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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Identifying building rooftops in hyperspectral imagery using CNN with pure pixel index

Yao Li, Chengming Ye, Yonggang Ge, Jos é Marcato Junior, *Member, IEEE,* Wesley Nunes Gon çalvese, *Member, IEEE,* and Jonathan Li, *Senior Member, IEEE*

***Abstract*—Deep learning and traditional machine learning algorithms have been widely applied to enhance the classification accuracy in remote sensing images. However, due to the variety and changeability of buildings, identifying building rooftops based on remote sensing images is still a challenge. Taking advantage of hyperspectral remote sensing imagery and spectroscopy, we propose a deep Convolutional Neural Networks (CNN) approach with Pure Pixel Index (PPI) constraints, named CNNP, to identify building rooftops materials. The framework, which accepts two kinds of data cubes as input data, extract spectral and spatial information by using 1D CNN and 3D CNNs with different kernel size, respectively. After the feature extraction, aiming to identify different building materials, the output of the top layer is the input to a classifier in a ratio decided upon by the PPI of the central pixel. To verify the effectiveness, we use Hyperion and Push-broom Hyperspectral Imager (PHI) data sets that represent high and low spatial resolution images to compare our proposed method with other traditional remote sensing image classification approaches, such as: Support Vector Machine (SVM); Stacked Auto-Encoders (SAE); Deep Belief Network (DBN); 1D CNN; and 2D CNN; 3D CNN; MiniGCN. The quantitative and qualitative results show that compared to other representative methods, CNNP achieves better performance, for both kinds of data, on Hyperion and PHI data sets with Overall Accuracy (OA) of 98.83% and 99.82%, respectively. And, the proposed method also provides an innovative idea for constructing other frameworks of hyperspectral image classification.**

***Index Terms*—Deep Learning, Building rooftops, CNN, Pure pixel index, Hyperspectral imagery**

I. Introduction

THE social and economic development, especially in developing countries, contribute to rapid urbanization, which accelerates the formation of mega-cities [1], [2]. Despite the research carried out to support urban expansion, due to the lack of basic information about urban buildings, there still exist some problems in urban planning and construction that prevent residents from seeking a comfortable living environment [3], [4], [5]. At the same time, this information is vital to risk elements detection, pre-disaster risk assessment, and post-disaster damage assessment. Conventional field mapping provides highly accurate results, but at a considerable cost in manpower and material resources [6], [7]. Another potential problem is that, when widespread disasters such as earthquakes, floods, tsunamis, etc. occur, field mapping does not satisfy the requirement for timeliness [8], [9], [10], [11], [12]. To reduce the time to acquire building material information, remote sensing technology is an effective approach that has the advantage of providing higher spatial scale and temporal resolution. Over the last two decades, remote sensing has made remarkable progress in sensor and data processing methods, generating ground surface information from qualitative to quantitative methods [13], [14], [15]. Also, a large variety of satellite imaging sensors enable to record multi-resolution and full spectrum data of buildings [16]. Furthermore, with data such as that from optical satellite images, single buildings and urban areas that are different in scale, can be extracted using spectral, textural and spatial features, which are designed based on expert knowledge and experience [17], [18], [19]. Likewise, several studies have focused on creating a model to estimate building height and damage after disasters, using microwave remote sensing, such as Synthetic Aperture Radar (SAR) and Interferometric SAR (InSAR), detecting the scattering properties of individual buildings [9], [11]. In particular, with the development of Light Detection and Ranging (LiDAR) technology, extremely dense point cloud data can be obtained, and a group of methods has been applied to detect the 3D information of buildings [6]. However, due to the lack of enough detection data in some vital bands, the identification of the building type material is still a tough topic [20], [21]. A hyper-spectral remote sensing sensor can generate images with hundreds of spectral bands; therefore, the obtained data contains spectral and spatial information that provides a basis for the classification of building materials [22].

This work was supported in part by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) [Grant No. 2019QZKK0902], the Strategic Priority Research Program of the Chinese Academy of Sciences [XDA23090203], the Major International (Regional) Joint Research Project [Grant No. 41520104002] from the National Natural Science Foundation of China, the Key Research Program of Frontier Sciences, CAS [Grant No. QYZDY-SSW-DQC006], the Major Program of National Natural Science Foundation of China [Grant No. 41790432], National Council for Scientific and Technological Development, CNPq [Grant No. 304173/2016-9], and the CAPES/Print [Grant No. 88881.311850/2018-01].

Y. Li, and Y Ge are with Key Laboratory of Mountain Hazards and Earth Surface Process, Chinese Academy of Sciences, Chengdu 610041, China, with University of the Chinese Academy of Sciences, Beijing 100049, China, with Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu 610041, China (e-mail: YaoLiCD@hotmail.com; gyg@imde.ac.cn).

M. Ye is with Key Laboratory of Earth Exploration and Information Technology of Ministry of Education, Chengdu University of Technology, Chengdu 610059, China (e-mail: rsgis@sina.com).

J. M. Junior is with the Faculty of Engineering, Architecture and Urbanism, and Geography, Federal University of Mato Grosso do Sul, Brazil. (e-mail: jose.marcato@ufms.br).

W. N. Gon çalves is with the Faculty of Computer Science and Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Brazil. (e-mail: wesley.goncalves@ufms.br)

J. Li is with Departments of Geography and Environmental Management and Systems Design Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Labeling every pixel on hyperspectral images is one of the main research topics in the field of remote sensing imagery processing [23]. Formerly, several conventional machine learning algorithms (so-called "shallow" methods), such as the Support Vector Machine (SVM), nearest neighbor, maximum likelihood, minimum distance, and decision tree methods were applied to HSI classification [24], [25], [26]. Among those methods, SVM is considered the state-of-the-art shallow approach that presents strong resistance to noise and the Hughes Phenomenon [27]. Although conventional methods have achieved remarkable performance, because of the lack of multiple feature mapping layers and complex spatial features, there is no room to obtain results of higher accuracy [28], [29]. Nowadays, hyper-spectral remote sensing image classification pays more attention to the combination of spectral and spatial information and high-level features. Hence, to acquire better classification results, some deep learning frameworks have been introduced to this field [30].

Basic models of deep learning (DL), which are stacked as deep learning frameworks in different ways, include the following: Restricted Boltzmann machine (RBM), Auto-Encoder (AE), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and their derivate [31], [32], [33], [34]. In the past decade, deep learning models have made remarkable achievements in many domains containing natural language processing, image recognition, and big data information extraction, thanks to the strong ability of DL in abstract feature representation that is significant for pattern recognition in a massive dataset. In computer vision (including remote sensing image classification), CNN is the most popular and versatile network architecture [35], [36]. In earlier research applied to hyper-spectral remote sensing image classification, Stacked Auto-Encoder (SAE) and Deep Belief Network (DBN) were employed, which were expected to extract the ground objects' spectral and spatial information [37], [28]. Unfortunately, by flattening the 3-D data cube to a 1-D vector, spatial information is lost. For this reason, researchers turned their attention to CNN, which can extract multi-scale spatial features and enhance the accuracy of the results [38]. But, a general problem is that mixed pixels, particularly at coarse spatial resolutions, exist in the majority of HSI, and researchers rarely consider the different contribution of spectral and spatial information to HSI classification (dense classification).

For instance, when it comes to labeling a pure pixel, greater importance should be given to the spectral information [39]. To the best of our knowledge, no study focused thus far on deep learning-based methods for the classification of rooftops materials using HSI. To overcome this, we propose a method (CNNP) to identify the building materials in hyper-spectral images based on Convolutional Neural Network (CNN) with Pure Pixel Index (PPI) constraints. Multi-scale 3-D CNN and 1-D CNN are applied to extract multi-scale spatial and spectral objects features. Then, those features, in a ratio decided by the PPI of the central pixel, are set as the input of the classifier at the top layer of CNNP. The main contribution of this paper consists of the following three aspects:

(1) We propose a deep learning framework to represent the features of building materials in HSI, which saves time and human resources to achieve high accuracy in extracting building material information and updating them on a large scale compared with field investigation.

(2) the scale effects widely exist in dense remote sensing images classification, especially for building rooftops identification. Here, the proposed framework uses convolution kernels with different sizes to synchronously extract multi-scale information, which is similar to transformation in Gaussian scale-space.

(3) In terms of feature fusion, PPI is used to decide the ratio of spectral and spatial features, which indicates that the extracted features contribute differently in identifying pixels in his and overcome the model overfitting when it faces small sample data of buildings.

The remainder of the paper is organized as follows: Section 2 introduces a detailed description of the proposed method. Section 3 describes the dataset of the study area, and the experimental results and discussion are presented in Section 4. Finally, Section 5 presents the conclusions.

II. Method and Experiments data

In this section, we briefly introduce CNN and PPI that have been used in this paper. Then we explain the idea why CNN and PPI are combined to construct the CNNP. Finally, we present the flow chart of the CNNP.

A. Convolutional Neural Network

As an important branch of the deep learning family, deep CNNs have been proved to significantly enhance the accuracy of image recognition compared with conventional machine learning methods [40], [41], [42]. In the past decade, in order to solve the bottleneck of remote sensing image classification, CNN is widely applied to extract effective features of remote sensing image and accordingly improve classification performance [43].

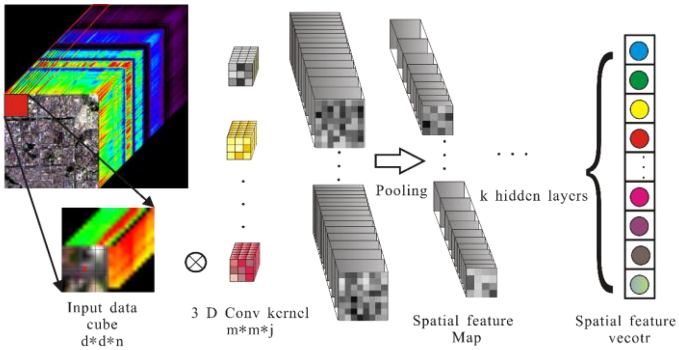

Fig. 1. 3-D Convolution neural network. The spatial and spectral information could be extracted from a 3-D data cube.

A base architecture of CNNs comprises an input, convolution, nonlinearity, pooling, and output layers Fig.1. Suppose that the input data cube is a subset $\mathbf{X} \in R^{h \times w \times d}$ from preprocessed images, where (h,w) and b represent the spatial size of the input data and the number of spectral bands. The output at position (x, y) of the j th feature map in the i th convolution layer is given as follows:

$$\mathbf{F}_{x,y}^{ij} = f(\sum_{m}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\mathbf{W}_{k_1k_2}^{ijm}\mathbf{F}_{(x-(K_1-1)/2+k_1)(y-(K_2-1)/2+k_2)}^{(i-1)jm}+\mathbf{B}_{ij}) \quad (1)$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

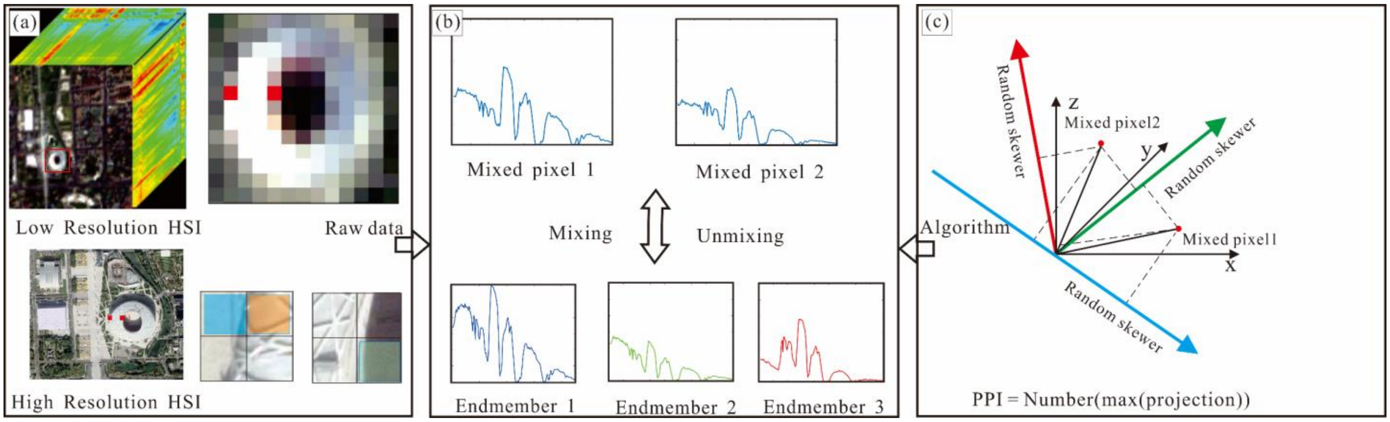

Fig. 2. Illustration of mixed pixels and PPI. (a) raw data from high and low resolution HIS. (b) The spectral characteristics of mixed pixels were composed by pure pixels linearly or nonlinearly. (c) Algorithm for calculating the projection value of each pixel on a random vector. Then $PPI_i = PPI_i + 1$, if the projection value of i th pixel is maximum compared with other pixels.

$$\mathbf{F}^{11} = f(\sum_{b}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\mathbf{W}_{k_1k_2}^{b}\mathbf{X}_{(x-(K_1-1)/2+k_1)(y-(K_2-1)/2+k_2)}^{b} + \mathbf{B}_{11}) \quad (2)$$

where $\mathbf{F}^{(i-1)jm}$ denotes the m th feature map in the $(i-1)$ th layer that connected to the j th feature map of the i th layer. K_1 and K_2 are the height and width of the convolution kernel, respectively; $\mathbf{w}_{k_1k_2}^{ijm}$ and $\mathbf{B}_{ij}$ represent the weights and bias of the j th feature map in the i th convolution layer, respectively. The $f(\cdot)$ is an activation function (such as Sigmoid function) that responsible for representing the complex and abstract nonlinear relationship in the data [44], [45], [46].

A large number of variables result in increased memory consumption and risk of over-fitting. To improve the robustness of the model, a pooling layer is designed to down-sample the $\mathbf{F}^{ij}$ in a specified window size after convolution processing. Down-sampling and sharing weights are two tricks that provide the CNNs the ability to extract spatial features unaffected by shift, scale, and distortion invariants in the images.

Although with those features, the performance in image recognition is improved at a relatively low computational cost. But, hyper-spectral sensors, which collect hundreds of narrow bands of ground objects, generate three-dimensional data cubes containing spatial and spectral information that helps to obtain higher accuracy in classification. Thus, it is extremely ineffective if CNNs are used only to extract the spatial information of the ground objects, without extracting spectral information.

Considering the characteristics of HSI, several innovative models, which convolutional filters could be 1-D, 2-D, or 3-D, are successfully introduced to hyper-spectral data set processing and classification. For instance, 1-D CNNs receive $b \times 1$ input vectors that consist of spectral vectors of each pixel, which means that classification is completed in the spectral domain. Different from 1-D CNNs, 2-D CNNs, which aim to obtain the spatial features of the central pixel, use a patch of $h \times h$ neighboring pixels as the input data.

However, before the extraction of the spatial information of the 2-D CNNs, there is always a reduction in dimensionality, which results in loss of information, especially in the spectral domain. Owing to the reason mentioned above, some hybrid frameworks, that combine 1-D and 2-D CNNs, are proposed to extract spatial and spectral information of unclassified pixels, respectively. In addition, 3-D CNNs, that receive raw data cubes created by stacking neighboring pixels on every band, are employed to obtain both spectral and spatial information simultaneously [40], [36].

B. Pure Pixel Index (PPI)

In general, there are inevitably some mixed pixels in a hyper-spectral image resulted from the limitation of the sensors and blurred boundaries between ground objects. Those mixed pixels consist of information of two or more kinds of objects, which are extremely common in urban areas (Fig. 2) [47], [48], [49]. In contrast to mixed pixels, pure pixels always contain information about one ground object and are considered the basic part of mixed pixels [50], [51], [52]. For several decades, hyper-spectral mixing models have been divided into two categories: linear and nonlinear mixture, described as follows:

$$\mathbf{x}_{lin} = \sum_{d=1}^{D}\mathbf{E}_d\mathbf{w}_d + \boldsymbol{\tau} = \mathbf{EW} + \boldsymbol{\tau}, s.t.\ \mathbf{w_d} \geq 0, \sum_{d=1}^{D}\mathbf{w}_d = 1 \quad (3)$$

$$\mathbf{x}_{non} = L(\sum_{d=1}^{D}\mathbf{E}_d\mathbf{w}_d) + \boldsymbol{\tau} \quad (4)$$

where $\mathbf{x} = [x_1, x_2, ..., x_n] \in \Re^n$ represents a pixel of a Hyper-Spectral Image with n bands; $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_d] \in \Re^{n\times d}$ is the endmembers matrix; d is the endmembers number; $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d] \in \Re^d$ is the weights, and $\boldsymbol{\tau}$ is assumed to be noise. For nonlinear mixture models, $L(\cdot)$ is a nonlinear kernel function.

To improve the classification accuracy of HSI, unmixing processing could be considered [53]. As mentioned above, there are two kinds of unmixing models. Among them, linear unmixing models are widely applied because of their low computational cost and explicit physical meaning. For linear hyper-spectral unmixing, extracting endmembers and estimating the abundance are two critical steps. Endmember

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

extraction can be classified into two categories: geometry-based and statistical-based. The geometry-based methods, including PPI, N-FINDR, and Simplex Shrink-Wrap Algorithm (SSWA), consider the endmember as the vertex of the convex [28], [54]. While the statistical-based methods including Nonnegative Matrix Factorization (NMF) and Independent Component Analysis (ICA) consider the eigenvectors as endmembers.

PPI, one of the pioneer linear unmixing models, which finds endmembers via a set of random vectors (skewers), has been very popular [55]. It is noteworthy that original hyper-spectral data should be submitted to the dimensionality reduction that has the ability to reduce the data redundancy and eliminating the noise interference [56], [54]. Fortunately, Chang and Du (2004) and Wang and Chang (2006) provide an effective approach for deciding the number of virtual dimensionalities that influence the information content of the processed data [57], [58].

C. Proposed Methods

Ground surface objects of urban areas are characterized by multi-scale and high density, which leads to the complexity and diversity of hyper-spectral data. There are many mixed pixels in HSI that correspond to small buildings or boundaries of different objects. On the contrary, pure pixels are mainly interior regions of large buildings. Considering a complex urban environment, we propose a CNN with PPI constraints (CNNP) to identify building materials, as shown in the Fig. 3. We innovatively adopt the PPI to adjust the ratio of the spectral and spatial information that are included in the classifier. We divide the CNNP into the following three steps:

(1) PPI ratio: PPI is a dimensionless parameter that cannot be input directly into the framework. For this reason, a normalization method is used to normalize the PPI [0, 1] by a linear projection. The normalization processing is given by the following equation:

$$PPI_r = PPI/N_K \tag{5}$$

where N_K is the number of skewers.

(2) Deep Spectral Feature Extraction: In the proposed model, a 1-D CNN, which is superior for extracting the features of 1-D vectors, is applied to extract spectral information of building materials. Different from other CNN architecture, 1-D CNN uses only the spectral values of a pixel as input data $\mathbf{x}_i = \left\{x_{i,1}, x_{i,2}, \ldots x_{i,n}\right\} \in R^n$, where n is the number of spectral bands, and i denotes the i_{th} pixel in the HSI, $X \in R^{h\times w\times n}$. To get the feature nonlinearly, the vector composed of the pixel's spectral reflectance is calculated as follows:

$$\mathbf{f}^1_{i,j,k} = f(\sum_{l=1}^{L} \mathbf{x}_{i,j-1,l} \times \mathbf{w}_{k,l} + b_k) \tag{6}$$

where j denotes the j_{th} spectral feature map; i denotes the i_{th} layer; $\mathbf{f}^1_{i,j,k}$ is the output of the K_{th} kernel; L is the depth of the convolution kernel; $\mathbf{W}$ is the weight vector; b is the bias parameter, and $f(\cdot)$ denotes the activation function. At the end of the convolution layer, a down-sampling pooling layer is used to provide sparse representation for the spectral information of images. In this way, convolution and pooling layers are alternately stacked to compose a deep architecture.

(3) Deep Spatial Feature Extraction: In recent years, especially with the sensor technology improvement, a large number of image classifiers have highlighted the importance of spatial information. To extract spatial information, many researchers have recently focused on adopting 2-D CNN, whose input data is a two-dimensional matrix consisting of the

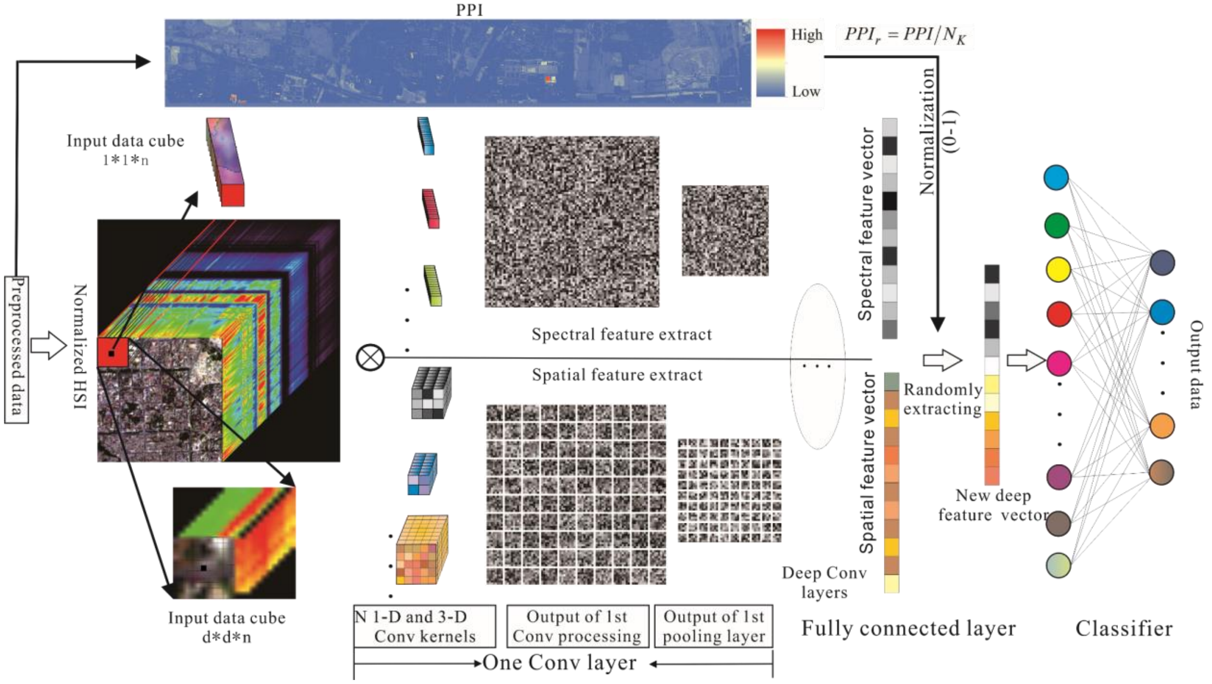

Fig. 3. CNNP framework. The spatial and spectral information is obtained by 1-D and 3-D convolution neural networks, then PPI act as constrains to decide the contributions of spatial and spectral information in objects identification.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

neighborhood values of the pixels in every band. However, there is some redundancy between the adjacent spectral bands in HSI. To enhance the efficiency of the feature extraction, Principal Component Analysis (PCA) or other dimension reduction algorithms are used to decrease the redundancy of raw data, which results in loss of information. We choose a 3-D CNN to complete the spatial feature extraction, whose input data is a raw data cube.

A single 3-D CNN layer consists of a convolution layer that takes a data cube as input data and a down-sampling pooling layer. In general, we select a 3-D CNN with a sized kernel to extract the features of one kind of ground objects, or group several kinds of 3-D CNN frameworks with different sizes of kernels for complex ground-surfaces. Given a data cube with the size of $d \times d \times n$, where n is the number of spectral bands, and d is the neighborhood size of the center pixel, the convolution layer output is formulated as follows:

$$\mathbf{F}_{i,j,k}^{2} = f\left(\sum_{m=1}^{M} \mathbf{x}_{i,j-1} \otimes \mathbf{w}_k + b_k\right) \qquad (7)$$

where j denotes the j_{th} spatial feature map; i denotes the i_{th} layer; $\mathbf{F}_{i,j,k}^{2}$ is the output of the k_{th} kernel; M is the depth of the convolution kernel; w is the weight vector; b is the bias parameter; $f(\cdot)$ denotes the activation function; $\otimes$ is the convolution calculation whose stride is "1" in every dimension.

As shown in Fig. 3, there are full connection layers at the ends of 1-D CNN and 3-D CNN that are flattened. Traditionally, feature vectors from full connection layers are the input to the classifiers, such as softmax, logistic regression, and SVM, that are responsible for classifying every pixel into a label. Different from other strategies, our proposed framework combines the spectral and spatial features in a proportion decided by PPI. The top feature layer combinations are formulated as follows:

$$\mathbf{F}_i = \left\{\mathbf{f}^1 \bullet dropout(1 - PPI_r), \mathbf{f}^2 \bullet dropout(PPI_r)\right\} \qquad (8)$$

where $\mathbf{F}_i$ denotes the i_{th} pixel's feature vector including spectral and spatial information; $\mathbf{f}^1$ and $\mathbf{f}^2$ are denotes the spectral and spatial feature vectors respectively. Note that two kinds of vectors are of the same length. It is a small trick and we do not pay special attention to it here; $dropout(\cdot)$ is an operation that randomly select the extracted feature.

After feature extraction, we choose softmax as the classifier, which is written as:

$$p(y_i \mid \mathbf{F}) = e^{\mathbf{F}_i} \Big/ \sum_{j=1}^{M} e^{\mathbf{F}_i} \qquad (9)$$

where $p(y_i \mid \mathbf{F})$ is probability that the pixel was labeled as y_i; $y_i \in (0,1)$ is the ground truth of a training sample.

Then, cross entropy is applied to construct the loss function and update the parameter $(\mathbf{W}, \mathbf{b})$ of the network, which is written as $j = -\sum_{i=1}^{K} y_i \log(p(y_i \mid \mathbf{F}))$, where K is the number of categories. Finally, Stochastic Gradient Descent (SGD) is used to train the parameters [59].

D.Data Set Description

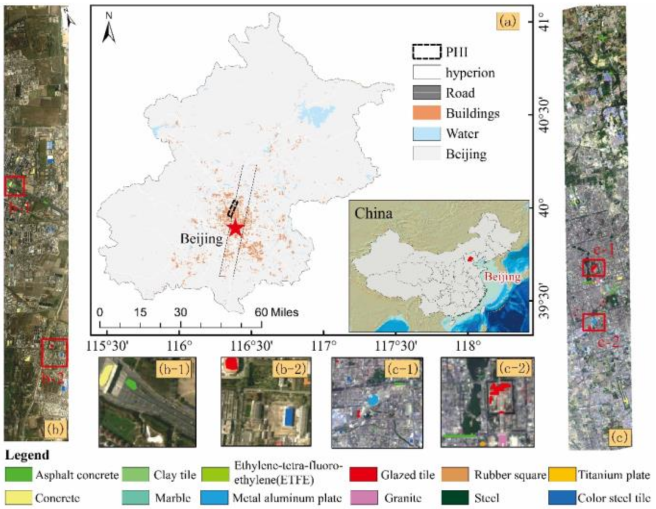

Fig. 4. Hyperion and PHI data sets, and samples. (a) Location of hyper-spectral images, (b) PHI data set, (c) Hyperion data set, (b-1)-(c-2) Zoom in on parts of two data sets.

In this section, two hyper-spectral data sets representing high and low resolution HSI are used to explore an optimal framework setting.

The first data set was acquired by the Earth Observing 1 (EO-1) Hyperion instrument that became operational on November 21, 2000, and has stopped operating on February 22, 2017 [60]. For sixteen years, the hyper-spectral data from EO-1 provided valuable material to research on remote sensing and scientific communities, and contributed significantly to the development of methods for dealing with hyper-spectral [61], [62], [63]. Hyperion sensors provide highly accurate radiometric images with 220 spectral bands in the range between 0.4 and 2.5 μm with 30 m of spatial resolution. The data used for this study was collected on May 10, 2017, in Beijing (Fig. 4). After the preprocessing, which includes atmospheric correction and removal of water absorption bands, 179 bands were retained for the follow-up analysis. To assess the model, by means of field investigation and visual interpretation with high resolution images, we labeled 845 pixels including 10 building classes (Table I). Figs.5a and 5b show the spectra of Color steel and Glazed tile in the Hyperion image.

TABLE I
TRAIN AND VALIDATION IN THE HYPERION DATA SET

No.	Class	Training (pixels)	Testing (pixels)
1	Asphalt concrete	36	144
2	Clay Tile	13	52
3	Color steel tile	13	49
4	Concrete	28	112
5	Ethylene-tetra-fluoro-ethylene(ETFE)	6	24
6	Glazed tile	24	94
7	Marble	7	25
8	Metal aluminum plate	23	92
9	Steel-Frame Construction	16	64
10	Titanium plate	5	18
Total		171	674

The second data set, also collected in Beijing, was acquired using an airborne push-broom Hyper-spectral Imager (PHI) that

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

contains 224 spectral bands in the range between 0.4 and 0.85 μm and with 536×3629 pixels with a spatial resolution of 1.2 m. Finally, a total of 3,097 pixels labeled in 15 kinds of building materials were used to train and test the model [64] (Table II). Figs. 5c and 5d show the spectra of the asphalt concrete and cement concrete in the PHI images.

TABLE II
TRAIN AND VALIDATION IN THE PHI DATA SET

No.	Class	Train(pixels)	Test(pixels)
1	Asphalt concrete	67	267
2	Color steel tile	97	386
3	Color read steel tile	25	100
4	Rubber square	10	40
5	Greenhouse	11	44
6	Glazed tile	37	147
7	Building 1	16	62
8	Building 2	46	181
9	Building 3	39	152
10	Concrete	97	386
11	Steel tile	119	473
12	Building 4	16	64
13	Clay Tile	23	91
14	Building 5	17	64
15	Building 6	4	15
Total		624	2472

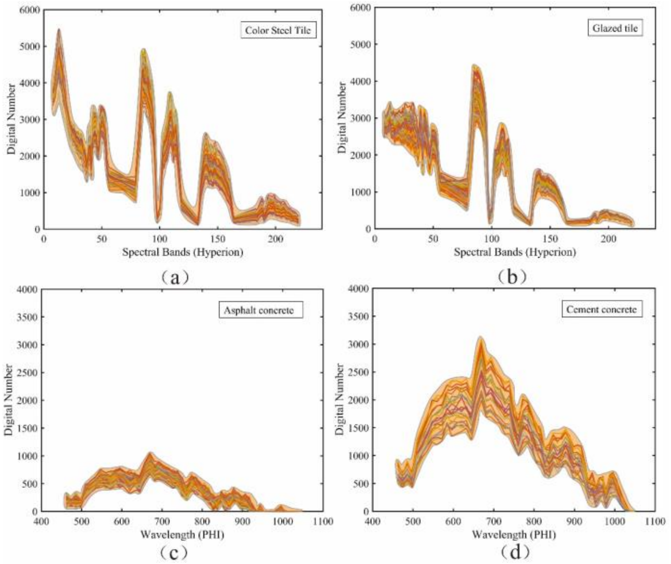

Fig. 5. Spectral of some building materials. (a) Color steel tile on Hyperion data set, (b) Glazed tile on Hyperion data set, (c) Asphalt concrete on PHI data set, (d)Cement concrete on PHI data set.

III. RESULTS AND DISCUSSION

In this section, our model was evaluated by using classification metrics (such as overall accuracy and Kappa coefficient). To improve the reliability of the results, we repeated each group of experiments 20 times with randomly selected training and testing data, and used the mean and deviation to represent the performance of the generated model. Then, we applied the optimal framework setting to evaluate the performance of the proposed models. And the final identification results are compared with some representative methods. For the performance metrics, we used the overall accuracy of all classes, denoted as OA, and the Kappa coefficient, denoted as Kappa.

A. Pure Pixel Index result

There are inevitable stripe noises that exist in some bands of the hyperspectral image, because the sensor is designed to collect the spectral reflectance of ground objects in a narrow band (bandwidth less than 10 nm), which makes the sensor overly susceptible to environment. These kinds of noises have an influence on the accuracy of PPI. Although a denoising algorithm is adopted in image preprocessing, the processed image still can't be used to deduce the PPI directly. Therefore, the minimum noise fraction transform is employed to reduce dimensionality and improve the signal-to-noise ratio of reserved components [65]. The vectors of pixels that are generated from reserved components, are then projected on random skewers and produced PPI. It should be noted that the frameworks' emphasis on the use of the difference among pixels, not highlighting the purity of the pixel. So, only 1000 random skewers are generated in an iterative process, which could avoid excessively discrete distribution of PPI (Fig 6).

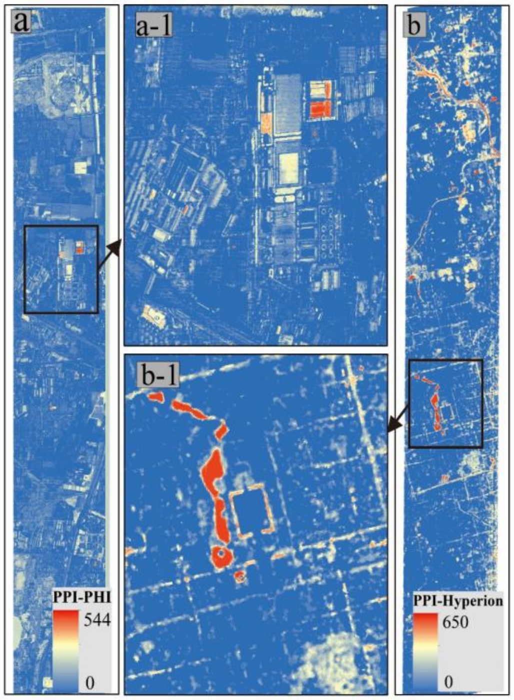

Fig. 6. PPI result. (a) PHI image, (b) Hyperion image.

B. Hyper-parameter Optimization

After designing the CNNP framework, we conducted several control experiments to comprehensively analysis the framework parameters, including the spatial size of the input data cubes, the setting of the convolution kernel (e.g., number and depth), and the learning rate. In order to give an objective and quantitative evaluation, the average, minimum and maximum Overall Accuracy (OA) are calculated over 100 repeated experiments with randomly selected samples. The optimal configurations of the model are preserved when the results get the highest average OA. We choose a 20%-80% training-test partition for the two data sets.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

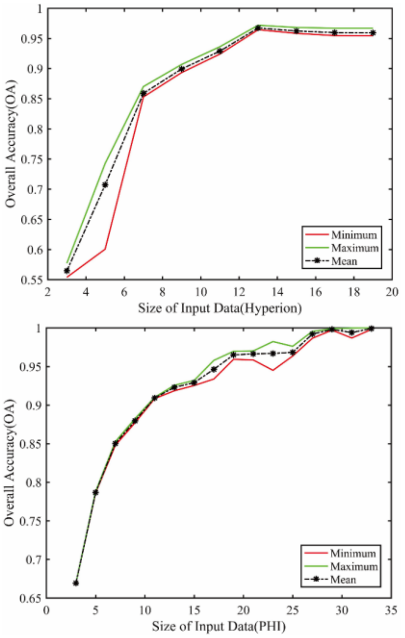

Fig. 7. OA (%) of framework with different size of input data.

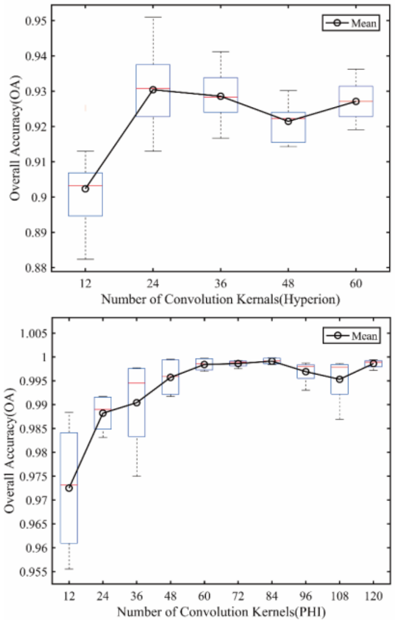

Fig. 8. OA (%) of different number of Convolution Kernel (CK).

The spatial size of the input data cube is an important factor of the framework because it decides the information for the input pixel and influences the efficiency of feature extraction. To analyze the relationship between the spatial size and the overall accuracy, we conduct a group of experiments with the same settings, except for the spatial size. Considering the resolution of data sets, we change the sizes from 3×3 to 19×19 for Hyperion data set, while the sizes ranged from 3×3 to 33×33 for PHI data set. All the results are shown in Fig. 7. From the figure, we can conclude that the spatial size of 13×13 is more suitable for Hyperion data. But, for PHI data, it is easier to obtain better performance in material identification with an input data cube with size 29×29. The different optimal sizes of the input data cube reveal that spatial and spectral information play different roles in material identification. In general, spatial information is more important for high resolution hyperspectral images classification, while spectral information for low resolution hyperspectral images.

On the one hand, the number and depth of the convolution kernels (CK) directly decide the number of CNNP parameters and how easily the framework is over-fitting. Specifically, more parameters of deep learning frameworks contribute to the framework for over-fitting with small data samples. On the other hand, the shallow depth and less number of CK also limit the ability of the frameworks. Therefore, it is unwise to decrease the depth and number of CK without considering the information extraction that further influences the accuracy of the identification. Two data sets are used to analyze the performance of the model in different CK numbers ranged from 12 to 120 in intervals of 12. All the results are shown in Fig. 8. It is clear to verify that the framework with 24 convolution kernels achieves the highest identification accuracy in the Hyperion data. For the PHI data set, the optimal number is 60. In terms of optimal depth of the convolution kernel, we carried out a series of experiments in different depths ranged from 2 to 14 in intervals of 2. As shown in Table III, the best depths of CK for the Hyperion and PHI data sets are 8 and 4, respectively. As shown in Table III and Fig. 8, the average OA increases along with the depth and number of CK at the beginning, because more parameters of CK contribute to learn more information to enhance the performance of identification. But, when the number of parameters reaches a certain level, the average OA tends to reduce, which means that the model encounters over-fitting problem.

TABLE III
OA (%) OF DIFFERENT DEPTH OF CONVOLUTION KERNEL (CK)

Depth of CK	Hyperion	PHI
2	89.07±0.51	94.85±032
4	89.58±0.48	**96.21±0.45**
6	90.56±0.34	95.67±0.56
8	**90.80±0.46**	95.94±0.41
10	89.70±0.38	95.69±0.38
12	88.60±0.52	94.38±0.42
14	89.52±0.56	94.53±0.48

Considering the efficiency of the learning process and avoiding local optimal solutions, we conducted several groups of experiments to determine the optimal learning rate by the grid research method. Firstly, the learning rate vectors, ranging from 0-1 and including 0.1, 0.01, 0.001, 0.0001, 0.00001 that are in different orders of magnitude, are applied to determine the optimal range. Then, we adopt the dichotomy with five iterations to obtain the optimal learning rates for Hyperion and PHI. Results suggest that 0.1 and 0.01 are supposedly the best learning rates for Hyperion and PHI, respectively. Finally, the framework with optimal setting is shown in Fig 9.

C. Building Rooftops Identification

We compared CNNP with the following traditional machine learning method (Support Vector Machine) and deep learning methods: Stacked Auto-Encoder [28], Deep Belief Network [37], 1D-CNN, 2D-CNN [66], 3D-CNN [67], and MiniGCNN [68]. The reasons why we choose those methods as competitors can be concluded as follows: (1) Because of their ability to classify high-dimensional considering small sample-size data sets. For example, SVM was recognized as the state-of-the-art model two decades ago. (2) SAE, DBN, 1D-CNN are good for

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

extracting spectral information from pixels in hyper-spectral image classification. (3) 2D-CNN, 3D-CNN and MiniGCN improve classification accuracy by using spatial information of objects.

The OAs and Kappa coefficients are presented in Tables IV and V for the two data sets. On one hand, CNNP performs better than the other methods on both data sets. And all the deep learning methods generate better results than SVM that is the representative traditional machine learning method. On the other hand, the Hyperion data provided lower accuracy for materials identification in all methods, indicating that the low-resolution HSI suffered from the disturbance of mixing pixels. This indicates a more challenging work to dense pixel-wise classification of low-resolution imagery compared to the high-resolution one.

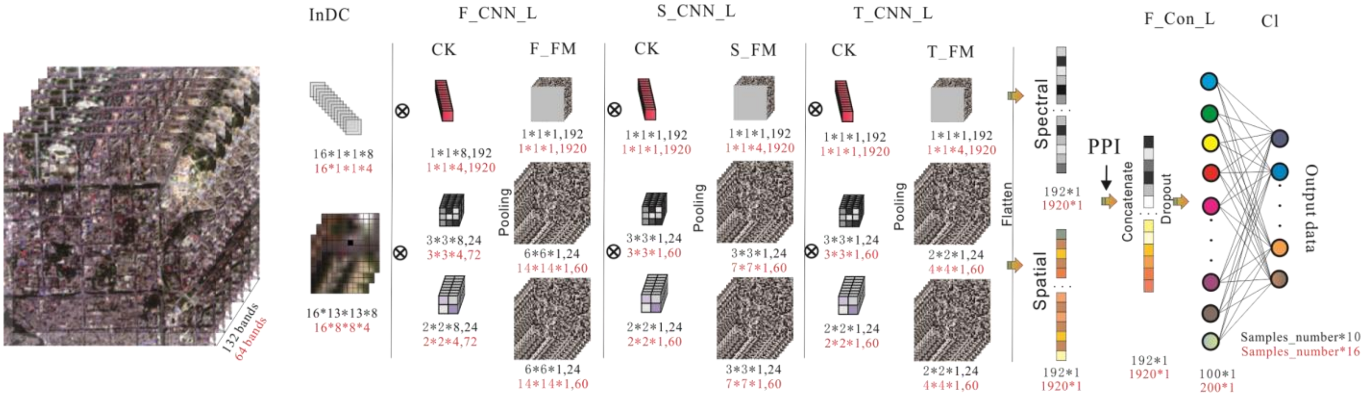

Fig. 9. CNNP framework with optimal setting. InDC (input data cube), CK(convolution kernel), F_FM (first feature map), F_CNN_L(first Convolution layer), F_Con_L (fully connected layer), CL (classifier). The activation function of CNNP is ReLu. The red label indicate framework setting on PHI data set, black labels for framework setting on Hyperion data set.

TABLE IV
QUANTITATIVE COMPARISON OF DIFFERENT ALGORITHMS IN TERMS OF OA, AND KAPPA COEFFICIENT ON THE HYPERION DATASET. THE BEST ONE IS SHOWN IN BOLD

No.	**SVM**	**SAE**	**DBN**	**1D-CNN**	**2D-CNN**	**3D-CNN**	**MiniGCN**	**CNNP**
Asphalt concrete	89.34	96.06	95.88	99.06	**99.4**	97.67	97.91	98.89
Clay Tile	87.97	81.48	84.78	86.22	82.76	89.26	86.22	**97.87**
Color steel tile	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.01	97.78	**100.00**
Concrete	88.16	93.65	95.38	89.26	96.55	96.86	96.98	**97.87**
Ethylene-tetra-fluoro-ethylene(ETFE)	**100.00**	**100.00**	**100.00**	**100.00**	93.10	99.01	**100.00**	**100.00**
Glazed tile	**100.00**	98.75	97.52	**100.00**	96.75	98.89	91.64	99.42
Marble	84.31	90.21	92.34	**100.00**	96.97	98.78	92.17	**100.00**
Metal aluminum plate	89.52	90.61	91.91	86.67	96.23	94.71	92.53	**98.58**
Steel-Frame Construction	88.78	81.26	80.22	**100.00**	90.91	88.55	89.24	**97.96**
Titanium plate	92.13	82.03	91.28	93.33	**100.00**	99.21	95.23	**100.00**
OA(%)	91.54	92.58	93.32	94.95	95.79	96.02	94.20	**98.85**
Kappa	0.9086	0.9202	0.9264	0.9393	0.9497	0.9522	0.9293	0.**9829**

TABLE V
QUANTITATIVE COMPARISON OF DIFFERENT ALGORITHMS IN TERMS OF OA, AND KAPPA COEFFICIENT ON THE PHI DATASET. THE BEST ONE IS SHOWN IN BOLD

No.	**SVM**	**SAE**	**DBN**	**1D-CNN**	**2D-CNN**	**3D-CNN**	**MiniGCN**	**CNNP**
Asphalt concrete	89.93	97.04	97.63	95.23	90.76	89.56	90.86	**98.90**
Color steel tile	97.29	**100.00**	**100.00**	99.79	**100.00**	**100.00**	**100.00**	**100.00**
Color read steel tile	95.17	96.27	96.07	94.92	99.21	95.42	**100.00**	**100.00**
Rubber square	92.96	100	100.00	97.83	**100.00**	**100.00**	**100.00**	**100.00**
Greenhouse	95.24	85.51	92.26	**100.00**	**100.00**	**100.00**	90.91	**100.00**
Glazed tile	97.28	**100.00**	**100.00**	98.91	**100.00**	**100.00**	96.28	**100.00**
Building 1	93.63	**100.00**	**100.00**	92.81	**100.00**	97.50	92.81	**100.00**
Building 2	92.66	94.12	94.18	98.63	76.1	85.58	82.59	**99.89**
Building 3	89.94	97.17	98.11	83.59	98.66	97.45	98.5	**99.48**
Concrete	98.95	92.8	92.76	96.55	97.51	96.29	94.31	**99.12**
Steel tile	97.01	**100.00**	99.31	90.87	**100.00**	**100.00**	99.04	99.62
Building 4	67.86	84.21	89.47	90.00	77.59	86.19	84.63	**100.00**
Clay Tile	83.46	90.54	82.59	87.78	85.02	87.67	85.00	**98.96**
Building 5	95.10	98.54	72.16	**100.00**	**100.00**	**100.00**	88.33	**100.00**
Building 6	95.24	84.55	82.14	75	**100.00**	**100.00**	88.50	**100.00**
OA(%)	94.28	96.65	95.89	94.71	95.62	96.02	94.68	**99.82**
Kappa	0.9402	0.9696	0.9512	0.9418	0.9491	0.9498	0.9351	**0.9921**

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

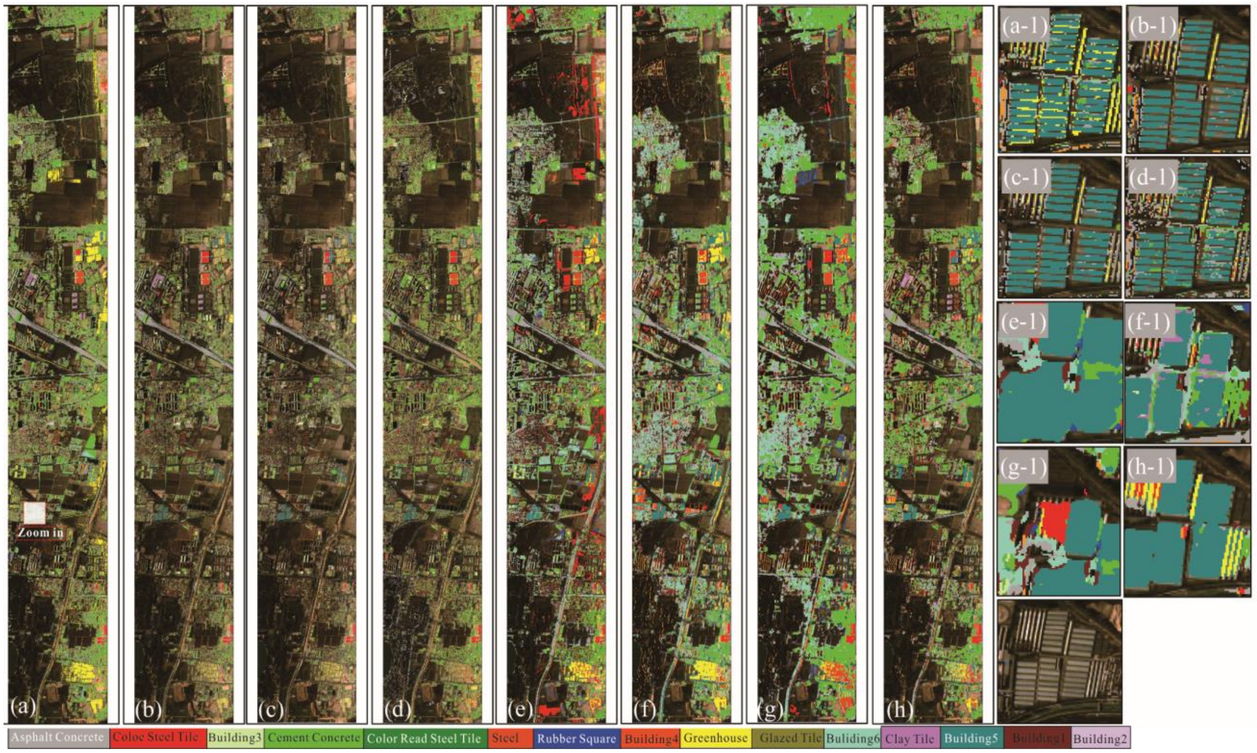

Fig. 10. The identification results of different methods on PHI data set. (a) SVM, (b) SAE, (c) DBN, (d) 1-D CNN, (e) 2-D CNN, (f) 3D-CNN, (g) MiniGCN, (h) CNNP, (a-1) - (h-1) Zoom in on a part of results. The enlarged area is full of a group of buildings that are neatly arranged. The identification result indicates that the methods depending only on spectral information (including SVM, SAE, DBN, 1-D CNN) are prone to missing some pixels in buildings. On contrary, the methods using spatial information is easy to expand the scope of identification. In conclusion, our proposed method has the best performance.

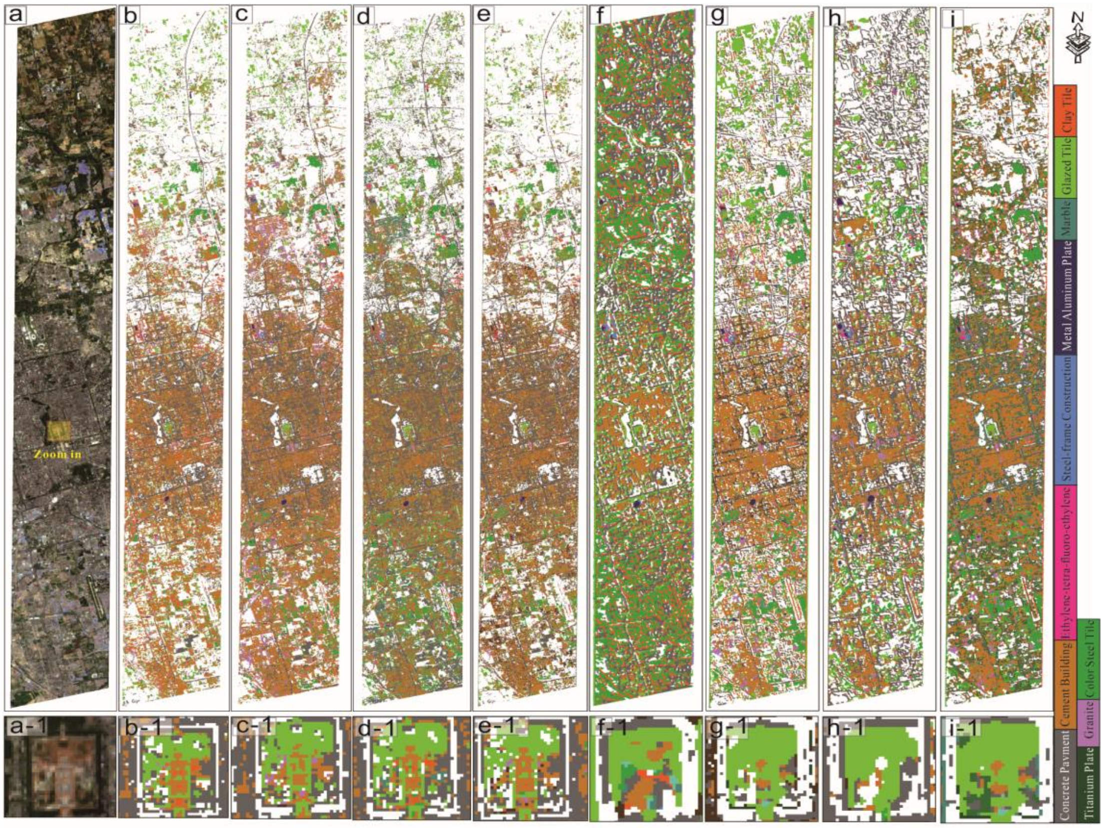

Fig. 11. The identification results of different methods on Hyperion data set. (a) SVM, (b) SAE, (c) DBN, (d) 1D CNN, (e) 2D CNN, (f) 3D-CNN, (g) MiniGCN, (h) CNNP, (a-1)-(f-1) Zoom in on a part of results. The enlarged area is the building besides the Tiananmen Square in Beijing. The most of pixels in this area are mixing pixels because the low resolution of image. The identification result indicates there is a significant difference between the methods depending only on spectral information (including SVM, SAE, DBN, 1-D CNN) and spectral-spatial used methods. In general, our proposed method can better identify the building completely compared with the other 7 methods.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Figs. 10 and 11 show the results for a whole image based on the best-trained frameworks and the true color images of the raw HSI. In both situations, when considering SVM, SAE, DBN, and 1D-CNN, which use only spectral information, the salt-and-pepper phenomenon adds noise to the classification maps. Although 2D-CNN applies spatial information to the identification and enhances the accuracy of the results, it also enlarges the area of the labeled image patches and decreases the robustness of the model. The identification result indicates that the methods depending only on spectral information (including SVM, SAE, DBN, 1-D CNN) are prone to missing some pixels in buildings. On contrary, the methods using spatial information extracted by one single size convolution kernel is easy to expand the scope of identification. CNNP, which combines high-level spectral and multi-scale spatial information, achieves the highest OA of 98.85% and 99.82%, respectively. Furthermore, because the constraints decided by the PPI provide a reasonable allocation of spectral and spatial information to the model, the integrity of the building objects, extracted by CNNP, is maintained. For example, the model will adjust the proportion of the spatial information to produce a better result when it comes to label a mixing pixel in a building object, to take the advantage of the spatial information of the neighboring pixels.

D. Train-Test Split Evaluation

A limited or imbalanced sample problem is very common in building rooftops identification, because there are various buildings to support traveling, shopping, entertainment, and sports of citizen, therefore, which call for an effective model to represent data and enhance classification accuracy. In this paper, we want to examine how the performance of the proposed CNNP on the limited dataset. To this end, we carried out several experiments with different a number of training samples from 10% to 90%, and reported the OA achieved by all methods. From Fig. 12, there are two results that could be observed. Firstly, except for CNNP, the classification accuracies of other methods drop dramatically or are inconsistent when the training samples are less than 30%, especially for 3D-CNN. It proves that although 3D-CNN has an advantage for spectral-spatial information extraction, it needs more training samples to obtain better classification performance because lacking of feature selection. On the other hands, CNNP has a better performance when facing small samples set, which probably because it decides the ratio of spectral and spatial information based on PPI, thus could pick up more representative feature to classification. The second result is that CNNP gets the highest classification accuracy in the situation of all kinds of train-test sample ratio. Therefore, all these observations indicate that the proposed CNNP is more effective than the baselines when sufficient training samples are provided.

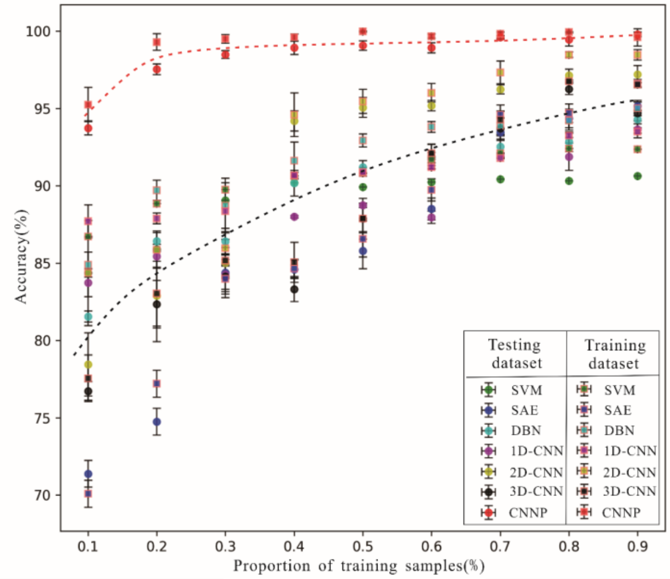

Fig. 12. Impact of train-test split ratio. The red and black dash line represents the changing trend of accuracy with increasing of training samples proportion on CNNP and traditional methods respectively.

E. Framework with PPI Avoid Overfitting

In the case of the same number of features, the effectiveness of the features determines the classification accuracy and robustness of the model, so as to avoid the overfitting phenomenon. With a small trick of dropout that could mitigate model overfitting, we conduct comparative test proves that the model added to PPI can further avoid the model overfitted. All Hyper-parameter parameters in the model used for comparison are the same as CNNP, including the number of top-level features. As shown in Fig. 13, the accuracy of the CNNP increases slowly and steadily, as the number of training epoch increases. However, the model without the PPI index as a constraint, is troubled by overfitting during the training process. Therefore, it is proved that PPI has made a contribution in feature selection, so as to avoid the influence of overfitting.

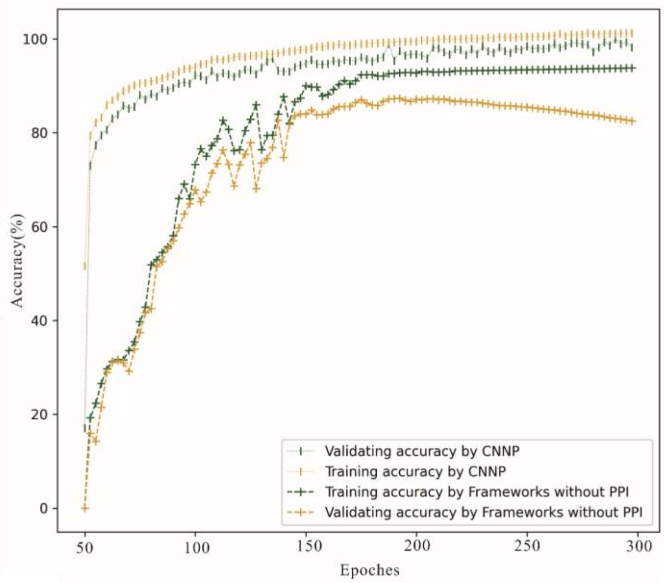

Fig. 13. The accuracy of CNNP and its same framework without PPI. The result show that PPI have a mitigation on overfitting.

IV. Conclusion

In this paper, we proposed a novel framework that contains a Convolution Neural Network (CNN) with Pure Pixel Index (PPI) constraints (CNNP) to identify building materials in the megacity. Firstly, the 1D-CNNs and 3D-CNNs are used to generate discriminative spectral and spatial information of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

buildings. Then, considering the negative impact of mixing pixels that exist widely in HSI, PPI is used as an index to decide the proportion of spectral and spatial information, which reflects the different contributions of the features for labeling a pixel. Experimental results demonstrate that CNNP obtains the highest identification accuracy in high and low resolution hyper-spectral images compared with other state-of-the-art high dimensional data classification methods.

There is no doubt that the reflectance spectrum can be regarded as an indicator of ground surface objects, especially when the spectral resolution is high enough. However, atmospheric perturbations and a complex near-surface environment magnify the variability of the spectral signatures. The deep learning method, which automatically extracts high-level spectral and spatial information from HSI without feature engineering, shows considerable success in representing data nonlinearly. Therefore, deep learning-based approaches perform better than the traditional shallow machine-learning algorithm in the two assessed data set. Different from conventional image recognition, identifying building materials in HSI is a dense pixel-wise mapping procedure, which poses the challenge of multi-scale effects. Thus, we adopted multi-scale 3D CNNs with different convolution kernel sizes to extract spatial information of buildings on different scales. Ultimately, two data sets representing high and low-resolution hyper-spectral images were applied in the experiments to validate the effectiveness of the CNNP. The results on both data set showed great potential for CNNP to identify building materials in other HSIs. The proposed method also provides an innovative idea for constructing other frameworks of hyperspectral image classification.

REFERENCES

[1]H. Taubenböck, T. Esch, A. Felbier, M. Wiesner, A. Roth, and S. Dech, "Monitoring urbanization in mega cities from space," *Remote Sens. Environ.,* vol. 117, pp. 162-176, Feb 2012.

[2]R. M. Garland *et al.*, "Aerosol optical properties in a rural environment near the mega-city Guangzhou, China: implications for regional air pollution, radiative forcing and remote sensing," *Atmos. Chem. Phys.*, vol. 8, no. 17, pp. 5161-5186, Sep 2008.

[3]E. Kalnay and M. Cai, "Impact of urbanization and land-use change on climate," *Nature.*, vol. 423, no. 6939, pp. 528-31, May 2003.

[4]H. Tran, D. Uchihama, S. Ochi, and Y. Yasuoka, "Assessment with satellite data of the urban heat island effects in Asian mega cities," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 8, no. 1, pp. 34-48, 2006.

[5]X. Xu, J. E. González, S. Shen, S. Miao, and J. Dou, "Impacts of urbanization and air pollution on building energy demands — Beijing case study," *Applied Energy*, vol. 225, pp. 98-109, 2018.

[6]T. Hermosilla, L. A. Ruiz, J. A. Recio, and J. Estornell, "Evaluation of Automatic Building Detection Approaches Combining High Resolution Images and LiDAR Data," *Remote Sens.*, vol. 3, no. 6, pp. 1188-1210, Jun 2011.

[7]L. Ma, M. C. Li, X. X. Ma, L. Cheng, P. J. Du, and Y. X. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 277-293, Aug 2017.

[8] R. Anniballe *et al.*, "Earthquake damage mapping: An overall assessment of ground surveys and VHR image change detection after L'Aquila 2009 earthquake," *Remote Sens. Environ.*, vol. 210, pp. 166-178, Jun 2018.

[9] J. M. Bodoque, C. Guardiola-Albert, E. Aroca-Jimenez, M. A. Eguibar, and M. L. Martinez-Chenoll, "Flood Damage Analysis: First Floor Elevation Uncertainty Resulting from LiDAR-Derived Digital Surface Models," *Remote Sens.*, vol. 8, no. 7, p. 604, Jul 2016.

[10] S. Ghaffarian, N. Kerle, and T. Filatova, "Remote Sensing-Based Proxies for Urban Disaster Risk Management and Resilience: A Review," *Remote Sens.*, vol. 10, no. 11, p. 1760, Nov 2018.

[11] Y. Q. Ji, J. T. S. Sumantyo, M. Y. Chua, and M. M. Waqar, "Earthquake/Tsunami Damage Assessment for Urban Areas Using Post-Event PolSAR Data," *Remote Sens.*, vol. 10, no. 7, p. 1088, Jul 2018.

[12] C. S. Zhang, "Towards an operational system for automated updating of road databases by integration of imagery and geodata," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, no. 3-4, pp. 166-186, Jan 2004.

[13] H. Ghassemian, "A review of remote sensing image fusion methods," *Infor. Fusion.*, vol. 32, pp. 75-89, Nov 2016.

[14] C. Toth and G. Jozkow, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 22-36, May 2016.

[15] F. D. van der Meer *et al.*, "Multi- and hyperspectral geologic remote sensing: A review," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 14, no. 1, pp. 112-128, 2012.

[16] A. F. H. Goetz, "Three decades of hyperspectral remote sensing of the Earth: A personal view," *Remote Sens. Environ.*, vol. 113, pp. S5-S16, Sep 2009.

[17] B. Huang, B. Zhao, and Y. M. 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, Sep 2018.

[18] L. Sun, Y. Q. Tang, and L. P. Zhang, "Rural Building Detection in High-Resolution Imagery Based on a Two-Stage CNN Model," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 1998-2002, Nov 2017.

[19] H. L. Yang, J. Y. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 8, pp. 2600-2614, Aug 2018.

[20] C. M. Ye, P. Cui, J. Li, and S. Pirasteh, "A method for recognising building materials based on hyperspectral remote sensing," *Mater. Res. Innovations.*, vol. 19, no. sup10, pp. S10-90-S10-94, Dec 2015.

[21] J. Zhu, L. Zhou, and D. Zhang, "Identification for building surface material based on hyperspectral remote sensing," in *2011 19th International Conference on Geoinformatics*, 2011: IEEE, pp. 1-5.

[22] Q. Tong, Y. Xue, and L. Zhang, "Progress in hyperspectral remote sensing science and technology in China over the past three decades," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 1, pp. 70-91, July 2013.

[23] P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu, "Multiple classifier system for remote sensing image classification: a review," *Sensors (Basel)*, vol. 12, no. 4, pp. 4764-92, Apr 2012.

[24] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *GeoSci. Front.*, vol. 7, no. 1, pp. 3-10, Jan 2016.

[25] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247-259, May 2011.

[26] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217-222, Jan 2005.

[27] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression With Active Learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085-4098, Nov 2010.

[28] Y. S. Chen, Z. H. Lin, X. Zhao, G. Wang, and Y. F. Gu, "Deep Learning-Based Classification of Hyperspectral Data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 6, pp. 2094-2107, Jun 2014.

[29] X. X. Zhu *et al.*, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8-36, Dec 2017.

[30] W. W. Song, S. T. Li, L. Y. Fang, and T. Lu, "Hyperspectral Image Classification With Deep Feature Fusion Network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173-3184, Jun 2018.

[31] S. Bernhard, P. John, and H. Thomas, "Greedy Layer-Wise Training of Deep Networks," in *Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference*: MIT Press, 2007, pp. 153-160.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016, pp. 770-778.

[33] A. Khan and N. Wahab, "Deep Residual Learning," 2016.

[34] T. Kuremoto, S. Kimura, K. Kobayashi, and M. Obayashi, "Time series forecasting using a deep belief network with restricted Boltzmann machines," *Neurocomputing.*, vol. 137, no. 15, pp. 47-56, Aug 2014.

[35] M. Zhang, W. Li, and Q. Du, "Diverse Region-Based CNN for Hyperspectral Image Classification," *IEEE Trans Image Process*, vol. 27, no. 6, pp. 2623-2634, Jun 2018.

[36] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847-858, Oct 2018.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

[37] Y. Chen, X. Zhao, and X. Jia, "Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief Network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 6, pp. 2381-2392, Jan 2015.

[38] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120-147, Dec 2018.

[39] X. Y. Wang, Y. F. Zhong, Y. Xu, L. P. Zhang, and Y. Y. Xu, "Saliency-Based Endmember Detection for Hyperspectral Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3667-3680, Jul 2018.

[40] Y. Li, H. K. Zhang, and Q. Shen, "Spectral-Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network," *Remote Sens.*, vol. 9, no. 1, p. 67, Jan 2017.

[41] M. Liang, L. Jiao, S. Yang, F. Liu, B. Hou, and H. Chen, "Deep Multiscale Spectral-Spatial Feature Fusion for Hyperspectral Images Classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 8, pp. 2911-2924, 2018.

[42] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised Spectral–Spatial Feature Learning via Deep Residual Conv–Deconv Network for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391-406, Oct 2018.

[43] X. R. Ma, A. Y. Fu, J. Wang, H. Y. Wang, and B. C. Yin, "Hyperspectral Image Classification Based on Deep Deconvolution Network With Skip Architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4781-4791, Aug 2018.

[44] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active Learning With Convolutional Neural Networks for Hyperspectral Image Classification Using a New Bayesian Approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6440-6461, Nov 2018.

[45] K. Jarrett, K. Kavukcuoglu, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *2009 IEEE 12th international conference on computer vision*, May 2009: IEEE, pp. 2146-2153.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807-814.

[47] R. Fernandez-Beltran, A. Plaza, J. Plaza, and F. Pla, "Hyperspectral Unmixing Based on Dual-Depth Sparse Probabilistic Latent Semantic Analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6344-6360, Nov 2018.

[48] J. Gao, Y. Sun, B. Zhang, Z. Chen, L. Gao, and W. Zhang, "Multi-GPU Based Parallel Design of the Ant Colony Optimization Algorithm for Endmember Extraction from Hyperspectral Images," *Sensors (Basel)*, vol. 19, no. 3, p. 598, Jan 2019.

[49] H. L. Li, J. Liu, and H. C. Yu, "An Automatic Sparse Pruning Endmember Extraction Algorithm with a Combined Minimum Volume and Deviation Constraint," *Remote Sens.*, vol. 10, no. 4, p. 509, Apr 2018.

[50] R. Arablouei, "Spectral Unmixing With Perturbed Endmembers," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 194-211, Dec 2019.

[51] M. Tang, B. Zhang, A. Marinoni, L. Gao, and P. Gamba, "Multiharmonic Postnonlinear Mixing Model for Hyperspectral Nonlinear Unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1765-1769, Aug 2018.

[52] B. Yang and B. Wang, "Band-Wise Nonlinear Unmixing for Hyperspectral Imagery Using an Extended Multilinear Mixing Model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6747-6762, Nov 2018.

[53] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process Mag.*, vol. 19, no. 1, pp. 44-57, Jan 2002.

[54] R. Heylen, M. Parente, and P. Gader, "A Review of Nonlinear Hyperspectral Unmixing Methods," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 6, pp. 1844-1868, Jun 2014.

[55] J. W. Boardman, F. A. Kruse, and R. O. Green, "Mapping target signatures via partial unmixing of AVIRIS data," presented at the Summaries 5th JPL Airborne Earth Science Workshop, 1995.

[56] C. I. Chang and C. C. Wu, "Design and Development of Iterative Pixel Purity Index," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 6, pp. 2676-2695, Jun 2015.

[57] C. I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608-619, Mar 2004.

[58] J. Wang and C. I. Chang, "Applications of Independent Component Analysis in Endmember Extraction and Abundance Quantification for Hyperspectral Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2601-2616, Aug 2006.

[59] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*, Heidelberg, Y. Lechevallier and G. Saporta, Eds., Sep 2010: Physica-Verlag HD, pp. 177-186.

[60] M. Marshall and P. Thenkabail, "Advantage of hyperspectral EO-1 Hyperion over multispectral IKONOS, GeoEye-1, WorldView-2, Landsat ETM plus , and MODIS vegetation indices in crop biomass estimation," *ISPRS J. Photogramm. Remote Sens.*, vol. 108, pp. 205-218, Oct 2015.

[61] L. Liu, J. Zhou, L. Han, and X. Xu, "Mineral mapping and ore prospecting using Landsat TM and Hyperion data, Wushitala, Xinjiang, northwestern China," *Ore Geol. Rev.*, vol. 81, pp. 280-295, Mar 2017.

[62] C. F. Waigl, A. Prakash, M. Stuefer, D. Verbyla, and P. Dennison, "Fire detection and temperature retrieval using EO-1 Hyperion data over selected Alaskan boreal forest fires," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 81, pp. 72-84, Sep 2019.

[63] Q. Zhang *et al.*, "Integrating chlorophyll fAPAR and nadir photochemical reflectance index from EO-1/Hyperion to predict cornfield daily gross primary production," *Remote Sens. Environ.*, vol. 186, pp. 311-321, Dec 2016.

[64] H. Shao, Y. Xue, and J. Wang, "Development of Chinese pushbroom hyperspectral imager (PHI)," in *Imaging System Technology for Remote Sensing*, 1998, vol. 3505: International Society for Optics and Photonics, pp. 108-116.

[65] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65-74, Jan 1988.

[66] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232-6251, 2016.

[67] H. Gong *et al.*, "Multiscale Information Fusion for Hyperspectral Image Classification Based on Hybrid 2D-3D CNN," *Remote Sens.*, vol. 13, no. 12, p. 2268, 2021.

[68] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2020.

Yao Li received the B.Sc. degree in saptial information and digital technology and M.Sc. degree in math from Chengdu Univeristy of Technology, China in 2015 and 2018, respectively. He is pursuing the Ph.D. degree in geotechnical engineering at the Institute of Mountain Hazards and Environment, CAS, Chengdu, China.

His research interests include geo-hazards mechanism and risk management, machine learning, SAR image processing, landslide prediction.

Chengming Ye received the Ph.D. degree in Earth Exploration and Information Technology from Chengdu University of Technology, China in 2011. He is currently an Associate Professor with the Departments of Geophysical, Chengdu University of Technology, Chengdu, China.

His main research interests comprise Geo-hazard remote sensing applications, ecological remote sensing, and LiDAR data processing.

Yonggsngi Ge received the Ph.D. degree in physical geography from the Institute of Mountain Hazards and Environment, CAS in 2009. He is currently an Professor with Key Laboratory of Mountain Hazards and Earth Surface Process, Chinese Academy of Sciences, Chengdu 610041, China.

His research interests include Geographic Information System and Geo-hazard risk assessment.

José Marcato Junior (M’17) received the Ph.D. degree in cartographic science from the Sao Paulo State University, Brazil. He is currently a Professor with the Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil.

His current research interests include UAV photogrammetry and deep neural networks for object detection, classification and segmentation. He has published more than 30 in refereed journals and over 70 in conferences,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2021.3127728, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

including papers published in ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Wesley Nunes Gonçalves (M’19) received the Ph.D. degree in computational physics from the University of Sao Paulo, Brazil. He is currently a Professor with the Faculty of Computer Science and Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil. His current research interests include computer vision, machine learning, deep neural networks for object detection, classification and segmentation. He has co-authored over 90 research papers, including Pattern Recognition, Pattern Recognition Letters.

Jonathan Li (Senior Member, IEEE) received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa. He is currently a Professor and the Head of the Mobile Sensing and Geodata Science Group, Department of Geography and Environmental Management, University of Waterloo, Canada. He has co-authored more than 400 publications, more than 200 of which were published in refereed journals, including the *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS)*, the *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS)*, the *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS)*, *ISPRS-JPRS*, and *RSE*.

His research interests include information extraction from LiDAR point clouds and from earth observation images. He is the Chair of the ISPRS WG I/2 on LiDAR, Air- and Space-borne Optical Sensing from 2016 to 2020 and the ICA Commission on Sensor-Driven Mapping for the period of 2019–2023, and the Associate Editor of the *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, and *Canadian Journal of Remote Sensing*.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/