



# Extracting Urban Impervious Surface from WorldView-2 and Airborne LiDAR Data Using 3D Convolutional Neural Networks

Zhongchang Sun<sup>1,2</sup> · Xiangwei Zhao<sup>3</sup>  · Mengfan Wu<sup>4</sup> · Cuizhen Wang<sup>5</sup>

Received: 19 July 2018 / Accepted: 15 November 2018 / Published online: 7 December 2018  
© The Author(s) 2018

## Abstract

The urban impervious surface has been recognized as a key quantifiable indicator in assessing urbanization and its environmental impacts. Adopting deep learning technologies, this study proposes an approach of three-dimensional convolutional neural networks (3D CNNs) to extract impervious surfaces from the WorldView-2 and airborne LiDAR datasets. The influences of different 3D CNN parameters on impervious surface extraction are evaluated. In an effort to reduce the limitations from single sensor data, this study also explores the synergistic use of multi-source remote sensing datasets for delineating urban impervious surfaces. Results indicate that our proposed 3D CNN approach has a great potential and better performance on impervious surface extraction, with an overall accuracy higher than 93.00% and the overall kappa value above 0.89. Compared with the commonly applied pixel-based support vector machine classifier, our proposed 3D CNN approach takes advantage not only of the pixel-level spatial and spectral information, but also of texture and feature maps through multi-scale convolutional processes, which enhance the extraction of impervious surfaces. While image analysis is facing large challenges in a rapidly developing big data era, our proposed 3D CNNs will become an effective approach for improved urban impervious surface extraction.

**Keywords** WorldView-2 · Airborne light detection and ranging (LiDAR) · Impervious surface · Convolutional neural networks (CNNs) · Support vector machine (SVM)

**JEL Classification** 42020

## Introduction

Impervious surfaces are usually defined as the entirety of impermeable surfaces such as roads, buildings, parking lots, and other urban infrastructures, where water cannot infiltrate through the ground (Sun et al. 2011). Urbanization results in the increase in impervious surfaces, which in turn casts great impacts on urban environmental problems such as increased urban heat islands (Ma et al. 2016), surface runoff (Sun et al. 2014), water contamination (Kim et al. 2016), and air pollution (Touchaei et al. 2016). Facing rapid urbanization all over the world, these environmental concerns have triggered a surge of interest in impervious surface studies.

Satellite imagery with various spatial and temporal resolutions has been widely employed to map impervious surfaces. Medium- and low-spatial-resolution images, including Landsat, MODIS data, have rich spectral information and high temporal resolution, which is suitable for

---

✉ Xiangwei Zhao  
zhaowchina@sdust.edu.cn; tlzwx1696@163.com

<sup>1</sup> Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China  
<sup>2</sup> Key Laboratory Earth Observation Hainan Province, Sanya Institute of Remote Sensing, Sanya 572029, Hainan, China  
<sup>3</sup> Shandong Province “3S” Engineering Research Center, Shandong University of Science and Technology, Qingdao 266590, China  
<sup>4</sup> Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China  
<sup>5</sup> Department of Geography, University of South Carolina, Columbia, SC 29208, USA

large-scale impervious surface mapping (Xu et al. 2018a, b; Zhang et al. 2018). However, challenges for mixed pixels confuse us to extract impervious surfaces in the large-area mapping. High-spatial-resolution images produce detailed land-cover and land-use information, but the spectral similarity of different objects and shadows of tall buildings or large trees limit the impervious surface extraction (Guo et al. 2014). Hyper-spectral images solve the problem of spectral similarity of different objects and spectral heterogeneity of the same land class, but small map coverage and high price obstruct the application of hyper-spectral images (Weber et al. 2018). SAR image extracts land information free from cloud, and SAR image is helpful for extracting impervious surfaces under the large tree crowns (Guo et al. 2014). However, the coherent noise of SAR image is a significant problem for impervious surface extraction. Therefore, the single-source imagery has various restrictions on urban impervious surface mapping. More recently, integration of multiple datasets from different image acquisition mechanisms has been considered promising to address these uncertainties (Xu et al. 2017), including medium–low-spatial-resolution images and high-spatial-resolution images (Parece and Campbell 2013a, b), optical images and SAR images (Zhang et al. 2014), high-spatial-resolution images and light detection and ranging (LiDAR) data (Im et al. 2012). In particular, LiDAR data can improve impervious surface extraction by providing the height information that significantly distinguishes between objects with similar spectral characteristics (Im et al. 2012). For example, although the buildings, roads, and bare soil usually have similar spectral features, the height difference is large. Therefore, LiDAR height information is helpful for dealing with the different objects with similar spectrum. What is more, the roof of buildings is flatter than the tree crowns. The LiDAR height variance is helpful for distinguishing buildings and trees.

Various approaches have been developed to extract and quantify urban impervious surfaces from satellite images (Sun et al. 2017). Index analysis approach is applied usually on large-scale mapping and estimation of impervious surfaces. Different indices demand different image pre-processing, bands, and time. Index analysis is sensitive to thresholds (Sun et al. 2017). Regression model is also applied to large-scale estimation of impervious surfaces from different remote sensing images (Akpona et al. 2018). Spectral mixture analysis method certainly solves the problem of mixed pixels. However, it is often used for medium–low-spatial-resolution images (Wang et al. 2018). Decision tree method is also generally applied to impervious surface extraction, including CART algorithm and rule-based decision method (Xu 2013). Decision tree method is a weak learning method and it is sensitive to the noise of data and the error of training samples. Moreover,

the rule-based method is uncertain due to different indexical thresholds. Classification method from machine learning has a wide application on impervious surface extraction, including artificial neural network (Hu and Weng 2009), support vector machine (Sun et al. 2011), and random forest (Xu et al. 2018a, b). These machine learning approaches have a good performance. However, these swallow learning methods train from raw images and manually extracted features, which consist of some superfluous and useless information. More recently, deep learning (Hinton and Salakhutdinov 2006; LeCun et al. 2015) has become a hot topic in many research areas, including urban remote sensing. As a typical deep learning model, the convolutional neural networks (CNNs) employ a set of trainable filters and local neighbourhood pooling operations on raw images, resulting in a hierarchy of increasingly complex features (Ji et al. 2013). Compared with other neural networks, the CNNs take advantage of weight sharing and local connections, which help to reduce the total number of trainable parameters and leads to more efficient training and more effective recognition of patterns (Chen et al. 2016a). Another benefit of the CNNs is the use of pooling, which results in slightly translational and rotational invariant features, a desirable property for natural signals (Långkvist et al. 2016). Therefore, the CNNs have demonstrated excellent performance in image classification (Maggiori et al. 2016; Marmanis et al. 2016; Chen et al. 2016b; Scott et al. 2017; Kussul et al. 2017) and target recognition (Sevo and Avramovic 2016; Ding et al. 2016; Cheng et al. 2016; Zhang et al. 2017). However, most of this research used the one-dimensional (1D) or two-dimensional (2D) CNNs to conduct image classification. Advantageously, the three-dimensional (3D) CNNs, due to their 3D convolutional operation, have the ability to simultaneously model spatial, textural, spectral, and other information. Presently, 3D CNNs are beginning to be applied for videos or volumetric images (Ji et al. 2013; Chen et al. 2016b; Tran et al. 2015), and their performance has not been illustrated in extracting urban impervious surfaces from satellite imagery.

The objective of this study is first to explore the potential of the 3D CNNs in extracting urban impervious surface from high-resolution (HR) imagery. To test its effectiveness, its results are compared with the classification outputs from the commonly applied support vector machine (SVM) method and 2D CNN methods. We also illustrate the influence of different parameters of the proposed 3D CNN model on impervious surface extraction. Two HR images are utilized to examine the effects of single-source (WorldView-2, WV-2) and multi-source (WV-2 and LiDAR) datasets in this study.

## Methodology

### 3D Convolutional Neural Networks

The CNNs are multi-stage feed-forward neural networks that hold the state-of-the-art performance in remote sensing. Although the CNNs were proposed many years ago, only in recent years it has become possible to train and implement large CNNs in remote sensing with computational progress such as high-performance GPUs, rectified linear units (ReLU) to improve much faster training, and dropout or data augmentation techniques to reduce overfitting (Chen et al. 2016a). In this study, we propose to perform 3D convolutions using a 3D kernel in the CNNs to compute various features from HR WV-2 and LiDAR data.

Typically, the 3D CNN layers consist of convolutional, nonlinear, and pooling operators. The input and output of each layer are called feature maps. Given an input image  $x \in R^{m \times m \times c \times d}$  having number of lines ( $m$ ), columns ( $m$ ), channels ( $c$ ), and depth ( $d$ ), the output image  $y \in R^{m' \times m' \times c' \times t}$  is composed of a number of lines ( $m'$ ), columns ( $m'$ ), output features ( $t$ ), and depth ( $c'$ ). The convolution of the input  $x$  with a 3D kernel  $w \in R^{n \times n \times c \times t}$  is calculated as (Ji et al. 2013):

$$y_{i'j'r'}^{lh} = f \left( b_{lh} + \sum_g \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^c W_{ijr}^{lhg} \times X_{S_n(i'+i-P_m, S_n(j'+j-P_m, r))}^{g(l-1)} \right) \tag{1}$$

where  $y_{i'j'r'}^{lh}$  is the neuron at position  $(i'j'r')$  of the  $h$ th feature map in the  $l$ th layer;  $b_{lh}$  is the bias for the  $h$ th feature map in the  $l$ th layer;  $n$  expresses the height and width of the spatial convolution kernel;  $c$  denotes the size of the 3D kernel along the spectral and elevation dimension (same as the number of channels of input image);  $g$  is the feature map in the  $(l-1)$ th layer connected to the current ( $h$ th) feature map;  $W_{ijr}^{lhg}$  represents the weight value of the position  $(i, j, r)$  connected to the  $g$ th feature map;  $P_m$  denotes padding of the input image;  $S_n$  is the sub-sampling stride of the output image;  $X_{S_n(i'+i-P_m, S_n(j'+j-P_m, r))}^{g(l-1)}$  denotes the neuron connected to the neuron of  $y_{i'j'r'}^{lh}$  in the  $(l-1)$  layer; and  $f(\cdot)$  is a nonlinear activation function. Generally, the typical activation function for CNNs is hyperbolic tangent or rectified linear units (ReLU). Here, the ReLU activation function (Eq. 2) (Ding et al. 2016) is used, which can reach CNNs' best performance with precisely supervised training on large labelled datasets.

$$y_{i,j,r} = \max(0, x_{i,j,r}) \tag{2}$$

The pooling layer decreases the resolution of the output features to make them less sensitive to the input shift and

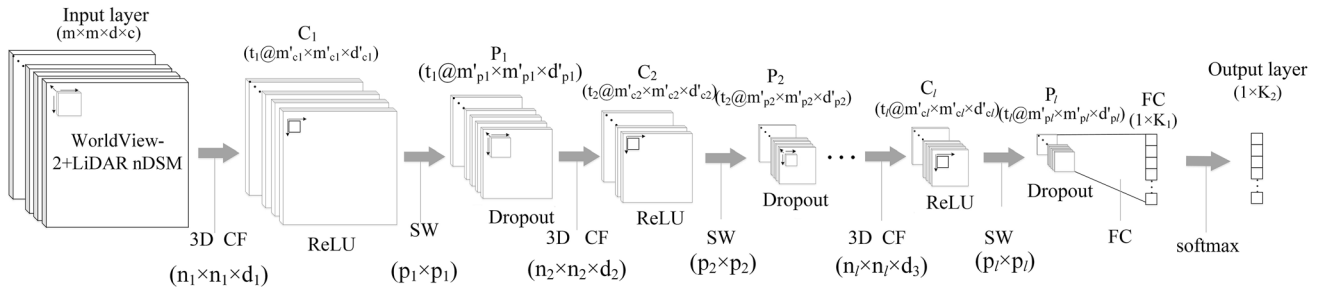
distortions. We adopt the max-pooling layer method in this study. It computes the maximum response of each image channel in a  $p \times p$  sub-window, which acts as a sub-sampling operation. The max-pooling layer can be written as (Vedaldi and Lenc 2015):

$$y_{i',j',r} = \max_{1 \leq i \leq p, 1 \leq j \leq p} (x_{S_n(i'+i-P_m, S_n(j'+j-P_m, r))} \tag{3}$$

### Architecture of 3D CNNs for Image Classification

The architecture of 3D CNNs for extracting impervious surfaces is depicted in Fig. 1. It is composed of one input layer, multiple convolutional layers, multi-max-pooling layers, and one softmax-loss layer (output layer). The size of input layer is  $m \times m \times c \times d$ , where  $c = 10$  represents the number of input image channels (including eight WV-2-fused multi-spectral bands, one panchromatic WV-2 band, and the LiDAR-derived nDSM band). The number of CNN convolutional layers  $L$  is designed for extracting features from the image. For each CNN layer, the size of convolutional kernel is  $n_l \times n_l \times d_l$  ( $l = 2, 3, 4, \dots, L$ ),  $d_l$  denotes the depth ( $d$ ) of convolutional kernel in the  $l$ th layer, and the size of the sub-window in the max-pooling layer is  $p_l \times p_l$  ( $l = 2, 3, 4, \dots, L$ ).  $t_l$  is the number of output feature maps in the  $l$ th layer. The optimal selection of CNN parameters ( $m, n, p, L, t$ ) will be discussed in Sect. 2.3. In our study, the ReLU are used to reduce training time in each convolutional layer, and the recently proposed dropout is employed for reducing the overfitting in the network training after each max-pooling layer (Chen et al. 2016a). The dropout rate is 0.5. During the architecture construction of the 3D CNNs, the padding parameter  $P_m$ , and the stride parameter  $S_n$  of the steps are set as:  $P_m = 0$  and  $S_n = 1$  in the convolutional layer, and  $P_m = 0$  and  $S_n = 2$  in the sub-sample layer. In addition, Vedaldi and Lenc (2015) suggested that the optimizer is Adam, the value of the learning rate is set to 0.0001, the number of training epochs is set to 100, and the mini-batch of input images for each epoch is set to 8. After the CNN convolution, a full-connection (FC) layer is built, which is similar to the convolutional layer. However, the size of output map in the FC layer must be  $1 \times K$ , where  $K$  of the output layer is the number of land-cover classes.

The size of the output layer corresponds to the total number of land-cover classes in our study area: buildings, roads/other impervious surfaces, trees, grasslands, and bare soils. The softmax nonlinearity is employed to conduct multi-class logistic regression for the output layer. Its output is a  $K$ -dimensional vector, in which each element corresponds to the probability of each class. Within a mini-batch ( $B$ ) of input samples  $\{(x^{(1)}, y^{(1)}), (x^{(1)}, y^{(1)}), \dots, (x^{(B)}, y^{(B)})\}$ , for an input sample  $(x^{(b)}, y^{(b)})$ ,  $b \in$



**Fig. 1** Architecture of 3D CNNs with input features of elevation, spectral, and spatial information for impervious surface extraction (CF—convolution filter; SW—sub-window; and FC—full connection)

$\{1, 2, \dots, B\}$ , its label is  $y^{(b)} \in \{1, 2, \dots, K\}$ . The probability of the  $k$ th ( $k \in \{1, 2, \dots, K\}$ ) class is estimated as follows (Chen et al. 2016b):

$$P(y^{(b)} = k | x^{(b)}; \theta) = \frac{\exp(\theta_k x^{(b)})}{\sum_{i=1}^K \exp(\theta_i x^{(b)})}. \tag{4}$$

$\theta$  is the model parameter that needs to be adjusted by constructing the softmax-loss function  $\mathcal{L}(\theta)$ .  $\mathcal{L}(\theta)$  is used to compare a prediction  $P(y^{(b)} = k | x^{(b)}; \theta)$  with a ground-truth class label  $k$ . The focal loss function is selected for reacting to the imbalance of training samples of each class. Focal loss is an effective development from cross-entropy loss. The focal loss adds the weight of each class to reduce the weight of easy-classified samples and major samples and raise the weight of hard-classified samples and minor samples. Easy-classified samples have larger  $P(y^{(b)} = k | x^{(b)}; \theta)$ , and hard-classified samples have smaller  $P(y^{(b)} = k | x^{(b)}; \theta)$ . The weight is  $(1 - \alpha^{(b)})(1 - P(y^{(b)} = k | x^{(b)}; \theta))^\gamma$ . The  $1 - \alpha^{(b)}$  is a coefficient for adjusting the weight of imbalance samples. For more samples of class  $b$ ,  $1 - \alpha^{(b)}$  is smaller and the  $\alpha^{(b)}$  is larger. In our study, designed  $\alpha^{(b)}$  is equal to the ratio between the number of samples of class  $b$  and the sum of all samples. In our study,  $\alpha^{(0)} = 0.56, \alpha^{(1)} = 0.38, \alpha^{(2)} = 0.06$ .  $\gamma$  is used for adjusting the weight of easy-classified and hard-classified samples. With the increase in  $\gamma$ , the weight of easy-classified samples is smaller than the weight of hard-classified samples. The formula of focal loss is computed as follows (Chen et al. 2016; Lin et al. 2017):

$$\mathcal{L}(\theta) = -\frac{1}{B} \left(1 - \alpha^{(b)}\right) \left(1 - P\left(y^{(b)} = k | x^{(b)}; \theta\right)\right)^\gamma \times \left[ \sum_{b=1}^B \sum_{k=1}^K 1\{y^{(b)} = k\} \log P\left(y^{(b)} = k | x^{(b)}; \theta\right) \right]. \tag{5}$$

### Optimal Selection of the 3D CNN Hyper-Parameters

The initial parameters of the 3D CNN model have an important influence on information extraction and classification results. Therefore, the optimal selection of CNN hyper-parameters is a key step for generating better training curacy of the CNN model. In this study, the selection of hyper-parameters comes from the training and validation samples. The five hyper-parameters, including the size of input image  $m$ , convolution kernel size  $n$ , pooling dimension  $p$ , the number of feature maps  $t$ , and the number of CNN layers  $L$ , are evaluated to assess their influences on classification accuracy. The optimal hyper-parameters are further used to construct the optimal 3D CNN model.

First, given the input image layers  $m \times m \times c \times d$  ( $c = 10$ ), when  $m = 25, L = 1$ , and  $t = 50$ , we examine the accuracies of the 3D CNNs using different convolutional kernel sizes  $n = [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]$  and pooling dimensions  $p = [2, 4, 6, 8, 10, 12]$ . By way of epoch, the optimal combination of  $n$  and  $p$  parameters is identified. Second, based on the optimal parameters  $n$  and  $p$  and a predetermined parameter  $L = 1$ , we further examine the accuracies using different sizes of input images  $m = [5, 15, 25, 35, 45]$  and different numbers of output features  $t = [50, 100, 150, 200]$ . The optimal parameters  $m$  and  $t$  are

**Table 1** Training, validation, and test sample pixels of each class in our study area

Classes	Samples		
	Training	Validation	Test
Impervious surfaces	2348	200	500
Vegetation	1517	200	500
Bare soil	311	200	500
Total	4176	600	1500

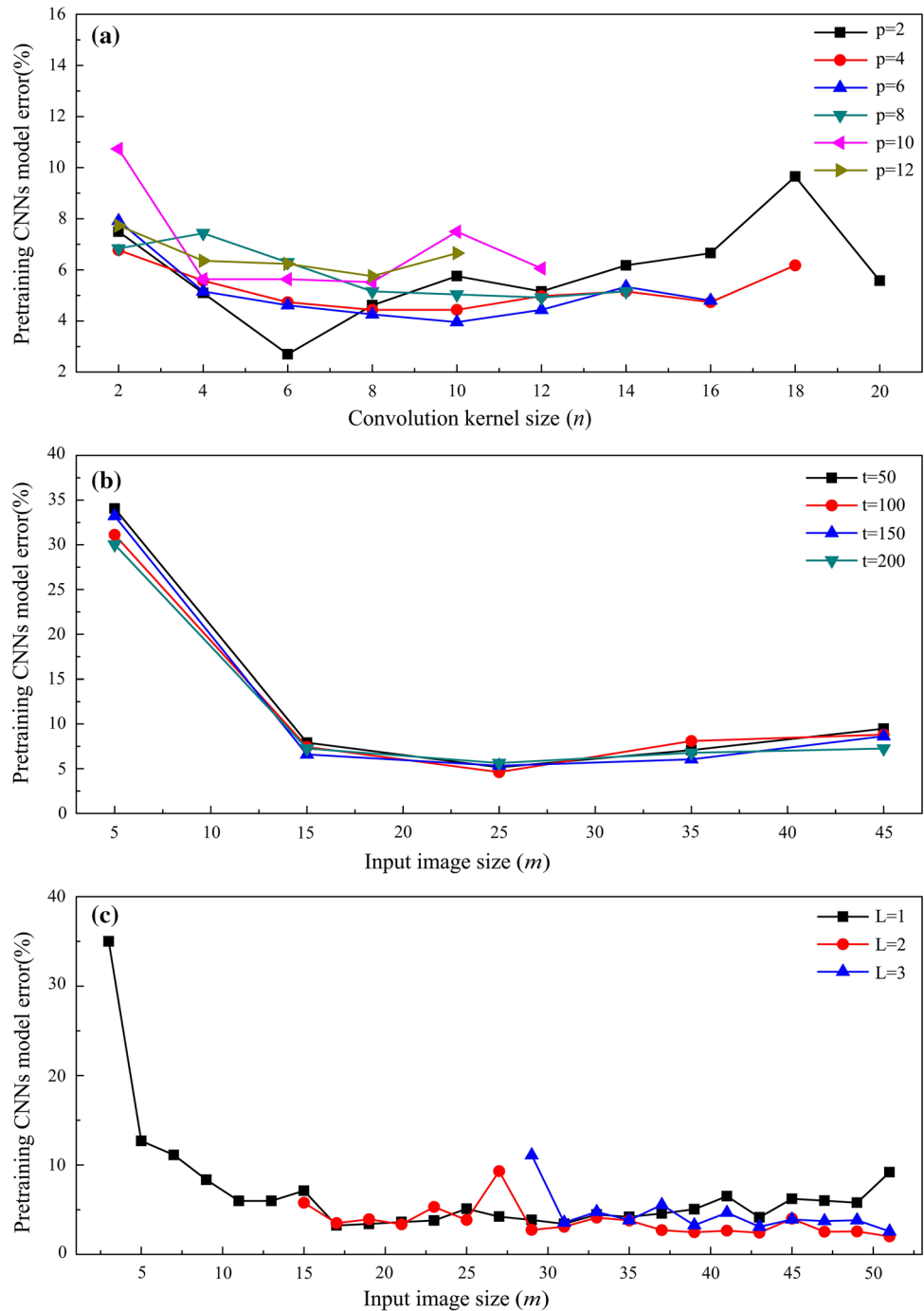
thus iteratively determined. Finally, based on the optimal combination of  $n$  and  $p$ ,  $m$  and  $t$ , we further evaluate the influence of the different sizes of input images  $m = [3, 5, 7, \dots, 47, 49, 51]$  and the numbers of CNN layers  $L = [1, 2, 3]$  on classification accuracy. The optimal parameters  $m$  and  $L$  are iteratively determined.

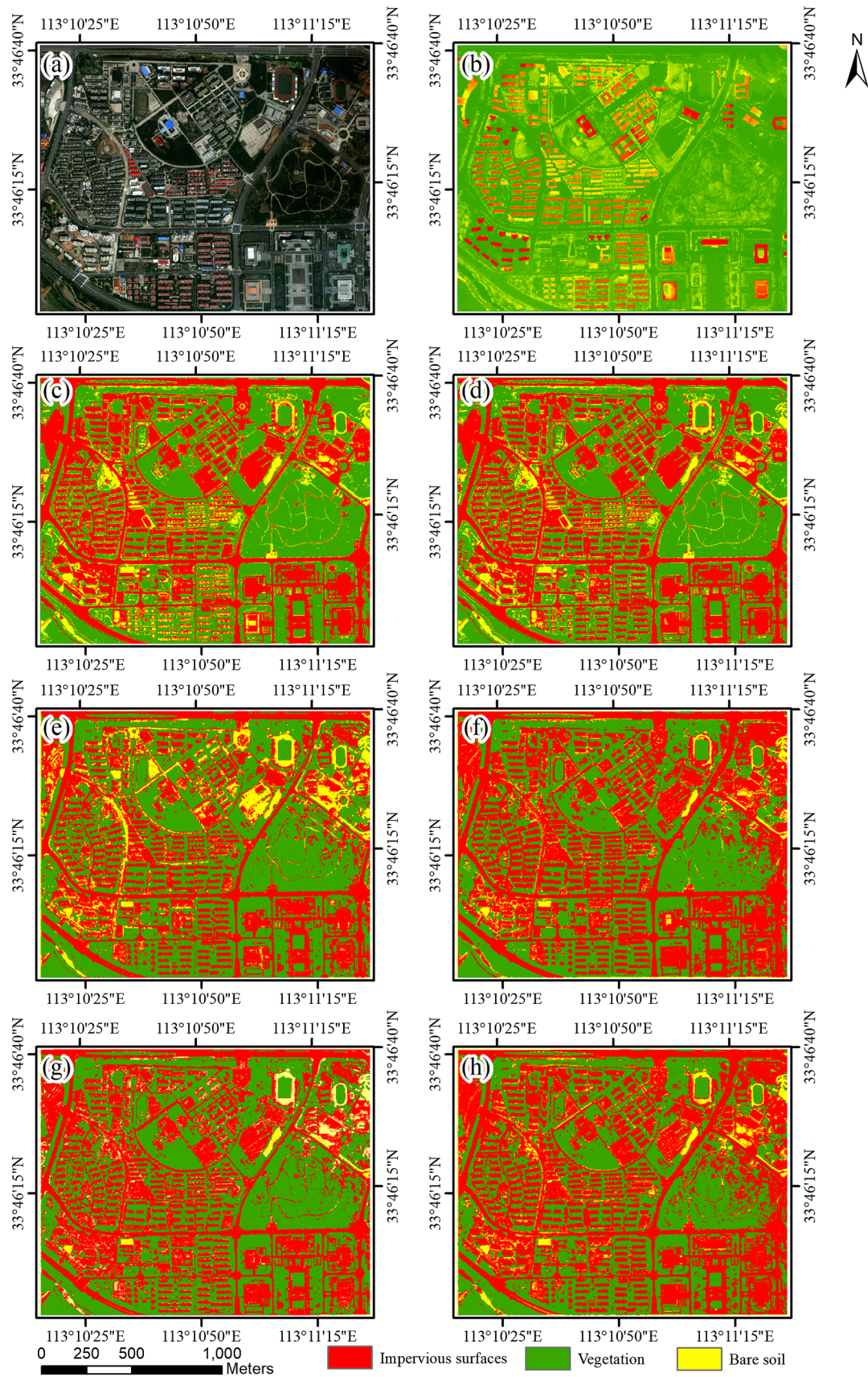
## Experiments and Discussion

### Datasets

The experimental site of this study is located in Pingdingshan City, Henan Province, China. A WV-2 scene was acquired in August 2014, and the airborne LiDAR data

**Fig. 2** Influence of the different 3D CNN model parameters on pre-training CNN model accuracies. **a** Influence of the parameters  $n$  and  $p$ ; **b** performance of the parameters  $m$  and  $t$ ; **c** performance of the parameters  $m$  and  $L$ . (The unit of image input size is pixel)





◀ **Fig. 3** WV-2 image (a) and airborne LiDAR nDSM (b); extracted ISA distributions from WV-2 only using SVM (c), 2D CNNs (e) and 3D CNNs (g); and from WV-2 + LiDAR using SVM (d), 2D CNNs (f) and 3D CNNs (h)

were taken in August 2013. The 11-bit WV-2 image consists of one panchromatic (PAN) band with 0.5 m pixel size and eight multi-spectral (MS) bands (Costal Blue,

Blue, Yellow, Red, Red Edge, NIR1, NIR2) across spectral regions from 400 to 1040 nm with 2 m spatial resolution. The WV-2 MS bands were pan-sharpened with the PAN band to reach the 0.5 m pixel size using the haze- and ratio-based (HR) fusion scheme. The airborne LiDAR datasets were provided in ASCII format, including the X, Y, and Z coordinates of first return points and their intensities. The average point cloud densities were 23.52 points  $m^{-2}$ . The

**Table 2** Error matrix and accuracy comparison between the SVM, 2D CNN and 3D CNN classifications with WV-2 only and WV-2 + LiDAR

SVM WV-2 only		Classified classes				PA (%)
		Impervious surfaces	Vegetation	Bare soil	Total	
Reference classes	Impervious surfaces	835	18	147	1000	83.50
	Vegetation	33	921	46	1000	92.10
	Bare soil	29	0	471	500	94.20
	Total	897	939	664	2500	
UA (%)		93.09	98.08	70.93		
OA (%)		89.08				
OK (%)		83.28				
SVM WV-2 + LiDAR		Classified classes				PA (%)
		Impervious surfaces	Vegetation	Bare soil	Total	
Reference classes	Impervious surfaces	873	17	110	1000	87.30
	Vegetation	31	923	46	1000	92.30
	Bare soil	10	0	490	500	98.00
	Total	914	940	646	2500	
UA (%)		95.51	98.19	75.85		
OA (%)		91.44				
OK (%)		86.86				
2D CNN WV-2 + LiDAR		Classified classes				PA (%)
		Impervious surfaces	Vegetation	Bare soil	Total	
Reference classes	Impervious surfaces	869	110	21	1000	86.90
	Vegetation	40	953	7	1000	95.30
	Bare soil	21	18	461	500	92.20
	Total	930	1081	489	2500	
UA (%)		93.44	88.16	94.27		
OA (%)		91.32				
OK (%)		86.42				
3D CNN WV-2 only		Classified classes				PA (%)
		Impervious surfaces	Vegetation	Bare soil	Total	
Reference classes	Impervious surfaces	900	65	35	1000	90.00
	Vegetation	36	955	9	1000	95.50
	Bare soil	17	11	472	500	94.40
	Total	953	1031	516	2500	
UA (%)		94.44	92.63	91.47		
OA (%)		93.08				
OK (%)		89.21				

**Table 2** (continued)

3D CNN WV-2 + LiDAR		Classified classes				PA (%)
		Impervious surfaces	Vegetation	Bare soil	Total	
Reference classes	Impervious surfaces	912	76	12	1000	91.20
	Vegetation	50	944	6	1000	94.40
	Bare soil	15	11	474	500	94.80
	Total	977	1031	492		
UA (%)		93.35	91.56	96.34		
OA (%)		93.20				
OK (%)		89.36				
2D CNN WV-2 only		Classified classes				PA (%)
		Impervious surfaces	Vegetation	Bare soil	Total	
Reference classes	Impervious surfaces	773	80	147	1000	77.30
	Vegetation	101	887	12	1000	88.70
	Bare soil	19	13	468	500	93.60
	Total	893	980	627	2500	
UA (%)		86.56	90.51	74.64		
OA (%)		85.12				
OK (%)		77.11				

DSM and the DEM were generated from LiDAR point cloud using ENVI software (version 5.1). Then, the nDSM image with 0.5 m spatial resolution was produced by subtracting the DEM values from DSM. Finally, the nDSM together with eight WV-2-fused multi-spectral bands and one panchromatic WV-2 band is inputted to 3D CNN model.

We set three land classes in the whole study area, including impervious surfaces, vegetation, and bare soil, based on V–I–S model. For 3D CNN classification, selected manually and randomly training, validation, and test sample pixels are listed in Table 1. In our study, the number of samples is equal to that of other papers about the deep learning methods for remote sensing image classification (Yue and Ma 2016; Gao et al. 2018), such as about 2000–7000 training samples of other studies. The training samples are used for the training of 3D CNN model, the validation samples are used for adjusting hyper-parameters, and the test samples are used for final accuracy evaluation.

### Optimal 3D CNN Hyper-Parameters

To identify the optimal 3D CNN parameters for model development, their performances were evaluated by using randomly selected training and validation samples (Fig. 2).

Figure 2a shows the influence of the hyper-parameters  $n$  and  $p$  on the CNN model accuracy in the case of  $m = 25$ ,  $L = 1$ , and  $t = 50$ . It can be observed that: (1) with the increase in parameter  $n$ , the CNN model errors (the ratio of

wrong classification to total classification at the 300th epoch in the test samples) generally decrease and then increase; (2) when  $n = [6, 8, 10]$ , relatively lower errors can be acquired with a different parameter  $p$ ; (3) when  $n < p$ , the model errors are high; when  $n \in [2p, 3p]$ , the errors become the lowest. Therefore, the optimal combination of  $n$  and  $p$  is  $n \in [2p, 3p]$ .

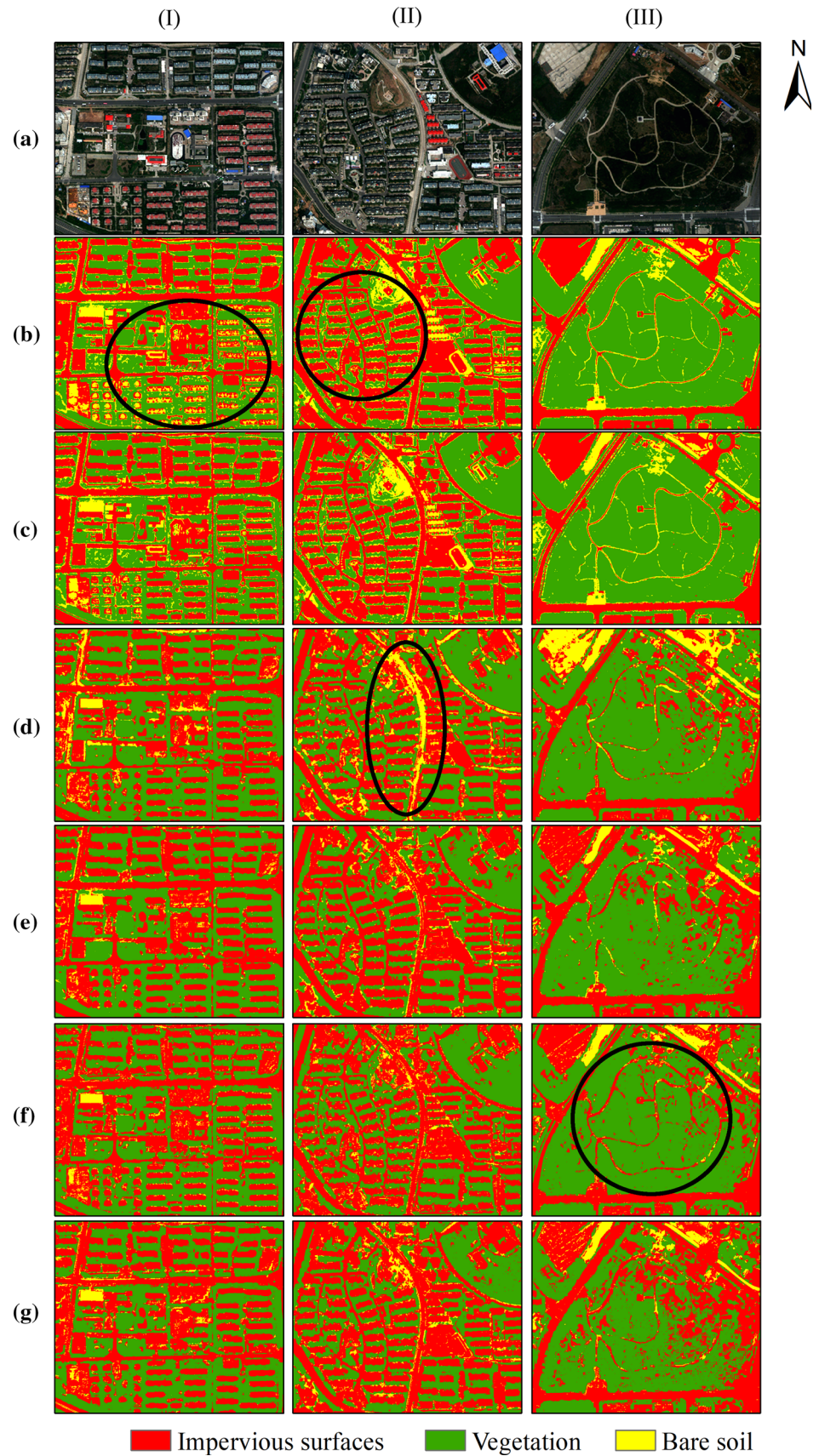
Figure 2b examines the performance of the hyper-parameters  $m$  and  $t$  in the case of  $L = 1$ . When  $m = 5$ , the model errors are the highest; when  $m = [25, 35, 45]$ , the errors are less than 10%. In addition, with the same parameter  $m$ , the number of the output feature  $t$  does not affect the CNN model errors.

Figure 2c examines the hyper-parameters  $m$  and  $L$ . When  $m \leq 15$ , the model errors are high. However, a larger  $m$  requires lower computation time. From Fig. 2b, c, the parameter  $m$  performs better in the range of 20–40. When  $L = 1$ , a larger or smaller  $m$  value results in higher model error. Generally speaking, when  $L = 2$  and  $L = 3$ , the pre-training CNN model errors are fairly stable. The model accuracies with  $L = 2$  are superior to those with  $L = 1$  and  $L = 3$ .

In short, for the 3D CNN model in this study, the optimal size of input image is set to  $m = 37$ ,  $d = 3$ ,  $c = 10$  and the number of layer  $L = 2$ . For the first layer, the size of the 3D convolutional kernel is  $8 \times 8 \times 2$ , and the size of sub-sample kernel is set to  $4 \times 4$ . For the second layer, the sizes of the 3D convolutional kernel and sub-sample kernel are  $3 \times 3 \times 2$  and  $2 \times 2$ , respectively. Additionally, the



**Fig. 4** Impervious surface (IS) distribution of three subsets in the study area. **a** The true colour WV-2 image is displayed; **b** IS results derived from WV-2 only using SVM; **c** IS results derived from WV-2 only using 2D CNNs; **d** IS results derived from WV-2 only using 3D CNNs; **e** IS results derived from WV-2 + LiDAR using SVM; **f** IS results derived from WV-2 + LiDAR using 2D CNNs; and **g** IS results derived from WV-2 + LiDAR using 3D CNNs



numbers of input features are set to 50 and 100, respectively. Similarly, to compare results of 2D CNNs and 3D CNNs, the hyper-parameters of 2D CNNs are equal to those of 3D CNNs, except the convolution kernels have two dimensions. That is, for the first layer, the size of the 2D convolutional kernel is  $8 \times 8$ , and the size of sub-sample kernel is set to  $4 \times 4$ . For the second layer, the sizes of the 2D convolutional kernel and sub-sample kernel are  $3 \times 3$  and  $2 \times 2$ , respectively. Our 3D CNN and 2D CNN implementation is based on the Pytorch 0.4 platform and NVIDIA GTX 1080Ti GPU. The SVM method is used for the comparison, whose hyper-parameters are consistent with the paper (Guo et al. 2014). The penalty coefficient  $C$  is 100, and the gamma is 0.1 for WorldView-2 + LiDAR image (10 bands) and 0.11 for WorldView-2 image (9 bands).

### Urban Impervious Surface Extraction

Figure 3 visually compares the extracted impervious surface area (ISA) with the 3D CNNs, 2D CNN and SVM methods from the WV-2 only and WV-2 + airborne LiDAR, respectively. The error matrix of Accuracy assessment is conducted by using the error matrix approach. The accuracy measures, i.e. producer's accuracy (PA), user's accuracy (UA), overall accuracy (OA), and overall kappa coefficient (OK), are calculated (Table 2).

As shown in Table 2, the 3D CNNs provide significantly higher accuracy for impervious surface extraction than SVM and 2D CNNs. With the WV-2 image only, the OA and OK in 3D CNNs are 93.08% and 0.89 compared with 89.08% and 0.83 in SVM, and with 85.12% and 0.77 in 2D CNNs, respectively. With the WV-2 + LiDAR, 3D CNNs reach an OA of 93.20% and OK of 0.89, while SVM has an OA of 91.44% and OK of 0.87, and 2D CNNs have an OA of 91.32% and OK of 0.86. Even for the WV-2 image only, the 3D CNNs have higher accuracies than 2D CNNs and SVM using WV-2 + LiDAR. When both datasets are considered, the improvements in 2D CNNs and SVM performance are higher than that of 3D CNNs.

The 3D CNN method can automatically extract multi-scale spectral, spatial, texture, and elevation features by a series of convolution steps, which results in higher in-class similarity and higher divisibility among different classes. Therefore, it is superior to the pixel-based SVM classifier. Furthermore, the correlation among bands is better considered in 3D CNNs than in 2D CNN by using three-dimensional convolution to extract the features of three bands.

For the producer's accuracy and user's accuracy of each class, the producer's accuracy of the impervious surfaces and the user's accuracy of bare soil for SVM and 2D CNN method are lower than 3D CNN method, especially

WorldView-2 only classification. The main reason is that many impervious surfaces are misclassified as bare soil due to lacking of height information. When adding the LiDAR height features, 3D CNNs have a better representation capacity to features than SVM and 2D CNNs.

To better demonstrate the effectiveness of our proposed 3D CNNs in impervious surface mapping, the detailed results in various local areas (Fig. 4) are further compared. For high-resolution optical images, because of the complex landscape types in urban areas, the same objects with different material (e.g. the rooftops) have different spectral signatures. In Fig. 4I, II, some building roofs with different materials are misclassified as bare soil from the WV-2 image only using the SVM method (Fig. 4b). In Fig. 4c, height information extracted from LiDAR data resolves this problem. In Fig. 4d, 2D CNNs improve the extraction of buildings through training the spectral and spatial features from WorldView-2 image. However, some roads are misclassified as bare soil. After adding the height information, the mixture gets the improvement (Fig. 4e). For the 3D CNNs, with both WV-2 image only and WV-2 + LiDAR, rooftops are completely identified (Fig. 4f, g), because 3D CNNs not only extract spectral, spatial, and height features, but also consider the correlation of neighbourhood spectrum. Moreover, in Fig. 4III, for the narrow roads in the park, SVM and 2D CNN method cannot extract completely. The best performance is 3D CNNs for WorldView-2 images. The 3D CNNs for WV-2 + LiDAR have some overestimation for the impervious surfaces in tree crown areas.

### Conclusions

In this study, a 3D CNN approach is proposed and employed for extracting urban impervious surfaces from HR WV-2 and airborne LiDAR datasets. We further evaluate the influences of different 3D CNN parameters on impervious surface extraction. Several findings are summarized.

Via deep learning, our proposed 3D CNN method can automatically extract spectral, spatial, textural, and elevation features via multi-step convolutional, ReLU, and pooling operators, which result in better extraction performance of impervious surfaces (especially for building roofs and roads).

Our results show that different 3D CNN parameters have significant effects on impervious surface extraction. The optimal combination of convolutional kernel size  $n$  and pooling dimension  $p$  is  $n \in [2p, 3p]$ . For input image size  $m$ , a smaller  $m$  results in lower accuracy, while a larger  $m$  requires more computation time. The range from 20 to 40 is an optimal choice for parameter  $m$ .

Additionally, the performance of impervious surface extraction is the most stable when the number of CNN layers  $L$  is set to 2.

We often rely on multi-source datasets such as WV-2 and LiDAR to improve the performance of pixel-based classifiers. The improvement by using multi-source data is less dramatic for the 3D CNN method. Even with only a single-source HR image, the 3D CNN model is able to extract impervious surfaces with high accuracy.

**Acknowledgements** The authors would like to thank the anonymous reviewers and the editor for their constructive comments and suggestions for this paper. This work was supported by the National Key Research and Development Program No. (2016YFA0600302), the National Natural Science Foundation of China No. (41201357), the Technology Cooperation Project of Sanya No. (2015YD18), and the Open Research Funded by Key Laboratory of Satellite Mapping Technology and Application, National Administration of Surveying No. (KLMSTA-201605).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Akpona, O., Frank, C., Sam, D. C., Jeroen, D., et al. (2018). Generalizing machine learning regression models using multi-site spectral libraries for mapping vegetation-impervious-soil fractions across multiple cities. *Remote Sensing of Environment*, 216, 482–496. <https://doi.org/10.1016/j.rse.2018.07.011>.
- Chen, Y. S., Jiang, H. L., Li, C. Y., Jia, X. P., & Ghamisi, P. (2016a). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10), 6232–6251. <https://doi.org/10.14358/PERS.80.1.91>.
- Chen, S. Z., Wang, H. P., Xu, F., & Jin, Y. Q. (2016b). Target classification using the deep convolutional networks for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8), 4806–4817. <https://doi.org/10.1109/TGRS.2016.2551720>.
- Cheng, G., Zhou, P. C., & Han, J. W. (2016). Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), 7405–7415. <https://doi.org/10.1109/TGRS.2016.2601622>.
- Ding, J., Chen, B., Liu, H., & Huang, M. (2016). Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and Remote Sensing Letters*, 13(3), 364–368. <https://doi.org/10.1109/LGRS.2015.2513754>.
- Gao, Q., Lim, S., & Jia, X. (2018). Hyperspectral image classification using convolutional neural networks and multiple feature learning. *Remote Sensing*, 10, 1–18. <https://doi.org/10.3390/rs10020299>.
- Guo, H. D., Yang, H. N., Sun, Z. C., Li, X. W., & Wang, C. Z. (2014). Synergistic use of optical and PolSAR imagery for urban impervious surface estimation. *Photogrammetric Engineering and Remote Sensing*, 80(1), 91–102. <https://doi.org/10.14358/PERS.80.1.91>.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507. <https://doi.org/10.1126/science.1127647>.
- Hu, X., & Weng, Q. (2009). Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks. *Remote Sensing of Environment*, 113(10), 2089–2102. <https://doi.org/10.1016/j.rse.2009.05.014>.
- Im, J. H., Lu, Z. Y., Rhee, J. Y., & Quackenbush, L. J. (2012). Impervious surface quantification using a synthesis of artificial immune networks and decision/regression trees from multi-sensor data. *Remote Sensing of Environment*, 117(1), 102–113. <https://doi.org/10.1016/j.rse.2011.06.024>.
- Ji, S. W., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 221–231. <https://doi.org/10.1109/TPAMI.2012.59>.
- Kim, H., Jeong, H., Jeon, J., & Bae, S. J. (2016). The impact of impervious surface on water quality and its threshold in Korea. *Water*, 8(4), 1–9. <https://doi.org/10.3390/w8040111>.
- Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778–782. <https://doi.org/10.1109/LGRS.2017.2681128>.
- Långkvist, M., Kiselev, A., Alirezaie, M., & Loutfi, A. (2016). Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, 8(4), 1–21. <https://doi.org/10.3390/rs8040329>.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>.
- Lin, T., Goyal, P., Girshick, R. B., He, K., & Doll, P. (2017). Focal loss for dense object detection. *CoRR*, abs/1708.02002, 1–10.
- Ma, Q., Wu, J. G., & He, C. Y. (2016). A hierarchical analysis of the relationship between urban impervious surfaces and land surface temperatures: Spatial scale dependence, temporal variations, and bioclimatic modulation. *Landscape Ecology*, 31(5), 1139–1153. <https://doi.org/10.1007/s10980-016-0356-z>.
- Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2016). Convolutional neural networks for large-scale remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), 7405–7415. <https://doi.org/10.1109/TGRS.2016.2612821>.
- Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2016). Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1), 105–109. <https://doi.org/10.1109/LGRS.2015.2499239>.
- Parece, T. E., & Campbell, J. B. (2013a). Landsat and high resolution aerial photography. *Remote Sensing*, 5(10), 4942–4960. <https://doi.org/10.3390/rs5104942>.
- Parece, T. E., & Campbell, J. B. (2013b). Comparing urban impervious surface identification using.
- Scott, G. J., England, M. R., Starns, W. A., Marcum, R. A., & Davis, C. H. (2017). Training deep convolutional neural networks for land-cover classification of high-resolution imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(4), 549–553. <https://doi.org/10.1109/LGRS.2017.2657778>.
- Sevo, I., & Avramovic, A. (2016). Convolutional neural network based automatic object detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 13(5), 740–744. <https://doi.org/10.1109/LGRS.2016.2542358>.
- Sun, Z. C., Guo, H. D., Li, X. W., Lu, L. L., & Du, X. P. (2011). Estimating urban impervious surfaces from landsat-5 TM imagery using multilayer perceptron neural network and support vector machine. *Journal of Applied Remote Sensing*, 5(1), 053501. <https://doi.org/10.1117/1.3539767>.

- Sun, Z. C., Li, X. U., Fu, W. X., Li, Y. K., & Tang, D. S. (2014). Long-term effects of land use/land cover change on surface runoff in urban areas of Beijing, China. *Journal of Applied Remote Sensing*, 8(1), 084596. <https://doi.org/10.1117/1.JRS.8.084596>.
- Sun, Z. C., Wang, C. Z., Guo, H. D., & Shang, R. R. (2017). A modified normalized difference impervious surface index (MNDISI) for automatic urban mapping from landsat imagery. *Remote Sensing*, 9(9), 942. <https://doi.org/10.3390/rs9090942>.
- Touchaei, A. G., Akbari, H., & Tessum., C. W. (2016). Effect of increasing urban albedo on meteorology and air quality of Montreal (Canada)—Episodic simulation of heat wave in 2005. *Atmospheric Environment*, 132(1), 188–206. <https://doi.org/10.1016/j.atmosenv.2016.02.033>.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *IEEE international conference on computer vision (ICCV)*, Chile, December 7–13.
- Vedaldi, A. & Lenc, K. (2015). MatConvNet: Convolutional neural networks for MATLAB. In *Proceedings of the 23rd ACM international conference on multimedia*, Brisbane, Australia, 26–30 October 2015 (pp. 689–692). <https://doi.org/10.1145/2733373.2807412>.
- Wang, J., Wu, Z., Wu, C., Cao, Z., et al. (2018). Improving impervious surface estimation: an integrated method of classification and regression trees (CART) and linear spectral mixture analysis (LSMA) based on error analysis. *GIScience and Remote Sensing*, 55(4), 583–603. <https://doi.org/10.1080/15481603.2017.1417690>.
- Weber, C., Aguejdad, R., Briottet, X., et al. (2018). Hyperspectral imagery for environmental urban planning. In *IGARSS 2018*.
- Xu, H. (2013). Rule-based impervious surface mapping using high spatial resolution imagery. *International Journal of Remote Sensing*, 34(1), 27–44. <https://doi.org/10.1080/01431161.2012.703343>.
- Xu, J., Zhao, Y., Zhong, K., Zhang, F., Liu, X., & Sun, C. (2018a). Measuring spatio-temporal dynamics of impervious surface in Guangzhou, China, from 1988 to 2015, using time-series Landsat imagery. *Science of the Total Environment*, 627, 264–281. <https://doi.org/10.1016/j.scitotenv.2018.01.155>.
- Xu, R., Zhang, H. S., & Lin, H. (2017). Urban impervious surfaces estimation from optical and SAR imagery: a comprehensive comparison. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(9), 4010–4021. <https://doi.org/10.1109/JSTARS.2017.2706747>.
- Xu, Z., Chen, J., Xia, J., Du, P., et al. (2018b). Multisource earth observation data for land-cover classification using random forest. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 789–793. <https://doi.org/10.1109/LGRS.2018.2806223>.
- Yue, Q., & Ma, C. (2016). Deep learning for hyperspectral data classification through exponential momentum deep convolution neural networks. *Journal of Sensor*, 2016, 1–9. <https://doi.org/10.1155/2016/3150632>.
- Zhang, L., Zhang, M., & Yao, Y. (2018). Mapping seasonal impervious surface dynamics in Wuhan urban agglomeration, China from 2000 to 2016. *International Journal of Applied Earth Observation and Geoinformation*, 70, 51–61. <https://doi.org/10.1016/j.jag.2018.04.005>.
- Zhang, P., Niu, X., Dou, Y., & Xia, F. (2017). Airport detection on optical satellite images using deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(8), 1183–1187. <https://doi.org/10.1109/LGRS.2017.2673118>.
- Zhang, Y., Zhang, H., & Lin, H. (2014). Improving the impervious surface estimation with combined use of optical and SAR remote sensing images. *Remote Sensing of Environment*, 141(2), 155–167. <https://doi.org/10.1016/j.rse.2013.10.028>.