This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Semi-MCNN:A Semi-supervised Multi-CNN Ensemble Learning Method for Urban Land Cover Classification Using Sub-meter HRRS Images

Runyu Fan, Ruyi Feng, Lizhe Wang, Jining Yan, and Xiaohan Zhang

***Abstract*—Sub-meter high-resolution remote sensing (HRRS) image land cover classification could provide significant help for urban monitoring, management, and planning. Deep learning (DL) based models have achieved remarkable performance in many land cover classification tasks through end-to-end supervised learning. However, the excellent performance of DL-based models relies heavily on a large number of well-annotated samples, which is impossible in practical land cover classification scenarios. Additionally, the training set could contain all of the different land cover types. To overcome these problems, in this paper a semi-supervised multiple-CNN ensemble learning method, namely Semi-MCNN, is proposed to solve the land cover classification problem. Considering the lack of labelled samples, a semi-supervised learning strategy was adopted to leverage large amounts of unlabelled data. In the proposed approach, an automatic sample selection method called an ensembled teacher model dataset generation (EMDG) was adopted to select samples and generate a dataset from large amounts of unlabelled data automatically. To tackle the error-propagation problem, an important strategy was adopted to correct the errors by pretraining on the selected unlabelled data and finetuning on the labelled data. Moreover, the semi-supervised idea together with the multi-CNN ensemble framework were integrated into an end-to-end architecture. This could significantly improve the generalization ability of the semi-supervised model, as well as the classification accuracy. Experiments were conducted on Shenzhen's land cover data (ShenzhenLC) and two other public remote sensing datasets. These experiments confirmed the superior performance of the proposed Semi-MCNN compared to the state-of-the-art land cover classification models.**

***Index Terms*—land cover classification, semi-supervised, deep learning, remote sensing**

I. Introduction

Land cover classification is a fundamental task of intelligent interpretation of remote sensing imagery, which aims to classify each pixel into a pre-defined land cover category. With the rapid development of earth observation and remote sensing technology, more and more sub-meter high spatial resolution remote sensing (HRRS) images have been acquired. Compared with low spatial resolution remote sensing imagery, sub-meter high spatial-resolution remote sensing imagery can capture more details of urban objects, which makes it capable of urban monitoring [1], planning [2], [3], and management [4] with a higher level of discrimination.

Lizhe Wang, Ruyi Feng, Jining Yan, Runyu Fan and Xiaohan Zhang are with School of Computer Science, China University of Geosciences, Wuhan 430074, PR China and Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, PR China.(e-mail: fengry@cug.edu.cn).

Corresponding author: Ruyi Feng.

Currently, the widely used remote sensing imagery land cover classification methods are based on supervised learning, which train the classifier according to the training samples and then classify the testing data into presetting categories with the trained classifier. Since AlexNet [5] was proposed in 2012, many DL-based models have been proposed to address the task of land cover classification. These methods use different DL-based models, such as autoencoder-based models [6], multilevel DL-based model [7], semi-transfer learning-based model [8], and multi-scale DL-based model [9]. By utilizing the spatial and spectral information comprehensively, and the powerful learning and feature conversion capabilities of deep models, these methods have achieved very significant results in land cover classification related tasks.

Through the end-to-end supervised learning manner, these DL-based models can automatically learn and transform high-level features from numerous labelled samples. These can also achieve state-of-the-art performance without any statistical information from spectral or spectral-spatial features. However, in real urban land cover classification scenarios, as the resolution of remote sensing imagery increases, more spatial details can be captured. With cities becoming more fragmented, there is significant heterogeneity in urban land cover types. The highly heterogeneous land cover types pose a considerable challenge for land cover classification tasks. Based on a data-driven idea, the DL-based methods utilize a large number of labelled samples in order to train highly nonlinear neural networks to distinguish highly heterogeneous land cover types. Whereas in real scenarios, the lack of labelled samples leads to poor classification results, especially for DL-based models.

In real-world land cover classification scenario, it is much easier to obtain unlabeled images than manually annotated data sets. Hence, many semi-supervised learning-based methods are proposed to use the raw unlabeled images to generate a labeled data set, which effectively makes up for the lack of well-labeled samples. These methods combine semi-supervised learning strategies (such as self-training approach [10] and pseudo-labeling strategies [11]) with SVM [12], graph-based method [13], GANs [14], transferable deep models [15], discriminative adversarial learning [16], and other models. By using a large amount of unlabeled data, the effect of land cover classification is significantly improved. However, the wrong choice of samples always occurs in the sample selection process of the semi-supervised method. These errors will be propagated to the downstream land cover classification tasks,

reducing the final land cover classification accuracy. Hence, most semi-supervised models reduce the impact of error propagation by manually setting certain thresholds (confidence parameters). However, since it is impossible to prevent the propagation of errors, the accuracy of land cover classification is restricted.

Inspired by the significant progress made in semi-supervised learning, a semi-supervised multi-CNN ensemble learning method (Semi-MCNN) for land cover classification is proposed in this work. To solve the lack of labelled samples problem, a semi-supervised learning strategy that started with a small labelled dataset and then spread the label to other unlabelled samples was used to leverage the unlabelled sub-meter high spatial-resolution remote sensing images. In this process, the wrong selection of samples might bring about an error propagation and affect the classification accuracy. To tackle this problem, a multi-model ensemble learning method was used to fuse multiple trained CNNs to reduce the errors in the sample selection. In addition, an important strategy was adopted to correct the errors by pretraining on the selected unlabelled data and finetuning on the labelled data. As a result, the proposed Semi-MCNN approach integrated a semi-supervised manner and the multi-CNN ensemble strategy into an end-to-end architecture. This effectively compensates for the poor generalization performance of a single model and the error propagation of a semi-supervised model.

To validate the effectiveness of the proposed method, a dataset built from the sub-meter high resolution remote sensing images of Shenzhen city, the ShenzhenLC dataset, was used for land cover classification. Experimental results with the ShenzhenLC dataset and two other widely used public remote sensing datasets showed the superior performance of the proposed Semi-MCNN model compared to state-of-the-art land cover classification methods. The main contributions of this paper are summarized as follows:

1) Focusing the lack of labelled samples problem on sub-meter urban land cover classification, we have proposed a simple semi-supervised method (Semi-MCNN), which integrates multiple CNNs to enhance the sample evaluation and uses pretrain-finetuning for error correction.
2) Taking Shenzhen city as the research area, we have constructed a large-scale land cover dataset (ShenzhenLC).
3) The proposed Semi-MCNN (10% training data) achieved similar accuracy close to supervised learning-based methods (60% training data) on the ShenzhenLC dataset. A sub-meter land cover classification map of Shenzhen city in 2018 has been obtained.

The rest of this paper is organized as follows: Section 2 presents the related work. In Section 3, the proposed method, a semi-supervised multi-model ensemble learning model for classification, is introduced in detail. Section 4 contains a description of the experimental datasets and an analysis of the experimental results. In Section 5, the conclusions are drawn.

II. Related Work

Land cover classification is a fundamental task in remote sensing imagery. It can be classified into two categories according to whether the samples are fully or partially labelled, i.e. supervised learning-based methods and semi-supervised learning-based methods. The supervised learning-based methods usually require a large number of labelled samples to train a robust classifier, while the semi-supervised learning-based methods often use small labelled samples. These methods also aim to label or learn more samples in order to tackle the small sample problem. In recent years, the sample generation method represented by Generative Adversarial Network (GAN) has also made impressive achievements in solving the lack of labelled samples problem. Therefore, in this section, the related work will be introduced from the following three aspects:

A. Supervised learning-based methods

Currently, the widely used remote sensing land cover classification methods are supervised learning-based methods, which train the classifier based on the prepared training samples, and then utilize the classifier to label the pixels into different land cover categories. Recently, the probability statistical-based methods, traditional machine learning-based methods, and DL-based methods have become the focus of current trends of supervised learning-based land cover classification.

The probabilistic statistical-based methods, including the nearest neighbour classification [17], [18], and the maximum likelihood method [19], usually assume that the data should obey normal distribution. When this condition cannot be met, classification results are unpredictable and the performance might be poor. To improve the classification results, machine learning-based methods such as SVM [20], [21], [22], artificial neural networks (ANN) [23], [24], [25], and decision trees (DT) [26], [27], [28] have been proposed. These methods have achieved satisfying results. However, with these methods the manually defined features all play an important role in the final result. Selecting and combining these defined features (including spatial features and spectral features) is time-consuming and limits the generalization of these models.

To reduce the dependence on the manually defined features, DL-based models have been intensely studied. Gong Cheng et al. [6] proposed a novel a single-hidden-layer autoencoder and a single-hidden-layer neural network to train coarse-to-fine shared intermediate representations. Considering that satellite images are usually multi-temporal and multi-sourced, Nataliia Kussul et al. [7] described a multilevel DL-based architecture that targeted land cover and crop type classification from multi-temporal and multi-source satellite imagery. To effectively use the rich spectral information, inspired by transfer learning, Huang et al. [8] proposed a semi-transfer deep convolutional neural network approach for multispectral remote sensing imagery. Some studies treated land cover classification as a semantic segmentation task, and the representative methods are a series of variants of FCN [29], [30], [31], [32] and Unet [33], [9].

DL-based methods can learn discriminative representation of different features through an end-to-end supervised process. However, there are still some problems that need to be solved for these DL-based methods. Traditionally, the excellent

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

performance of DL-based models relies on a large number of well-labelled samples. For a large study area such as Shenzhen city, there are significant differences in the visual representation of land cover types. Manually labelling a complete dataset containing all of the different visual representations of land cover types is expensive and time-consuming. As a result, reducing the dependence on labelled land cover samples is an important research topic.

B. GAN-based methods

Self-supervised learning-based high-resolution sample generation is another way to solve the lack of labelled samples problem. Since the GAN [34] can generate a large number of high-quality samples in a self-supervised manner, many GAN-based works have been studied in depth for sample generation to promote the solution to the lack of labelled samples problem. To generate high-quality remote sensing imagery samples, Xu et al. [35] applied the scaled exponential linear units into the GAN. Ma et al. [36] designed the SiftingGAN to generate more numerous and diverse labeled samples for data augmentation. Yu et al. [37] introduced the Attention GANs, which integrates the attention mechanism into GANs for aerial scene classification. More recently, Han et al. [38] proposed the GANRSIGM, which integrates the Wasserstein distance into GAN to create high-resolution samples for scene classification. In terms of semi-supervised learning combined with generative models, Zhan et al.'s [14] research shows that GANs combined with semi-supervised learning can also have a good hyperspectral image classification performance. Zhu et al. [16] proposed a semi-supervised centre-based discriminative adversarial learning framework for a cross-domain scene-level land cover classification of aerial images.

Most GAN-based methods use GAN for sample generation or feature learning. Compared with CNN-based methods, currently, the performance of GAN-based methods is inferior to CNN-based methods [39]. The possible reasons can be concluded as follows: Firstly, most GAN-based classification methods cannot be trained end-to-end, because they usually require labels to train an additional classifier. Secondly, the training of GAN is challenging. In the case of a few samples, the performance cannot be guaranteed (for example, spatial information and spectral information may be lost). Thirdly, the GAN-based methods generate samples from a certain distribution. In a real scene, the samples may not completely conform to the distribution or not only conform to the distribution. However, due to the powerful self-supervised feature learning ability of GAN, the GAN-based methods still provide a promising future direction for land cover classification.

C. Semi-supervised learning-based methods

To reduce the reliance on a large number of labelled samples, many semi-supervised learning-based methods for land cover classification were studied in-depth. The semi-supervised learning-based methods mainly include a semi-supervised strategy and a learning model. Self-training [10] is a widely used semi-supervised learning strategy. Focusing on obtaining an expanded annotation dataset, the self-training strategy annotates the most reliable predictions to unlabeled instances. In self-training, the classifier first trains on a limited number of labeled samples and then merges the most reliable instances with the initial samples into an expanded data set. The model is finally optimized on the expanded data set. However, existing self-training methods are based on handcrafted features, which rely heavily on manual design, and cannot guarantee effectiveness. Pseudo-Label [11] technology is a simple and efficient semi-supervised learning strategy for deep neural networks. During Pseudo-Label, the model is trained in a supervised process with well-labeled data. For unlabeled data, the class which has the maximum predicted probability is picked up as if they were true labels. Self-training and Pseudo-Label show good results on semi-supervised tasks, but these methods can not guarantee reliability since they assume that the selected samples are all correct.

Since the semi-supervised learning-based methods have good performance in the case of limited samples, in recent years, many semi-supervised learning-based methods have been proposed for remote sensing image classification tasks. Camps-Valls et al. [13] designed a semi-supervised graph-based method to handle the special characteristics of hyperspectral images for hyperspectral image classification. In terms of semi-supervised learning combined with discriminative models, Liu et al. [12] proposed a novel semi-supervised SVM model that utilized the self-training approach in order to address the problem of remote sensing land cover classification. Han et al. presented the [40] SSGA-E based on cotraining strategy and deep learning. Inspired by transfer learning, recently, Tong et al. [15] proposed a semi-supervised learning-based pseudo-labelling and sample selection scheme to train transferable deep models for land-use classification with HRRS images. IZ Yalniz [41] presented a web-scale semi-supervised teacher/student paradigm pipeline with large convolutional networks to leverage billions of images on the Internet to enhance image and video classification. Due to the dependence on manual preset parameters and thresholds, these methods still have room for improvement in real land cover classification scenarios.

These semi-supervised learning-based methods successfully solved the problem of reducing the dependence on a large number of labelled samples. However, they just only consider using a single feature space to evaluate samples. A single feature space has insufficient ability to distinguish samples. Thus it is easy to select the wrong samples, which may ultimately affect the classification accuracy. Especially when dealing with highly heterogeneous land cover types, the overfitting problems and the poor generalization performance are prone to occur. In this paper, we use multiple CNN ensembles to obtain different feature spaces through different neural networks. In this way, the error-propagation problem in semi-supervised sample selection can be largely alleviated. In addition, though these semi-supervised methods can reduce the error-propagation by optimising the sample selection method, there were still different levels of error-propagation problems. To solve these problems, an automatic sample selection method called EMDG is proposed. The proposed EMDG can reduce the propagation of errors in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

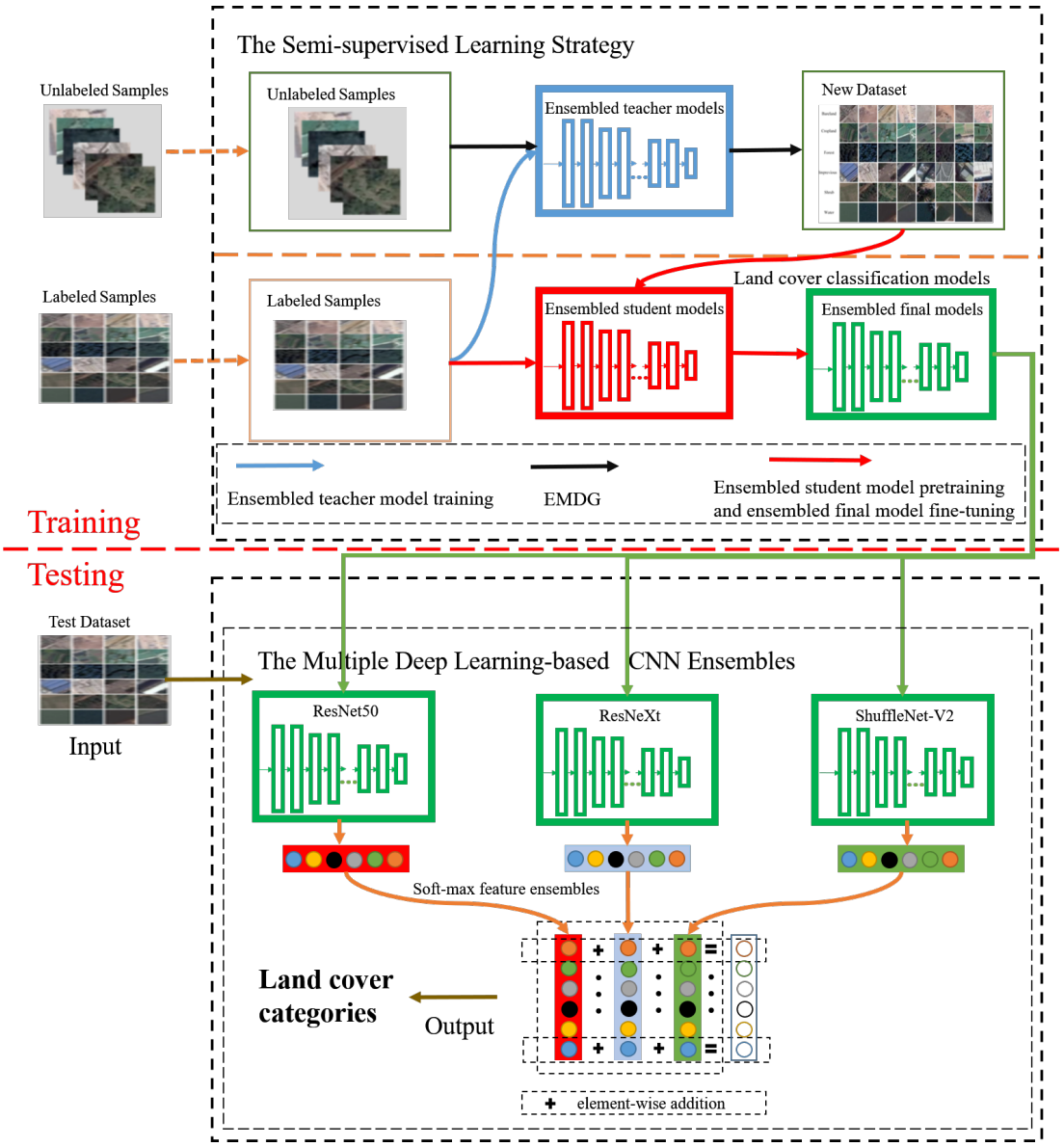


Fig. 1: The flow of the proposed Semi-MCNN.

semi-supervised sample selection by automatically selecting samples with higher accuracy, without any manual setting of thresholds and parameters. Moreover, a strategy was used to further correct the propagated errors by pretraining on the selected unlabelled data and finetuning on the labelled data. In Section III, more details will be introduced for the proposed algorithm.

III. The Semi-Supervised Multi-CNN Ensemble Learning Method (Semi-MCNN)

To tackle the lack of labelled samples problem, as well as to enhance the generalization of the classification models, a semi-supervised multi-model ensemble learning model (namely Semi-MCNN) is proposed in this paper. This method aimed to identify the land cover types for sub-meter high resolution remote sensing imagery. Formally, an HRRS image $I = \{x_{ij}\}$ was given, where i, j is the width and length of image I, respectively. The output $Y = \{y_{ij}\}$ with y_{ij} was the land cover label of x_{ij}.

The folowchart of the proposed Semi-MCNN is shown in Fig. 1. It consists of a semi-supervised learning strategy and ensembles multiple DL-based CNN models. Specifically, the proposed method contains two processes: training and testing. During the training process, the model utilizes the small samples and trains them with the semi-supervised strategy. Then, the initial classification model can be obtained. For the testing process, the test dataset is input into the initial model and output with the probability value through the ensembled multiple DL-based model. The final label for each pixel is decided with the maximum probability value in the probability distribution. The details are introduced in the following two subsections.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

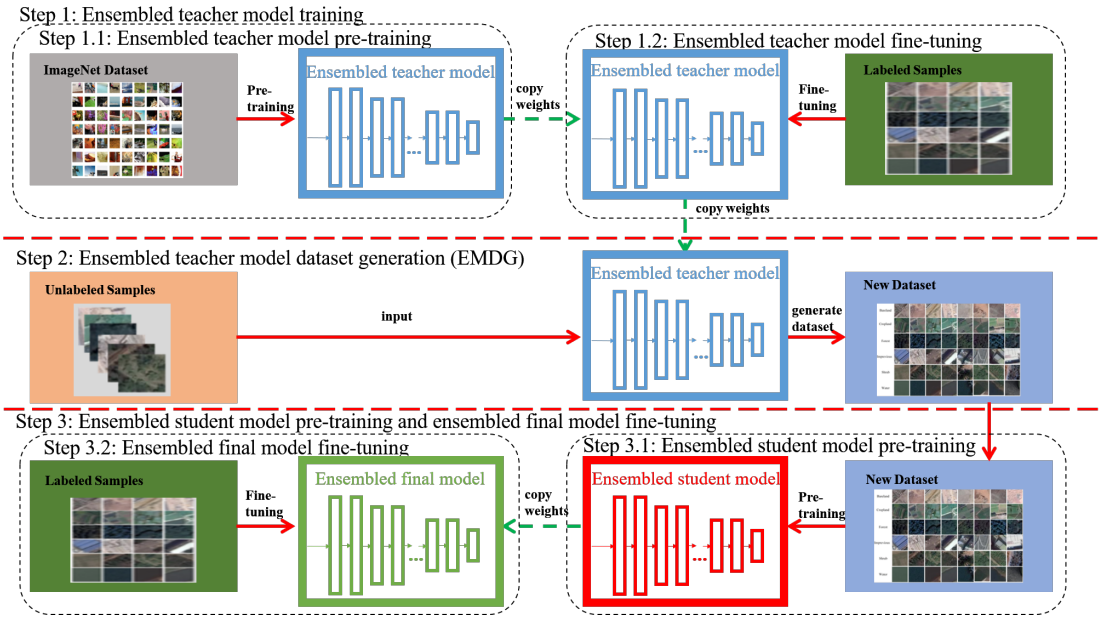


Fig. 2: The Semi-supervised Learning Strategy

A. The Semi-supervised Learning Strategy

In the proposed algorithm, to label the unlabelled samples more efficiently and precisely, a semi-supervised learning strategy was adopted, shown in Fig. 2. The semi-supervised strategy was based on the idea of the teacher-student model. First, an ensembled teacher model was used to train the training samples initially to obtain a preliminary sample discrimination. Then, the fine-tuned ensembled teacher model was applied to select samples and generate a new dataset. Finally, the newly generated dataset was used to train a student model to get the final model. The generated dataset used by the student model was selected by the teacher model from unlabelled samples, and would cause an error-propagate problem. Therefore, a fine-tuning operation was adopted by the student model with a labelled dataset in order to optimize the final results. The semi-supervised strategy is described as follows:

1) Ensembled teacher model training: In the semi-supervised learning strategy, the first step is to train the teacher model. A good teacher model will significantly reduce the error transmission problem in the semi-supervised process and directly affect the final performance of the proposed method. Therefore, two strategies are used to guarantee a better teacher model. The first one is the use of transfer learning. The parameters trained by the ImageNet dataset are adopted in order to initialize the teacher model. Next, a fine-tuning operation is carried out for the teacher model with the training samples. The second strategy is the multi-model ensemble work. Because a single model can easily fall into a local optimum, and the generalization ability is not enough, in the Semi-MCNN a variety of different CNN models are used for the ensemble methods. The features extracted and transformed by different CNN models are ensembled to obtain different features. Through the above two strategies, a good ensembled teacher model is obtained for the semi-supervised

Algorithm 1 Ensembled teacher model dataset generation (EMDG)

Input: The lowest single category accuracy rate in the ensembled teacher model: $acc_{teacher}$, Unlabelled data: $U = \{u_1, u_2, ..., u_I\}$, Ensembled teacher model: T

Output: $NewDataSet\ D = \{d_1, d_2, ..., d_K\}$

function EMDG($acc_{teacher}$, U, T)
  $D \leftarrow \varnothing$
  for $i = 1$ to I **do**
    Input u_i into the model T and calculate the corresponding category and its accuracy, and denote it as acc
    if $acc \geq acc_{teacher}$ **then**
      Add u_i to D according to the corresponding category
    end if
  end for
  Calculate the number of samples in each subset in D. The number of samples in the smallest subset are called the minimum number N
  Sort the samples in each subset in D in decreasing order according to acc.
  Keep the samples within the top N of each subset in D, and discard the remaining samples.
  return D
end function

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

learning-based classification.

2) Ensembled teacher model dataset generation (EMDG): After the teacher model is obtained, labelling the large unlabelled dataset and generating an augmented training sample dataset is another task. Currently, the sample generation methods mostly rely on thresholds, which are preset manually. These include the top-K [41] and the SSGA-E [40] method, and the performance might be greatly affected by the manually preset thresholds. To improve these circumstances, a semi-supervised dataset generation method—namely the EMDG—is presented in this paper.

The process of EMDG is shown in Algorithm 1. Here, the lowest single category accuracy rate in the ensembled teacher model ($acc_{teacher}$), the unlabelled dataset: $U = \{u_1, u_2, ..., u_I\}$, and the ensembled teacher model T are the input. The purpose of the EMDG is to obtain a generated dataset: $NewDataSet\ D = \{d_1, d_2, ..., d_K\}$. First, the unlabelled sample is input into the ensembled teacher model T to calculate the category and the corresponding accuracy rate, which is denoted as acc. To reduce the error of the sample selection with teacher model T, samples with a higher confidence value are selected (lines 5-8 in Algorithm 1). In the newly generated dataset, the number of samples in each category might be inconsistent, which will cause the category imbalance problem and affect the final land cover classification accuracy. To avoid this problem, the number of samples in each category, which are denoted as N, are calculated to ensure the minimum number for each category. Then, the first N samples in descending order of acc for each class are selected to obtain the final dataset D (lines 9-12 in Algorithm 1).

3) Ensembled student model pretraining and ensembled final model fine-tuning: Finally, the ensembled student model is trained with this newly generated training dataset. Because the newly generated dataset D is automatically selected from the unlabelled samples, it will inevitably contain some wrong samples, which will result in error propagation problems. To solve this problem, the weights of the ensembled student model are inherited and used to initialize the final model. Then, the final model is fine-tuned with the original labelled dataset. By pretraining on the selected unlabelled data and finetuning on the labelled data, the error propagation problems can be better solved. Through progressive training from coarse (the ensembled teacher model) to fine (the ensembled final model), the performance can be greatly improved.

B. The Multiple Deep Learning-based CNN Ensembles

To enhance the generalization of the semi-supervised learning-based classification methods, multiple CNN models are to be ensembled. Many works have shown that different structures of CNN can learn variant features, and some are good at extracting spatial features. Others are more skilled at extracting spectral features. Different components of CNN have different feature perception capabilities, such as kernel size, activation function, depth, and the number of hidden units. Therefore, the features generated by different CNNs can be combined to improve the land cover classification performance. Among them, residual networks [42] have efficient feature extraction capabilities, and the ResNet50 network is widely used for various image classification tasks due to its strong feature extraction and generalization capabilities. The ResNet50 network uses skip connections, which enables the network to be trained under a very deep architecture (hundreds of layers), and significantly improves the accuracy of the model. The ResNeXt [43], as a supplement and an upgraded version of residual networks, has multi-scale analysis capabilities. The RexNeXt network proposes the aggregated residual transformations combined with the deep residual network to more effectively utilize the parameters. Compared with the residual network, RexNeXt has better classification ability with fewer parameters. ShuffleNet-V2 [44] is a lightweight network with efficient feature extraction capability and few parameters, which is very suitable for model fusion. Therefore, the ResNet50, ResNeXt, and ShuffleNet-V2 models are suitable to apply to remote sensing imagery for feature extraction.

The proposed semi-supervised method, shown in Fig. 2, is first used to train the above three models independently and save the optimal model for each network. The above models are used to extract the softmax feature, which was the output vector of the softmax layer (denoted as $output = \{a_1, a_2, ..., a_k\}$, and k is the number of categories). Where

$$a_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, i \in [1, k] \tag{1}$$

z is the input vector of the softmax layer.

Finally, the softmax features are ensembled to calculate the final output by normalizing the element-wise addition of these softmax features, which are defined as follows:

$$output_{final} = \frac{\sum_{i=1}^{N} output_i}{N} \tag{2}$$

where N is the number of models. The category corresponding to the largest value in output is the predicted land cover type.

IV. Experiment and analysis

To validate the effectiveness of the proposed framework, in this section ResNet50 [42], ResNeXt [43], and ShuffleNet-V2 [44] are adopted and compared to the proposed method. More information about the study area, as well as the classification system, will be introduced in Subsection IV-A. Details about the experimental settings are presented in Subsection IV-B. The analysis on the experimental results is shown in Subsection IV-C, Subsection IV-D, Subsection IV-E, respectively.

A. The Shenzhen data and classification system

1) Classification system: In this paper, the Chinese Land Use Classification Criteria (GB/T21010-2017) is referred as our land cover classification system, and the Shenzhen data is classified into six categories: bare land, cropland, impervious, shrub, water, and forest, shown as Fig. 3.

2) Shenzhen data: In this study, Shenzhen is used as an example of urban land cover classification because it is a typical international city that attracts industry, trade, tourism, and finance. The city is China's first special economic zone, which covers an area of about 2,020.5 km^2, and is located

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Fig. 3: Samples of land cover categories in Shenzhen city

in the southern part of Guangdong Province, China. Shenzhen city borders Hong Kong in the south, Huizhou in the north and northeast, Dongguan in the north and northwest, the Pearl River in the west, and Dapeng Bay in the east. Over the past three decades, Shenzhen city has been one of the fastest-growing cities in the world, and its land use and land cover have undergone tremendous changes. The houses, natural growth, and artificially planted vegetation have different visual appearances. Urban planning, construction, and development have made the land cover fragmented and heterogeneous. Affected by factors such as light, temperature, and humidity, land covers in different regions also show significant differences. These will increase the heterogeneity of land cover in Shenzhen city and affect its classification accuracy. Therefore, to better estimate the land cover types of Shenzhen city, in this study Shenzhen is treated as a research area for land cover classification research.

To better analyze the land cover situation in Shenzhen city, we used sub-meter high resolution remote sensing images of Shenzhen city obtained from Google earth, and manually constructed a sub-meter ($0.59m$) land cover classification dataset of Shenzhen (ShenzhenLC). The dataset containes six classes of land cover types, and each class contains 1,000 images, as shown in Table 1. In DL-based land cover classification experiments, the images are usually reshaped to the size of 28, 56, 128, 224, or 256 pixels. In this study, considering that the image patches should contain enough information (not too small), and also should not bring too many parameters to the model (not too large), all of the image patches are reshaped to 56×56.

B. Experimental settings

To assess the performance of the proposed framework, ResNet50 and ResNeXt, together with ShuffleNet-V2, were adopted to analyze the land cover situation of Shenzhen city using the ShenzhenLC dataset. In addition, to prove the effectiveness and generalization of the proposed Semi-MCNN, two experiments are also designed based on the public benchmark remote sensing datasets, i.e. the NWPU-RESISC45 dataset [45] and the aerial image dataset (AID) [46].

The semi-supervised strategy in the proposed Semi-MCNN model includes three training steps: Ensembled teacher model training, Ensembled teacher model dataset generation (EMDG), and Ensembled student model pretraining and ensembled final model fine-tuning. In the ensembled teacher model training step, only 10% of the labelled samples are used to train models, without any unlabelled samples. In the ensembled teacher model dataset generation (EMDG) step, we pick up unlabelled samples using the proposed EMDG algorithm. The number of samples picked for the ensembled student model pretraining are 337 per class ($337 \div 1000 = 33.7\%$) for the ShenzhenLC dataset, 61 per class ($61 \times 30/10000 = 18.3\%$) for the AID dataset and 171 per class ($171 \div 700 = 24.2\%$) for the NWPU-RESISC45 dataset. In the ensembled final model fine-tuning step (the final training process), only 10% of the labelled samples (without any unlabelled samples) are used.

The parameters of the model were finally determined after many attempts. During the training, the Adam optimization method [47] was used. The learning rate was set to 0.0001 for the pretraining period, and 0.00001 for the finetuning period. While the decay was set to 0.000001, the batch size was set to 32 and a total of 100 epochs were trained. In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE I: Statistics for the land cover classification dataset of Shenzhen city (ShenzhenLC)

Land cover categories	bareland	cropland	forest	impervious	shrub	water
Number of samples per category	1000	1000	1000	1000	1000	1000

order to facilitate a comparison to other models, the following strategy was used to split the dataset: 10% as the training set, 50% as the unlabelled set, and 40% as the testing set. All implementations were based on Pytorch (Version 1.5.0) and an NVIDIA RTX 2080Ti GPU. In order to evaluate the proposed method, we assessed the experimental results using the overall accuracy (OA) and the confusion matrix, which are widely used evaluation criteria.

C. Experiment (1): Land cover classification using the ShenzhenLC dataset

Table II shows the OA of different models for the ShenzhenLC dataset. It can be seen that the proposed method had the highest OA, with nearly 3% improvement on the OA when using just 10% of the training set compared to other models. It is worth mentioning that the OA of the proposed method using only 10% of the training set outperformed the other models. This occurred even when 50% of the training set was used. It was also close to the other methods that used 60% of the training set. The experimental results using the ShenzhenLC dataset indicate that the proposed approach can effectively generate samples using a semi-supervised learning strategy, and that the multiple DL-based CNN ensembles can simultaneously improve the accuracy of land cover classification.

To test the effectiveness of the semi-supervised strategy in the proposed model, models with different training stages were used. The final results with the OA values are shown in Table III. It can be seen that the proposed Semi-MCNN (final) model achieved the highest OA value. The proposed Semi-MCNN (final) model achieved an improvement of nearly 2% on OA compared to the Semi-MCNN (teacher) model, and a nearly 1% improvement compared to the Semi-MCNN (student) model. Still, the results of the additional experiment indicate that the proposed approach can effectively reduce the error propagation problem and enhance the classification accuracy. The classification map of Shenzhen city in 2018 is shown in Figure 5. The original classification maps exceeds $40GB$. In order to show it in the paper, we resample it to 3508×2480 pixels.

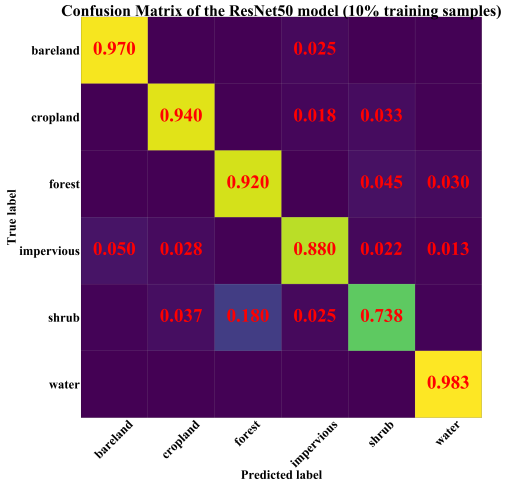
TABLE IV shows the time cost of different models as well as the proposed Semi-MCNN on the ShenzhenLC dataset. As can be seen from TABLE IV, for training time per epoch, the time cost of the proposed Semi-MCNN model (6.40s per epoch) on a single GPU is approximately the sum of the other three models. To reduce the overall time cost, the three different CNN networks were independently trained on different machines or GPUs. When the proposed model is trained on three GPUs separately, the time consumption (2.40s per epochs) is close to the training time of these single models (ResNet50, ResNeXt and ShuffleNet-V2). For the ShenzhenLC test data, the time cost of the proposed model on the entire test data set (8.73s) is about twice of the single models. Considering that for the neural networks, the time cost of training is much greater than the time cost of testing, so we believe that the twice of testing time will not have a great impact on the real application outcome of the proposed Semi-MCNN.

The confusion matrices obtained by ResNet50 [42], ResNeXt [43], ShuffleNet-V2 [44] and the proposed method, are shown in Fig. 4 (a-f, respectively). It can be seen that for all models, the forest and shrub categories were heavily confused, perhaps because of the difficulty in identifying different visual features, such as texture and structural information. For water, impervious surfaces, and bare land, all models had a high classification score. It demonstrated that visual features can help classify these categories well. It can be seen that the proposed method had the highest accuracy rate in all of the categories except for the shrub category, which obtained the second-highest OA accuracy.

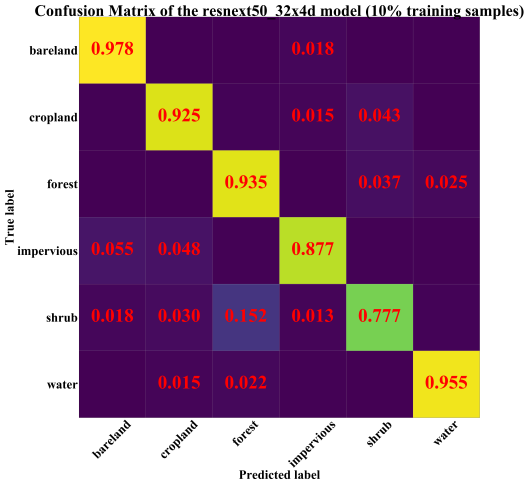
D. Experiment (2): The NWPU-RESISC45 dataset

The NWPU-RESISC45 dataset is a large-scale dataset that was used for the remote sensing imagery scene classification task, which contains 31,500 remote sensing images with 45 categories. The images were obtained from satellite as well as aerial photography. The spatial resolution varies from about 0.2 to 30 meters. Due to the diversity of different perspectives, occlusions, poses, spatial resolutions, this dataset contains certain similarities between the different categories, as well as certain differences within the same categories. Recently, the NWPU-RESISC45 dataset has been widely used to evaluate the performance of different scene classification models. Therefore, in this study it was also used to assess the proposed model with other state-of-the-art models.

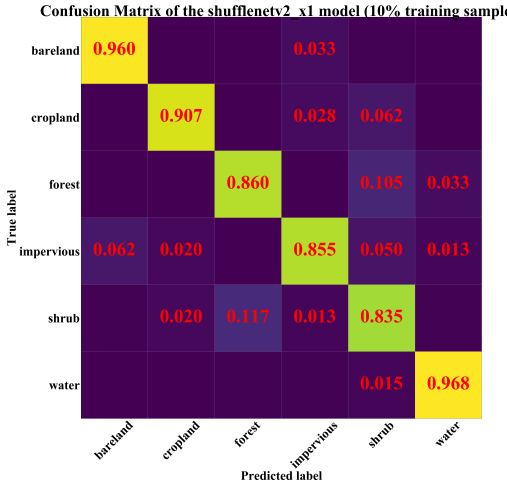
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing



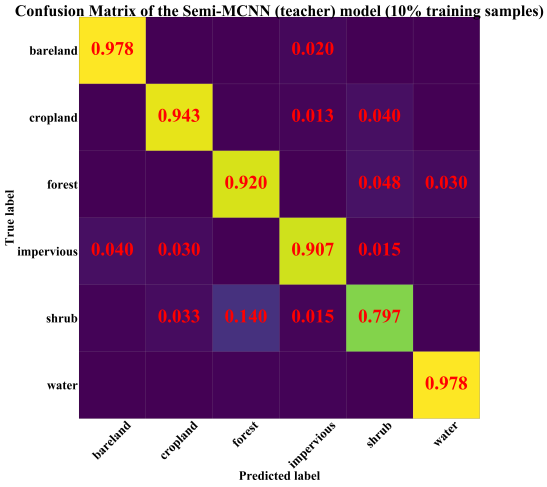
(a) The Confusion Matrix of ResNet50.



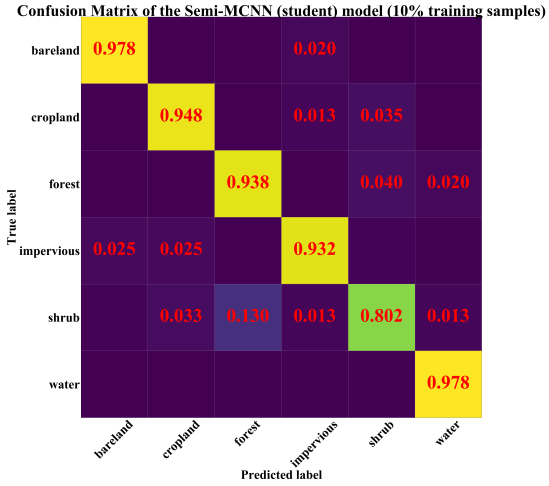
(b) The Confusion Matrix of ResNeXt.



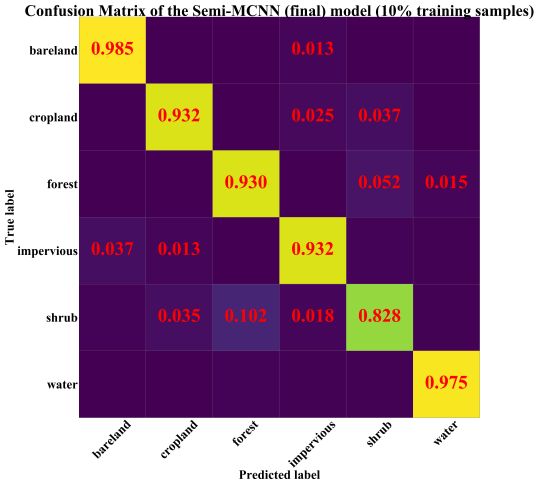
(c) The Confusion Matrix of ShuffleNet-V2.



(d) The Confusion Matrix of Semi-MCNN (teacher model).



(e) The Confusion Matrix of Semi-MCNN (student model).



(f) The Confusion Matrix of Semi-MCNN (final model).

Fig. 4: The confusion matrix of the mentioned methods for the ShenzhenLC dataset

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE II: Results for the proposed models on the ShenzhenLC dataset

	Model Name	OA (%)		
		10% training set	50% training set	60% training set
Supervised Methods	ResNet50	90.50	92.26	92.54
	ResNeXt	90.79	92.73	**93.17**
	ShuffleNet-V2	89.75	92.83	92.88
		OA (%)		
		10% training set and 50% unlabeled set		
Semi-MCNN	Semi-MCNN (ensembled teacher model)	92.04		
	Semi-MCNN (ensembled student model)	92.92		
	Semi-MCNN (ensembled final model)	**93.04**		

TABLE III: Results for different models of the proposed Semi-MCNN on the ShenzhenLC dataset

	Model Name	OA (%)
Semi-MCNN (ensembled teacher model)	ResNet50	90.50
	ResNeXt	90.79
	ShuffleNet-V2	89.75
Semi-MCNN (ensembled student model)	ResNet50	91.75
	ResNeXt	91.83
	ShuffleNet-V2	91.67
Semi-MCNN (ensembled final model)	ResNet50	**92.29**
	ResNeXt	**92.67**
	ShuffleNet-V2	**92.50**

TABLE IV: Time cost of different models and the proposed Semi-MCNN on the ShenzhenLC dataset

Model	Training Time per Epochs (s)	Time cost of the ShenzhenLC test data (s)
ResNet50	2.20	4.90
ResNeXt	2.40	5.12
ShuffleNet-V2	2.35	4.96
Semi-MCNN (One GPU)	6.40	8.73
Semi-MCNN (Three GPU)	2.40	-

Table V shows the quantitative results of the different models in our experiments. It can be seen that the proposed Semi-MCNN outperformed the other semi-supervised methods, with an improvement in OA ranging from 5% to 12%. The proposed Semi-MCNN had the highest OA value compared to all of the supervised learning-based models, with an improvement of nearly 1% to 3.5% in the case of the 10% training set. It can be seen that the Semi-MCNN, in the case of the 10% training set, had a higher OA value than some of the supervised learning-based models in the case of the 20% training set. It was also close to HW-CNN [51] and Hygra [52] in the case of the 20% training set, which are the current outstanding models for remote sensing image classification. The results also indicated that the proposed framework had a very competitive performance with small samples.

The confusion matrix of the proposed method for the NWPU-RESISC45 dataset is present in Fig. 6. It can be seen that some pairs of classes, e.g. church and palace, mountain and desert, rectangular farmland and terrace, lake and wetland, were slightly confused. The reason for this is because there is a certain visual similarity between these categories. Some categories, such as church, freeway, commercial area, medium

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

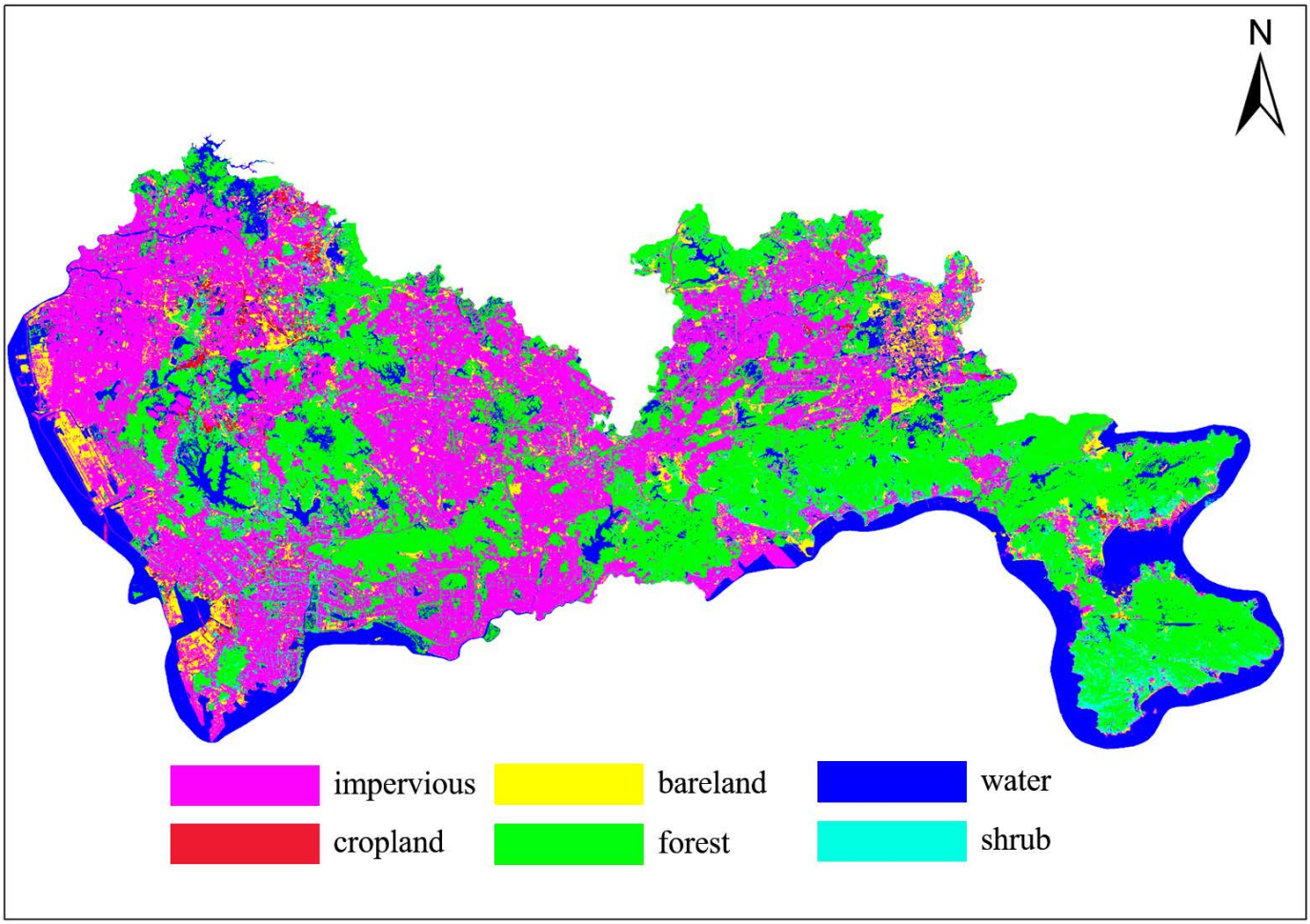


Fig. 5: The land cover classification map of Shenzhen city

residential, palace, railway station, and thermal power station, where the OA was lower than 90%, were difficult to correctly classify due to their high inter-class diversity.

E. Experiment (3): Aerial image dataset (AID)

The AID is different from the NWPU-RESISC45 dataset, as it is a large-scale aerial image dataset collected from Google Earth images, which consists of 30 categories of aerial scenes with 10,000 images. The images in the AID are multi-sourced, as these come from different sensors. They are acquired at different locations, times and seasons, and so the different imaging conditions lead to a certain diversity within each category. This can bring certain challenges to different scene classification algorithms. In this section, the AID dataset was also used to evaluate the proposed method.

Table VI shows the quantitative results of different algorithms with the AID. Unlike the previous experiments on the NWPU-RESISC45 where the OA only reached 94.51%, on the AID, the OA of the best methods is higher than 95% and 97% with the labeled training set ratio of 20% and 50%. Compared with the NWPU-RESISC45, the AID contains fewer categories and smaller diversities and variations, which enables DL-based models to achieve better classification results. It can be observed that the Semi-MCNN outperformed the other semi-supervised methods with an improvement rate of 7% to 14%. Different from the results for other semi-supervised comparison methods, the supervised learning-based methods generally achieved an OA of higher than 90%, due to the use of more samples. Among them, the Hygra [52] achieved the best OA in the case of 20% and 50% ratio of training sets. The Semi-MCNN in the case of the 10% training set achieved the highest OA value. It had an improvement rate of nearly 0.2% to 5% in the case of the 20% training set for supervised methods. In addition, in the case of the 10% training set, the proposed final model reached a higher OA than some of the supervised learning-based models in the case of the 50% training set. When the proposed method used the 10% training set, the results were only 1% to 1.7% lower than the results of the current best classification methods ((D-CNN [48], SF-CNN with VGGNet [49], HW-CNN [51] and Hygra [52]) with the 50% training set.

The confusion matrix of the Semi-MCNN on the AID dataset is provided in Fig. 7. It can be observed that the resort and park were easily misclassified, and that the categories of school, resort, commercial, and square were slightly confused, as these have a high intra-class similarity.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE V: Results for the proposed models on the NWPU-RESISC4 dataset

	Model Name	OA (%)	
		10% training set	20% training set
Supervised Methods	D-CNN [48]	89.22 ± 0.50	91.89 ± 0.22
	SF-CNN with GoogleNet [49]	87.43 ± 0.13	90.51 ± 0.13
	SF-CNN with VGGNet [49]	89.89 ± 0.16	92.55 ± 0.14
	Inception-v3-CapsNet [50]	89.03 ± 0.21	92.6 ± 0.11
	HW-CNN [51]	-	94.38 ± 0.16
	ResNet [52]	89.24 ± 0.75	91.96 ± 0.71
	Hygra: an ensemble of CNNs [52]	$\mathbf{92.44 \pm 0.34}$	$\mathbf{94.51 \pm 0.21}$
		OA (%)	
		10% training set and 50% unlabeled set	
Semi-supervised Methods	Self-training (VGG-S) [40]	81.46 ± 0.68	
	Self-training (ResNet) [40]	85.82 ± 1.30	
	Co-training [40]	87.25 ± 0.95	
	SSGA-E [40]	88.60 ± 0.95	
Semi-MCNN	Semi-MCNN (ensembled teacher model)	92.28	
	Semi-MCNN (ensembled student model)	91.13	
	Semi-MCNN (ensembled final model)	**93.48**	

TABLE VI: Result of proposed models on the AID.

	Model Name	OA (%)		
		10% training set	20% training set	50% training set
Supervised Methods	D-CNN [48]	-	90.82 ± 0.16	96.89 ± 0.10
	SF-CNN with GoogleNet [49]	-	91.83 ± 0.11	95.53 ± 0.09
	SF-CNN with VGGNet [49]	-	93.60 ± 0.12	96.66 ± 0.11
	VGG-16-CapsNet [50]	-	91.63 ± 0.19	94.64 ± 0.17
	HW-CNN [51]	-	-	96.98 ± 0.33
	Hygra: an ensemble of CNNs [52]	-	$\mathbf{95.50 \pm 0.27}$	$\mathbf{97.40 \pm 0.10}$
		OA (%)		
		10% training set and 50% unlabeled set		
Semi-supervised Methods	Self-training (VGG-S) [40]	81.46 ± 0.68		
	Self-training (ResNet) [40]	85.82 ± 1.30		
	Co-training [40]	87.25 ± 0.95		
	SSGA-E [40]	88.60 ± 0.95		
Semi-MCNN	Semi-MCNN (ensembled teacher model)	0.9410		
	Semi-MCNN (ensembled student model)	0.9490		
	Semi-MCNN (ensembled final model)	**0.9573**		

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

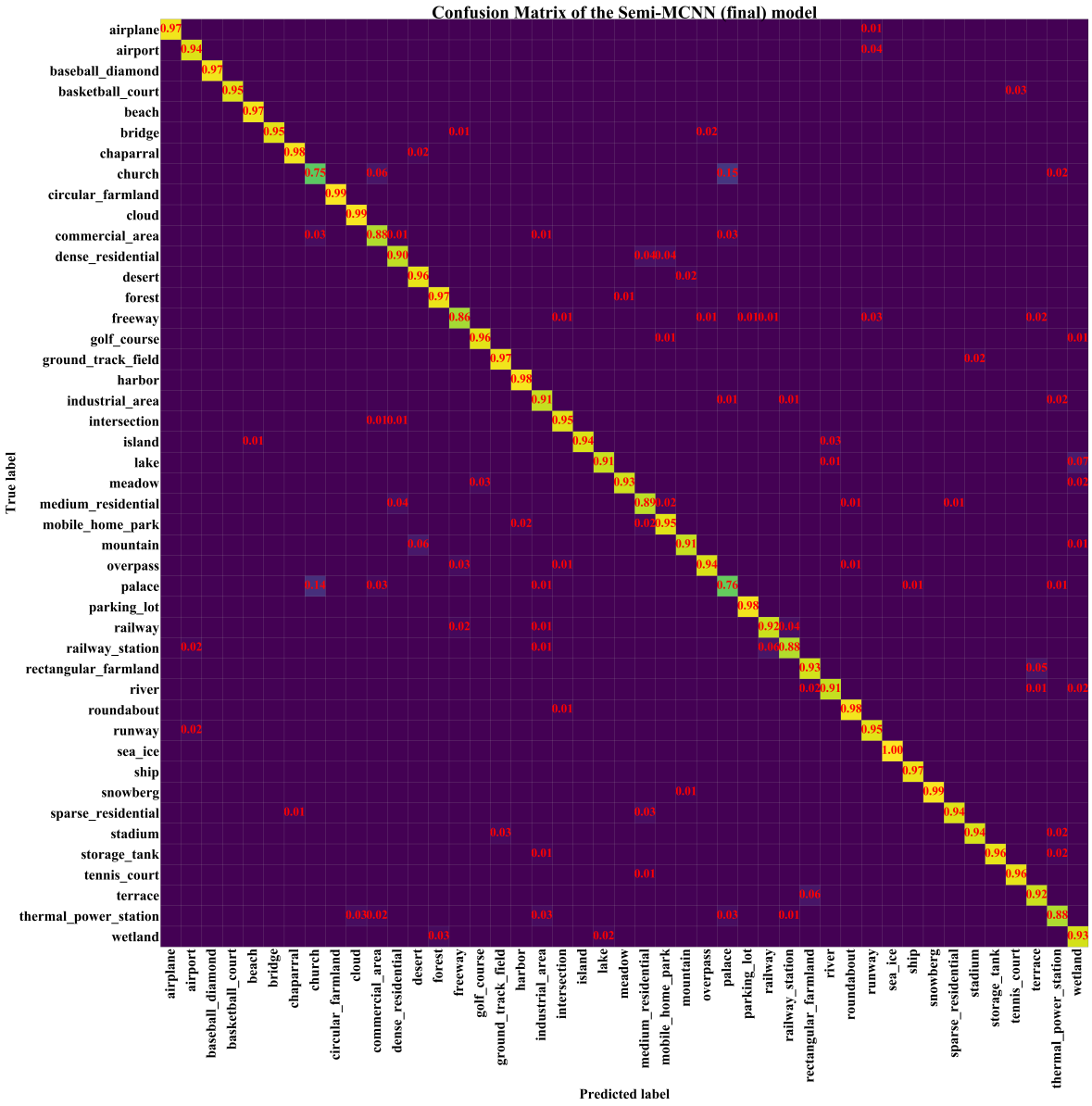


Fig. 6: The Confusion Matrix of The Proposed Semi-MCNN on the NWPU-RESISC45 dataset

V. Conclusions

In this paper, a semi-supervised multi-CNN ensemble learning method, namely the Semi-MCNN, was proposed for urban land cover classification, which integrated a semi-supervised strategy and a multi-CNN ensemble strategy into an end-to-end architecture. A semi-supervision strategy was adopted to leverage lots of unlabelled images to labelled samples, and the EMDG was proposed to automatically select samples and generate a dataset from unlabelled data without any manual setting thresholds. In addition, pretraining on the selected unlabelled data and finetuning on the labelled data operation were taken in the proposed framework and used to solve the error propagation problem. To evaluate the performance of the proposed Semi-MCNN method, several state-of-the-art land cover classification models were compared to the ShenzhenLC dataset, together with the public high-resolution remote sensing scene classification benchmarks, i.e. the NWPU-RESISC45 dataset and the AID. All of the experimental results demonstrated a consistent conclusion that the proposed Semi-MCNN had better performance results compared to other state-of-the-art models in both quality and quantity. In addition, it can achieve comparative accuracy with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3019410, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

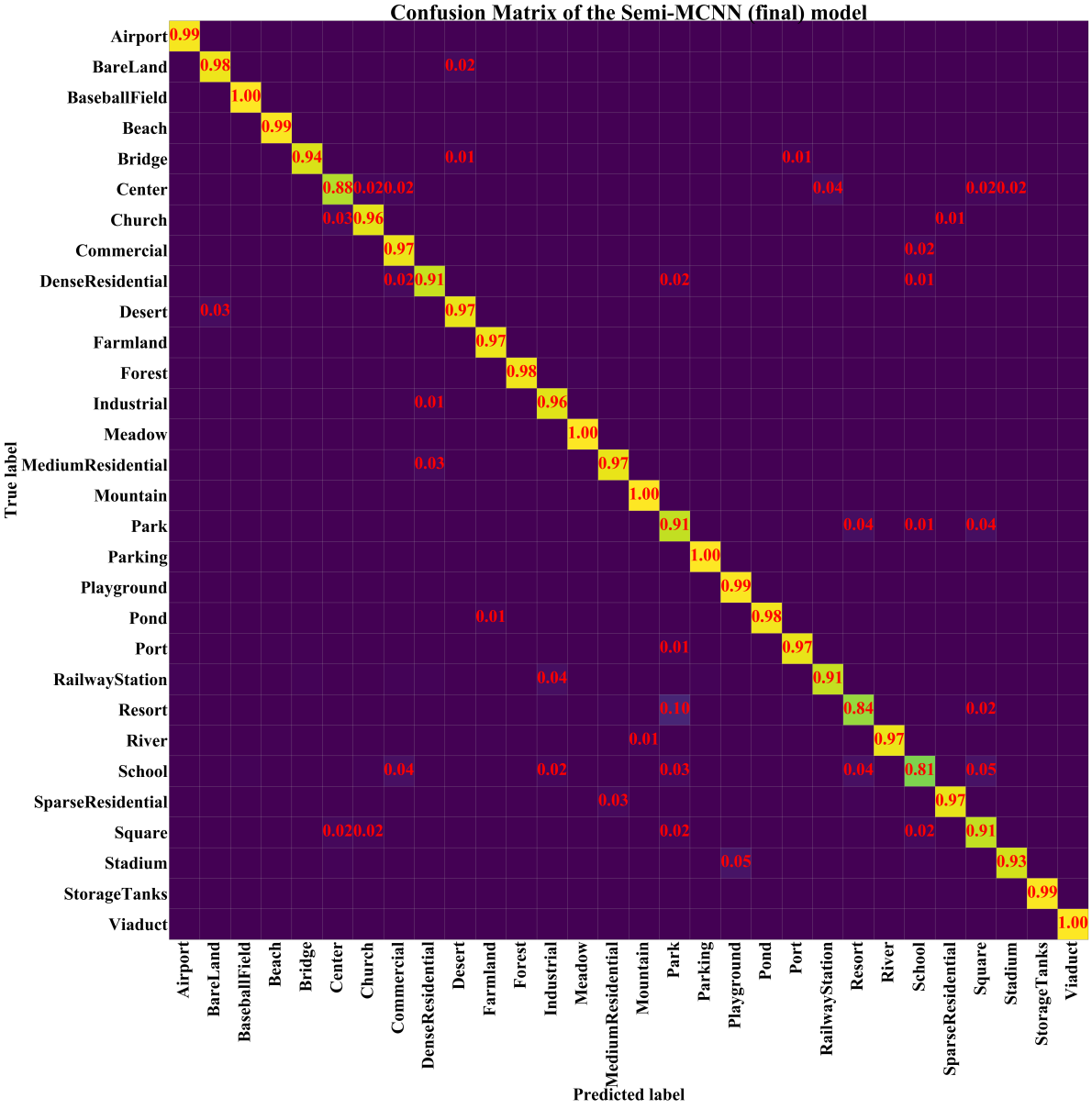


Fig. 7: The Confusion Matrix of The Proposed Semi-MCNN on the AID

small samples compared to the supervised model with big training sample sets, and also proves the superiority of the semi-supervised strategy and the multi-CNN fusion strategy of the Semi-CNN in handling land cover classification tasks.

ACKNOWLEDGMENT

This paper is funded by National Natural Science Foundation of China (No. U1711266, No. 41925007 and No.41701429).

REFERENCES

[1] P. Treitz and J. Rogan, "Remote sensing for mapping and monitoring land-cover and land-use change-an introduction," *Prog. Plan.*, vol. 61, no. 4, pp. 269–279, 2004.

[2] S. Pauleit, R. Ennos, and Y. Golding, "Modeling the environmental impacts of urban land use and land cover change—a study in merseyside, UK," *Landsc. Urban Plan.*, vol. 71, no. 2-4, pp. 295–310, 2005.

[3] E. López, G. Bocco, M. Mendoza, and E. Duhau, "Predicting land-cover and land-use change in the urban fringe: a case in morelia city, Mexico," *Landsc. Urban Plan.*, vol. 55, no. 4, pp. 271–285, 2001.

[4] W. Zhou, G. Huang, and M. L. Cadenasso, "Does spatial configuration matter? understanding the effects of land cover pattern on land surface temperature in urban landscapes," *Landsc. Urban Plan.*, vol. 102, no. 1, pp. 54–63, 2011.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 53, no. 8, pp. 4238–4249, 2015.

[7] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote. Sens. Lett.*, vol. 14, no. 5, pp. 778–782, 2017.

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

[9] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li, and S. Zhang, "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," *Sensors*, vol. 18, no. 11, p. 3717, 2018.

[10] R. Mihalcea, "Co-training and self-training for word sense disambiguation," in *Proc. Conf. Comp. Nat. Langu. Learn., CoNLL 2004, Boston, Massachusetts, USA, May 6-7, 2004*, pp. 33–40.

[11] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn., ICML 2013, Atlanta, Georgia, USA, 2013*, vol. 3, no. 2, 2013.

[12] Y. Liu, B. Zhang, L. Wang, and N. Wang, "A self-trained semisupervised SVM approach to the remote sensing land cover classification," *Comput. Geosci.*, vol. 59, pp. 98–107, 2013.

[13] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 45, no. 10, pp. 3044–3054, 2007.

[14] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 2, pp. 212–216, 2018.

[15] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," 2018.

[16] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images," *ISPRS-J. Photogramm. Remote Sens.*, vol. 155, pp. 72–89, 2019.

[17] P. T. Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2018.

[18] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k -nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 48, no. 11, pp. 4099–4109, 2010.

[19] J. R. Otukei and T. Blaschke, "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 12, no. Supplement-1, pp. S27–S31, 2010.

[20] M. Pal, "Support vector machine-based feature selection for land cover classification: a case study with dais hyperspectral data," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 2877–2894, 2006.

[21] T. Kavzoglu and I. Colkesen, "A kernel functions analysis for support vector machines for land cover classification," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 11, no. 5, pp. 352–359, 2009.

[22] C. Sukawattanavijit, J. Chen, and H. Zhang, "GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data," *IEEE Geosci. Remote. Sens. Lett.*, vol. 14, no. 3, pp. 284–288, 2017.

[23] D. L. Civco, "Artificial neural networks for land-cover classification and mapping," *Int. J. Geogr. Inf. Sci.*, vol. 7, no. 2, pp. 173–186, 1993.

[24] T. Kavzoglu and P. Mather, "The use of backpropagating artificial neural networks in land cover classification," *Int. J. Remote Sens.*, vol. 24, no. 23, pp. 4907–4938, 2003.

[25] X. Song, Z. Duan, and X. Jiang, "Comparison of artificial neural networks and support vector machine classifiers for land cover classification in northern china using a spot-5 hrg image," *Int. J. Remote Sens.*, vol. 33, no. 10, pp. 3301–3320, 2012.

[26] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sens. Environ.*, vol. 86, no. 4, pp. 554–565, 2003.

[27] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, 1997.

[28] R. De Fries, M. Hansen, J. Townshend, and R. Sohlberg, "Global land cover classifications at 8 km spatial resolution: the use of training data derived from landsat imagery in decision tree classifiers," *Int. J. Remote Sens.*, vol. 19, no. 16, pp. 3141–3168, 1998.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR 2015, Boston, MA, USA, June, 2015*, pp. 3431–3440.

[30] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016.

[31] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS-J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, 2018.

[32] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote. Sens.*, vol. 9, no. 5, p. 498, 2017.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[34] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[35] S. Xu, X. Mu, D. Chai, and X. Zhang, "Remote sensing image scene classification based on generative adversarial networks," *Remote Sens. Lett.*, vol. 9, no. 7, pp. 617–626, 2018.

[36] D. Ma, P. Tang, and L. Zhao, "Siftinggan: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro," *IEEE Geosci. Remote. Sens. Lett.*, vol. 16, no. 7, pp. 1046–1050, 2019.

[37] Y. Yu, X. Li, and F. Liu, "Attention gans: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 1, pp. 519–531, 2020.

[38] W. Han, L. Wang, R. Feng, L. Gao, X. Chen, Z. Deng, J. Chen, and P. Liu, "Sample generation based on a supervised wasserstein generative adversarial network for high-resolution remote-sensing scene classification," *Inf. Sci.*, vol. 539, pp. 177–194, 2020.

[39] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[40] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS-J. Photogramm. Remote Sens.*, vol. 145, pp. 23–43, 2018.

[41] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," 2019.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR 2016, Las Vegas, NV, USA, June, 2016*, pp. 770–778.

[43] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR 2017, Honolulu, HI, USA, July, 2017*, pp. 5987–5995.

[44] N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet V2: practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis., ECCV 2018, Munich, Germany, September, 2018*, pp. 122–138.

[45] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[46] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 55, no. 7, p. 3965–3981, Jul 2017.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent., ICLR 2015, San Diego, CA, USA, May, 2015,*.

[48] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geosci. Remote. Sens.*, vol. 56, no. 5, pp. 2811–2821, 2018.

[49] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 57, no. 9, pp. 6916–6928, 2019.

[50] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using cnn-capsnet," *Remote. Sens.*, vol. 11, no. 5, p. 494, 2019.

[51] Y. Liu, C. Y. Suen, Y. Liu, and L. Ding, "Scene classification using hierarchical wasserstein CNN," *IEEE Trans. Geosci. Remote. Sens.*, vol. 57, no. 5, pp. 2494–2509, 2019.

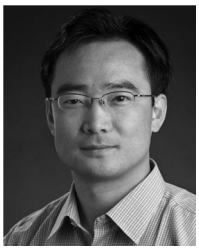
[52] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 57, no. 9, pp. 6530–6541, 2019.

Runyu Fan received the B.S. degree in computer science and technology from China University of Geosciences, Wuhan, China, in 2017, where he is currently working towards the Ph. D. degree in Geoscience Information Engineering from China University of Geosciences.

His research interests include time-series multivariate spatial data analysis, urban land surface change analysis and prediction, and remote sensing data analysis.

Runyi Feng received the B.S. degree in geographic information system from Hunan Normal University, Changsha, China, in 2011, and the M.S. degree in surveying and mapping engineering and the Ph. D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013 and 2016, respectively.

She has been with the School of Computer Science, China University of Geosciences (Wuhan) since 2016 and is currently an associate Professor. Her research interests include sparse representation, deep learning, hyperspectral image analysis, high-resolution remote sensing understanding, and intelligent interpretation of remote sensing imagery.



Lizhe Wang is the Dean and "ChuTian" Chair Professor at School of Computer Science, China University of Geosciences. He received B.E. and M.E from Tsinghua University in 1998 and 2001 and D.E. from Univeristy of Karlsruhe (Magna Cum Laude), Germany in 2007.

His research interests include remote sensing data processing, Digital Earth, Big Data Computing.

He is a fellow of the Institution of Engineering and Technology and British Computer Society and Associate Editor of Remote Sensing, International Journal of Digital Earth, ACM Computing Surveys, IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Sustainable Computing, etc. He is the recipient of Distinguished Young Scholars of NSFC, National Leading Talents of Science and Technology Innovation, 100-Talents Program of Chinese Academy of Sciences.

Jining Yan received his PhD in signal and information processing in the University of Chinese Academy of Sciences. He is an associate professor of School of Computer Science, China University of Geoscience.

His research is focused on remote sensing data processing, time-series analysis and change detection, cloud computing in remote sensing and applied oceanography.

Xiaohan Zhang received the B.S. degree in remote sensing science and technology from China University of Geosciences, Wuhan, China, in 2019, where he is currently working towards the Ph. D. degree in Geoscience Information Engineering from China University of Geosciences.

His research interests include machine learning, deep learning, unsupervised representation, and high-resolution remote sensing understanding.