similar samples from the source domain $\boldsymbol{D_S}$. We use the deep features extracted from full-connected layer of the pre-trained ResNet-50 for retrieval. The similarities between $\hat{\mathbf{x}}_j$ and the source domain samples are measured by the Euclidean distance.

Then, we use the existing labels of the source domain to determine the confidence of the pseudo-labels. If the top δ retrieved results from the source domain have the same label g, and g is the same as the pseudo-label $\hat{l}_j$ of the query patch $\hat{\mathbf{x}}_j$ (*i.e.* $\hat{l}_j = g$), $\hat{\mathbf{x}}_j$ is considered to be a relevant sample. Otherwise, $\hat{\mathbf{x}}_j$ is removed from the set $\mathbf{U}_2$. After sample screening, the remainders of $\mathbf{U}_2$ form a new target domain set $\mathbf{U}_{tg}$. Finally, the image patches along with their corresponding pseudo-labels in the set $\mathbf{U}_{tg}$ are used to fine-tune a CNN model that is specific to the target image.

3.3. A hybrid land-cover classification algorithm

Land-cover classification aims to assign pixels in a RS image with land-cover category labels. Both the category and boundary information of the ground objects is essential for accurate classification. We therefore propose a hybrid algorithm, which combines patch-wise classification and hierarchical segmentation through a majority voting strategy, as shown in Fig. 3.

3.3.1. Patch-wise classification

Since the ground objects show different characteristics in different spatial resolutions, it is difficult to capture sufficient information of objects from the single-scale observation field. To exploit the attributes of the objects and their spatial distributions, we propose to utilize multi-scale contextual information for classification, which is illustrated in Fig. 4.

The target image χ_T is partitioned into non-overlapping patches $\mathbf{U} = \{\mathbf{x}_i\}_{i=1}^{I}$ by grid with the size of $s_1 \times s_1$ pixels (s_1 is the minimum value in the succession of scales). For each patch $\mathbf{x}_i$, its center pixel is regarded as a reference pixel $\mathbf{z}$. Around $\mathbf{z}$, a series of patches with sizes of $s_2 \times s_2$, ..., $s_N \times s_N$ pixels are sampled, so that each reference pixel possesses N multi-scale samples. Then, these multi-scale patches are uniformly resized to 224 × 224 and are input to the ResNet-50 model. After forward propagation, the classification probability vector $\mathbf{p}_{s_n}(\mathbf{z})$ of scale s_n at pixel $\mathbf{z}$ is obtained from softmax layer:

$$\mathbf{p}_{s_n}(\mathbf{z}) = \{p_{s_n,1}(\mathbf{z}), p_{s_n,2}(\mathbf{z}), ..., p_{s_n,K}(\mathbf{z})\},\ \mathbf{p}_{s_n}(\mathbf{z}) \in \mathbb{R}^K \tag{3}$$

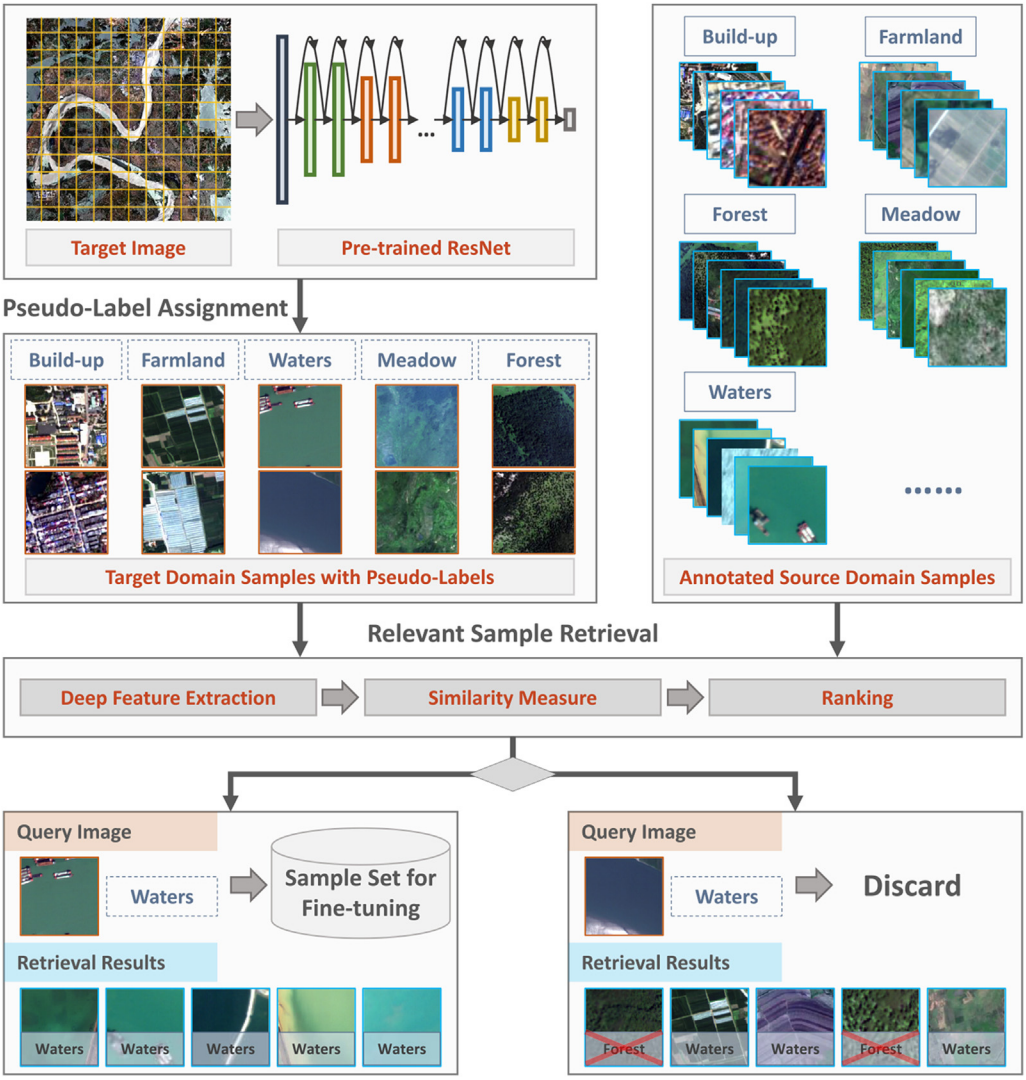

Fig. 2. Sample selection for model fine-tuning.

where $n \in \{1,...,N\}$, $p_{s_n,k}(\mathbf{z})$ represents the probability that $\mathbf{z}$ belongs to class k at the n-th scale.

Contextual information of multi-scale patches is aggregated using a weighted fusion strategy. The specificity measure (Lu et al., 2016), which describes the certainty of classification result, is employed as the weight:

$$W_{s_n}(\mathbf{z}) = \sum_{k=1}^{K-1} \frac{1}{k} \cdot \left(\hat{p}_{s_n,k}(\mathbf{z}) - \hat{p}_{s_n,k+1}(\mathbf{z})\right) \tag{4}$$

where $\{\hat{p}_{s_n,1}(\mathbf{z}), \hat{p}_{s_n,2}(\mathbf{z}),...,\hat{p}_{s_n,K}(\mathbf{z})\}$ is the descending order of the vector $\mathbf{p}_{s_n}(\mathbf{z})$. The value of $W_{s_n}(\mathbf{z})$ ranges from 0 to 1, and the higher value signifies the higher categorization confidence. The weighted probability

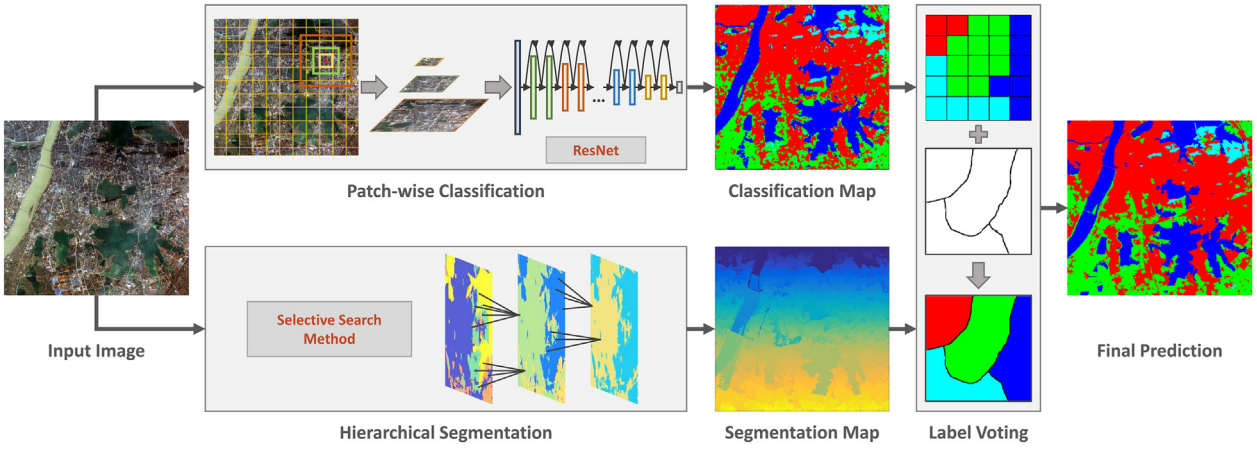

Fig. 3. The proposed land-cover classification approach.

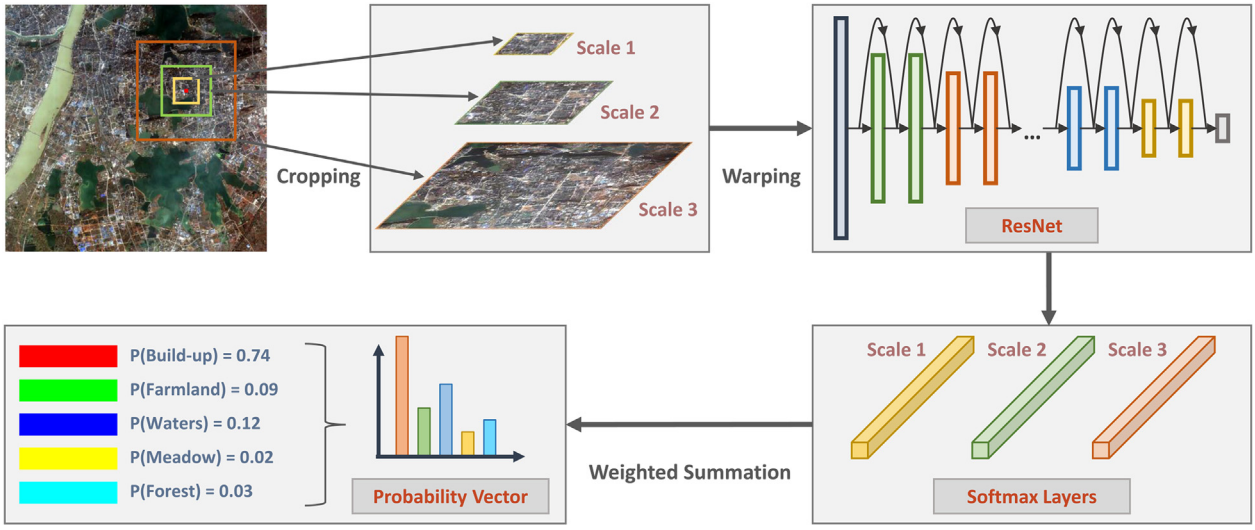

Fig. 4. Multi-scale contextual information aggregation.

$\tilde{\mathbf{p}}_k(\mathbf{z})$ is expressed as:

$$\tilde{\mathbf{p}}_k(\mathbf{z}) = \frac{\sum_{n=1}^{N} W_{s_n}(\mathbf{z}) \cdot p_{s_n,k}(\mathbf{z})}{\sum_{n=1}^{N} W_{s_n}(\mathbf{z})} \tag{5}$$

where $\tilde{\mathbf{p}}_k(\mathbf{z}) \in [0,1]$ indicates the probability that the reference pixel $\mathbf{z}$ belongs to class k. The aggregated probabilities of all categories can constitute a new classification probability vector. Then the reference pixel $\mathbf{z}$ is classified as:

$$l(\mathbf{z}) = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \tilde{\mathbf{p}}_k(\mathbf{z}) \tag{6}$$

where $l(\mathbf{z})$ is the category label of the pixel $\mathbf{z}$. Then, we assign the label $l(\mathbf{z})$ to each pixel in the patch $\mathbf{x}_i$. After classifying all the patches in the entire RS image, a patch-wise classification map $\chi_{\mathbf{c}}$ is therefore acquired.

3.3.2. Object-based voting

To obtain precise boundary information of the objects, we utilize segmentation map generated by selective search method (Uijlings et al., 2013) to refine the preliminary classification map. Selective search is a hierarchical segmentation method. It exploits a graph-based approach (Felzenszwalb and Huttenlocher, 2004) to produce a series of initial regions in different color spaces, and then uses the greedy algorithm to iteratively merge small regions. The color, texture, size and fill similarities are employed to control the merging level. Since various image features are considered in the process of initial segmentation and iterative merging, selective search can produce high-quality segmentation results.

After obtaining classification and segmentation maps by patch-wise classification and selective search, we integrate the category and boundary information using a majority voting strategy. Let $\mathbf{V} = \{\mathbf{y}_f\}_{f=1}^{F}$ denote the homogeneous regions in the segmentation map $\chi_{\mathbf{s}}$ generated from the target image χ_T. And $\hat{\mathbf{y}}_f$ is the corresponding area of $\mathbf{y}_f$ in the classification map $\chi_{\mathbf{c}}$. The number of pixels contained by $\hat{\mathbf{y}}_f$ is $M = |\mathbf{y}_f|$, and category label of the m-th pixel is l_m, $m \in \{1,\ldots,M\}$. Then the number of pixels belonging to each class in $\hat{\mathbf{y}}_f$ is counted, and the most frequent label $T(\mathbf{y}_f)$ is assigned to all pixels in $\mathbf{y}_f$:

$$T(\mathbf{y}_f) = \underset{r \in \{1,\ldots,K\}}{\operatorname{argmax}} \sum_{m=1}^{M} sign(l_m = r) \tag{7}$$

where $sign(\cdot)$ is an indicator function, $sign(true) = 1$, $sign(false) = 0$, and r denotes the possible class label. For all segmented objects, the same voting scheme is applied, and the final classification result is then acquired.

4. GID: a well-annotated dataset for land-cover classification

We construct a large-scale land-cover dataset with Gaofen-2 (GF-2) satellite images. This new dataset, which is named as Gaofen Image Dataset (GID), has superiorities over the existing land-cover dataset because of its large coverage, wide distribution, and high spatial resolution. GID consists of two parts: a large-scale classification set and a fine land-cover classification set. The large-scale classification set contains 150 pixel-level annotated GF-2 images (see Fig. 5), and the fine classification set is composed of 30,000 multi-scale image patches (see Fig. 6) coupled with 10 pixel-level annotated GF-2 images. More details are shown in Table 1. The training and validation data with 15 categories is collected and re-labeled based on the training and validation images with 5 categories, respectively.

The training images of GID are utilized to pre-train a CNN model with strong generalization ability specific to RS domain. In addition, it can serve as data resource to advance the state-of-the-art in land-cover classification task. GID and its reference annotations have been provided online at http://captain.whu.edu.cn/GID/.

Furthermore, to validate the transferability of our method on multi-source HRRS images, we annotate high-resolution images acquired by different sensors, including Gaofen-1, Jilin-1, Ziyuan-3, Sentinel-2A satellite images, and Google Earth platform data. GID and multi-source images are introduced in Section 4.1 and Section 4.2, respectively.

4.1. Gaofen image dataset

4.1.1. Gaofen-2 satellite images

Gaofen-2 (GF-2) is the second satellite of High-definition Earth Observation System (HDEOS) promoted by China National Space Administration (CNSA). Two panchromatic and multispectral (PMS) sensors with effective spatial resolution of 1 m panchromatic (pan)/4 m multispectral (MS) are onboard the GF-2 satellite, with a combined swath of 45 km. The resolution of the sub-satellite point is 0.8 m pan/3.24 m MS, and the viewing angle of a single camera is 2. 1°. GF-2 satellite realizes global observation within 69 days and repeat observations within 5 days.

The multispectral image we used to establish GID provide a spectral range of blue (0.45–0.52 μm), green (0.52–0.59 μm), red (0.63–0.69 μm) and near-infrared (0.77–0.89 μm), and a spatial dimension of 6800 × 7200 pixels covering a geographic area of 506 km^2.

GF-2 images achieve a combination of high spatial resolution and

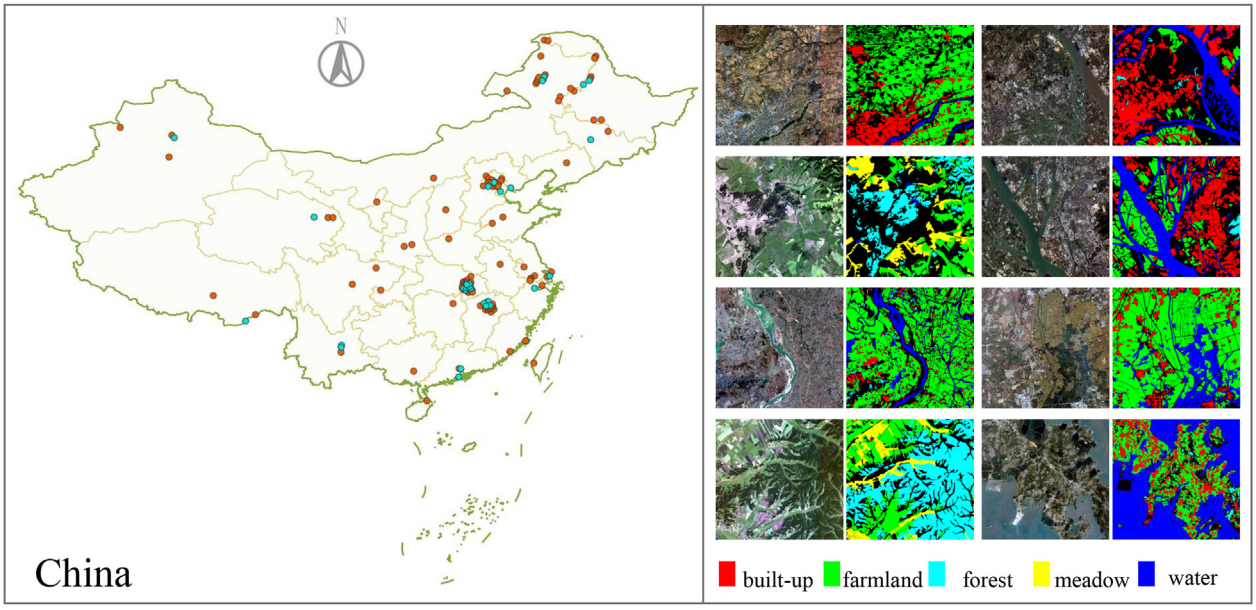

Figure 5. Left: Distribution of the geographical locations of images in GID. The large-scale classification set contains 120 training images and 30 validation images, which are marked with orange and cyan. Right: Examples of GF-2 images and their corresponding ground truth.

wide field of view, allowing the observation of detailed information over large areas. Since launched in 2014, GF-2 has been made use of for land-cover surveys, environmental monitoring, crop estimation, construction planning and other important applications.

4.1.2. Land-cover types

We refer to Chinese Land Use Classification Criteria (GB/T21010-2017) to determine a hierarchical category system. In the large-scale classification set of GID, 5 major categories are annotated: *built-up*, *farmland*, *forest*, *meadow*, and *water*, which are pixel-level labeled with five different colors: red, green, cyan, yellow, and blue, respectively. Areas not belonging to the above five categories and clutter regions are labeled as background, which is represented using black color. The fine land-cover classification set is made up of 15 sub-categories: *paddy field*, *irrigated land*, *dry cropland*, *garden land*, *arbor forest*, *shrub land*, *natural meadow*, *artificial meadow*, *industrial land*, *urban residential*, *rural residential*, *traffic land*, *river*, *lake*, and *pond*. Its training set contains 2000 patches per class, and validation images are labeled in pixel level. Examples of training and validation data with 15 categories are demonstrated in Fig. 6. The two parts of GID constitute a hierarchical classification system, and the affiliation of them is shown in Fig. 7.

4.1.3. Dataset properties

Widely distributed: GID contains 150 high-quality GF-2 images acquired from more than 60 different cities in China, which is shown in Fig. 5. It is widely distributed over the geographic areas covering more than 50,000 km^2. Due to the extensive geographical distribution, GID represents the distribution information of ground objects in different areas.

Multi-temporal: The images obtained at different time from the same location or overlapping areas are included in GID. For example, Fig. 8(a)-(b) are images acquired at Xiantao, Hubei Province on September 2, 2015 and June 14, 2016 respectively. Fig. 8(c)–(d) are images captured at Wuhan, Hubei Province on September 2, 2015 and June 14, 2016. Fig. 8(e)–(f) are images acquired around Nanchang, Jiangxi Province on August 12, 2016 and January 3, 2015. Fig. 8(g)–(h) are images acquired around Langfang, Hebei Province on August 27, 2016 and June 9, 2016. The spectral responses of the ground objects in the same area emerge in distinct differences due to seasonal changes.

Therefore, GID presents rich diversity of the ground objects in spectral response and morphological structure.

4.2. Gaofen-1, Jilin-1, Ziyuan-3, Sentinel-2A images, and Google Earth data

Newly acquired RS images may come from different sensors, the classification of the multi-source images is therefore of great significance. We also verify the effectiveness of our algorithm on HRRS images captured by other sensors, including Gaofen-1 (GF-1), Jilin-1 (JL-1), Ziyuan-3 (ZY-3), Sentinel-2A (ST-2A), and Google Earth platform data. The introduction of these images is as follows.

Gaofen-1 Satellite Images: GF-1 satellite configures with two PMS and four wide field of view (WFV) sensors. The resolution of PMS is 2 m pan/8 m MS, and the swath is 60 km. Two GF-1 multispectral images that were captured in Wuhan, Hubei Province on July 25, 2016, and in Jiujiang, Jiangxi Province on October 16, 2015 are employed in the experiments. We denote them as GF-1(1) and GF-1(2), as shown in Fig. 9.

Jilin-1 Satellite Images: The resolution of JL-1 satellite is 0.72 m pan/2.88 m MS, and it has only three bands of red, green, and blue. Two JL-1 images that were respectively captured around Ha Noi, Vietnam on June 11, 2016, and around Tokyo, Japan on June 3, 2016 are used in the experiments. We denote them as JL-1(1) and JL-1(2), as shown in Fig. 9.

Ziyuan-3 Satellite Images: ZY-3 satellite configures with three panchromatic time delay integration (TDI) charge coupled device (CCD) sensors and a multispectral scanner (MSS). The resolution of MSS is 5.8 m, and the swath is 52 km. Two ZY-3 multispectral images that were respectively captured in Fuzhou, Jiangxi Province on August 28, 2016, and in Shangrao, Jiangxi Province on August 28, 2016 are utilized in the experiments. We denote them as ZY-3 (1) and ZY-3 (2), as shown in Fig. 9.

Sentinel-2A Satellite Images: ST-2A satellite carries a PMS sensor that covers 13 spectral bands. The spatial resolution of visible and near-infrared bands is 10 m. ST-2A satellite has a very wide swath of 290 km. Two ST-2A multispectral images that were respectively captured around La Rochelle, France on October 21, 2018, and around Orleans, France on October 18, 2018 are used in the experiments. We denote them as ST-2A(1) and ST-2A(2), as shown in Fig. 9.

Google Earth Platform Data: To confirm the practicality of our

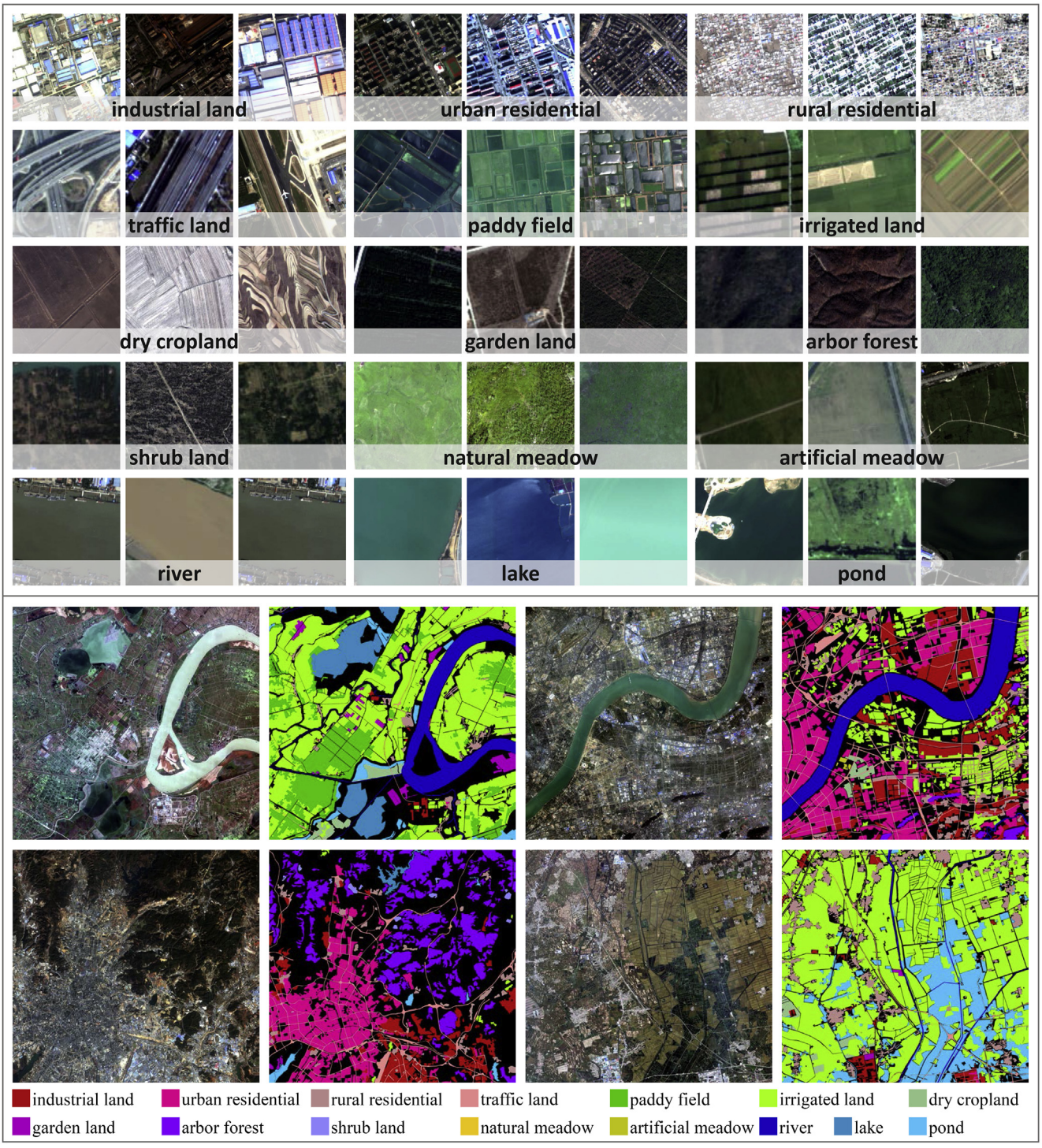

Fig. 6. Top: Training samples in the fine land-cover classification set. Three examples of each category are shown. There are 30,000 images within 15 classes. Bottom: Examples of validation images and their corresponding ground truth.

Table 1
Composition of GID dataset.

Set	Class	Training	Size	Validation
Large-scale Classification	5	120 GF-2 images	6800 × 7200	30 GF-2 images
Fine Land-cover Classification	15	30,000 patches	56 × 56, 112 × 112, 224 × 224	10 GF-2 images

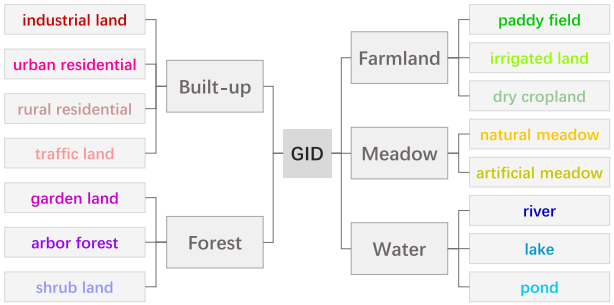

Fig. 7. Classification criteria for GID dataset.

algorithm for application, we conduct land-cover classification in Wuhan, Hubei province, China. Google Earth platform data captured on December 9, 2017 from Wuhan area are utilized. They have the resolution of 4.78 m and contains only three bands of red, green, and blue. We refer to these images as GE-WH, as shown in Fig. 10.

We demonstrate the heterogeneity of multi-source data in Fig. 11. Sub images with size of 400 × 400 are cropped from 6 different data. It can be observed that, besides the significant difference in spatial resolution, the morphology of *farmland* and *built-up* varies greatly due to the diversity of geographic location. In addition, because of seasonal changes, part of *farmland* is fallow and the other part is covered by crops. The heterogeneity of the source and target domains brings challenges to transfer learning.

5. Experimental results

We test our algorithm and analyse the experimental results in this section. Two types of land-cover classification issues are examined: 1) transferring deep models to classify HRRS images captured with the same sensor and under different conditions, 2) transferring deep models to classify multi-source HRRS images. For performance comparison, several object-based land-cover classification methods are utilized. The implementation details, comparison methods, and evaluation metrics are introduced in Section 5.1. Section 5.2 presents the experimental results of Gaofen-2 (GF-2) images. Section 5.3 tests the transferability of our algorithm on multi-source images.

5.1. Experimental setup

Pre-processing: For pre-processing, we re-quantize the responses of GF-2, GF-1, JL-1, ST-2A, and ZY-3 images to 8-bit with the optimized linear stretch function embedded in ENVI software. For GE-WH, we perform no pre-processing. In particular, to classify 3-band images (JL-1, GE-WH), we remove the near-infrared band of GF-2 images to train the model with the input size of 3 channels. When annotating label masks, we use the lasso tool in Adobe Photoshop to mark the areas of each land-cover category in the images.

Model Training: ResNet-50 models are pre-trained on the training images of GID, and our algorithm is tested on the validation images of GID as well as multi-source images. For the large-scale classification set, we train the models using patches with multiple scales to exploit the multi-scale contextual information. Patches of size 56 × 56, 112 × 112, and 224 × 224 are randomly sampled on each training image. If more than 80% pixels in a patch are covered by the same category, this patch is considered as a training sample. For each of the 5 categories, 10,000 samples are selected for each scale. Thus, a total of 10,000 × 3 × 5 = 150,000 patches are collected. Then, they are uniformly resized to 224 × 224 to pre-train a ResNet-50 model. For the fine land-cover classification set, we directly use the 30,000 image patches to fine-tune the ResNet-50 model pre-trained on 5 categories. In addition, image augmentation strategies (Krizhevsky et al., 2012) are adopted to avoid overfitting.

For multi-source images, we separately partition them into candidate patch sets with multi-scale sliding windows. We set different window sizes for different data according to their spatial resolution. For GF-1, the window sizes are 28 × 28, 56 × 56, and 112 × 112. For JL-1, the window sizes are 78 × 78, 156 × 156, and 312 × 312. For ZY-3, the window sizes are 38 × 38, 76 × 76, and 152 × 152. For ST-2A, the window sizes are 22 × 22, 44 × 44, and 88 × 88. For GE-WH, the window sizes are 46 × 46, 92 × 92, and 184 × 184. And the length of the stride is set as half the size of the window.

The parameters of ResNet-50 are initialized with ImageNet (Deng et al., 2009), and the softmax layer is initialized by Gaussian distribution. The last three bottlenecks and softmax layer of ResNet-50 are trained. In particular, when pre-training ResNet-50 for 4-band data, the new conv1 layer is initialized by Gaussian distribution, and is trained along with the last three bottlenecks and softmax layer. To train a deep model with multi-scale patches, patches with multiple scales are uniformly resized to 224 × 224. Considering the differences in feature

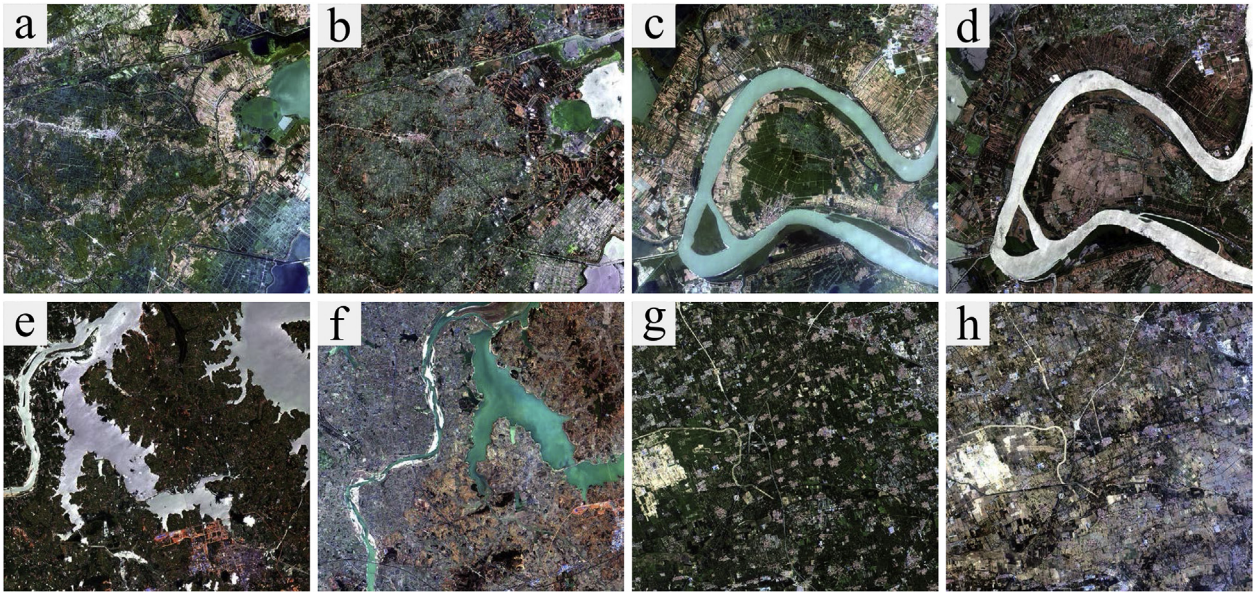

Fig. 8. Multi-temporal images captured from the same locations or overlapping areas.

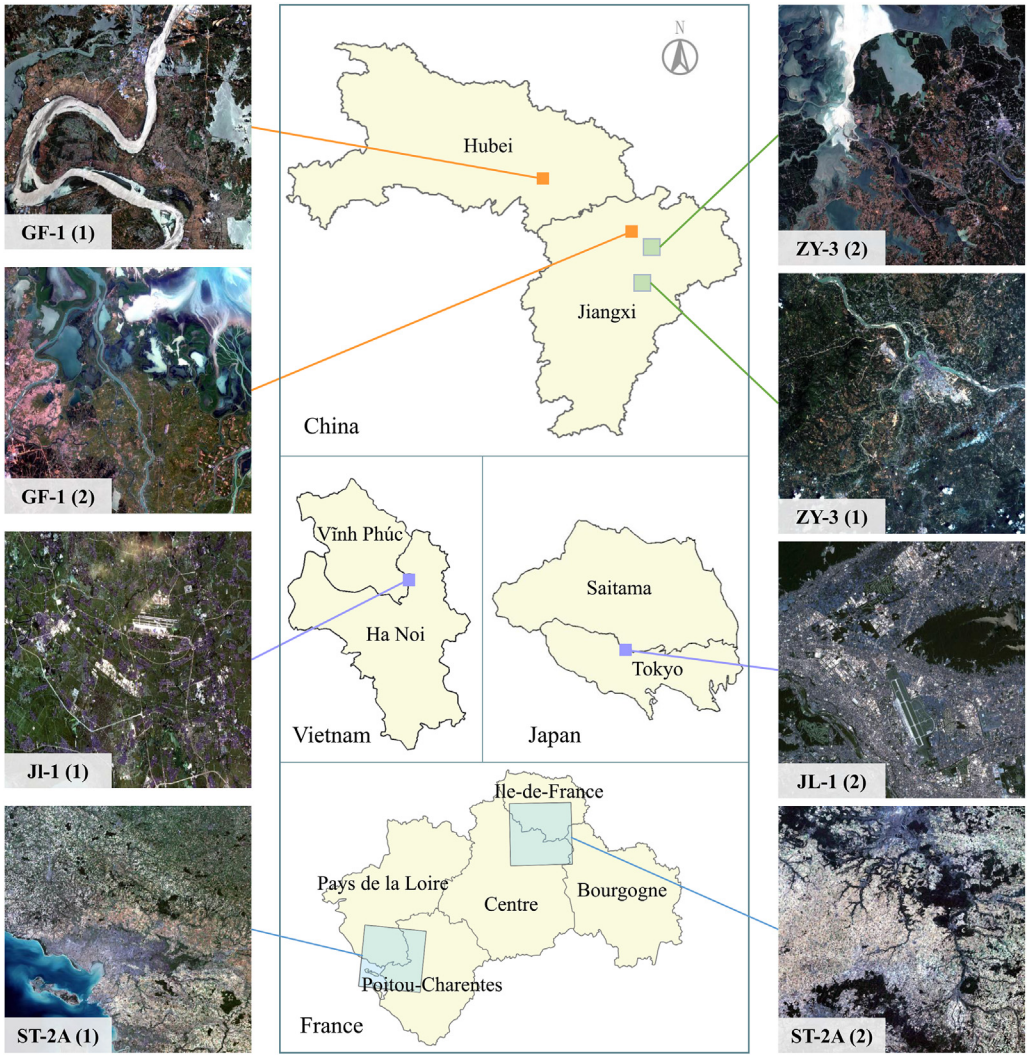

Fig. 9. GF-1, JL-1, ZY-3, ST-2A satellite images and their acquisition locations.

distribution between multi-source images, we select $\mathbf{U}_{tg}$ set and fine-tune ResNet-50 model for each target image separately to prevent different data sources from interfering with each other. The hyper-parameters for training are set as follows: batch size is 32, epoch number is 15, momentum value is 0.9, and initial learning rate is 0.1. When the error rate stops decreasing, we divide the learning rate by 10 and use the new value to update the parameters.

Comparison Methods: We compare our algorithm with several object-based classification methods. Selective search method is used to segment the image into homogeneous objects. Specifically, a set of four different features are exploited, including spectral feature, gray-level co-occurrence matrix (GLCM) (Haralick et al., 1973), differential morphological profiles (DMP) (Benediktsson et al., 2005), and local binary patterns (LBP) (Ojala et al., 2002). Moreover, we consider multi-feature fusion strategy, which aggregates the above features by normalization and vector concatenation. Maximum likelihood classification (MLC), random forest (RF), support vector machine (SVM), and multi-layer perceptron (MLP) are employed as classifiers.

The parameters of the comparison methods are set to the optimal values. The window size is set to 7×7 pixels for GLCM. The radius of the structural elements for DMP is set to 4 pixels. And for LBP, the filter size is 5×5 pixels. The number of trees for RF is 500. The kernel function of SVM is radial basis function (RBF) kernel. MLP has 4 hidden layers with 20 nodes per layer. The initial segmentation size is set to 400 for selective search. We randomly select 15,000 multi-scale patches from GID's training set to train comparison classifiers. After training, the classifiers are directly used to classify the target data.

Evaluation Metrics: To evaluate our algorithm, we assess the experimental results with Kappa coefficient (Kappa), overall accuracy (OA), and user's accuracy (Olofsson et al., 2014).

We test our algorithm on the validation set of GID and the multi-source data. The classification accuracy is assessed on all of the labeled pixels (except for the background) in the test images. Let P_{ab} denote the number of pixels of class a predicted to belong to class b, and let $t_a = \sum_b P_{ab}$ be the total number of pixels belong to class a, let $t_b = \sum_a P_{ab}$ be the total number of pixels predicted to class b. The metrics are defined as follow:

1) Kappa coefficient: Kappa is a statistic that measures the agreement between the prediction and the ground truth.

$$Kappa = \frac{P_o - P_c}{1 - P_c} \tag{8}$$

where

$$P_o = \frac{\sum_a P_{aa}}{\sum_a t_a} \tag{9}$$

$$P_c = \frac{\sum_k (\sum_b P_{kb} \cdot \sum_a P_{ak})}{\sum_a t_a \cdot \sum_a t_a} \tag{10}$$

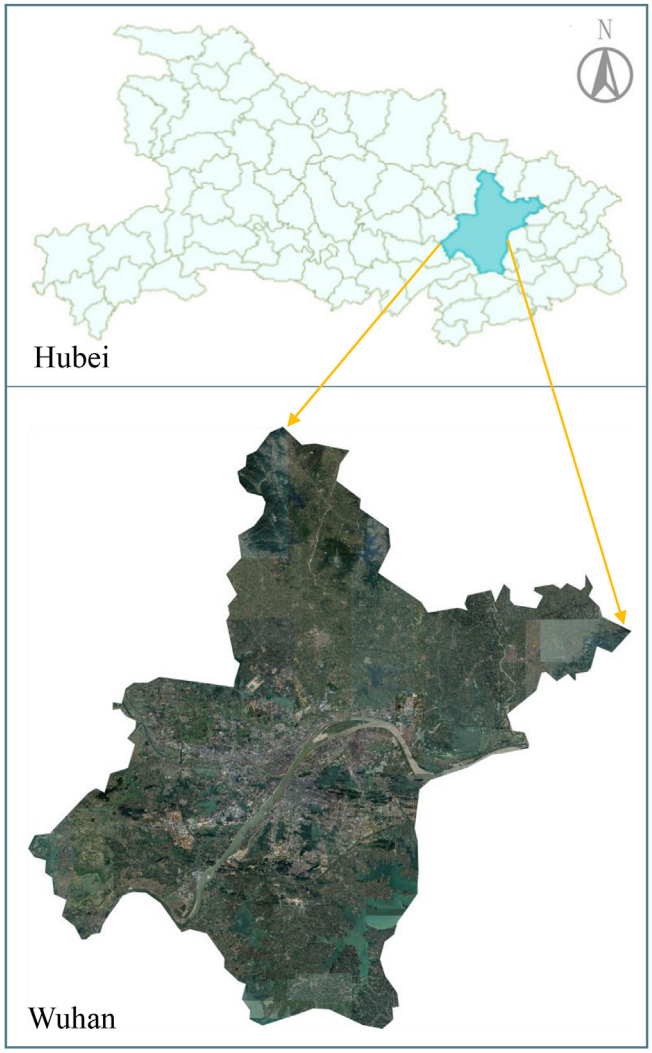

Fig. 10. Google Earth platform data in Wuhan, Hubei province, China.

where $k \in [1, K]$, and K is the number of categories.

2) Overall accuracy: OA is the percentage of correctly classified pixels and all pixels in the entire image.

$$OA = \frac{\sum_a P_{aa}}{\sum_a t_a} \tag{11}$$

3) User's accuracy: User's accuracy of class b is the proportion of correctly classified pixels in all pixels predicted to class b.

$$user's\ accuracy = \frac{P_{bb}}{t_b} \tag{12}$$

The values of Kappa, OA, and user's accuracy are in the range of 0 to 1, and the higher value indicates the better classification performance.

5.2. Experiments on Gaofen-2 images

To test the effectiveness of CNN model pre-trained on GID, which is denoted as PT-GID, we compare our approach with object-based methods. Although the acquisition location and time are diverse, the objects belong to the same class have similar spectral response in the images captured by the same sensor (*i.e.* GF-2 satellite). Hence we directly use ResNet-50 pre-trained on GID's training set to classify the validation images. Multi-scale information aggregation is exploited here, and the initial segmentation size is 400 pixels. The experimental results of our algorithm and the comparison methods are shown in Table 2.

It can be seen that PT-GID achieves the highest Kappa and OA of 0.924 and 96.28% on 5 classes, and of 0.605 and 70.04% on 15 classes. This experimental phenomenon shows the strong transferability of PT-GID. However, among the comparison methods, the optimal results are given by RF + Fusion, yielding Kappa and OA of 0.641 and 78.45% on 5 classes, and of 0.237 and 33.70% on 15 classes, respectively. This shows that conventional classifiers and features lack generalization capabilities for data shift.

To demonstrate the classification results more intuitively, we display the land-cover classification maps. Fig. 12(a)-(b) show a GF-2 image belonging to the large-scale classification set, which is obtained in Dongguan, Guangdong Province on January 23, 2015, and its ground truth. Fig. 12 (c)–(g) are the results generated by MLC + Fusion, RF + Fusion, SVM + Fusion, MLP + Fusion, and PT-GID. It can be seen that *farmland* is the most difficult class to be recognized in this image. *Meadow*, *forest* and *farmland* categories are seriously confused by the comparison methods. Compared to the comparison methods, our algorithm generates the best classification performance for *built-up* and *farmland* categories. However, our method misclassified some paddy fields into *water*.

Fig. 13(a) displays a GF-2 image belonging to the fine land-cover classification set, which is acquired in Wuhan, Hubei Province on April 11, 2016, and Fig. 13(b) is its ground truth. Fig. 13 (c)–(g) show the classification results produced by MLC + Fusion, RF + Fusion, SVM + Fusion, MLP + Fusion, and PT-GID. It can be observed that the comparison methods misclassify large areas of *urban residential* into *rural residential* and *industrial*. This is because the spectral responses of these classes are similar. When the labeling information of the target data is unavailable, it is difficult for the conventional feature extraction methods to represent the contextual properties of these ground objects. Whereas, our scheme generates smooth classification maps that close to the ground truth.

5.3. Experiments on multi-source images

This section focuses on validating the effectiveness of the proposed algorithm on multi-source data. Two deep models are utilized to classify each target image: 1) ResNet-50 pre-trained on the source domain data, 2) ResNet-50 fine-tuned with FT-$\mathbf{U}_{tg}$, which are denoted as PT-GID and FT-$\mathbf{U}_{tg}$, respectively. For learning transferable models, we set the parameters σ, δ, μ to 0.8, 5, 4000 for 5 classes, and to 0.7, 5, 2000 for 15 classes, respectively. Multi-scale information aggregation is utilized, and the initial segmentation size is 400 pixels. We compare our algorithm with object-based classification methods. The fusion of spectral feature, GLCM, DMP, and LBP is used to represent the characteristics of images. The classifiers used are MLC, RF, SVM, and MLP.

5.3.1. Results of Gaofen-1, Jilin-1, Ziyuan-3, and Sentinel-2A images

The experimental results of RS images captured by different sensors are shown in Table 3, where OA is used for performance assessment. The accuracy of our algorithm is obviously higher than the comparison methods. For all of the target images, the best OA values of both 5 classes and 15 classes are achieved by FT-$\mathbf{U}_{tg}$. These results show that the relevant samples selected from the target domain can strengthen the transferability of CNN models. Therefore, our sample selection and fine-tuning scheme is very effective for multi-source HRRS images.

When our algorithm is applied to images acquired by different sensors, FT-$\mathbf{U}_{tg}$ can boost the performance compared to PT-GID. Especially for JL-1(2), compared with the results of PT-GID, the OA of FT-$\mathbf{U}_{tg}$ increases by 16.57% on 5 classes, and by 4.94% on 15 classes. This experimental phenomenon indicates that, if the spectral responses of the target domain and the source domain are similar, the information learned from the source domain samples can benefit the interpretation

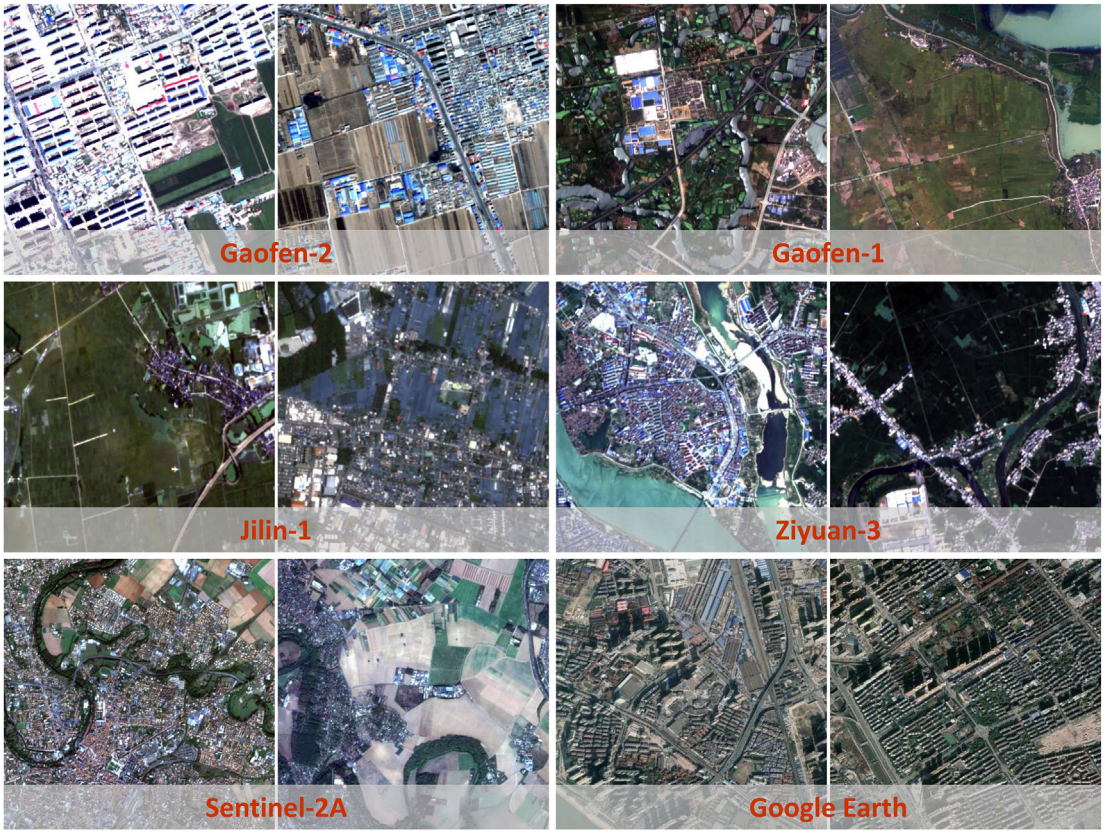

Fig. 11. The heterogeneity of multi-source data.

Table 2
Comparison of different land-cover classification methods on GID.

Methods	5 Classes							15 Classes	
	User's Accuracy (%)								
	Kappa	OA(%)	built-up	farmland	forest	meadow	water	Kappa	OA(%)
MLC + spectral	0.504	65.48	58.28	66.65	33.25	4.23	82.30	0.134	22.65
MLC + GLCM	0.389	56.22	66.65	73.56	20.55	0.05	63.50	0.092	16.07
MLC + DMP	0.512	65.82	59.05	68.04	33.91	4.19	82.53	0.028	23.26
MLC + LBP	0.195	37.31	39.77	52.39	25.08	0.96	66.82	0.084	16.28
MLC + Fusion	0.606	74.53	61.81	67.38	36.15	2.92	82.09	0.145	23.61
RF + spectral	0.526	68.73	55.48	67.25	34.76	2.22	84.25	0.164	23.79
RF + GLCM	0.426	61.83	65.73	70.71	18.84	0.68	67.54	0.119	19.05
RF + DMP	0.512	67.04	56.38	67.08	34.97	3.22	83.40	0.173	24.52
RF + LBP	0.365	58.76	46.14	59.42	23.19	1.40	68.30	0.063	11.60
RF + Fusion	0.641	78.45	62.61	71.12	36.10	3.94	84.29	0.237	33.70
SVM + spectral	0.103	46.11	54.68	42.86	41.82	1.02	62.11	0.024	22.72
SVM + GLCM	0.456	62.15	72.67	68.64	20.41	0.91	67.13	0.096	16.65
SVM + DMP	0.125	47.12	54.14	44.47	40.78	0.13	70.97	0.130	19.01
SVM + LBP	0.293	51.94	44.65	59.02	20.40	1.89	45.83	0.027	12.28
SVM + Fusion	0.488	66.88	61.28	72.27	23.01	2.26	54.18	0.148	23.92
MLP + spectral	0.442	60.93	52.42	58.26	29.21	1.36	84.11	0.082	14.19
MLP + GLCM	0.440	61.21	74.31	68.94	19.95	1.00	66.48	0.082	16.65
MLP + DMP	0.480	63.05	55.81	65.25	41.38	1.55	82.53	0.162	26.06
MLP + LBP	0.220	41.30	38.90	49.75	15.70	0.82	61.03	0.104	18.63
MLP + Fusion	0.616	75.81	58.69	72.40	32.86	2.67	83.99	0.199	30.57
PT-GID	**0.924**	**96.28**	**88.42**	**91.85**	**79.42**	**70.55**	**87.60**	**0.605**	**70.04**

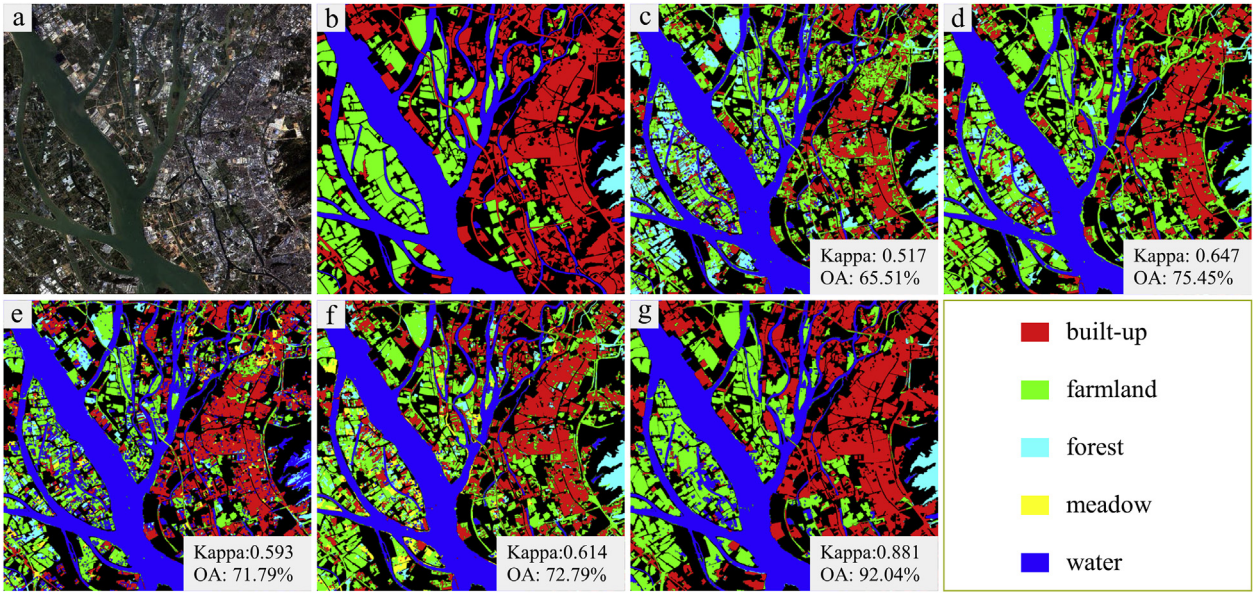

Fig. 12. Land-cover classification maps of a GF-2 image obtained in Dongguan, Guangdong Province on January 23, 2015. (a) The original image. (b) Ground truth. (c)–(g) Results of MLC + Fusion, RF + Fusion, SVM + Fusion, MLP + Fusion, and PT-GID.

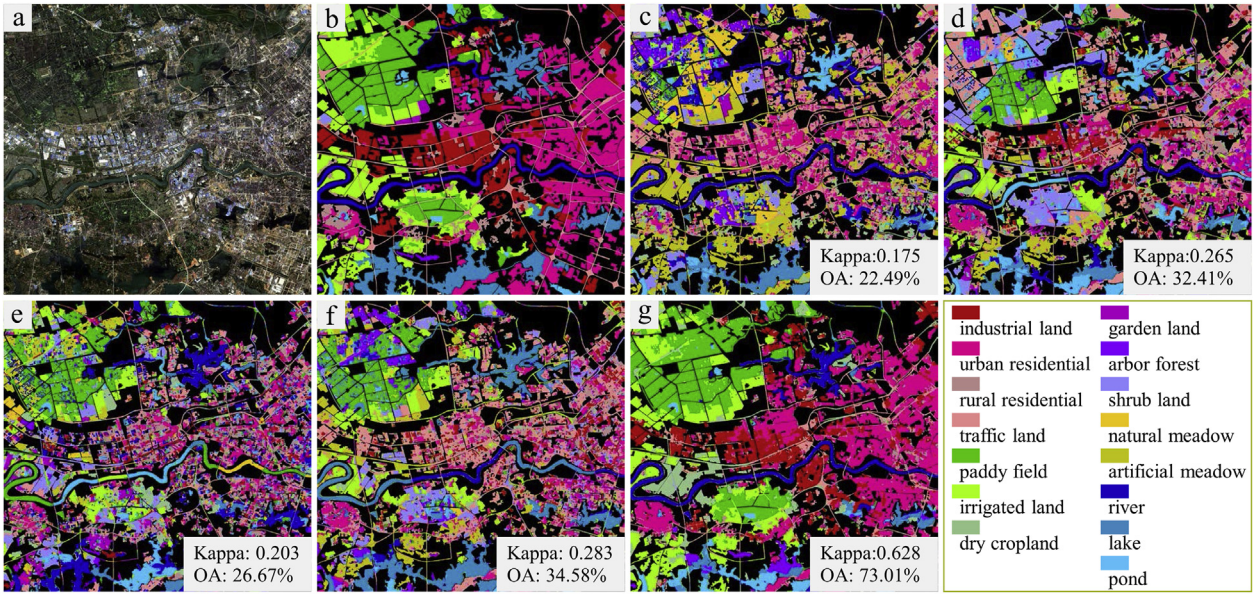

Fig. 13. Land-cover classification maps of a GF-2 image acquired in Wuhan, Hubei Province on April 11, 2016. (a) The original image. (b) Ground truth. (c)–(g) Results of MLC + Fusion, RF + Fusion, SVM + Fusion, MLP + Fusion, and PT-GID.

of the target domain. Conversely, if the spectral responses of the target and source domain are very different (*e.g.* obtained by different sensors), the supervision information of the source domain is not reliable for the target domain.

Fig. 14(a)-(b) show JL-1(2) and its corresponding ground truth with 15 categories. Fig. 14 (c)–(h) are the results of MLC + Fusion, RF + Fusion, SVM + Fusion, MLP + Fusion, PT-GID, and FT-$\mathbf{U}_{tg}$, respectively. The performance of the comparison methods is unsatisfactory. This is because the distributions of the target domain and the source domain are quite different, and the conventional classifiers do not have sufficient transferability. As shown in Fig. 14 (g), *urban residential* and *rural residential* are confused, while in Fig. 14 (h), *urban residential* is correctly classified. The results show that after model fine-tuning, CNN has learnt the distribution of the target domain, which proves that our relevant sample selection scheme can search reliable samples from the target domain.

Fig. 15(a)-(b) display a sub image sized 1000 × 1250 pixels cropped from ZY-3(1) and its ground truth with 15 categories. The classification results of PT-GID and FT-$\mathbf{U}_{tg}$ are demonstrated in Fig. 15(c)–(d), respectively. These results show the effect of fine-tuning schemes on the classification performance. Compared to Fig. 15(c), less *urban residential* area is mistakenly classified as *rural residential* in Fig. 15(d). This is because the spectral responses and the textures of these categories are similar. Whereas, transferred model can learn structural information and spatial relationship of the objects specific to the target domain. These experimental phenomena further validate the robustness and transferability of our approach for diverse HRRS images.

5.3.2. Results of Google Earth platform data in Wuhan

To test the performance of our method for large-scale land-cover classification, we conduct experiment on GE-WH, which is partially annotated with 5 classes for accuracy assessment, as shown in

Table 3
OA (%) of different methods on images captured by different sensors.

Image		MLC	RF	SVM	MLP	PT-GID	FT-$\mathbf{U}_{tg}$
GF-1(1)	5 classes	62.89	61.95	43.06	68.13	82.62	**89.84**
	15 classes	13.09	21.80	24.33	8.01	57.72	**60.07**
GF-1(2)	5 classes	84.87	84.69	55.45	78.99	92.40	**95.38**
	15 classes	14.08	33.46	35.11	25.04	74.90	**76.89**
JL-1(1)	5 classes	66.78	67.86	28.26	74.26	88.96	**90.37**
	15 classes	4.49	8.28	24.90	11.57	50.12	**52.86**
JL-1(2)	5 classes	80.67	61.14	65.25	67.31	72.51	**89.08**
	15 classes	6.89	23.60	17.47	23.87	66.15	**71.09**
ST-2A(1)	5 classes	49.40	76.96	58.10	81.06	97.08	**97.38**
	15 classes	8.30	35.08	8.67	21.11	61.14	**66.61**
ST-2A(2)	5 classes	55.34	53.37	45.24	81.58	94.46	**95.46**
	15 classes	9.28	11.22	11.09	2.15	28.96	**56.89**
ZY-3(1)	5 classes	63.37	69.24	58.38	62.18	85.62	**89.21**
	15 classes	24.34	34.38	48.18	10.52	80.97	**82.50**
ZY-3(2)	5 classes	55.91	72.94	66.74	64.22	92.75	**94.36**
	15 classes	17.27	13.86	30.18	19.32	49.47	**54.95**

Fig. 16(c). Table 4 displays the classification performance of the different methods. Overall, our method produces the most satisfactory results. The best result of the comparison methods is generated by RF + Fusion, only reaching Kappa and OA value of 0.490 and 61.43%, respectively. However, FT-$\mathbf{U}_{tg}$ achieves the overall highest Kappa and OA of 0.924 and 94.56%. Compared to PT-GID, FT-$\mathbf{U}_{tg}$ increases Kappa and OA values by 0.205 and 14.24%, respectively.

Fig. 16(a) shows the original GE-WH. Fig. 16(b) displays the intact classification map produced by FT-$\mathbf{U}_{tg}$. Fig. 16(c)–(d) are the partially annotated label mask and the classification result of FT-$\mathbf{U}_{tg}$ in the labeled areas, respectively. It can be seen that some areas in Fig. 16(d) are misclassified, for example, *water* area in the middle of Wuhan is identified as *farmland*. However, in the absence of labeling information of the target image, our method achieves Kappa and OA values that exceed 90%. These results show that our algorithm has the ability to cope with large-coverage HRRS images. Moreover, they also demonstrate the potential of our algorithm to interface with Google Earth and to be applied for practical land-cover mapping.

6. Sensitivity analysis

In the former section, the experimental results show the promising performance of the proposed method. However, some parameters have impact on the classification results. In this section, we analyse and discuss these factors through additional experiments, including analysis on patch size, segmentation method, and thresholds of transfer learning scheme.

6.1. Analysis on multi-scale information fusion

To verify the effectiveness of multi-scale information fusion strategy, we compare the classification performance of the single- and multi-scale methods on the dataset with 5 classes. Image patches of three different sizes are tested: 56 × 56, 112 × 112, and 224 × 224. For selective search segmentation, the initial segmentation size is set to 400. The classification accuracies obtained with different patch sizes are shown in Table 5.

For single scales, the optimal result is achieved by the smallest patch size 56 × 56. This is because our classification method is based on the image patches generated from non-overlapping grid partition, and all pixels in a patch are assigned with the same label. If the patch size is too large, the object details in the patches will be lost. Compared with the best results given by the single-scale approaches, multi-scale information fusion strategy attains the highest Kappa and OA of 0.924 and 96.28%, respectively. These results indicate that ground objects in HRRS images show great variations of contextual information in different scales. And combining image information of different scales helps to characterize the spatial distributions of the ground objects.

6.2. Analysis on segmentation

To analyse the influence of the segmentation scale, we test five different initial segmentation sizes of selective search method,

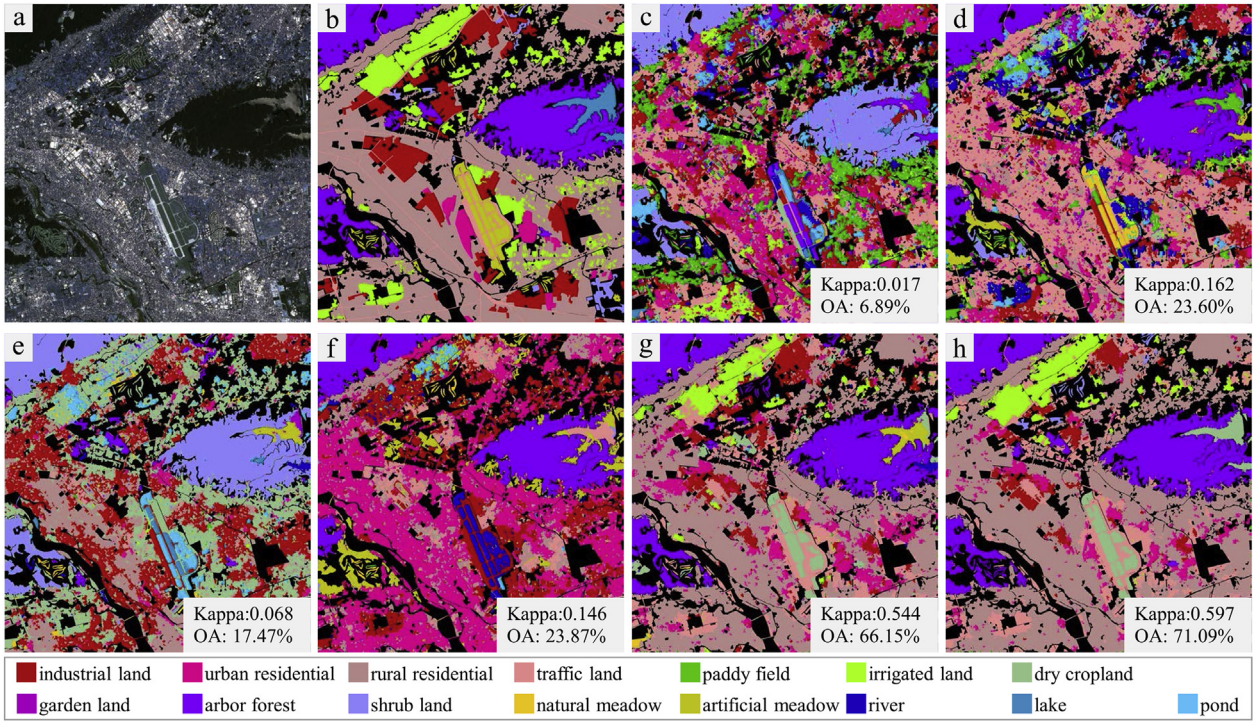

Fig. 14. Classification results of JL-1(2). (a) The original image. (b) Ground truth. (c)–(h) Results of MLC + Fusion, RF + Fusion, SVM + Fusion, MLP + Fusion, PT-GID, and FT-$\mathbf{U}_{tg}$.

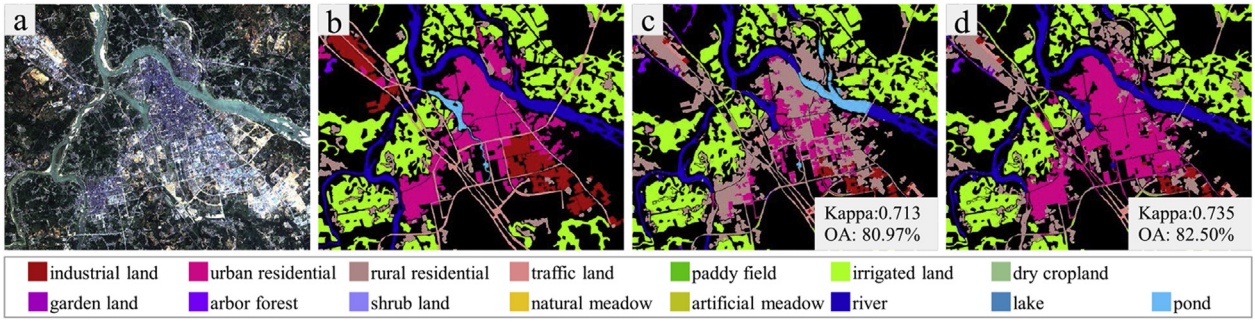

Fig. 15. Classification results of a sub image in ZY-3(1). (a) The original sub image. (b) Ground truth. (c)–(d) Results of PT-GID and FT-$\mathbf{U}_{tg}$.

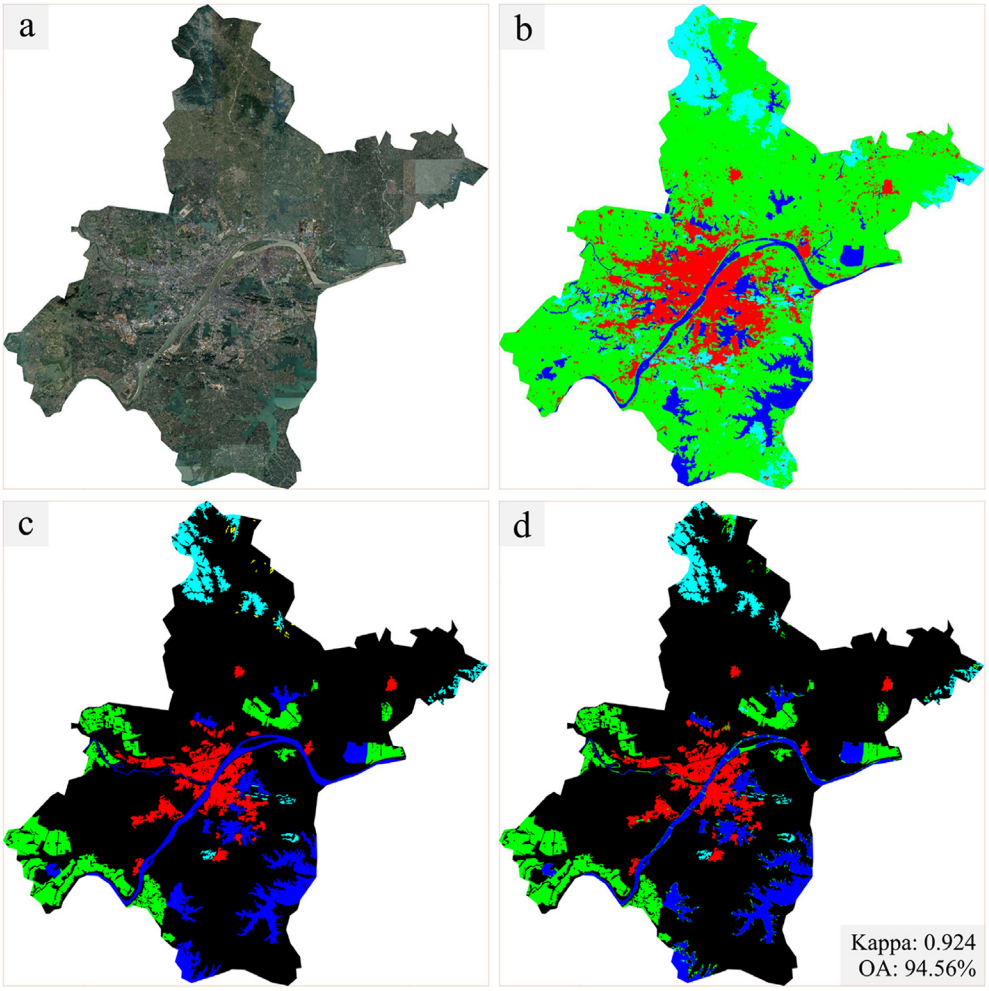

Fig. 16. (a) GE-WH image. (b) Classification map produced by FT-$\mathbf{U}_{tg}$. (c) Partially labeled ground truth. (d) Classification result of FT-$\mathbf{U}_{tg}$ in the labeled areas.

Table 4
Comparison of different land-cover classification methods on GE-WH.

Evaluation Metrics		MLC	RF	SVM	MLP	PT-GID	FT-U_{tg}
Kappa		0.132	0.490	0.253	0.424	0.719	0.924
OA(%)		27.05	61.43	41.40	58.07	80.32	94.56
User's Accuracy	Built-up	58.32	96.05	34.73	80.04	**99.38**	98.27
(%)	Farmland	77.51	62.27	61.00	63.49	62.65	**87.20**
	Forest	13.89	82.26	32.99	74.90	**97.34**	97.06
	Meadow	0.02	0.28	00.30	0.19	**72.69**	0
	Water	47.46	74.45	64.46	74.19	99.92	**99.97**

Table 5
Comparison of different patch sample scales.

Patch	Kappa	OA	User's Accuracy (%)				
Size		**(%)**	built-up	farmland	forest	meadow	water
56 × 56	0.915	95.41	80.02	89.47	76.83	**78.87**	89.12
112 × 112	0.892	94.36	71.33	87.10	76.31	77.84	**89.71**
224 × 224	0.880	92.67	72.13	86.41	75.95	71.54	87.75
Multi-scale	**0.924**	**96.28**	**88.42**	**91.85**	**79.42**	70.55	87.60

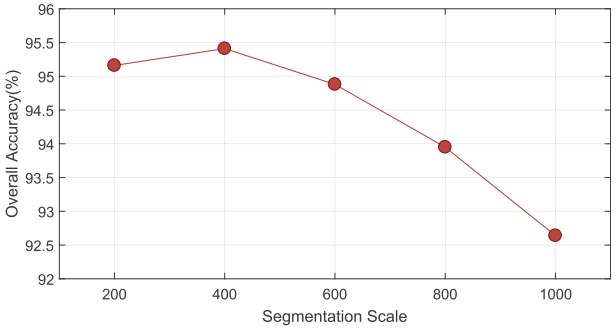

Fig. 17. Sensitivity analysis for the segmentation scale.

including 200, 400, 600, 800, and 1000, on the validation set with 5 classes. The image patches of size 56×56 are used for model pre-training and patch-wise classification. The mean OA values generated by each segmentation scale are illustrated in Fig. 17.

It can be seen that, the mean OA value first slightly increases and then continuously decreases with the increase of the segmentation scale. When the segmentation scale is set to 400, the best result is yielded. In general, for object-based land-cover classification, the most suitable segmentation scale depends on the spatial resolution of the RS image and the characteristics of the ground objects. If the image is over-segmented, more noise will be introduced into the classification results, and if the image is under-segmented, the classification map will lose a lot of details.

In addition, we compare the performance of selective search method with multi-resolution segmentation on 5 classes. We utilize the multi-resolution segmentation function embedded in eCognition, which is a professional and efficient software for RS image analysis, to generate segmentation maps. The scale parameter is set to 400, the same with that of selective search method. And the patch size is set to 56×56. The results are shown in Table 6.

Selective search provides slightly higher Kappa and OA value than the results of multi-resolution segmentation. On *farmland and water*, selective search behaves better, while multi-resolution segmentation generates higher accuracy on *built-up*, *forest* and *meadow* categories. The experimental results show that the performance of the two methods is comparable. Since this procedure is not the key factor to determine the accuracy of our approach, one may employ either selective search or multi-resolution segmentation method.

6.3. Analysis on transfer learning parameters

To investigate how the σ, δ, and μ parameters affect the proposed transfer learning scheme, we study on every parameter separately, i.e. at each time experimenting with one parameter and fixing the other two. Tests are conducted on GF-1, JL-1, ZY-3 and ST-2A images, and the mean OA value of these images are calculated to indicate the impact of the parameters. The relationship between the mean OA value and σ, δ, μ is presented in Fig. 18.

Parameter σ is varied from 0.5 to 0.9, with an interval of 0.1. δ and μ are set to 5 and 4000 for 5 classes, 5 and 2000 for 15 classes, respectively. σ is the threshold for pseudo-label assignment. In the transfer learning scheme, candidate patches with classification probability greater than σ are assigned the pseudo-labels. When σ is set to 0.8 and 0.7, the best performance is achieved on 5 classes and 15 classes respectively. This is because smaller σ value leads to unreliable pseudo-labels, while larger σ value may result in very few candidate samples, which are insufficient to train a well-performed deep model. Furthermore, when the classes are finer, the same category of different data sources have a larger difference in characteristics, and higher confidence may filter out valuable samples.

Parameter δ is varied from 1 to 9, with an interval of 2. σ and μ are set to 0.8 and 4000 for 5 classes, 0.7 and 2000 for 15 classes, respectively. δ is utilized to filter out the pseudo-labels that do not match the annotation information of the source domain. If the true-labels of the top δ retrieved samples are identical to the pseudo-label of the query patch, this query patch is retained, otherwise it is removed. The δ with value of 5 provides the highest mean OA value. The reason for this phenomenon is that smaller δ value can not guarantee the consistency of the pseudo-labels and the true-labels, while larger δ is too strict to extract sufficient candidate patches from the target domain.

Parameter μ is varied from 1000 to 5000, with an interval of 1000 on 5 classes, and from 500 to 2500, with an interval of 500 on 15 classes. σ and δ are set to 0.8 and 5 for 5 classes, 0.7 and 5 for 15 classes, respectively. μ is employed to limit the number of target domain samples. It can be observed that, on 5 classes, at the beginning, the classification accuracy increases with the increase of the μ value, and when μ is greater than 2000, the mean OA value rises more gently. The highest accuracy is achieved when μ is equal to 4000. This shows that μ of 2000 is enough to select sufficient samples for training a well-performed model, although it does not provide the best performance. And when μ is larger than 4000, there is redundancy between the selected target domain samples, which causes the deep model to be biased toward the redundant information. In addition, excessive training samples distinctly reduce the model training efficiency. On 15 categories, our approach behaves best when μ is equal to 1500. This is because the 5 major categories are subdivided in the fine land-cover classification set, and each sub-class contains fewer samples.

The high diversity of the samples in the source domain (i.e. GID) enables the pre-trained CNN to be discriminating. Therefore, diverse target samples can be correctly identified and selected by the models. To ensure that sufficient samples are extracted for model fine-tuning, we have not adopted diverse criterion to deal with data redundancy. Due to the large volume of CNN's parameters and the variety of target samples, appropriate data redundancy does not have a significant impact on model performance. In the future research, we are interested in further investigating how to address the redundancy problem with diverse criterions.

7. Discussion

Land-cover classification is closely tied to the ecological condition of the Earth's surface and have significant implications for global ecosystem health, water quality, and sustainable land management. Most studies on large-scale land-cover classification generally use the low-/medium-spatial resolution RS images, however, due to the lack of spatial information, these images are insufficient for detailed mapping for high heterogeneous areas (Hu et al., 2013). By contrast, high-spatial resolution images provide rich texture, shape, and spatial distribution

Table 6
Comparison of selective search and multi-resolution segmentation.

Segmentation	Kappa	OA	User's Accuracy (%)				
Method		**(%)**	built-up	farmland	forest	meadow	water
Selective Search Method	**0.915**	**95.41**	80.02	**89.47**	76.83	78.87	**89.12**
Multi-resolution Segmentation	0.907	94.62	**82.35**	87.50	**81.89**	**83.94**	88.01

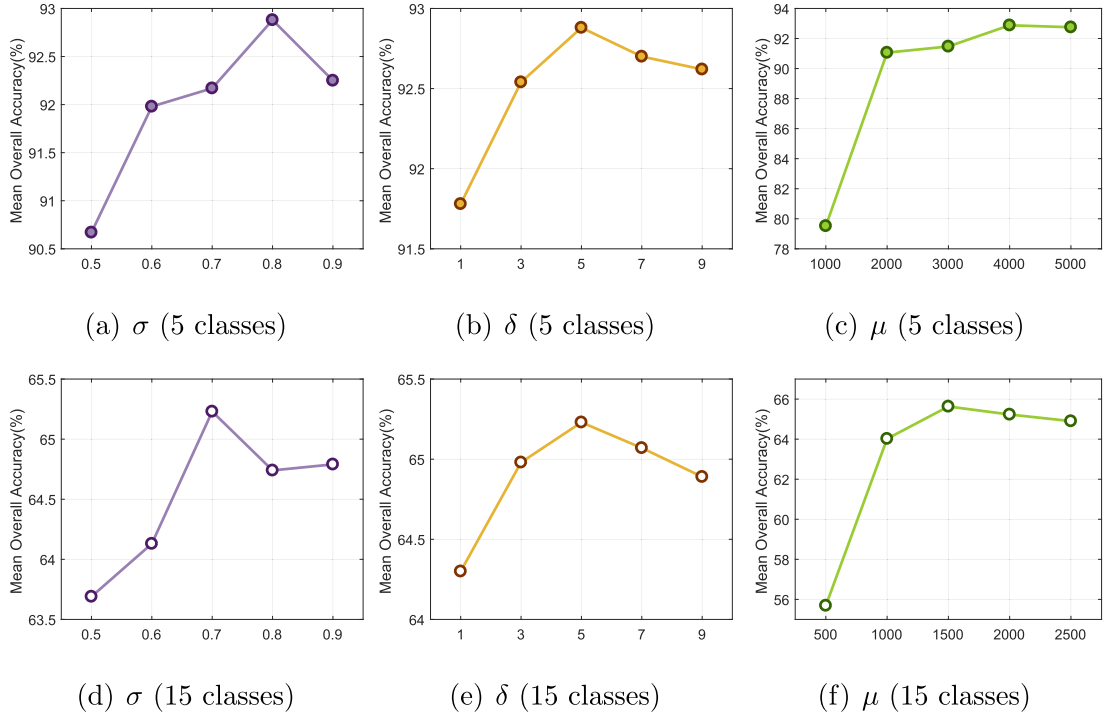

Fig. 18. Sensitivity analysis for σ, δ, and μ parameters.

information of ground objects, which contribute significantly to distinguish categories with similar spectral characteristics. Nevertheless, because of the narrow spatial coverage and high economic costs, high-spatial resolution images are commonly employed in land-cover classification for some specific small regions. In addition, even if a mass of HRRS images are available, in the case where accurate annotation is difficult to quickly obtain, the classifiers will have insufficient adaptability to data and connot be used in practical applications. Therefore, it is highly demanded to develop robust and transferable algorithms to achieve high-precision land-cover classification at large-scale.

Considering that the structure and spatial relationship of the ground objects do not change with the acquisition conditions, we relate the intrinsic characteristics of the objects in the multi-source data through the high-level deep features of the images. Our approach is inspired by pseudo-label assignment (Wu and Yap, 2006; Lee, 2013) and joint fine-tuning (Xue et al., 2007; Ge and Yu, 2017) methods. However, compared with these two methods, our semi-supervised algorithm does not need any annotation information of the target domain, and the reliability of our method is improved by the constraint of feature similarity. Our approach achieves complete automatic classification for the unlabeled target images, providing new possibilities for real-time land-cover classification applications. Although our approach proves to be effective in experiments and presents remarkable performance on 5 classes, for more complex categories, the classification accuracy still has room for improvement (see Table 2). We use confusion matrices to analyse the behavior of our method on different fine categories.

As shown in Fig. 19(a)–(c) are the confusion matrices of classification results of PT-GID on GID, PT-GID on multi-source images, and FT-$\mathbf{U}_{tg}$ on multi-source images, respectively. It can be seen that the fine categories of the same major class are seriously confused, for instance,

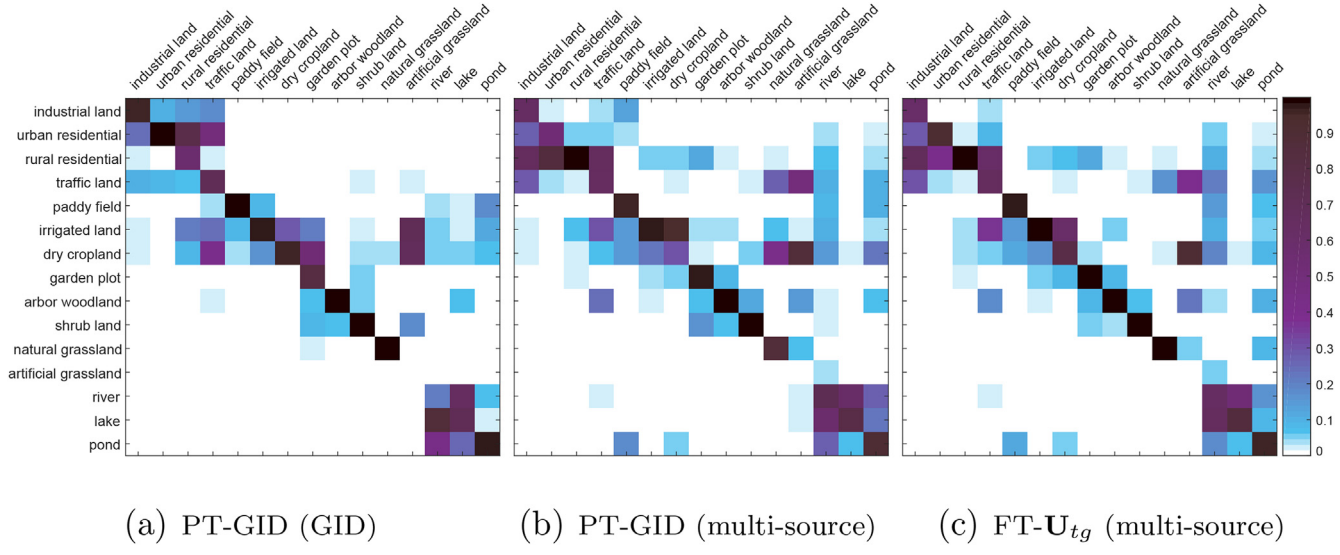

Fig. 19. Confusion matrices of land-cover classification results on GID and multi-source images with 15 categories.

river, *lake*, and *pond* are misclassified. In addition, *farmland* and *meadow* are severely confused. This is because that they are difficult to be recognized based on spectral, texture and structure information. Furthermore, the shapes of *water* and *farmland* areas are diverse, and they present different appearances in different seasons. In this case, multi-temporal analysis may provide the information that distinguishes between these categories (Shao et al., 2016; Sheng et al., 2016; Kussul et al., 2015; Vuolo et al., 2018). However, multi-temporal analysis generally requires explicit labels of the samples as supervised information for multi-temporal feature learning. Currently, our method cannot guarantee that the pseudo-labels of the selected samples are sufficiently accurate. Therefore, in our future work, it is of great interest to us to further study multi-supervised multi-temporal analysis.

8. Conclusion

We present a land-cover classification algorithm that can be applied to classify multi-source HRRS images. The proposed algorithm has the following attractive properties: 1) it automatically selects training samples from the target domain based on the contextual information extracted from deep model. In consequence, it does not require new manual annotation or algorithm adjustment when being applied to multi-source images. 2) it uses multi-scale contextual information for classification. Therefore, the spatial distributions of the objects are characterized, and the transferability of deep models for RS images with different resolutions is strengthen. 3) it combines patch-wise classification and hierarchical segmentation. The accurate category and boundary information is simultaneously obtained, and the classification noise is reduced in the classification map.

In addition, we constructed a large-scale land-cover dataset, *i.e.* GID, with 150 high-resolution GF-2 images. It well represents the true distribution of land-cover categories and can be used to train CNN model specific to RS data. We conduct experiments on multi-source HRRS images, including Gaofen-2 (GF-2) images in GID, coupled with Gaofen-1 (GF-1), Jilin-1 (JL-1), Ziyuan-3 (ZY-3), Sentinel-2A (ST-2A), and Google Earth (GE-WH) platform data. The proposed algorithm shows encouraging classification performance. To benefit researchers, GID have been provided online at http://captain.whu.edu.cn/GID/.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grants 61922065, 61771350, 61871299 and 41820104006, in part by the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Science LSU-SJLY-2017-01, the Outstanding Youth Project of Hubei Province under Contract 2017CFA037.

References

Ardila, J.P., Tolpekin, V.A., Bijker, W., Stein, A., 2011. Markov random field-based super-resolution mapping for identification of urban trees in vhr images. ISPRS J. Photogrammetry Remote Sens. 66 (6), 762–775.

Audebert, N., Le Saux, B., Lefevre, S., 2016. How useful is region-based classification of remote sensing images in a deep learning framework? In: IEEE International Geoscience and Remote Sensing Symposium, pp. 5091–5094.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R., 2005. Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. 43 (3), 480–491.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. ISPRS J. Photogrammetry Remote Sens. 58 (3–4), 239–258.

Blaschke, T., 2001. What's wrong with pixels? some recent developments interfacing remote sensing and gis. GeoBIT/GIS 6, 12–17.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogrammetry Remote Sens. 65 (1), 2–16.

Bruzzone, L., Carlin, L., 2006. A multilevel context-based system for classification of very high spatial resolution images. IEEE Trans. Geosci. Remote Sens. 44 (9), 2587–2600.

Bruzzone, L., Chi, M., Marconcini, M., 2006. A novel transductive svm for semisupervised classification of remote-sensing images. IEEE Trans. Geosci. Remote Sens. 44 (11), 3363–3373.

Bruzzone, L., Persello, C., 2009. A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. IEEE Trans. Geosci. Remote Sens. 47 (9), 3180–3191.

Burnett, C., Blaschke, T., 2003. A multi-scale segmentation/object relationship modelling methodology for landscape analysis. Ecol. Model. 168 (3), 233–249.

Casals-Carrasco, P., Kubo, S., Madhavan, B.B., 2000. Application of spectral mixture analysis for terrain evaluation studies. Int. J. Remote Sens. 21 (16), 3039–3055.

Chakraborty, S., Balasubramanian, V., Sun, Q., Panchanathan, S., Ye, J., 2015. Active batch selection via convex relaxations with guaranteed solution bounds. IEEE Trans. Pattern Anal. Mach. Intell. 37 (10), 1945–1958.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Demir, B., Bovolo, F., Bruzzone, L., 2012. Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification. IEEE Trans. Geosci. Remote Sens. 50 (5), 1930–1941.

Demir, B., Minello, L., Bruzzone, L., 2014. Definition of effective training sets for supervised classification of remote sensing images by a novel cost-sensitive active learning method. IEEE Trans. Geosci. Remote Sens. 52 (2), 1272–1284.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Duro, D.C., Franklin, S.E., Dubé, M.G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery. Remote Sens. Environ. 118, 259–272.

Fauvel, M., Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Tilton, J.C., 2013. Advances in spectral-spatial classification of hyperspectral images. Proc. IEEE 101 (3), 652–675.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Int. J. Comput. Vis. 59 (2), 167–181.

Ge, W., Yu, Y., 2017. Borrowing treasures from the wealthy: deep transfer learning through selective joint fine-tuning. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 6.

Gerke, M., Rottensteiner, F., Wegner, J.D., Sohn, G., 2014. Isprs semantic labeling contest. In: Photogrammetric Computer Vision, . http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

Giada, S., De Groeve, T., Ehrlich, D., Soille, P., 2003. Information extraction from very high resolution satellite imagery over lukole refugee camp, Tanzania. Int. J. Remote Sens. 24 (22), 4251–4266.

Gómez-Chova, L., Camps-Valls, G., Munoz-Mari, J., Calpe, J., 2008. Semisupervised image classification with laplacian support vector machines. IEEE Geosci. Remote Sens. Lett. 5 (3), 336–340.

Gong, P., Marceau, D.J., Howarth, P.J., 1992. A comparison of spatial feature extraction algorithms for land-use classification with spot hrv data. Remote Sens. Environ. 40 (2), 137–151.

Haralick, R.M., Shanmugam, K., et al., 1973. Textural features for image classification. IEEE Trans. on Systems, Man, and Cybernetics (6), 610–621.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015a. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7 (11), 14680–14707.

Hu, F., Xia, G.-S., Hu, J., Zhong, Y., Xu, K., 2016. Fast binary coding for the scene classification of high-resolution remote sensing imagery. Remote Sens. 8 (7), 555.

Hu, F., Xia, G.-S., Zhang, L., 2017. Deep sparse representations for land-use scene classification in remote sensing images. In: IEEE International Conference on Signal Processing, pp. 192–197.

Hu, J., Xia, G.-S., Hu, F., Zhang, L., 2015b. A comparative study of sampling analysis in the scene classification of optical high-spatial resolution remote sensing imagery. Remote Sens. 7 (11), 14988–15013.

Hu, Q., Wu, W., Xia, T., Yu, Q., Yang, P., Li, Z., Song, Q., 2013. Exploring the use of google earth imagery and object-based methods in land use/cover mapping. Remote Sens. 5 (11), 6026–6042.

Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 214, 73–86.

Izquierdo-Verdiguier, E., Laparra, V., Gomez-Chova, L., Camps-Valls, G., 2013. Encoding invariances in remote sensing image classification with svm. IEEE Geosci. Remote Sens. Lett. 10 (5), 981–985.

Jensen, J.R., Lulla, K., 1986. Introductory digital image processing: a remote sensing perspective. Geocarto Int. 2 (1) 65–65.

Jiang, T.-B., Xia, G.-S., Lu, Q.-K., Shen, W.-M., Jul, 2017. Retrieving aerial scene images with learned deep image-sketch features. J. Comput. Sci. Technol. 32 (4), 726–737.

Jun, G., Ghosh, J., 2011. Spatially adaptive classification of land cover with remote sensing data. IEEE Trans. Geosci. Remote Sens. 49 (7), 2662–2673.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp. 1097–1105.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci. Remote Sens. Lett. 14 (5), 778–782.

Kussul, N., Skakun, S., Shelestov, A., Lavreniuk, M., Yailymov, B., Kussul, O., 2015. Regional scale crop mapping using multi-temporal satellite imagery. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 40 (7), 45.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.
Lee, D.-H., 2013. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML. vol. 3. pp. 2.
Li, X., Zhang, L., Du, B., Zhang, L., Shi, Q., 2017. Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification. IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing 10 (5), 2022–2035.
Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W., Munteanu, A., 2017. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. Remote Sens. 9 (6), 522.
Lu, Q., Huang, X., Li, J., Zhang, L., 2016. A novel mrf-based multifeature fusion for classification of remote sensing images. IEEE Geosci. Remote Sens. Lett. 13 (4), 515–519.
Lu, Q., Ma, Y., Xia, G.-S., 2017. Active learning for training sample selection in remote sensing image classification using spatial information. Remote Sensing Letters 8 (12), 1210–1219.
Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017a. A review of supervised object-based land-cover image classification. ISPRS J. Photogrammetry Remote Sens. 130, 277–293.
Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017b. A review of supervised object-based land-cover image classification. ISPRS J. Photogrammetry Remote Sens. 130, 277–293.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. High-resolution Semantic Labeling with Convolutional Neural Networks. arXiv preprint arXiv:1611.01962.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017a. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: IEEE International Symposium on Geoscience and Remote Sensing, pp. 3226–3229.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017b. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55 (2), 645–657.
Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using imagenet pretrained networks. IEEE Geosci. Remote Sens. Lett. 13 (1), 105–109.
Matasci, G., Volpi, M., Kanevski, M., Bruzzone, L., Tuia, D., 2015. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 53 (7), 3550–3564.
Mathieu, R., Freeman, C., Aryal, J., 2007. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. Landsc. Urban Plan. 81 (3), 179–192.
Mattyus, G., Wang, S., Fidler, S., Urtasun, R., 2015. Enhancing road maps by parsing aerial images around the world. In: IEEE International Conference on Computer Vision, pp. 1689–1697.
Mnih, V., 2013. Machine Learning for Aerial Image Labeling. University of Toronto, Canada Ph.D. thesis.
Moser, G., Serpico, S.B., Benediktsson, J.A., 2013. Land-cover mapping by markov modeling of spatial–contextual information in very-high-resolution remote sensing images. Proc. IEEE 101 (3), 631–651.
Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sens. Environ. 115 (5), 1145–1161.
Napoletano, P., 2018. Visual descriptors for content-based retrieval of remote-sensing images. Int. J. Remote Sens. 39 (5), 1343–1376.
Ojala, T., Pietikäinen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 971–987.
Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.
Othman, E., Bazi, Y., Melgani, F., Alhichri, H., Alajlan, N., Zuair, M., 2017. Domain adaptation network for cross-scene classification. IEEE Trans. Geosci. Remote Sens. 55 (8), 4441–4456.
Ozdarici-Ok, A., Ok, A.O., Schindler, K., 2015. Mapping of agricultural crops from single high-resolution multispectral imagesdata-driven smoothing vs. parcel-based smoothing. Remote Sens. 7 (5), 5611–5638.
Pacifici, F., Chini, M., Emery, W.J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113 (6), 1276–1292.
Paisitkriangkrai, S., Sherrah, J., Janney, P., Hengel, V.-D., et al., 2015. Effective semantic pixel labelling with convolutional networks and conditional random fields. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 36–43.
Paisitkriangkrai, S., Sherrah, J., Janney, P., van den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing 9 (7), 2868–2881.
Persello, C., Bruzzone, L., 2012. Active learning for domain adaptation in the supervised classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. 50 (11), 4468–4483.
Persello, C., Bruzzone, L., 2014. Active and semisupervised learning for the classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. 52 (11), 6937–6956.
Persello, C., Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in vhr images. IEEE Geosci. Remote Sens. Lett. 14 (12), 2325–2329.
Shao, W., Yang, W., Xia, G.-S., 2013. Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification. Int. J. Remote Sens. 34 (23), 8588–8602.
Shao, Y., Lunetta, R.S., Wheeler, B., Iiames, J.S., Campbell, J.B., 2016. An evaluation of time-series smoothing algorithms for land-cover classifications using modis-ndvi multi-temporal data. Remote Sens. Environ. 174, 258–265.
Sheng, Y., Song, C., Wang, J., Lyons, E.A., Knox, B.R., Cox, J.S., Gao, F., 2016. Representative lake water extent mapping at continental scales using multi-temporal landsat-8 imagery. Remote Sens. Environ. 185, 129–141.
Sherrah, J., 2016. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. arXiv preprint arXiv:1606.02585.
Shi, H., Chen, L., Bi, F.-k., Chen, H., Yu, Y., 2015. Accurate urban area detection in remote sensing images. IEEE Geosci. Remote Sens. Lett. 12 (9), 1948–1952.
Tarabalka, Y., Chanussot, J., Benediktsson, J.A., 2010a. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics) 40 (5), 1267–1279.
Tarabalka, Y., Fauvel, M., Chanussot, J., Benediktsson, J.A., 2010b. Svm-and mrf-based method for accurate classification of hyperspectral images. IEEE Geosci. Remote Sens. Lett. 7 (4), 736–740.
Tong, X.-Y., Lu, Q., Xia, G.-S., Zhang, L., 2018. Large-scale Land Cover Classification in Gaofen-2 Satellite Imagery. arXiv preprint arXiv:1806.00901.
Tuia, D., Camps-Valls, G., 2016. Kernel manifold alignment for domain adaptation. Public Library of Science 11 (2), e0148655.
Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: an overview of recent advances. IEEE Geoscience and Remote Sensing Magazine 4 (2), 41–57.
Tuia, D., Ratle, F., Pozdnoukhov, A., Camps-Valls, G., 2010. Multisource composite kernels for urban-image classification. IEEE Geosci. Remote Sens. Lett. 7 (1), 88–92.
Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. Int. J. Comput. Vis. 104 (2), 154–171.
Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55 (2), 881–893.
Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., Ng, W.-T., 2018. How much does multi-temporal sentinel-2 data improve crop type classification? Int. J. Appl. Earth Obs. Geoinf. 72, 122–130.
Wu, K., Yap, K.-H., 2006. Fuzzy svm for content-based image retrieval: a pseudo-label support vector machine framework. IEEE Comput. Intell. Mag. 1 (2), 10–16.
Xia, G., Delon, J., Gousseau, Y., 2010a. Shape-based invariant texture indexing. Int. J. Comput. Vis. 88 (3), 382–403.
Xia, G., Liu, G., Bai, X., Zhang, L., 2017a. Texture characterization using shape co-occurrence patterns. IEEE Trans. Image Process. 26 (10), 5005–5018.
Xia, G., Tong, X., Hu, F., Zhong, Y., Datcu, M., Zhang, L., 2017b. Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation. CoRR Abs/1707. , 07321.
Xia, G., Yang, W., Delon, J., Gousseau, Y., Sun, H.P.H., Maitre, H., 2010b. Structural High-Resolution Satellite Image Indexing. in: ISPRS TC VII Symposium C 100 Years ISPRS, Vienna, Austria.
Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. Dota: a large-scale dataset for object detection in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition.
Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017c. Aid: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55 (7), 3965–3981.
Xue, Y., Liao, X., Carin, L., Krishnapuram, B., 2007. Multi-task learning for classification with dirichlet process priors. J. Mach. Learn. Res. 8 (Jan), 35–63.
Yan, G., Mas, J.-F., Maathuis, B., Xiangmin, Z., Van Dijk, P., 2006. Comparison of pixel-based and object-oriented image classification approachesa case study in a coal fire area, wuda, inner Mongolia, China. Int. J. Remote Sens. 27 (18), 4039–4055.
Yang, H.L., Crawford, M.M., 2016. Domain adaptation with preservation of manifold geometry for hyperspectral image classification. IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing 9 (2), 543–555.
Yang, W., Yin, X., Xia, G.-S., 2015. Learning high-level features for satellite image classification with limited labeled samples. IEEE Trans. Geosci. Remote Sens. 53 (8), 4472–4482.
Yu, H., Yang, W., Xia, G.-S., Liu, G., 2016. A color-texture-structure descriptor for high-resolution satellite image classification. Remote Sens. 8 (3), 259.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, pp. 818–833.
Zhang, C., Kovacs, J.M., 2012. The application of small unmanned aerial systems for precision agriculture: a review. Precis. Agric. 13 (6), 693–712.
Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., Atkinson, P.M., 2018a. A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification. ISPRS J. Photogrammetry Remote Sens. 140, 133–144.
Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018b. An object-based convolutional neural network (ocnn) for urban land use classification. Remote Sens. Environ. 216, 57–70.
Zhang, L., Huang, X., Huang, B., Li, P., 2006. A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 44 (10), 2950–2961.
Zhao, B., Huang, B., Zhong, Y., 2017. Transfer learning with fully pretrained deep convolution networks for land-use classification. IEEE Geosci. Remote Sens. Lett. 14 (9), 1436–1440.
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 Trans. Geosci. Remote Sens. 54 (4), 2108–2123.
Zhao, W., Du, S., 2016. Learning multiscale and deep representations for classifying

remotely sensed imagery. ISPRS J. Photogrammetry Remote Sens. 113, 155–165.
Zhao, W., Guo, Z., Yue, J., Zhang, X., Luo, L., 2015. On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. Int. J. Remote Sens. 36 (13), 3368–3379.
Zhong, Y., Wu, S., Zhao, B., 2017. Scene semantic understanding based on the spatial context relations of multiple objects. Remote Sens. 9 (10), 1030.
Zhong, Y., Zhao, J., Zhang, L., 2014. A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 52 (11), 7023–7037.
Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5 (4), 8–36.