Joint Deep Learning for land cover and land use classification

Ce Zhang [a, *], Isabel Sargent [b], Xin Pan [c, d], Huapeng Li [d], Andy Gardiner [b], Jonathon Hare [e], Peter M. Atkinson [a, *]

[a] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK; [b] Ordnance Survey, Adanac Drive, Southampton SO16 0AS, UK; [c] School of Computer Technology and Engineering, Changchun Institute of Technology, 130021 Changchun, China; [d] Northeast Institute of Geography and Agroecology, Chinese Academic of Science, Changchun 130102, China; [e] Electronics and Computer Science (ECS), University of Southampton, Southampton SO17 1BJ, UK

Abstract Land cover (LC) and land use (LU) have commonly been classified separately from remotely sensed imagery, without considering the intrinsically hierarchical and nested relationships between them. In this paper, for the first time, a highly novel joint deep learning framework is proposed and demonstrated for LC and LU classification. The proposed Joint Deep Learning (JDL) model incorporates a multilayer perceptron (MLP) and convolutional neural network (CNN), and is implemented via a Markov process involving iterative updating. In the JDL, LU classification conducted by the CNN is made conditional upon the LC probabilities predicted by the MLP. In turn, those LU probabilities together with the original imagery are re-used as inputs to the MLP to strengthen the spatial and spectral feature representations. This process of updating the MLP and CNN forms a joint distribution, where both LC and LU are classified simultaneously through iteration. The proposed JDL method provides a general framework within which the pixel-based MLP and the patch-based CNN provide mutually complementary information to each other, such that both are refined in the classification process through iteration. Given the well-known complexities associated with the classification of very fine spatial resolution (VFSR) imagery, the effectiveness of the proposed JDL was tested on aerial photography of two large urban and suburban areas in Great Britain (Southampton and Manchester). The JDL consistently demonstrated greatly increased accuracies with increasing iteration, not only for the LU classification, but for both the LC *and* LU classifications, achieving by far the greatest accuracies for each at around 10 iterations. The average overall classification accuracies were 90.18% for LC and 87.92% for LU for the two study sites, far higher than the initial accuracies and consistently outperforming benchmark comparators (three each for LC and LU classification). This research, thus, represents the first attempt to unify the remote sensing classification of LC (state; what is there?) and LU (function; what is going on there?), where previously each had been considered separately only. It, thus, has the potential to transform the way that LC and LU classification is undertaken in future. Moreover, it paves the way to address effectively the complex tasks of classifying LC and LU from VFSR remotely sensed imagery via joint reinforcement, and in an automatic manner.

Keywords: multilayer perceptron; convolutional neural network; land cover and land use classification; VFSR remotely sensed imagery; object-based CNN

1. Introduction

Land cover and land use (LULC) information is essential for a variety of geospatial applications, such as urban planning, regional administration, and environmental management (Liu et al., 2017). It also serves as the basis for understanding the constant changes on the surface of the Earth and associated socio-ecological interactions (Cassidy et al., 2010; Patino and Duque, 2013). Commensurate with the rapid development in sensor technologies, a huge amount of very fine spatial resolution (VFSR) remotely sensed imagery is now commercially available, opening new opportunities for LULC information extraction at a very detailed level (Pesaresi et al., 2013; Zhao et al., 2016). However, classifying land cover (LC) from VFSR images remains a difficult task, due to the spectral and spatial complexity of the imagery. Land use (LU) classification is even more challenging due to the indirect relationship between LU patterns and the spectral responses recorded in images. This is further complicated by the heterogeneity presented in urban and suburban landscapes as patterns of high-level semantic functions, in which some identical low-level ground features or LC classes are frequently shared amongst different LU categories (C. Zhang et al., 2018c). This complexity and diversity in LU characteristics cause huge gaps between identifiable low-level features and the desired high-level functional representations with semantic meaning.

Over the past decade, tremendous effort has been made in developing automatic LU and LC classification methods using VFSR remotely sensed imagery. For LC, traditional classification approaches can broadly be divided into pixel-based and object-based methods depending on the basic processing units, either per-pixel or per-object (Salehi et al., 2012). Pixel-based methods are used widely to classify individual pixels into particular LC categories based purely on spectral reflectance, without considering neighbouring pixels (Verburg et al., 2011). These methods often have limited classification accuracy due to speckle noise and increased inter-class variance compared with coarse or medium resolution remotely sensed imagery. To overcome the weakness of pixel-based approaches, some post-classification approaches have been introduced (e.g. Hester et al., 2008; McRoberts, 2013). However, these techniques may eliminate small objects of a few pixels such as houses or small areas of vegetation. Object-based methods, under the framework of object-based image analysis (OBIA), have dominated in LC classification using VFSR imagery over the last decade (Blaschke et al., 2014). These OBIA approaches are built upon relatively homogeneous objects that are composed of similar pixel values across the image, for the identification of LCs through physical properties (such as spectra, texture, and shape) of ground components. The major challenges in applying these object-based approaches are the selection of segmentation scales to obtain objects that correspond to specific LC types, in which over- and under-segmentation commonly exist in the same image (Ming et al., 2015). To date, no effective solution has been proposed for LC classification using VFSR remotely sensed imagery.

Similar to LC classification, traditional LU classification methods using VFSR data can generally be categorised into three types; pixel-based, moving window-based, and object-based. The pixel-level approaches that rely purely upon spectral characteristics are able to classify LC, but are insufficient to distinguish LUs that are typically composed of multiple LCs, and this limitation is particularly significant in urban settings (Zhao et al., 2016). Spatial texture information (Herold et al., 2003; Myint, 2001) or spatial context (Wu et al., 2009) have been incorporated to analyse LU patterns through moving windows or kernels (Niemeyer et al., 2014). However, it could be argued that both pixel-based and moving window-based methods are based on arbitrary image structures, whereas actual objects and regions might be irregularly shaped in the real world (Herold et al., 2003). Therefore, the OBIA framework has been used to characterise LU based on spatial context. Typically, two kinds of information within a spatial partition are utilised, namely, within-object information (e.g. spectra, texture, shape) and between-object information (e.g. connectivity, contiguity, distances, and direction amongst adjacent objects). Many studies applied OBIA for LU classification using within-object information with a set of low-level features (such as spectra, texture, shape) of the land features (e.g. Blaschke, 2010; Blaschke et al., 2014; Hu and Wang, 2013). These OBIA methods, however, might overlook semantic functions or spatial configurations due to the inability to use low-level features in semantic feature representation. In this context, researchers have developed a two-step pipeline, where object-based LCs were initially extracted, followed by aggregating the objects using spatial contextual descriptive indicators on well-defined LU units, such as cadastral fields or street blocks. Those descriptive indicators are commonly derived by means of spatial metrics to quantify their morphological properties (Yoshida and Omae, 2005) or graph-based methods that model the spatial relationships (Barr and Barnsley, 1997; Walde et al., 2014). Yet, the ancillary geographic data for specifying the LU units might not be available at some regions, and the spatial contexts are often hard to be described and characterised as a set of “rules”, even though the complex structures or patterns might be recognisable and distinguishable by human experts (Oliva-Santos et al., 2014; C. Zhang et al., 2018c).

The major issue of the above-mentioned methods is the adoption of shallow structured classification models with hand-crafted features that are domain-specific and require a huge amount of effort in feature engineering. Recent advances in pattern recognition and machine learning have demonstrated a resurgence in the use of multi-layer neural networks to model higher-level feature representations without human-designed features or rules. This is largely driven by the wave of excitement in deep learning, where the most representative and discriminative features are learnt end-to-end, and hierarchically (Arel et al., 2010). Deep learning methods have achieved huge success not only in classical computer vision tasks, such as target detection, visual recognition and robotics, but also in many other practical applications (Hu et al., 2015; Nogueira et al., 2017). Convolutional neural networks (CNNs), as a well-established and popular deep learning method, have made considerable improvements beyond the state-of-the-art records in image analysis, and have attracted great interest in both academia and industrial communities (Krizhevsky et al., 2012). Owing to its superiority in higher-level feature representation, the CNN has demonstrated great potential in many remotely sensed tasks such as vehicle detection (Chen et al., 2014; Dong et al., 2015), road network extraction (Cheng et al., 2017), remotely sensed scene classification (Othman et al., 2016), and semantic segmentation (Zhao et al., 2017).

Translational invariance is a major advantage introduced by CNNs through a patch-wise procedure, in which a higher-level object within an image patch can be recognised even if the objects are shifted a few and/or geometrically distorted. Such translational invariance can help detect objects with higher order features, such as LU or functional sites. However, this characteristic becomes a major weakness in LC and LU classification for pixel-level differentiation, which introduces artefacts on the border of the classified patches and often produces blurred boundaries between ground surface objects (Zhang et al., 2018a, 2018b), thus, introducing uncertainty into the LC/LU classification. Previous research has, therefore, developed improved techniques for adapting CNN models to the LU/LC classification task. For example, Zhang et al. (2018a) fused deep CNN networks with the pixel-based multilayer perceptron (MLP) method to solve LC classification with spatial feature representation and pixel-level differentiation; Zhang et al. (2018b) proposed a regional fusion decision strategy based on rough set theory to model the uncertainties in LC classification of the CNN, and further guide data integration with other algorithms for targeted adjustment; Pan and Zhao, (2017) developed a central-point-enhanced CNN network to enhance the weight of the central pixels within image patches to strengthen the LC classification with precise land-cover boundaries. Besides, a range of research has explored the pixel-level Fully Convolutional Networks (FCN) and its extensions for remotely sensed semantic segmentations (e.g. Maggiori et al., 2017; Paisitkriangkrai et al., 2016; Volpi and Tuia, 2017), in which low-level LC classes, such as buildings, grassland, and cars, are classified with relatively high accuracy, although boundary distortions still exist due to the insufficient contextual information at up-sampling layers (Fu et al., 2017). With respect to LU classification, Zhang et al., (2018c) recently proposed a novel object-based CNN (OCNN) model that combines the OBIA and CNN techniques to learn LU objects through within-object and between-object information, where the semantic functions were characterised with precise boundary delineations. However, these pioneering efforts in CNN classification can only classify the image at a single, specific level, either LC or LU, whereas the landscape can be interpreted at different semantic levels simultaneously in a landscape hierarchy. At its most basic level this hierarchy simultaneously comprises LC at a lower, state level (what is there?) and LU at a higher, functional level (what is going on there?). Thus, both LC and LU cover the same geographical space, and are nested with each other hierarchically. The LUs often consist of multiple LC classes, and different spatial configurations of LC could lead to different LU classes. These two classification hierarchies are, thus, intrinsically correlated and are realised at different semantic levels.

The fundamental conceptual contribution of this paper is the realisation that the spatial and hierarchical relationships between LC (defined as a low-order state) and LU (defined as a higher-order semantic representation capturing function) might be learnt by characterising both representations at different levels with a *joint distribution*. In this paper, the first joint deep learning framework is proposed and demonstrated for LC and LU classification. Specifically, an MLP and Object-based CNN were applied iteratively and conditionally dependently to classify LC and LU *simultaneously*. The effectiveness of the proposed method was tested on two complex urban and suburban scenes in Great Britain.

The remainder of this paper is organised as: Section 2 introduces the key components of the proposed methods. Section 3 specifies the study area and data sources. The results are presented in section 4, followed by a discussion in section 5. The conclusions are drawn in the last section.

2. Method

2.1 multilayer perceptron (MLP)

A multilayer perceptron (MLP) is a network that maps from input data to output representations through a feedforward manner (Atkinson and Tatnall, 1997). The fundamental component of a MLP involves a set of computational nodes with weights and biases at multiple layers (input, hidden, and output layers) that are fully connected (Del Frate et al., 2007). The weights and biases within the network are learned through backpropagation to approximate the complex relationship between the input features and the output characteristics. The learning objective is to minimise the difference between the predictions and the desired outputs by using a specific cost function.

2.2 Convolutional Neural Networks (CNN)

As one of the most representative deep neural networks, convolutional neural network (CNN) is designed to process and analyse large scale sensory data or images in consideration of their stationary characteristics at local and global scales (LeCun et al., 2015). Within the CNN network, convolutional layers and pooling layers are connected alternatively to generalise the features towards deep and abstract representations. Typically, the convolutional layers are composed of weights and biases that are learnt through a set of image patches across the image (Romero et al., 2016). Those weights are shared by different feature maps, in which multiple features are learnt with a reduced amount of parameters, and an activation function (e.g. rectified linear units) is followed to strengthen the non-linearity of the convolutional operations (Strigl et al., 2010). The pooling layer involves max-pooling or average-pooling, where the summary statistics of local regions are derived to further enhance the generalisation capability.

2.3 Object-based Convolutional Neural Networks (OCNN)

An object-based CNN (OCNN) was proposed recently for the urban LU classification using remotely sensed imagery (Zhang et al., 2018c). The OCNN is trained as for the standard CNN model with labelled image patches, whereas the model prediction labels each segmented object derived from image segmentation. For each image object (polygon), a minimum moment bounding box was constructed by anisotropy with major and minor axes (Zhang and Atkinson, 2016). The centre point intersected with the polygon and the bisector of the major axis was used to approximate the central location of each image patch, where the convolutional process is implemented once per object. Interested readers are referred to a theoretical description on convolutional position analysis for targeted sampling on the centre point of image objects (C. Zhang et al., 2018c). The size of the image patch was tuned empirically to be sufficiently large, so that the object and spatial context were captured jointly by the CNN network. The OCNN was trained on the LU classes, in which the semantic information of LU was learnt through the deep network, while the boundaries of the objects were retained through the process of segmentation. The CNN model prediction was recorded as the predicted label of the image object to formulate a LU thematic map. Here, the predictions of each object are assigned to all of its pixels.

2.4 LC-LU Joint Deep Learning Model

The assumption of the LC – LU joint deep learning (LC-LU JDL) model is that both LC and LU are manifested over same geographical space and are nested with each other in a hierarchical manner. The LC and LU representations are considered as two random variables, where the probabilistic relationship between them can be modelled through a joint probability distribution. In this way, the conditional dependencies between these two random variables are captured via an undirected graph through iteration (i.e. formulating a Markov process). The joint distribution is, thus, factorised as a product of the individual density functions, conditional upon their parent variables as

$$p(x) = \prod_{v=1}^{k} p(x_v \mid x_{pa(v)}) \qquad (1)$$

where x_v represents a specific random variable, that is, either LC or LU class, and the $x_{pa(v)}$ denotes the parent variable of x_v. For example, x_v represents the LC class, and the $x_{pa(v)}$ in this case corresponds to the LU class.

Specifically, let C_{LC} = {C_{LC1}, C_{LC2}, …, C_{LCi} …, C_{LCm}} ($i \in [1, m]$), where C_{LCi} denotes the set of LC samples of the *i*th class, and *m* represents the number of LC classes; C_{LU} = { C_{LU1}, C_{LU2}, …, $C_{LU\mathrm{j}}$ …, C_{LCn}}($j \in [1, n]$), where C_{LUj} denotes the set of LU samples of the *j*th class and *n* indicates the number of LU classes. Both LC and LU classifications rely on a set of feature vectors *F* to represent their input evidence, and the predicted LC/LU categories are assigned based on the maximum *a posteriori* (MAP) criterion. Thus, the classification output of *m* LC classes or *n* LU classes can be derived as

$$C^* = \arg\max_{C_i} p(C_i \mid F) \tag{2}$$

where *i* corresponds to the specific LC/LU class during iteration.

Through the Bayes' theorem

$$p(C_i \mid F) = \frac{p(C_i)\,p(F \mid C_i)}{p(F)} \tag{3}$$

The classification result *C** is obtained as

$$C^* = \arg\max_{C_i} p(C_i)\,p(F \mid C_i) \tag{4}$$

In which $p(F)$ is the same at all states of C_i.

The $p(C_i)$ describes the prior probability distribution of each LC/LU class. In this research, we do not specify any priors for the classification, meaning that the joint distribution is equivalent to the modelled conditional distribution. The conditional probability $p(F \mid C_i)$ for the LC is initially estimated by the probabilistic MLP at the pixel level representing the membership association. Those LC conditional probabilities are then fed into the OCNN model to learn and classify each LU category. The estimated LU probabilities together with the original images are then re-used as input layers for LC classification using MLP in the next iteration. This iterative process can obtain both LC and LU classification results simultaneously at each iteration. Figure 1 illustrates the general workflow of the proposed LC and LU joint deep learning (LC-LU JDL) model, with key components including the JDL inputs, the Markov Process to learn the joint distribution, and the classification outputs of LC and LU at each iteration. Detailed explanation is given as follows.

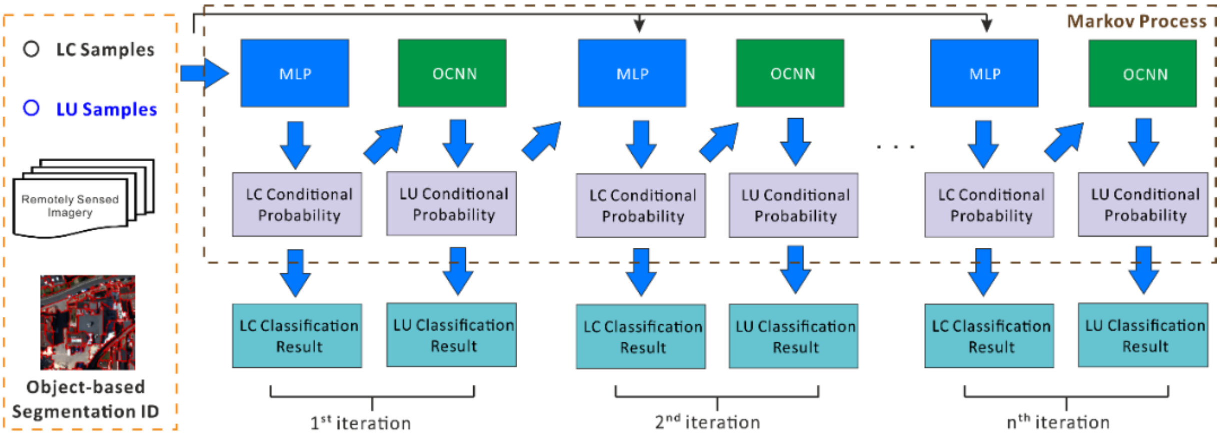

Figure 1 The general workflow of the land cover (LC) and land use (LU) joint deep learning (JDL).

JDL input involves LC samples with pixel locations and the corresponding land cover labels, LU samples with image patches representing specific land use categories, together with the remotely sensed imagery, and the object-based segmentation results with unique identity for each segment. These four elements were used to infer the hierarchical relationships between LC and LU, and to obtain LC and LU classification results through iteration.

Markov Process models the joint probability distribution between LC and LU through iteration, in which the joint distributions of the *i*th iteration are conditional upon the probability distribution of LC and LU derived from the previous iteration (*i*-1):

$$P(\text{LandCover}^{i},\text{LandUse}^{i}) = P(\text{LandCover}^{i},\text{LandUse}^{i} \mid \text{LandCover}^{i-1},\text{LandUse}^{i-1}) \tag{5}$$

where the LandCover^{i} and LandUse^{i} at each iteration update each other to approximate a complex hierarchical relationship between LC and LU.

Assume the complex relationship formulates a function *f*, equation (5) can be expressed as:

$$P(\text{LandCover}^{i},\text{LandUse}^{i}) = f(\text{LandCover}^{i-1},\text{LandUse}^{i-1},\text{Image},\text{SegmentImage},C_{LC},C_{LU}) \tag{6}$$

where the LandCover$^{i\text{-}1}$ and LandUse$^{i\text{-}1}$ are the LC and LU classification outputs at the previous iteration (*i*-1). The LandUse^{0} is an empty image with null value. Image here represents the original remotely sensed imagery, and SegmentImage is the label image derived from object-based segmentations with the same ID for each pixel within a segmented object. The C$_{LC}$ and C$_{LU}$ are LC and LU samples that record the locations in the image with corresponding class categories. All these six elements form the input parameters of the *f* function. Whereas the predictions of the *f* function are the joint distribution of LandCover^{i} and LandUse^{i} as the classification results of the *i*th iteration.

Within each iteration, the MLP and OCNN are used to derive the conditional probabilities of LC and LU, respectively. The input evidence for the LC classification using MLP is the original image together with the LU conditional probabilities derived from the previous iteration, whereas the LU classification using OCNN only takes the LC conditional probabilities as input variables to learn the complex relationship between LC and LU. The LC and LU conditional probabilities and classification results are elaborated as follows.

Land cover (LC) conditional probabilities are derived as:

$$P(\mathrm{LandCover}^{i}) = P(\mathrm{LandCover}^{i} \mid \mathrm{LandUse}^{i-1}) \tag{7}$$

where the MLP model is trained to solve equation (7) as:

$$MLPModel^{i} = TrainMLP(concat(\mathrm{LandUse}^{i-1}, \mathrm{Image}), C_{LC}) \tag{8}$$

The function *concat* here integrates LU conditional probabilities and the original images, and the LC samples C_{LC} are used to train the MLP model. The LC classification results are predicted by the MAP likelihood as:

$$LandCover^{i} = MLPModel^{i}.predict\ (concat(\mathrm{LandUse}^{i-1}, \mathrm{Image}) \tag{9}$$

Land use (LU) conditional probabilities are deduced as:

$$P(\mathrm{LandUse}^{i}) = P(\mathrm{LandUse}^{i} \mid \mathrm{LandCover}^{i}) \tag{10}$$

where the OCNN model is built to solve equation (10) as:

$$OCNNModel^{i} = TrainCNN(\mathrm{LandCover}^{i}, C_{LU}) \qquad (11)$$

The OCNN model is based on the LC conditional probabilities derived from MLP as its input evidence. The C_{LU} is used as the training sample sites of LU, where each sample site is used as the centre point to crop an image patch as the input feature map for training the CNN model. The trained CNN can then be used to predict the LU membership association of each object as:

$$LandUse^{i} = CNNModel^{i}.predict(\mathrm{cast}(\mathrm{LandCover}^{i}, \mathrm{SegmentImage}) \qquad (12)$$

where the function *cast* denotes the cropped image patch with LC probabilities derived from LandCover^{i}, and the predicted LU category for each object was recorded in SegmentImage, in which the same label was assigned for all pixels of an object.

Essentially, the Joint Deep Learning (JDL) model has four key advantages:

1. The JDL is designed for joint land cover and land use classification in an automatic fashion, whereas previous methods can only classify a single, specific level of representation.
2. The JDL jointly increases the accuracy of both the land cover and land use classifications through mutual complementarity and reinforcement.
3. The JDL accounts explicitly for the spatial and hierarchical relationships between land cover and land use that are manifested over the same geographical space at different levels.
4. The JDL increases model robustness and generalisation capability, which supports incorporation of deep learning models (e.g. CNNs) with a small training sample size.

3. Experimental Results and Analysis

3.1 Study area and data sources

In this research, two study areas in the UK were selected, namely Southampton (S1) and Manchester (S2) and their surrounding regions, lying on the Southern coast and in North West England, respectively (Figure 2). Both study areas involve urban and suburban areas that are highly heterogeneous and distinctive from each other in both LC and LU characteristics and are, therefore, suitable for testing the generalisation capability of the joint deep learning approach.

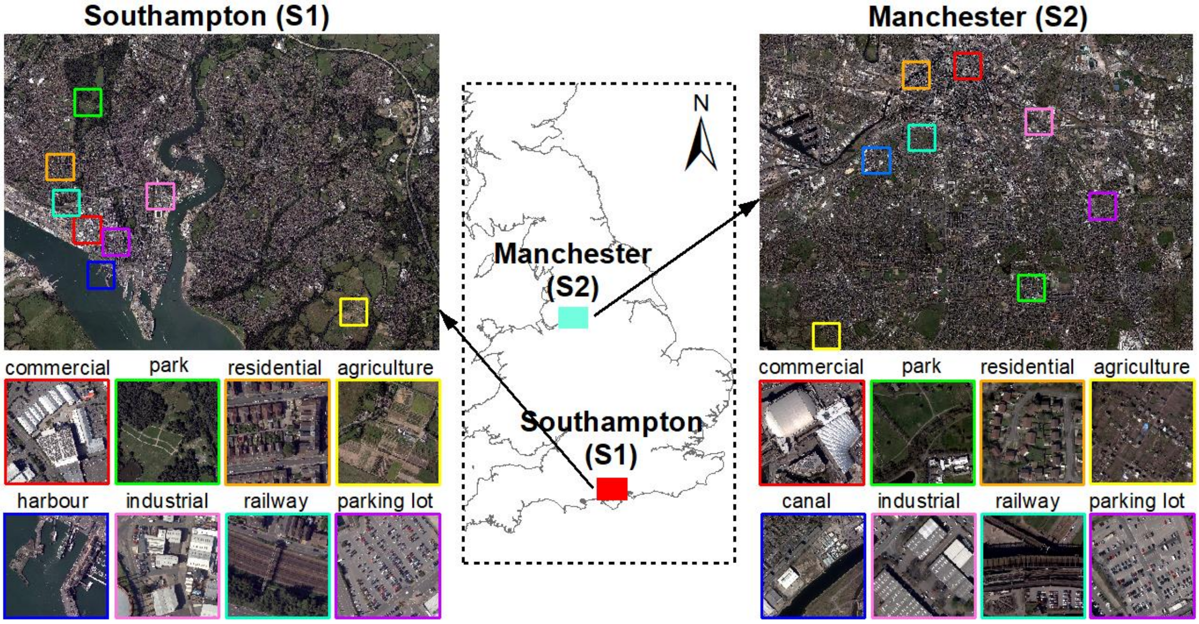

Figure 2 The two study areas: S1 (Southampton) and S2 (Manchester) with highlighted regions representing the majority of land use categories.

Aerial photos of S1 and S2 were captured using Vexcel UltraCam Xp digital aerial cameras on 22/07/2012 and 20/04/2016, respectively. The images have four multispectral bands (Red, Green, Blue and Near Infrared) with a spatial resolution of 50 cm. The study sites were subset into the city centres and their surrounding regions with spatial extents of 23250×17500 pixels for S1 and 19620×15450 pixels for S2, respectively. Besides, digital surface model (DSM) data of S1 and S2 with the same spatial resolution as the imagery were also acquired, and used for image segmentation only. 10 dominant LC classes were identified in both S1 and S2, comprising *clay roof*, *concrete roof*, *metal roof*, *asphalt*, *rail*, *bare soil*, *woodland*, *grassland*, *crops*, and *water* (Table 1). These LCs represent the physical properties of the ground surface recorded by the spectral reflectance of the aerial images. On the contrary, the LU categories within the study areas were characterised based on human-induced functional utilisations. 11 dominant LU classes were recognised in S1, including *high-density residential*, *commercial*, *industrial*, *medium-density residential*, *highway*, *railway*, *park and recreational area*, *agricultural area*, *parking lot*, *redeveloped area*, and *harbour and sea water*. In S2, 10 LU categories were found, including *residential*, *commercial*, *industrial*, *highway*, *railway*, *park and recreational area*, *agricultural areas*, *parking lot*, *redeveloped area*, and *canal* (Table 1). The majority of LU types for both study sites are highlighted and exemplified in Figure 2. These LC and LU classes were defined based on the Urban Atlas and CORINE land cover products coordinated by the European Environment Agency (https://land.copernicus.eu/), as well as the official land use classification system designed by the Ministry of Housing, Communities and Local Government (MHCLG) of the UK government. Detailed descriptions for LU and the corresponding sub-classes together with the major LC components in both study sites are summarised in Table 1.

Table 1. The land use (LU) classes with their sub-class descriptions, and the associated major land cover (LC) components across the two study sites (S1 and S2).

LU	Study site	Sub-class descriptions	Major LC
(High-density) residential	S1, S2	Residential houses, terraces, green space	Buildings, Grassland, Woodland
Medium-density residential	S1	Residential flats, green space, parking lots	Buildings, Grassland, Asphalt
Commercial	S1, S2	Shopping centre, retail parks, commercial services	Buildings, Asphalt
Industrial	S1, S2	Marine transportation, car factories, gas industry	Buildings, Asphalt
Highway	S1, S2	Asphalt road, lane, cars	Asphalt
Railway	S1, S2	Rail tracks, gravel, sometimes covered by trains	Rail, Bare soil, Woodland
Parking lot	S1, S2	Asphalt road, parking line, cars	Asphalt
Park and recreational area	S1, S2	Green space and vegetation, bare soil, lake	Grassland, Woodland
Agricultural area	S1, S2	Pastures, arable land, and permanent crops	Crops, Grassland
Redeveloped area	S1, S2	Bare soil, scattered vegetation, reconstructions	Bare soil, Grassland
Harbour and sea water	S1	Sea shore, harbour, estuaries, sea water	Water, Asphalt, Bare soil

Canal	S2	Water drainage channels, canal water	Water, Asphalt

The ground reference data for both LC and LU are polygons collected by local surveyors and digitised manually by photogrammetrists in the UK, covering the majority of the study areas (over 80%). These reference polygons with well-defined labelling protocols are specified in Table 1. The polygons were split randomly into a 50% subset for training and calibration and the other 50% subset for validation, to avoid spatial correlation in the sample distributions. Unbiased sample sets were generated for each class, proportional to the total area of the reference polygons corresponding to a specific class, through a stratified random sampling scheme. The sample sizes for specific classes with sparse spatial coverage (e.g. railways) were increased so as to obtain a sample distribution that was comparable in size. The training sample size for LCs was approximately 600 per class to allow the MLP to learn the spectral characteristics over the relatively large sample size. The LU classes consist of over 1000 training sample sites per class, in which deep CNN networks could sufficiently distinguish the patterns through data representations. These LU and LC sample sets were checked and cross referenced with the MasterMap Topographic Layer produced by Ordnance Survey (Regnauld and Mackaness, 2006), and Open Street Maps, together with field survey to ensure the precision and validity of the sample sets. The sampling probability distribution was further incorporated into the accuracy assessment statistics (e.g. overall accuracy) to ensure statistically unbiased validation (Olofsson et al., 2014).

3.2 Model structure and parameter settings

The model structures and parameters were optimised in S1 through cross validation and directly generalised into S2 to test the robustness and the transferability of the proposed methods in different experimental environments. Within the Joint Deep Learning approach, both MLP and

OCNN require a set of predefined parameters to optimise the accuracy and generalisation capability. Detailed model structures and parameters were clarified as below.

3.2.1 MLP Model structure and parameters

The initial input of the MLP classifier is the four multi-spectral bands at the pixel level, where the prediction is the LC class that each pixel belongs to. Followed by the suggestions of Mas and Flores (2008) and Zhang et al., (2018a), one, two and three hidden layers of MLPs were tested, with different numbers of nodes {4, 8, 12, 16, 20, and 24} in each layer. The learning rate was optimised as 0.2 and the momentum was optimally chosen as 0.7. The number of epochs for the MLP network was tuned as 800 to converge at a stable stage. The optimal parameters for the MLP were chosen by cross validating among different numbers of nodes and hidden layers, in which the best accuracy was reported with two hidden layers and 16 nodes at each layer.

3.2.2 Object-based Segmentation parameter settings

The Object-based Convolutional Neural Network (OCNN) requires the input image to be pre-processing into segmented objects through object-based segmentation. A hierarchical step-wise region growing segmentation algorithm was implemented through the Object Analyst Module in PCI Geomatics 2017. A series of image segmentations was performed by varying the scale parameter from 10 to 100, while other parameters (shape and compactness) were fixed as default. Through cross validation with trial-and-error, the scale parameter was optimised as 40 to produce a small amount of over-segmentation and, thereby, mitigate salt and pepper effects simultaneously. A total of 61,922 and 58,408 objects were obtained from segmentation for S1 and S2, respectively. All these segmented objects were stored as both vector polygons in an ArcGIS Geodatabase and raster datasets with the same ID for all pixels in each object.

3.2.3 OCNN model structure and parameters

For each segmented object, the centre point of the object was taken as the centre of the input image patch, where a standard CNN was trained to classify the object into a specific LU category. In other words, a targeted sampling was conducted once per object, which is different from the standard pixel-wise CNNs that apply the convolutional filters at locations evenly spaced across the image. The model structure of the OCNN was designed similar to the AlexNet (Krizhevsky et al., 2012) with eight hidden layers (Figure 3) using a large input window size (96×96), but with small convolutional filters (3×3) for the majority of layers except for the first one (which was 5×5). The input window size was determined through cross validation on a range of window sizes, including {32×32, 48×48, 64×64, 80×80, 96×96, 112×112, 128×128, 144×144} to sufficiently cover the contextual information of objects relevant to their LU semantics. The filter number was tuned as 64 to extract deep convolutional features effectively at each level. The CNN network involved alternating convolutional (conv) and pooling layers (pool) as shown in Figure 3, where the maximum pooling within a 2×2 window was used to generalise the feature and keep the parameters tractable.

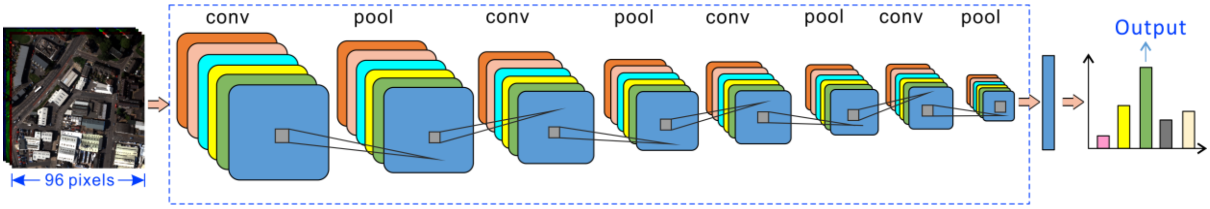

Figure 3 Model architectures and structures of the CNN with 96×96 input window size and eight-layer depth.

All the other parameters were optimised empirically on the basis of standard practice in deep network modelling. For example, the number of neurons for the fully connected layers was set as 24, and the output labels were predicted through softmax estimation with the same number of LU categories. The learning rate and the epoch were set as 0.01 and 600 to learn the deep features through backpropagation.

3.2.4 Benchmark approaches and parameter settings

To validate the classification performance of the proposed Joint Deep Learning for LC and LU classification, three existing methods (i.e. multilayer perceptron (MLP), support vector machine (SVM), and Markov Random Field (MRF)) were used as benchmarks for LC classification, and three methods, MRF, object-based image analysis with support vector machine (OBIA-SVM), and the pixel-wise CNN (CNN), were used for benchmark evaluation of the LU classification. Detailed descriptions and parameters are provided as follows:

MLP: The model structures and parameters for the multilayer perceptron were kept the same as the MLP model within the proposed Joint Deep Learning, with two hidden layers and 16 nodes for each layer. Such consistency in parameter setting makes the baseline results comparable.

SVM: A penalty value *C* and a kernel width σ within the SVM model are required to be parameterised. As suggested by Zhang et al., (2015), a wide parameter space (*C* and σ within [2^{-10}, 2^{10}]) was used to exhaustively search the parameters through a grid-search with 5-fold cross validation. Such settings of parameters should result in high accuracies with support vectors formulating optimal hyperplanes among different classes.

MRF: The Markov Random Field, a spatial contextual classifier, was taken as a benchmark comparator for both the LC and LU classifications. The MRF was constructed by the conditional probability formulated by a support vector machine (SVM) at the pixel level, which was parameterised through grid search with a 5-fold cross validation. Spatial context was incorporated by a neighbourhood window (7×7), and a smoothness level γ was set as 0.7. The simulated annealing was employed to optimise the posterior probability distribution with iteration.

OBIA-SVM: Multi-resolution segmentation was implemented initially to segment objects through the image. A range of features were further extracted from these objects, including spectral features (mean and standard deviation), texture (grey-level co-occurrence matrix) and geometry (e.g. perimeter-area ratio, shape index). In addition, the contextual pairwise similarity that measures the similarity degree between an image object and its neighbouring objects was deduced to account for the spatial context. All these hand-coded features were fed into a parameterised SVM for object-based classification.

Pixel-wise CNN: The standard pixel-wise CNN was trained to predict each pixel across the entire image using densely overlapping image patches. The most crucial parameters that influence directly the performance of the pixel-wise CNN are the input patch size and the network depth (i.e. number of layers). As discussed by Längkvist et al., (2016), the input patch size was chosen from {28×28, 32×32, 36×36, 40×40, 44×44, 48×48, 52×52 and 56×56} to test the influence of contextual area on classification results. The optimal input image patch size for the pixel-wise CNN was found to be 48×48 to leverage the training sample size and the computational resources (e.g. GPU memory). The depth configuration of the CNN network is essential in classification accuracy since the quality of the learnt features is influenced by the levels of representations and abstractions. Followed by the suggestions from Chen et al. (2016), the number of layers for CNN network was set as six with three convolutional layers and three pooling layers to balance the complexity and the robustness of the network. Other CNN parameters were empirically tuned through cross validation. For example, the filter size was set to 3×3 of the convolutional layer with one stride, and the number of convolutional filters was set to 24. The learning rate was chosen as 0.01, and the number of epochs was set as 600 to learn the features fully with backpropagation.

3.3 Classification results and analysis

The classification performance of the proposed Joint Deep Learning using the above-mentioned parameters was investigated in both S1 (experiment 1) and S2 (experiment 2). The LC classification results (JDL-LC) were compared with benchmarks, including the multilayer perceptron (MLP), support vector machine (SVM) and Markov Random Field (MRF); whereas, the LU classification results (JDL-LU), were benchmarked with MRF, Object-based image analysis with SVM (OBIA-SVM), and standard pixel-wise CNN. Visual inspection and quantitative accuracy assessment, with overall accuracy (OA) and the per-class mapping accuracy, were adopted to evaluate the classification results. In addition, two recently proposed indices, including quantity disagreement and allocation disagreement, instead of the Kappa coefficient, were used to summarise comprehensively the confusion matrix of the classification results (Pontius and Millones, 2011).

3.3.1 LC-LU JDL Classification Iteration

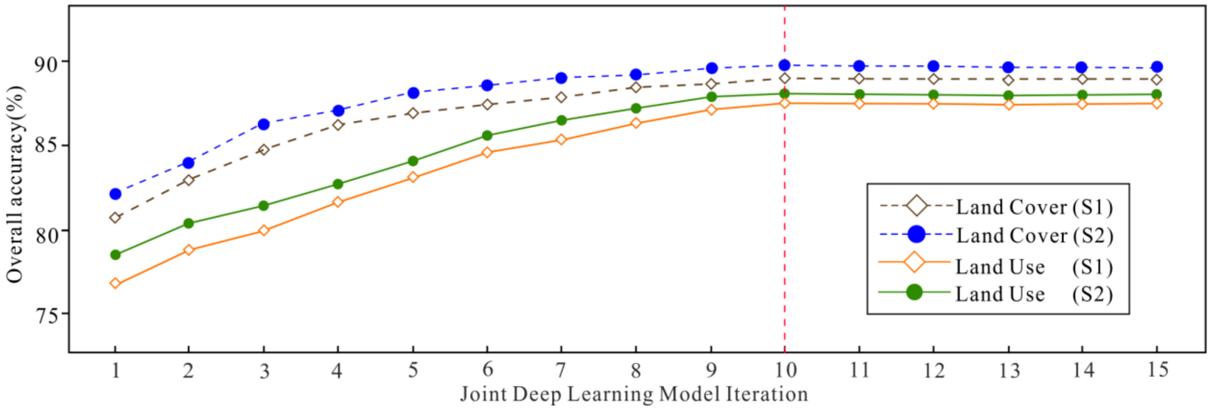

Figure 4 The overall accuracy curves for the Joint Deep Learning iteration of land cover (LC) and land use (LU) classification results in S1 and S2. The red dash line indicates the optimal accuracy for the LC and LU classification at iteration 10

The proposed LC-LU JDL was implemented through iteration. For each iteration, the LC and LU classifications were implemented 10 times with 50% training and 50% testing sample sets split randomly using the Monte Carlo method, in which the testing samples of each run did not involve the pixels that have been used during the training process. The average overall accuracy (OA) of each iteration (each repeated 10 times) was reported to demonstrate how the accuracy evolves during the iterative process. Figure 4 demonstrates the average OA of both S1 and S2 through accuracy curves from iteration 1 to 15. It can be seen that the accuracies of LC classified by MLP in both S1 and S2 start from around 81%, and gradually increase along the process until iteration 10 with a tendency of being closer to each other, and reach the highest OA up to around 90% for both sites. After iteration 10 (i.e. from iteration 10 to 15), the OA tends to be stable (i.e. around 90%). A similar trend is found in LU classifications in the iterative process, with a lower accuracy than the LC classification at each iteration. Specifically, the OAs in S1 and S2 start from around 77.5% and 78.1% at iteration 1, and keep increasing and getting closer at each iteration, until reaching the highest (around 87%) accuracy at iteration 10 for both study sites, and demonstrate convergence at later iterations (i.e. being stable from iteration 10 to 15). Therefore, iteration 10 was found to provide the optimal solution for the joint deep learning model between LC and LU.

3.3.2 JDL Land cover (JDL-LC) classification iteration

LC classification results in S1 and S2, obtained by the JDL-Land cover (JDL-LC) through iteration, are demonstrated in Figures 5 and 6, respectively, with the optimal classification outcome (at iteration 10) marked by blue boxes. In Figure 5, four subsets of S1 at different iterations (1, 2, 4, 6, 8, and 10) are presented to provide better visualisation, with yellow and red circles highlighting correct and incorrect classification, respectively. The classification in iteration 1 was affected by the shadow cast in the images. For example, the shadows of the woodland on top of grassland demonstrated in Figure 5(*a*) (the red circle on the right side) were misclassified as Rail due to the influence of illumination conditions and shadow contaminations in the imagery. Also, misclassification between bare soil and asphalt appeared in the result of iteration 1, caused by within-class variation in the spectral reflectance of bare land (red circles in Figure 5(*a*) and 5(*c*)). Further, salt and pepper effects were found in iteration 1 with obvious confusion between different roof tiles and asphalt, particularly the misclassification between Concrete roof and Asphalt (red circles in Figure 5(*b*)), due to the huge spectral similarity between different physical materials and characteristics. Besides, the noisy effects were also witnessed in rural areas, such as the severe confusion between Woodland and Grassland, and the misclassifications between Crops and Grassland in agricultural areas (Figure 5(*d*)). These problems were gradually solved by the introduction of spatial information at iteration 2 and thereafter, where the relationship between LC and LU was modelled using a joint probability distribution which helped to introduce spatial context, and the misclassification was reduced through iteration. Clearly, the shadow (red circles in Figure 5(*a*)) was successively modified and reduced throughout the process (iteration 2 – 8) with the incorporation of contextual information, and was completely eliminated in iteration 10 (yellow circle in Figure 5(*a*)). At the same time, the classifications demonstrated obvious salt-and-pepper effects in the early iterations (red circles in iteration 2 – 8 of Figure 5(*b*)), but the final result appeared to be reasonably smooth with accurate characterisation of asphalt road and clay roof (yellow circles in Figure 5(*b*) of iteration 10). In addition, confusion between metal roof and concrete roof (iteration 1 – 8 with red circles in Figure 5(*c*)) was rectified step-by-step through iteration, with the entire building successfully classified as metal roof at iteration 10 (yellow circle in Figure 5(*c*)). Moreover, the crops within Figure 5(*d*) was smoothed gradually from severe salt-and-pepper effects in iteration 1 (red circles in Figure 5(*d*)) to sufficiently smoothed representations in iteration 10 (yellow circle in Figure 5(*d*)). In short, a desirable result was achieved at iteration 10, where the LC classification was not only free from the influence of shadows and illuminations, but also demonstrated smoothness while keeping key land features well maintained (yellow circles in Figure 5(*a*-*d*)). For example, the small path within the park was retained and classified as Asphalt at iteration 10, and the Grassland and Woodland were distinguished with high accuracy (yellow circle in Figure 5(*d*)).

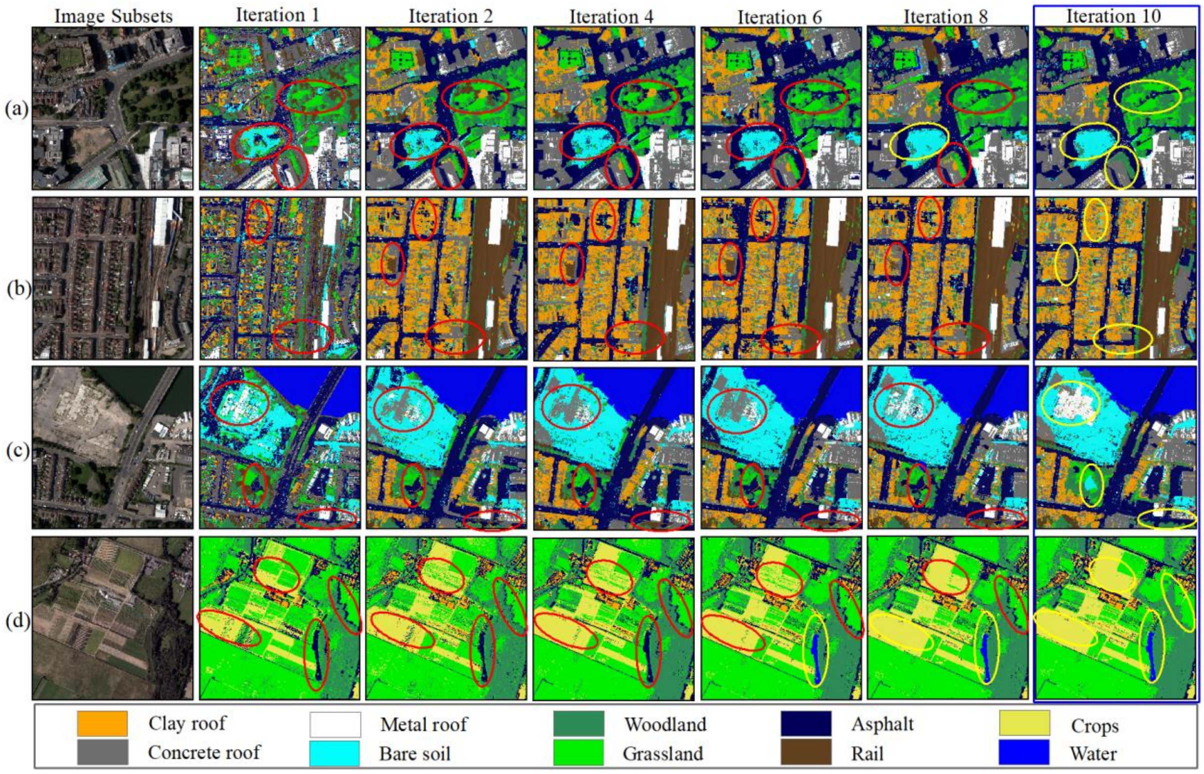

Figure 5 Four subset land cover classification results in S1 using Joint Deep Learning – Land cover (JDL-LC), the best results at iteration 10 were highlighted with blue box. The circles in yellow and red represent the correct and incorrect classification, respectively.

In S2, the LC classification results demonstrated a similar trend as for S1, where iteration 10 achieved the classification outputs with highest overall accuracy (Figure 4) and best visual appeal (Figure 6). The lowest classification accuracy was achieved in iteration 1, with obvious misclassification caused by the highly mixed spectral reflectance and the scattering of peripheral ground objects, together with salt-and-pepper effects throughout the classification results (Figure 6(*c*)). Such problems were tackled with increasing iteration (Figure 6(*d*-*h*)), where spatial context was gradually incorporated into the LC classification. The greatest improvement demonstrated with increasing iteration was the removal of misclassified shadows within the classified maps. For example, the shadows of the buildings were falsely identified as water due to the similar dark spectral reflectance (Figure 6(*c*)). Such shadow effects were gradually reduced in Figure 6(*d*-*g*) and completely eliminated in Figure 6(*h*) at iteration 10, which was highlighted by blue box as the best classification result in JDL-LC (Figure 6(*h*)). Other improvements included the clear identification of Rail and Asphalt through iteration and the reduced noisy effects, for example, the misclassified scatter (asphalt) in the central region of bare soil was successfully removed in iteration 10.

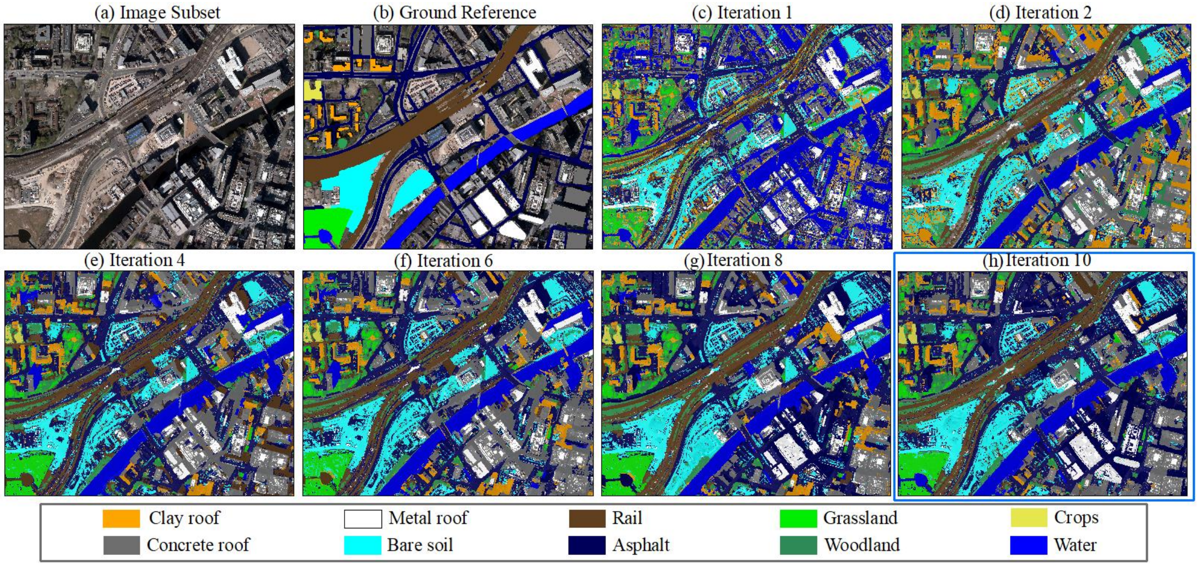

Figure 6 The land cover classification results in S2 using Joint Deep Learning – Land cover (JDL-LC), the best results at (h) iteration 10 were highlighted with blue box.

3.3.3 JDL-Land use (JDL-LU) classification Iteration

LU classifications from the JDL-Land use (JDL-LU) are demonstrated in Figures 7 and 8 for S1 (four subsets) and S2 (one subset), respectively, for iterations 1, 2, 4, 6, 8, and 10. Overall, the LU classifications in iteration 10 for both S1 and S2 are the optimal results with precise and accurate LU objects characterised through the joint distributions (in blue boxes), and the iterations illustrate a continuous increase in overall accuracy until reaching the optimum as shown by the dashed red line in Figure 4.

Specifically, in S1, several remarkable improvements have been achieved with increasing iteration, as marked by the yellow circles in iteration 10. The most obvious performance improvement is the differentiation between parking lot and highway. For example, a highway was misclassified as parking lot in iterations 1 to 4 (red circles in Figure 7(*a*)), and was gradually refined through the joint distribution modelling process with the incorporation of more accurate LC information (yellow circles in iteration 6 – 10). Such improvements can also be seen in Figure 7(*c*), where the misclassified parking lot was allocated to highway in iterations 1 to 8 (red circles), and was surprisingly rectified in iteration 10 (yellow circle). Another significant modification gained from the iteration process is the differentiation between agricultural areas and redeveloped areas, particularly for the fallow or harvested areas without pasture or crops. Figure 7(*d*) demonstrates the misclassified redeveloped area within the agricultural area from iterations 1 to 8 (highlighted by red circles), which was completely rectified as a smoothed agricultural field in iteration 10. In addition, the adjacent high-density residential areas and highway were differentiated throughout the iterative process. For example, the misclassifications of residential and highway shown in iteration 1 – 6 (red circles in Figure 7(*b*)) were mostly rectified in iteration 8 and were completely distinguished in iteration 10 with high accuracy ((yellow circles in Figure 7(*b*)). Besides, the mixtures between complex objects, such as commercial and industrial, were modified throughout the classification process. For example, confusion between commercial and industrial in iterations 1 to 8 (red circles in Figure 7(*a*)) were rectified in iteration 10 (yellow circle in Figure 7(*a*)), with precise LU semantics being captured through object identification and classification. Moreover, some small objects falsely identified as park and recreational areas at iterations 1 to 6, such as the high-density residential or railway within the park (red circles in Figure 7(*a*) and 7(*c*)), were accurately removed either at iteration 8 (yellow circle in Figure 7(*a*)) or at iteration 10 (yellow circle in Figure 7(*c*)).

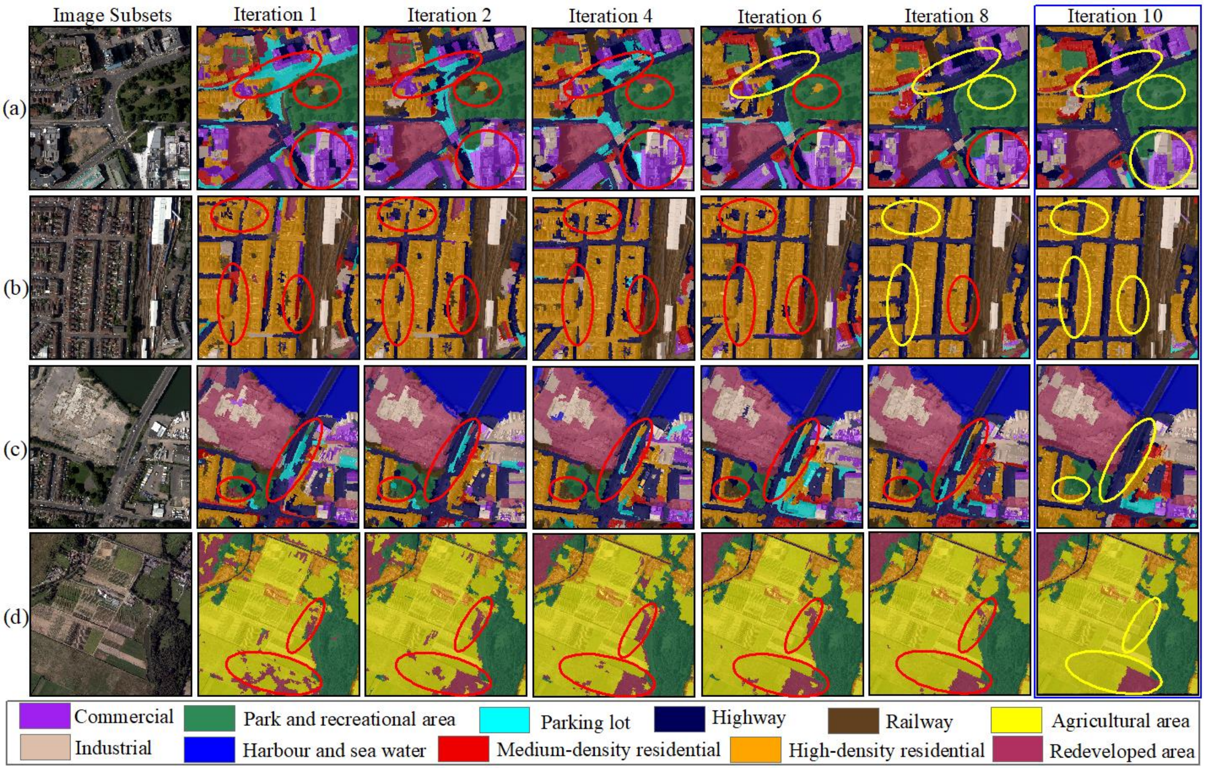

Figure 7 Four subset land use classification results in S1 using Joint Deep Learning – Land use (JDL-LU), the best results at iteration 10 were highlighted with blue box. The circles in yellow and red represent the correct and incorrect classification, respectively.

In S2, the iterative process also exhibits similar improvements with iteration. For example, the mixture of commercial areas and industrial areas in S2 (Figure 8(*c*)) was gradually reduced through the process (Figure 8(*d*-*g*)), and was surprisingly resolved at iteration 10 (Figure 8(*h*)), with the precise boundaries of commercial buildings and industrial buildings as well as the surrounding configurations identified accurately. Besides, the misclassification of parking lot as highway or redeveloped area was rectified through iteration. As illustrated in Figure 8(*c*-*g*), parts of the highway and redeveloped area were falsely identified as parking lot, but were accurately distinguished at iteration 10 (Figure 8(*h*)). Moreover, a narrow highway that was spatially adjacent to the railway, that was not identified at iteration 1 (Figure 8(*c*)), was identified at iteration 10 (Figure 8(*h*)), demonstrating the ability of the proposed JDL method to differentiate small linear features.

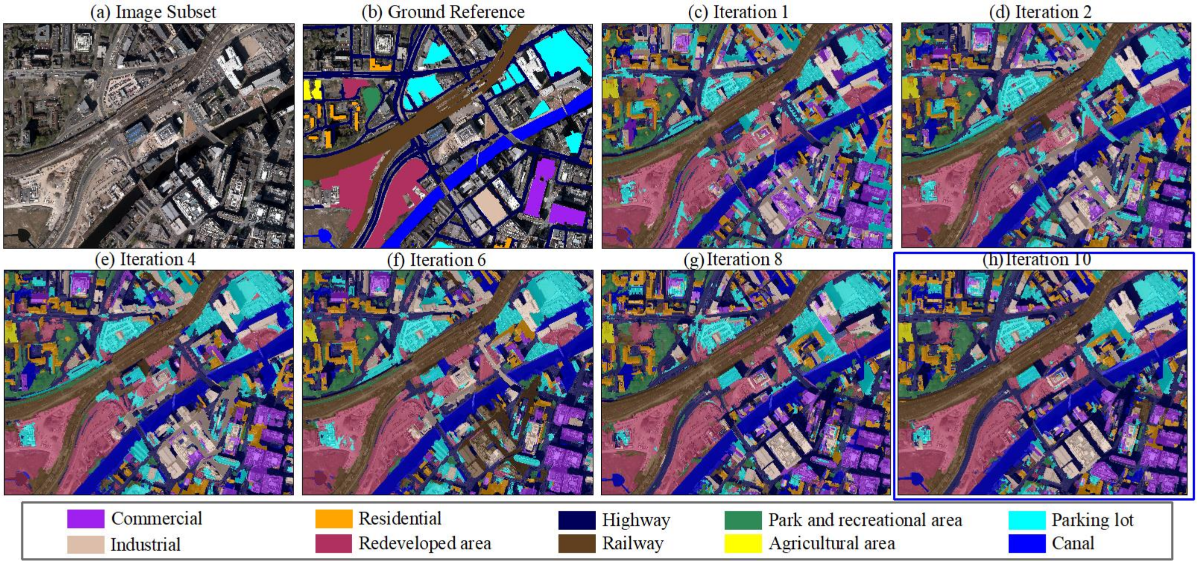

Figure 8 The land use classification results in S2 using Joint Deep Learning – Land use (JDL-LU), the best results at (h) iteration 10 were highlighted with blue box.

3.3.4 Benchmark comparison for LC and LU classification

To further evaluate the LC and LU classification performance of the proposed JDL method with the best results at iteration 10, a range of benchmark comparisons were presented. For the LC classification, a multilayer perceptron (MLP), support vector machine (SVM) and Markov Random Field (MRF) were benchmarked for both S1 and S2; whereas the LU classification took the Markov Random Field (MRF), Object-based image analysis with SVM classifier (OBIA-SVM) and a standard pixel-wise convolutional neural network (CNN) as benchmark comparators. The benchmark comparison results for overall accuracies (OA) of LC and LU classifications were demonstrated in Figure 9(*a*) and Figure 9(*b*), respectively. As shown by Figure 9(*a*), the JDL-LC achieved the largest OA of up to 89.64% and 90.72% for the S1 and S2, larger than the MRF of 84.78% and 84.54%, the SVM of 82.38% and 82.26%, and the MLP of 81.29% and 82.22%, respectively. For the LU classification in Figure 9(*b*), the proposed JDL-LU achieved 87.58% and 88.26% for S1 and S2, higher than those of CNN

(84.08% and 83.32%), OBIA-SVM (80.26% and 80.42%), and MRF (79.38% and 79.26%) respectively.

In addition to the OA, the proposed JDL method achieved consistently the smallest values for both Quantity and Allocation Disagreement, respectively. From Table 2 and 3, the JDL-LC has the smallest disagreement in terms of LC classification, with an average of 6.93% and 6.73% for S1 and S2 accordingly, which is far smaller than for any of the three benchmarks. Similar patterns were found in LU classification (Table 4 and 5), where the JDL-LU acquired the smallest average disagreement in S1 and S2 (9.98% and 9.16%), much smaller than for the MRF (20.32% and 19.11%), OBIA-SVM (18.59% and 16.82%), and CNN (14.23% and 13.99%).

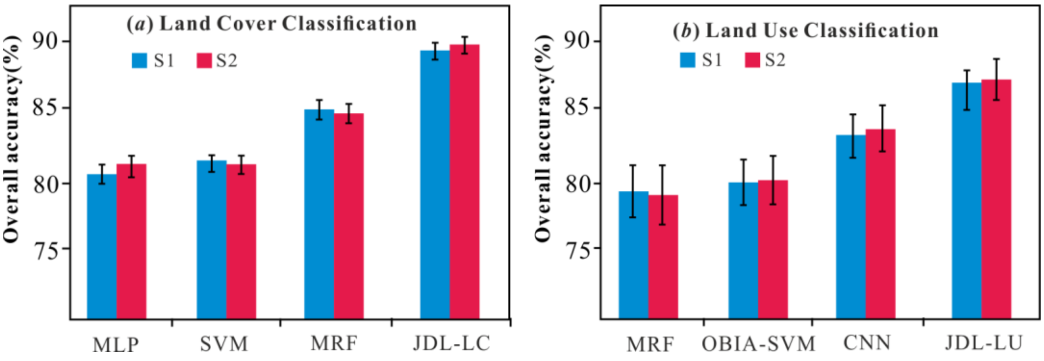

Figure 9 Overall accuracy comparisons among the MLP, SVM, MRF, and the proposed JDL-LC for land cover classification, and the MRF, OBIA-SVM, CNN, and the proposed JDL-LU for land use classification.

Per-class mapping accuracies of the two study sites (S1 and S2) were listed to provide detailed comparison of each LC (Table 2 and Table 3) and LU (Table 4 and Table 5) category. Both the proposed JDL-LC and the JDL-LU constantly report the most accurate results in terms of class-wise classification accuracy highlighted in bold font within the four tables.

For the LC classification (Table 2 and Table 3), the mapping accuracies of Clay roof, Metal roof, Grassland, Asphalt and Water are higher than 90%, with the greatest accuracy obtained by water in S1 (98.52%) and S2 (98.33%), respectively. The most remarkable increase in accuracy can be seen in Grassland with an accuracy of up to 90.05% and 90.63%, respectively, much higher than for the other three benchmarks, including the MRF (75.53% and 75.45%), the SVM (73.06% and 73.56%), and the MLP (70.63% and 72.22%). Another significant increase in accuracy was found in Woodland through JDL-LC with the mapping accuracy of 88.52% (S1) and 88.23% (S2), dramatically higher than for the MRF of 76.28% and 75.32%, SVM of 70.52% and 70.22%, and MLP of 69.02% and 69.59%, respectively. Likewise, the Concrete roof also demonstrated an obvious increase in accuracy from just 69.46% and 70.58% classified by the MLP to 79.47% and 79.27% in S1 and S2, respectively, even though the mapping accuracy of the Concrete roof is still relatively low (less than 80%). In addition, moderate accuracy increases have been achieved for the classes of Rail and Bare soil with an average increase of 5.28% and 5.51%, respectively. Other LC classes such as Clay roof, Metal roof, and Water, demonstrate only slight increases using the JDL-LC method in comparison with other benchmark approaches, with an average of 1% to 3% accuracy increases among them.

Table 2. Per-class and overall land cover accuracy comparison between MRF, OBIA-SVM, Pixel-wise CNN, and the proposed JDL-LC method for S1. The quantity disagreement and allocation disagreement are also shown. The largest classification accuracy and the smallest disagreement are highlighted in bold font.

Land Cover Class (S1)	MLP	SVM	MRF	JDL-LC
Clay roof	89.58%	89.33%	89.18%	**92.38%**
Concrete roof	69.46%	69.79%	73.23%	**79.47%**
Metal roof	89.35%	90.74%	90.16%	**91.58%**
Woodland	69.02%	70.52%	76.28%	**88.52%**
Grassland	70.63%	73.06%	75.53%	**90.05%**
Asphalt	88.42%	88.29%	89.42%	**91.22%**
Rail	82.05%	82.42%	83.56%	**87.26%**
Bare soil	80.12%	80.23%	82.44%	**85.72%**

Crops	84.14%	84.64%	86.59%	**89.64%**
Water	97.18%	97.45%	98.36%	**98.52%**
Overall Accuracy (OA)	81.29%	82.38%	84.78%	**89.64%**
Quantity Disagreement	17.18%	16.94%	11.28%	**7.63%**
Allocation Disagreement	16.26%	16.41%	13.47%	**6.23%**

Table 3. Per-class and overall land cover accuracy comparison between MRF, OBIA-SVM, Pixel-wise CNN, and the proposed JDL-LC method for S2. The quantity disagreement and allocation disagreement are also shown. The largest classification accuracy and the smallest disagreement are highlighted in bold font.

Land Cover Class (S2)	MLP	SVM	MRF	JDL-LC
Clay roof	90.06%	90.24%	89.55%	**92.85%**
Concrete roof	70.58%	70.42%	74.21%	**79.27%**
Metal roof	90.12%	90.85%	90.09%	**91.32%**
Woodland	69.59%	70.22%	75.32%	**88.23%**
Grassland	72.22%	73.56%	75.45%	**90.63%**
Asphalt	89.46%	89.53%	89.42%	**91.64%**
Rail	83.18%	83.14%	84.36%	**88.52%**
Bare soil	80.21%	80.36%	82.25%	**85.63%**
Crops	85.01%	85.28%	87.84%	**90.79%**
Water	97.54%	97.25%	98.02%	**98.33%**
Overall Accuracy (OA)	82.22%	82.26%	84.54%	**90.72%**
Quantity Disagreement	16.31%	16.41%	11.32%	**7.24%**
Allocation Disagreement	15.79%	15.93%	12.15%	**6.22%**

Table 4. Per-class and overall land use accuracy comparison between MRF, OBIA-SVM, Pixel-wise CNN, and the proposed JDL-LU method for S1. The quantity disagreement and allocation disagreement are also shown. The largest classification accuracy and the smallest disagreement are highlighted in bold font.

Land Use Class (S1)	MRF	OBIA-SVM	CNN	JDL-LU
Commercial	70.06%	72.84%	73.24%	**82.42%**
Highway	77.24%	78.06%	76.15%	**79.65%**
Industrial	67.25%	69.03%	71.21%	**84.73%**

High-density residential	81.56%	80.38%	80.02%	**86.45%**
Medium-density residential	82.71%	84.37%	85.24%	**88.57%**
Park and recreational area	91.02%	93.12%	92.33%	**97.06%**
Agricultural area	85.08%	88.55%	87.43%	**90.94%**
Parking lot	78.04%	80.12%	83.75%	**91.86%**
Railway	88.05%	90.63%	86.53%	**91.89%**
Redeveloped area	89.08%	90.07%	89.24%	**90.62%**
Harbour and sea water	97.32%	98.38%	98.51%	**98.44%**
Overall Accuracy (OA)	79.38%	80.26%	84.08%	**87.58%**
Quantity Disagreement	20.66%	18.35%	14.37%	**10.28%**
Allocation Disagreement	19.97%	18.82%	14.08%	**9.67%**

Table 5 Per-class and overall land use accuracy comparison between MRF, OBIA-SVM, Pixel-wise CNN, and the proposed JDL-LU method for S2. The quantity disagreement and allocation disagreement are also shown. The largest classification accuracy and the smallest disagreement are highlighted in bold font.

Land Use Class (S2)	MRF	OBIA-SVM	CNN	JDL-LU
Commercial	71.06%	72.43%	74.13%	**82.67%**
Highway	81.41%	79.22%	80.57%	**84.25%**
Industrial	72.53%	72.08%	74.85%	**83.22%**
Residential	78.37%	80.42%	80.52%	**84.91%**
Parking lot	79.64%	82.05%	84.36%	**92.07%**
Railway	85.91%	88.17%	88.31%	**91.49%**
Park and recreational area	88.45%	89.52%	90.78%	**94.57%**
Agricultural area	84.62%	87.12%	86.54%	**91.43%**
Redeveloped area	82.54%	84.14%	87.09%	**93.74%**
Canal	90.62%	92.27%	94.16%	**98.72%**
Overall Accuracy (OA)	79.26%	80.42%	83.32%	**88.26%**
Quantity Disagreement	19.45%	17.08%	14.29%	**9.84%**
Allocation Disagreement	18.76%	16.55%	13.68%	**8.48%**

With respect to the LU classification, the proposed JDL-LU achieved excellent classification accuracy for the majority of LU classes at both S1 (Table 4) and S2 (Table 5). Five LU classes, including Park and recreational area, Parking lot, Railway, Redeveloped area in both study sites, as well as Harbour and sea water in S1 and Canal in S2, achieved very high accuracy using the proposed JDL-LU method (larger than 90% mapping accuracy), with up to 98.44% for Harbour and sea water, 98.72% for Canal, and an average of 95.82% for the Park and recreational area. In comparison with other benchmarks, significant increases were achieved for complex LU classes using the proposed JDL-LU method, with an increase in accuracy of 12.36% and 11.61% for the commercial areas, 17.48% and 10.69% for industrial areas, and 13.82% and 12.43% for the parking lot in S1 and S2, respectively. Besides, a moderate increase in accuracy was obtained for the class of park and recreational areas and the residential areas (either high-density or medium-density), with around 6% increase in accuracy for both S1 and S2. Other LU classes with relatively simple structures, including highway, railway, and redeveloped area, demonstrate no significant increase with the proposed JDL-LU method, with less than 3% accuracy increase relative to other benchmark comparators.

3.3.5 Model Robustness with Respect to Sample Size

To further assess the model robustness and generalisation capability, the overall accuracies for both LC and LU classifications at S1 and S2 were tested using reduced per-class training set sample sizes of 10%, 30%, and 50% (Figure 10), with the boxplots showing the mean classification accuracy with a 95% confidence internal. The average overall accuracy (i.e. the mean value of the boxplot) for each training set was reported through a repetition of 10 different training samples, to demonstrate statistical robustness. Similar patterns in overall accuracy as a function of sample size reduction were observed for S1 and S2. From Figure 10, it is clear that JDL-LC and JDL-LU are the least sensitive methods to reduced sample size, with no significant decrease in terms of overall accuracies while 50% of the training samples were used.

Thus, the proposed JDL method demonstrates the greatest model robustness and the least sample size requirement in comparison with other benchmark approaches (Figure 10).

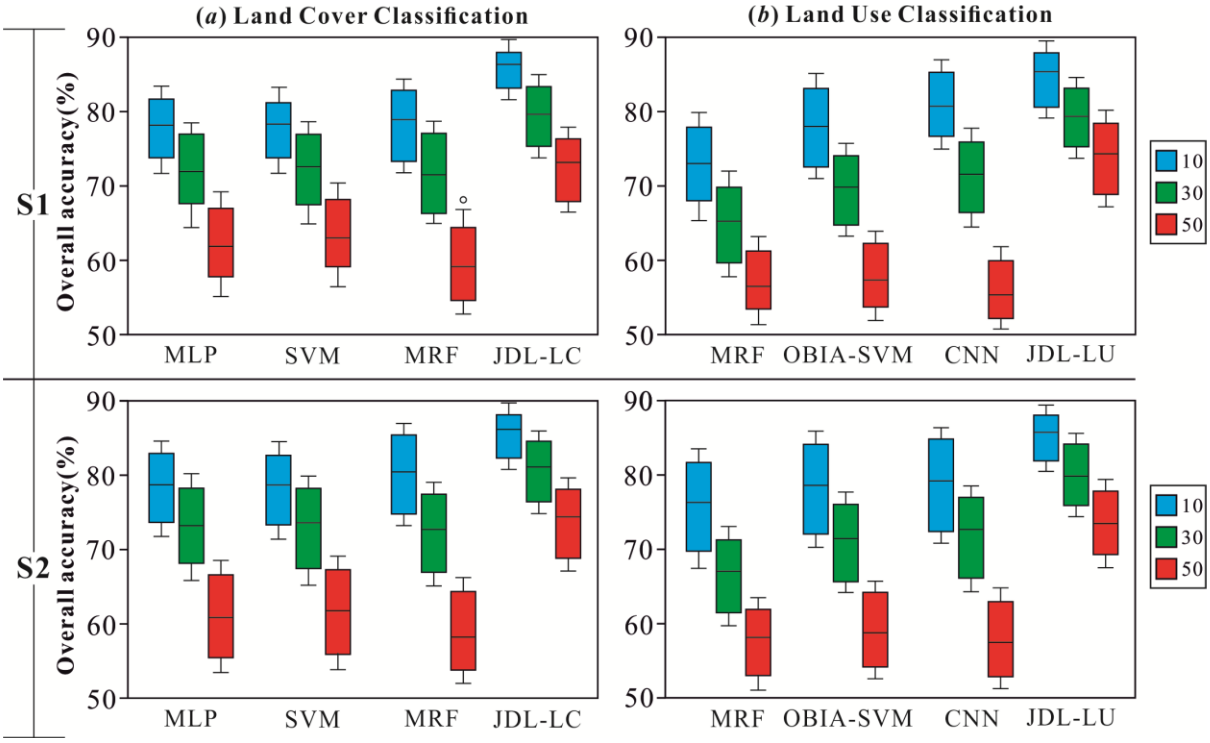

Figure 10 The effect of reducing sample size (50%, 30%, and 10% of the original training sample size per class) on the accuracy of (*a*) land cover classification (JDL-LC) and (*b*) land use classification (JDL-LU), and their respective benchmark comparators at study sites S1 and S2. The boxplot here represents the mean classification accuracy with a 95% confidence interval.

For the LC classification (Figure 10(*a*)), the accuracy distributions of the MLP and SVM were similar, although the SVM was slightly less sensitive to sample size reduction than the MLP, with about 1% higher OA for the 50% sample size reduction. The MRF was the most sensitive method to LC sample reduction, with less than 60% in OA for both S1 and S2 in terms of 50% sample size. The JDL-LC was the least sensitive to the reduction of training sample size, with an average around 88%, 80%, and 73% in the two study areas for the 10%, 30%, and 50% of sample size reduction, respectively, far outperforming the benchmarks in terms of model robustness (Figure 10(*a*)).

In terms of the LU classification (Figure 10(*b*)), the CNN was most sensitive to sample size reduction, with the lowest OA (53% and 56%) when 50% samples were used in S1 and S2, respectively. MRF and OBIA-SVM were less sensitive to sample size reduction than the CNN, with an OA close to 60% in average while reducing the sample size to 50%. The JDL-LU, however, demonstrated the most stable performance with respect to sample size reduction, achieving a high overall accuracy in average at study sites S1 and S2, with about 85.5%, 80%, and 73% for the sample size reduction of 10%, 30%, and 50%, respectively.

4. Discussion

This paper proposed a Joint Deep Learning (JDL) model to characterise the spatial and hierarchical relationship between LC and LU. The complex, nonlinear relationship between two classification schemes was fitted through a joint probability distribution such that the predictions were used to update each other iteratively to approximate the optimal solutions, in which both LC and LU classification results were obtained with the highest classification accuracies (iteration 10 in our experiments) for the two study sites. This JDL method provides a general framework to jointly classify LC and LU from remotely sensed imagery in an automatic fashion without formulating any 'expert rules' or domain knowledge.

4.1 Joint deep learning model

The joint deep learning was designed to model the joint distributions between LC and LU, in which different feature representations were bridged to characterise the same reality. Figure 11(*a*) illustrates the distributions of LC (in red) and LU (in blue) classifications, with the conditional dependency captured through joint distribution modelling (in green) to infer the underlying causal relationships. The probability distribution of the LC within the JDL framework was derived by a pixel-based MLP classifier as $P(C_{LC}|LU\text{-}Result, Image)$; that is, the LC classification was conditional upon the LU results together with the original remotely sensed images. In contrast, the distribution of LU deduced by the CNN model (object-based

CNN) was represented as a conditional probability, $P(C_{LU}|LC\text{-}Result)$, associated with the LU classification and the conditional probabilities of the LC result. The JDL method was developed based on Bayesian statistics and inference to model the spatial dependency over geographical space. We do not consider any prior knowledge relative to the joint probability distribution, and the conditional probabilities were deduced by MLP and CNN for joint model predictions and decision-making. Increasing trends were demonstrated for the classification accuracy of both LC and LU in the two distinctive study sites at each iteration (Figure 4), demonstrating the statistical fine-tuning process of the proposed JDL. To the best of our knowledge, the joint deep learning between LC and LU developed in this research is completely novel in the remote sensing community and is a profound contribution that has implications for the way that LU-LC classification should be performed in remote sensing and potentially in other fields. Previously in remote sensing only a single classification hierarchy (either LC or LU) was modelled and predicted, such as via the Markov Random Field with Gibbs joint distribution for LC characterisation (e.g. Schindler, 2012; Zheng and Wang, 2015; Hedhli et al., 2016). They are essentially designed to fit a model that can link the land cover labels x to the observations y (e.g. satellite data) by considering the spatial contextual information (through a local neighbourhood) (Hedhli et al., 2016). Our model follows the same principle of Markov theory, but aims to capture the latent relationships between LC classification ($y1$) and LU classification ($y2$) through their joint distribution. The JDL model was applied at the pixel level and classification map level to connect effectively the ontological knowledge at the different levels (e.g. LC and LU in this case). Essentially, the deep learning method (CNN) plays a fundamental role within the JDL framework formulated as part of an iterative Markov process, where the spatial patterns are characterised through hierarchical feature representations. Some previous work has recognised that an iterative classification process could potentially lead to high accuracy, for example, the multi-process classification using spatial context (road structure, morphology) (Mountrakis and Luo, 2011), and the iterative OBIA (spectra, texture and shape) by integrating bottom-up classification and top-down feedback (Zhang et al., 2018). Their methods are, however, based on traditional human-designed features or rules that are subject to user knowledge and expertise, whereas this JDL model incorporates deep learning to automatically extract spatial and hierarchical features, and to model the classification hierarchies through the joint distribution. The proposed methodology offers a new outlook and an important contribution to the remote sensing community by integrating the deep learning method (CNN), as the most appropriate approach to higher-order land use classification, into the iterative joint modelling framework.

4.2 Mutual Benefit of MLP and CNN Classification

The pixel-based multilayer perceptron (MLP) has the capacity to identify pixel-level LC class purely from spectral characteristics, in which the boundary information can be precisely delineated with spectral differentiation. However, such a pixel-based method cannot guarantee high classification accuracy, particularly with fine spatial resolution, where single pixels quickly lose their thematic meaning and discriminative capability to separate different LC classes (Xia et al., 2017). Spatial information from a contextual neighbourhood is essential to boost classification performance. Deep convolutional neural networks (CNN), as a contextual-based classifier, integrate image patches as input feature maps, with high-level spatial characteristics derived through hierarchical feature representations, which are directly associated with LU with complex spatial structures and patterns. However, CNN models are essentially patch-wise models applied across the entire image and are dependent upon the specific scale of representation, in which boundaries and small linear features may be either blurred or completely omitted throughout the convolutional processes. Therefore, both the pixel-based MLP and patch-based CNN exhibit pros and cons in LC and LU classification.

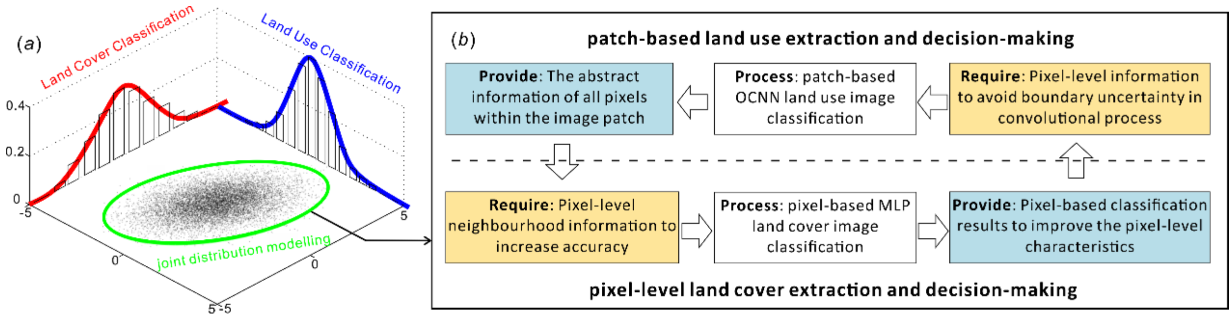
(a)
Land Cover Classification
Land Use Classification
joint distribution modelling
0.4
0.2
0
-5
0
5
-5
0
5
(b)
patch-based land use extraction and decision-making
Provide: The abstract information of all pixels within the image patch
Process: patch-based OCNN land use image classification
Require: Pixel-level information to avoid boundary uncertainty in convolutional process
Require: Pixel-level neighbourhood information to increase accuracy
Process: pixel-based MLP land cover image classification
Provide: Pixel-based classification results to improve the pixel-level characteristics
pixel-level land cover extraction and decision-making

Figure 11 Joint deep learning with joint distribution modelling (*a*) through iterative process for pixel-level land cover (LC) and patch-based land use (LU) extraction and decision-making (*b*).

The major breakthrough of the proposed JDL framework is the interaction between the pixel-based LC and patch-based LU classifications, realised by borrowing information from each other in the iterative updating process. Within the JDL, the pixel-based MLP was used for spectral differentiation amongst distinctive LCs, and the CNN model was used to identify different LU objects through spatial feature representations. Their complementary information was captured and shared through joint distribution modelling to refine each prediction through iteration, ultimately to increase classification accuracy at *both* levels. This iterative process is illustrated in Figure 11(*b*) as a cyclic graph between pixel-level LC and patch-based LU extractions and decision-making. The method starts with pixel-based classification using MLP applied to the original image to obtain the pixel-level characteristics (LC). Then this information (LC conditional probabilities) was fed into the LU classification using the CNN model as part of modelling the joint distributions between LC and LU, and to infer LU categories through patch-based contextual neighbourhoods. Those LU conditional probabilities learnt by the CNN and the original image were re-used for LC classification through the MLP classifier with spectral and spatial representations. Such refinement processes are mutually beneficial for both classification levels. For the LU classes predicted by the CNN model, the JDL is a bottom-up procedure respecting certain hierarchical relationships which allows gradual generalisation towards more abstract feature representations within the image patches.

This leads to strong invariance in terms of semantic content, with the increasing capability to represent complex LU patterns. For example, the parking lot was differentiated from the highway step-by-step with increasing iteration, and the commercial and industrial LUs with complex structures were distinguished through the process. However, such deep feature representations are often at the cost of pixel-level characteristics, which give rise to uncertainties along the boundaries of objects and small linear features, such as small paths. The pixel-based MLP classifier was used here to offer the pixel-level information for the LC classification within the neighbourhood to reduce such uncertainties. The MLP within the JDL incorporated both spectral (original image) and the contextual information (learnt from the LU hierarchy) through iteration to strengthen the spatial-spectral LC classification and produce a very high accuracy. For example, the misclassified shadows in the image were gradually removed with increasing iteration via contextual information, and the huge spectral confusion amongst different LCs, such as between concrete roof and asphalt, was successively reduced through the JDL. Meanwhile, an increasingly accurate LC classification via increasing iteration was (re)introduced into the CNN model, which re-focused the starting point of the CNN within the Joint Deep Learning back to the pixel level before convolving with small convolutional filters (3×3). As a consequence, ground features with diverse scales of representations were characterised, in which small features and boundary information were preserved in the LU classification. For example, the canal (a linear feature) was clearly identified in S2 (Figure 8).

From an artificial intelligence perspective, the JDL mimics the human visual interpretation, combining information from different levels to increase semantic meaning via joint and automatic reinforcement. Such joint reinforcement through iteration has demonstrated reduced sample size requirement and enhanced model robustness compared with standard CNN models (Figure 10), which has great generalisation capability and practical utility. There are some other techniques such as Generative Adversarial Networks (GANs) that are developed for continuous adversarial learning to enhance the capability of deep learning models, but in a competitive fashion. Therefore, the joint reinforcement in JDL has great potential to influence the future development of AI and machine learning, and the further application in machine vision.

5. Conclusions

Land cover (LC) and land use (LU) are intrinsically hierarchical representing different semantic levels and different scales, but covering the same continuous geographical space. In this paper, a novel joint deep learning (JDL) framework, that involves both the MLP and CNN classification models, was proposed for *joint* LC and LU classification. In the implementation of this JDL, the spatial and hierarchical relationships between LC and LU were modelled via a Markov process using iteration. The proposed JDL framework represents a new paradigm in remote sensing classification in which the previously separate goals of LC (state; what is there?) and LU (function; what is going on there?) are brought together in a single unifying framework. In this JDL, the pixel-based MLP low-order representation and the patch-based CNN higher-order representation interact and update each other iteratively, allowing the refinement of both the LC *and* LU classifications with mutual complementarity and joint improvement.

The classification of LC and LU from VFSR remotely sensed imagery remains a challenging task due to high spectral and spatial complexity of both. Experimental results in two distinctive urban and suburban environments, Southampton and Manchester, demonstrated that the JDL achieved by far the most accurate classifications for both LC *and* LU, and consistently outperformed the benchmark comparators, which is a striking result. In particular, complex LC classes covered by shadows that were extremely difficult to characterise were distinguished precisely, and complex LU patterns (e.g. parking lots) were recognised accurately. Therefore, this research effectively addresses the complex LC and LU classification task using VFSR remotely sensed imagery in a joint and automatic manner.

The MLP- and CNN-based JDL provides a general framework to jointly learn hierarchical representations at a range of levels and scales, not just at the two levels associated with LC and LU. For example, it is well known that LC can be defined at multiple levels as a set of states nested within each other (e.g. woodland can be split into deciduous and coniferous woodland). Likewise, and perhaps more interestingly, LU can be defined at multiple levels nested within each other to some degree. For example, a golf course is a higher-order and larger area representation than a golf shop and golf club house, both of which are LUs but nest within the golf course. The JDL proposed here should be readily generalisable to these more complex ontologies. In the future, we also aim to expand the JDL framework to other data sources (e.g. Hyperspectral, SAR, and LiDAR data) and to further test the generalisation capability and model transferability to other regions. The corresponding accuracy assessment framework would be consolidated by designing and implementing a fully generalisable approach. It is also of interest to place the JDL framework in a time-series setting for LC and LU change detection and simulation. These topics will be the subject of future research.

Acknowledgements

This research was funded by UK PhD Studentship "Deep Learning in massive area, multi-scale resolution remotely sensed imagery", sponsored by Ordnance Survey and Lancaster University (NO. EAA7369). The authors thank three anonymous referees for their constructive comments on this manuscript. The authors would also like to thank Dr Tiejun Wang from Faculty of ITC, University of Twente for discussions on accuracy assessment framework.

Reference

Arel, I., Rose, D.C., Karnowski, T.P., 2010. Deep machine learning - A new frontier in artificial intelligence research. IEEE Comput. Intell. Mag. 5, 13–18. https://doi.org/10.1109/MCI.2010.938364

Atkinson, P.M., Tatnall, A.R.L., 1997. Introduction Neural networks in remote sensing. Int. J. Remote Sens. 18, 699–709. https://doi.org/10.1080/014311697218700

Barr, S.L., Barnsley, M.J., 1997. A region-based, graph- theoretic data model for the inference of second-order thematic information from remotely-sensed images. Int. J. Geogr. Inf. Sci. 11, 555–576. https://doi.org/10.1080/136588197242194

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis - towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87, 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014

Cassidy, L., Binford, M., Southworth, J., Barnes, G., 2010. Social and ecological factors and land-use land-cover diversity in two provinces in Southeast Asia. J. Land Use Sci. 5, 277–306. https://doi.org/10.1080/1747423X.2010.500688

Chen, X., Xiang, S., Liu, C.-L., Pan, C.-H., 2014. Vehicle detection in satellite images by hybrid deep Convolutional Neural Networks. IEEE Geosci. Remote Sens. Lett. 11, 1797–1801. https://doi.org/10.1109/LGRS.2014.2309695

Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P., 2016. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. IEEE Trans. Geosci. Remote Sens. 54, 6232–6251. https://doi.org/10.1109/TGRS.2016.2584107

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C., 2017. Automatic road detection and centerline extraction via cascaded end-to-end Convolutional Neural Network. IEEE

Trans. Geosci. Remote Sens. 55, 3322–3337. https://doi.org/10.1109/TGRS.2017.2669341

Del Frate, F., Pacifici, F., Schiavon, G., Solimini, C., 2007. Use of neural networks for automatic classification from high-resolution images. IEEE Trans. Geosci. Remote Sens. 45, 800–809. https://doi.org/10.1109/TGRS.2007.892009

Dong, Z., Pei, M., He, Y., Liu, T., Dong, Y., Jia, Y., 2015. Vehicle type classification using unsupervised Convolutional Neural Network. IEEE Trans. Intell. Transp. Syst. 16, 2247–2256. https://doi.org/10.1109/ICPR.2014.39

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sens. 9. https://doi.org/10.3390/rs9050498

Hedhli, I., Moser, G., Zerubia, J., Serpico, S.B., 2016. A New Cascade Model for the Hierarchical Joint Classification of Multitemporal and Multiresolution Remote Sensing Data. IEEE Trans. Geosci. Remote Sens. 54, 6333–6348. https://doi.org/10.1109/TGRS.2016.2580321

Herold, M., Liu, X., Clarke, K.C., 2003. Spatial Metrics and Image Texture for Mapping Urban Land Use. Photogramm. Eng. Remote Sens. 69, 991–1001. https://doi.org/10.14358/PERS.69.9.991

Hester, D.B., Cakir, H.I., Nelson, S. a C., Khorram, S., 2008. Per-pixel Classification of High Spatial Resolution Satellite Imagery for Urban Land-cover Mapping. Photogramm. Eng. Remote Sens. 74, 463–471.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep Convolutional Neural Networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7, 14680–14707. https://doi.org/10.3390/rs71114680

Hu, S., Wang, L., 2013. Automated urban land-use classification with remote sensing. Int. J. Remote Sens. 34, 790–803. https://doi.org/10.1080/01431161.2012.714510

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep Convolutional Neural Networks, in: NIPS2012: Neural Information Processing Systems. Lake Tahoe, Nevada, pp. 1–9.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., Hong, Y., 2017. Classifying urban land use by integrating remote sensing and social media data. Int. J. Geogr. Inf. Sci. 31, 1675–1696. https://doi.org/10.1080/13658816.2017.1324976

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional Neural Networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657. https://doi.org/10.1109/TGRS.2016.2612821

Mas, J.F., Flores, J.J., 2008. The application of artificial neural networks to the analysis of remotely sensed data. Int. J. Remote Sens. 29, 617–663. https://doi.org/10.1080/01431160701352154

McRoberts, R.E., 2013. Post-classification approaches to estimating change in forest area using remotely sensed auxiliary data. Remote Sens. Environ. 151, 149–156. https://doi.org/10.1016/j.rse.2013.03.036

Ming, D., Li, J., Wang, J., Zhang, M., 2015. Scale parameter selection by spatial statistics for GeOBIA: Using mean-shift based multi-scale segmentation as an example. ISPRS J. Photogramm. Remote Sens. 106, 28–41. https://doi.org/10.1016/j.isprsjprs.2015.04.010

Mountrakis, G., Luo, L., 2011. Enhancing and replacing spectral information with intermediate structural inputs: A case study on impervious surface detection. Remote Sens. Environ. 115, 1162–1170. https://doi.org/10.1016/j.rse.2010.12.018

Myint, S.W., 2001. A robust texture analysis and classification approach for urban land-use and land-cover feature discrimination. Geocarto Int. 16, 29–40. https://doi.org/10.1080/10106040108542212

Niemeyer, J., Rottensteiner, F., Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. ISPRS J. Photogramm. Remote Sens. 87, 152–165. https://doi.org/10.1016/j.isprsjprs.2013.11.001

Nogueira, K., Penatti, O.A.B., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognit. 61, 539–556. https://doi.org/10.1016/j.patcog.2016.07.001

Oliva-Santos, R., Maciá-Pérez, F., Garea-Llano, E., 2014. Ontology-based topological representation of remote-sensing images. Int. J. Remote Sens. 35, 16–28. https://doi.org/10.1080/01431161.2013.858847

Olofsson, P., Foody, G.M., Herold, M., Stehman, S. V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Othman, E., Bazi, Y., Alajlan, N., Alhichri, H., Melgani, F., 2016. Using convolutional features and a sparse autoencoder for land-use scene classification. Int. J. Remote Sens. 37, 2149–2167. https://doi.org/10.1080/01431161.2016.1171928

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. https://doi.org/10.1109/JSTARS.2016.2582921

Pan, X., Zhao, J., 2017. A central-point-enhanced convolutional neural network for high-resolution remote-sensing image classification. Int. J. Remote Sens. 38, 6554–6581. https://doi.org/10.1080/01431161.2017.1362131

Patino, J.E., Duque, J.C., 2013. A review of regional science applications of satellite remote sensing in urban settings. Comput. Environ. Urban Syst. 37, 1–17. https://doi.org/10.1016/j.compenvurbsys.2012.06.003

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M.A., Ouzounis, G.K., Scavazzon, M., Soille, P., Syrris, V., Zanchetta, L., 2013. A global human settlement layer from optical HR/VHR RS data: Concept and first results. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 2102–2131. https://doi.org/10.1109/JSTARS.2013.2271445

Pontius, R.G., Millones, M., 2011. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. 32, 4407–4429. https://doi.org/10.1080/01431161.2011.552923

Regnauld, N., Mackaness, W. a., 2006. Creating a hydrographic network from its cartographic representation: a case study using Ordnance Survey MasterMap data. Int. J. Geogr. Inf. Sci. 20, 611–631. https://doi.org/10.1080/13658810600607402

Romero, A., Gatta, C., Camps-valls, G., Member, S., 2016. Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 54, 1349–1362. https://doi.org/10.1109/TGRS.2015.2478379.

Salehi, B., Zhang, Y., Zhong, M., Dey, V., 2012. A review of the effectiveness of spatial information used in urban land cover classification of VHR imagery. Int. J.

Geoinformatics 8, 35–51.

Schindler, K., 2012. An Overview and Comparison of Smooth Labeling Methods for Land-Cover Classification. Geosci. Remote Sensing, IEEE Trans. 50, 4534–4545. https://doi.org/10.1109/TGRS.2012.2192741

Strigl, D., Kofler, K., Podlipnig, S., 2010. Performance and scalability of GPU-based Convolutional Neural Networks, in: 2010 18th Euromicro Conference on Parallel, Distributed and Network-Based Processing. pp. 317–324. https://doi.org/10.1109/PDP.2010.43

Verburg, P.H., Neumann, K., Nol, L., 2011. Challenges in using land use and land cover data for global change studies. Glob. Chang. Biol. 17, 974–989. https://doi.org/10.1111/j.1365-2486.2010.02307.x

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55, 881–893. https://doi.org/10.1109/TGRS.2016.2616585

Walde, I., Hese, S., Berger, C., Schmullius, C., 2014. From land cover-graphs to urban structure types. Int. J. Geogr. Inf. Sci. 28, 584–609. https://doi.org/10.1080/13658816.2013.865189

Wu, S.S., Qiu, X., Usery, E.L., Wang, L., 2009. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. Ann. Assoc. Am. Geogr. 99, 76–98. https://doi.org/10.1080/00045600802459028

Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55, 3965–3981.

https://doi.org/10.1109/TGRS.2017.2685945

Yoshida, H., Omae, M., 2005. An approach for analysis of urban morphology: methods to derive morphological properties of city blocks by using an urban landscape model and their interpretations. Comput. Environ. Urban Syst. 29, 223–247. https://doi.org/10.1016/j.compenvurbsys.2004.05.008

Zhang, C., Atkinson, P.M., 2016. Novel shape indices for vector landscape pattern analysis. Int. J. Geogr. Inf. Sci. 30, 2442–2461. https://doi.org/10.1080/13658816.2016.1179313

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., Atkinson, P.M., 2018a. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. ISPRS J. Photogramm. Remote Sens. 140, 133–144. https://doi.org/10.1016/j.isprsjprs.2017.07.014

Zhang, C., Sargent, I., Pan, X., Gardiner, A., Hare, J., Atkinson, P.M., 2018b. VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images. IEEE Trans. Geosci. Remote Sens. 56, 4507–4521. https://doi.org/10.1109/TGRS.2018.2822783

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018c. An object-based convolutional neural networks (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70. https://doi.org/https://doi.org/10.1016/j.rse.2018.06.034

Zhang, C., Wang, T., Atkinson, P.M., Pan, X., Li, H., 2015. A novel multi-parameter support vector machine for image classification. Int. J. Remote Sens. 36, 1890–1906. https://doi.org/10.1080/01431161.2015.1029096

Zhang, X., Du, S., Wang, Q., 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sens. Environ. 212, 231–248. https://doi.org/10.1016/j.rse.2018.05.006

Zhao, B., Zhong, Y., Zhang, L., 2016. A spectral-structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 116, 73–85. https://doi.org/10.1016/j.isprsjprs.2016.03.004

Zhao, W., Du, S., Wang, Q., Emery, W.J., 2017. Contextually guided very-high-resolution imagery classification with semantic segments. ISPRS J. Photogramm. Remote Sens. 132, 48–60. https://doi.org/10.1016/j.isprsjprs.2017.08.011

Zheng, C., Wang, L., 2015. Semantic Segmentation of Remote Sensing Imagery Using Object-Based Markov Random Field Model With Regional Penalties. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 1924–1935.

Figure Captions

Figure 1 The general workflow of the land cover (LC) and land use (LU) joint deep learning (JDL).

Figure 2 The two study areas: S1 (Southampton) and S2 (Manchester) with highlighted regions representing the majority of land use categories.

Figure 3 Model architectures and structures of the CNN with 96×96 input window size and eight-layer depth.

Figure 4 The overall accuracy curves for the Joint Deep Learning iteration of land cover (LC) and land use (LU) classification results in S1 and S2. The red dash line indicates the optimal accuracy for the LC and LU classification at iteration 10.

Figure 5 Four subset land cover classification results in S1 using Joint Deep Learning – Land cover (JDL-LC), the best results at iteration 10 were highlighted with blue box. The circles in yellow and red represent the correct and incorrect classification, respectively.

Figure 6 The land cover classification results in S2 using Joint Deep Learning – Land cover (JDL-LC), the best results at (h) iteration 10 were highlighted with blue box.

Figure 7 Four subset land use classification results in S1 using Joint Deep Learning – Land use (JDL-LU), the best results at iteration 10 were highlighted with blue box. The circles in yellow and red represent the correct and incorrect classification, respectively.

Figure 8 The land use classification results in S2 using Joint Deep Learning – Land use (JDL-LU), the best results at (h) iteration 10 were highlighted with blue box.

Figure 9 Overall accuracy comparisons among the MLP, SVM, MRF, and the proposed JDL-LC for land cover classification, and the MRF, OBIA-SVM, CNN, and the proposed JDL-LU for land use classification.

Figure 10 The effect of reducing sample size (50%, 30%, and 10% of the original training sample size per class) on the accuracy of (*a*) land cover classification (JDL-LC) and (*b*) land use classification (JDL-LU), and their respective benchmark comparators at study sites S1 and S2. The boxplot here represents the mean classification accuracy with a 95% confidence interval.

Figure 11 Joint deep learning with joint distribution modelling (*a*) through iterative process for pixel-level land cover (LC) and patch-based land use (LU) extraction and decision-making (*b*).