Received January 12, 2019, accepted February 27, 2019, date of publication March 5, 2019, date of current version April 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903127

Deep Learning-Based Classification Methods for Remote Sensing Images in Urban Built-Up Areas

WENMEI LI[1,2,3], (Member, IEEE), HAIYAN LIU[2], YU WANG[2], (Student Member, IEEE), ZHUANGZHUANG LI[2], YAN JIA[1,3], AND GUAN GUI[2], (Senior Member, IEEE)

[1]School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2]College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[3]Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, Nanjing 210023, China

Corresponding author: Guan Gui (guiguan@njupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701258 and Grant 41401480, in part by the Jiangsu Specially Appointed Professor Program under Grant RK002STP16001, in part by the Summit of the Six Top Talents Program of Jiangsu under Grant XYDXX-010, in part by the Program for High-Level Entrepreneurial and Innovative Talents Introduction under Grant CZ0010617002, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170906 and Grant BK20180765, in part by the Natural Science Foundation of Jiangsu Higher Education Institutions under Grant 17KJB510044, in part by the Scientific Research Fund of Nanjing University of Posts and Telecommunications (NUPTSF) under Grant 218085, in part by the Nanjing Technology Innovation Foundation for Selected Overseas Scientists under Grant 20180075, and in part by the 1311 Talent Plan of Nanjing University of Posts and Telecommunications.

ABSTRACT Urban areas have been focused recently on the remote sensing applications since their function closely relates to the distribution of built-up areas, where reflectivity or scattering characteristics are the same or similar. Traditional pixel-based methods cannot discriminate the types of urban built-up areas very well. This paper investigates a deep learning-based classification method for remote sensing images, particularly for high spatial resolution remote sensing (HSRRS) images with various changes and multi-scene classes. Specifically, to help develop the corresponding classification methods in urban built-up areas, we consider four deep neural networks (DNNs): 1) convolutional neural network (CNN); 2) capsule networks (CapsNet); 3) same model with a different training rounding based on CNN (SMDTR-CNN); and 4) same model with different training rounding based on CapsNet (SMDTR-CapsNet). The performances of the proposed methods are evaluated in terms of overall accuracy, kappa coefficient, precision, and confusion matrix. The results revealed that SMDTR-CNN obtained the best overall accuracy (95.0%) and kappa coefficient (0.944) while also improving the precision of parking lot and resident samples by 1% and 4%, respectively.

INDEX TERMS Deep learning, convolution neural network, urban built-up area, capsule network, model ensemble, high resolution remote sensing classification.

I. INTRODUCTION

In recent years, deep learning has been applied in many applications, such as computer vision and wireless communications [1]–[16]. Compared with traditional pixel-based methods (e.g., minimum distance supervision classification [17], iterative self-organization (ISO) cluster unsupervised classification [18], support vector machine (SVM) classification [19], random forest classification [20]), deep learning is considered to be an effective method for extracting multi-layer features that often contain abstract and semantic information [10]. Hence, deep learning plays an important role in the field of target detection and classification. Current deep learning models have managed to offer a baseline for the use of deep learning in high spatial resolution remote sensing (HSRRS) image applications. One of the representative algorithms in deep learning is the neural network, which includes the deep belief network (DBN) [21], recurrent neural network (RNN) [22], and convolutional neural network (CNN) [23]. Many CNN algorithms have been successfully applied as powerful information extractors in computer vision, natural language processing (NLP), and medical and remote sensing image processing [8], [24], [25].

Remote sensing scene classification (RSSC) can provide a series of semantic classes which can assist in land cover and land use classification. HSRRS images with higher

The associate editor coordinating the review of this manuscript and approving it for publication was Yi-Zhe Song.

spatial resolution are typical categories for RCCS. Hence, HSRRS is often applied in urban mapping, target detection, precision agriculture, and natural resource management. Over recent years, extensive efforts have been made in developing feature representations and classifiers for the task of HSRRS image scene classification in broader areas of application. Urban areas have been focused on recently in remote sensing applications. Urban land cover classification, urban green space detection [26], hard target detection [27], urban flood [28], urban water and gas pollution [29], and so on have emerged with the occurrence and development of HSRRS imaging [30]. With the development of remote sensing technology, there are a number of HSRRS images, such as the UC Merced land use dataset [31], the SAT-4 and SAT-6 airborne datasets [32], the SpaceNet dataset [33], the remote sensing image classification benchmark (RSI-CB) dataset [34], and the Aerial Image Dataset (AID) [19]. The HSRRS image dataset displays texture and color information more clearly because of its higher spatial resolution. HSRRS images contain multiple scene classes, various changes compared with traditional remote sensing images, and are very hard to recognize with traditional pixel-based methods. Deep learning enables the object-level recognition and classification of HSRRS images, and has the potential to better understand the contents of HSRRS images at the semantic level.

Deep learning algorithms can automatically learn features from inputted raw data using deep architecture neural networks (e.g., CNN, CapsNet), and generate powerful deep learning features directly [6]. These algorithms have achieved a lot of results in scene classification and object detection. Urban built-up area detection and classification is one of the important applications, and its application with HSRRS images is of significance in practical applications. The classification of traditional remote sensing images in urban areas focusing on land cover, gives little or no consideration to urban functions [35], [36]. Meanwhile, urban functions relate with types of land use. For example, both residents and city roads belong to the same construct, but belong to different functional areas. In addition, urban planning and emergency response are connected with the functional partition of itself. Therefore, the classification of urban built-up areas is important for urban planning, urban ecological environment evaluation, urban emergency response, and so on.

In this paper, to demonstrate the different function and features of built-up area, two efficient approaches have been proposed to classify the urban built-up areas into different function area with HSRRS images based on deep learning. The result of classification will support the function analysis and provide ideas for urban functional zoning refinement. Meanwhile, in order to improve the precision of single category, two algorithms based on multi-model ensemble were put forward. The flowchart of our work is shown in Fig. 1. First, we apply both CNN and CapsNet architectures to find an approach that performs well with HSRRS images for urban built-up area scene classification. Furthermore, multi-model ensemble based methods were proposed to design a procedure with stable performance and extensible capabilities that could be extended into other area types or fields. The contributions of this paper can be summarized as below:

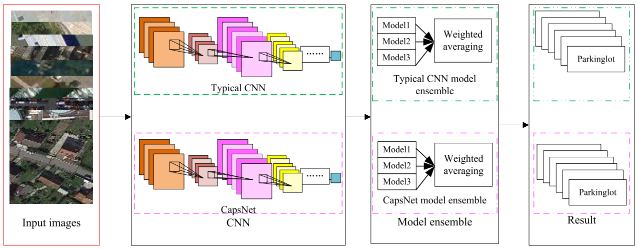


FIGURE 1. **The flowchart of urban built-up area scene classification.**

1. *Deep learning-based scene classification algorithm:* Considering the limited capability of traditional classification algorithms, especially in complex and large-scale built-up scenes based on HSRRS images, two deep learning-based methods, CNN and CapsNet, were proposed for more effective and powerful identification performance, because of their automatic multi-scale feature extraction abilities.
2. *Multi-model ensemble scene classification algorithm:* Two multi-model ensemble based algorithms have been put forward for better identification ability, due to the limited performances of CNN and CapsNet in single category recognition.

II. RELATED WORKS AND BACKGROUND

Land use and land cover classification with remote sensing images are based on the hypothesis that same things have the same or similar spectrums and different things display foreign spectrums. The traditional classification approaches applied spectral information in pixel scale to obtain a map of classes. Spatial information was then combined with spectral information for land use classification [37], subsequently improving classification accuracy. With the improvement in the spatial resolution of remote sensing images, semantic level information has obtained more attention in land use or land cover classification.

Over recent years, deep learning has been extensively studied and used due to the efficient feature extraction and performance improvements in computer vision and pattern recognition. Recently, CNN has been extended into remote sensing field, including target recognition and detection, land use and land cover classification with satellite or spaceborne remote sensing images. A large amount of research has been conducted on object or target detection using HSRRS images in urban areas. Region-based CNN (R-CNN) [38], Fast R-CNN [39], Faster R-CNN [40], and region-free methods have been proposed to extract features for linear SVM and then achieve the target category. Airplanes, building blocks, green spaces, and other objects have been focused on in urban area based on HSRRS images. Tayara and Chong *et al.* [41] proposed a uniform one-stage model for object detection based on CNN and obtained a better mean average precision and computation time. Zhai *et al.* [42] proposed a method

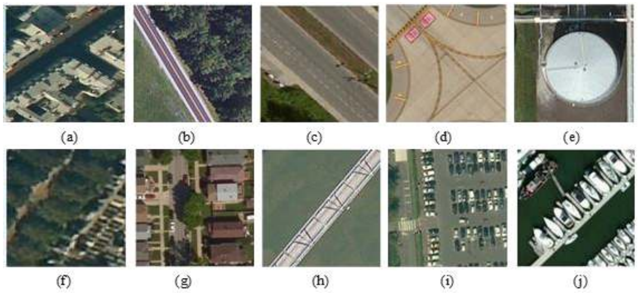


FIGURE 2. The HSRRS image dataset. (a)-(j) Building, avenue, road, airport, storeroom, roadside-tree, residents, bridge, parkinglot and marina, respectively.

based on a position-sensitive balancing (PSB) framework and residual network that takes full advantage of the fully connected network to detect 10-class objects. Although these methods could obtain object categories with higher precision and speed [41], only significant differences in characteristics classes were extracted. Meanwhile, when required to obtain many categories with the same or similar characteristics based on CNN, the traditional classifiers would be added. Usually, CNN would be used to extract features and then SVM or other classifiers followed to classify types of land use or cover (e.g., forest, farm, structure, grass, wasteland, water). These approaches are suitable for classifying small amounts of data and cannot discriminate or obtain finer function types in urban built-up areas. CNN as "end-to-end" models has proven to be more suitable for finer classification in urban built-up areas based on HSRRS images.

III. METHODOLOGY

The HSRRS image dataset for scene classification in urban built-up areas was constructed by collecting building, avenue, road, airport, storeroom, roadside tree, residents, bridge, parking lot, and marina images from the RSI-CB and UC Merced datasets. The size of one image is 128 × 128 pixels in 3 channels, and some examples of the dataset are shown in Fig. 2. It shows that some of the classes are similar in shape (e.g., roads and bridges, avenues and bridges), some are similar in color (e.g., residents and roadside trees, residents and buildings), and some are similar in texture (e.g., roadside trees and avenues, bridges and parking lots). All of these similarities make scene classification difficult in urban built-up areas.

A. CNN

Deep learning is characterized by "end-to-end" learning (e.g., feature learning, feature abstraction, model learning) and depends on a multi-layer task module to achieve the final goal. A CNN is a typical deep learning algorithm that is good at computer vision and image classification. In this paper, the typical structure of CNN will be described.

A CNN consists of one or several convolutional layers which are followed by fully connected layers that output the results. The structure of the typical CNN could be used for the 2-dimensional image. In our experiments, an input image is 128 × 128 × 3, which means the width and height of the image is 128 and 3 is the number of color channels. A convolutional layer has ρ kernels of size $m\ \times m\ \times\ h$, where m is smaller than 128 and h is equal to or smaller than 3. The kernels can generate a local connection structure and then convolve with the images to construct ρfeature maps with a size of $128 - m + 1$. A non-linearity function will then be performed on each feature map between the convolutional and pooling layers. After that, every feature map will be subsampled with maximum pooling over $\eta \times \eta$ local regions, and η is usually smaller than 5. Following the convolutional layers, some fully-connected layers will be placed to output the classification results. The overall architecture of a typical CNN with one convolution layer is shown in Fig. 3.

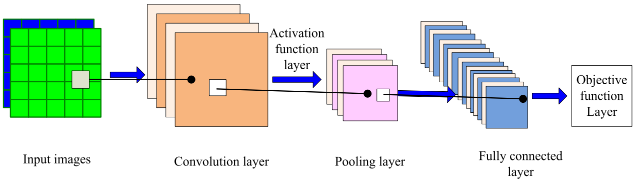


FIGURE 3. Typical CNN architecture.

The convolutional layer is the basic process of CNN, and is a local operation. Therefore, the local information of images can be obtained by putting certain size kernels on local images. Kernels (also named filters) are usually trained via network learning, and all kinds of kernels are focused on the basic patterns (e.g., boundaries, colors, shapes, textures) contained in a complex enough deep CNN. In addition, the "notation" representation will be abstracted by composing these kernels and followed by conducting the network operations. After that, basic and general patterns are replaced by conceptual representations, which are connected with the specific sample categories.

The pooling layer usually contains two types of pooling operation, one is average-pooling and the other is max-pooling. The pooling layer has no parameters to learn, and is only needed to assign super-parameters such as pooling type, kernel size, and stride. The pooling operation is a type of down-sampling and is also considered as a nonlinear convolution operation with p-norm.

The non-linearity mapping layer (also called the activation function layer) is introduced to increase the expression capability of the whole network. Rectified linear unit (ReLu) is one of the most popular activation functions, and is a segment function that helps with the fast convergence of the random gradient descent methods.

The fully connected layers act as classifiers, which map the learned feature representations to the sample labeled space. The role of the objective function layer is to measure the error between the predicted value and the labeled real sample. Usually, the cross-entropy loss function is applied in classification issues.

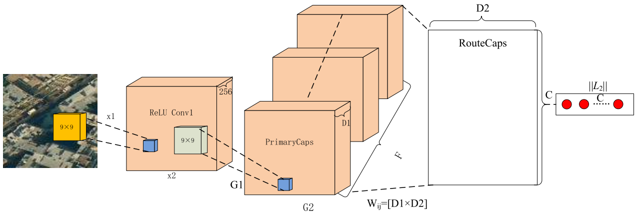


FIGURE 4. CapsNet architecture.

B. CapsNet

Fig. 4 shows the CapsNet architecture [43]. The first convolutional layer produces 256 feature maps with a 9 × 9 kernel and valid padding. In the PrimaryCaps layer, x1 and x2 depend entirely on the size of the input image. The size of each capsule is G1 × G2, which are computed based on x1 and x2. F represents the number of channels in the primary capsule, D1 and D2 are the dimensions of the output vector in the primary and touting capsules, respectively. C represents the number of classes.

As shown in Fig. 4, the input of the CapsNet fully connected layer is the linear weighted summation combined with the coupling coefficient, which can be obtained using the following formula:

$$s_j = \sum_i co_{ij}\hat{u}_{j/i}$$
$$\text{with } \hat{u}_{j/i} = We_{ij}u_i,$$
$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \tag{1}$$

where u is the output of the up-level CapsNet; We_{ij} is the weight coefficient; co_{ij} is the coupling coefficient; b depends on u; the initial of b is set as 0; the next layer of S can be obtained with b, u, and W_{ij}; and v_j is the activation function which is given by,

$$v_j = \frac{\left\|s_j\right\|^2}{1+\left\|s_j\right\|^2}\frac{s_j}{\left\|s_j\right\|} \tag{2}$$

C. ENSEMBLE LEARNING

The deep learning model ensemble usually contains "Data level" and "Model level" aspects. The data level ensemble focused on the augmentation in training, and this process could solve the problems of unbalanced samples.

Model level ensembles can contain single models or multi-models. A multi-layer ensemble containing single models is one of the most important model ensembles. As the semantic information of different layers can complement each other, a multi-layer ensemble could be used for semantic segmentation, image classification, and so on. Importantly, it links different layers in order to improve discrimination accuracy.

The multi-model ensemble approach contains four methods: simple averaging, weighted averaging, voting, and stacking. In our experiment, the weighted average generated by the same model with different training rounds (SMDTR) is applied as,

TABLE 1. Confusion matrix.

		True label				
		1	2	…	n	Total
Predicted label	1	C_{11}	C_{12}	…	C_{1n}	C_{1+}
	2	C_{21}	C_{22}	…	C_{2n}	C_{2+}
	…	…	…	…	…	…
	n	C_{n1}	C_{n2}	…	C_{nn}	C_{n+}
	Total	C_{+1}	C_{+2}	…	C_{+n}	N

$$S = \frac{\sum_{i=1}^{N} w_i s_i}{N}$$
$$\text{with } w_i \geq 0$$
$$\sum_{i=1}^{N} w_i = 1 \tag{3}$$

where w_i and s_i represent the weight and score of the i-th ($i = 1, 2, 3$) model, respectively. The w_i is defined as the ratio between the training accuracy of i-th model and the sum of the three model.

D. CLASSIFICATION EVALUATION INDICATORS

A confusion matrix is the most commonly used indicator in classification. As indicated in Table 1 (n classes as example), the horizontal is the predicted label and the vertical direction is the true label. The diagonal element is the number correctly classified.

Overall accuracy is another indicator for classification, and is applied to evaluate the proportion correctly classified. Hence, the overall accuracy can be given as,

$$\mathrm{OA} = \frac{1}{N}\sum_{i=1}^{n} C_{ii}$$
$$\text{with } i = 1, 2, \cdots, n \tag{4}$$

where C_{ii} is the number correctly classified for class i; n is the number of categories, and N is the number of the total sample. The kappa coefficient calculated from the confusion matrix is used to check consistency and evaluate classification precision. As indicated in formula (5), it not only considers the overall accuracy but also considers the variations in the number of samples in each category,

$$k = \frac{p_0 - p_e}{1 - p_e}$$
$$\text{with } p_0 = \frac{1}{N}\sum_{i=1}^{n} C_{ii}$$
$$p_e = \frac{1}{N^2}\sum_{i=1}^{n}(C_{i+} \times C_{+i})$$
$$i = 1, 2, \cdots, n \tag{5}$$

The precision is an indicator for measuring the accuracy of each category, and represents the number classified into class i by the model, which actually belong to the true class I ($i = 1, 2, \ldots\ldots, n$). It is also calculated from the confusion

TABLE 2. The number of each object category for training and testing.

Object category \ Number of samples	Training	Testing
Airport	578	100
Avenue	444	100
Bridge	369	100
Building	914	100
Roadside tree	321	100
Road	267	100
Marina	266	100
Parking lot	367	100
Residents	710	100
Storeroom	1207	100

TABLE 3. The types of data augmentation: image normalization, scaling, rotation, and shift.

Data augmentation	Value
Image normalization	1/255
Scaling factor	[0.9,1.2]
Rotation	[90°, 180°]
Width shift	0.1
Height shift	0.1

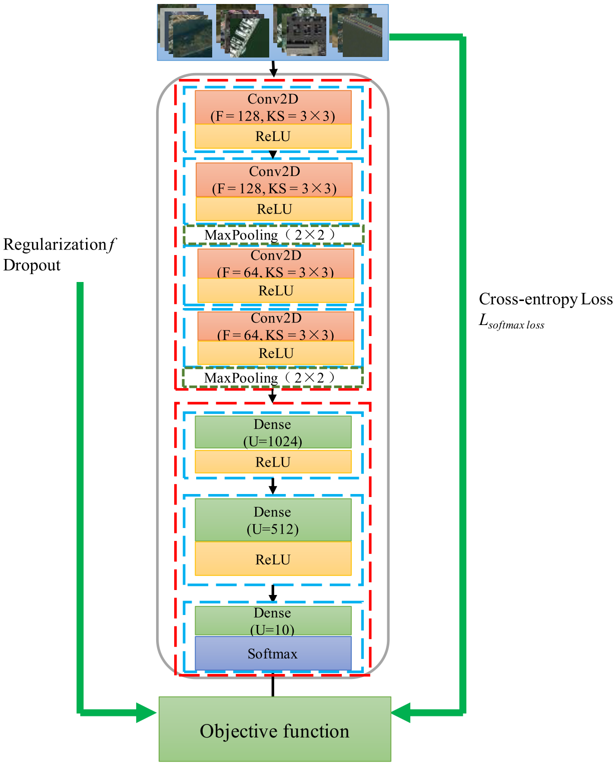


FIGURE 5. CNN architecture for HSRRS image classification.

matrix, and for the *i*-th class, we can obtain the precision for each category as

$$P_i = \frac{C_{ii}}{C_{+i}} \quad (i = 1, 2, \cdots, n) \tag{6}$$

IV. EXPERIMENTS

To obtain the best performance in terms of accuracy and stability, we designed several approaches for built-up area classification. There are 10 objects needing to be classified in our experiment, the numbers of each object for training and testing are shown in Table 2.

A. DATA AUGMENTATION

Effective data augmentation not only enlarges the number of training samples, but also increases their diversity. On one hand, it could avoid over fitting. On the other hand, it could improve the performance of the model [44]. As the number of our samples is within the range of [266, 1207], making it a limited sample, geometric transformations are proposed for data augmentation (Table 3). The types of data augmentation consisted of image normalization, scaling, rotation, width shift, and height shift. As the HSRRS image dataset collected by airborne or satellite is a little inclined, the performed shifts are on width and height, and both the rotation and scale factors are lower.

B. CNN-BASED CLASSIFICATION ARCHITECTURE

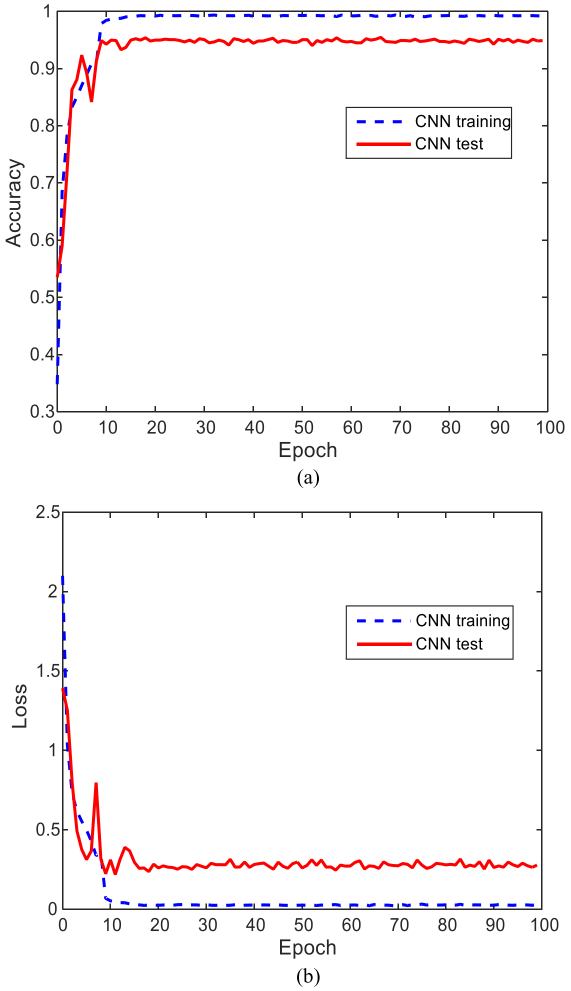
The HSRRS image scene classification architecture based on CNN is shown in Fig. 5. The purpose of the classification architecture is to preserve local details and extract semantic information. Under the consideration of the above, the designed architecture consists of three parts. The first part is the HSRRS image input, and the size of each image is 128 × 128 pixels, and the number of channels is 3. In the second part, four convolutional layers are applied to extract features. Meanwhile, two layers of max-pooling (2 × 2) are inserted into the CNN layers to reduce the parameters and keep useful features. The third part is a fully connected layer for image classification.

1) FEATURE EXTRACTION

CNN-based feature extraction consists of three aspects: convolutional layers, activation function layers, and pooling layers. As shown in Fig. 5, there are four convolutional layers in our architecture. That is two 3 × 3 kernels with a dimension of 128, and 3 × 3 kernels with a dimension of 64. The ReLU activation function is applied in our experiment as it uses a threshold to get the activation value without other operations. So it is faster in the speed of convergence of the network based on ReLU. To avoid over-fitting and reduce the number of parameters, a max-pooling layer with 2 × 2 kernels is applied after each convolutional layer.

2) CLASSIFICATION STRATAGEM

The fully connected layers are applied to combine the features with previous edges. In our work, we used three fully connected layers with 1024, 512, and 10 (number of category) neurons, respectively, to connect with the next convolutional layer. The Softmax model is usually exploited to calculate the probability of each category.



(a)

(b)

FIGURE 6. **Variations in (a) accuracy and (b) loss during training and testing epochs of CNN.**

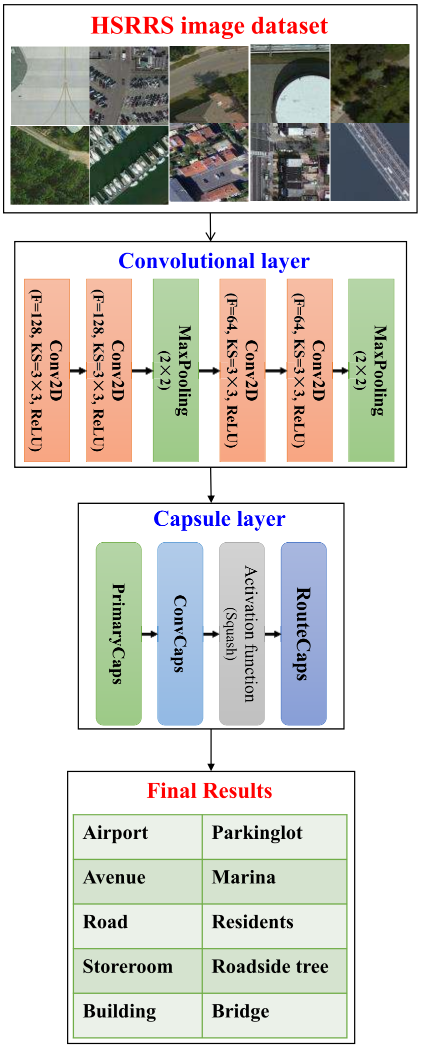
3) LOSS FUNCTION AND REGULARIZATION

The cross-entropy loss (also called Softmax loss) function is the most popular objective classification function in CNN. It is defined as:

$$L_{cross-entropy\ loss} = L_{soft\max loss} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{h_{y_i}}}{\sum_{j=1}^{C} e^{h_j}}\right)$$
$$\text{with } y_i \in \{1, 2, \ldots\ldots, C\}$$
$$h = (h_1, h_2, \ldots\ldots, h_C)^T \tag{7}$$

where C is the number of categories, y_i is the real label, and h is the final output of the network, which is also called the forecast result of the sample i.

Dropout is the most commonly used regularization in CNN, which equips with the fully connected layer. It not only reduces the complexity of the network, but is also an effective ensemble learning method in deep learning models. The principle of dropout is that the weight of the neuron is set to 0 randomly with probability p for each neuron in every layer in training, and all of the neurons are active with their weight multiplied by $(1 - p)$ to ensure the weights of the training and testing possess the same expectations. The variations in accuracy and loss in training and testing epochs based on CNN are shown in Fig. 6 (a) and (b), respectively. It shows that the accuracy of testing is lower than that of training, and the loss of testing is higher than that of training when the model is convergent. Moreover, both accuracy and loss fluctuated abruptly around the 7-9 testing epoch.



Airport	Parkinglot
Avenue	Marina
Road	Residents
Storeroom	Roadside tree
Building	Bridge

FIGURE 7. **CapsNet architecture for HSRRS image classification.**

C. CapsNet-BASED CLASSIFICATION ARCHITECTURE

Fig. 7 shows the CapsNet architecture for HSRRS image dataset classification in urban built-up areas. The architecture can be divided into two main layers: (1) a convolutional layer (including convolutional, activation function, and pooling layers), and (2) a capsule layer (including PrimaryCaps, ConvCaps, activation function, and RouteCaps). The function of the capsule layer is similar to the fully connected layer in CNN, but it can get vectors instead of scalars.

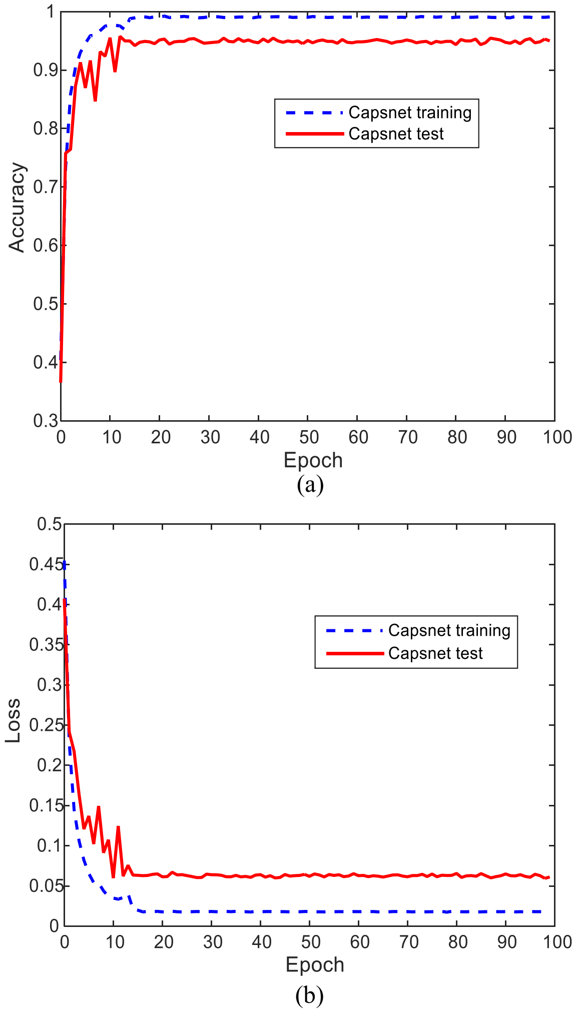


FIGURE 8. Variations in (a) accuracy and (b) loss during training and testing epochs of CapsNet.

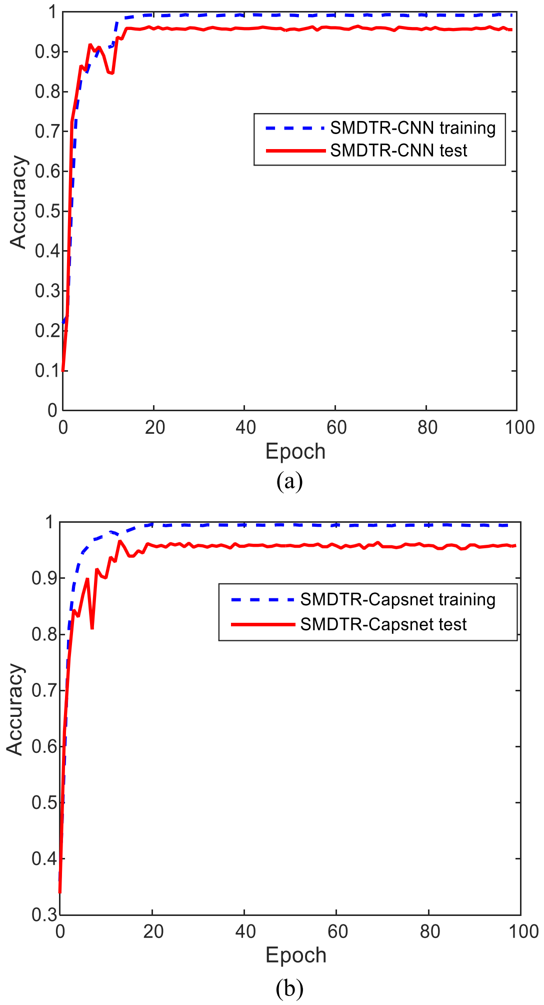


FIGURE 9. Iterations of training and testing accuracy changes on the HSRRS image dataset using the (a) SMDTR-CNN method and (b) SMDTR-CapsNet method.

D. MODEL ENSEMBLE ARCHITECTURE

The variations in accuracy and loss during training and testing epochs are shown in Fig. 8. It is easy to see that the accuracy and loss of training is better than that of testing when the model is convergent. In addition, both accuracy and loss fluctuated abruptly around the 5-10 testing epoch. This phenomenon tells us that the CapsNet model we trained and tested is suitable for finer classification in urban built-up areas.

In order to improve robustness, reduce the randomness of the model, and increase the accuracy of scene classification, a model ensemble architecture is considered based on CNN (cf. SMDTR-CNN) and CapsNet (cf. SMDTR-CapsNet). The flowchart is shown in Fig. 1. SMDTR was used to collect several models produced during the training epochs, with weighted averaging then being performed on selected models to construct a new model for urban built-up area classification. In Fig. 9(a), we can see the variations in accuracies in training and testing epochs using SMDTR-CNN and SMDTR-CapsNet approaches.

V. RESULTS AND ANALYSIS

The results can be divided into two parts: (1) an evaluation of the robustness and practicability of the model, and (2) an evaluation of scene classification precision in urban built-up areas.

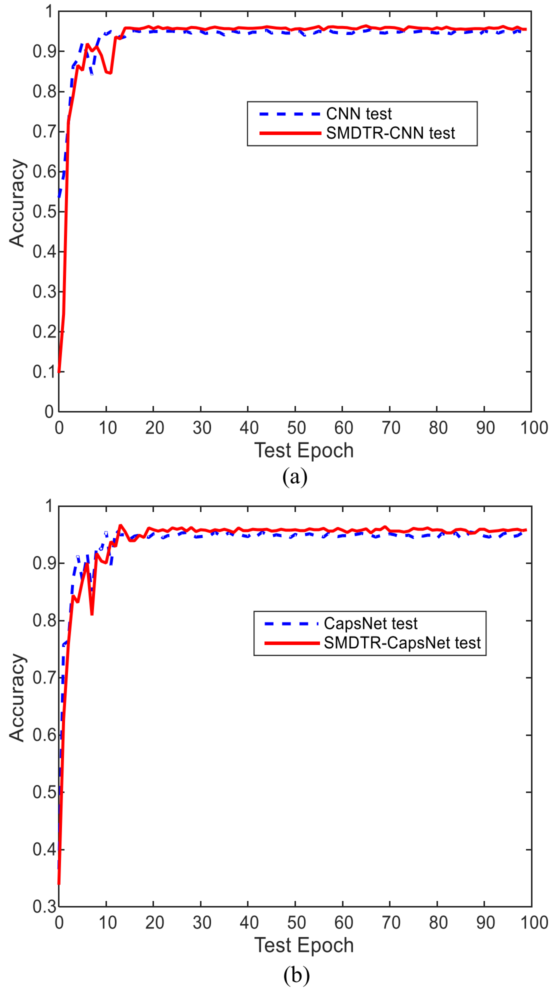
A. MODEL VALIDATION

Fig. 10 shows the variations in validation accuracy during the test epoch based on CNN and CapsNet. It can be seen that after about 20 epochs the accuracies of CNN and CapsNet converge. Although the accuracies of CNN and CapsNet fluctuated abruptly during the 3-15 epochs, with a worse accuracy for CNN, the convergent accuracy of CNN is higher than that of CapsNet. After model ensemble learning, the overall accuracy and kappa coefficient have been improved by about 0.2% and 0.002 for CNN, respectively.

B. CLASSIFICATION ACCURACY EVALUATION

1) CONFUSION MATRIX

The confusion matrices based on CNN, CapsNet, SMDTR-CNN, and SMDTR-CapsNet are shown in Fig. 11. The larger



(a)

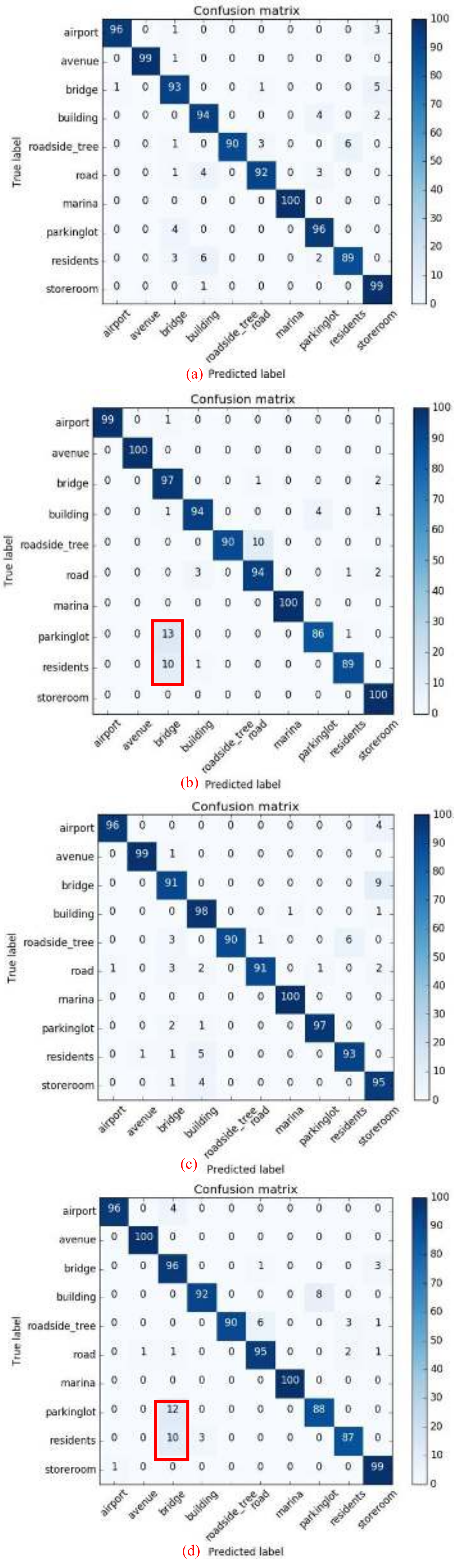
(b)

FIGURE 10. Validation accuracies of CNN, CapsNet, SMDTR-CNN, and SMDTR-CapsNet.

the number of diagonal elements, the better the classification is. From the figure, we can see that the parking lot and residents samples often divided into other classes by mistake in CNN, CapsNet, and SMDTR-CapsNet methods. 13% of parking lot samples and 10% of resident samples were mistakenly divided into bridge in the CapsNet method. This may correlate with the principle of CapsNet, which focuses on the spatial position relationship, as the spatial direction of the parking lot and residents samples seems similar to the bridge samples.

2) ACCURACY AND KAPPA COEFFICIENT

Table 4 shows the overall accuracy and kappa coefficient of the methods. It can be seen that the accuracy of SMDTR-CNN is 0.2% more than CNN. In addition, the accuracy of SMDTR-CapsNet is 0.6% lower than CapsNet. The three ensemble models in SMDTR-CapsNet are produced by searching local maximum, and there is one model produced with 7 training epoch which may not be appropriate for the whole dataset.



(a)

(b)

(c)

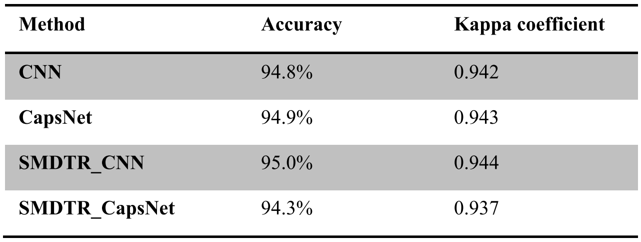
(d)

FIGURE 11. The confusion matrices based on (a) CNN, (b) CapsNet, (c) SMDTR-CNN, and (d) SMDTR-CapsNet.

3) PRECISION

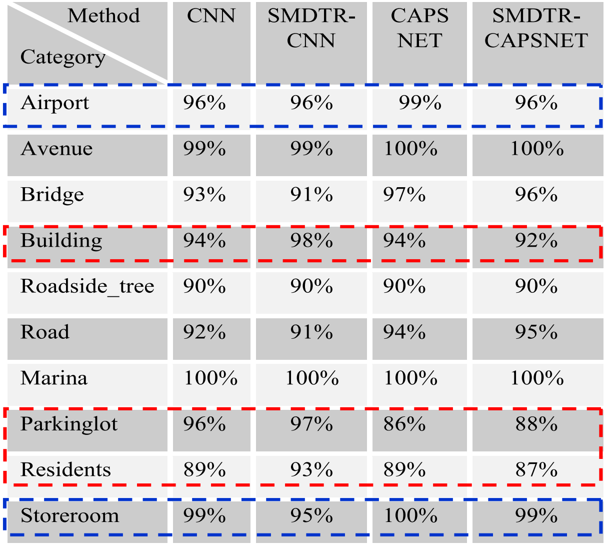
The precision of each category obtained by the four methods is displayed in Table 5. It indicates that the model ensemble

TABLE 4. The overall accuracy and kappa coefficient of the four methods.



Method	Accuracy	Kappa coefficient
CNN	94.8%	0.942
CapsNet	94.9%	0.943
SMDTR_CNN	95.0%	0.944
SMDTR_CapsNet	94.3%	0.937

TABLE 5. Category precision of the four methods.



Method Category	CNN	SMDTR-CNN	CAPS NET	SMDTR-CAPSNET
Airport	96%	96%	99%	96%
Avenue	99%	99%	100%	100%
Bridge	93%	91%	97%	96%
Building	94%	98%	94%	92%
Roadside_tree	90%	90%	90%	90%
Road	92%	91%	94%	95%
Marina	100%	100%	100%	100%
Parkinglot	96%	97%	86%	88%
Residents	89%	93%	89%	87%
Storeroom	99%	95%	100%	99%

based on CNN (cf. SMDTR-CNN) improved the precision of parking lot and residents samples by 1% and 4%, respectively, while reducing the precision of storeroom samples from 99% to 95%. The model ensemble based on CapsNet (cf. SMDTR-CapsNet) increased the precision of parking lot samples from 86% to 88%, while reducing the precision of resident samples from 89% to 87%, and airport samples from 99% to 96%. Our work indicates that it is shown that four DNN-based methods achieve different performances on different categories of images. It is possible to design more flexible ensemble learning method to finish the classification task in the future work.

VI. CONCLUDING REMARKS

This paper has developed a deep learning-based classification method for HSRRS images with various changes and multi-scene classes. In order to develop corresponding classification methods in urban built-up areas, we adopted four DNN, i.e., CNN, CapsNet, SMDTR-CNN, and SMDTR-CapsNet. The performances of the proposed methods have been confirmed in terms of accuracy, kappa coefficient, precision, and a confusion matrix. The results revealed that SMDTR-CNN obtained the best overall accuracy (95.0%) and kappa coefficient (0.944), while also improving the precision of parking lot and resident samples by 1% and 4%, respectively.

There are still some problems to be studied further, for example just the SMDTR is applied to improve the capability of CNN and CapsNet method, and the ensemble of different models could be considered for better classification task further. What's more, more complex net such as AlexNet, VGGNet could be used in HSRRS image classification task.

REFERENCES

[1] Z. Rui, W. Ziyu, M. Zhanyu, W. Guijin, and X. Jing-Hao, "LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test," *Pattern Recognit. Lett.*, vol. 116, pp. 36–42, Dec. 2018.

[2] Z. Ma and A. Leijon, "Human skin color detection in RGB space with Bayesian estimation of beta mixture models," in *Proc. Eur. Signal Process. Conf.*, Jun. 2014, pp. 1204–1208.

[3] X. Ma, J. Zhang, Y. Zhang, and Z. Ma, "Data scheme-based wireless channel modeling method: Motivation, principle and performance," *J. Commun. Inf. Netw.*, vol. 2, no. 3, pp. 41–51, 2017.

[4] X. Li *et al.*, "Supervised latent Dirichlet allocation with a mixture of sparse softmax," *Neurocomputing*, vol. 312, pp. 324–335, 2018.

[5] F. Zhu *et al.*, "Image-text dual neural network with decision strategy for small-sample image classification," *Neurocomputing*, vol. 328, pp. 182–188, Feb. 2019.

[6] 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, May 2018.

[7] C. Chen, W. Gong, Y. Chen, and W. Li, "Learning a two-stage CNN model for multi-sized building detection in remote sensing images," *Remote Sens. Lett.*, vol. 10, no. 2, pp. 103–110, 2019.

[8] Y. Yu and F. Liu, "Aerial scene classification via multilevel fusion based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 287–291, Feb. 2018.

[9] B. Pan, Z. Shi, and X. Xu, "MugNet: Deep learning for hyperspectral image classification using limited samples," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 108–119, Nov. 2018.

[10] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2359–2367.

[11] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[12] H. Huang, J. Yang, Y. Song, H. Huang, and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, Sep. 2018.

[13] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, 2018.

[14] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018.

[15] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 852–855, Oct. 2018.

[16] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA based heterogeneous IoT with imperfect SIC," *IEEE Internet Things J.*, to be published. doi: 10.1109/JIOT.2018.2876152.

[17] M. E. Hodgson, "Reducing the computational requirements of the minimum-distance classifier," *Remote Sens. Environ.*, vol. 25, no. 1, pp. 117–128, 1988.

[18] K.-Y. Huang, "The use of a newly developed algorithm of divisive hierarchical clustering for remote sensing image analysis," *Int. J. Remote Sens.*, vol. 23, no. 16, pp. 149–168, 2006.

[19] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1, pp. 89–97, 2004.

[20] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2007.

[21] R. Zand *et al.*, "R-DBN: A resistive deep belief network architecture leveraging the intrinsic behavior of probabilistic devices," in *Proc. ACM Great Lakes Symp. VLSI (GLSVLSI)*, 2018, pp. 1–8.

[22] M. Y. Miao and M. Gowayyed, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. Autom. Speech Recognit. Understand.*, Dec. 2016, pp. 167–174.

[23] S.-C. B. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Netw.*, vol. 8, nos. 7–8, pp. 1201–1214, 1995.

[24] A. Ben Hamida, A. Benoit, P. Lambert, and C. Ben Amar, "Three dimensional deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[25] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[26] A. Canetti, M. C. Gárrastazu, P. P. de Mattos, E. M. Braz, and S. P. Netto, "Understanding multi-temporal urban forest cover using high resolution images," *Urban Forestry Urban Greening*, vol. 29, pp. 106–112, Nov. 2018.

[27] A. Milan, "An integrated framework for road detection in dense urban area from high-resolution satellite imagery and Lidar data," *J. Geograph. Inf. Syst.*, vol. 10, no. 2, pp. 175–192, 2018.

[28] Y. Wang, A. S. Chen, G. Fu, S. Djordjevi, C. Zhang, and D. A. Savić, "An integrated framework for high-resolution urban flood modelling considering multiple information sources and urban features," *Environ. Model. Softw.*, vol. 107, pp. 85–95, Sep. 2018.

[29] Z. Tane, D. Roberts, A. Koltunov, S. Sweeney, and C. Ramirez, "A framework for detecting conifer mortality across an ecoregion using high spatial resolution spaceborne imaging spectroscopy," *Remote Sens. Environ.*, vol. 209, pp. 195–210, May 2018.

[30] Y. Li and D. Ye, "Greedy annotation of remote sensing image scenes based on automatic aggregation via hierarchical similarity diffusion," *IEEE Access*, vol. 6, pp. 57376–57388, 2018.

[31] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.

[32] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: A learning framework for satellite imagery," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, vol. 37, 2015, pp. 1–10.

[33] D. Lindenbaum and T. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 1–10.

[34] Y. J. Liu and C. C. Hsu, "Exploring anxiety in Ignoring read messages of line-comparison in four stages of romance relationship," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage.*, Dec. 2016, pp. 1795–1799.

[35] W. Zhang, W. Li, C. Zhang, D. M. Hanink, X. Li, and W. Wang, "Parcel feature data derived from Google Street View images for urban land use classification in Brooklyn, New York Cityfor urban land use classification in Brooklyn, New York Cityretain," *Data Brief*, vol. 12, pp. 175–179, Jun. 2017.

[36] C. Zhang *et al.*, "An object-based convolutional neural network (OCNN) for urban land use classification," *Remote Sens. Environ.*, vol. 216, pp. 57–70, Oct. 2018.

[37] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Jun. 2018.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 580–587.

[39] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2015, pp. 1440–1448.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[41] H. Tayara and K. T. Chong, "Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network," *Sensors*, vol. 18, no. 10, p. 3341, 2018.

[42] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 235–253, Aug. 2018.

[43] S. Sabour, C. V. Nov, and G. E. Hinton. (2017). "Dynamic routing between capsules." [Online]. Available: https://arxiv.org/abs/1710.09829

[44] D. Duarte, F. Nex, N. Kerle, and G. Vosselman, "Multi-resolution feature fusion for image classification of building damages with convolutional neural networks," *Remote Sens.*, vol. 10, no. 10, p. 1636, 2018.

WENMEI LI (M'18) received the M.S. degree from Nanjing University, in 2010, and the Ph.D. degree from the Chinese Academy of Forestry, in 2013. She is currently an Associate Professor with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, where she has been working for her postdoctoral studies, since 2018. Her research interests include deep learning, optimization, image reconstruct, and their application in land remote sensing.

HAIYAN LIU received the B.E. degree in electronic information engineering from Hefei Normal University, Hefei, China, in 2018. She is currently pursuing the master's degree with the Nanjing University of Posts and Telecommunications, Nanjing, China. Her research interests include deep learning, optimization, and its application in remote sensing image processing.

YU WANG (S'18) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, optimization, and its application in wireless communications.

ZHUANGZHUANG LI received the B.E. degree in communication engineering from Xi'an Technological University, China, in 2018. He is currently pursuing the master's degree with the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include machine learning, optimization, and its application in remote sensing image processing.

YAN JIA received the double M.S. degree in telecommunications engineering and computer application technology from the Politecnico di Torino, Turin, Italy, and Henan Polytechnic University, in 2013, and the Ph.D. degree in electronics engineering from the Politecnico di Torino, in 2017. She is currently with the Nanjing University of Posts and Telecommunications. Her research interests include microwave remote sensing, soil moisture retrieval, global navigation satellite system reflectometry (GNSS-R) applications to land remote sensing, and antenna design.

GUAN GUI (M'11–SM'17) received the Dr.Eng. degree in information and communication engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2012.

From 2009 to 2014, he was a Research Assistant and a Postdoctoral Research Fellow with the Wireless Signal Processing and Network Laboratory (Prof. Fumiyuki Adachi laboratory), Department of Communications Engineering, Graduate School of Engineering, Tohoku University. From 2014 to 2015, he was an Assistant Professor with the Department of Electronics and Information System, Akita Prefectural University. Since 2015, he has been a Professor with the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China. He is currently engaged in the research of deep learning, compressive sensing, and advanced wireless techniques. He has published more than 200 international peer-reviewed journal/conference papers. He received the Member and Global Activities Contributions Award at the IEEE ComSoc and seven best paper awards at the ICC 2014, VTC 2014-Spring, ICC 2017, ADHIP 2018, CSPS 2018, ICNC 2018, and ICEICT 2019. He was selected as the Jiangsu Specially-Appointed Professor, in 2016, the Jiangsu High-level Innovation and Entrepreneurial Talent, in 2016, the 1311 Talent Plan of NJUPT, in 2017, the Jiangsu Six Top Talent, in 2018, and the Nanjing Youth Award, in 2018. He was an Editor of *Security and Communication Networks*, from 2012 to 2016. He has been an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, since 2017, *KSII Transactions on Internet and Information Systems*, since 2017, IEEE ACCESS, since 2018, and *Journal of Communications*, since 2019, and the Editor-in-Chief of the *EAI Transactions on Artificial Intelligence*, since 2018.

• • •