Remote Sensing of Environment 214 (2018) 73–86

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery

Bo Huang[a,b,c], Bei Zhao[a,*], Yimeng Song[a]

[a] Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong
[b] Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, PR China
[c] Institute of Space and Earth Information and Science, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Deep convolutional neural networks
Transfer learning
Skeleton extraction
Street block
Land-use mapping
Classification
High spatial resolution
Multispectral remote sensing image

ABSTRACT

Urban land-use mapping is a significant yet challenging task in the field of remote sensing. Although numerous classification methods have been developed for obtaining land-use information in urban areas, the accuracy and efficiency of these methods are insufficient to meet the requirements of real-world applications such as urban planning and land management. In recent years, deep learning techniques, especially deep convolutional neural networks (DCNN), have achieved an astonishing level of performance in image classification. However, the traditional DCNN methods do not focus on multispectral remote sensing images with more than three channels, and they are limited by their training samples. In addition, these methods uniformly decompose large images into small processing units, which chop up the land-use patterns and produce land-use maps with obvious "blocks". In this study, a semi-transfer deep convolutional neural network (STDCNN) approach is proposed to overcome these weaknesses. The proposed STDCNN has three parts: one part involves a transferred DCNN with deep architecture; another part is designed to analyze multispectral images; and the final part fuses the first two parts into a classification layer. Moreover, a skeleton-based decomposing method using street block data is devised to maintain the integrity of the land-use patterns. In two case studies, the proposed method is used to generate urban land-use maps from a WorldView-3 image of a 143 km^2 area of Hong Kong and a WorldView-2 image of a 25 km^2 area of Shenzhen. The results show that the proposed STDCNN obtains an overall accuracy (OA) of 91.25% and a Kappa coefficient (Kappa) of 0.903 for Hong Kong land-use classification, and an OA of 80% and a Kappa of 0.780 for Shenzhen land-use classification. In addition, due to the proposed skeleton-based decomposition method, the proposed method can produce better land-use maps for real-world urban applications.

1. Introduction

Urban land-use mapping is a fundamental method for recognizing and locating land-uses for different purposes, such as industrial, residential, institutional and commercial areas. Urban land-use maps have great value for urban environment monitoring, planning and designing (Voltersen et al., 2014; Wu et al., 2017), In particular, they are useful for the study of phenomena, such as urban heat island effects (Chen et al., 2006), urban transport (Geurs and Van Wee, 2004) and house rents (Fujita, 1989). At present, methods for updating of urban land-use maps depend on the interpretation of aerial photos and field surveys, both of which are laborious and time consuming. With the development of remote sensing technologies, a large number of high spatial resolution (HSR) remote sensing images covering an urban area can be obtained by sensors installed on aircraft or satellites. Although it would be useful to extract land-use information from such HSR remote sensing imagery, a single land parcel used for one purpose (e.g. a residential, commercial, or industrial area) often contains multiple types of land-cover with distinct spatial/spectral/geometric characteristics, e.g. a single residential area may contain trees, buildings, and waterbodies, which makes the automatically mapping of land usage more challenge (Zhao et al., 2016b).

Traditionally, land-use classification methods have operated at the pixel level and assessed the geometrical, textural, and contextual features surrounding the focal pixels. However, such methods are not suitable for urban land-use classification based on HSR imagery. As urban land-use class labels are usually assigned at the land parcel and each land parcel contains a variety of land-covers. HSR images of land parcels are too complex to be categorized at the pixel level (Wu et al., 2009; Zhao et al., 2016b). Object-based classification methods

* Corresponding author.
E-mail address: zhaoys@whu.edu.cn (B. Zhao).

https://doi.org/10.1016/j.rse.2018.04.050
Received 29 October 2017; Received in revised form 20 April 2018; Accepted 28 April 2018
Available online 26 May 2018
0034-4257/

(Blaschke et al., 2014) can have similar problems, as they derive land-use descriptions through the application of co-occurrence (Aksoy et al., 2005), neighborhood-graph-based (Voltersen et al., 2014; Walde et al., 2014) or geometric measure (Huang et al., 2015) methods. These methods can incorporate land-cover information and are compatible with many existing land-cover classification methods. The performance of these methods, however, depends heavily on the selected land-cover classification system and on the accuracy of the land-cover classification. These methods also require knowledge of land-covers and information on land-uses.

Another approach is to use per-field classification methods to directly extract and classify the low-level features of the fields (e.g., spectral, textural, geometrical and contextual features) with predetermined boundaries. This approach has many advantages over the per-pixel or object-based methods of urban land-use classification (Hu and Wang, 2013; Wu et al., 2009). Due to the complexity of the land-use images, per-field classification methods should extract numerous features and exhaustively select an optimal subset of features to obtain the satisfied classification process. To reduce the difficulty and complexity of the feature extraction, a bag-of-visual-words (BOVW) model can be used, which views each land-use image as a bag of "visual terms" (Quelhas et al., 2007; van Gemert et al., 2010), where each term identifies a small aspect of the overall land-use, and captures a simple biophysical characteristic. Instead of the low-level features, the BOVW model represents the land-use images by mid-level features which are obtained by coding the low-level features with a learned "dictionary" of visual terms. The "dictionary" is often from the low-level features through an unsupervised learning algorithm such as the *k*-means (Chen and Tian, 2015; Yang and Newsam, 2011; Zhao et al., 2014), Gaussian mixture model (Perronnin et al., 2010), spectral clustering (Hu et al., 2015a), part-lets detector (Cheng et al., 2015) and sparse dictionary learning (Yang et al., 2009) algorithms. The commonly used feature coding methods (Bosch et al., 2008; Huang et al., 2013) include hard-voting (Zhao et al., 2016b), spare coding (Cheriyadat, 2014; Zheng et al., 2013), fisher coding (Zhao et al., 2016c), and the probabilistic topic models (Bahmanyar et al., 2015; Lienou et al., 2010; Luo et al., 2014; Zhao et al., 2016a, 2013; Zhong et al., 2015). Due to the adoption of the BOVW model and the effective organization of low-level features, the mid-level-feature-based methods can obtain better classification accuracy than low-level-feature-based methods.

All of the aforementioned per-pixel, object-based and per-field land-use classification methods are based on shallow architectures and hand-craft feature descriptors, which fail to capture the fine features of the complex land-use images used for generalization. Consequently, none of thee methods achieve the level of accuracy required by practical applications. In an urban land-use scheme, land-use can be described at many levels, including pixel intensities, edges, object parts, objects (building, trees, roads, etc.), and land parcels, all of which can be represented efficiently with deep architectures. Deep learning is a process through which a set of machine learning algorithms attempt to model high-level abstractions of data by using deep architectures composed of multiple nonlinear transformations (LeCun et al., 2015). As deep learning is able to model the hierarchical representations of features and as urban land-use schemes can be described by such features, the deep learning model is a very promising avenue to address urban land-use classification problems. Among the various deep learning techniques, the deep convolutional neural networks (DCNN) method has achieved an astonishing level of performance in the land-use classification of HSR images (Jia et al., 2015; Zhang and Du, 2016). DCNNs are composed of multiple convolutional layers, and are able to learn high-level abstract features from the original pixel values of land-use images. However, the increase in the number of layers increases the number of parameters in the DCNN, which creates the requirement for a large amount of training samples. To reduce the required number of training land-use samples, some researchers have proposed using transfer DCNNs (Castelluccio et al., 2015; Hu et al., 2015b; Marmanis et al., 2016; Penatti et al., 2015; Zhao et al., 2017) or small DCNNs with only a few layers (Zhang et al., 2016) to prevent the overfitting of trained networks. However, the traditional transfer DCNNs only incorporate the grey or RGB images and fail to adapt to HSR multispectral images (which have more than three channels), whereas the small DCNNs are unable to take advantage of deep architecture. Therefore, a new method is needed that can use transfer DCNNs and small DCNNs to determine land-use classification from HSR multispectral images. Moreover, existed developed land-use classification methods tend to evenly split the large HSR images into small processing units of fixed sizes through the uniform decomposition method (Zhang et al., 2014, 2016). This method chop up the patterns of land-use and generates a land-use map with obvious "blocks". As a result, the land-use maps obtained by traditional DCNNs do not meet the standards needed for practical applications.

To solve these problems, this study proposes a semi-transfer deep convolutional neural networks (STDCNN) method of land-use classification for urban land-use mapping based on HSR multispectral images. This STDCNN system consists of three parts. The first part of the system is a DCNN that is transferred from the pretrained AlexNet model. This model has been trained with a large set of natural images, including more than 1.2 million images and 1000 classes of topographical features (Krizhevsky et al., 2012), and is available for free on the Internet. This transferred DCNN allows the proposed STDCNN to acquire a deep architecture. The second part of the method is a small DCNN with only a few layers, which is designed for interpreting multispectral images. This small DCNN needs to be trained on the HSR multispectral images with randomly initialized parameters. The third part of the STDCNN method contains a fully connected layer and a softmax layer; the former layer fuses the first two parts, and the latter layer generates a final confidential vector of the land-use image. The proposed network can be trained through a semi-transfer process. Due to the small DCNN and the transfer DCNN, the proposed STDCNN can obtain good outcomes using the limited number of training samples derived from HSR multispectral imagery.

To obtain an urban land-use map, a street block is recommended as the minimum land-use mapping unit (Hu and Wang, 2013; Voltersen et al., 2014; Walde et al., 2014; Wu et al., 2009; Zhang and Du, 2015). However, street blocks are always irregular in shape and are not suitable for input into a DCNN model. Therefore, a skeleton-based decomposition method that incorporates the street block data is proposed to adaptively split every mapping unit (or street block) into processing units with regular shapes. The proposed decomposition method maintains a better integrity of the mapping units than the uniform decomposition method. Case studies with a WorldView-3 image covering 143 km^2 of Hong Kong and a WorldView-2 image covering 25 km^2 of Shenzhen demonstrate that the proposed STDCNN can obtain better land-use classification accuracies, and can produce more practical land-use maps.

2. Study area and classification system

2.1. Study area and data collection

Hong Kong is one of the world's most significant financial centers, and one of the most popular destinations for visitors. The city's service sector-dominated economy is characterized by free trade and low taxation, and Hong Kong has been consistently listed as the freest market economy in the world. The second study area is Shenzhen which is a major city in Guangdong Province, China and one of the five largest and wealthiest cities of China. The city is located immediately north of Hong Kong Special Administrative Region with more than 10 million population in 2015. Land-use mapping of these areas are valuable for better understanding and analysis of the cities. The complicated spatial arrangement and the various types of land-use in these areas make them worthwhile to generate the land-use map automatically. In this study, the area under investigation covers the metropolitan areas of both Hong

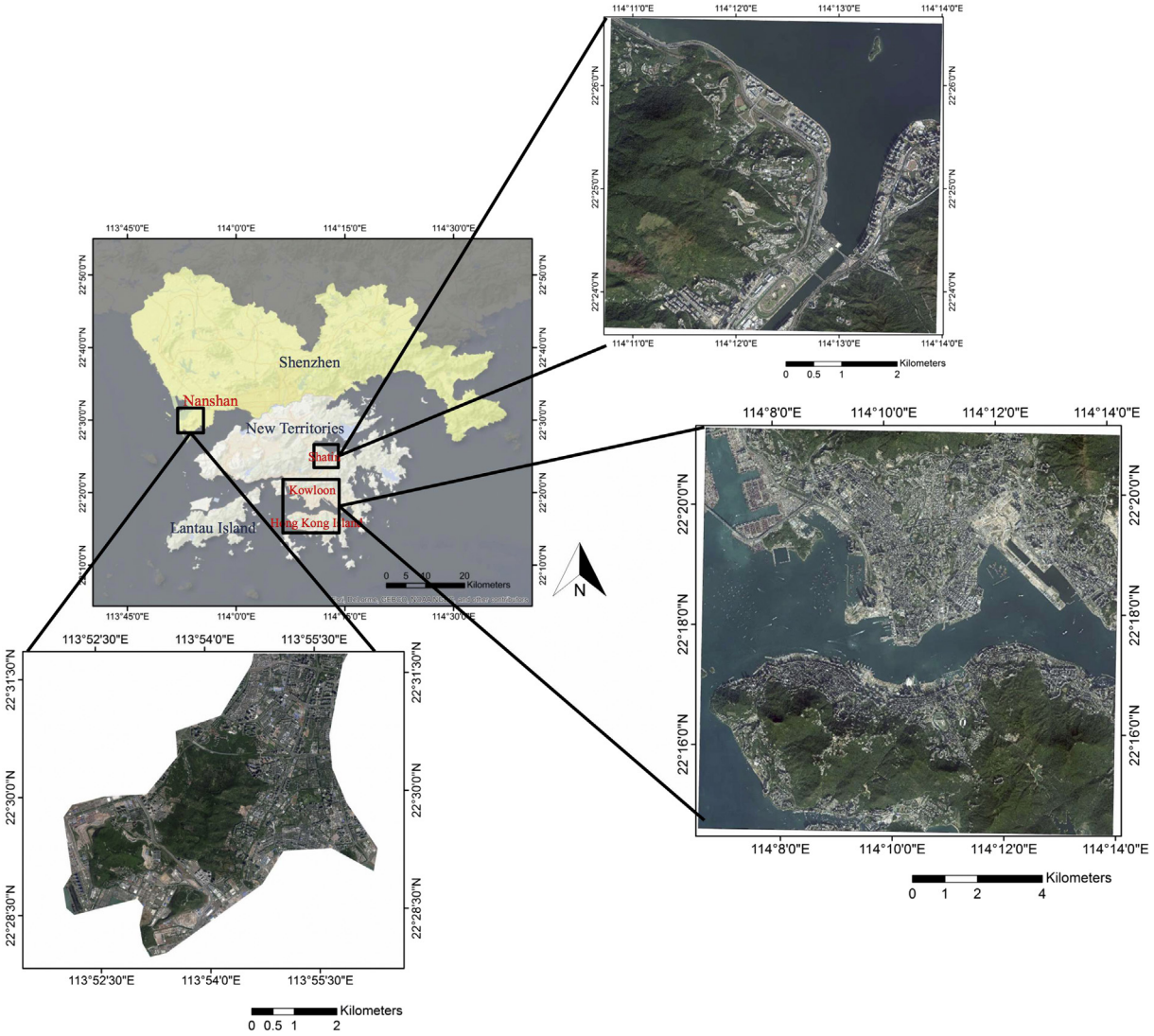

Fig. 1. Study areas and image data including the WorldView-3 images of Hong Kong and the WorldView-2 image of Shenzhen.

Kong and Shenzhen, including two areas of Hong Kong (Kowloon Peninsula, the northern edge of Hong Kong Island and the Shatin area in the New Territories) and one area of Shenzhen (Nanshan area in the southwest of Shenzhen) (Fig. 1).

For the Hong Kong area, there are two WorldView-3 images with a resolution of 1.24 m, which were acquired for land-use mapping on October 15, 2015. One WorldView-3 image covers Kowloon and Hong Kong Island with a size of 9472 × 9728, and the other image covers Shatin with a size of 4352 × 4352. Both images contain 8 channels covering 400–1040 nm spectral bands, including the coastline, blue, green, yellow, red, red edge, near infrared 1 and near infrared 2. The vector data on street blocks and roads are provided by the Planning Department of Hong Kong, and the update time for these data is 2013. Although the WorldView-3 images and the vector data were obtained at different times, there were few changes in land-use boundaries during these two years, and the potential influence of differences between these times is limited and can be ignored.

For Shenzhen, the WorldView-2 image was acquired on April, 2015 with a spatial resolution of 0.5 m, a size of 14,546 × 17,361 and four spectral bands (blue, green, red, and near infrared). The vector data on street blocks and roads are provided by the Urban Planning, Land & Resources Commission of Shenzhen Municipality, and the update time for these data is 2015.

2.2. Classification system and samples

Investigation of the Hong Kong area shows that 11 types of image structures can be used to characterize different land-use classes, including the commercial, institutional, port, dense residential, sparse residential, woodland, water, open space, vacant, industrial and container terminal classes (Fig. 2). In Shenzhen, the image structures of land-use classes appear very different resulted from both the rise of spatial resolution of the WorldView-2 image and different cultural and geographical condition. For Shenzhen, another 11 types of image structures are explored to represent the following land-use classes: commercial, institutional, dense residential, sparse residential, woodland, open space, vacant, industrial, container terminal, and road (Fig. 2). Compared to the Hong Kong area, the image structure of the road class is added for the Shenzhen area, because the highway, intersection, and overpass occupy a large number of pixels due to the 0.5 m spatial resolution which should not be neglected in the process of classification. The pairs of image structure types and the main land-use classes are listed in Table 1. In addition, the corresponding urban functions of image structures are also described in Table 1.

In the Hong Kong area, the training and testing samples for the image structure types are collected from the WorldView-3 image. There are 36 samples per land-use class collected for training, each with a size of 256 × 256. In addition, the following numbers of testing samples

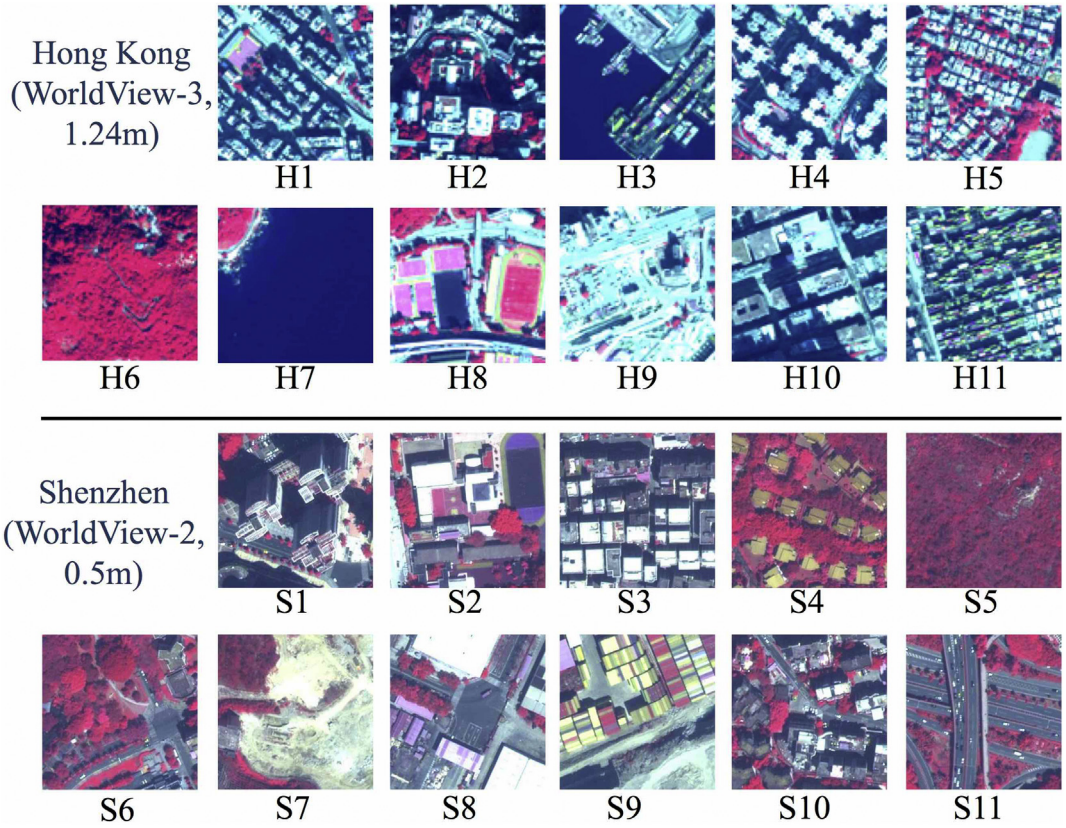

Fig. 2. Images of different structural types with false color (R: near infrared 2; G: red; B: green) for both the Hong Kong and Shenzhen area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Pairs of the image structure types and land-use classes, and their corresponding description of urban functions.

Structure	Main land-use	Urban function
H1, S1	Commercial	Catering, hotel, shop, residential, office
H2, S2	Institutional	Education, church, scientific research, office
H3	Port	Port, ship, water
H4, S3	Dense residential (D_Resid)	Residential with dense population
H5, S4	Sparse residential (S_Resid)	Village, villa, shop, old residential
H6, S5	Woodland	Woods, grass, park building
H7	Water	Water, ship
H8, S6	Open space	Park, sports field, leisure square
H9, S7	Vacant	Unused or developing area
H10, S8	Industrial	Factory, warehouse
H11, S9	Container terminal (C_Termi)	Container, terminal square
S10	Medium residential (M_Resid)	Residential with medium dense population
S11	Road	Highway, overpass, and intersection

(having the same image sizes as the training samples) are collected for the following land-use types: 15 commercial, 25 institutional, 15 port, 25 dense residential, 25 sparse residential, 25 woodland, 25 water, 25 open space, 20 vacant, 25 industrial and 15 container terminal samples. For Shenzhen, the training and testing image samples are extracted from the WorldView-2 image. There are 40 training samples and 30 testing samples per land-use class with a size of 256 × 256 to test the performance of the classification.

3. Methodology

To produce a land-use map from a HSR multispectral image, a STDCNN-based land-use mapping method is proposed. In the proposed method, the HSR image and vector data on the street block and roads are pre-processed, and a STDCNN model is trained. The large HSR multispectral image is then decomposed into processing units using a skeleton-based decomposition method. Subsequently, the trained STDCNN is used to classify the processing units into different land-use classes. Finally, the land-use labels of the processing units are merged into the large land-use map according to the restrictions provided by the street block data. The details of the proposed land-use mapping method are described in the following three subsections.

3.1. Pre-processing HSR images and vector data

The HSR images and the vector data of the street blocks and roads are both projected into the same geo-referenced coordinate system (UTM/WGS84 in this study). The vector data are then co-registered with the HSR images in this coordinate system. For the image data, all of the channels of images are normalized using z-score method. In the normalized image, the values lower than −1 are set to −1, and values higher than 1 are set to 1. Finally, the images are stretched to [0, 255] linearly. This normalization method not only reduces the amount of data, but also keeps enough spatial/spectral information for urban land-use classification of HSR imagery.

3.2. STDCNN

3.2.1. Structure of STDCNN

The structure of the proposed STDCNN, shown in Fig. 3, includes a transfer DCNN, a small DCNN with a limited number of layers, and a fully connected layer that fuses the first two parts. The proposed STDCNN model includes a trainable multilayer architecture that contains a number of convolutional, pooling and fully connected layers.

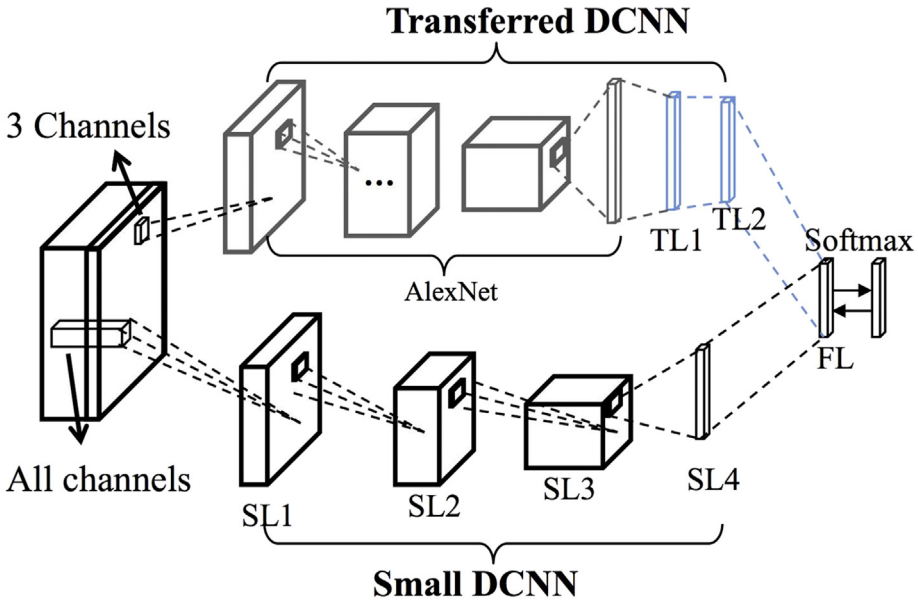

Fig. 3. Structure of the proposed STDCNN.

3.2.1.1. Convolutional layer. The input to each convolutional layer is an $m \times n \times r$ feature map, where r is the number of channels of the feature map, and $m \times n$ is the size of the feature map. The convolutional layer has K filters of size $l \times l \times r'$, where the size of $l \times l$ is smaller than the size of the input feature map, $l < \min m, n$, and r' can either be equal to the number of channels r, or less. The output is an $m' \times n' \times K$ feature map with K channels and a size of $m' \times n'$. Let W_k be the k-th filter, and $\mathbf{X}$ be the input feature map. The output k-th feature map, z_k, can be computed by the equation $z_k = W_k * \mathbf{X} + b_k$, where * is the 2-D discrete convolutional operator, and b is the trainable bias. Following the convolutional operator, a nonlinearity function $f(\bullet)$ is often applied to the output feature map, for which the rectified linear unit, $f(x) = \max(0, x)$, is commonly used.

3.2.1.2. Pooling layer. The pooling layer simply takes the activations within small spatial regions of each feature map and then use the maximum or average operator to extract values for the spatial regions. The pooling layers using the maximum operator or average operator are called the maximum pooling or the average pooling, respectively.

3.2.1.3. Fully connected layer. The fully connected layer is very similar to the convolutional layer. However, unlike the convolutional layer, the size of the filters of the fully connected layer, $m \times n \times r$, are the same as those of the input feature map. Therefore, given K fully connected filters, the output of the fully connected layer is a K-dimension vector.

The components in the network structure of the STDCNN are given in Table 2, where B is the number of channels for the input images, and L is the number of land-use classes.

For the transfer DCNN in the STDCNN, the related parameters are initialized by a pretrained DCNN. In the field of image recognition, there have been several successful modern DCNN architectures, such as AlexNet (Krizhevsky et al., 2012), CaffeNet (Jia et al., 2014), GoogLeNet (Szegedy et al., 2014) and VGGNet (Simonyan and Zisserman, 2014). Among these architectures, the AlexNet is highly regarded and is often used as a baseline DCNN. Therefore, the transfer DCNN in the proposed STDCNN uses AlexNet as the basic model. As the pretrained AlexNet is trained by the images with only RGB channels, the RGB channels of input HSR multispectral data are used as the input for the transfer DCNN. The layers of AlexNet that are transferred into the STDCNN include the five convolutional layers and the first fully connected layer, which contain 96, 256, 384, 384, 256 and 4096 hidden units, respectively. Following these transferred layers, two fully connected layers, namely TL1 (with 512 hidden units) and TL2 (with 256 hidden units), are designed to adapt to the new data.

Table 2
Configuration of the proposed STDCNN.

Layer name		Layer type	Input size	Filters	Stripe
Input		Image data	$227 \times 227 \times B$	–	–
Five convolution layers and the first fully connected layer of AlexNet					
Transfer DCNN	TL1	Fully connected	$1 \times 1 \times 4096$	512	–
	TL2	Fully connected	$1 \times 1 \times 512$	256	–
Small DCNN	SL1	Convolution	$5 \times 5 \times B$	64	2
	Pool1	Max pooling	3×3	–	2
	SL2	Convolution	$3 \times 3 \times 64$	128	1
	Pool2	Max pooling	3×3	–	2
	SL3	Convolution	$3 \times 3 \times 128$	256	1
	Pool3	Ave pooling	28×28	–	–
	SL4	Fully connected	$1 \times 1 \times 256$	256	–
Concatenation of TL2 and SL4					
Fusion	FL	Fully connected	$1 \times 1 \times 512$	L	–
	Softmax				

For the small DCNN in the STDCNN, there are three convolutional layers, namely SL1, SL2 and SL3, and one fully connected layer, SL4. The SL1 and SL2 layers are followed by the maximum pooling, but the SL3 layer is operated by the average pooling. The local response normalization is also conducted after the pooling of the first and second convolutional layers (Krizhevsky et al., 2012). The small DCNN is designed for multispectral images without limitations on the number of input channels.

Finally, the outputs of the transfer DCNN and the small DCNN are fused into a single layer, which is the fully connected layer, or FL. After the softmax function, the output vector of the FL layer gives more confident identifications of L land-use classes. If the size of the input image is larger than 227×227, the random cropping and mirroring operator is applied to ensure that the image size is suitable for feeding to the STDCNN (Jia et al., 2014).

3.2.2. Back-propagation (BP) learning

The entire network can be trained by the BP of a loss function (Rumelhart et al., 1986). Let $(\mathbf{X}_i, \mathbf{y}_i)$ be the i-th training sample, where $\mathbf{X}_i$ is the image data, $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \cdots, y_{i,L})$, $y_{i,k} \in \{0, 1\}$ is the true land-use label vector. $\mathbf{W}$ be the parameters of STDCNN; $\Phi(\mathbf{X}_i, \mathbf{y}_i, \mathbf{W})$ be the loss on the i-th sample. The loss function $J(\mathbf{W})$ throughout all of the training

samples can be calculated with Eq. (1), where N is the number of training samples, $r(\mathbf{W})$ is the regularization term, and λ is a weight decay coefficient.

$$J(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N}\Phi(\mathbf{X}_i, \mathbf{y}_i, \mathbf{W}) + \lambda r(\mathbf{W}). \quad (1)$$

The parameters $\mathbf{W}$ can then be updated by minimizing the loss function. N can be very large, so in practice we can use a stochastic approximation of this objective in each BP iteration. A mini-batch of $n \ll N$ training samples can be drawn from the N training samples. To minimize the loss function, a stochastic gradient descent (SGD) (Bottou, 2010) can be used to update the parameters $\mathbf{W}$.

During the training procedure, a step-by-step training strategy is recommended. Before training the STDCNN, the small DCNN is randomly initialized and trained using samples that have all of the channels. To make it possible to obtain the L dimension vector and compute the loss function, a new fully connected layer with a size of $1 \times 1 \times 256$ and L hidden units is concatenated to the SL4. Then, the five convolutional layers and the first fully connected layer of the transfer DCNN are transferred from the pretrained AlexNet. Next, the layers of the small DCNN (SL1, SL2, SL3 and SL4) are initialized by the trained small DCNN model, and the other layers, (TL1, TL2 and FL) are initialized randomly. Finally, the STDCNN is fine-tuned through SGD optimization to obtain the trained STDCNN, which can be used to classify the processing units of the HSR multispectral imagery.

3.3. Skeleton-based decomposition method with the street blocks

To deal with the large HSR multispectral image, a decomposition method should be used to split the large image into small processing units. Previous studies (Voltersen et al., 2014; Zhang and Du, 2015) have proposed dividing large images into parcels (mapping units) based on street blocks to generate a practical land-use map. However, the shapes of these mapping units are often irregular, or have different scales, which may make them unsuitable as input for the DCNN model. This study proposes a skeleton-based decomposition method that samples every mapping unit and represents it with a set of processing units of a regular size. To represent the mapping units more accurately, the center of the union of the processing units should be placed at the center of the mapping unit for emphasis. In a mapping unit, the skeleton can maintain a unit's the geometric form and locate the unit's central area [Fig. 4 (a)]. Therefore, it is reasonable to use the skeleton of the mapping unit to derive the set of candidate centers of the processing units.

3.3.1. Morphological skeleton extraction

In the proposed method, the morphological operator (Haralick and Shapiro, 1992) is used to extract the skeletons of the mapping units. For each mapping unit, the minimum bounding box is used to generate a binary image X where the values of pixels located in the mapping unit are set to 1, and the others are set to 0. Given a structuring element $\mathbf{B}$, a family of shapes $\{n\mathbf{B}\}_{n=1}^{\infty}$ can be constructed, where, $n\mathbf{B} = \underbrace{\mathbf{B} \oplus \cdots \oplus \mathbf{B}}_{n \text{ times}}$, n is the size of the structuring element of n $\mathbf{B}$. The skeleton $S(X)$ can be obtained by the morphological Eq. (2), where INF is an infinite number, $\ominus$ is the binary erosion operator and $\circ$ is the binary opening operator. With the increase of n, the operator $(X \ominus n\,\mathbf{B})$ shrinks the areas with the value 1 until all values in X become 0. At that time, the process of skeleton extraction is completed for the mapping unit.

$$S(X) = \bigcup_{n=1}^{INF} S_n(X), \quad S_n(X) = (X \ominus n\mathbf{B}) - (X \ominus n\mathbf{B}) \circ \mathbf{B}. \quad (2)$$

3.3.2. Determination of the centers of processing units

Using the morphological skeleton operator, it is possible to obtain a skeleton with a connected line that is 1 pixel thick [Fig. 2 (a)]. The points on this skeleton are the candidates for being the centers of the processing units. To select the final centers from these candidates, an iteration algorithm is tested. Given N candidates $\mathbf{P}_{can}^{(t)} = \{\mathbf{p}_i^{(t)}\}_{i=1}^{N}$, $\mathbf{p}_i^{(t)}$ is the location of the i-th candidate at the t-iteration, and the center of the mass O of the candidates is $\mathbf{p}_O^{(t)} = \sum_{i=1}^{N}\mathbf{p}_i^{(t)}/N$. The first center is selected as the point with the minimum distance to $\mathbf{p}_O^{(t)}$, $\mathbf{p}_c^{(t=1)} = \min_{\mathbf{p}^{(t=1)} \in \mathbf{p}_i^{(t=1)N}_{i=1}} D(\mathbf{p}^{(t=1)} - \mathbf{p}_O^{(t=1)})$ [Fig. 2 (a)]. The other candidates in the processing unit that correspond to the first center $\mathbf{p}_c^{(t=1)}$ are removed, and the points on the border of the processing unit $\overline{\mathbf{p}}_c^{(t)}$ are added into the next generation set of candidates: $\mathbf{P}_{can}^{(t+1)} = \mathbf{P}_{can}^{(t)}/\mathbf{p}_c^{(t)} \bigcup \overline{\mathbf{p}}_c^{(t)}$ [Fig. 2 (b)]. The next center is found in the set of the remaining candidates, $\mathbf{P}_{can}^{(t+1)}$, according to the same minimum distance rule applied to find the first center. The iteration of selecting, removing and adding steps is not stopped until the set of candidates $\mathbf{P}_{can}$ becomes empty [Fig. 2 (d)]. Following the selection of centers from the first generation of candidates, a similar selection is conducted on the next generation of candidates [Fig. 2 (e)]. When no candidate remains, the center selection step for the processing units ends.

For example, in Fig. 2, point 1 is selected as the first center of the processing unit, because it is nearest to the center of the mass of candidates O. The candidates in the first processing unit are removed from the set of candidates [Fig. 2 (b)]. Among the remaining candidates, point 2 is the nearest point, and points 3, 4 and 5 are gradually selected to be centers [Fig. 2 (c) and (d)]. As candidates are removed from the processing unit, the points on the white border line are added to the next generation of candidates. After the candidate set of the former generation (within the yellow lines) becomes empty [Fig. 2 (d)], the center selection processing continues by using the candidates in the next generation (white lines) [Fig. 2 (e)]. When there is no candidate remaining in any generation of candidate sets, the center selection process ends [Fig. 2 (f)].

The result of the skeleton-based decomposition of a small HSR image is shown in Fig. 5. The unified decomposition selects centers with fixed spacing, without considering the street blocks. Compared to the unified decomposition. The skeleton-based decomposition method performs better than the unified decomposition method in obtaining the centers for each mapping unit, and this process ensures that the centers are located in the main areas of the mapping units.

3.4. Land-use mapping

The processing units generated by the decomposition method can then be classified by the trained STDCNN. After classification, an L dimension vector $\mathbf{y} = (y_1, \cdots, y_L)$ can be obtained, where the value in the k-th component, y_k, indicates the level of confidence of the identification of the corresponding k-th land-use class. To get the land-use map for a mapping unit, all of the processing units with centers in that mapping unit should be combined with the following two steps: calculation of a confidential vector for each pixel and combine the confidential vectors of pixels in the mapping unit.

3.4.1. Calculation of a confidential vector for each pixel

For each processing unit, there can be three types of regions, R_1, R_2 and R_3, and each one needs to be treated differently (Fig. 6). R_1 is the region in the mapping unit that is overlapped by other processing units; R_2 is the regions covered by only one processing unit, and R_3 is a region of a processing unit that is not in a mapping unit. The pixels in R_3 are ignored, because they are not in the mapping unit.

Let the confidential vector of the pixel p in the processing unit be $\mathbf{y}_p$. The calculation of $\mathbf{y}_p$ can be calculated using Eq. (3). In Eq. (3), $\{\mathbf{y}_i\}_{i=1}^{n}$ indicates the confidential vectors of the processing units covering R_1 or R_2, and n is the number of related processing units, where $n = 1$ for R_2.

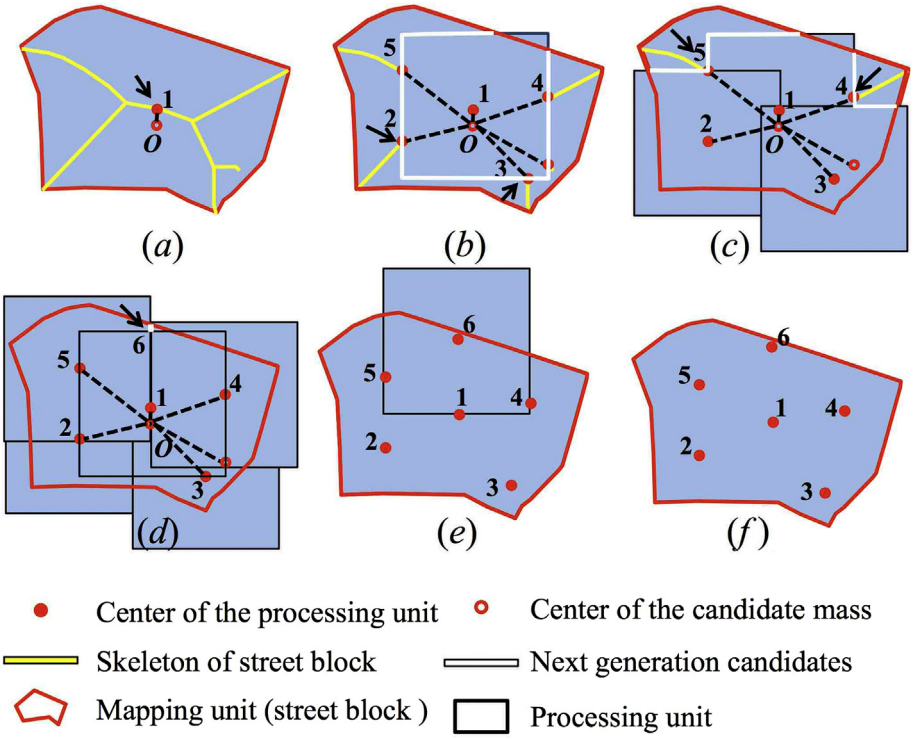

Fig. 4. Decomposition of the HSR image, based on the skeleton of the street block; (a) extract the skeleton, and find the center of the first processing unit; (b)–(d) remove the candidates in the selected center, and find the other centers of processing units; (e) find the centers from the candidates of the next generation; (f) find the final centers of the processing units.

$$\mathbf{y}_p = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i, & \text{if } p \in R_1 \cup R_2 \\ \phi, & \text{otherwise} \end{cases}. \quad (3)$$

3.4.2. Combination of the confidential vectors of pixels

Let $\{\mathbf{y}_p\}_{p=1}^{M}$ be the confidential vector of pixels in the mapping unit, and M be the number of pixels in the mapping unit. The confidential vector of the mapping unit can be computed by the equation $\mathbf{y} = \sum_{p=1}^{M} \mathbf{y}_p$. In addition, the land-use map can be acquired by labeling the mapping units according to the max. confidential rule, $s = \arg\max_{k\in[1,L]} y_k$.

The procedure for the land-use mapping method based on STDCNN is illustrated in Fig. 7. In the first step, the training samples are collected to train the STDCNN. The large HSR multispectral image is split into processing units by the skeleton-based decomposition method. The processing units are then classified by the trained STDCNN model.

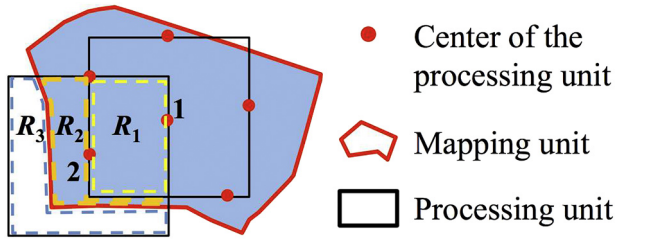

Fig. 6. Post-classification by trimming the processing unit with the mapping unit.

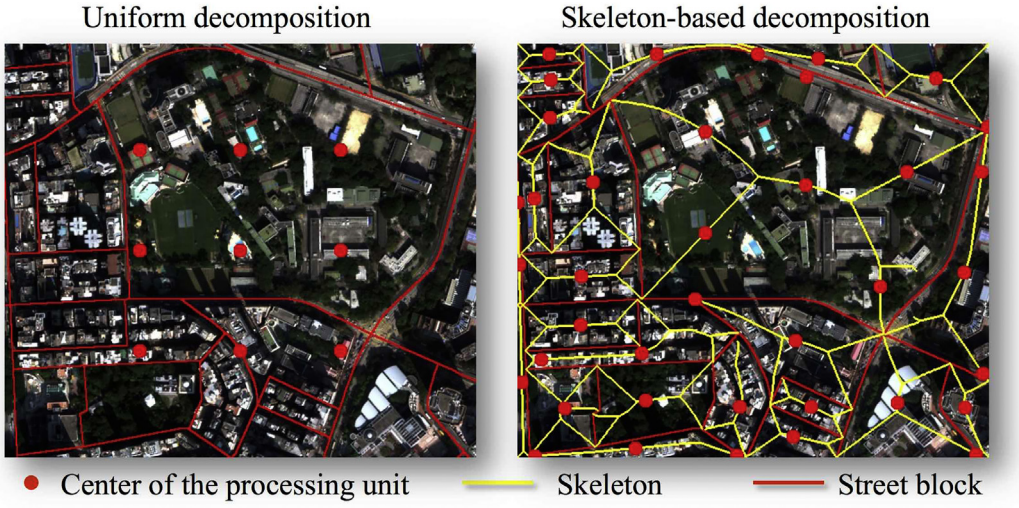

Fig. 5. Skeleton-based decomposition results of an HSR image with the street blocks.

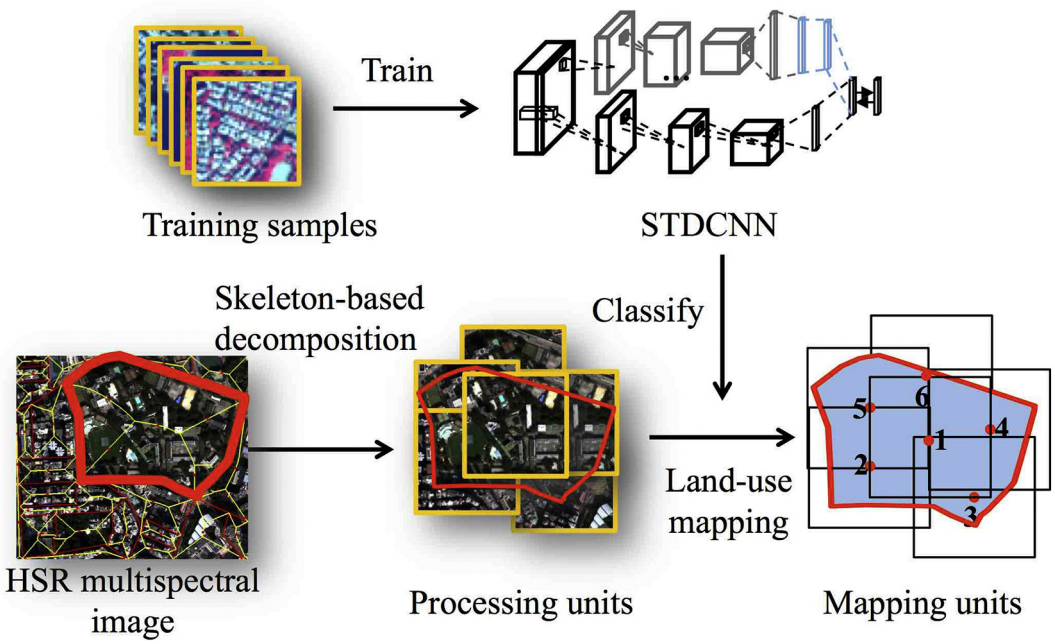

Fig. 7. Land-use mapping method based on STDCNN.

Finally, the processing units are combined to generate a land-use map.

4. Experimental evaluation and analyses

To obtain land-use maps of the study areas, two steps are conducted, namely the land-use classification and the land-use mapping. The land-use classification involves classifying the processing units, i.e., the small land-use images, and the land-use mapping involves decomposing the large HSR images and combining the classification results of the processing units to generate a final land-use map. The confusion matrix, overall accuracy (OA), and Kappa coefficients (Kappa) are used to evaluate the performance of the land-use classification; the McNemar's test is also applied to evaluate the accuracy of different methods. In addition, the land-use maps are evaluated through a visual interpretation of the image and the points of interest.

4.1. Evaluation of land-use classification

In the evaluation experiments, the transfer DCNN (Zhao et al., 2017) and the small DCNN of the STDCNN are used to classify the land-use images by adding a fully connected layer with a size of $1 \times 1 \times 256$ and 11 hidden units. If the parameters of any layer in the three DCNNs are pretrained (before the DCNN training), the learning rate of that layer is set to 0.001, and the other layers are set to 0.01. For the SGD optimization algorithm, the batch size is set to 66, the weight decay is set to 0.0005, the momentum is set to 0.9, and the number of iterations is set to 10,000. These parameters are set at levels that let the algorithm perform well. These experiments are conducted with the Caffe platform (Jia et al., 2014) on the Ubuntu 12.04 operation system with 8 Intel Xeon(R) E5620 @ 2.40 GHz CPU and an NVIDIA GTX 670 GPU with 4 GB memory. The accuracies of the land-use classification produced by the three DCNNs are reported in Table 3, and the confusion matrices are shown in Fig. 8. The classification accuracies of the methods mentioned in Yang and Newsam (2011) and Zhao et al. (2013) are also shown in Table 3. The McNemar's tests are performed to assess the differences between the STDCNN and other methods (Table 4).

As shown in Table 3, the proposed STDCNN achieves the highest OA (91.25%/80.00%) and Kappa (0.903/0.780) for the two datasets. A comparison of the transfer DCNN (OA 87.91%/71.52% and Kappa 0.867/0.687) and the small DCNN (OA 85.83%/76.06% and Kappa 0.843/0.737) shows that the STDCNN improves the OA by more than 3%/8% and 5%/3.5%, and the Kappa by more than 0.03/0.09 and 0.06/0.04 over the transfer DCNN and the small DCNN for the Hong Kong dataset and the Shenzhen dataset, respectively.

Table 3
Classification accuracies of the proposed STDCNN.

Method		Hong Kong	Shenzhen
Yang and Newsam (2011)	OA(%)	80.00	71.82
	Kappa	0.779	0.690
Zhao et al. (2013)	OA (%)	77.08	70.00
	Kappa	0.747	0.670
Transfer DCNN (Zhao et al., 2013)	OA (%)	87.91	71.52
	Kappa	0.867	0.687
Small DCNN	OA (%)	85.83	76.06
	Kappa	0.843	0.737
STDCNN	OA (%)	**91.25**	**80.00**
	Kappa	**0.903**	**0.780**

The data with bold format indicate that the data is better than the other data without bold format.

The *p*-values given in Table 4 suggest that the differences in accuracy between the STDCNN and transfer DCNN in the Hong Kong dataset (0.0801) and between the STDCNN and small DCNN in the Shenzhen dataset (0.0736) are greater than 0.05, whereas the other differences are less than 0.05. These results indicate that the results of the STDCNN are significantly better than those of the other methods, except for the transfer DCNN in the Hong Kong dataset. The results achieved by the STDCNN in the Shenzhen dataset are significantly better than those of the other methods, except for the small DCNN.

The confusion matrices (Fig. 8) suggest that the application of the proposed STDCNN to the Hong Kong dataset achieves accuracies that are equal to or more than 80% for all classes, and more than 90% accurate for six of the land-use classes, namely commercial, port, woodland, water, industrial and container terminal. Compared to the transfer DCNN, the STDCNN more accurately identifies commercial (93.3%), institutional (84.0%), port (93.3%), dense residential (84.0%), open space (88.0%) and industrial (100.0%) areas. The STDCNN also gives a better classification performance than the small DCNN for open space, vacant space and industrial areas. For the Shenzhen dataset, the proposed STDCNN is over 90% accurate in the classification of vacant,

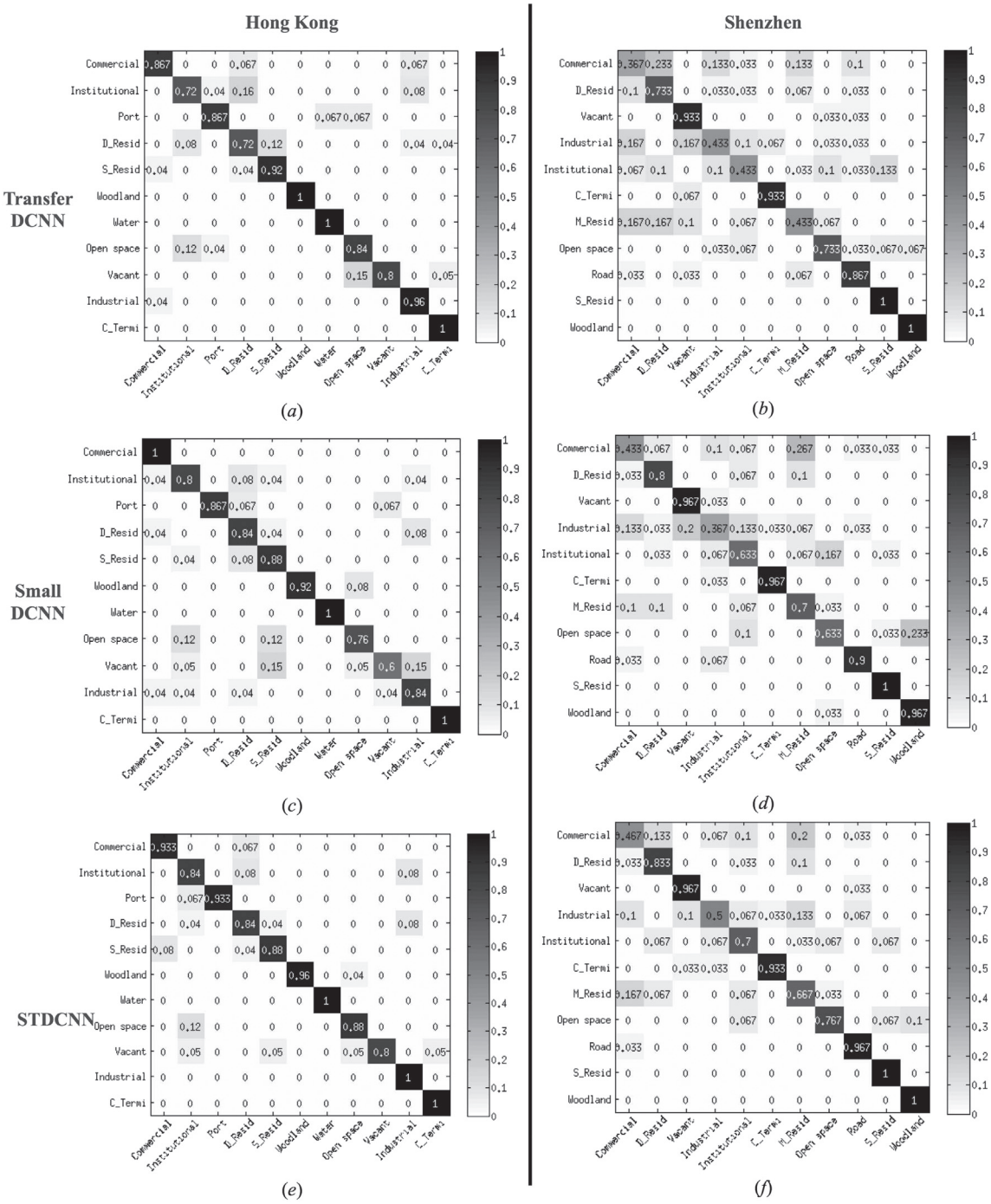

Fig. 8. Confusion matrices of land-use classification with different methods for the Hong Kong and Shenzhen datasets: (a) and (b) transfer DCNN; (c) and (d) small DCNN; (e) and (f) STDCNN; (a), (c), and (e) Hong Kong; and (b), (d), and (f) Shenzhen.

Table 4
p-Values of McNemar's tests between the proposed STDCNN with other classification methods for the Hong Kong and Shenzhen dataset.

Methods	*p*-Value	
	Hong Kong	Shenzhen
Yang and Newsam (2011)	0.0000	0.0059
Zhao et al. (2013)	0.0000	0.0007
Transfer DCNN (Zhao et al., 2017)	0.0801	0.0003
Small DCNN	0.0163	0.0736

container terminator, road, sparse residential, and woodland areas. Compared to the transfer DCNN and the small DCNN, the proposed STDCNN more accurately classifies the dense residential, institutional, open space, and road areas. However, both the proposed STDCNN and the other two methods do not obtain satisfactory accuracy in the classification of commercial, industrial, and medium residential, primarily due to the difficulty in discriminating between these land-use classes in HSR images without the auxiliary data on human activities.

A comparison of the accuracy curves during the training process for the transfer DCNN, small DCNN and STDCNN in Fig. 9 shows that the training and testing curves of the transfer DCNN and STDCNN rise sharply in the first several hundred iterations and then remain

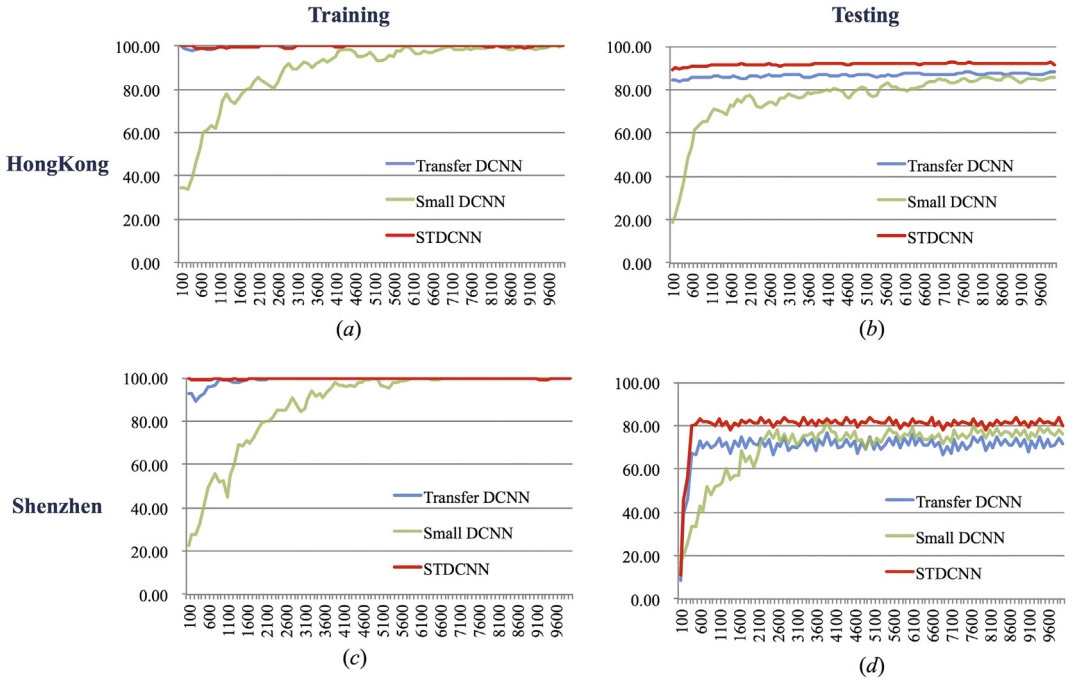

Fig. 9. Accuracy curves of the different methods during the training process: (a) and (c) training accuracies; (b) and (d) testing accuracies; (a) and (b) Hong Kong; (c) and (d) Shenzhen.

relatively stable for both datasets, whereas the training and testing curves of the small DCNN increase relatively slowly at first, and then remain stable after several thousand iterations for both datasets. This pattern indicates that the transfer DCNN and STDCNN trained with the partially pretrained parameters of the networks converge faster, during the training process. The testing curve of the STDCNN stays above that of the other methods, indicating that the proposed DCNN obtains a higher accuracy with relative stability.

These results indicate that the performance of a transfer DCNN can be improved, because the transfer DCNN uses only the RBG channels of the multispectral images. The small DCNN, however, learns the features of the images without a deep architecture, despite using all of the multispectral images' channels. The proposed STDCNN can combine the advantages of the transfer DCNN and the small DCNN to obtain a better performance than either of the other two methods alone.

4.2. Evaluation of land-use mapping

After training, the trained STDCNN model is used to classify the processing units obtained by either the uniform decomposition or the skeleton-based decomposition. In the uniform decomposition method, the processing units are obtained by splitting the HSR multispectral image into equal sections of 256 × 256 with a spacing of 128; this ensures that each pair of two 4-connected neighborhood units has a 50% overlap. This setting is consistent with the skeleton-based decomposition process. The confidential vectors of the pixels in the overlapped areas are computed by averaging the confidential vectors of processing units contained in each overlapped area. By applying the maximum rule, the land-use labels of pixels can be derived from the confidential vectors. Figs. 10 and 11 display the four land-use maps obtained by the STDCNN with uniform and skeleton-based decomposition methods for the Hong Kong and the Shenzhen area, respectively. Overall, the land-use map obtained using the proposed skeleton-based decomposition method are more elaborate than those obtained using the uniform decomposition method.

Four small areas of the maps are magnified for closer comparison. The amplified images are generated by overlapping the original HSR multispectral image with the land-use map, with 40% transparency. The amplified images shown in Fig. 10 show that the industrial area in image 1 and the dense residential area in image 2 are split by the uniform decomposition method, but are well-preserved in the maps using the skeleton-based decomposition method. In addition, the sparse residential area in image 3 and the dense residential area in image 4 are misclassified in the process using the uniform decomposition method, because the center of the residential area is not located in the processing unit, and is therefore easily misclassified as a different land-use class (institutional). This problem is solved by the use of the skeleton-based decomposition method, which ensures that the center of the street block is placed at the center of the processing unit. Similarly, the amplified images in Fig. 11 shows that a commercial area in area 1, an open space area in area 2, and an institutional area in area 3 are clearly recognized by the proposed decomposition method, but are misclassified by the uniformly decomposition method. Therefore, the proposed skeleton-based decomposition method provides a better performance than the commonly used uniform decomposition method.

The final land-use map can be acquired after applying the constraint of the street block data,. Let a mapping unit (street block) be $O = \{\mathbf{p}_1, \cdots, \mathbf{p}_N\}$, where $\mathbf{p}_j$ is the confidential vector of the j-th pixel in this unit. The confidential vector of the mapping unit can be calculated by $\mathbf{p}_O = \sum_{j=1}^{N} \mathbf{p}_j / N$. Let the major land-use class of a street block be the class of the street block indicated by the use of $\arg\min_i p_{O,i}$, where $p_{O,i}$ is a component of $\mathbf{p}_O$. The final land-use maps can then be obtained, as shown in Fig. 12. Due to the lack of a true land-use map, a visual evaluation is used to assess the final land-use maps. This evaluation is done by overlapping the final land-use map with the original HSR image (Fig. 12).

The comparison of the original HSR images with the final land-use maps obtained by the proposed method shows that the obtained land-use map generated by the proposed method comes close to labeling the different urban areas correctly. Among the various land-use classes, the commercial, institutional, residential, open space and industrial areas are difficult to classify correctly. However, three amplified areas containing these types of land-use(Fig. 12) demonstrate that the proposed method can produce satisfactory land-use maps of these areas.

It is worth noting that the confidential vectors of the street blocks contain mixed information from different land-use classes, as these

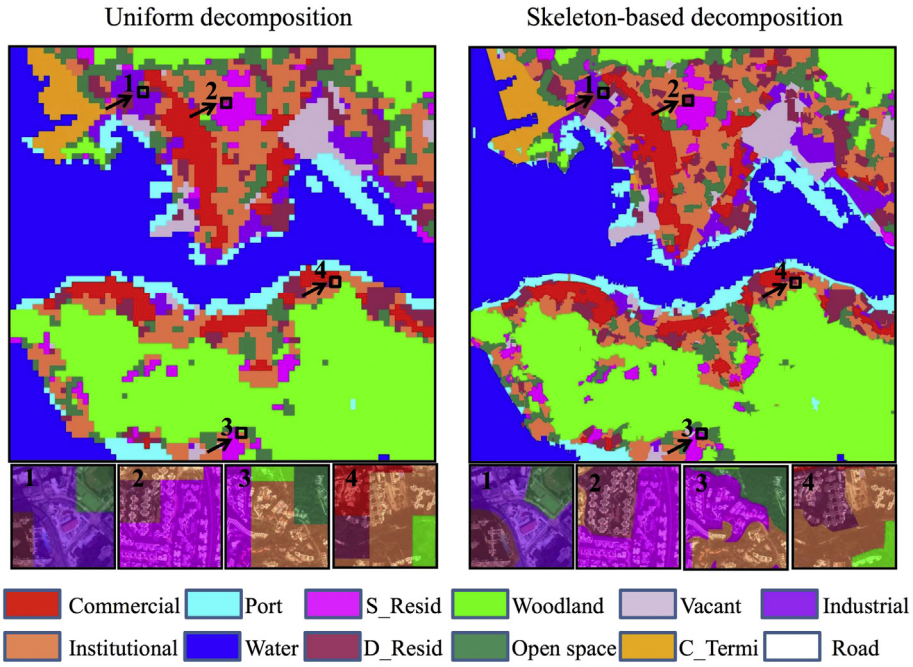

Fig. 10. STDCNN land-use maps with different decomposition methods before the street block restriction, from the WorldView-3 image of Kowloon and Hong Kong Island.

street blocks can consist of multiple parts of processing units with different land-use classes. Therefore, each component of the final confidential vector of the street map contains the fractional information for the corresponding land-use classes. Combining each component of the confidential vectors of all of the street blocks generates a map that displays the spatial arrangement of the land-use class represented by those components. Fig. 13 shows the component maps of all of the land-use classes obtained by the STDCNN land-use mapping based on skeleton-based decomposition after the restriction of street block data. The spatial arrangement of the different land-use classes can be inferred from this figure.

4.3. Discussions

The proposed STDCNN land-use mapping method automatically generates an urban land-use maps based on HSR multispectral remote sensing images and street block data. However, the performance of this proposed mapping method is highly dependent on the quality of the STDCNN land-use classification, which is based solely on the HSR remote sensing image without any other data sources. In some well-designed and developed urban areas, this method can rapidly obtain an accurate land-use map. However, in some areas, land-use classes such as the commercial, residential, and industrial areas are not recognizable

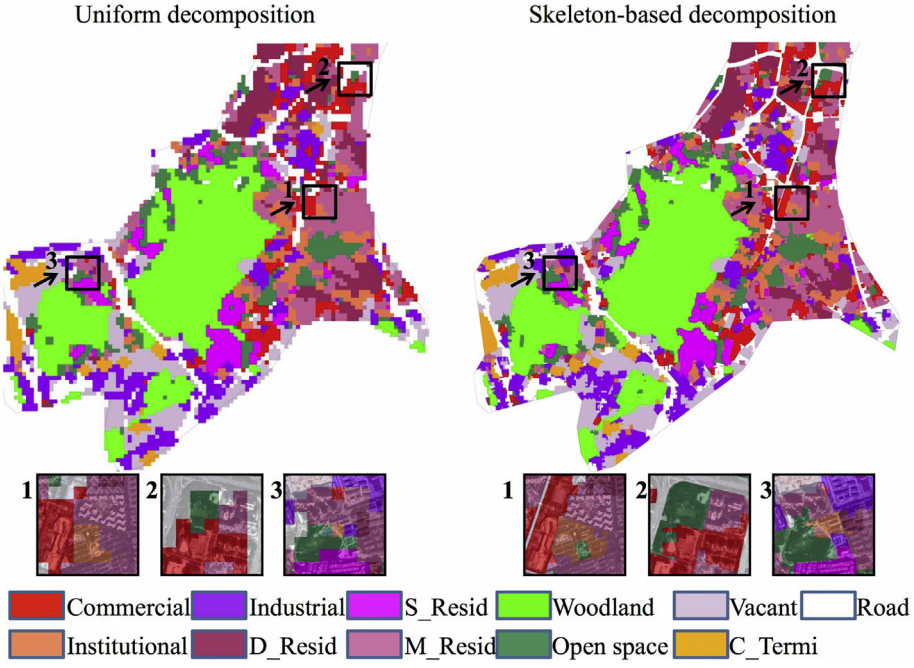

Fig. 11. STDCNN land-use maps with different decomposition methods before the street block restriction, from the WorldView-2 image of the Nanshan district of Shenzhen.

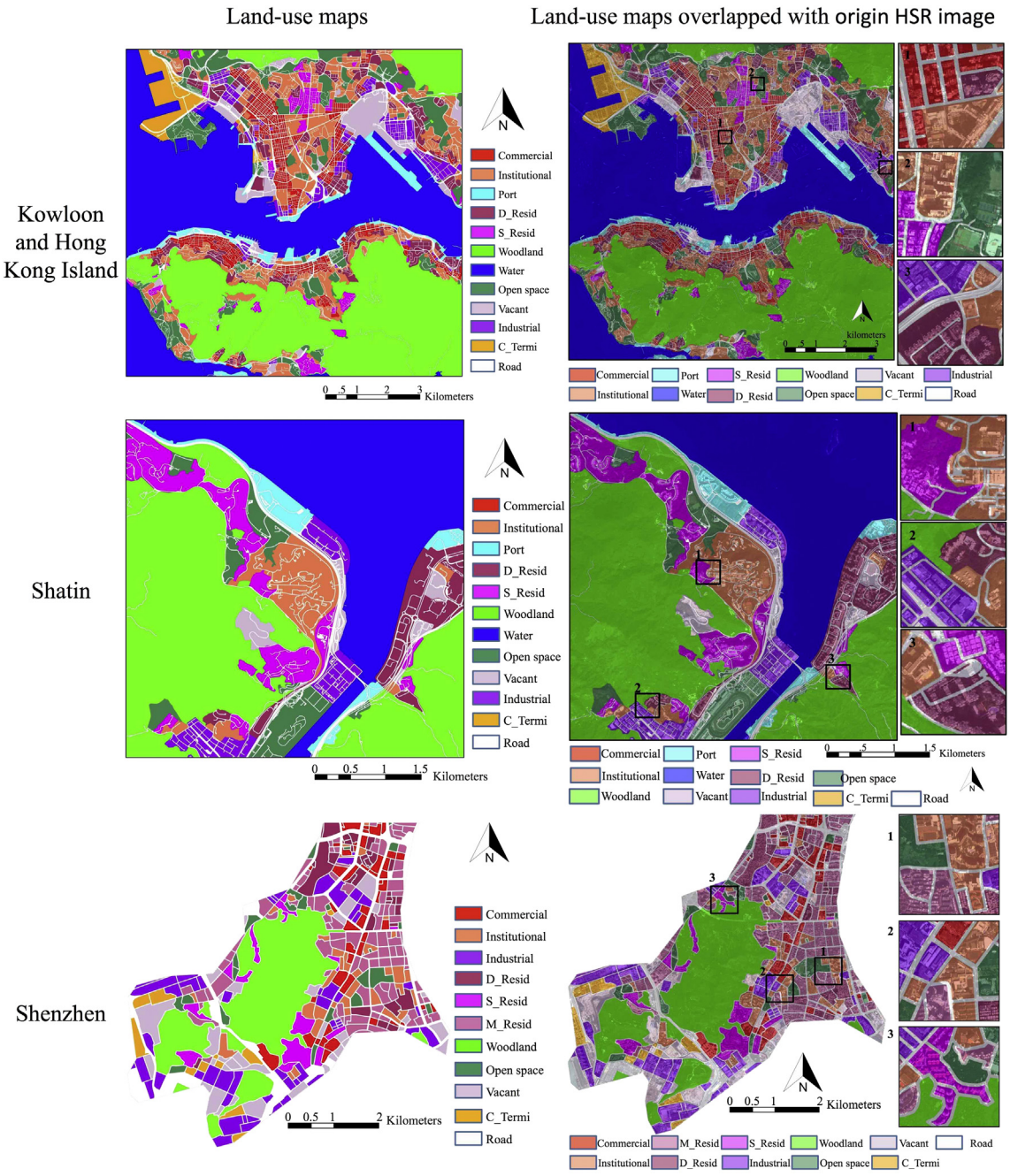

Fig. 12. Land-use maps obtained by STDCNN based on the skeleton-based decomposition method and street block restriction for Kowloon and Hong Kong Island, Shatin, and Shenzhen. The new land-use maps overlapped with the corresponding original HSI images are also displayed.

in airborne or satellite images, as these land-use types may have similar textures, structures, shapes, or other features. For these complex situations, the land-use mapping technique should be combined with other types of data, such as the height of buildings, and other GIS data to generate a more reliable map. Moreover, as the land-use of urban areas is highly related to socioeconomic factors, human activities have a large effect on the types of land-use. Therefore, some data on the distribution of human activity can also be included to produce a more accurate land-use map for practical application.

5. Conclusions

This study presents a new STDCNN method for the land-use classification of high spatial resolution (HSR) multispectral remote sensing images. One part of the STDCNN (the transfer DCNN) is transferred from AlexNet and used to deepen the structure of the STDCNN. Another part of the STDCNN is a small DCNN with a limited number of layers, which is designed to analyze multispectral images with more than three channels. These two parts of the STDCNN are then fused into a fully connected layer that combines the advantages of the transfer DCNN and the small DCNN. This new method overcomes the weaknesses of the traditional DCNN (i.e. it cannot deal with HSR multispectral images with more than three channels) and of the traditional small DCNN (i.e. it does not involve the deep architecture of the DCNN). The experimental results from an analysis of WorldView-3 images of Hong Kong and a WorldView-2 image of Shenzhen indicate that the proposed STDCNN land-use classification method can obtain 91.25%/80.00% OA, and 0.903/0.780 Kappa for the two datasets. These results exceed

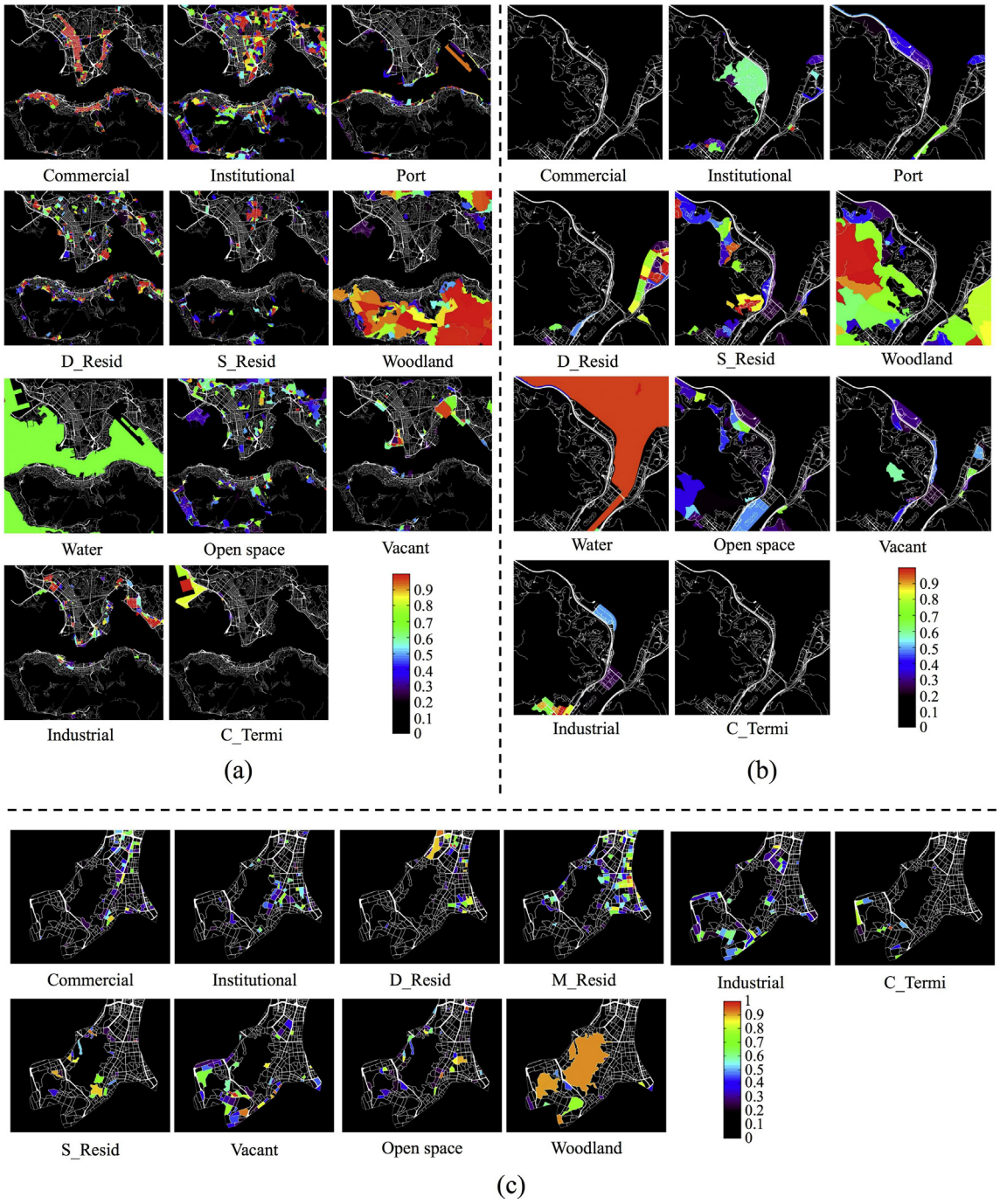

Fig. 13. STDCNN land-use maps of each land-use class based on the skeleton-based decomposition method and street block restriction. (a) Kowloon and Hong Kong Island. (b) Shatin. (c) Shenzhen.

those of methods given in Yang et al. (2009) and Zhao et al. (2013), the transfer DCNN (Zhao et al., 2017), and the small DCNN.

Furthermore, to overcome the weakness of the uniform decomposition method used in previous DCNN methods, which commonly split the land-use pattern into pieces, a new skeleton-based decomposition method that incorporates street block data. The new method adaptively splits large images into small processing units in a way that maintains the integrity of the land-use patterns. The experimental results from the analysis of a large WorldView-3 image of 143 km^2 of Hong Kong and a large WorldView-2 image of 25 km^2 of Shenzhen show that the proposed skeleton-based decomposition method produces better land-use maps than those produced by the uniform decomposition method. Visual comparisons show that the proposed STDCNN land-use mapping method can obtain a satisfactory and practically useful land-use map. Furthermore, the proposed deep learning approach can be easily generalized to an automatic program.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No. 2017YFB0503605, and in part by the Hong Kong Research Grants Council under GRF No. 14606315. We also thank Prof. Bing Xu orcid = "0000-0001-9159-2512" and Dr. Bin Chen at Beijing Normal University for sharing their data with us.

References

Aksoy, S., Koperski, K., Tusk, C., Marchisio, G., Tilton, J.C., 2005. Learning Bayesian classifiers for scene classification with a visual grammar. IEEE Trans. Geosci. Remote Sens. 43 (3), 581–589.

Bahmanyar, R., Shiyong, C., Datcu, M., 2015. A comparative study of bag-of-words and bag-of-topics models of EO image patches. IEEE Geosci. Remote Sens. Lett. 12 (6), 1357–1361.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87 (0), 180–191.

Bosch, A., Zisserman, A., Muñoz, X., 2008. Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal. Mach. Intell. 30 (4), 712–727.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of International Conference on Computational Statistics. Physica-Verlag HD, Heidelberg, pp. 177–186.

Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L., 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. arXiv preprint arXiv:1508.00092.

Chen, S., Tian, Y., 2015. Pyramid of spatial relations for scene-level land use classification. IEEE Trans. Geosci. Remote Sens. 53 (4), 1947–1957.

Chen, X.-L., Zhao, H.-M., Li, P.-X., Yin, Z.-Y., 2006. Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. Remote Sens. Environ. 104 (2), 133–146.

Cheng, G., Han, J., Guo, L., Liu, Z., Bu, S., Ren, J., 2015. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Trans. Geosci. Remote Sens. 53 (8), 4238–4249.

Cheriyadat, A.M., 2014. Unsupervised feature learning for aerial scene classification. IEEE Trans. Geosci. Remote Sens. 52 (1), 439–451.

Fujita, M., 1989. Urban Economic Theory: Land Use and City Size. Cambridge University Press, New York.

Geurs, K.T., Van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. J. Transp. Geogr. 12 (2), 127–140.

Haralick, R.M., Shapiro, L.G., 1992. Computer and Robot Vision. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Hu, F., Xia, G., Wang, Z., Huang, X., Zhang, L., Sun, H., 2015a. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 8 (5), 2015–2030.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015b. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7 (11), 14680.

Hu, S., Wang, L., 2013. Automated urban land-use classification with remote sensing. Int. J. Remote Sens. 34 (3), 790–803.

Huang, X., Liu, H., Zhang, L., 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 53 (7), 3639–3657.

Huang, Y., Wu, Z., Wang, L., Tan, T., 2013. Feature coding in image classification: a comprehensive study. IEEE Trans. Pattern Anal. Mach. Intell. 36 (3), 493–506.

Jia, S., Liu, H., Sun, F., 2015. Aerial scene classification with convolutional neural networks. In: Advances in Neural Networks. vol. 9377. Springer International Publishing, pp. 258–265.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of ACM International Conference on Multimedia. ACM, pp. 675–678.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lienou, M., Maitre, H., Datcu, M., 2010. Semantic annotation of satellite images using latent Dirichlet allocation. IEEE Geosci. Remote Sens. Lett. 7 (1), 28–32.

Luo, W., Li, H., Liu, G., Zeng, L., 2014. Semantic annotation of satellite images using author-genre-topic model. IEEE Trans. Geosci. Remote Sens. 52 (1), 1356–1368.

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using imagenet pretrained networks. IEEE Geosci. Remote Sens. Lett. 13 (1), 105–109.

Penatti, O.A., Nogueira, K., dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 44–51.

Perronnin, F., SÃąnchez, J., Mensink, T., 2010. Improving the Fisher kernel for large-scale image classification. In: European Conference on Computer Vision (ECCV). vol. 6314. Springer, Berlin Heidelberg, pp. 143–156.

Quelhas, P., Monay, F., Odobez, J., Gatica-Perez, D., Tuytelaars, T., 2007. A thousand words in a scene. IEEE Trans. Pattern Anal. Mach. Intell. 29 (9), 1575–1589.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going Deeper with Convolutions. arXiv preprint arXiv:1409.4842.

van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M., 2010. Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. 32 (7), 1271–1283.

Voltersen, M., Berger, C., Hese, S., Schmullius, C., 2014. Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level. Remote Sens. Environ. 154, 192–201.

Walde, I., Hese, S., Berger, C., Schmullius, C., 2014. From land cover-graphs to urban structure types. Int. J. Geogr. Inf. Sci. 28 (3), 584–609.

Wu, C., Zhang, L., Du, B., 2017. Kernel slow feature analysis for scene change detection. IEEE Trans. Geosci. Remote Sens. 55 (4), 2367–2384.

Wu, S.-S., Qiu, X., Usery, E.L., Wang, L., 2009. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. Ann. Assoc. Am. Geogr. 99 (1), 76–98.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1794–1801.

Yang, Y., Newsam, S., 2011. Spatial pyramid co-occurrence for image classification. In: IEEE International Conference on Computer Vision (ICCV), pp. 1465–1472.

Zhang, F., Du, B., Zhang, L., 2014. Saliency-guided unsupervised feature learning for scene classification. IEEE Trans. Geosci. Remote Sens. 53 (4), 2175–2184.

Zhang, F., Du, B., Zhang, L., 2016. Scene classification via a gradient boosting random convolutional network framework. IEEE Trans. Geosci. Remote Sens. 54 (3), 1793–1802.

Zhang, L., Du, B., 2016. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4 (2), 22–40.

Zhang, X., Du, S., 2015. A linear Dirichlet mixture model for decomposing scenes: application to analyzing urban functional zonings. Remote Sens. Environ. 169, 37–49.

Zhao, B., Huang, B., Zhong, Y., 2017. Transfer learning with fully pretrained deep convolution networks for land-use classification. IEEE Geosci. Remote Sens. Lett. 17 (9), 1436–1440.

Zhao, B., Zhong, Y., Xia, G.S., Zhang, L., 2016a. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 54 (4), 2108–2123.

Zhao, B., Zhong, Y., Zhang, L., 2013. Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. Remote Sens. Lett. 4 (12), 1204–1213.

Zhao, B., Zhong, Y., Zhang, L., 2016b. A spectral-structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 116, 73–85.

Zhao, B., Zhong, Y., Zhang, L., Huang, B., 2016c. The Fisher kernel coding framework for high spatial resolution scene classification. Remote Sens. 8 (2), 157.

Zhao, L., Tang, P., Huo, L., 2014. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 7 (12), 4620–4631.

Zheng, X., Sun, X., Fu, K., Wang, H., 2013. Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint. IEEE Geosci. Remote Sens. Lett. 10 (4), 652–656.

Zhong, Y., Zhu, Q., Zhang, L., 2015. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 53 (11), 6207–6222.