ISPRS Journal of Photogrammetry and Remote Sensing 143 (2018) 3–14

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data

Ying Sun[a], Xinchang Zhang[b,*], Qinchuan Xin[a,c,*], Jianfeng Huang[a]

[a] Department of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China
[b] School of Geographical Sciences, Guangzhou University, Guangzhou 510006, China
[c] Guangdong Key Laboratory for Urbanization and Geo-simulation, Guangzhou 510275, China

ARTICLE INFO

Keywords:
LiDAR
High-resolution imagery
Multi-modal fusion
Multi-resolution segmentation
Semantic segmentation

ABSTRACT

Semantic segmentation of LiDAR and high-resolution aerial imagery is one of the most challenging topics in the remote sensing domain. Deep convolutional neural network (CNN) and its derivatives have recently shown the abilities in pixel-wise prediction of remote sensing data. Many existing deep learning methods fuse LiDAR and high-resolution aerial imagery towards an inter-modal mode and thus overlook the intra-modal statistical characteristics. Additionally, the patch-based CNNs could generate the salt-and-pepper artifacts as characterized by isolated and spurious pixels on the object boundaries and patch edges leading to unsatisfied labelling results. This paper presents a semantic segmentation scheme that combines multi-filter CNN and multi-resolution segmentation (MRS). The multi-filter CNN aggregates LiDAR data and high-resolution optical imagery by multi-modal data fusion for semantic labelling, and the MRS is further used to delineate object boundaries for reducing the salt-and-pepper artifacts. The proposed method is validated against two datasets: the ISPRS 2D semantic labelling contest of Potsdam and an area of Guangzhou in China labelled based on existing geodatabases. Various designs of data fusion strategy, CNN architecture and MRS scale are analyzed and discussed. Compared with other classification methods, our method improves the overall accuracies. Experiment results show that our combined method is an efficient solution for the semantic segmentation of LiDAR and high-resolution imagery.

1. Introduction

With the rapid development of remote sensing technology, advanced airborne sensors such as optical sensors and Laser Detection and Ranging (LiDAR) could provide high-resolution remote sensing (HRRS) images at the sub-meter and even centimeter spatial resolution. As a research frontier in the field of remote sensing, classification of high-resolution images as well as LiDAR point cloud data plays an important role in a wide range of applications such as urban planning, environmental monitoring, forestry and agriculture management and land inventory. Among various classification approaches, semantic segmentation is a common one that interprets the remote sensing images at the pixel level (i.e., making a prediction for each individual pixel). However, the complexity of high-resolution images and LiDAR data in spatial and spectral patterns makes semantic segmentation a challenging task.

Machine learning approaches that use support vector machine (SVM, Vapnik and Vapnik, 1998), AdaBoost (Freund and Schapire, 1995), random forest (RF, Ho, 1998), and artificial neural network (ANN, Fukushima et al., 1983) have been widely developed for semantic segmentation. In these approaches, LiDAR point cloud data are commonly converted to range images and then concatenated with raw images to obtain images that are more informative than either of the data sources (Zhang, 2010). However, accurate feature representation for the concatenated images is essential for pixel-wise prediction in the above-mentioned machine learning approaches (Zhou et al., 2016). To address this issue, many hand-crafted intra-modal features based on spectral signals, geometry, height and texture have been extracted (Stumpf and Kerle, 2011), such as spectral indices that highlight certain objects and geometry features extracted by the morphological profile (Liao et al., 2015), scale-invariant feature transform (Lowe, 2004), local binary patterns (Ojala et al., 2002), histogram of oriented gradients (Dalal and Triggs, 2005), the bag of visual words model (Yang and Newsam, 2010) and sparse representation (Han et al., 2014). These intra-modal features extracted from the two data sources, i.e., feature level fusion, are superior to the raw data (Wang et al., 2007). However, due to the heterogeneous appearance and large intra-class variance characteristics of the high-resolution imagery and LiDAR data, the above studies only extracted shallow features that were low-level or middle-level, which are not representative enough. A higher

* Corresponding authors at: Department of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China (Q. Xin).
E-mail addresses: eeszxc@mail.sysu.edu.cn (X. Zhang), xinqinchuan@gmail.com (Q. Xin).

https://doi.org/10.1016/j.isprsjprs.2018.06.005
Received 27 September 2017; Received in revised form 18 February 2018; Accepted 6 June 2018
Available online 19 June 2018

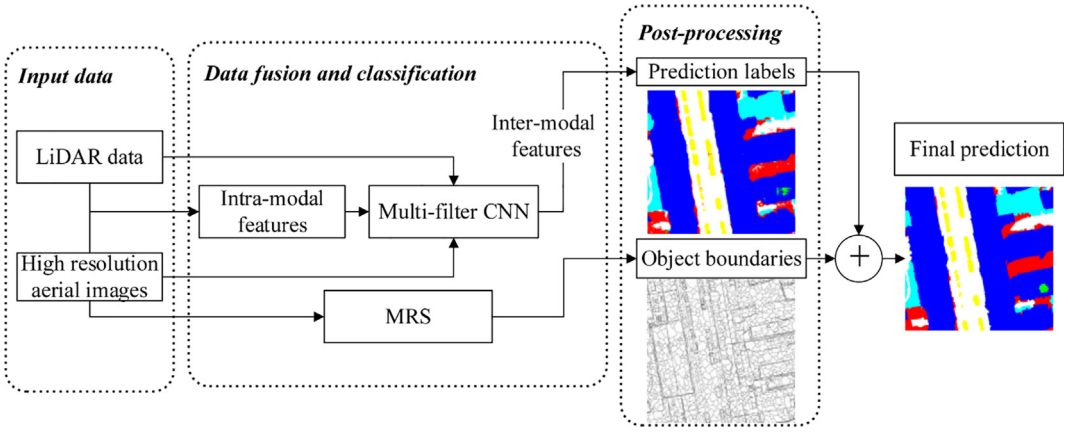

Fig. 1. The flowchart of the semantic segmentation method that consists of multi-filter CNN and multi-resolution segmentation.

level of abstract features is more discriminating and will be helpful for the improvement of semantic segmentation (Zhang et al., 2016).

Deep learning such as convolutional neural networks (CNN) has received an increasing amount of attention (Gupta et al., 2014; Long and Jin, 2008; Simonyan and Zisserman, 2014; Girshick et al., 2014), and has been applied in semantic labelling of remote sensing imagery (Längkvist et al., 2016; Islam et al., 2017; Lin et al., 2017; Zhao et al., 2017). CNN could fuse the high-resolution imagery and LiDAR data in the inter-modal way and extract high-level features that outperform hand-crafted intra-modal features. Volpi and Tuia (2017) presented an architecture with full patch labeling by learned upsampling (CNN-FPL) using the concatenated imagery and LiDAR data. Sherrah (2016) use the 3-band image data as input, and proposed a hybrid network that combines the pre-trained image features with DSM. Liu et al. (2017) proposed a decision-level scheme for data fusion of image and LiDAR, in which a CNN is trained based on the CIR images; and the inter-modal trained features and the hand-crafted features are combined in the final CRF framework. However, CNN with a fixed scale often limits the receptive field and makes feature extraction difficult. Unlike fixed CNNs, multi-scale CNNs consider multiple scales to capture different information for HRRS classification. They come in three flavors: (i) methods that use the same resolution input images with different patch sizes (Paisitkriangkrai et al., 2016); (ii) methods that use different resolution input images of the same geographical area (Zhao and Du, 2016); and (iii) methods that use CNN with different kernel sizes (Audebert et al., 2016). For the first two approaches, different kinds of input data should be prepared that cannot be directly used in the encoder-decoder CNN architecture as the input images and the corresponding labelled images are of different resolutions. For the third method, multi-scale CNN with different kernels is trained separately for earth observation data classification, and the losses of the three CNNs are averaged for error propagation. In general, there remain two defects that need improvement: (1) Loss averaging may introduce errors from single-kernel CNN and thus influence the correct weights updating; and (2) the existing multi-scale CNNs only use the inter-modal features extracted based on CNN, while intra-modal structures inferred preciously can often help the higher-level features to be mined more accurately. In addition, although the encoder-decoder CNN architecture up-samples the low-resolution features derived from pooling layers to the input resolution (Badrinarayanan et al., 2015; Chen et al., 2016; Long et al., 2015), the object boundaries are blurred irreversibly because the upsampling layers reconstruct the appearance of the object rather than the shape. Whereas CNNs often use patch-based images for classification owing to the computational ability, there is a lack of contextual information for pixels near the patch edges, resulting in the salt-and-pepper artifacts near the patch edges when mosaicking images (Mnih, 2013).

To overcome these problems, we develop a method that combines multi-filter CNN and MRS post-processing for semantic segmentation of high-resolution aerial imagery and LiDAR data. The multi-filter CNN employs three parallel CNNs with filters of different spatial context size, and a two-route loss function is employed for weight updating. LiDAR data and imagery are fused in the multi-filter CNN for multi-modal features extraction and classification. MRS is then used to delineate boundaries of objects and eliminate the salt-and-pepper artifacts. Two datasets, i.e., the Potsdam dataset in the ISPRS 2D labelling contest and the Guangzhou dataset in China, are used for method assessment.

2. Methods

The proposed method of semantic segmentation, as shown in Fig. 1, mainly consists of multi-filter CNN and multi-resolution segmentation. An end-to-end multi-filter CNN is first built for the classification of fused high-resolution aerial imagery and LiDAR data, and the method of multi-resolution segmentation is then applied to delineate object boundaries and refine the results.

2.1. A brief description of the convolution neural network

The convolutional neural network (CNN) is typically comprised of several convolutional stages. Each convolutional stage consists of multiple layers such as the convolutional layer, the activation function layer (usually a rectified linear unit, ReLU), the pooling layer, and the optional layer of batch normalization (BN). In this paper, we use a SegNet-like CNN which has a convolutional-deconvolutional structure, in which the deconvolutional process is to up-sample the input feature maps that are down-sampled by the pooling layers in the convolutional stage. Each deconvolutional stage is usually composed of the upsampling layer, the convolutional layer, and the optional layer of batch normalization.

2.2. A multi-filter convolutional neural network

The size of the receptive field determines the observation scale greatly and affects the prediction results accordingly. As traditional CNN adopts fixed filter sizes and hence limited observation scales (Zhao and Du, 2016), ensemble methods that apply the multi-scale technique are favorable in practice. To explore the multi-resolution features of local and global contexts, we develop a multi-filter CNN that consists of three different filters (i.e., 3×3, 5×5, and 7×7 in parallel) for data fusion and semantic segmentation.

As both high-resolution aerial images and LiDAR data are involved, the method first learns the intra-modal features from each individual data source and then extracts the inter-modal features using the multi-

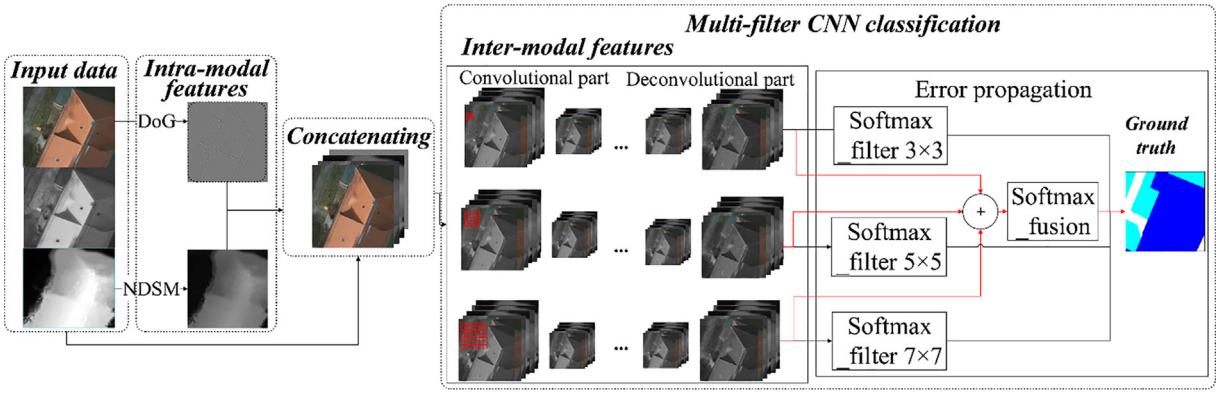

Fig. 2. Multi-modal fusion of LiDAR data and high-resolution aerial imagery based on multi-filter CNN for semantic segmentation.

filter CNN (Fig. 2). The intra-modal features of both LiDAR and high-resolution aerial images are considered. In the LiDAR data source, the normalized digital surface model (NDSM) is chosen as the intra-modal feature, which is generated by subtracting the digital elevation model (DEM) from the digital surface model (DSM) to distinguish objects by elevation. For imagery, we introduce difference of Gaussians (DoG) to produce accurate boundary localization (Davidson and Abramowitz, 2006). DoG performs edge detection based on two different Gaussian kernels. Let $f(x, y)$denote the source image and $G_{\sigma_1}(x, y)$, $G_{\sigma_2}(x, y)$ denote Gaussian kernels with standard deviation ofσ_1, σ_2; DoG is defined as follows:

$$DoG{*}f(x,y) = G_{\sigma_1}(x,y){*}f(x,y) - G_{\sigma_2}(x,y){*}f(x,y) = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma_1}\exp\left(-\frac{x^2+y^2}{2\sigma_1^2}\right) - \frac{1}{\sigma_2}\exp\left(-\frac{x^2+y^2}{2\sigma_2^2}\right)\right){*}f(x,y) \quad (1)$$

where DoG detects edges by the values that cross zero.

After that, the two intra-modal features are concatenated with LiDAR data (intensity, number of returns) and source imagery to produce an input map of multiple layers. Then, they are fed to the convolutional network for inter-modal feature learning.

The multi-modal fusion method sufficiently exerts the advantages of the intra-modal properties of each data source and inter-modal correlations in multi-filter CNN, preventing either of them from being overlearned.

The multi-filter CNN architecture has three parallel neural networks with five-layer convolutional part and a symmetrical deconvolutional part (see Fig. 3). Each convolutional layer is made up of a convolutional layer of 64 output features, a BN layer, a ReLU layer and a max pooling layer. The corresponding deconvolutional layer has an upsampling layer, a convolutional layer and a BN layer. Accordingly, inter-modal feature maps are derived within the CNN. To achieve translation invariance, the pooling layer is used and the deep feature maps are downsampled by a factor of 2. Symmetrically, the upsampling layers upsample the input feature maps by a factor of 2. Finally, the feature maps extracted by single filter in the parallel network are sent to a softmax classifier separately to calculate the loss. Meanwhile, they are concatenated and fed to another softmax classifier to produce class probabilities for each pixel independently. Different from the inception networks (Szegedy et al., 2016) which concatenate the different filters' features in each convolutional part, we do not concatenate the three filters' features until they are fed to the softmax classifier.

In the model training phase, we use a two-route loss calculating method, which includes the loss of the three single-filters' output as well as the fused output (as is shown in the error propagation part of Fig. 3). The cross-entropy loss function is employed for loss calculation:

$$Loss_{\text{single}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} z_j^i \log\left(\frac{\exp(y_j^{i,s})}{\sum_{j=1}^{k}\exp(y_j^{i,s})}\right) \quad (2)$$

$$Loss_{fuse} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} z_j^i \log\left(\frac{\exp(y_j^{i,f})}{\sum_{j=1}^{k}\exp(y_j^{i,f})}\right) \quad (3)$$

where S denotes the total number of filters in multi-filter CNN, N the total number of pixels in an image, k the total class, $y_j^{i,s}$ the predictions of a pixel at index i for the j-th class at the s-th scale, $y_j^{i,f}$ the predictions of the i-th pixel for the j-th class after the feature concatenated, and z_j^i denotes the label. In our architecture, the above loss is minimized using the stochastic gradient descent (Bottou, 2010) technique for weights updating.

2.3. Post-processing using multi-resolution segmentation

Due to the spatial coherence of objects, the labels of nearby image pixels are strongly correlated, and hence, structured knowledge is helpful for classification (Mnih, 2013). The multi-resolution segmentation (MRS) algorithm (Baatz, 2000) is a standard structured prediction approach that could delineate homogenous objects based on segmentation.

MRS is a bottom-up region merging method. Images are segmented to small objects, and then homogenous objects adjacent to each other are merged into larger objects based on the parameters of scale, shape and compactness. Scale parameter is the most important parameter in the MRS algorithm, which determines the size of the segment objects and the corresponding homogeneity. Small scale results in over-segmentation with a small size of objects and high homogeneity in an object. MRS delineated objects often have sharp class boundaries. By comparison, object boundaries in the CNN result normally appear to be coarse. Therefore, MRS can be used as a supplement of pixel-wise CNN classification to delineate boundary information and clean up predictions. In this paper, the MRS result is only used as a post-processing method to eliminate classification noise. The multi-filter CNN classified pixels are smoothed within each object using the majority rule. In essence, the label of a given object is determined by the label that occurs most often in the CNN classification result.

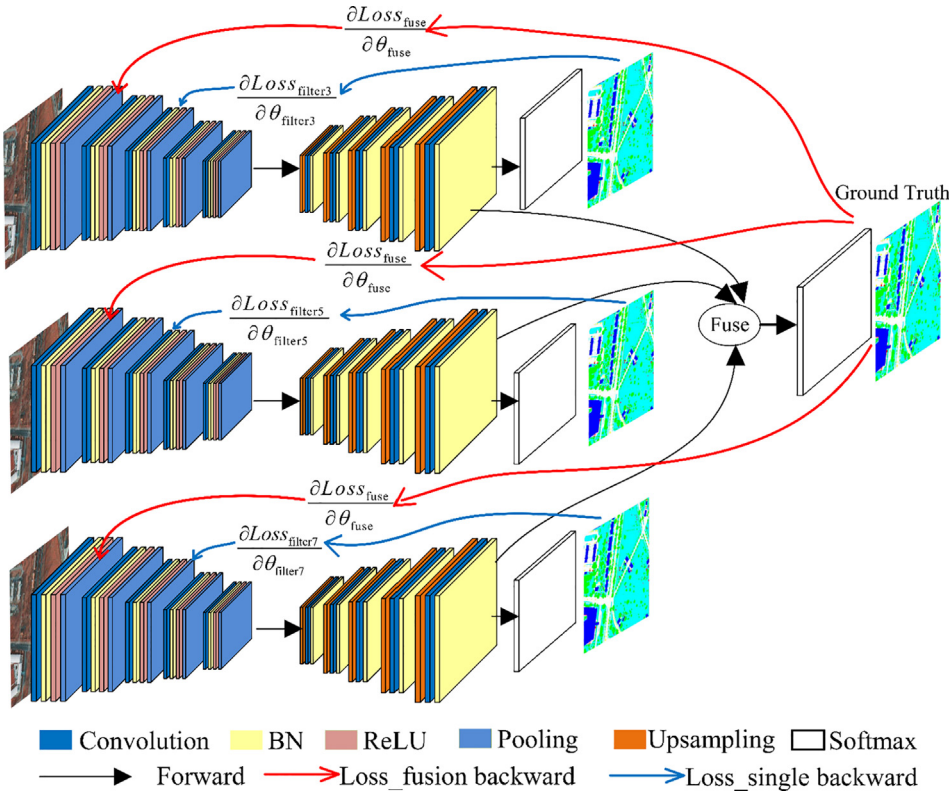

Fig. 3. An illustrative scheme for the architecture of the developed multi-filter CNN.

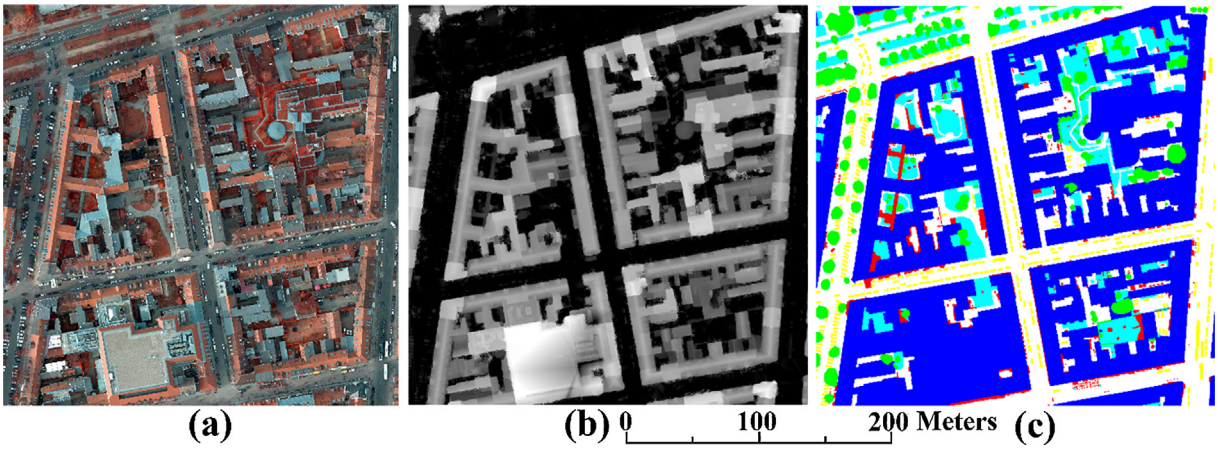

Fig. 4. Images for the Potsdam dataset are shown for (a) the false-color composite, (b) NDSM, and (c) the labelled ground reference.

3. Experiment design

3.1. Study materials

Two datasets are used for tests; one is the Potsdam dataset as obtained from the ISPRS 2D semantic labelling benchmark, and the other is from Guangzhou City in China. Both datasets have high-resolution aerial images and LiDAR data. In the Potsdam dataset, the high-resolution aerial images have four channels from red[1], green, blue, and near infrared bands, whereas the LiDAR data include DSM and NDSM as provided in the raster format. The ground sampling distance of the Potsdam dataset is approximately 5 cm per pixel. The Potsdam dataset contains 38 patches in total, of which 24 have already been labelled, and each patch has an image size of 6000 × 6000. Fig. 4 shows the test images for the Potsdam dataset.

The Guangzhou dataset is acquired by an airborne system that provides simultaneous RGB images and LiDAR measurements. As the optical sensor in airborne system does not have a near infrared band and LiDAR uses the near infrared light to measure the surfaces, the LiDAR intensity and the number of returns are used as surrogates of the near infrared band to help distinguish vegetation and impervious surfaces in the Guangzhou dataset. The spatial resolution of the RGB images is 10 cm per pixel, and the LiDAR point cloud is approximately 5 points per square meter. To fuse the high-resolution aerial images and LiDAR data efficiently, the vector point clouds were transformed into raster data by the following processes. First, non-ground points are separated from the ground points. Second, the DEM and DSM are derived by rasterizing three-dimensional ground and non-ground point cloud data based on the inverse-distance-weighted method, whereas

[1] For interpretation of color in Figs. 4 and 5, the reader is referred to the web version of this article.

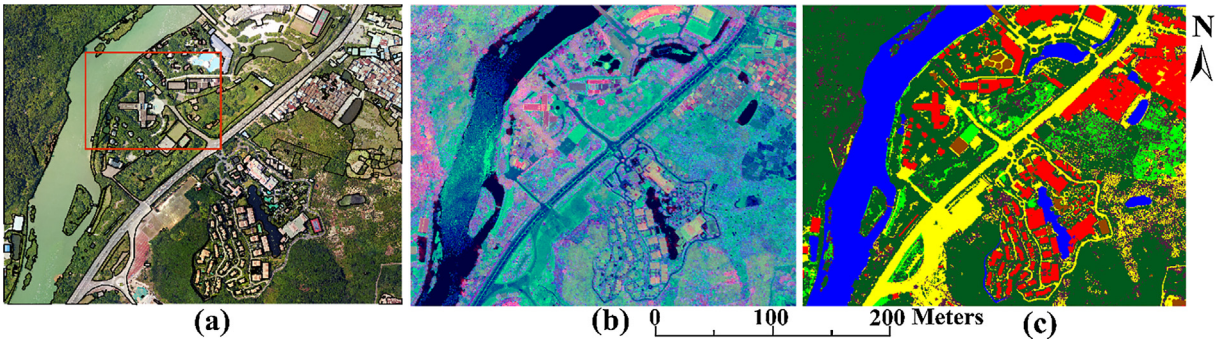

Fig. 5. Images for the Guangzhou dataset are shown for (a) DOM overlaid by the existing geodatabase, (b) the false-color composite of NDSM, LiDAR intensity and the number of returns, and (c) the labelled ground truth.

NDSM is simply derived by subtracting DEM from DSM. Third, LiDAR intensity data are normalized using the method of Höfle and Pfeifer (2007), which derives the normalized intensity based on the standard range, measured intensity, and measured range between sensor and surface (Eqs. (4) and (5). Last, the point intensity and the number of returns of LiDAR are converted to the raster format. As shown in Fig. 5, the entire Guangzhou dataset has an image size of 10000 × 8000, from which a representative area that covers all classes as marked by the red rectangle is chosen for validation and the rest is used for training.

$$\begin{cases} i_N(R) = i_{R_s} f(R), f(R) < f(R + \Delta R)\ \forall\ \Delta R > 0 \\ f(R_s) = 1 \end{cases} \tag{4}$$

$$f(R) = \frac{1}{aR^2 + bR + (1 - R_s^2 a - R_s b)} \tag{5}$$

where R_s denotes the standard range, i denotes the measured intensity, R denotes the measured range between sensor and surface, i_N denotes normalized intensity, and a and b are the function parameters.

3.2. Image labelling

The training of deep CNN requires a large number of labelled images. The Potsdam dataset has already been labelled by the ISPRS research community and contains the major classes of impervious surfaces, buildings, low vegetation, trees, and cars in addition to the minor class of background. However, the Guangzhou dataset has to be labelled before use, and thus a semi-automatic image labelling approach was applied based on existing geodatabases constructed for nearly the same time period by the Guangzhou Urban Planning and Design Survey Research Institute. The fundamental geodatabase contains objects such as polygons, lines, and points in the vector format. The polygon objects include the classes of building, road, infrastructure (such as swimming pool and plastic track), ground, grass, and water body (such as rivers, ponds, and lakes). The line objects include pathways and small streams and the point objects include independent trees in addition to large blocks of vegetation that are represented by polygons. Although polygons with distinct boundaries can be easily converted to pixel-based labels, labelling image pixels based on line and point objects requires additional processes. The lines and points are first rasterized for the crossed and centered pixel and then manually extended in width to match the high-resolution aerial images. As both trees and shrubs are mixed in the vegetation polygons, a threshold of 2 m in NDSM was applied to separate them. All raster maps as labelled based on polygons, lines, and points are eventually merged into one ground reference map that is made up of seven land cover classes, including trees, buildings, shrubs, ground, infrastructure, grass, and water.

3.3. Method implementation

For tests using the Potsdam dataset, the training samples include 21 labelled images, which are cropped into 8400 patch-based images with a window size of 300 × 300 pixels, and the validation samples are the fully labelled Potsdam 5_12 image. Moreover, we trained all the 24 labelled images, and tested the remaining 14 unlabeled images in the Potsdam dataset. For tests with the Guangzhou dataset, the training samples include 3320 patch-based images with a window size of 300 × 300 pixels, of which 830 are from the raw data and the rest are derived based on data augmentation. Data augmentation is adopted by many studies and shows to be an effective way to increase the training dataset for deep learning while avoiding the problem of over-fitting (Mnih, 2013). The required training samples are augmented by rotating and reflecting in both vertical and horizontal directions. The validation samples are fully labelled as marked by the red rectangle in Fig. 5a.

There are six classes in the Potsdam dataset and seven classes as labelled in the Guangzhou dataset. As some classes could have much lower percentages in total areas than others, which likely influences feature distributions as well as classification results, each land cover class is then weighted with the class-balance method (Eigen and Fergus, 2015) as follows:

$$\begin{cases} weight_i = f_{i\_median}/f_i \\ f_i = pix_i/(img\_count_i * M * N) \end{cases} \tag{6}$$

where pix_i is the number of pixels of class i; img_count_i is the number of images where class i is present; M and N are the width and height of the input images; f_i is the pixel frequencies of class i; f_{i_median} is the median of all f_i.

Table 1
Data fusion strategies for high-resolution imagery and LiDAR data.

Data fusion strategies		Modality 1	Modality 2	Modality 3	Modality 4	Modality 5
Raw data	RGB images	√	√	√	√	√
	Near infrared band		√		√	√
	DSM	√	√	√	√	
Intra-modal features	NDVI			√		
	NDSM				√	√
	DoG					√
Inter-modal features	CNN features	√	√	√	√	√

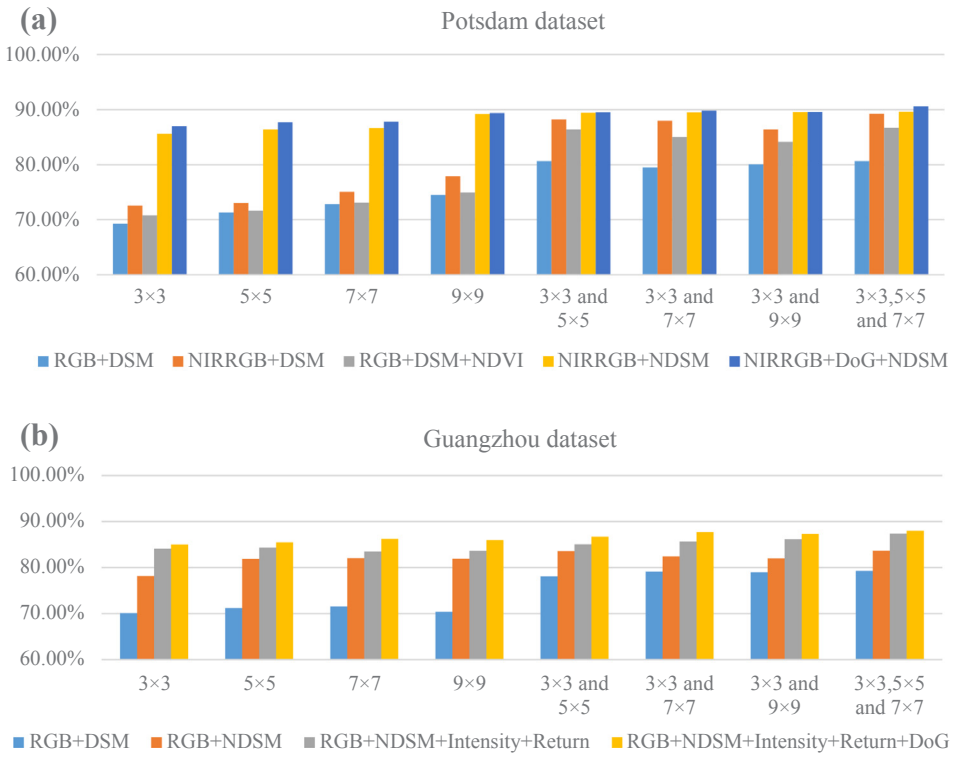

Fig. 6. Overall classification accuracies with different CNNs and data fusion strategies. (a) Results of Potsdam dataset, (b) results of Guangzhou dataset.

As both the filter size and the CNN numbers in parallel influence the computational complexity as well as the classification accuracy, comparative studies are conducted to evaluate the performance of both single-filter and multi-filter CNN on semantic segmentation. Results from single-filter CNNs with the filter size of 3 × 3, 5 × 5, 7 × 7, and 9 × 9 are compared with that from multi-filter CNNs with a combination of the filter size 3 × 3 and 5 × 5, 3 × 3 and 7 × 7, 3 × 3 and 9 × 9, 3 × 3, 5 × 5, and 7 × 7 in parallel. As shown in Table 1, five different data fusion strategies as implemented in the multi-filter CNN are also compared for extracting features across modalities. The first two modalities concatenate the raw data of RGB + DSM and NIRRGB + DSM respectively for inter-modal feature extraction. Modality 3 uses intra-modal feature NDVI and source imagery of RGB + DSM for inter-modal feature extraction. Modality 4 combines intra-modal feature NDSM and the source imagery of NIRRGB for inter-modal feature extraction. Our method (modality 5) incorporates intra-modal feature DoG, NDSM and source imagery of NIRRGB for inter-modal feature extraction.

Our architecture is running on NVIDIA TITAN X, and we compute the computational time of multi-filter CNN over 100 iterations with mini batch size 1. The average forward pass time is 143.119 ms, and the average backward pass time is 181.087 ms. To evaluate our method, we present the classification results on the Potsdam 5_12 image and Guangzhou dataset and compare our method with popular sophisticated machine learning algorithms, including SVM, RF, Gated Feedback Refinement Network (G-FRNet, Islam et al., 2017), SegNet and multi-scale (MS) CNN (Audebert et al., 2016). Furthermore, the results of benchmark test of Potsdam are qualitative compared with some of the existing state-of-the-art methods in the official leaderboard. The training datasets of SVM and RF are selected from the test images with a stratified random sampling method. However, the latter CNNs use training data from other images for model training and transfer to the test images. The metrics we use are the confusion matrices, overall accuracy (OA), and *F1 score* (*F1*) which defined as the harmonic mean of precision and recall:

$$F_1 = 2\cdot\frac{\text{precision}\cdot\text{recall}}{\text{precision} + \text{recall}},\ \textit{where},\ \ \text{precision} = \frac{T_p}{T_p + F_p},\ \text{recall} = \frac{T_p}{T_p + F_n}$$
$$OA = T_p/N \tag{7}$$

where T_p is true positive, F_p is false positive, F_n is false negative, and N is the total population number.

4. Experiment results

4.1. Influence of CNN filter and multi-modal fusion on classification

The overall accuracies of classification for the two datasets using different CNNs and data fusion strategies are shown in Fig. 6. The filter size impacts the classification performance for all data fusion strategies. Overall, multi-filter CNNs achieve higher classification accuracies than single-filter CNNs in both datasets. For single-filter CNNs, using larger filter size, such as 5 × 5, 7 × 7 and 9 × 9, performs better than using a small filter size of 3 × 3. Compared with the filter size of 3 × 3, the overall accuracies using large filters are improved by 4.16% on average for the Potsdam dataset (Fig. 6a) and 1.85% on average for the Guangzhou dataset (Fig. 6b), depending on the data fusion strategy. Regarding the filter numbers, three-filter CNN shows superiority to two-filter CNN with an average accuracy improvement of 1.55% in the Potsdam dataset and 1.61% in the Guangzhou dataset, respectively.

The data fusion strategy is also a key factor that impacts the classification accuracies. For the Potsdam dataset, the first two data fusion strategies (RGB + DSM and NIRRGB + DSM) in Table 1 only extract the inter-modal features based on multi-filter CNN, providing relatively low accuracies. Upon the multi-modal data fusion strategy, using modality 3 in Table 1 (RGB + DSM + NDVI) does not improve the accuracy as compared to using modality 2 of NIRRGB + DSM. However, modality 4 of NIRRGB + NDSM shows higher accuracy than that of modality 2 (NIRRGB + NDSM). Among all data fusion strategies, the highest accuracy is achieved for the Potsdam dataset when using modality 5 with multi-modal features (NIRRGB + DoG + NDSM) in Table 1, whereas similar results are obtained for the Guangzhou dataset as shown in Fig. 6b.

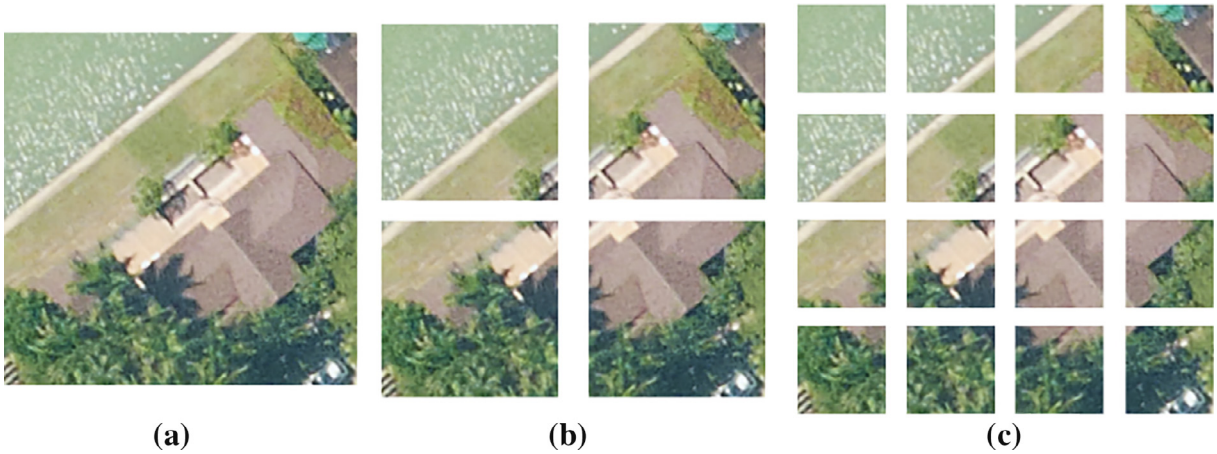

Fig. 7. Three kinds of patches tested in our experiment. (a) Patch size of 300 × 300 pixels, (b) patch size of 150 × 150 pixels, and (c) patch size of 75 × 75 pixels.

Table 2
Accuracies of different patch sizes in Guangzhou dataset.

Patch size (F1 score :%)	Tree	Building	Shrub	Ground	Infrastructure	Water	Grass	OA
75	91.41	79.40	35.91	78.82	72.38	96.10	72.84	84.36
150	92.18	82.20	43.29	79.93	71.20	96.51	74.47	85.64
300	93.38	90.05	48.08	81.55	78.82	97.63	77.36	87.88

4.2. Effect of the patch size

The training samples and validation samples in our experiment are in the form of patch. In order to evaluate the effect of the patch size for classification performances, we train three sizes of patches with the same image resolution. Considering the computation load, the largest patch size we use is 300 × 300 pixels, and what follows are 150 × 150 pixels and 75 × 75 pixels. Unlike Paisitkriangkrai et al. (2016), the three kinds of patches are not centered at the same pixel. We obtain the smaller size of patches by cropping the larger ones directly, as shown in Fig. 7.

Experimental results are reported in Table 2. The patch size of 300 × 300 pixels shows an improvement over small patch sizes in terms of overall accuracies. The larger patch size has wider scope than the smaller size, which can provide more information about neighbors, and thus multi-filter CNN generates more significant features with the neighbors. Benefiting from this, the classification results of shrub and infrastructure class with low prevalence are greatly improved, as is shown in Fig. 8. From the middle right of the test image, we can see that shrubs are finely extracted with the patch size of 300 × 300 pixels. However, the classification accuracy of the ground drops greatly, as

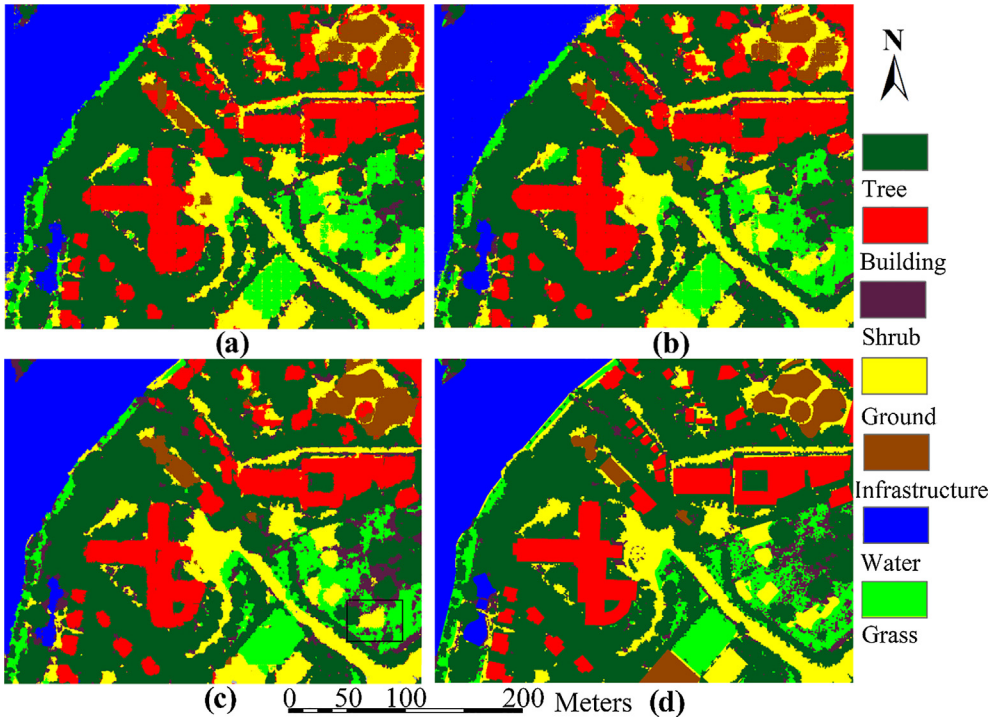

Fig. 8. Classification results of Guangzhou dataset with different patch sizes. (a) The classification map based on a patch size of 75 × 75 pixels, (b) the classification map based on a patch size of 150 × 150 pixels, (c) classification map based on a patch size of 300 × 300 pixels, and (d) the ground truth.

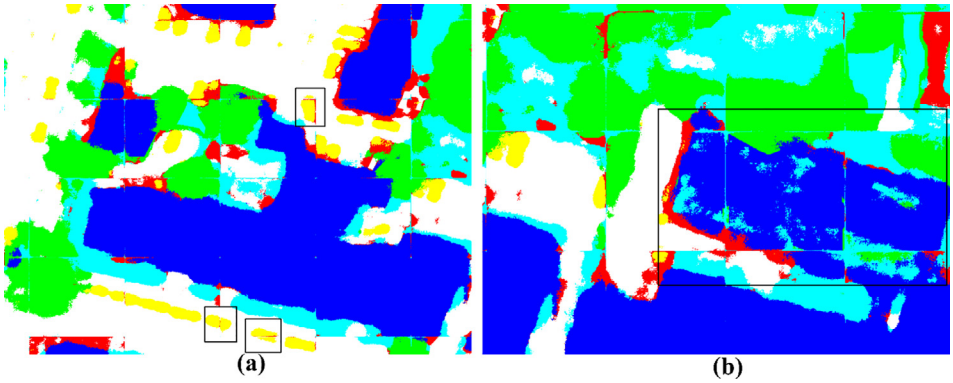

Fig. 9. The salt-and-pepper artifacts in multi-filter CNN classification results.

Table 3
Accuracies of MRS post-processing in Potsdam with three different scales.

F1 score (%)	Imp_surf	Building	Low_veg	Tree	Car	OA
Multi-filter CNN	90.89	96.94	76.01	73.84	86.93	90.62
Multi-filter CNN + MRS (Scale 10)	90.87	96.96	76.52	70.81	86.51	90.55
Multi-filter CNN + MRS (Scale 30)	90.94	96.98	76.32	73.37	88.55	90.65
Multi-filter CNN + MRS (Scale 50)	90.55	97.01	76.05	69.00	83.43	90.27

indicated by the test image; many ground pixels are identified as grass. This is largely due to the lack of near infrared channel, and the mixed ground pixels and grass pixels with similar height are considered as texture features in larger patches, and thus the accuracy is reduced. Generally, a patch size of 300 × 300 pixels performs the best in overall accuracies.

4.3. MRS based post-processing

Although intra-modal feature DoG provides edge information as compensation for the loss of spatial resolution, multi-filter CNN still needs further promotion in predicting the object boundaries and the patch edges of the input images, e.g., the cars (as shown in Fig. 9a) and the building (see Fig. 9b), which are divided into several parts, and many pixels are predicted as low vegetation in the building. MRS based post-processing is performed to let predictions at a given pixel influence that of nearby pixels and form meaningful image objects. Over-segmentation ensures a higher homogeneity in an object. According to this foundation, we evaluate the effects of MRS post-processing in three different scales.

Table 3 reports the results of MRS-based post-processing in three different scales for the Potsdam dataset. The results show that the classification accuracy increased slightly from 90.62% to 90.65% in terms of overall accuracies. The classification results indicate that the overall accuracies for scale 10, scale 30 and scale 50 are similar, of which scale 30 achieves the highest accuracies. However, per class accuracies vary with the scales. Buildings achieve better results in larger scales, while a smaller scale is beneficial for small objects such as cars.

The effects of MRS post-processing can be seen in Fig. 10. Fig. 10a–c show the classification results of MRS post-processing with a scale of 10, 30 and 50 respectively. The salt-and-pepper artifacts on the object boundaries and patch edges have been well eliminated in all three scales on the basis of over-segmentation. In terms of object shapes, scale 30 and scale 50 (see Fig. 10b and c) regularize the edges of buildings significantly. Although scale 50 does not achieve the highest overall accuracy, it improves the shapes of buildings compared to that of other two scales. By contrast, a small scale of 10 does not eliminate the salt-and-pepper artifacts well owing to the small object size, and thus, irregular edges are generated.

4.4. Classification and accuracy assessments

Fig. 11 shows the classification results of Potsdam 5_12. As compared with the ground reference in Fig. 4c, our method could produce an accurate classification map (Fig. 11f). SVM (Fig. 11a) and RF (Fig. 11b) generate results that contain apparent errors due to the salt-and-pepper artifacts. SVM overestimates and RF underestimates the class of background, while neither performs well on the class of cars. In Fig. 11c, G-FRNet incorrectly classifies the building in to trees at the bottom of the image. Both SegNet (Fig. 11d) and MS-CNN (Fig. 11e) misclassify trees along the roads as low vegetation, whereas SegNet also tends to largely misclassify the class of clutter. By comparison, our method improves the mapping of the tree class largely because MRS implemented in our method helps refine object delineation based on the coarse boundaries of patch-based CNN. Overall, the CNN-based methods perform better than traditional classifiers like SVM and RF. As our multi-filter CNN is an architecture based on VGG16, we compared it with four methods which are also VGG16 architecture in the official leaderboard of ISPRS Potsdam dataset. Fig. 12 exhibits the qualitative comparisons with the four methods. It is easy to see that our method can obtain accurate results and capture the edges of features better than the other ones.

To understand the robustness of the algorithms, the classification maps of an experimental site located in Guangzhou, China as produced from different methods are shown in Fig. 13. Similar to the results for the benchmarking dataset of Potsdam, there remains distinct salt-and-pepper artifacts in the classes of river, buildings and trees in the maps as obtained using SVM (Fig. 13a) and RF (Fig. 13b). In Fig. 13c, G-FRNet over-estimates the class of water, and many other classes such as building, grass and tree are misclassified as water. SegNet (Fig. 13d) and MS-CNN (Fig. 13e) eliminate the salt-and-pepper artifacts but often misclassify trees if located near buildings or infrastructures. Our method (Fig. 13b) improves the classification results by providing distinct building boundaries as well as reducing the salt-and-pepper artifacts near the patch edges; however, the classes of grass and ground are occasionally misclassified.

We use the no_boundary labelled ground truth for the Potsdam dataset assessment like other studies in the official leaderboard of ISPRS Potsdam dataset. That is, the accuracy metrics are the global accuracy on each class pixel except the boundaries. Table 4 shows the confusion matrix for Potsdam 5_12. In terms of the five main classes excluding clutter, the largest confusion occurs between the classes of trees and low vegetation as both have similar spectral reflectance and are located closely to each other. There is also misclassification of trees and impervious surfaces, probably because tree canopies with few leaves allow

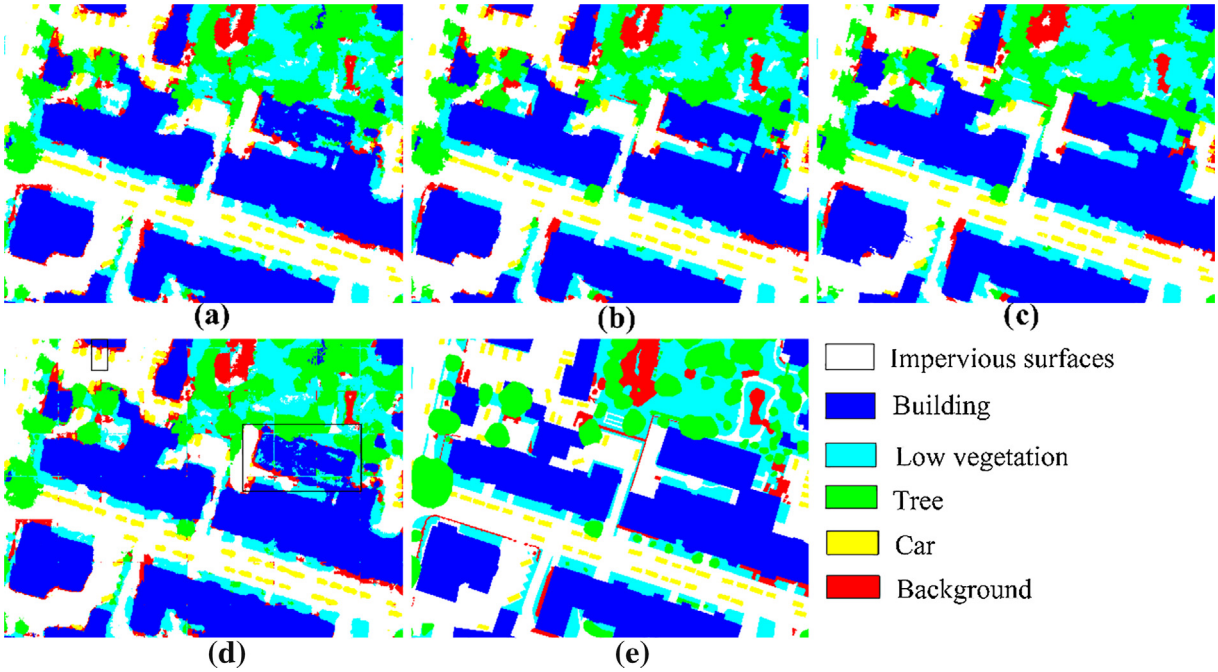

Fig. 10. MRS post-processing results of different scales. (a) Result of scale 10, (b) result of scale 30, (c) result of scale 50, (d) result of multi-filter CNN, and (e) the ground truth.

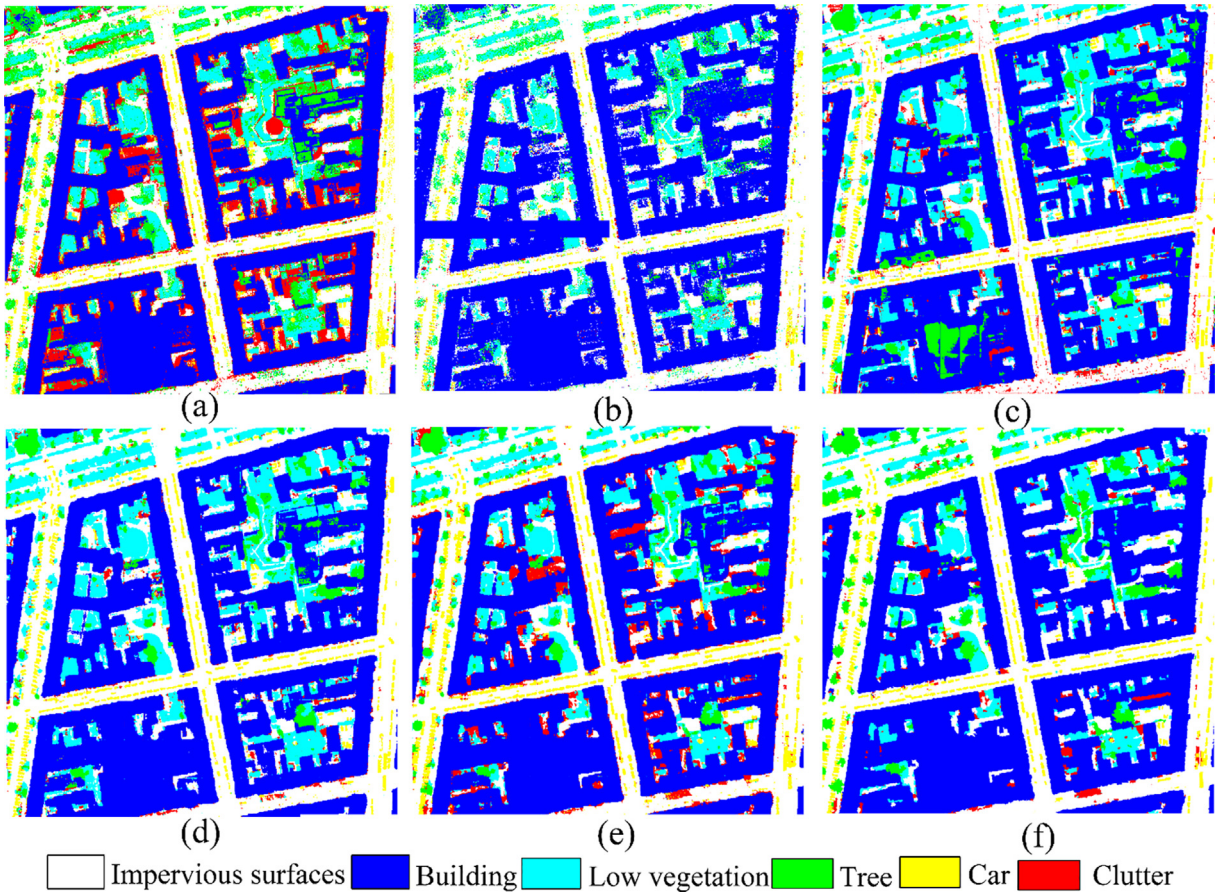

Fig. 11. Comparisons with the classification results of Potsdam dataset using the method of (a) SVM, (b) RF, (c) G_FRNet, (d) SegNet, (e) MS-CNN, and (f) multi-filter CNN.

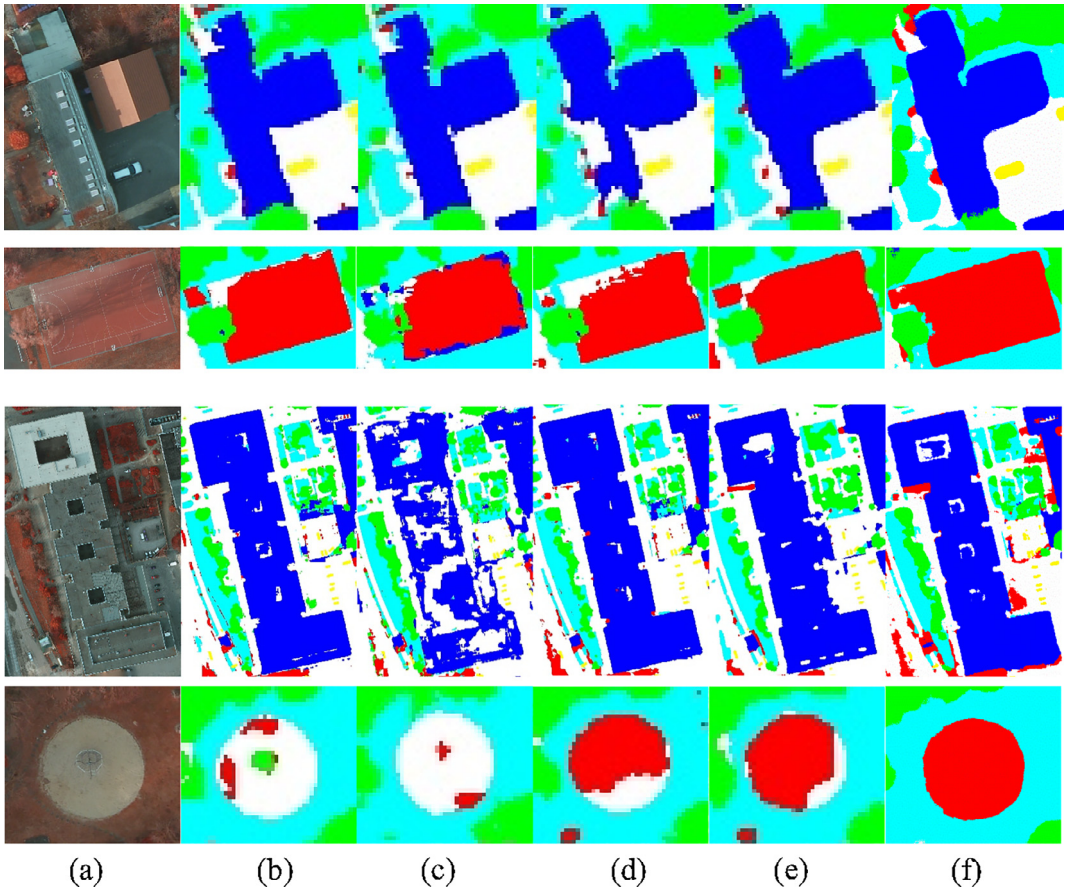

Fig. 12. Comparisons with the methods in the official leaderboard of ISPRS Potsdam dataset. The labels are the same as Fig. 11. (a) Image, (b) BKHN_1, (c) RIT_L1, (d) RIT6, (e) DST6, and (f) multi-filter CNN.

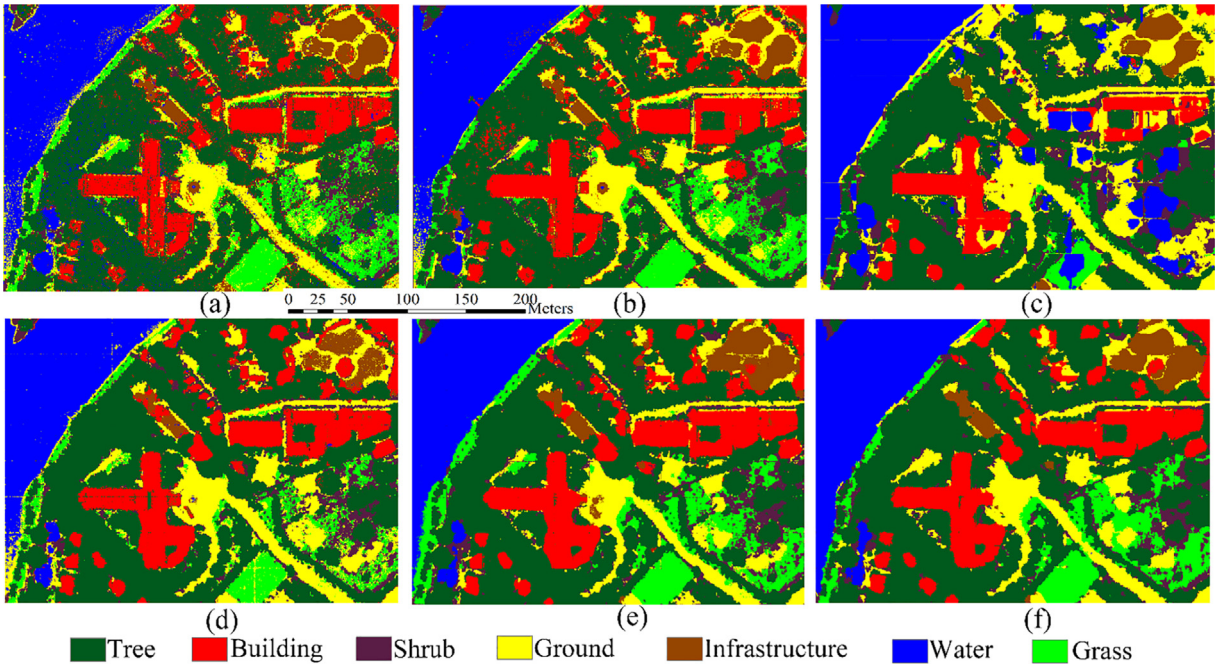

Fig. 13. The same as Fig. 11 but showing the results for the Guangzhou dataset.

for light penetration such that the underneath background can be directly seen from above. Table 5 exhibits the classification accuracies for the Potsdam dataset as obtained using different approaches. The multi-filter CNN as developed in this study achieves the highest overall accuracy.

Taking all the pixels for accuracy metrics, Table 6 shows comparisons of classification accuracies for the Guangzhou dataset. Our method achieves the highest OA and performs the best in terms of the

Table 4
The confusion matrix for the Potsdam dataset.

Classified (%)	Reference					
	Impervious	Building	Low vegetation	Tree	Car	Clutter
Impervious	90.17	0.72	9.22	13.36	3.33	28.34
Building	3.12	98.33	4.62	3.75	0.05	39.73
Low vegetation	3.13	0.36	74.76	9.44	0.04	7.88
Tree	0.55	0.20	9.51	69.07	0.50	0.74
Car	0.64	0.00	0.01	3.16	93.1	3.67
Clutter	2.38	0.39	1.88	1.22	2.98	19.63
F1	90.94	96.98	76.32	73.37	88.55	23.89

Table 5
The classification accuracies of the Potsdam dataset using different methods.

F1 score (%)	Imp_surf	Building	Low_veg	Tree	Car	OA
SVM	83.09	91.30	67.11	43.49	56.21	78.33
Random forest	85.23	93.75	65.14	38.06	66.89	84.36
G-FRNet	87.79	94.89	66.21	50.65	88.92	85.30
SegNet	88.70	95.92	67.29	52.73	84.5	87.09
Multi-scale CNN	89.64	96.83	72.58	67.88	85.41	89.43
Multi-filter CNN	90.94	96.98	76.32	73.37	88.55	90.65

classes of trees, buildings, ground and water. RF achieves the highest accuracies in the classes of shrubs and grass, whereas the deep networks of G-FRNet performs well on the classes of infrastructure.

5. Discussion

Real-world objects appear different from one receptive field size to another. CNN with single filter often extracts fixed predictions from the fixed receptive field. As shown in Fig. 6, ensemble learning of multi-filter CNN achieves the highest overall accuracies. Because multi-filter is helpful to derive more significant abstracted features at different scales, the larger filter size considers much more information about neighbors than small ones, which tends to overcome the salt-and-pepper artifacts; and the smaller filter size can capture details of objects, which is also beneficial for classification. Beyond this, the filter parameters in our network are optimized according to the two kinds of back propagation losses, which can leverage the benefits of the single- and multi-filters. As opposed to Audebert et al. (2016), the two-route loss calculating methods outperform the averaging prediction of the three single-filter losses as reported in Tables 5 and 6, which is because the fusion features can avoid errors in the single-filter CNN. Moreover, we concatenate the intra-modal features and raw data first instead of training separately, thus reducing the parameter numbers to be optimized. Compared to the work of Zhao and Du (2016), we do not prepare multi-scale input data, which is time consuming. Therefore, among these multi-scale classification methods, it is advisable to ensemble different sizes of filters for multi-modal data fusion and high accuracy predictions with low computational complexity.

In terms of the data fusion, different strategies result in different levels of features, which may accommodate heterogeneities that affect the final predictions (Lahat et al., 2015). As shown in Fig. 6 above, multi-modal data fusion is more effective than inter-modal features. The first two data fusion strategies in Table 1 consider the inter-modal features extracted by multi-filter CNN; however, they neglect the useful intra-modal features from raw imagery and LiDAR data. Although modality 3 in Table 1 uses both the intra-modal feature NDVI extracted from imagery and inter-modal features, its performance is poorer than that of the strategy in modality 2 probably because that NDVI leaves some useful information in the NIR channel out. Accordingly, appropriate features are essential, and intra-modal features of less significance may introduce uncertainties, which affect further classification. The data fusion strategy in modality 4 uses the intra-modal feature NDSM from LiDAR and inter-modal features and performs better than the first three, which indicates that NDSM removes noisy information contained in DSM. Our multi-modal data fusion strategy in modality 5 achieves the highest accuracy, and the reason for this phenomenon is obvious. Multi-modal features contain richer and more effective information than the other four data fusion strategies above.

As a pixel-wise classification method, CNNs are efficient in predicting complex senses based on the multi-modal deep features, and they show better performance than transitional methods. However, there remains some amount of salt-and-pepper artifacts on the object boundaries and the patch edges. As reported in Section 4.3, MRS post-processing improves the accuracies of our multi-filter CNN by delineating the images into relative homogeneous objects based on color, shape and texture information of pixels. The delineation scale parameter is crucial and determines the accuracy of post-processing. It seems that a slightly larger scale on the basis of over-segmentation is preferred for the larger objects, and a smaller scale is effective for smaller objects and irregularly shaped objects (as reported in Table 3). In summary, it is helpful to perform MRS post-processing to improve the classification results of patch-based CNN, and the scale is application dependent.

In the Potsdam dataset, the classes of buildings, impervious surfaces and cars achieve higher accuracies than that of vegetation. We believe that the vegetation cover in the test data influences the result. The vegetation cover in Potsdam is much more fragmented and in various shapes and sizes with the spectral features not obvious enough, thus increasing difficulties for classification. Interestingly, our method delivers a good classification of the tree class compared with other methods, and we believe that useful features are extracted by using multi-filter CNNs and multi-modal data fusion. Comparatively, the vegetation scores are higher in the Guangzhou dataset than that of Potsdam, which is largely due to the dense vegetation cover and the obvious features contained in LiDAR intensity and number of returns. In the Guangzhou dataset, our method performs worse on the class of shrubs and grass than RF does due to the biases of the training and validation datasets. RF uses the exact image for sampling by stratified randomization, and this reduces the biases.

6. Conclusion

In this study, we present a semantic segmentation method combined multi-filter convolutional neural network with MRS based on the fusion of LiDAR data and high-resolution aerial imagery. To evaluate the

Table 6
The classification accuracies of the Guangzhou dataset using different methods.

F1 score (%)	Tree	Building	Shrub	Ground	Infrastructure	Water	Grass	OA
SVM	92.33	76.97	53.26	77.20	66.59	94.92	76.32	84.66
Random forest	90.69	72.02	86.33	74.61	72.08	90.88	78.08	84.39
G-FRNet	89.95	68.19	33.76	62.27	83.42	82.68	39.38	75.89
SegNet	92.76	81.73	58.16	76.91	72.36	96.02	67.92	85.28
Multi-scale CNN	92.51	85.08	56.55	78.52	79.20	96.00	74.85	86.49
Multi-filter CNN	93.38	90.05	48.08	81.55	78.82	97.63	77.36	87.88

effectiveness of the proposed method, the ISPRS dataset of Potsdam and an area of Guangzhou, China, which is labelled by an existing geodatabase, are tested. First, the choices for different numbers of filters in a CNN framework are evaluated. Compared to a single filter CNN, the multi-filter CNN framework aggregates features at different scales, resulting in higher classification accuracies, of which three-filter CNN achieves the highest overall accuracies. Meanwhile, different data fusion strategies also influence the classification result, of which multi-modal fusion works best. Furthermore, MRS with appropriate scale can obtain effective edge information so that the classification results of multi-filter CNN are smoothed out by the MRS objects. Experimental results show a slight improvement of overall accuracies, but it eliminates salt-and-pepper artifacts on the objects boundaries and the patch edges. In particular, it is beneficial for objects with regular shapes, such as buildings and cars. In future work, there is a need to investigate how different data sources form other sensors are fused in deep CNN for further classification improvement.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This research was supported by the National Natural Science Foundation of China (grant no. 41431178), the Natural Science Foundation of Guangdong Province, China (grant no. 2016A030311016), the Key Projects for Young Teachers at Sun Yat-sen University (grant no. 17lgzd02), and the National Administration of Surveying, Mapping and Geoinformation of China (grant no. GZIT2016-A5-147).

References

Audebert, N., Saux, B.L., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, pp. 180–196.

Baatz, M., 2000. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. Angew geogr informationsverarbeitung 12–23.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, Physica-Verlag HD. p. 177–186.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, pp. 886–893.

Davidson, W., Abramowitz, M., 2006. Molecular expressions microscopy primer: digital image processing-difference of gaussians edge enhancement algorithm. Olympus America Inc., and Florida State University.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV, Chile. p. 2650–2658.

Freund, Y., Schapire, R.E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting, European conference on computational learning theory. Springer, Barcelona, Spain, pp. 23–37.

Fukushima, K., Miyake, S., Ito, T., 1983. Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Trans. Syst. Man Cybernet. 826–834.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, pp. 580–587.

Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from RGB-D images for object detection and segmentation. In: European Conference on Computer Vision, Springer, pp. 345–360.

Höfle, B., Pfeifer, N., 2007. Correction of laser scanning intensity data: data and model-driven approaches. ISPRS J. Photogramm. Remote Sens. 62, 415–433.

Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., Bu, S., Wu, J., 2014. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. ISPRS J. Photogramm. Remote Sens. 89, 37–48.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intel. 20, 832–844.

Islam, M.A., Rochan, M., Bruce, N.D., Wang, Y., 2017. Gated feedback refinement network for dense image labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, pp. 4877–4885.

Längkvist, M., Kiselev, A., Alirezaie, M., Loutfi, A., 2016. Classification and segmentation of satellite orthoimagery using convolutional neural networks. Remote Sens. 8, 329.

Lahat, D., Adali, T., Jutten, C., 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. Proc. IEEE 103, 1449–1477.

Liao, W., Pižurica, A., Bellens, R., Gautama, S., Philips, W., 2015. Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. IEEE Geosci. Remote Sens. Lett. 12, 552–556.

Lin, G., Milan, A., Shen, C., Reid, I. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, pp. 1925–1934.

Liu, Y., Piramanayagam, S., Monteiro, S. T., Saber, E., 2017. Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fullyconvolutional neural networks and higher-order crfs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, pp. 76–85.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,CVPR, Boston, MA, USA, pp. 3431–3440.

Long, T., Jin, L., 2008. Building compact MQDF classifier for large character set recognition by subspace distribution sharing. Pattern Recogn. 41, 2916–2925.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.

Mnih, V., 2013. Machine learning for aerial image labeling. University of Toronto.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intel. 24, 971–987.

Paisitkriangkrai, S., Sherrah, J., Janney, P., van den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 2868–2881.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv preprint arXiv: 1606.02585.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. Remote Sens. Environ. 115, 2564–2577.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, pp. 2818–2826.

Vapnik, V.N., Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55 (2), 881–893.

Wang, Z., Boesch, R., Ginzler, C., 2007. Arial images and LiDAR fusion applied in forest boundary detection. In: 7th WSEAS International Conference on Signal, Speech and Image Processing. Beijing, China. pp. 130–135.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. ACM, pp. 270–279.

Zhang, J., 2010. Multi-source remote sensing data fusion: status and trends. Int. J. Image Data Fusion 1, 5–24.

Zhang, L., 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, 22–40.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, pp. 2881–2890.

Zhao, W., Du, S., 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 113, 155–165.

Zhou, H., Zhang, J., Lei, J., Li, S., Tu, D., 2016. Image semantic segmentation based on FCN-CRF model, Image, Vision and Computing (ICIVC), International Conference on. IEEE, pp. 9–14.