

# Urban Zoning Using Higher-Order Markov Random Fields on Multi-View Imagery Data

Tian Feng<sup>1,\*</sup>, Quang-Trung Truong<sup>2,\*</sup>, Duc Thanh Nguyen<sup>3</sup>, Jing Yu Koh<sup>2</sup>,  
Lap-Fai Yu<sup>4</sup>, Alexander Binder<sup>2</sup>, and Sai-Kit Yeung<sup>5</sup>

<sup>1</sup> University of New South Wales, Australia

<sup>2</sup> Singapore University of Technology and Design, Singapore

<sup>3</sup> Deakin University, Australia

<sup>4</sup> University of Massachusetts Boston, USA

<sup>5</sup> Hong Kong University of Science and Technology, Hong Kong

*\*Co-first author*

**Abstract.** Urban zoning enables various applications in land use analysis and urban planning. As cities evolve, it is important to constantly update the zoning maps of cities to reflect urban pattern changes. This paper proposes a method for automatic urban zoning using higher-order Markov random fields (HO-MRF) built on multi-view imagery data including street-view photos and top-view satellite images. In the proposed HO-MRF, top-view satellite data is segmented via a multi-scale deep convolutional neural network (MS-CNN) and used in lower-order potentials. Street-view data with geo-tagged information is augmented in higher-order potentials. Various feature types for classifying street-view images were also investigated in our work. We evaluated the proposed method on a number of famous metropolises and provided in-depth analysis on technical issues.

**Keywords:** Urban zoning, street-view images, satellite images, higher-order Markov random fields

## 1 Introduction

Urban zoning is a common practice adopted by many developed countries for urban planning [1]. The primary purpose of urban zoning is to segregate an urban area into distinct zones with regard to the use of spaces (e.g., residential, commercial, industrial), while specifying the height and bulk of structures, the lot dimensions, and open space requirements [2]. Urban planners, administrators, and policy makers rely on urban zoning maps to analyze, predict, and plan for urban development.

Conventional urban zoning approaches [2] require tremendous manual efforts; hence, they are time-consuming, prone to error, and non-scalable. Routine processes in updating a zoning map typically require several months of intensive labor work. Therefore, an automatic approach for urban zoning is highly favorable and deserves in-depth studies.

Existing efforts (e.g., [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]) have been classifying land use and land cover using remotely sensed data (i.e., satellite and aerial images). Several methods (e.g., [3], [9], [10], [11]) applied image segmentation techniques on aerial images. However, manually segmenting those images is laborious and challenging: human efforts are needed for visually interpreting

every pixel from single or multiple band(s). Hence, automatic semantic segmentation techniques (e.g., [4], [5], [6], [7], [8], [14], [16]) have also been proposed.

Thanks to the rise of social networking services (e.g., Flickr, Facebook), an enormous amount of street-view photos with geo-tagged information are publicly shared. This sort of data carry detailed semantic information about different places and thus could help to interpret zoning information. In this paper, we explore the use of street-view photos and satellite images for automatic urban zoning. Specifically, we propose an urban zoning method using multi-source data including top-view satellite images and street-view photos. This multi-source data is fused into a higher order Markov random fields (HO-MRF) model. In the model, a multi-scale deep convolutional neural network (MS-CNN) is built to segment the top-view data and used in lower-order potentials while the street-view photos are classified and added in higher-order potentials. We conducted extensive experiments to investigate various aspects of our proposed solution. In particular, we investigated different features and classifiers that could be used for classifying street-view photos. We compared the use of multi-source vs single-source data. We also compared our proposed HO-MRF model with conventional MRF and our deep neural network with existing network architectures.

It is important to note that urban zoning conceptually differs from land cover or land use despite their correlation. Land cover refers to the observed physical cover on the earth surface. Land use refers to the activities people undertake on a certain type of land cover to change or maintain it, or to produce [17]. Urban zoning, on the other hand, refers to segregating an urban area into distinct zones by the use of buildings and spaces within a zone. It provides a convenient mean to visualize patterns of social and economic developments.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Our proposed method is presented in Section 3. Datasets and experiments are described in section 4. Section 5 concludes the paper and provides remarks.

## 2 Related Work

### 2.1 Land use and Land cover Classification

Early works (e.g., [18], [19]) have successfully applied satellite sensor technology for monitoring agricultural land use, which motivate recent attempts on applying similar technologies for analyzing land use and land cover. Barnsley and Barr [20] proposed to extract land use information using land cover classification from multi-spectral images captured by a satellite sensor. However, the degree of sensitivity between land use patterns and the accuracy of initial land cover classification is yet to be determined.

Brown et al. [3] analyzed the relationship between land use and land cover from a spatial-temporal perspective using a Markov transition probability model. Porway et al. [12], [13] proposed a hierarchical and contextual model for aerial image understanding. Lienou et al. [15] annotated satellite images using semantic concepts related to land use. The annotation task combined the classification of image patches and the integration of the spatial information between these patches. Rozenstein and Karnieli [4] introduced Geographical Information Systems (GIS) to facilitate land use classification based on

hybrid supervised and unsupervised learning on remote sensing data. Hu and Wang [14] used a decision tree with remote-sensing data to classify urban land use classes. Banerjee et al. [7] applied cluster ensemble techniques to support self-training-based, unsupervised land cover classification on satellite images, to overcome the challenge of limited information in data distribution. Luus et al. [16] introduced multi-view deep learning to the field of land use classification.

There are also recent works on applying semantic segmentation techniques to land use and land cover classification. For example, Frohlich et al. [6] used iterative context forests to classify land cover from satellite images by considering contextual information. Volpi and Ferrari [9] segmented satellite images using conditional random fields. Albert et al. [11] proposed a method for the simultaneous classification of land cover and land use, taking the consideration of spatial context.

## 2.2 Urban Understanding from Street-View Photos

The abundance of street-view photos provides new opportunities for computer vision research on understanding urban areas. For example, previous works have demonstrated using such data for city identification [21], [22], geo-informative social attributes prediction [23] and urban perception [24], [25]. Recently, Dubey et al. quantified the perception of urban environment at the global scale by training a convolutional neural architecture on a new crowd-sourced dataset [26].

An early attempt was made by Leung and Newsam [27] on using street-view photos for classifying land use and land cover. To measure social development, they formulated the problem as supervised binary classification. Oba et al. [28] proposed text features in addition to visual features to improve land cover classification. Frey et al. [29] applied unsupervised learning to automatically characterize large geographical regions into location types.

## 2.3 Correlation between Top-View and Street-View Imagery Data

Recently the correlation between top-view and street-view imagery data has been exploited for scene understanding. For example, Lin et al. [30] proposed to learn features using deep networks for cross-view image matching. In this work, the geo-location of a street-view query image on an aerial image is determined via feature matching. In [31], Mátyus et al. proposed an automatic road segmentation method for vehicles using both aerial and ground-view imagery data. Specifically, ground-view images of a road are captured using a stereo camera built in vehicles and paired with aerial imagery obtained from GPS to reasoning the road surface and perform road segmentation. In [32], functions of buildings were classified using both ground-level and overhead images. Convolutional neural networks (CNNs) were used to learn visual features at both ground and overhead levels. The ground-level feature map for an input overhead image was then constructed by applying kernel regression on the ground-level features extracted by the CNNs on ground-level images.

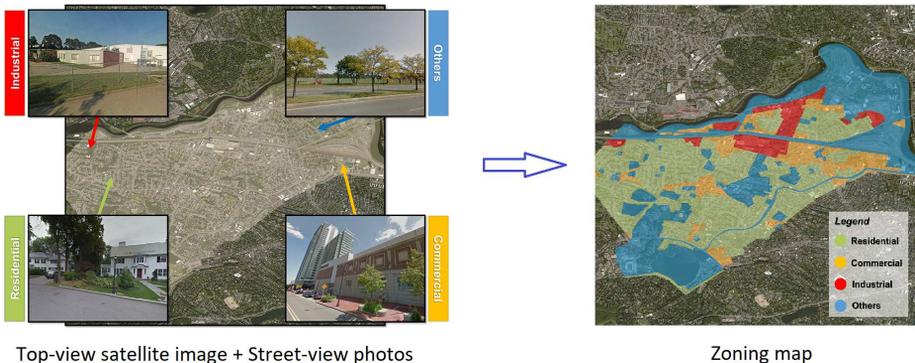


Fig. 1: Our approach infers a reliable urban zoning map from top-view satellite images and street-view photos.

### 3 Proposed Method

#### 3.1 Higher-order Markov Random Fields

The problem of urban zoning can be described as follows. Given a satellite image  $S$  covering an urban area  $U$  and a set of randomly downloaded street-view photos  $G = \{g_i\}$  located within  $U$  and associated with geo-tagged information, the problem is to infer possible zoning maps of  $U$  at metropolis-level. Fig. 1 illustrates the proposed urban zoning system.

We formulate the problem of urban zoning as segmenting the satellite image  $S$  into a number of regions, called *zones*, and identifying the zone types of the regions. The zone type of each region is determined by the visual information extracted from the pixels of that region and the associated street-view photos.

According to the definition of the uses of urban buildings and spaces [2], we categorize urban zones into 4 types: *Residential*, *Commercial*, *Industrial*, and *Others* in this paper. This categorization ensures the generality of the problem. Let  $Z$  be the set of zone types.  $|Z| = 4$  in our case. Fig. 2 shows some examples of street-view photos under different zone types.

Technically, the problem stated above can be regarded as *semantic segmentation* of the satellite image  $S$ . Following this idea, we propose a higher-order Markov random fields (HO-MRF) model to solve the problem. In our HO-MRF model, unary terms are computed from visual features extracted on the satellite image  $S$  via a deep convolutional neural network. The relationships between the satellite image  $S$  and its associated street-view photos  $G$  are encoded in higher order potentials and augmented to the HO-MRF model. Fig. 3 summarizes the workflow of this solution.

The HO-MRF model is constructed as follows. The input satellite image  $S$  is represented as a lattice of pixels. Each pixel  $\mathbf{p}_i \in S$  is considered as a node and its label is denoted as  $l_i$  taking value in  $Z$ . Akin to fully connected MRFs [33], each pixel is connected to all other pixels.



Fig. 2: Examples of street-view photos from New York, San Francisco, and Boston.

The zoning problem is equivalent to finding the best configuration  $\mathcal{L} = (l_1, l_2, \dots, l_{|S|})$  for  $|S|$  pixels of the satellite image  $S$ . In particular, we minimize the energy function,

$$\mathcal{L}^* = \arg \min_{\mathcal{L} \in \mathbf{L}^{|S|}} \left[ \sum_{i \in S} \psi_i(l_i) + \sum_{(i,j), i < j} \psi_{i,j}(l_i, l_j) + \sum_{i \in S} \varphi(l_i, G) + \sum_{g \in G} \varphi(g) \right], \quad (1)$$

The unary potential  $\psi_i(l_i)$  in (1) is defined as,

$$\psi_i(l_i = z) \propto -\log C(S, \mathbf{p}_i | l_i = z), \quad (2)$$

where  $C$  is the classification scores of assigning the zone type of pixel  $\mathbf{p}_i$  (i.e.,  $l_i$ ) to  $z$  based on the input satellite image  $S$ . The computation of  $C$  will be presented in details in Section 3.2.

The pairwise potential  $\psi_{i,j}(l_i, l_j)$  is defined as a mixture of Gaussians of the location and color information of image pixels on  $S$ . In particular,

$$\psi_{i,j}(l_i, l_j) = \mu_{ij} \left[ \exp \left( -\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{2\alpha^2} - \frac{\|\mathbf{c}_i - \mathbf{c}_j\|^2}{2\beta^2} \right) + \exp \left( -\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{2\gamma^2} \right) \right] \quad (3)$$

where  $\mathbf{c}_i/\mathbf{c}_j$  is the color vector at pixel  $\mathbf{p}_i/\mathbf{p}_j$  and  $\mathbf{p}_i$  is the location vector (i.e.,  $x$ - and  $y$ -coordinate) of  $\mathbf{p}_i$ ,  $\mu_{ij}$  is the Pott label compatibility function [33], e.g.,

$$\mu_{ij} = \begin{cases} -1, & \text{if } l_i = l_j \\ 1, & \text{otherwise.} \end{cases} \quad (4)$$

In the HO-MRF model, we introduce higher-order potentials (e.g.,  $\varphi(g)$ ) capturing the relationships between  $S$  and  $G$ . The term  $\varphi(l_i, G)$  encodes the zone consistency

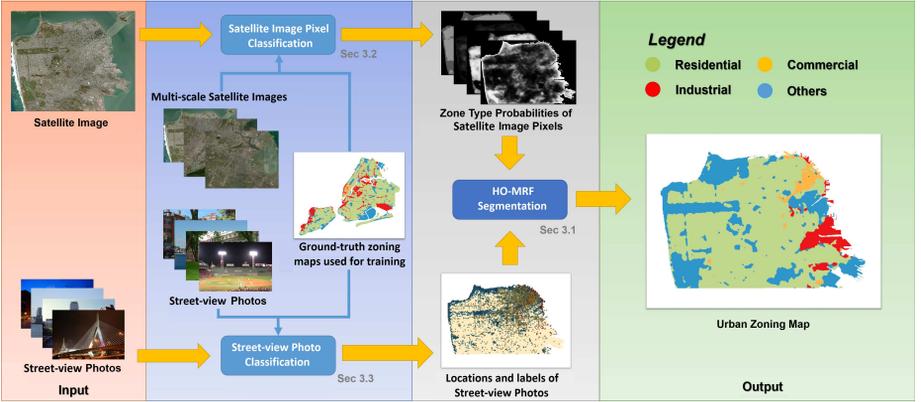


Fig. 3: The workflow of our approach. From left to right: Input data (1st column), Satellite image pixel classification and street-view photo classification (2nd column), zoning using HO-MRF (3rd column), output (4th column).

between a point  $\mathbf{p}_i$  on  $S$  and its nearest street-view photo in  $G$ . Note that since every street-view photo is associated with a geo-location, the distance between a pixel  $\mathbf{p}_i \in S$  to a street-view photo can be determined. In particular, we define,

$$\varphi(l_i, G) \propto \min_{g \in G, f(g)=l_i} \left\{ -\log \left[ \frac{1}{d(\mathbf{p}_i, g)} \right] \right\}, \quad (5)$$

where  $f(g)$  is a function returning the zone type of  $g$  and is described in Section 3.3;  $d(\mathbf{p}_i, g)$  is the spatial distance between  $\mathbf{p}_i$  and  $g$ . Intuitively,  $\varphi(l_i, G)$  is inverse to the distance from  $\mathbf{p}_i$  to its closest street-view photo whose zone type is  $l_i$ . In other words, the zone type of  $\mathbf{p}_i$  would be more biased by its nearest street-view photo.

Note that  $\varphi(l_i, G)$  needs to be computed for every  $\mathbf{p}_i$ . To save the computational cost, the Distance Transform proposed in [34] is applied on grids formed by the locations of street-view photos  $G$ . In particular, the zone type of each street-view photo is first obtained (see Section 3.3). For every zone type  $z \in Z$ , a Distance Transform  $D_z$  is applied on the street-view photos that have been classified as zone type  $z$ . The potential  $\varphi(l_i, G)$  can then be rewritten as,

$$\varphi(l_i, G) \propto -\log \frac{1}{D_{l_i}(\mathbf{p}_i)}, \quad (6)$$

where  $D_{l_i}(\mathbf{p}_i)$  is the value of the Distance Transform  $D_{l_i}$  at location  $\mathbf{p}_i$ .

The term  $\varphi(g)$  represents the zone consistency of pixels in a local image region (on  $S$ ), at which the street-view photo  $g$  could be captured. Specifically, given  $g$ , its geo-location on  $S$  can be obtained and, at this geo-location, a local image region  $R(g)$  of size  $W \times W$  is extracted. In our implementation,  $W$  is set to 46, which is also the size of image patches used in classifying pixels on the satellite image  $S$  (see Section 3.2). We then construct a probability distribution  $P_g(l_k)$  over the labels  $l_k \in \mathcal{L}$  conditioned

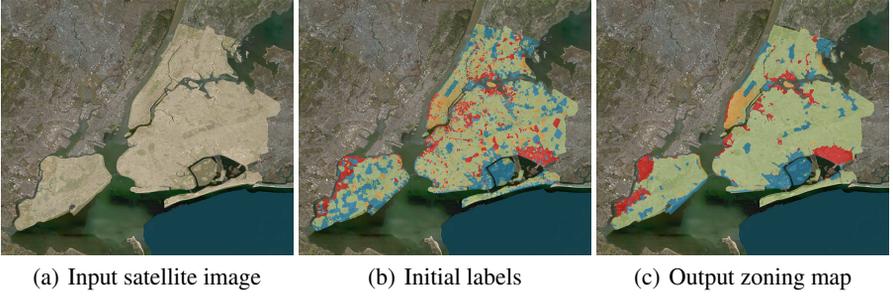


Fig. 4: Zoning of New York City. Initial label of each pixel is determined by the zone type of its nearest street-view photo.

on pixels  $k$  inside  $R(g)$ . The cost  $\varphi(g)$  is then computed from the entropy of  $P_g(l_k)$  as,

$$\varphi(g) \propto - \sum_{l_k \in \mathcal{L} | \mathbf{p}_k \in R(g)} P_g(l_k) \log P_g(l_k) \quad (7)$$

The optimization problem in (1) can be solved using the variational mean field method [35], [36]. In particular, (1) is equivalent to finding a maximum of a posteriori (MAP)  $p(L|S = \{\mathbf{p}_i\})$ . In variational mean field, this can be approximated by a variational distribution  $Q$  which is fully factorized, i.e.  $Q(L) = \prod_i Q_i(l_i)$ . The solution of (1) is finally achieved by iteratively updating  $Q_i(l_i)$  as follows,

$$Q_i(l_i = z) \leftarrow \frac{1}{Z_i} \exp \left[ - \psi_i(l_i = z) - \sum_{z' \in Z} \sum_{j \neq i} Q_j(l_j = z') \psi_{ij}(l_i, l_j) - \varphi(l_i = z, G) - \sum_{g \in G} \sum_{\{l_j | \mathbf{p}_j \in R(g), l_i = z\}} \mathbf{Q}(R(g) - \mathbf{p}_i) \varphi(g) \right] \quad (8)$$

where  $R(g)$  is an image patch on  $S$  that is centered at the location of  $g$ ,  $\mathbf{Q}$  is the variational distribution of the higher order terms [37],  $R(g) - \mathbf{p}_i$  is the set of pixels in  $R(g)$  except  $\mathbf{p}_i$ , and  $Z_i$  is the partition function. We compute the higher order term  $\sum_{\{l_j | \mathbf{p}_j \in R(g), l_i = z\}} \mathbf{Q}(R(g) - \mathbf{p}_i) \varphi(g)$  as,

$$\begin{aligned} & \sum_{\{l_j | \mathbf{p}_j \in R(g), l_i = z\}} \mathbf{Q}(R(g) - \mathbf{p}_i) \varphi(g) = \\ & - \sum_{l_k \in \mathcal{L} | \mathbf{p}_k \in R(g)} P_g(l_k) \log P_g(l_k) \left[ \prod_{\substack{\mathbf{p}_m \in R(g), \\ m \neq i}} Q_m(l_m = z) \right] \\ & - \sum_{l_k \in \mathcal{L} | \mathbf{p}_k \in R(g)} P_g(l_k) \log P_g(l_k) \left[ 1 - \prod_{\substack{\mathbf{p}_m \in R(g), \\ m \neq i}} Q_m(l_m = z) \right] \end{aligned} \quad (9)$$

In our implementation, the label of each pixel on the satellite image  $S$  was initialized by the label of its nearest street-view photo. Fig. 4 shows an example of zone segmentation of New York city.

### 3.2 Classifying Satellite Image Pixels

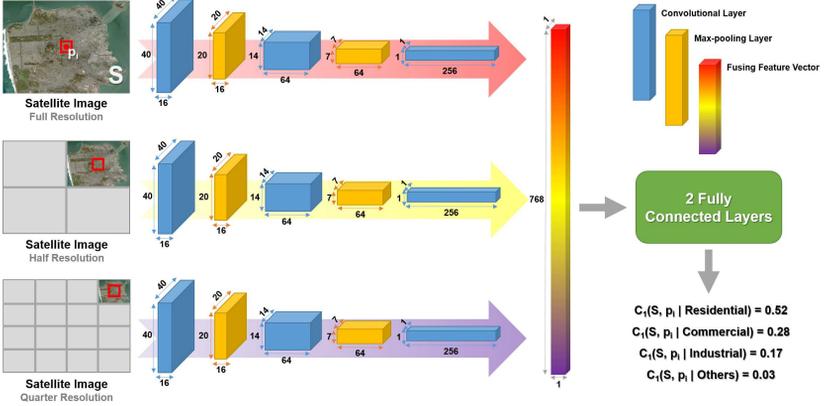


Fig. 5: MS-CNN for classifying satellite image pixels.

This section describes the classification of the zone type of pixels  $\mathbf{p}_i$  on the satellite image  $S$ , i.e., estimation of  $C(S, \mathbf{p}_i | l_i)$  in (2). Like the work by Farabet et al. [38], our network receives input the YUV image of  $S$ . A pyramid including three scales ( $S^0 = S$ ,  $S^1 = \frac{1}{2}S$ ,  $S^2 = \frac{1}{2}S^1$ ) is then created. For each pixel  $\mathbf{p}_i$ , three local image patches  $I_i^0$ ,  $I_i^1$ , and  $I_i^2$  of size  $46 \times 46$ -pixels centered at  $\mathbf{p}_i$  on  $S^0$ ,  $S^1$ , and  $S^2$  are extracted.

A multi-scale convolutional neural network (MS-CNN) is then constructed to learn the appearance features and classify local image patches at various scales. Utilizing the multi-scale approach would enable learning scale-invariant features. In addition, multi-scale regions at a pixel would allow the incorporation of context information and thus could be useful for the classification. In the MS-CNN, for each scale, a 5-layer sub-CNN is formed with the interchange of convolutional and max-pooling layers. For example, the 1st and 3rd layers are obtained from the banks of  $16 \times 7 \times 7$  filters, of which 10 filters connect to the Y channel and the other 6 filters connect to the U and V channels. The 2nd and 4th are results of  $16 \times 2 \times 2$  and  $64 \times 2 \times 2$  max-pooling operations respectively. The 5th layer is formed by  $256 \times 7 \times 7$  filters. The outputs of all sub-CNNs are then concatenated to a layer fed to a 2 fully connected layer structure for classification. The first layer of this fully connected network includes 1024 nodes and the second layer contains 4 nodes corresponding to 4 different zone types. Fig. 5 illustrates the MS-CNN. In the network, *softmax* is used as the activation function for the last layer of the sub-CNNs and the last layer of the fully connected network, and *tanh* is used for all other layers.

The MS-CNN is trained via optimizing a cross-entropy loss and using stochastic gradient descent method. The batch size is set to 100 and the learning rate is set to

$10^{-3}$ . We found that these settings achieved the best performance. Since the training data may be biased, the loss function is weighted relatively to the proportion of the class labels in the training dataset. In addition, all sub-CNNs share the same parameters including weight and bias. As indicated by Farabet et al. [38], imposing complete weight sharing across scales is a natural way of forcing the network to learn scale-invariant features, and at the same time, to reduce the chances of over-fitting. Given the MS-CNN,  $C(S, \mathbf{p}_i | l_i = z)$  is computed as classification score of pixel  $\mathbf{p}_i$  to the zone type  $z$ , i.e., the response of the MS-CNN at the zone type  $z$ .

### 3.3 Classifying Street-View Photos

Recall that in (5), the HO-MRF requires a function  $f(g)$  that returns the zone type of a street-view photo  $g$ . Intuitively, if the geo-location of each photo could be paired with the corresponding zone type, we would have a reasonable zoning map. To recognize street-view photos, features need to be determined. Well-known features such as GIST [39], HOG [40], local binary patterns (LBP) [41] have been proposed in the literature. However, those features are handcrafted and thus require the specific knowledge in particular fields. Recently, deep neural networks have been used for automatically learning features and training classifiers [42].

Inspired by the relevance of the information conveyed by deep networks for scene recognition, we choose the *Places-CNN*, a CNN trained on Places Database [43]. Places-CNN is able to recognize 205 object types and scenes from daily taken photos. We note that the object types and scenes of interest in the Places-CNN (e.g., buildings, skyscrapers, warehouses) are semantically close to the basic concepts of urban zoning (e.g., commercial, industrial). In our implementation, a street-view photo  $g$  is passed to the Places-CNN and the output is a vector  $\mathbf{x}(g) = \{x(g)_1, \dots, x(g)_{205}\}$ , representing how probably the photo  $g$  contains 205 types of objects and scenes.

Given the features, different classifiers could be used. However, different types of features make different impacts on the performance of a classifier. In practice, extensive experiments are often conducted for classifier selection. In this paper, we consider Random Forests (RF) [44] as the classifier for a number of reasons. First, RF is known for its capability of effectively exploiting and integrating multiple classifiers, where each classifier has its own favor on particular feature types and label classes. Second, as proven in our experimental results, compared with other classifiers, RF based on the features generated by the Places-CNN works well on our problem.

Our RF classifier  $f$  includes multiple weak classifiers [45]. The predicted label of  $f$  given an input photo  $g$  is a weighted majority vote of the predictions made by individual classifiers. Specifically, let  $h_1, \dots, h_M$  be  $M$  weak classifiers. The predicted label of a photo  $g$  is denoted as  $f(g)$  and can be computed as,

$$f(g) = \arg \max_{z \in Z} \left( \sum_{k=1}^M w_k I(h_k(\mathbf{x}(g)) = z) \right), \quad (10)$$

where  $\mathbf{x}(g)$  is the 205-dimensional feature vector extracted using the Places-CNN,  $I(\cdot)$  is the indicator function, and  $w_k$  is the weight of the  $k$ -th weak classifier  $h_k$ .

Table 1: Numbers of street-view photos.

	Residential	Commercial	Industrial	Others	Total
<b>New York</b>	59,906	77,193	25,109	56,570	218,778
<b>San Francisco</b>	49,543	46,454	11,553	41,140	148,690
<b>Boston</b>	32,412	37,588	10,179	42,682	122,861

Table 2: Proportion of zones in ground-truth zoning maps.

	Residential	Commercial	Industrial	Others
<b>New York</b>	64.21%	4.90%	14.15%	16.74%
<b>San Francisco</b>	53.82%	8.88%	6.49%	30.81%
<b>Boston</b>	40.91%	12.07%	3.42%	43.60%

## 4 Experiments

### 4.1 Dataset

Our experiments were conducted on the map data of three metropolises: **New York**, **San Francisco** and **Boston**.

*Satellite Data.* Satellite images are our top-view imagery data that could be obtained without expensive devices or professional expertise. We downloaded the satellite image tiles for each city by its corresponding geographical boundary from the National Map service operated by U.S. Geology Survey [46].

*Street-View Data.* Popular social network services provide public access to many of their street-view photos and associated geo-tagged information. We collected 490,329 photos of the three metropolises from Flickr and Panoramio. Specifically, we queried the URLs of outdoor photos by calling the APIs of Flickr and Panoramio within the geographical boundary (the minimum bounding box) of each city. Photos within each city were downloaded after their coordinates were verified by GIS. Table 1 summarizes the number of street-view photos used for each metropolis. Fig. 6 shows the distribution of street-view photos in New York.

*Urban Zoning Maps.* We collected urban zoning maps from the websites of local governments, which were stored in the standard GIS format, SHAPEFILE [47]. The zone types include: *Residential*, *Commercial*, *Industrial* and *Others*. The mentioned maps served as the ground-truth for evaluations. Table 2 shows the percentage of each type of zones in the ground truth zoning maps.

### 4.2 Classifying Street-View Photos

We first evaluated street-view photo classification. In this experiment, we investigated different feature types and classifiers that are well-known for scene classification. Specifically, for features, we evaluated the GIST [39], HOG [40], and Places-CNN [43].

For classifiers, we compared RBF-kernel SVM,  $k$ -nearest neighbors, and RF. We found that RBF-kernel SVM significantly outperformed linear SVM. For the  $k$ -nearest neighbors, we set  $k$  to 10 which gave the best performance on this technique. To measure the similarity between samples, we used the Euclidean distance. For the RF, 15 trees were used. All the classifiers were evaluated using 3-fold cross validation.

Table 3 summarizes the results of this experiment. As shown, in most cases, the combination of Places-CNN with RF achieves the best performance. It is also worthwhile to notice that compared with handcrafted features, the features learned by the Places-CNN often gain higher classification accuracy and achieve the best overall performance irrespective of the classifiers.

We also evaluated the use of neural network for street-view photo classification. Specifically, we directly connected the outputs of Places-CNN to a fully connected neural network (with one hidden layer and 4 outputs corresponding to 4 zone types). The results of this experiment are reported in Table 4. As shown in both Table 3 and Table 4, applying RF on-top of Places-CNN achieves the best overall performance.

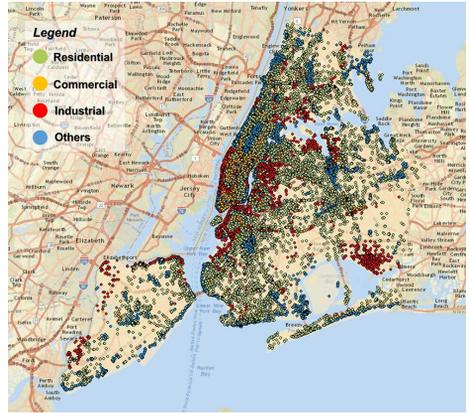


Fig. 6: The plot map of street-view photos in New York city. Each photo is marked with a color corresponding to its zone type.

Table 3: Accuracy of street-view photo classification.

	Random Forest			SVM with RBF kernel			k-Nearest Neighbors		
	Places-CNN	GIST	HOG	Places-CNN	GIST	HOG	Places-CNN	GIST	HOG
<b>New York</b>									
Residential	<b>79.04%</b>	70.56%	71.75%	71.32%	64.12%	65.33%	65.21%	60.59%	61.44%
Commercial	<b>81.65%</b>	65.19%	69.22%	73.74%	62.90%	64.78%	67.31%	57.57%	62.16%
Industrial	71.75%	66.21%	<b>72.80%</b>	68.46%	54.31%	62.10%	61.43%	57.53%	54.12%
Others	<b>81.26%</b>	67.78%	69.13%	72.97%	68.49%	66.31%	67.06%	61.53%	58.73%
Overall	<b>79.76%</b>	67.51%	70.35%	72.33%	62.32%	65.07%	66.05%	59.47%	60.20%
<b>San Francisco</b>									
Residential	<b>78.19%</b>	69.25%	70.67%	74.19%	67.39%	68.11%	73.33%	65.38%	67.14%
Commercial	<b>82.55%</b>	72.54%	72.34%	71.52%	68.10%	67.55%	65.71%	60.44%	59.76%
Industrial	70.88%	65.81%	<b>70.96%</b>	73.73%	63.54%	68.91%	64.32%	63.92%	65.88%
Others	<b>86.18%</b>	80.55%	83.48%	76.26%	74.62%	76.39%	69.09%	66.79%	68.42%
Overall	<b>81.19%</b>	73.13%	74.76%	73.89%	69.31%	70.29%	69.07%	64.11%	65.10%
<b>Boston</b>									
Residential	<b>79.88%</b>	69.18%	71.57%	75.08%	64.11%	70.02%	72.57%	66.53%	67.82%
Commercial	<b>82.25%</b>	71.20%	70.88%	70.26%	68.29%	69.23%	68.45%	65.78%	65.93%
Industrial	<b>86.45%</b>	70.09%	69.47%	74.31%	67.54%	65.33%	69.13%	66.91%	65.02%
Others	<b>82.39%</b>	72.34%	73.59%	77.29%	69.51%	74.87%	73.29%	67.52%	70.19%
Overall	<b>82.08%</b>	70.96%	71.77%	74.01%	67.58%	70.22%	71.21%	66.60%	67.67%

Table 4: Accuracy of street-view photo classification using Places-CNN and fully connected neural network.

	Residential	Commercial	Industrial	Others	Overall
<b>New York</b>	79.97%	44.22%	27.89%	66.11%	60.89%
<b>San Francisco</b>	74.01%	45.69%	24.84%	78.43%	68.13%
<b>Boston</b>	58.24%	25.82%	9.41%	80.95%	65.38%

### 4.3 Zoning

Since zoning is formulated as semantic segmentation of satellite images, its performance can be evaluated via the segmentation accuracy, i.e., the accuracy of classifying of satellite image pixels using the proposed MS-CNN. In our implementation, the MS-CNN was trained/tested in 3-fold cross fashion (i.e. two cities were used for training while the other was used for testing). To increase the data for training/testing the MS-CNN, for each city, 6 satellite images were created, 5 of which were obtained by rotating the original one by a multiple of  $60^\circ$ . Each created satellite image was also scaled (with 3 different scales) and densely scanned by a  $46 \times 46$ -pixel window. The process resulted in 1,265,572 satellite image windows at each scale. The performance of zoning across cities is reported in Table 5 (the last column). Zoning results are presented in Fig. 7.

Table 5 also compares the use of single data source, e.g. satellite data only or street-view photos only, with multi-source data (i.e., our proposed method). Note that the solely use of satellite data is equivalent to the 1-st term in our energy formulation in (1). When only street-view photos are used, the zone type of each pixel on the satellite image is decided by the zone type of its closest street-view photo (i.e., only the 3-rd term in (1) is used).

Recall that we propose the HO-MRF to perform zone segmentation in which a deep neural network is used to obtain low-level MRF segmentation and street-view photos are augmented via higher order potentials (see (1)). Therefore, we compare the HO-MRF with the conventional MRF (i.e., using only the first two terms in (1)). As shown in Table 5, the use of satellite data outperformed that of street-view photos and our proposed HO-MRF with the combination of both the data sources achieved the best performance on all the cities. Specifically, the combination of both sources improved up to 10.17% and 7.12% compared with the single use of top-view data or street-view data respectively. The HO-MRF boosted up 8.76% compared with conventional MRF.

We also compared our deep network (i.e., MS-CNN) used to compute the unary terms in (1) with the network recently proposed by Volpi and Tuia [10]. Since the

Table 5: Zoning performance.

	Satellite pixels	Street-view photos	Conventional MRF	HO-MRF
<b>New York</b>	66.70%	60.50%	66.87%	<b>70.33%</b>
<b>San Francisco</b>	64.68%	61.63%	62.99%	<b>71.80%</b>
<b>Boston</b>	51.33%	48.08%	47.95%	<b>57.09%</b>

Table 6: Comparison of different networks in classifying satellite image pixels.

	New York	San Francisco	Boston
<b>Volpi and Tuia[10]</b>	20.50%	10.10%	22.50%
<b>Our MS-CNN</b>	<b>66.70%</b>	<b>64.68%</b>	<b>51.33%</b>

method in [10] requires the digital surface model for each satellite image, that is unavailable in our data, we adapted the network in [10] by changing its input and output layers to meet with our data while maintaining other layers and settings (e.g. activation functions) same. The adapted network was then re-trained and tested on the same data with our network. As shown in Table 6, our network significantly outperformed the one proposed in [10].

#### 4.4 Discussion

We have found that larger improvement with the HO-MRF was obtained on Boston and San Francisco compared to only using satellite imagery as opposed to New York probably because New York contains more industrial regions (e.g., as shown in Table 2 industrial takes 14.15% in New York, compared with 6.49% and 3.42% in San Francisco and Boston). However, as shown in Table 3, industrial street-view photos are recognized with lower accuracy compared with other zone types.

Experimental results (e.g., Fig. 7) also show our method fails to segment tiny/thin regions. This is because those regions occupy small portions in local image windows and thus are biased by nearby larger regions. We have also found that the method may fail at regions with less street-view images captured. Note that street-view photos are captured sparsely and non-uniformly (see Fig. 6). We believe that, 3D data such as digital surface models, digital elevation models (if available or achievable approximately) would be useful to resolve these issues.

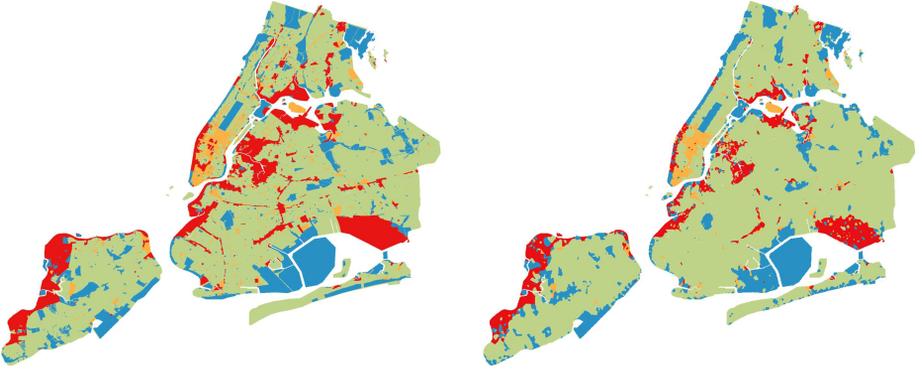
## 5 Conclusion

We proposed a higher-order Markov random fields model for urban zoning based on multi-view imagery data including top-view satellite and street-view images. We also developed a multi-scale deep convolutional neural network used for classifying satellite image pixels. By integrating different sources of imagery data, our approach can achieve urban zoning automatically, and hence overcome the scalability bottleneck faced by the conventional practice of creating zoning maps manually. We investigated various implementation strategies including feature types, classification models, deep architectures, and conducted extensive experiments and comparisons to verify our approach.

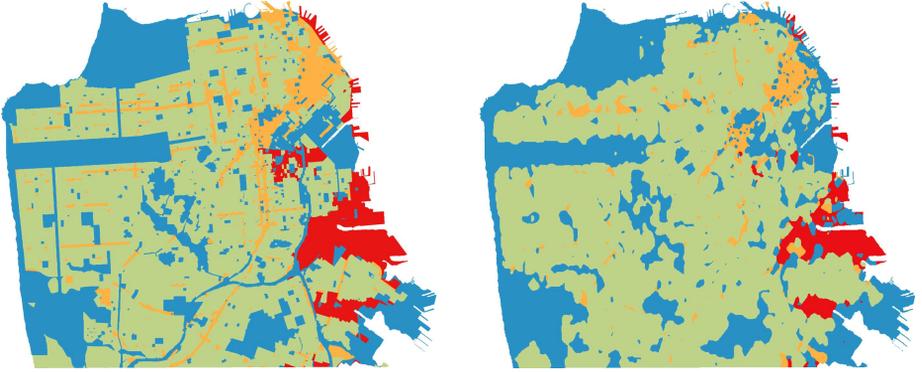
## 6 Acknowledgement

Quang-Trung Truong was supported by Singapore MOE Academic Research Fund MOE2016-T2-2-154. Tian Feng was supported by the UNSW Art & Design Faculty Research Grant PS49003. Lap-Fai Yu was supported in part by the National Science Foundation under award number 1565978. Alexander Binder was supported by the SUTD grant SRIS15105.

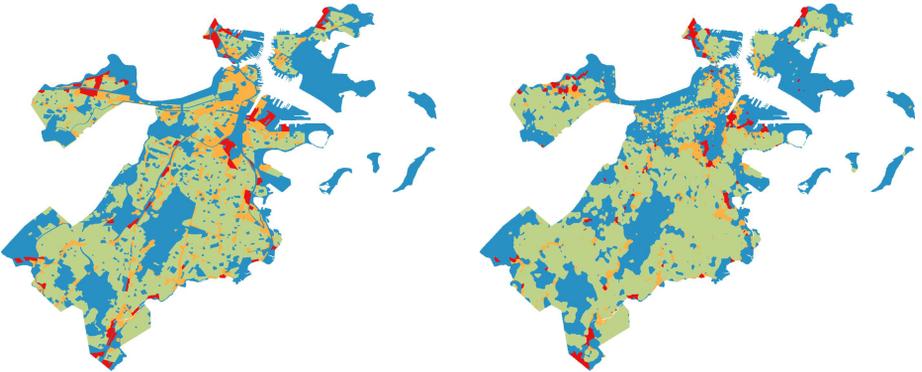
This research was partially funded by the Heritage Research Grant of the National Heritage Board, Singapore NRF under its IDM Futures Funding Initiative and Virtual Singapore Award No. NRF2015VSGAA3DCM001-014, and by an internal grant from HKUST (R9429).



(a) New York



(b) San Francisco



(c) Boston

Fig. 7: Zoning results. For each city, the ground-truth is shown on the left and the zoning map result is shown on the right.

## References

1. Whitnall, G.: History of zoning. *Annals of the American Academy of Political and Social Science* **155** (1931) 1–14
2. GoLdberg, M., Horwood, P., Block, W.: *Zoning: Its Costs and Relevance for the 1980s*. Fraser Institute's housing and land economics series. Fraser Institute (1980)
3. Brown, D., Pijanowski, B., Duh, J.: Modeling the relationships between land use and land cover on private lands in the upper midwest, usa. *Journal of Environmental Management* **59**(4) (2000) 247–263
4. Rozenstein, O., Karnieli, A.: Comparison of methods for land-use classification incorporating remote sensing and gis inputs. *Applied Geography* **31**(2) (2011) 533–544
5. Tuia, D., Muñoz-Marí, J., Kanevski, M., Camps-Valls, G.: Structured output svm for remote sensing image classification. *Journal of Signal Processing Systems* **65**(3) (2011) 301–310
6. Frohlich, B., Bach, E., Walde, I., Hese, S., Schmullius, C., Denzler, J.: Land cover classification of satellite images using contextual information. *Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* (1) (2013) 1–6
7. Banerjee, B., Bovolo, F., Bhattacharya, A., Bruzzone, L., Chaudhuri, S., Mohan, B.: A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy. *IEEE Geoscience and Remote Sensing Letters* **12**(4) (2015) 741–745
8. Tokarczyk, P., Wegner, J., Walk, S., Schindler, K.: Features, color spaces, and boosting: New insights on semantic classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **53**(1) (2015) 280–295
9. Volpi, M., Ferrari, V.: Semantic segmentation of urban scenes by learning local class interactions. In: *Proc IEEE Computer Vision and Pattern Recognition Workshops*. (2015)
10. Volpi, M., Tuia, D.: Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* **55**(2) (2017) 881–893
11. Albert, L., Rottensteiner, F., Heipke, C.: An Iterative Inference Procedure Applying Conditional Random Fields for Simultaneous Classification of Land Cover and Land Use. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* (August 2015) 369–376
12. Porway, J., Wang, K., Yao, B., Zhu, S.C.: A hierarchical and contextual model for aerial image understanding. In: *Proc IEEE International Conference on Computer Vision and Pattern Recognition*. (2008) 1–8
13. Porway, J., Wang, Q., Zhu, S.C.: A hierarchical and contextual model for aerial image parsing. *International Journal of Computer Vision* **88**(2) (2010) 254–283
14. Hu, S., Wang, L.: Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing* **34**(3) (2013) 790–803
15. Lienou, M., Maitre, H., Datcu, M.: Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation. *IEEE Geoscience and Remote Sensing Letters* **7** (January 2010) 28–32
16. Luus, F.P.S., Salmon, B.P., van den Bergh, F., Maharaj, B.T.J.: Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters* **12**(12) (Dec 2015) 2448–2452
17. Food and Agriculture Organization of the United Nations: *Land Cover Classification System (LCCS): classification concepts and user manual*. (2000)
18. Forster, B.C.: An examination of some problems and solutions in monitoring urban areas from satellite platforms. *International Journal of Remote Sensing* **6**(1) (1985) 139–151
19. Gong, P., Howarth, P.J.: Land-use classification of spot hrv data using a cover-frequency method. *International Journal of Remote Sensing* **13**(8) (1991)

20. Barnsley, M.J., Barr, S.L.: Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. *Photogrammetric Engineering and Remote Sensing* **62**(8) (1996) 949–958
21. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? *ACM Transactions on Graphics* **31**(4) (2012) 101:1–101:9
22. Zhou, B., Liu, L., Oliva, A., Torralba, A.: Recognizing city identity via attribute analysis of geo-tagged images. In: *Proc European Conference on Computer Vision*. (2014) 519–534
23. Lee, S., Zhang, H., Crandall, D.J.: Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In: *Proc IEEE International Winter Conference on Applications of Computer Vision*. (2015) 550–557
24. Ordonez, V., Berg, T.L.: Learning high-level judgments of urban perception. In: *Proc European Conference on Computer Vision*. (2014) 494–510
25. Quercia, D., O’Hare, N.K., Cramer, H.: Aesthetic capital: What makes london look beautiful, quiet, and happy? In: *Proc 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, New York, NY, USA, ACM (2014) 945–955
26. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A. In: *Deep Learning the City: Quantifying Urban Perception at a Global Scale*. Springer International Publishing, Cham (2016) 196–212
27. Leung, D., Newsam, S.: Proximate sensing: Inferring what-is-where from georeferenced photo collections. In: *Proc IEEE International Conference on Computer Vision and Pattern Recognition*. (2010) 2955–2962
28. Oba, H., Hirota, M., Chbeir, R., Ishikawa, H., Yokoyama, S.: Towards better land cover classification using geo-tagged photographs. In: *Proc IEEE International Symposium on Multimedia*. (2014) 320–327
29. Frey, N., Torralba, A., Stauffer, C.: Unsupervised non-parametric geospatial modeling from ground imagery. In: *Proc IEEE International Winter Conference on Applications of Computer Vision*. (2014) 698–705
30. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: *Proc IEEE International Conference on Computer Vision and Pattern Recognition*. (2015) 5007–5015
31. Mátyus, G., Wang, S., Fidler, S., Urtasun, R.: Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In: *Proc IEEE International Conference on Computer Vision and Pattern Recognition*. (2016) 3611–3619
32. Workman, S., Zhai, M., Crandall, D.J., Jacobs, N.: A unified model for near and remote sensing. In: *Proc IEEE International Conference on Computer Vision*. (2017) 2688–2697
33. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: *Proc Conference on Neural Information Processing Systems*. (2011) 109–117
34. Felzenszwalb, P.F., Huttenlocher, D.P.: Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science (2004)
35. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.: An introduction to variational methods for graphical models. *Machine Learning* (1999) 183–233
36. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**(1-2) (2008) 1–305
37. Vineet, V., Warrell, J., Torr, P.H.S.: Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *International Journal of Computer Vision* **110**(3) (2014) 290–307
38. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8) (2013) 1915–1929
39. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* **42**(3) (2001) 145–175

40. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc IEEE International Conference on Computer Vision and Pattern Recognition. (2005) 886–893
41. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition* **29**(1) (1996) 51–59
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc Conference on Neural Information Processing Systems. (2012) 1106–1114
43. Zhou, B., Lapedriza, À., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proc Conference on Neural Information Processing Systems. (2014) 487–495
44. Breiman, L.: Random forests. *Machine Learning* **45**(1) (2001) 5–32
45. Kuncheva, L.I.: *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience (2004)
46. USGS: The national map. <https://viewer.nationalmap.gov/basic/>
47. Esri Inc.: ESRI Shapefile Technical Description. Environmental Systems Research Institute, Inc. (July 1998)