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

Yao Li®, Chengming Ye®, Yonggang Ge, José Marcato Junior®, Member, IEEE, Wesley Nunes Gonçalves®, 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; MiniGCN. The quantitative and qualitative results show that compared to other representative methods, CNNP achieves higher classification accuracy in remote sensing images. However, due to the variety and changeability of 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].

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...
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 robustness 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 derivative [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.](image)

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 $X \in \mathbb{R}^{d \times w \times d}$ from preprocessed images, where $(b, d)$ 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:

$$F_{i,j}^{(x,y)} = f(\sum_{m} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} W_{i,j,m}^{(k_1,k_2)} x_{(i-k_1)/2+k_1} y_{(i-k_2)/2+k_2} + B_{i,j})$$ (1)
where $F^{(i-1)jm}$ 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; $w_{ij}^{mk}$ and $B_{ij}$ 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 $F^d$ 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 $h \times 1$ 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:

$$x_{lin} = \sum_{d=1}^{D} E_d w_d + \tau = E W + \tau, s.t. w_d \succeq 0, \sum_{d=1}^{D} w_d = 1$$

$$x_{non} = L(\sum_{d=1}^{D} E_d w_d ) + \tau$$

where $x = [x_1, x_2, ..., x_n] \in \mathbb{R}^n$ represents a pixel of a Hyper-Spectral Image with $n$ bands; $E = [e_1, e_2, ..., e_d] \in \mathbb{R}^{n \times d}$ is the endmembers matrix; $d$ is the endmembers number; $W = [w_1, w_2, ..., w_d] \in \mathbb{R}^d$ is the weights, and $\tau$ 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
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, especially with the sensor technology improvement, a large number of image classifiers have highlighted the importance of spatial and spectral information in objects identification.

C. Proposed Methods

(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:

\[ PPI' = \frac{PPI}{N_K} \]

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 \( x_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\} \in \mathbb{R}^n \), where \( n \) is the number of spectral bands, and \( i \) denotes the \( i_{th} \) pixel in the HSI, \( X \in \mathbb{R}^{h \times w \times n} \). To get the feature nonlinearly, the vector composed of the pixel’s spectral reflectance is calculated as follows:

\[ f_{i,j,k}^1 = f \left( \sum_{l=1}^{L} x_{i,j-l,1} \times w_{k,l} + b_k \right) \]

where \( j \) denotes the \( j_{th} \) spectral feature map; \( i \) denotes the \( i_{th} \) layer; \( f_{i,j,k}^1 \) is the output of the \( K_{th} \) kernel; \( L \) is the depth of the convolution kernel; \( W \) 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 spatial and spectral information in images. To divide the CNNP into the following three steps:

**Image Normalization:** 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.

**Ground surface objects of urban areas:**

The 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 the 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, especially with the sensor technology improvement, a large number of image classifiers have highlighted the importance of spatial and spectral information in objects identification.
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:

\[
F^2_{i,j,k} = f \left( \sum_{m=1}^{M} x_{i,j-1} \quad \otimes \quad w_{m} + b_{k} \right)
\]  

(7)

where \( j \) denotes the \( j_{th} \) spatial feature map; \( i \) denotes the \( i_{th} \) layer; \( F^{i}_{i,j,k} \) 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:

\[
F_i = \left\{ t^1 \cdot \text{dropout}(1 - \text{PPI}) , t^2 \cdot \text{dropout}(\text{PPI}) \right\}
\]  

(8)

where \( F_i \) denotes the \( i_{th} \) pixel’s feature vector including spectral and spatial information; \( t^1 \) and \( t^2 \) are 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; \( \text{dropout}(\cdot) \) is an operation that randomly select the extracted feature.

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

\[
p(y_j | F) = \frac{e^{F_j}}{\sum_{j=1}^{K} e^{F_j}}
\]

(9)

where \( p(y_j | F) \) is probability that the pixel was labeled as \( y_j : y_j \in \{0,1\} \) is the ground truth of a training sample.

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

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 \( \mu m \) 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).

![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.](image)

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>TRAIN AND VALIDATION IN THE HYPERION DATA SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>Class</td>
</tr>
<tr>
<td>1</td>
<td>Asphalt concrete</td>
</tr>
<tr>
<td>2</td>
<td>Clay Tile</td>
</tr>
<tr>
<td>3</td>
<td>Color steel tile</td>
</tr>
<tr>
<td>4</td>
<td>Concrete</td>
</tr>
<tr>
<td>5</td>
<td>Ethylene-tetra-fluoro-ethylene(ETF)</td>
</tr>
<tr>
<td>6</td>
<td>Glazed tile</td>
</tr>
<tr>
<td>7</td>
<td>Marble</td>
</tr>
<tr>
<td>8</td>
<td>Metal aluminum plate</td>
</tr>
<tr>
<td>9</td>
<td>Steel-Frame Construction</td>
</tr>
<tr>
<td>10</td>
<td>Titanium plate</td>
</tr>
<tr>
<td>Total</td>
<td>171</td>
</tr>
</tbody>
</table>

The second data set, also collected in Beijing, was acquired using an airborne push-broom Hyper-spectral Imager (PHI) that
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 (Table II). Figs. 5c and 5d show the spectra of the asphalt concrete and cement concrete in the PHI images.

### Table II

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

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.

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

### 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).

#### 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 quantiative 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.
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 $3 \times 3$ to $19 \times 19$. From the figure, we can conclude that the spatial size of $13 \times 13$ 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 $29 \times 29$. 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.

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...
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.

![Image](https://example.com/image.png)

**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**

<table>
<thead>
<tr>
<th>No.</th>
<th>SVM</th>
<th>SAE</th>
<th>DBN</th>
<th>1D-CNN</th>
<th>2D-CNN</th>
<th>3D-CNN</th>
<th>MiniGCN</th>
<th>CNNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt concrete</td>
<td>89.34</td>
<td>96.06</td>
<td>95.88</td>
<td>99.06</td>
<td>99.01</td>
<td>97.91</td>
<td>98.89</td>
<td></td>
</tr>
<tr>
<td>Clay Tile</td>
<td>87.97</td>
<td>81.48</td>
<td>84.78</td>
<td>86.22</td>
<td>89.26</td>
<td>86.22</td>
<td>97.87</td>
<td></td>
</tr>
<tr>
<td>Color steel tile</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>Concrete</td>
<td>99.18</td>
<td>96.55</td>
<td>96.86</td>
<td>96.86</td>
<td>96.86</td>
<td>96.86</td>
<td>96.86</td>
<td></td>
</tr>
<tr>
<td>Ethylene-tetrafluoro-ethylene(ETFE)</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>Glazed tile</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>99.42</td>
<td></td>
</tr>
<tr>
<td>Marble</td>
<td>84.31</td>
<td>90.21</td>
<td>92.34</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>Metal aluminum plate</td>
<td>95.79</td>
<td>96.23</td>
<td>94.71</td>
<td>92.53</td>
<td>92.53</td>
<td>92.53</td>
<td>92.53</td>
<td></td>
</tr>
<tr>
<td>Steel-Frame Construction</td>
<td>88.78</td>
<td>81.26</td>
<td>80.22</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>97.96</td>
<td></td>
</tr>
<tr>
<td>Titanium plate</td>
<td>92.13</td>
<td>92.03</td>
<td>91.28</td>
<td>93.33</td>
<td>93.33</td>
<td>93.33</td>
<td>93.33</td>
<td></td>
</tr>
<tr>
<td>OA(%)</td>
<td>91.54</td>
<td>92.58</td>
<td>93.32</td>
<td>94.95</td>
<td>95.79</td>
<td>96.02</td>
<td>94.20</td>
<td>98.85</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.9086</td>
<td>0.9202</td>
<td>0.9264</td>
<td>0.9393</td>
<td>0.9497</td>
<td>0.9522</td>
<td>0.9529</td>
<td>0.9529</td>
</tr>
</tbody>
</table>

### 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**

<table>
<thead>
<tr>
<th>No.</th>
<th>SVM</th>
<th>SAE</th>
<th>DBN</th>
<th>1D-CNN</th>
<th>2D-CNN</th>
<th>3D-CNN</th>
<th>MiniGCN</th>
<th>CNNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt concrete</td>
<td>89.93</td>
<td>97.04</td>
<td>97.63</td>
<td>95.73</td>
<td>90.76</td>
<td>89.56</td>
<td>90.86</td>
<td>98.90</td>
</tr>
<tr>
<td>Color steel tile</td>
<td>97.29</td>
<td>100.00</td>
<td>100.00</td>
<td>99.79</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Color read steel tile</td>
<td>95.17</td>
<td>96.27</td>
<td>96.07</td>
<td>94.92</td>
<td>99.21</td>
<td>95.45</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Rubber square</td>
<td>92.96</td>
<td>100</td>
<td>100</td>
<td>97.83</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Greenhouse</td>
<td>95.24</td>
<td>85.51</td>
<td>92.26</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Glazed tile</td>
<td>97.28</td>
<td>100.00</td>
<td>100.00</td>
<td>98.91</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Building 1</td>
<td>93.63</td>
<td>100.00</td>
<td>100.00</td>
<td>92.81</td>
<td>100.00</td>
<td>97.30</td>
<td>92.81</td>
<td>100.00</td>
</tr>
<tr>
<td>Building 2</td>
<td>92.66</td>
<td>94.12</td>
<td>94.18</td>
<td>98.63</td>
<td>76.1</td>
<td>85.38</td>
<td>82.59</td>
<td>99.89</td>
</tr>
<tr>
<td>Building 3</td>
<td>89.94</td>
<td>97.17</td>
<td>98.11</td>
<td>98.59</td>
<td>98.96</td>
<td>97.34</td>
<td>98.96</td>
<td>99.48</td>
</tr>
<tr>
<td>Concrete</td>
<td>98.95</td>
<td>92.8</td>
<td>92.76</td>
<td>96.55</td>
<td>97.51</td>
<td>96.29</td>
<td>94.31</td>
<td>99.12</td>
</tr>
<tr>
<td>Steel tile</td>
<td>97.01</td>
<td>100.00</td>
<td>99.31</td>
<td>90.87</td>
<td>100.00</td>
<td>100.00</td>
<td>99.04</td>
<td>99.62</td>
</tr>
<tr>
<td>Building 4</td>
<td>67.86</td>
<td>84.21</td>
<td>89.47</td>
<td>90.00</td>
<td>77.59</td>
<td>86.19</td>
<td>84.63</td>
<td>100.00</td>
</tr>
<tr>
<td>Clay Tile</td>
<td>83.46</td>
<td>90.54</td>
<td>82.59</td>
<td>87.78</td>
<td>85.02</td>
<td>87.67</td>
<td>85.00</td>
<td>98.96</td>
</tr>
<tr>
<td>Building 5</td>
<td>95.10</td>
<td>98.54</td>
<td>72.16</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Building 6</td>
<td>95.24</td>
<td>84.55</td>
<td>82.14</td>
<td>75</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>OA(%)</td>
<td>94.28</td>
<td>96.65</td>
<td>95.89</td>
<td>94.71</td>
<td>95.62</td>
<td>96.02</td>
<td>94.68</td>
<td>99.82</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.9402</td>
<td>0.9696</td>
<td>0.9512</td>
<td>0.9418</td>
<td>0.9491</td>
<td>0.9498</td>
<td>0.9351</td>
<td>0.9921</td>
</tr>
</tbody>
</table>
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 of 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.
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, 1D-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.

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.

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
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 sets show 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


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-IPRS, 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 I2 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.