Unsupervised Deep Feature Learning for Urban Village Detection from High-Resolution Remote Sensing Images

Article *in* Photogrammetric Engineering and Remote Sensing · August 2017

DOI: 10.14358/PERS.83.8.567

Yansheng Li

Wuhan University

Xin Huang

Wuhan University

Unsupervised Deep Feature Learning for Urban Village Detection from High-Resolution Remote Sensing Images

Yansheng Li, Xin Huang, and Hui Liu

Abstract

Urban villages (UVs) are a typical informal settlement in China resulting from the rapid urbanization in recent decades. Their formation and demolition are attracting increasing interest. In the remote sensing community, UVs have been detected based on hand-crafted features. However, the hand-crafted features just consider one or several characteristics of UVs, and ignore many effective cues hiding in the image. Recently, deep learning has been used to automatically learn suitable feature representations from a huge amount of data, without much expertise or effort in designing features. Motivated by its great success, this paper aims to use deep learning for detecting UVs. Because of the scarce labeled samples, this paper presents a novel unsupervised deep learning method to learn a data-driven feature. Experiments show the data-driven feature obtained with the proposed method outperform the existing unsupervised deep neural networks, and achieve results comparable to that obtained using the best hand-crafted features.

Introduction

As one of the by-products of urbanization, informal settlements (e.g., slums and shanty towns) are common in the developing cities of many countries (Kuffer *et al.*, 2016). While they have physical similarities to these other informal settlements, urban villages (UVs) are a unique product of the urbanization of China, and are common in the mega-cities of China. Different from other regions in urban areas, most spaces in UVs are occupied by small buildings, leaving little room for vegetation, streets, and bare ground. In addition, UVs are also known as "*chengzhongcun*" or "villages in the city" (Chuang, 2010). As a special type of urban settlement, UVs result from the complicated socio-economic development of China.

In the past few decades, large amounts of villages in the urban fringes have been progressively enveloped by expanding cities due to China's rapid urbanization. The original residential areas of these villages are left intact, but the farmland is used for urban development (Hao *et al.*, 2013). The original villagers legitimately own the residential areas, but they are not allowed to expand the land. During this period, the migration of large numbers of rural workers to cities has created a great demand for affordable housing, along with the rapid economic growth (Shen, 1995; Yang, 2000). Driven by the economic interest, many villagers have built additional dwellings in their residential areas and then rented them to migrant workers and the poor. The original villagers have become landlords and the enveloped villages have become the so-called UVs. However, the development of UVs is neither authorized nor planned. As a consequence, UVs suffer from poor sanitary conditions, absent infrastructure, and various social problems, including crime and environmental pollution. Accordingly, the UVs are preventing the mega-cities of China becoming recognized as international modern cities.

Because of the aforementioned problems induced by UVs, many cities of China have decided to dismantle and redevelop the UVs (Chuang, 2009; Chuang and Zhou, 2011). In order to guarantee that the UV redevelopment policy is fully implemented, an up-to-date UV map is necessary for planners and policymakers. However, such a UV map is often incomplete or unavailable. In reality, the identification of UVs largely relies on fieldwork and social investigation, and has rarely been addressed in the remote sensing community. However, as the mega-cities of China cover very large areas, timely and complete detection of UVs just using fieldwork is impossible.

In the literature, high-resolution remote sensing images have been successfully utilized in various applications, such as urban land-cover classification (Stavrakoudis *et al.*, 2011; Persello and Bruzzone, 2014), object matching and detection (Ma *et al.*, 2015; Cheng and Han, 2016), built-up areas detection (Li *et al.*, 2015a), central business district detection (Taubenbock *et al.*, 2013), private garden detection (Mathieu *et al.*, 2007), and so forth. Compared with the detection of these urban zones, the mapping of UVs using high-resolution remote sensing images suffers from additional difficulties, such as the large variance of the spectral reflectance and spatiotemporal patterns because of the unplanned development (Huang *et al.*, 2015). In addition, these difficulties also exist in the detection of informal settlements. In the existing studies, both object-based and segmentation-based methods (Hofmann *et al.*, 2008; Hofmann, 2001; Rhinane *et al.*, 2011; Kuffer *et al.*, 2016) have been proposed to identify informal settlements from urban areas using high-resolution remote sensing images. These approaches are certainly one option that might help in the mapping of UVs, but more intelligent approaches are also needed. Particularly, the advent of deep learning and scene-based analysis techniques opens up the possibility of establishing new UV detection approaches.

Yansheng Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (liyansheng414@163.com).

Xin Huang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; and the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China. xhuang@whu.edu.cn

Hui Liu is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China.

Photogrammetric Engineering & Remote Sensing
Vol. 83, No. 8, August 2017, pp. 567–579.
0099-1112/17/567–579

doi: 10.14358/PERS.83.8.567

Compared with the pixel-based and object-based classification methods (Khatami *et al.*, 2016; Blaschke *et al.*, 2014), scene-based classification methods have been acknowledged to be a more effective way to interpret high-resolution remote sensing images (Yang and Newsam, 2010; Chen and Tian, 2015), and they have benefited various applications, such as the recognition of urban functional zoning (Zhang and Du, 2015). For scene-based classification, the primary unit is the image block, and its semantic scene category is determined by not only the contained objects, but also their relative positions (Cheng *et al.*, 2014; Zhao *et al.*, 2016). Hence, scene-based classification methods are able to better distinguish complex categories. In order to overcome the aforementioned restrictions of the object-based methods for informal settlement detection, Huang *et al.* (2015) first proposed a scene-based classification method to detect UVs using high-resolution remote sensing images based on hand-crated features. However, the hand-crafted features just consider one or several characteristics of the image scenes, and they ignore many effective cues hiding in the data, such as edges, corners, junctions, object parts, and so forth.

In the past decade, deep learning has achieved remarkable performance improvements in various visual recognition tasks (Hinton *et al.*, 2006; Krizhevsky *et al.*, 2012). The intrinsic difference between deep learning methods and the traditional visual recognition methods is that the deep learning methods can autonomously learn feature representations from a large amount of data, without much expertise or effort in designing features. The deep learning methods can automatically learn hierarchical feature representations. In the hierarchical feature representations, a simple concept is first learned, and then complex concepts are successively built by composing simpler ones. This feature abstraction process also accords with the human visual cognition process. Motivated by the great success of deep learning, this paper attempts to utilize a data-driven feature which can be automatically learned from large amounts of data, to construct the scene-based classification model for the detection of UVs. Generally, the sufficient learning of a single deep convolutional neural network (CNN), which is one of the most influential deep learning models in the visual recognition domain, depends on millions of labeled training samples (Deng *et al.*, 2011). However, for the UV detection task, only a small amount of labeled training samples are available as the annotation of training samples needs fieldwork verification, which makes the annotation of a large number of training samples very expensive. Hence, a supervised learning based CNN is not applicable for the detection of UVs.

Although the supervised learning based CNN model (Krizhevsky *et al.*, 2012; Zhao and Du, 2016) has substantially improved the performance of various visual recognition tasks, unsupervised deep learning (Lee *et al.*, 2009; Coates *et al.*, 2011; Ngiam *et al.*, 2011; Romero *et al.*, 2015) still attracts undiminished attention, and is a potential development direction for deep learning (LeCun *et al.*, 2015) as large numbers of labeled training samples are often not available in many remote sensing applications. With the consideration that labeled UV data are a scarce resource, this paper exploits the unsupervised deep learning technique to produce a data-driven feature for constructing a scene-based UV detection method. More specifically, this paper proposes a novel unsupervised deep neural network (UDNN) which is composed of an unsupervised deep convolutional neural network (UDCNN) and an unsupervised deep fully connected neural network (UDFNN). In the context of unsupervised deep learning, UDCNNs have been widely discussed (Lee *et al.*, 2009; Dong *et al.*, 2014; Romero *et al.*, 2016; Li *et al.*, 2016a), but the merits of UDFNNs have been ignored. To better suit the UV detection, unsupervised multilayer feature learning (Li et al., 2016a) is advocated to construct the UDCNN, and the UDFNN is further implemented by stacking restricted Boltzmann machines (RBMs) to abstract the convolutional feature representation from a global perspective. As a consequence, the data-driven feature obtained with the proposed UDNN is more suitable for the UV detection task than the existing unsupervised deep neural networks (Dong *et al.*, 2014; Romero *et al.*, 2016). Based on the data-driven feature from the proposed UDNN, this paper presents a novel scene-based classification method for UV detection. In the experimental section, we utilize four temporal QuickBird remote sensing images of Shenzhen (i.e., a mega-city of China) and their UV manual annotation results to evaluate the proposed scene-based approach using unsupervised deep feature learning.

In following sections of this paper, we describe the study area and the data first, and then the proposed scene-based UV detection approach based on unsupervised deep feature learning. In the experiments, the proposed method is compared with the existing UDNNs and the state-of-the-art approaches using hand-crated features. Finally, we give the conclusions and future prospects.

Study Area and Data

As a relatively young city inhabited by many immigrants, Shenzhen, one of the mega-cities of China, has experienced severe problems associated with UVs in recent decades. Hence, the urban areas of Shenzhen were chosen for this study. Shenzhen, located in Guangdong province, was a small village on the Pearl River Delta before it became China's first special economic zone (SEZ) in 1979. Since then, Shenzhen has experienced rapid economic development and has become one of the biggest cities in China. During this period, Shenzhen's population has increased from less than 0.1 million in 1979 to over 10 million in 2010, and UVs have rapidly spread across the city.

It has recently been estimated that about half of Shenzhen's population live in UVs (Zacharias and Tang, 2010). However, the area of UVs is dramatically smaller than non-UVs. Under this unbalance distribution, the density of population and accommodation in UVs is extremely high as aforementioned. Remote sensing data from the QuickBird satellite acquired during 2003 to2010 were used in this study (Table 1), which had already been radiometrically calibrated. QuickBird data have a spatial resolution of 2.4 m, with an image size of 5,360 × 4,507 pixels. All the images were obtained between 114°3'E to 114°9'E and 22°30'N to 22°37'N, and cover about 91.84 km^2 of the SEZ of Shenzhen (see Figure 1).

As a special type of urban settlement in China, UVs mostly originate from villages. They are surrounded by planned urban areas and are usually adjacent to highways (see Figure 1b), undeveloped areas (see Figure 1c), skyscrapers (see Figure 1d), and other modern urban infrastructures. Because of the lack of urban management, UVs are densely settled and develop in a disorderly manner. Compared with other urban areas, buildings in UVs are much smaller and lower. These buildings are densely distributed in the UVs and occupy most of the available space. Accordingly, vegetation and public spaces, which are fundamental components of planned urban areas, are rarely found in UVs.

Table 1. Description of the Shenzhen dataset.

Acquisition date	Satellite	Spectral range	Spatial resolution
2003/01/17 2005/12/17 2007/12/10 2010/05/26	QuickBird	4 bands 450-900 nm	2.4 m

Methodology

In the context of high-resolution imagery, the main observable characteristics that distinguish UVs from formal residential areas are as follows. (a) Most spaces in UVs are occupied by small buildings, leaving little room for vegetation, streets, and bare ground; and (b) In contrast to formal residential areas, UVs tend to have a disordered layout. The proportions of objects of several major classes (i.e., buildings and vegetation), as well as the spatial configuration, are the key to detecting UVs.

In this paper, we carry out the detection at the scene level. A scene is an image block that usually contains various objects and belongs to some semantic categories. Generally, the scene-based method can simultaneously consider the containing objects and the relationships between objects; hence, it is competent for the semantic identification of complex man-made structures. In the literature, a number of different scene-level high-resolution remote sensing image classification approaches (Yang and Newsam, 2010; Chen and Tian, 2015; Li *et al.*, 2016a; Romero *et al.*, 2016; Hu *et al.*, 2015; Marmanis *et al.*, 2016) have been proposed and evaluated with the popular scene classification datasets (Yang and Newsam, 2010). In the existing approaches, the adopted features can be divided into three categories: hand-crafted features obtained using bag-of-visual-words (BoVW) representation (Yang and Newsam, 2010; Chen and Tian, 2015); data-driven features obtained using unsupervised deep feature learning (Li *et al.*, 2016a; Romero *et al.*, 2016); and data-driven features obtained using supervised deep feature learning (Hu *et al.*, 2015; Marmanis *et al.*, 2016). In this paper, rather than utilizing hand-crafted features (Huang *et al.*, 2015), we consider a data-driven feature for the UV detection task. With the consideration that the labeled UV samples are a scarce resource which makes the supervised deep feature learning approaches (Hu *et al.*, 2015; Marmanis *et al.*, 2016) unavailable, this paper presents a novel UDNN, which can be trained using unsupervised learning, to generate the data-driven features. Given a single input image scene, the corresponding feature representation can be automatically achieved using the UDNN. We then utilize support vector machine (SVM) (Mountrakis *et al.*, 2011) to discriminate the feature representations and undertake the scene classification (i.e., UVs or non-UVs), as SVM has been proven to show a good feature discrimination performance, even when the number of labeled samples is relatively small.

In the following sections, we specifically introduce the proposed UDNN first. Once the UDNN is trained, the feature extraction function is available. Then, we describe the complete UV detection method, where the training of the discriminative classifier is introduced, and the UV detection process is illustrated.

Unsupervised Deep Feature Learning

In this section, we present the unsupervised deep feature learning approach for training the UDNN (Figure 2). Here, we first introduce the unlabeled dataset. The unlabeled dataset is composed of the aforementioned image scenes or image blocks which can be randomly sampled from the large image. We let $\mathbf{U} = \{I^1, I^2, \ldots, I^M\}$ denote the unlabeled dataset, where I denotes the original image scene, and M is the volume of the unlabeled dataset. Using this unlabeled dataset, the UDNN, which is composed of the UDCNN and the UDFNN, can be trained in succession. In the following, Part 1 presents the training procedure for the UDCNN. Part 2 then introduces the training process for the UDFNN.

Unsupervised Deep Convolutional Neural Network (UDCNN)

In our implementation, unsupervised multilayer feature learning (Li *et al.*, 2016a; Li *et al.*, 2016b) is used to construct the UDCNN due to its superior performance in describing remote sensing image content. Compared with other UDCNNs (Dong *et al.*, 2014; Romero *et al.*, 2016), the proposed UDCNN is implemented using the plain *k*-means clustering method and can outperform other UDCNNs, especially when a small receptive field is applied. It should be noted that a smaller receptive field in the UDCNN is more favorable for identifying UVs as a small receptive field helps to distinguish the dense small buildings in the UVs and to further distinguish UVs from non-UVs. In the following, the proposed UDCNN is specifically introduced.

As illustrated in Figure 2, the UDCNN is composed of multiple feature extraction layers. Figure 2 gives an example where the UDCNN contains three feature layers, and the layer number of the UDCNN can be adjusted based on the need of the specific application. As depicted in Figure 2, each feature extraction layer includes three operations: the convolution operation, the local pooling operation, and the global pooling operation. Among these three operations, the local pooling operation and the global pooling operation aim at feature reduction in the local and global domains, which can be designed in advance, but the convolution operation works for

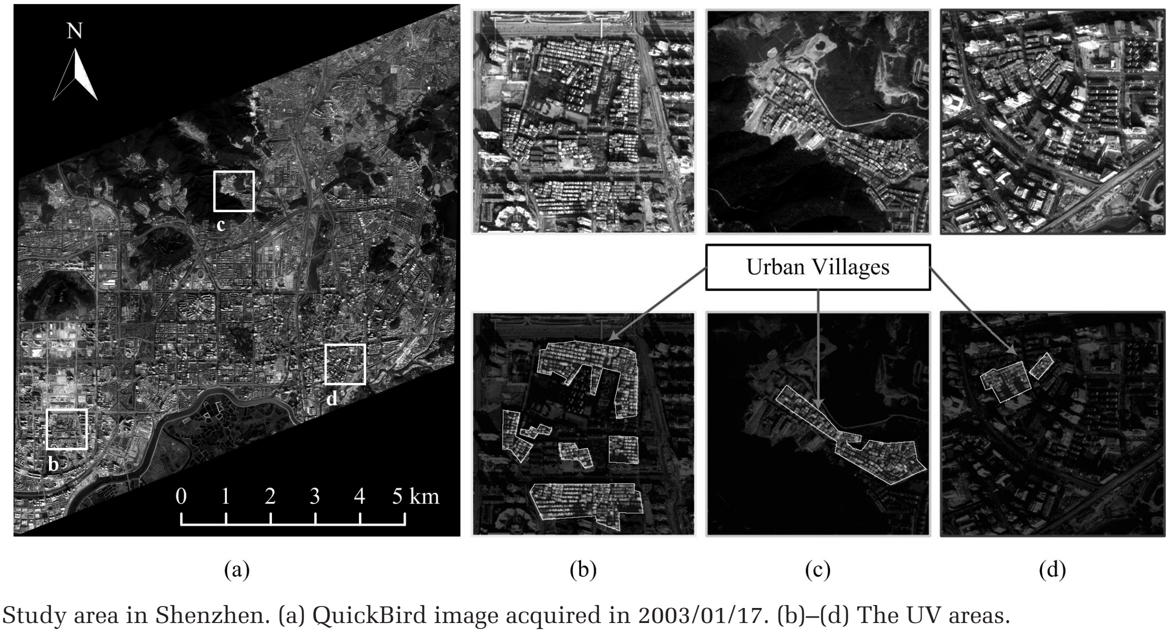


Figure 1. Study area in Shenzhen. (a) QuickBird image acquired in 2003/01/17. (b)–(d) The UV areas.

feature mapping, which need to be determined by unsupervised feature learning from the data.

The convolution operation is intuitively shown in Figure 2. More specifically, the feature mapping is defined under the constraint of function bases which can be generated by unsupervised feature learning. The *k*-means clustering approach is advocated to implement this function as the learned function bases can be highly analogous to the neuron responses of the human visual system (Li *et al*., 2016a). Compared with some of the recent unsupervised feature learning approaches (Ngiam *et al*., 2011; Romero *et al*., 2015), the performance superiority of the proposed UDCNN is verified in the experiments. Through *k*-means clustering, the image patches are aggregated into clusters whose centers are taken as the function bases. For clarification, we take the first feature layer as an example to demonstrate the generation process of the function bases. As the original image scenes are the input of the first feature layer, we randomly sample patches from the original image scenes $\mathbf{U} = \{I^1, I^2, \ldots, I^M\}$ to learn the function bases. Given the sampled patches with the dimension $w - d - d$, where w denotes the size of the receptive field, and d is the number of feature channels we can construct the feature set $\mathbf{X} = \{x^1, x^2, \ldots, x^N\}$, where $x_i \in R^D$ denotes the vectorization vector of the i^{th} patch, $D = w{\cdot}w{\cdot}d$ is the dimension of the feature vector, and N is the volume of the sampled patches. The initial feature set $\mathbf{X}$ is first preprocessed by intensity normalization and zero component analysis whitening. After preprocessing, the feature set $\mathbf{X}$ is clustered into K clusters by *k*-means clustering. The corresponding K centers $\mathbf{c} = \{c^1, c^2, \ldots, c^K\}$ are referred to as function bases, where $c^i \in R^D$.

Once the function bases of the first feature layer are determined, the convolution operation (i.e., feature mapping) of the first feature layer can be defined as follows. We let $\mathbf{x}$ denote the vectorization vector of one sliding patch of the input image. Based on the function bases $\mathbf{c} = \{c^1, c^2, \ldots, c^K\}$, the feature vector x of one given sliding patch can be mapped into the sparse feature vector $r \in R^K$ using Equation 1 which works for pursuing feature sparsity. It is noted that feature sparsity is considered a principle adopted by the early visual cortex as an efficient means of coding (Willmore and Tolhurst, 2011). The sparse feature mapping function in Equation 1 also contains the non-linear mapping operator which plays the role of the activation function in traditional CNNs.

$$r_k = \max\{0, u(z) - z_k\} \tag{1}$$

where $z_k = \| x - c^k \|_2$, and $u(z)$ denotes the mean value of vector z.

Let I denote the input of the first feature layer. In addition, the dimension of I is $h - h - d$, where h stands for the height and width of I, and d denotes the depth of I. Through the convolution operation, the input image of the first layer I is mapped into the convolution result of the first layer R with the dimension $(h - w) - (h - w) - K$.

The local pooling operation is intuitively illustrated in Figure 1, and works to keep the invariance to slight translation and rotation. Here, the local pooling operation is implemented by a local maximum operation. We let s denote the local window size of the local pooling operation, and the local pooling operation result P can be expressed by:

$$P(i/s, j/s, k) = \max(R(i - s/2: i + s/2, j - s/2: j + s/2, k \tag{2}$$

where $k = 1, 2, \ldots, K$, and the dimension of P is $((h - w)/s) - ((h - w)/s) - K$.

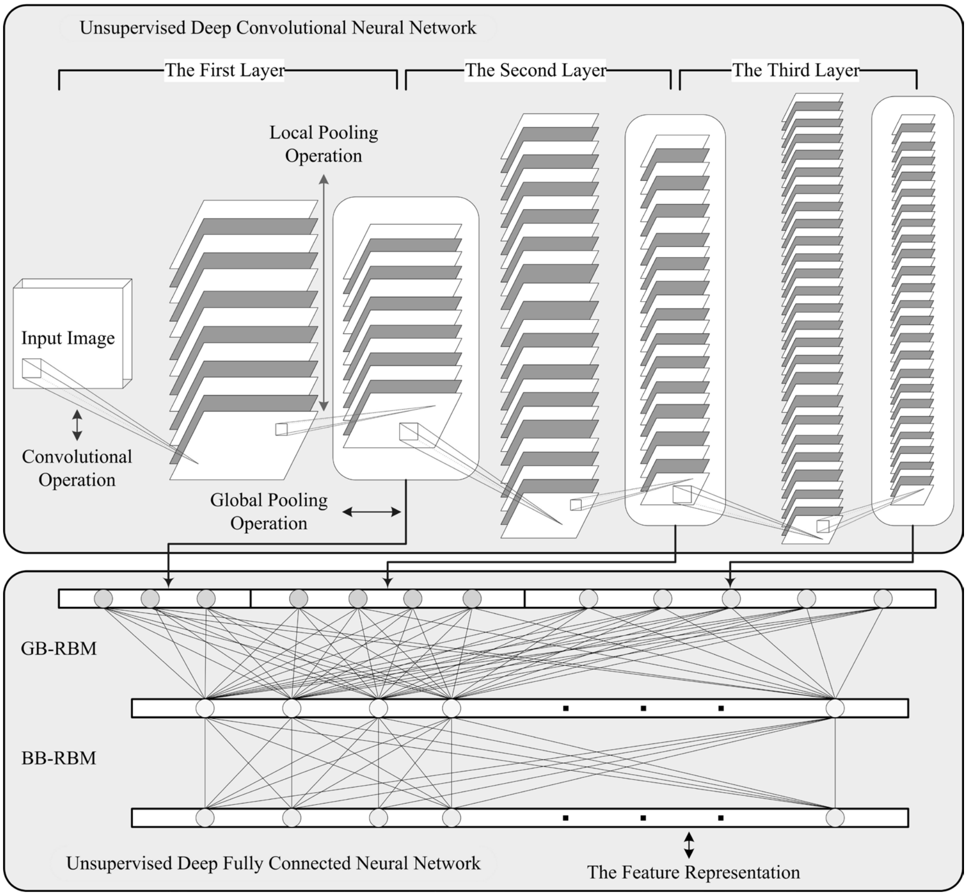


Figure 2. The overall architecture of the proposed unsupervised deep neural network.

As depicted in Figure 1, the global pooling operation is intuitively indicated. The global pooling operation works for feature reduction. In this implementation, the global pooling operation is implemented among the four quarters of the output *P* of the first layer. On the first quarter, the second quarter, the third quarter, and the fourth quarter, the corresponding feature vectors can be expressed as g^1, g^2, g^3, and g^4. More specifically, the corresponding feature vectors can be formulated as:

$$g_k^1\text{=mean}(\mathrm{P}(1{:}(h\text{–}w)/(2\cdot s),1{:}(h\text{–}w)/(2\cdot s),k \tag{3}$$

$$g_k^2\text{=mean}(\mathrm{P}(1{:}(h\text{–}w)/(2\cdot s),(h\text{–}w)/(2\cdot s)+1{:}(h\text{–}w)/s,k)) \tag{4}$$

$$g_k^3\text{=mean}(\mathrm{P}((h\text{–}w)/(2\cdot \mathrm{s})+1{:}(h\text{–}w)/\mathrm{s},1{:}(h\text{–}w)/(2\cdot \mathrm{s}),\mathrm{k}) \tag{5}$$

$$g_k^4\text{=mean}(\mathrm{P}((h\text{–}w)/(2\cdot s)+1{:}(h\text{–}w)/s,(h\text{–}w)/(2\cdot s)+1{:}(h\text{–}w)/s,k)) \tag{6}$$

where $k = 1, 2, \ldots, K$.

Furthermore, the integrated result $g = [g^1, g^2, g^3, g^4]$ from the four quarters is taken as the global pooling result of the first layer, where the dimension of g is $4\cdot K$.

Through implementing the convolution operation and the local pooling operation of the first layer on the original image scenes $\mathbf{U} = \{I^1, I^2, \ldots, I^M\}$, we can obtain the output $\mathbf{P} = \{P^1, P^2, \ldots, P^M\}$, which is the input of the second layer. Similar to the generation process of the function bases of the first layer, we can sample patches with the dimension $w - w - K$ from $\mathbf{P} = \{P^1, P^2, \ldots, P^M\}$ to generate the function bases of the second layer. Accordingly, the convolution operation of the second layer can be determined. In addition, the local pooling operation and the global pooling operation of the second layer are the same as the operations of the first layer. Furthermore, the operations of the backward layers can be similarly constructed.

Assuming that the UDCNN contains *L* feature layers, where the number of function bases of the *L* feature layers is $K_1, K_2, \ldots, K_L$, given one image scene I, the global pooling result from the *L* feature layers can be expressed by $f_1, f_2, \ldots, f_L$. Through integrating the results from the *L* feature layers, the feature representation of I using the UDCNN can be expressed as $f = [f_1, f_2, \ldots, f_L]$ with the dimension $4\cdot(K_1 + K_2 + \ldots + K_L)$.

*2. Unsupervised Deep Fully Connected Neural Network (*UDFNN*)*

To the best of our knowledge, the merits of fully connected layers have been ignored in the unsupervised deep learning methods (Dong *et al.*, 2014; Li *et al.*, 2016a; Romero *et al.*, 2016). On the basis of the UDCNN, the presented UDFNN is implemented by stacking the RBMs, and works to further abstract the feature representation from a global perspective.

The UDCNN mainly focuses on mining the local structures of the input image, and the UDFNN is designed to further mine the global semantic information. As reported in Hinton *et al.* (2006), deep belief nets can be trained in an unsupervised manner and can perform well in feature abstraction. Hence, in this paper, we utilize RBMs with Gaussian units as the input and Bernoulli units as the output (i.e., GB-RBM), and RBMs with Bernoulli units as the input and Bernoulli units as the output (i.e., BB-RBM), to implement the UDFNN.

Once the aforementioned UDCNN is constructed, it possesses the ability to automatically extract hierarchical features from the original image scenes. Given the original image scenes $\mathbf{U} = \{I^1, I^2, \ldots, I^M\}$, the UDCNN can output the corresponding feature representations $\mathbf{F} = [f_1, f_2, \ldots, f^M]$, which can be utilized to train the UDFNN.

We let f and h denote the visible and hidden layers of the RBMs. As the values of f are continuous, we utilize GB-RBM to model the abstraction between the visible layer f with Gaussian units and the hidden layer h with Bernoulli units. More specifically, the energy function can be expressed as:

$$E(\mathrm{f},\mathrm{h}) = \sum\nolimits_{i=1}^{n_\mathrm{f}} \frac{\left(\mathrm{f}_i - a_i^1\right)^2}{2\sigma_i^2} - \sum\nolimits_{j=1}^{n_\mathrm{h}} h_j b_j^1 - \sum\nolimits_{i=1}^{n_\mathrm{f}} \sum\nolimits_{j=1}^{n_\mathrm{h}} \frac{\mathrm{f}_i}{\sigma_i} h_j \mathrm{W}_{i,j}^1 \tag{7}$$

where n_f and n_h are the dimensions of the visible layer and the hidden layer. σ_i denotes the standard variation of the i^{th} feature component $[f_i^1, f_i^2, \ldots, f_i^M]$. $\mathrm{W}_{i,j}^1$ stands for the linking coefficients of the visible layer and the hidden layer; b_j^1 denotes the bias of the visible layer. stands for the bias of the hidden layer.

Using the training data $\mathbf{F} = [f_1, f_2, \ldots, f^M]$ and the coefficients $\mathrm{W}_{i,j}^1$, the biases a_i^1 and b_j^1 can be determined using the gradient descent algorithm, where the gradient can be approximately calculated by the famous contrastive divergence algorithm (Hinton, 2002). Given one image scene I^m, the corresponding feature representation using the UDCNN can be expressed as f^m. Based on the learned GB-RBM, f^m can be further mapped into h^m:

$$p(h_j^m = 1 | \mathrm{f}^m) = \varsigma(\sum\nolimits_{i=1}^{n_\mathrm{f}} \tfrac{\mathrm{f}_i}{\sigma_i} W_{i,j}^1 + b_j^1) \tag{8}$$

where $j = 1, 2, \ldots, n_\mathrm{h}$, and $\varsigma(x)=(1+e^{-x})^{-1}$ is the sigmoid function.

Using the learned weights $\mathrm{W}_{i,j}^1$, a_i^1, b_j^1, and of the first layer of the UDFNN, the feature set can be mapped into $\mathbf{F} = [f_1, f_2, \ldots, f^M]$, which is utilized to train the second layer of the UDFNN.

As the values of h are binary, the backward layers of the UDFNN can be modeled by BB-RBM. In the second layer of the UDFNN, h is taken as the visible layer. We let o with Bernoulli units denote the hidden layer. Thus, the corresponding energy function can be expressed as:

$$E(\mathrm{h},\mathrm{o}) = -\sum\nolimits_{i=1}^{n_\mathrm{h}} a_i^2 h_i - \sum\nolimits_{j=1}^{n_o} o_j b_j^2 - \sum\nolimits_{i=1}^{n_\mathrm{h}} \sum\nolimits_{j=1}^{n_o} h_i\, o_j\, W_{i,j}^2 \tag{9}$$

where n_h and n_o are the dimensions of the visible layer and the hidden layer. $\mathrm{W}_{i,j}^2$ stands for the linking coefficients of the visible layer and the hidden layer. a_i^2 denotes the bias of the visible layer. b_j^2 stands for the bias of the hidden layer.

Similar to the optimization process of GB-RBM, BB-RBM can also be optimized by the gradient descent algorithm. Once the coefficients $\mathrm{W}_{i,j}^2$, a_i^2, and b_j^2 are determined, the aforementioned feature vector h^m can be mapped into o^m:

$$p\left(\mathrm{o}_j^m = 1| \,\mathrm{h}^m\right) = \varsigma(\sum\nolimits_{i=1}^{n_\mathrm{f}} \mathrm{h}_i^m \mathrm{W}_{i,j}^2 + b_j^2) \tag{10}$$

where $j = 1, 2, \ldots, n_\mathrm{o}$, and $\varsigma(x) = (1 + e^{-x}$ is the sigmoid function.

In order to concisely depict the fully connected network, we just introduce two layers in the UDFNN. Actually, the UDFNN can be extended to include more layers through stacking more BB-RBM models at the end of this network.

As a whole, the proposed unsupervised feature extraction network is composed of UDCNN working for local structure extraction and UDFNN working for global semantic feature abstraction. On the training stage, for UDCNN, unsupervised learning works for generating the function bases $\mathbf{C} = \{c^1, c^2, \ldots, c^K\}$ of each feature extraction layer in UDCNN by a layer-wise way. As far as UDFNN, unsupervised learning works for updating the weights $W_{i,j}$, a_i, and b_j of each feature layer in UDFNN. To sum up, the proposed feature extraction network can be trained in a fully unsupervised way. On the feature extraction stage, the feature representation of one image scene can be extracted using UDCNN and UDFNN. Given one input image scene, its local convolutional feature $f = \{f_1, f_2, \ldots, f_L\}$ can be firstly mined using the adopted UDCNN. Using Equation 8, f can be mapped to the hidden feature vector h. Finally, h is mapped to the final feature representation o using Equation 10. In the following parts, the final feature representation o is taken as the input of the feature classifier to identify UV scenes from remote sensing images.

The Proposed Scene-Based Classification Approach for UV Detection

As illustrated in Figure 3, the proposed scene-based UV detection approach includes two stages (i.e., the offline stage and the online stage). The offline stage works for training the feature extraction network and the discriminative classifier, which are further utilized to detect the UVs in the online stage.

As mentioned before, the feature representation of one image scene can be acquired using the trained feature extraction neural network. In addition, the feature discrimination issue is specially discussed in the following. Generally, the classification performances of neural network classifiers highly depend on large amounts of labeled samples, but the support vector machine (SVM) can achieve a relatively ideal classification result even when the number of labeled samples is small. As aforementioned, the number of labeled UV samples and non-UV samples is very limited. With this consideration, this feature discrimination module is implemented by the linear SVM classifier which is widely adopted because of its computational efficiency (Romero *et al.*, 2016). In the feature discrimination module, the feature representation of a small number of labeled UV and non-UV samples are firstly extracted using the aforementioned neural network, and further utilized to train the feature discrimination classifier through SVM. In the SVM training stage, only one regularization parameter c is required to be tuned. The sensitivity of the regularization parameter c is reported in the experimental part.

In the online stage, the feature extraction neural network and the feature classifier are utilized to recognize the sliding window. In this stage, one sliding window is classified into either UVs or non-UVs. As the sliding step is smaller than the sliding window, several adjacent sliding windows may cover one region. The overlapping region is the primary unit used to detect the UVs in the high-resolution remote sensing images. The likelihood of one overlapping region belonging to a UV is voted for by the classification results of the covering sliding windows.

Results and Discussion

In this section, we first show the quantitative and visual results of the proposed approach on the study area. Then, the sensitivity of the crucial parameters in the proposed approach is analyzed. Next, we give a comparison with the existing approaches, including the existing UDNNs and the hand-crafted

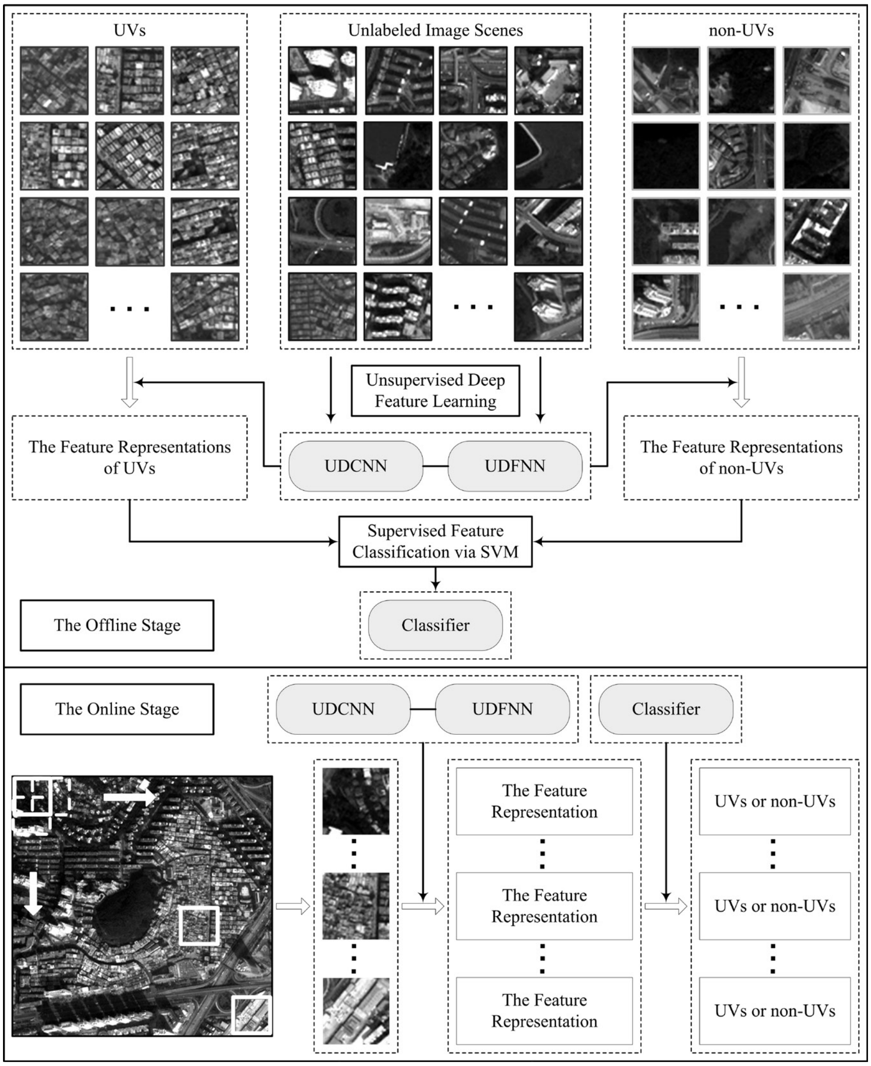

UVs
Unlabeled Image Scenes
non-UVs
. . .
Unsupervised Deep Feature Learning
The Feature Representations of UVs
UDCNN
UDFNN
The Feature Representations of non-UVs
Supervised Feature Classification via SVM
The Offline Stage
Classifier
The Online Stage
UDCNN
UDFNN
Classifier
The Feature Representation
The Feature Representation
The Feature Representation
UVs or non-UVs
UVs or non-UVs
UVs or non-UVs

Figure 3. The flowchart of the proposed UV detection approach.

features which are specifically designed for UV detection. Finally, we provide the discussions about the characteristics and potential applications of the presented approach.

The Overall Performance of the Proposed Approach

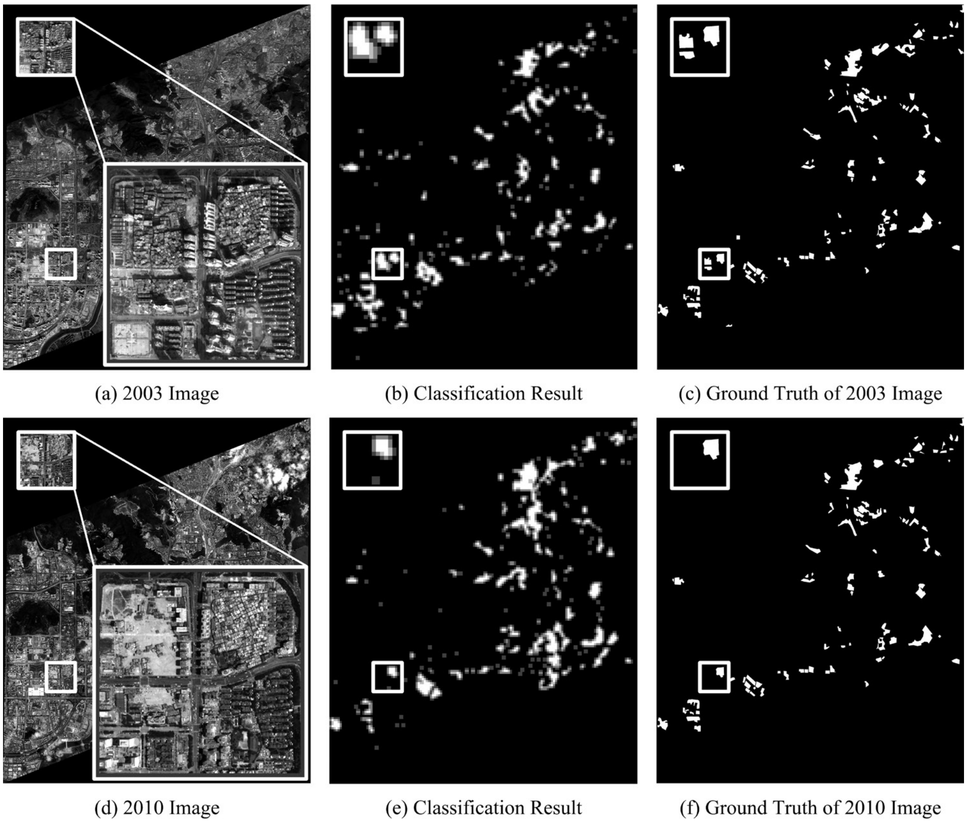
As suggested in (Huang *et al*., 2015), the scene size was set to 60 × 60 pixels, which corresponds to regions of 144 × 144 m on the ground in the adopted research case (i.e., Shenzhen). It is noted that the optimal scene size may vary along with the change of the ground spatial resolution and the research areas. Hence, the scene size should be empirically adjusted based on specific tasks. In addition, we utilized M = 100,000 image scenes without labels and N = 1,000,000 unlabeled image patches to train the proposed UDNN in a layer-wise style. The supervised training dataset was composed of 15 UV image scenes and 15 non-UV image scenes, which were utilized to train the discriminative classifier using linear SVM. In order to evaluate the performance of the UDNN and the discriminative classifier, we constructed a test dataset which was composed of 200 UV image scenes and 600 non-UV image scenes. More specifically, one scene is assigned to the UV image scene if more than 90 percent of it are occupied by UVs, and one scene is assigned to the non-UV scene if there does not exist any UVs in it. This setting was designed to simulate the actual test situation, as the area of the UVs is dramatically smaller than the area of non-UVs in reality. Furthermore, the overall classification accuracy (i.e., the OA) and the Kappa value were taken as the evaluation criteria. In addition to the quantitative evaluation results, in the following parts, we also report the visual UV detection result of the whole remote sensing image for facilitating readers' intuitive and global inspection.

The large amounts of experiments showed that the proposed UDNN can achieve the best performance when the UDCNN contains three layers and the UDFNN contains one layer. More specifically, the optimal numbers of function bases for the three layers are 100, 900, and 2500 in the UDCNN, and the optimal number for the hidden layer is 2500 in the UDFNN. We conducted four experiments on a 2003 QuickBird image, a 2005 QuickBird image, a 2007 QuickBird image, and a 2010 QuickBird image, respectively. In these experiments, the unlabeled image scenes came from the large test image to learn the feature extraction neural network in an unsupervised manner. In addition, the training dataset and the test dataset were constructed through randomly selecting image scenes from the same large test image. The statistical results of the five experiments on the four large temporal images are reported in Table 2. These positive results fully confirm the validity of the proposed data-driven feature in the UV detection task.

Table 2. The classification performance of the proposed approach on the four temporal images.

	2003 QuickBird	2005 QuickBird	2007 QuickBird	2010 QuickBird
OA	0.9855±0.0087	0.9830±0.0066	0.9820±0.0063	0.9680±0.0282
Kappa	0.9623±0.0221	0.9558±0.0170	0.9527±0.0158	0.9198±0.0674

In order to intuitively show the performance of the proposed approach, the visual classification results of the 2003 QuickBird image and the 2010 QuickBird image are shown in Figure 4. The visual results show that the scene-based classification approach is able to identify the UVs in the high-resolution remote sensing images. From the amplified results of the same locations on the 2003 QuickBird image and the 2010 QuickBird image, we can see that the proposed approach can not only discriminate the UVs under a very low omission rate, but it can also reflect the temporal variation of the UVs.



(a) 2003 Image (b) Classification Result (c) Ground Truth of 2003 Image

(d) 2010 Image (e) Classification Result (f) Ground Truth of 2010 Image

Figure 4. The visual classification results of the proposed approach.

The Parameter Analysis

In order to help the readers design neural networks based on their specific applications, this section systematically analyzes the sensitivity of the crucial parameters in the proposed UDCNN. All of the evaluation experiments were conducted on the 2003 QuickBird image.

We first analyze the influence of the size of the receptive field on the classification performance. Table 3 lists the different configurations of the proposed UDCNN under different sizes of receptive field when the proposed UDCNN only contains one feature extraction layer. The classification performance of the neural networks in Table 3 are shown in Figure 5a. As illustrated in Figure 5a, the smaller the size of the receptive field, the better the performance that the proposed UDCNN can achieve. As previously mentioned, the UV scenes are composed of dense small buildings. Accordingly, a small receptive field helps the feature extraction network to distinguish the small-scale structures from the tiny buildings, and to further discriminate UV scenes from non-UV scenes, which accords with the result shown in Figure 5a.

Fixing the size of receptive field, we further explore the influence of the size of the image scene on the classification performance. Let *s* denote the size of image scene, Table 4 lists different configurations of the proposed UDCNN under different sizes of image scene when the proposed UDCNN only contains one feature extraction layer. The classification performance of the neural networks in Table 4 are shown in Figure 5b. As shown in Figure 5b, the proposed UDCNN can achieve the best classification performance when the size of image scene equals 60.

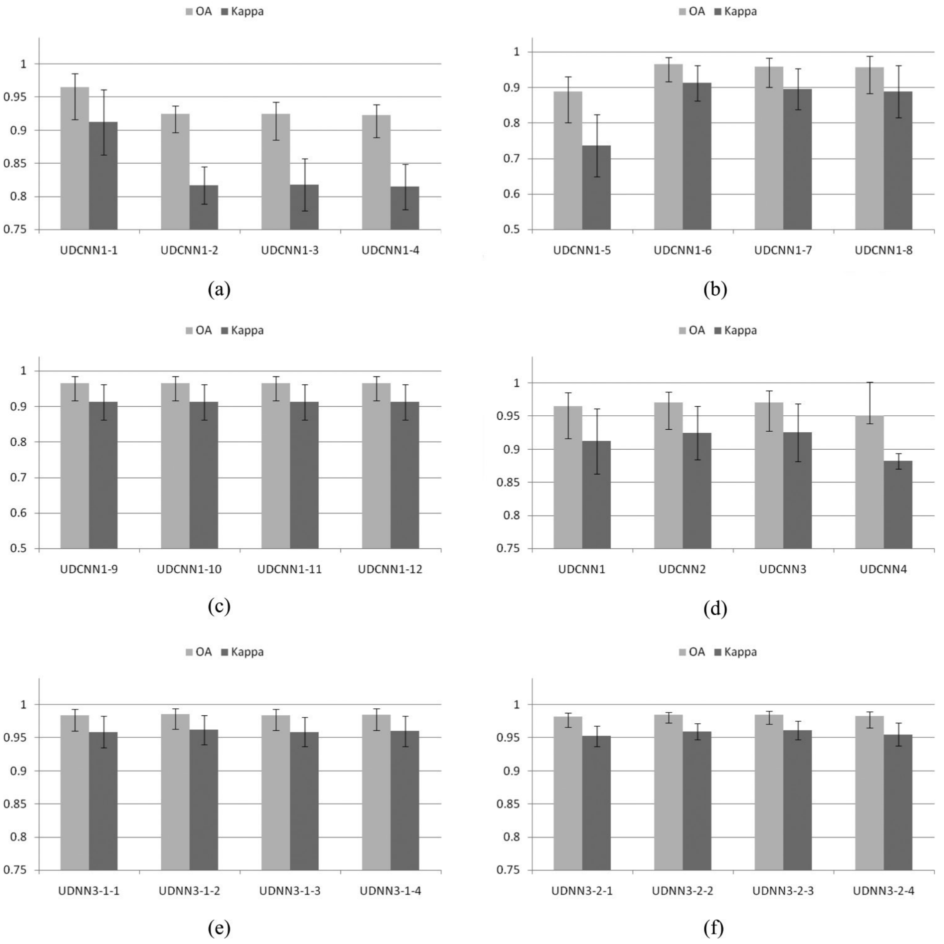
In order to analyze the sensitivity of the regularization parameter in the SVM training module, we conduct four experiments as defined in Table 5 where *c* stands for the regularization parameter. The corresponding classification results are summarized in Figure 5c. From the intuitive results in Figure 5c, it is not hard to draw the conclusion that the presented classification method is not very sensitive to the selection of

Table 3. The different configurations of the proposed UDCNN under different sizes of receptive field.

	Configuration
UDCNN1-1	UDCNN with 1 layer, $K_1 = 1024$, $w = 2$
UDCNN1-2	UDCNN with 1 layer, $K_1 = 1024$, $w = 4$
UDCNN1-3	UDCNN with 1 layer, $K_1 = 1024$, $w = 6$
UDCNN1-4	UDCNN with 1 layer, $K_1 = 1024$, $w = 8$

Table 4. The different configurations of the proposed UDCNN under different sizes of image scene.

	Configuration
UDCNN1-5	UDCNN with 1 layer, $K_1 = 1024$, $s = 30$
UDCNN1-6	UDCNN with 1 layer, $K_1 = 1024$, $s = 60$
UDCNN1-7	UDCNN with 1 layer, $K_1 = 1024$, $s = 90$
UDCNN1-8	UDCNN with 1 layer, $K_1 = 1024$, $s = 120$



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. The quantitative results of the proposed UDNN under different configurations: (a), (b), (c), (d), (e), and (f) denote the evaluation results of the 2003 QuickBird image.

the regularization parameter c. Hence, in all experiments, the regularization parameter c of the SVM classifier is set to 100.

In Table 6, the configurations of the neural networks are summarized when the proposed UDCNN contains different numbers of feature extraction layers. In addition, the classification performances of these neural networks are reported in Figure 5d. As illustrated in Figure 5d, the proposed UDCNN with three feature extraction layers can achieve the best performance, which shows that more layers are not always helpful for unsupervised deep convolutional neural networks. In addition, we constructed the UDCNN with $K_1 = 100$, $K_2 = 900$, and $K_3 = 2500$, which is denoted as UDCNN3-E. UDCNN3-E is an enhanced version of UDCNN3 that can achieve a better performance than UDCNN3.

Finally, we verified the effect of the layer number and the node number of the UDFNN. Fixing the UDCNN, Table 7 lists the neural networks with different numbers of layers and different numbers of nodes for each layer. The corresponding classification performance is summarized in Figure 5e and 5f and Table 8.

As depicted in Figure 5e, UDNN3-1-2 can achieve the best performance. In addition, UDNN3-2-3 can achieve the best performance in Figure 5f. As depicted in Table 8, UDNN3-1-2 can achieve a better performance than UDNN3-2-3. Hence, we can conclude that the proposed UDNN can achieve the best performance when it contains one fully connected layer. Compared with the UDCNN (i.e., UDCNN3-E), the proposed UDFNN achieves a nearly 3 percent improvement in Kappa.

Hence, the proposed UDCNN was constructed based on the parameter setting of UDCNN3-E, and the proposed UDNN was configured by the parameter setting of UDNN3-1-2. To determine the universality of the proposed approach, the achieved parameter setting was applied in all four temporal images. In order to show the superiority of the advocated UDNN compared with the intermediate UDCNN, the extracted feature using UDNN with 10,000 dimensions and the extracted feature using UDNN with 3,000 dimensions are utilized to train linear SVM classifiers for evaluation in the experiments, respectively.

Table 5. The different configurations of the proposed UDCNN under different regularization parameters in the SVM training module.

	Configuration
UDCNN1-9	UDCNN with 1 layer, $K_1 = 1024$, $C = 25$
UDCNN1-10	UDCNN with 1 layer, $K_1 = 1024$, $C = 50$
UDCNN1-11	UDCNN with 1 layer, $K_1 = 1024$, $C = 100$
UDCNN1-12	UDCNN with 1 layer, $K_1 = 1024$, $C = 200$

Table 6. The different configurations of the proposed UDCNN under different numbers of feature extraction layers.

	Configuration
UDCNN1	UDCNN with 1 layer, $K_1 = 1024$
UDCNN2	UDCNN with 2 layers, $K_1 = 100$, $K_2 = 1024$
UDCNN3	UDCNN with 3 layers, $K_1 = 64$, $K_2 = 100$, $K_3 = 1024$
UDCNN4	UDCNN with 4 layers, $K_1 = 36$, $K_2 = 64$, $K_3 = 100$, $K_4 = 1024$

Table 7. The different configurations of the proposed UDNN under different layer numbers and node numbers.

	Configuration
UDNN3-1-1	UDCNN3-E, UDFNN with 1 layer, $n_h = 2000$
UDNN3-1-2	UDCNN3-E, UDFNN with 1 layer, $n_h = 2500$
UDNN3-1-3	UDCNN3-E, UDFNN with 1 layer, $n_h = 3000$
UDNN3-1-4	UDCNN3-E, UDFNN with 1 layer, $n_h = 3500$
UDNN3-2-1	UDCNN3-E, UDFNN with 2 layers, $n_h = 2500$, $n_o = 1000$
UDNN3-2-2	UDCNN3-E, UDFNN with 2 layers, $n_h = 2500$, $n_o = 2000$
UDNN3-2-3	UDCNN3-E, UDFNN with 2 layers, $n_h = 2500$, $n_o = 3000$
UDNN3-2-4	UDCNN3-E, UDFNN with 2 layers, $n_h = 2500$, $n_o = 4000$

Table 8. The classification performance of the proposed UDNN under different configurations.

	UDCNN3	UDCNN3-E	UDNN3-1-2	UDNN3-2-3
OA	0.9708±0.0176	0.9755±0.0128	0.9855±0.0087	0.9852±0.0056
Kappa	0.9256±0.0435	0.9372±0.0318	0.9623±0.0221	0.9615±0.0143

Comparison with the Existing Approaches

1. Comparison with the Existing Unsupervised Deep Neural Networks

In order to show the superiority of the proposed UDCNN, we utilized the parameter setting of the proposed UDCNN to construct two UDCNNs using different unsupervised feature learning methods, i.e., sparse filtering (SF) (Ngiam *et al.*, 2011) and enforcing population and lifetime sparsity (EPLS) (Romero *et al.*, 2015). In fact, UDNNs using SF and EPLS have already been utilized in target detection (Dong *et al.*, 2014) and remote sensing image classification (Romero *et al.*, 2016).

The results obtained with the 2003 QuickBird image, the 2005 QuickBird image, the 2007 QuickBird image, and the 2010 QuickBird image are summarized in Figure 6a, 6b), 6c), and 6d, respectively. As illustrated in Figure 6, the proposed UDCNN can clearly outperform the UDCNN using SF (Ngiam *et al.*, 2011) and the UDCNN using EPLS (Romero *et al.*, 2015), in all four temporal images. In addition, the proposed UDNN can improve the performance of the UDCNN in all four temporal images.

2. Comparison with Hand-Crafted Features

As mentioned before, UVs contain more dense buildings and less vegetation than non-UV areas. With this in mind, Huang *et al.* (2015) utilized the morphological building index (MBI) and the normalized difference vegetation index (NDVI) to generate the scene descriptor for distinguishing UV samples from non-UV samples. It was found that the index-based scene descriptor can clearly outperform some of the traditional BoVW descriptors. In order to show the effectiveness of the proposed data-driven feature descriptor (i.e., the proposed UDNN), the proposed UDNN is compared with a hand-crafted feature (i.e., an index-based scene descriptor). In the following, we first give a visual comparison between the results obtained with the proposed data-driven feature and the MBI feature (Huang and Zhang, 2011).

The MBI feature is widely considered to be the best hand-crated feature for the interpretation of urban remote sensing imagery (Huang *et al.*, 2014). However, the MBI feature still encounters some problems. In the following, the second column of Figure 7 denotes the results of the MBI feature. As depicted in the second and third rows of Figure 7, the MBI feature misses some buildings when the buildings are densely distributed. When the buildings present a similar spectral reflectance, some of the adjacent buildings are merged by the MBI feature, as depicted in the third row of Figure 7. These situations have a negative influence on the index-based UV descriptor (Huang *et al.*, 2015), which is highly dependent on the density and size of the buildings. In contrast, the proposed data-driven feature can hierarchically extract the structures from a small scale to a large scale, as depicted in the (c) to (k) columns of Figure 7. More specifically, the first and second feature layers of the proposed feature extraction network can robustly determine the edges and boundaries of buildings in the UV scenes. In addition, the third feature layer can distinguish large buildings and roads. Hence, the proposed data-driven feature is competent for the UV discrimination task.

Furthermore, we give a quantitative comparison between hand-crafted features and the proposed data-driven feature. More specifically, the index-based scene descriptors are the index-based scene descriptor using the MBI and the index-based scene descriptor using both the MBI and the NDVI. In addition, the MBI-based descriptor and the MBI-NDVI-based descriptor were discriminated by the random forest classifier and the SVM classifier. Due to its superiority on a small set of samples, SVM was also adopted in the discrimination module of the proposed approach.

A comparison between the results obtained with the proposed UDNN and the index-based descriptors on the four temporal images is shown in Figure 8. As depicted in Figure 8, the proposed UDNN can achieve the best classification performance on the 2003 QuickBird image and the 2010 QuickBird image. On the 2005 QuickBird image, the proposed UDNN can achieve a classification performance that is comparable to the performance of the best hand-crafted feature (i.e., the MBI-NDVI-based descriptor). In addition, the proposed UDNN can achieve a comparable classification performance to the MBI-NDVI-based descriptor.

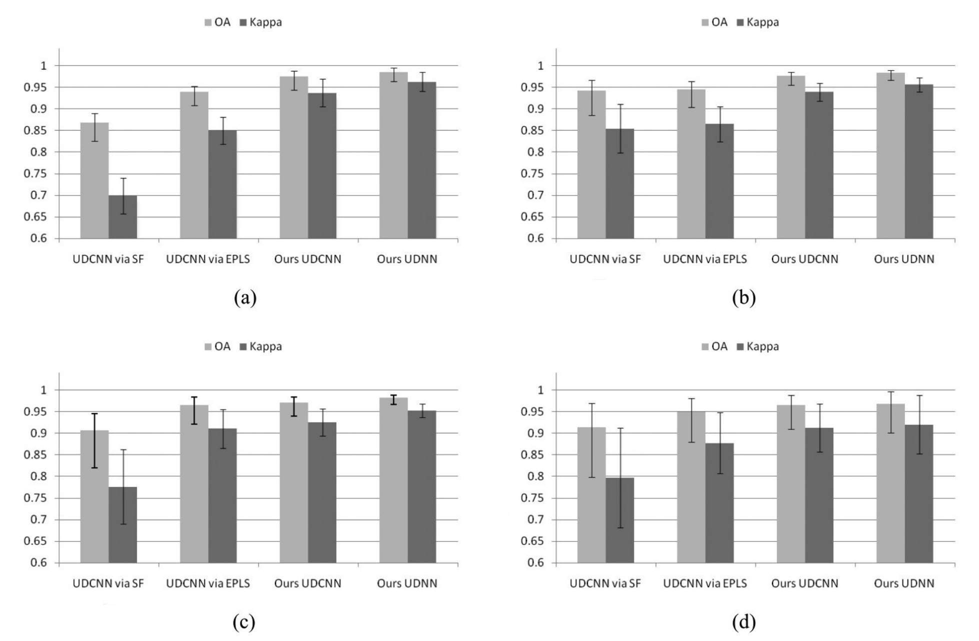


Figure 6. Comparison with the results of other existing UDNNs: (a), (b), (c), and (d) denote the evaluation results obtained with the 2003, 2005, 2007, and 2010 QuickBird images.

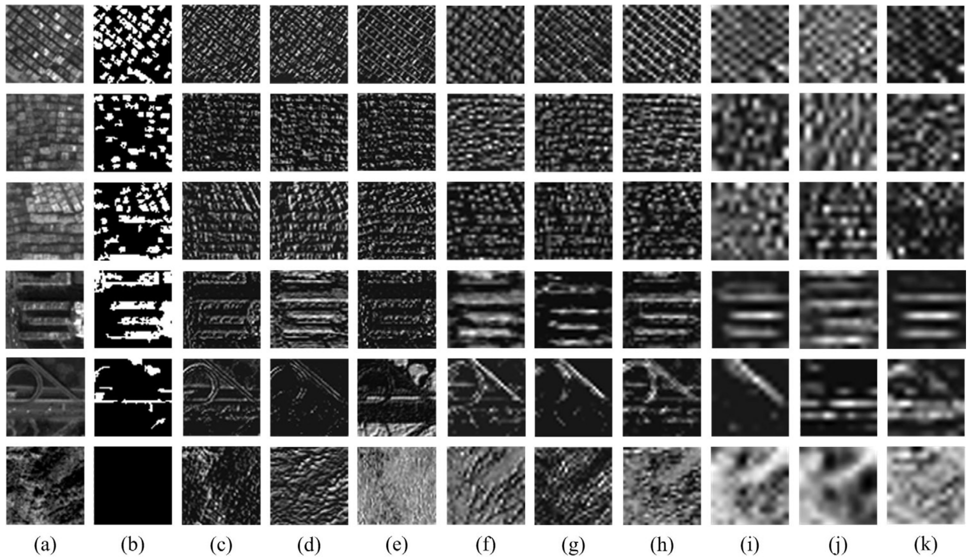


Figure 7. Visual comparison between the MBI feature and the proposed data-driven feature. The first three rows denote the UV scenes, and the last three rows represent the non-UV scenes. (a) The original image scene. (b) The feature map of the MBI index (Huang and Zhang, 2011). (c)–(e) Three random feature maps from the first feature layer of the proposed UDCNN. (f)–(h) Three random feature maps from the second feature layer of the proposed UDCNN. (i)–(k) Three random feature maps from the third feature layer of the proposed UDCNN.

In order to give an intuitive performance comparison, we show the visual classification results of the different methods on the 2010 QuickBird image. From the visual comparison, as depicted in Figure 9, we can see that the results obtained using the hand-crafted features show a low false alarm rate, but may miss some true areas. In contrast, the proposed UDNN can output a result with a very low omission rate. In general, the data-driven feature of the proposed UDNN can achieve a result that is comparable to the results obtained with the best hand-crafted features. It should be stressed that the data-driven feature can be automatically learned from the data in an unsupervised manner; however, hand-crafted features need to be painstakingly designed through the experience and observations of experts. Although the data-driven feature still cannot clearly outperform the existing hand-crafted features, the data-driven feature is useful as it avoids the need for further expertise and the effort of designing features for new specific tasks.

Discussion

1. Why the Proposed UDNN is Competent for the UV Detection Task

The proposed UDNN is composed of two crucial modules: the UDCNN and the UDFNN. As depicted in our previous works (Li *et al.*, 2016a; Li *et al.*, 2016b), the advocated UDCNN can hierarchically extract the local features, from simple to complex. More specifically, the local features include the simple edge response and the complex corner responses, which are important characteristics of buildings (Cote and Saeedi, 2013). In addition, the density of buildings is a crucial clue to distinguish UV samples from non-UV samples (Huang *et al.*, 2015). Hence, the output feature of the first module (i.e., the UDCNN) of the proposed UDNN has the primary ability to discriminate UV samples from non-UV samples. However, the

UDCNN just encodes the local features and ignores the extraction of the global concept of the image scene. In addition, the existing unsupervised deep learning methods (Romero *et al.*, 2016; Dong *et al.*, 2014) are also confronted with these limitations. With this consideration, for the first time, we utilize the UDFNN to mine the global concept of the primary feature exported by the UDCNN. The performance improvement stemming from the proposed UDFNN was quantitatively reported in Figure 6.

2. The Potential Applications in the Remote Sensing Community for the Proposed UDNN

This paper reveals the construction details of the UDNN, which may help the readers to design their neural networks based on specific applications in the remote sensing community. Compared with supervised deep feature learning methods depending on large numbers of labeled samples (Krizhevsky *et al.*, 2012), the adaptive range of the proposed UDNN is very wide as the proposed UDNN can be trained in a fully unsupervised manner. In addition to the UV detection task reported in this paper, the proposed UDNN could be extended to scene-based interpretation tasks, such as the recognition of central business districts (Taubenbock *et al.*, 2013) and land-cover classification (Yang and Newsam, 2010).

Conclusions and Future Prospects

As a special type of urban settlement, urban villages (UVs) result from the complicated socio-economic development of China, and have attracted a lot of research interest. Differing from the existing UV studies, which have mainly focused on social surveys, this paper has proposed a UV detection technique from the remote sensing perspective. In this paper, we have presented a novel UV detection approach using high-resolution remote sensing images, and we have attempted to construct an up-to-date map of UVs, which is necessary for the UV redevelopment policies of various cities in China.

Differing from our previous UV detection approach relying on hand-crafted features which are painstakingly designed by experts, this paper proposes a deep learning based UV detection approach in which the feature representation can be automatically learned from the data. As labeled samples are a scarce resource in the context of UV detection, the paper presents a novel unsupervised deep neural network (UDNN) to generate the data-driven feature, where the UDNN is composed of the unsupervised deep convolutional neural network (UDCNN) and the unsupervised deep fully connected neural network (UDFNN). More specifically, the UDCNN works to mine the local structures, from simple to complex, and the UDNN can further abstract the feature of the UDCNN from a global perspective.

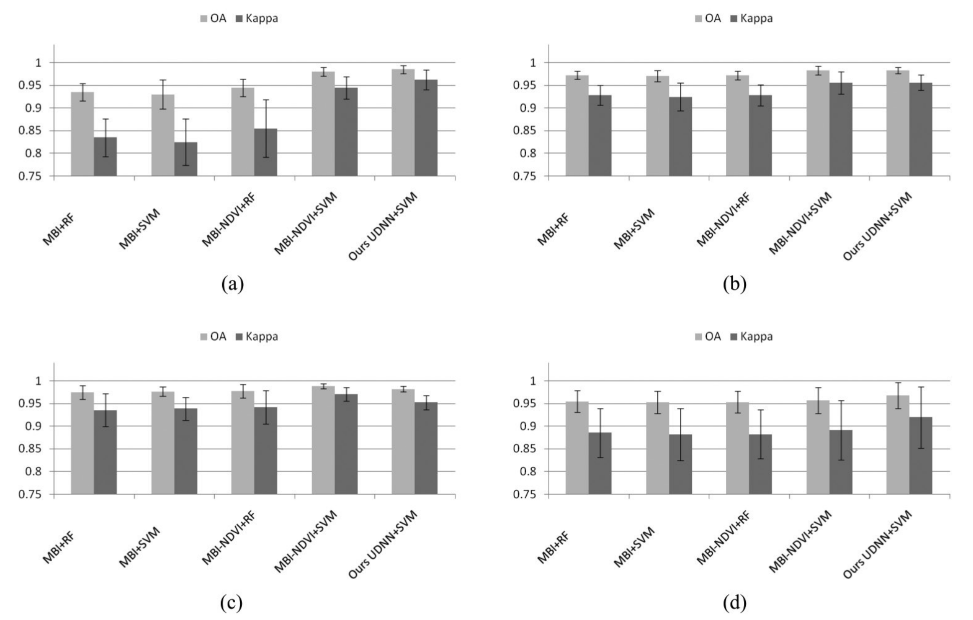
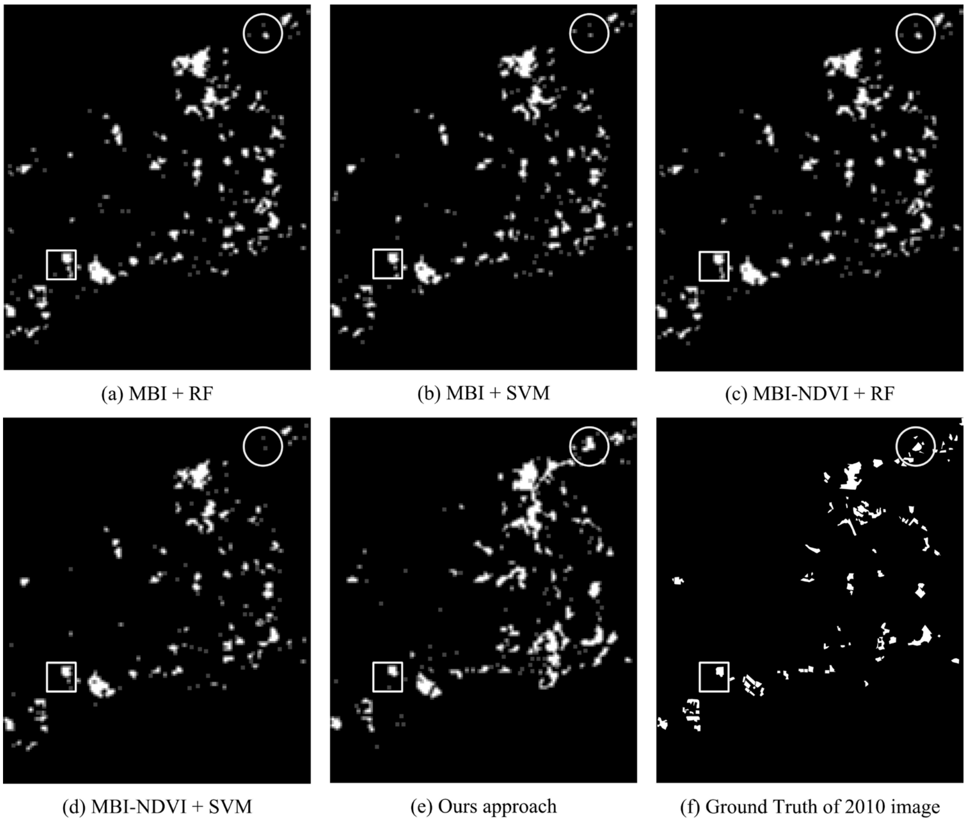


Figure 8. Comparison with the results obtained with hand-crafted features in the UV detection task: (a), (b), (c), and (d) denote the evaluation results obtained with the 2003, 2005, 2007, and 2010 QuickBird images, respectively.



(a) MBI + RF

(b) MBI + SVM

(c) MBI-NDVI + RF

(d) MBI-NDVI + SVM

(e) Ours approach

(f) Ground Truth of 2010 image

Figure 9. Visual comparison of the results obtained with the 2010 QuickBird image.

Compared with the existing unsupervised deep learning methods, the superiority of the proposed UDNN is reflected in two aspects: the proposed UDCNN outperforms the existing UDCNNs, and the UDFNN is utilized for the first time in the context of unsupervised deep learning. Large amounts of experiments with high-resolution remote sensing images confirmed that the proposed data-driven feature from the proposed UDNN can clearly outperform the existing data-driven feature generation methods, and it can achieve a result that is comparable to the results obtained with the best hand-crafted features, such as index-based features. Although the learned data-driven feature cannot clearly outperform the best hand-crafted features, the reported result is still impressive as the data-driven feature comes from a feature extraction network which is trained in a fully unsupervised manner. In our future work, we will consider fusing the proposed data-driven feature with the existing hand-crated features to implement UV detection in more cities of China. In addition, we will try to extend the proposed unsupervised deep feature learning approach to more tasks such as object detection (Cheng *et al.*, 2014; Li *et al.*, 2015b; Li *et al.*, 2016c), feature matching (Ma *et al.*, 2013; Ma *et al.*, 2014), image retrieval (Demir and Bruzzone, 2016), and image fusion (Ma *et al.*, 2016).

Acknowledgments

The work presented in this paper was supported in part by the National Key Research and Development Program of China under Grant No 2016YFB0501403, in part by China Science Fund for Excellent Young Scholars under Grant No. 41522110, and in part by the Foundation for the Author of National Excellent Doctoral Dissertation of PR China (FANEDD) under Grant 201348, in part by the National Natural Science Foundation of China under grant 41601352. The authors would like to thank the anonymous reviewers for their insightful and constructive comments, which significantly improved the quality of this paper.

References

Blaschke, T., G.J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R.Q. Feitosa, F. van der Meer, H. van der Werff, and F. van Coillie, 2014. Geographic object-based image analysis-towards a new paradigm, *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:180–191.

Chuang, H., 2009. The planning of 'villages-in-the-city' in Shenzhen, China: The significance of the new state-led approach, *International Planning Studies*, 14:253–273.

Chung, H., 2010. Building an image of villages-in-the-city: A clarification of China's distinct urban spaces, *International Journal of Urban and Regional Research*, 34:421–437.

Chuang, H., and S. Zhou, 2011. Planning for plural groups? Villages-in-the-city redevelopment in Guangzhou City, China, *International Planning Studies*, 16, 333–353.

Coates, A., U. Lee, and A. Ng, 2011. An analysis of single-layer networks in unsupervised learning, *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, Florida, pp. 215–223.

Cote, M., and P. Saeedi, 2013. Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution, *IEEE Transactions on Geoscience and Remote Sensing*, 51:313–328.

Cheng, G., and J. Han, 2016. A survey on object detection in optical remote sensing images, *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28.

Cheng, G., J. Han, P. Zhou, and L. Guo, 2014. Multi-class geospatial object detection and geographic image classification based on collection of part detectors, *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132.

Chen, S., and Y. Tian, 2015. Pyramid of spatial relations for scene-level land use classification, *IEEE Transactions on Geoscience and Remote Sensing*, 53:1947–1957.

Demir, B., and L. Bruzzone, 2016. Hashing-based scalable remote sensing image search and retrieval in large archives, *IEEE Transactions on Geoscience and Remote Sensing*, 54(2):892–904.

Deng, J., W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, 2009. Imagenet: A large-scale

hierarchical image database, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, Florida, pp. 248–255.

Dong, Z., M. Pei, Y. He, T. Liu, Y. Dong, and Y. Jia, 2014. Vehicle type classification using unsupervised convolutional neural network, *Proceedings of the International Conference on Pattern Recognition*, Stockholm, Sweden, pp. 172–177.

Hofmann, P., J. Strobl, T. Blaschke, and H. Kux, 2008. Detecting informal settlements from QuickBird data in Rio de Janeiro using an object based approach, *Object-Based Image Analysis* (T. Blaschke, S. Lang, and G. Hay, editors), New York, Springer-Verlag, pp. 531–553.

Hofmann, P., 2001. Detecting informal settlements from IKONOS image data using methods of object oriented image analysis - An example from Cape Town (South Africa), *Remote Sensing of Urban Areas* (C. Jugens, editor), Regensburger, Germany: Regensburger Geographische Schriften, pp. 41–42.

Huang, X., and L. Zhang, 2011. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery, *Photogrammetric Engineering & Remote Sensing*, 77(7):721–732.

Huang, X., H. Liu, and L. Zhang, 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 53:3639–3657.

Huang, X., Q. Lu, and L. Zhang, 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas, *ISPRS Journal of Photogrammetry and Remote Sensing*, 90:36–48.

Hinton, G., S. Osindero, and Y. Teh, 2006. A fast learning algorithm for deep belief nets, *Neural Computation*, 18:1527–1554.

Hinton, G., 2002. Training products of experts by minimizing contrastive divergence, *Neural Computation*, 14:1771–1880.

Hu, F., G. Xia, J. Hu, and L. Zhang, 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery, *Remote Sensing*, 7:14680–14707.

Hao, P., P. Hooimeijer, R. Sliuzas, and S. Geertman, 2013. What drives the spatial development of urban villages in China?. *Urban Studies*, 50, 3394–3411.

Khatami, R., G. Mountrakis, and S.V. Stehman, 2016. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research, *Remote Sensing of Environment*, 177:89–100.

Krizhevsky, A., I. Sutskever, and G. Hinton, 2012. Imagenet classification with deep convolutional neural networks, *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, Stateline, Nevada, Lake Tahoe, pp. 1097–1105.

Kuffer, M., K. Pfeffer, R.V. Sliuzas, and I. Baud, 2016. Extraction of slum areas from VHR imagery using GLCM variance, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(5):1830–1840.

Li, Y., Y. Tan, J. Deng, Q. Wen, and J. Tian, 2015a. Cauchy graph embedding optimization for built-up areas detection from high-resolution remote sensing images, *IEEE Journal of Selected Topics in Applied Observations and Remote Sensing*, 8:2078–2096.

Li, Y., Y.Tan, J. Yu, S. Qi, and J. Tian, 2015b. Kernel regression in mixed feature space for spatio-temporal saliency detection, *Computer Vision and Image Understanding*, 135:126–140.

Li, Y., C. Tao, Y. Tan, K. Shang, and J. Tian, 2016a. Unsupervised multilayer feature learning for satellite image scene classification, *IEEE Geoscience and Remote Sensing Letters*, 13:157–161.

Li, Y., Y. Zhang, C. Tao, and H. Zhu, 2016b. Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion, *Remote Sensing*, 8:709–723.

Li, Y., Y. Zhang, J. Yu, Y. Tan, and J. Tian, 2016c. A novel spatio-temporal saliency approach for robust DIM moving target detection from airborne infrared image sequences, *Information Sciences*, 369:548–563.

LeCun, Y., Y. Bengio, and G. Hinton, 2015. Deep learning, *Nature*, 521:436–444.

Lee, H., R. Grosse, R. Ranganath, and A.Ng, 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, *Proceedings of the 26th Annual Internal Conference on Machine Learning*, Montreal, Quebec, Canada, pp. 609–616.

Ma, J., C. Chen, C. Li, and J. Huang, 2016. Infrared and visible image fusion via gradient transfer and total variation minimization, *Information Fusion*, 31:100–109.

Ma, J., J. Zhao, J. Tian, X. Bai, and Z. Tu, 2013. Regularized vector field learning with sparse approximation for mismatch removal, *Pattern Recognition*, 46(12):3519–3532.

Ma, J., J. Zhao, J. Tian, A.L. Yuille, and Z. Tu, 2014. Robust point matching via vector field consensus, *IEEE Transactions on Image Processing*, 23(4):1706–1721.

Ma, J., H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, 2015. Robust feature matching for remote sensing image registration via locally linear transforming, *IEEE Transactions on Geoscience and Remote Sensing*, 53(12):6469–6481.

Mathieu, R., C. Freeman, and J. Aryal, 2007. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery, *Landscape and Urban Planning*, 81:179–192.

Marmanis, D., M. Datcu, T. Esch, and U. Stilla, 2016. Deep learning earth observation classification using ImageNet pretrained networks, *IEEE Geoscience and Remote Sensing Letters*, 13:105–109.

Mountrakis, G., J. Im, and C. OgoIe, 2011. Support vector machines in remote sensing: A review, *ISPRS Journal of Photogrammetry and Remote Sensing*, 66:247–259.

Ngiam, J., P. Koh, Z. Chen, S. Bhaskar, and A. Ng, 2011. Sparse filtering, *Proceedings of Advances in Neural Information Processing Systems*, Granada Congress and Exhibition Centre, Granada, Spain, pp. 1125–1133.

Persello, C., and L. Bruzzone, 2014. Active and semisupervised learning for the classification of remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):6937–6956.

Rhinane, H., A. Hilali, A. Berrada, and M. Hakdaoui, 2011. Detecting slums from SPOT data in Casablanca Morocco using an object based approach, *Journal of Geographic Information System*, 3:217–224.

Romero, A., P. Radeva, and C. Gatta, 2015. Meta-parameter free unsupervised sparse feature learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:1716–1722.

Romero, A., C. Gatta, and G. Camps-Valls, 2016. Unsupervised deep feature extraction for remote sensing image classification, *IEEE Transactions on Geoscience and Remote Sensing*, 54:1349–1362.

Shen, J., 1995. Rural development and rural to urban migration in China 1978-1990, *Geoforum*, 26:395–409.

Stavrakoudis, D.G., J.B. Theocharis, and G.C. Zalidis, 2011. A boosted genetic fuzzy classifier for land cover classification of remote sensing imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, 66:529–544.

Taubenbock, H., M.Klotz, M. Wurm, J. Schmieder, B. Wagner, M. Wouster, T. Esch, and S. Dech, 2013. Delineation of central business districts in mega city regions using remotely sensed data, *Remote Sensing of Environment*, 136:386-401.

Willmore, B.,and D.J. Tolhurst, 2001. Charactering the sparseness of neural code,. *Network*, 12:255–270.

Yang, X., 2000. Determinants of migration intentions in Hubei province, China: Individual versus family migration, *Environment and Planning A*, 32:769–788.

Yang, Y., and S. Newsam, 2010. Bag-of-visual-words and spatial extensions for land-use classification, *Proceedings of the 18th SIGSPATIAL International Conference on Advanced Geographic Information Systems*, San Jose, California, pp. 270–279.

Zacharias, J., and Y. Tang, 2010. Restructuring and repositioning Shenzhen, China's new mega city, *Progress in Planning*, 73, 209–249.

Zhang, X., and S. Du, 2015. A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings, *Remote Sensing of Environment*, 169: 37–49.

Zhao, W., and S. Du, 2016. Learning multiscale and deep representations for classifying remotely sensed imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, 113:155–165.

Zhao, B., Y. Zhong, and L. Zhang, 2016. A spectral-structural bag-of-features scene classifier for very high resolution remote sensing imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:73–85.