Deep Learning for Land Cover and Land Use Classification

Ce Zhang *BSc*, *MSc*, *MSc*

Lancaster Environment Centre, Lancaster University

This thesis is submitted in fulfilment of the regulations for the degree of

Doctor of Philosophy

August 2018

To my family

Abstract

Recent advances in sensor technologies have witnessed a vast amount of very fine spatial resolution (VFSR) remotely sensed imagery being collected on a daily basis. These VFSR images present fine spatial details that are spectrally and spatially complicated, thus posing huge challenges in automatic land cover (LC) and land use (LU) classification. Deep learning reignited the pursuit of artificial intelligence towards a general purpose machine to be able to perform any human-related tasks in an automated fashion. This is largely driven by the wave of excitement in deep machine learning to model the high-level abstractions through hierarchical feature representations without human-designed features or rules, which demonstrates great potential in identifying and characterising LC and LU patterns from VFSR imagery. In this thesis, a set of novel deep learning methods are developed for LC and LU image classification based on the deep convolutional neural networks (CNN) as an example. Several difficulties, however, are encountered when trying to apply the standard pixel-wise CNN for LC and LU classification using VFSR images, including geometric distortions, boundary uncertainties and huge computational redundancy. These technical challenges for LC classification were solved either using rule-based decision fusion or through uncertainty modelling using rough set theory. For land use, an object-based CNN method was proposed, in which each segmented object (a group of homogeneous pixels) was sampled and predicted by CNN with both within-object and between-object information. LU was, thus, classified with high accuracy and efficiency. Both LC and LU formulate a hierarchical ontology at the same geographical space, and such representations are modelled by their joint distribution, in which LC and LU are classified simultaneously through iteration. These developed deep learning techniques achieved by far the highest classification accuracy for both LC and LU, up to around 90% accuracy, about 5% higher than the existing deep learning methods, and 10% greater than traditional pixel-based and object-based approaches. This research made a significant contribution in LC and LU classification through deep learning based innovations, and has great potential utility in a wide range of geospatial applications.

Keywords: deep learning, land cover classification, land use classification, very fine spatial resolution, remotely sensed imagery, hierarchical representations

Declaration for Authorship

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisors Professor Peter M. Atkinson (Lancaster University), Dr Isabel Sargent (Ordnance Survey), and Dr Jonathon Hare (University of Southampton).

Excerpts of this thesis have been published in the following peer-reviewed journal articles as part of the project outcomes for the three-year PhD Studentship (2015 – 2018) titled: "Deep Learning in Massive Area, Multi-scale Resolution Remotely Sensed Imagery" funded by Ordnance Survey and Lancaster University (EAA7369).

1) **Ce Zhang ***, Peter M. Atkinson, 2016, Novel shape indices for vector landscape pattern analysis. *International Journal of Geographical Information Science*, 30(12): 2442-2461. https://doi.org/10.1080/13658816.2016.1179313

2) **Ce Zhang *,** Xin Pan, Shuqing Zhang, Huapeng Li, Peter M. Atkinson, 2017, A rough set decision tree based MLP-CNN for very high resolution remotely sensed image classification. *International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, XLII-2/W7, 1451-1454, doi:10.5194/isprs-archives-XLII-2-W7-1451-2017. https://doi.org/10.5194/isprs-archives-XLII-2-W7-1451-2017

3) **Ce Zhang ***, Xin Pan, Huapeng Li, Andy Gardiner, Isabel Sargent, Jonathon Hare, Peter M. Atkinson *, 2018a, A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140: 133-144. https://doi.org/10.1016/j.isprsjprs.2017.07.014 **(Chapter 3)**

4) **Ce Zhang ***, Isabel Sargent, Xin Pan, Andy Gardiner, Jonathon Hare, Peter M. Atkinson *, 2018b, VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4507-4521.

https://doi.org/10.1109/TGRS.2018.2822783 **(Chapter 4)**

5) **Ce Zhang ***, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, Peter M. Atkinson *, 2018c, An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sensing of Environment*, 216:57-70.

https://doi.org/10.1016/j.rse.2018.06.034 **(Chapter 5)**

6) **Ce Zhang ***, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, Peter M. Atkinson *, 2018d, Joint deep Learning for land cover and land use classification. *Remote Sensing of Environment*. (Under 2nd round review) **(Chapter 6)**

These published papers were produced in a strategic partnership with Lancaster University and Ordnance Survey (Britain's National Mapping Agency) as well as wider supervision team at University of Southampton.

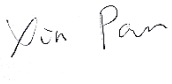
I declare that the publications are purely my own work as the first authorship. The aerial photography and ground supporting data used in this PhD research were provided by Ordnance Survey. The ideas were formulated by myself together with my primary supervisor Professor Peter M. Atkinson, and discussed and approved by other supervisors and research and development (R&D) team at Ordnance Survey. I designed and executed experiments, and drafted the entire papers. The experiments were partially supported by external collaboration with Dr Xin Pan from School of Computer Technology and Engineering, Changchun Institute of Technology, China, and Dr Huapeng Li and Professor Shuqing Zhang from Northeast Institute of Geography and

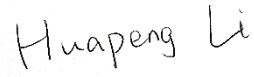
Agroecology, Chinese Academic of Science, China. All these papers were submitted by myself as the corresponding authors together with Professor Peter M. Atkinson. The co-authors of these publications have signed below to confirm this.

Yours sincerely

Signed:

Isabel Sargent - 

Xin Pan - 

Huapeng Li - 

Shuqing Zhang - 

Andy Gardiner - 

Jonathon Hare - 

Peter M. Atkinson - 

Acknowledgements

When I look back the past three years, I would say that doing a PhD is full of challenges and uncertainties, but finally I have found the way to handle it. Finding this way was a long journey. I am sincerely grateful to many people and organisations for all their support to help me pass through this journey.

First of all, I would like to thank to Ordnance Survey and Lancaster University for sponsoring my PhD with the fully-funded Studentship. Ordnance Survey is a great industrial partner to work with all the way throughout my PhD project. I worked with the Business Change & Innovation team for research and development (R&D), and discussed with so many people to strengthen the industrial and commercial value of my PhD. I would therefore like to take this opportunity to thank Ordnance Survey for supplying the aerial imagery and supporting ground data as well as their continuous support throughout the progression on this PhD research.

Foremost, I would like to express my greatest gratitude and appreciation to my primary supervisor, Professor Peter M. Atkinson for his valuable guidance, constructive feedbacks and comments from the very beginning till the completion of this PhD. During the past three years, working with Professor Pete is a wonderful experience. His wide knowledge, strong research enthusiasm and hard-working attitude have inspired me during my PhD, and will have a profound effect on my academic career forever. This PhD research could not have finished without his support, guidance, patience and encouragement all the way. I am extremely lucky to have such an excellent supervisor and mentor as well as an interesting friend. Pete, you trained me to become a highly technically competent researcher to produce top journal articles that are of the highest international quality with real-world impact. You always gave me critical and

constructive comments and feedbacks on my writing within a short amount of time. I still remember that you direct me to shift the research focus from pixel-based classification of land cover (flogging the *dead horse*) to the higher-order land use features (driving the *Ferrari*), and again re-think about the intrinsic relationship between these two classification levels in terms of hierarchical representations. These work has great potential to be the most important contributions in land cover and land use classification over the past decades. It was a real pleasure working with Pete and we will continue working together for my postdoctoral research and thereafter.

I would like to extend my gratitude to my industrial advisor Dr Isabel Sargent from Ordnance Survey, for her help, support, and trust on my ability to conduct research with significant commercial impact. She was able to provide constructive feedbacks and comments on my paper writing and enhance the quality of my work from geospatial machine learning perspectives. Izzy, I thank you very much for your mentoring and guidance on bridging academia and industry. Whenever I visit Ordnance Survey, you always host me like home and make time to me. I would like to thank Mr Andy Gardiner as my daily industrial advisor from Ordnance Survey to support me with aerial imagery and ground data. Andy is always kind to me and he particularly focuses on the industrial perspective to encourage me to embed my work into Ordnance Survey. I am grateful to all the useful discussions and catch-ups at Ordnance Survey and via Skype meetings.

I am deeply grateful to my previous colleagues (Dr Xin Pan, Dr Huapeng Li, and Professor Shuqing Zhang) from Northeast Institute of Geography and Agro-ecology, Chinese Academy of Science (CAS) for their constant supports and great academic networks. Dr Xin Pan, I thank you for supporting me with python code and all the online discussions we had. Dr Huapeng Li, thank you for visiting Lancaster University to work with Pete and myself, we had many productive and interesting conversations on both

research and daily life. I am grateful to Professor Shuqing Zhang for his continuing support. We had lots of email exchanges and Skype meetings to discuss ideas and share knowledge and expertise, which greatly inspired me to interpret the experimental results and research findings as well as scientific contributions in GIS and remote sensing. I also thank to Dr Jonathon Hare from University of Southampton. He brought many computer vision perspectives and challenged me to do engineering research in a structured manner. We had nice talks during my visits at University of Southampton. He always asked me to keep pace with the state-of-the-arts in deep learning.

Many thanks to my colleagues in Geospatial Data Science research group at Lancaster Environment Centre for the group meetings and discussions, particularly the support from other academic staff including Professor Alan Blackburn, Dr Duncan Whyatt, Dr Amber Leeson. I also thank to Xiaowei Gu from SCC for his closed collaboration and many other members in Data Science Institute that connect with me. I thank Ojoatre Sadadi for the collaboration and great discussions on the ground data in Africa to further extend my PhD work. I also would like to thank to Andy Harrod for his great administration and coordination in postgraduate research. Most importantly, I thank my girlfriend Ying Wang, who always trusts me and takes very good care of me. Thank you for being the light of my life.

Finally, my sincere gratitude goes to my parents for their moral support and unconditional love during the whole period of time. You have always special place in my heart.

Ce Zhang, Lancaster, August 2018.

List of Abbreviations

Land Cover (LC)

Land Use (LU)

Artificial Intelligence (AI)

Multilayer Perceptron (MLP)

Convolutional Neural Networks (CNN)

Very Fine Spatial Resolution (VFSR)

Markov Random Field (MRF)

Grey Level Co-occurrence Matrix (GLCM)

Support Vector Machine (SVM)

Object-based Image Analysis (OBIA)

Object-based CNN (OCNN)

Digital Surface Models (DSMs)

Moment Bounding Box (MB)

Variable Precision Rough Set (VPRS)

Joint Deep Learning (JDL)

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Project Background

Ordnance Survey, the British National Mapping Agency, acquires thousands of square kilometres of aerial photography each year to update their imagery and topographic products. For example, in 2015, 139, 000 images were acquired, covering 56, 000 km^2 of Great Britain. The volume of sensor data being flown is constantly growing in terms of increasing both spatial resolution (up to 25 cm) and temporal frequency (twice a year), owing to the improved instrumentation and the pressure from potential customers and end-users to increase data currency. The considerable majority of the ground features were captured manually through on-site survey and aerial photo interpretation, which are extremely labour-intensive and time-consuming. Arguably, this large archive of aerial imagery is highly under-utilised and could be 'mined' for much more information efficiently and effectively through modern geospatial artificial intelligence (AI) and machine learning.

1.2 Real-world Demands from Ordnance Survey

Great Britain is one of the most highly urbanised countries around the world (Bibby 2009). The urban areas in Britain developed over thousands of years of human habitation, and are still rapidly changing with fast-paced and poorly-planned urban growth (Dwyer 2011), posing grand challenges across the country: from environmental degradation and food insecurity, to unsustainable economy. Such demanding problems, caused by rapid urban development, require responsive plans and decisions from environmental planners, policy makers, and local government authorities (Hu and

Wang 2013). Accurate and up-to-date land cover and land use (LULC) information is, therefore, urgently needed to keep pace with the ever-changing urban environments and to support relevant policies and decision-making.

As a government-owned company, Ordnance Survey has experienced increasing demand from customers, including, but not excluding UK government, different stakeholders and end-users, for both more bespoke products and rapid development of new products. The expectations of customers are driven by the offerings of other geospatial players, such as Google, as well as broader technological advances such as those included under banners as "big data", "AI technologies" and "Smart Cities". One of the key requirements is to obtain reliable LULC information in a detailed, coherent and consistent approach that promotes automation as part of the Industry 4.0 revolution. Such requirements of customers are greatly motivated and emboldened by the increasing volume of sensor data, substantially improved computational resources, and state-of-the-art geospatial technologies, where the characteristics of, and changes in, urban LULC can be extracted and analysed by developing intelligent and automatic methods through technological innovations.

1.3 Broad Context and Academic Requirements

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 complex interactions between human activities and global environmental changes (Cassidy et al. 2010, Patino and Duque 2013). Many predictive models (e.g. ecosystem, hydrologic, and transportation models) involve LULC as their input variables to simulate natural and anthropogenic processes and the functioning of the Earth surface (Verburg et al.

2011). Earth observations from diverse sources, including satellite, airborne, *in situ* platforms and citizen observatories provide great opportunities to identify the characteristics of, and changes in, LULC across different scales (Anderson et al. 2017). With the rapid development of sensors and devices, large quantities of very fine spatial resolution (VFSR) remotely sensed imagery are now commercially available with sub-metre resolution, facilitating the acquisition of precise LULC information at fine spatial detail (Pesaresi et al. 2013, Zhao et al. 2016).

Land cover (LC) classification using VFSR remotely sensed data can be a very complicated task due to the spectral and spatial complexity of the imagery. Land use (LU) classification, however, is even more challenging due to the indirect relationship between land use patterns and the spectral responses recorded in images. These land uses are typically defined in terms of functions or socioeconomic activities rather than physical forms of land covers, which can only be inferred in directly through the interpretation of tone, texture or shapes of the image features (Li et al. 2016). Often, the same land use types (e.g. residential areas) are characterised by distinctive physical properties or land cover materials (e.g. composed of different roof tiles, residential gardens, etc.), and different land use categories might exhibit the same or similar reflectance spectra and textures (e.g. tarmac roads and parking lots) (Pan et al. 2013). As a consequence, such spectral and spatial complexity and heterogeneity make automatic LC and LU classification using VFSR images an extremely challenging task.

Over the past few decades, tremendous effort has been made in developing automatic LULC classification methods using VFSR remotely sensed data. These methods, particularly in terms of land cover settings, are designed primarily on the basis of spectral features reflected by the physical properties of the ground surface. Many per-

pixel classification approaches (e.g. Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Random Forest (RF), etc.) have been developed to learn the non-linear spectral feature space at the pixel level irrespective of its statistical properties (Pacifici et al. 2009, Zhang et al. 2015). However, these pixel-based methods cannot guarantee high classification accuracy, particularly at fine spatial resolution, due to the fact that single fine pixels can lose their thematic meanings and discriminative efficiency to separate different types of land covers (Xia et al. 2017). Object-based methods, under the framework of object-based image analysis (OBIA), have dominated in land cover classification using VFSR imagery over the last decade (Blaschke et al. 2014). Many studies applied OBIA approaches to obtain urban land cover information from VFSR images, by exploiting the use of spectral, textural, and geometrical information of image objects that are composed of relatively homogeneous neighbouring pixels (Myint et al. 2011). The major challenges of these object-based approaches, however, are the choice of segmentation scales to obtain objects that correspond to specific land cover types, in which over-segmentation and under-segmentation commonly exist within the same image (Ming et al. 2015). So far, existing techniques remain inadequate to analyse the data properly, and no effective solution has been proposed for land cover classification using VFSR remotely sensed imagery.

Land use (LU) classification, in comparison with land cover (LC) classification, is less explored due to the complexity in spatial composition and configuration, where the land use patterns are formed by high-level semantics or functions. For example, land cover objects can be recognised as buildings, grassland, woodland etc. based on the low-level feature descriptors (spectra, texture, shape etc.), whereas land use features are characterised as functional types with high-level semantics, such as residential, commercial, and industrial areas (Bratasanu et al. 2011, Zhong et al. 2015, Liu et al.

2017). Such disparity has resulted in a semantic gap between the 'information' coming from the data itself and the 'knowledge' specific to users and applications (Bratasanu et al. 2011). To bridge such a semantic gap, many researches have attempted to incorporate expert knowledge or ancillary data as spatial context for land use feature extraction. They have generally developed as two-step pipelines, in which object-based land covers were extracted initially, followed by aggregating these land cover objects using spatial contextual descriptive indicators defined on land use units, such as cadastral fields or street blocks (Hermosilla et al. 2012). Yet, the ancillary geographic data for specifying the land use units might not be available for some regions (Li et al. 2016), and the spatial contexts are often hard to describe and characterise as a set of "rules", even though the complex structures or patterns might be recognisable and distinguishable for human experts (Oliva-Santos et al. 2014).

1.4 Deep Learning in Remote Sensing

Recent advances in AI and machine learning, especially the emerging field of deep learning, have changed the way we process, analyse and manipulate geospatial sensor data. This is largely driven by the wave of excitement in deep machine learning, as a new frontier of AI, where the most representative and discriminative features are learnt end-to-end, 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). They have made considerable improvements beyond the state-of-the-art records in a variety of domains, and have attracted great interest in both academia and industrial communities.

The essence of deep learning is about representation learning or feature learning, where the most representative and discriminative features are learnt end-to-end, hierarchically (Chen et al. 2016). Unlike their shallow counterparts, such as support vector machine and multi-layer perceptron, deep learning methods do not rely on the prior feature extractions or human feature design, but rather learn the higher-level feature representations through models themselves to enhance the generalisation capabilities (Arel et al. 2010). In addition, the deep layers of representations have great potential to characterise robust features with complex patterns and semantics, such as land use, functional sites etc. One typical example is the railway station that is comprised of long thin platforms and long thin roofs together with a set of objects that surround it (e.g. railway lines, car park and multiple roads) (Tang et al. 2016). Deep learning methods are naturally a good fit to capture such kinds of feature representations with high-level semantics (Nogueira et al. 2017).

Over the past few years, deep learning and, in particular, deep convolutional neural networks (CNNs), have gained significant attention in the image analysis community (Krizhevsky et al. 2012, Yang et al. 2015). They were originally devised for *image* categorisation, where an image is assigned to a specific semantic category according to its content, such as natural scenes used in computer vision applications or remotely sensed land-use scenes, such as ‘airport’, ‘residential’ or ‘commercial’ (Maggiori et al. 2017). These scene-level land use classifications, however, do not meet the actual requirement of land use image classification, which expects all pixels in an entire image to be identified and labelled into land use categories.

In summary, although deep learning methods have their intrinsic advantages for learning hierarchical feature representations, their feasibility and actual utility in both

LC and LU image classification have not been explored until now, and the LU image classification, in particular, has not yet been solved due to the huge intra-class heterogeneity and inter-class similarity of land use features. In addition, both land cover and land use classifications are essentially the abstractions or generalisations of the real-world landscape. The classification systems are presented at different levels, nested within each other hierarchically over the same geographical space. Different ground features occur across different scales, and their mapping objectives are strongly dependent upon the applications of interest (Heydari and Mountrakis 2018). Until now, it is still an open question how to appropriately adopt or develop CNN-based methods to solve the complex LC and LU classification tasks using VFSR remotely sensed imagery.

1.5 Research Objectives and Questions

The main objective of this PhD thesis was to produce the most accurate land cover and land use maps that are most suitable for meeting the potential customer requirements from Ordnance Survey, such as to identify and understand land use changes in a fast changing urban environment. To reach this aim, the following specific objectives and questions are posed.

1) Develop a deep learning method for land cover classification using VFSR remotely sensed images.
   Research question: *Can a novel method be developed based on CNNs to solve complex land cover classification using VFSR remotely sensed images?*

2) Model the uncertainty in deep learning for VFSR land cover image classification.
   Research question: *Can a novel method be developed to quantify and model the uncertainty within the VFSR land cover classification using deep CNNs?*

3) Develop a deep learning method to solve the complex land use classification using VFSR remotely sensed imagery.
   Research question: *Can a novel method be developed based on CNNs to solve complex land use image classification problems using VFSR remotely sensed imagery?*

4) Develop a novel method for joint land cover and land use classification using VFSR remotely sensed imagery.
   Research question: *Can the complex hierarchical relationship between land cover and land use be modelled jointly?*
   Research question: *Can a novel method be developed for joint land cover and land use classification using VFSR remotely sensed imagery?*

1.6 Thesis Structure

This thesis is based on a number of journal articles, either already published or prepared for publication (Chapters 3-6):

Chapter 1 gives the general introduction of this thesis. It comes with commercial and academic needs and towards deep learning as the state-of-the-art methods, focusing on the challenges and opportunities for land cover and land use classification using VFSR remotely sensed imagery.

Chapter 2 provides a concise review of the traditional and deep learning based methods in land use and land cover classification using VFSR imagery, and discusses the pros and cons as well as future research directions.

Chapter 3 presents a hybrid MLP-CNN classifier for land cover image classification by using a rule-based fusion decision strategy. It was designed to solve the blurred

boundary issues in standard pixel-wise CNN methods by fusing with a pixel-based multilayer perceptron (MLP) classifier.

Chapter 4 presents uncertainty modelling for CNN-based land cover classification using rough set theory. A variable precision rough set (VPRS) model was proposed to quantify the uncertainty within the CNN classification map, and the uncertain regions were rectified by using a Markov random field (MRF) through precise segmentation and spectral differentiation.

Chapter 5 presents an object-based CNN (OCNN) for complex urban land use classification. The OCNN method was proposed to analyse a group of pixels as an object with geometry, and incorporate spatial context to classify different urban land use classes with high accuracy and efficiency.

Chapter 6 presents a Joint Deep Learning (JDL) for land cover and land use classification. Both land cover and land use classifications formulate a hierarchical ontology within the same geographical space, and such representations are modelled by their joint distribution as a Markov process, in which land cover and land use are classified simultaneously through iteration.

Chapter 7 summarises the results obtained from Chapters 3 – 6, and answers the research questions in Chapter 1, followed by reflections and future recommendations as well as concluding remarks of this thesis.

Chapter 2 Literature Review

The last decades have seen a constellation of sensors borne by diverse Earth observation platforms, including satellites, aerial systems and unmanned aerial vehicles (Othman et al. 2017). Such technological advances have resulted in significant growth in the availability of very fine spatial resolution (VFSR) imagery on a daily basis (Yao et al. 2016). Those VFSR images provide unprecedented spatial detail, which facilitate many practical applications, such as precision agriculture (Ozdarici-Ok et al. 2015), traffic monitoring (Larsen et al. 2013), and urban planning (Caccetta et al. 2016). Land cover (LC) and land use (LU) classifications are widely used to extract meaningful information for these purposes (Malinverni et al. 2011). However, LC classification from VFSR remotely sensed imagery is very complex, due to the huge within-class spectral heterogeneity caused by the differences in age, level of maintenance and composition as well as illumination conditions (Demarchi et al. 2014), which might be further complicated by the scattering of peripheral ground objects (Chen et al. 2014). Classification of land use (LU) is even more challenging, because of its indirect relationship with the physical characteristics of the Earth surface recorded by the VFSR images (Pan et al. 2013). LU refers to functions or human activities, which cannot be interpreted by using tone, texture or shapes of image features (Li et al. 2016). Indeed, the classification of LC and LU from VFSR remotely sensed images is still an open and unsolved task in the remote sensing community.

2.1 Traditional LC and LU Classification Approaches

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 be divided broadly 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 land cover categories purely based on spectral reflectance, without taking consideration of neighbouring pixels (Verburg et al. 2011). These methods are often limited in classification performance due to the speckle and increased inter-class variance in comparison with coarse or medium spatial resolution remotely sensed data. 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 could eliminate small classes with few pixels or single family classes such as houses or small scrubland areas. Object-based methods, under the framework of object-based image analysis (OBIA), have dominated in land cover 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 pixel values across the image, for the identification of land covers 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 land cover types, in which over-segmentation and under-segmentation commonly exist within the same image (Ming et al. 2015). As a result, the LC classification task is very challenging, especially for urban areas, which exhibit high intra-class variation with huge diversity of land cover objects (e.g. buildings with different roof tiles), and low inter-class disparity with similar visual

characteristics for different land cover types (e.g. buildings and roads). Meanwhile, these fine-structured objects often interact with each other through occlusions and cast shadows, which poses additional challenges to identify them precisely and accurately. To date, no effective solution has been proposed for land cover classification using VFSR remotely sensed imagery.

Similar to land cover classification, traditional land use classification methods using VFSR data can generally be categorised into three types, including pixels, moving windows, and objects. The pixel-level approaches that rely purely upon spectral characteristics are suitable for classifying land cover, but are insufficient to distinguish land uses that are typically composed of multiple land covers, and such problems are particularly significant in urban settings (Zhao et al. 2016). Spatial information, that is, texture (Myint 2001, Herold et al. 2003) or context (Wu et al. 2009), was incorporated to analyse land use patterns through moving kernel windows (Niemeyer et al. 2014). However, it could be argued that both pixel-based and moving window-based methods require predefinition of 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 land use based on spatial context. Typically, two kinds of information within a spatial partition were utilised, namely, the within-object information (e.g. spectral, texture, shape) and the between-object information (e.g. connectivity, contiguity, distances, and direction amongst adjacent objects). Many studies applied OBIA for land use 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, Hu and Wang 2013, Blaschke et al. 2014). 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 land covers were extracted initially, followed by aggregating the objects using spatial contextual descriptive indicators on well-defined land use units, such as cadastral fields or street blocks. Those descriptive indicators were 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). Nevertheless, the ancillary geographic data for specifying the land use units might not be available for some regions, and the spatial contexts are often hard to describe and characterise as a set of "rules", even though the complex structures or patterns might be recognizable and distinguishable for human experts (Oliva-Santos et al. 2014).

2.2 Problems in Traditional LC and LU Classification Approaches

Most traditional classification methods for both LC and LU, are hand engineered in feature design coupled with classifiers in shallow structure and architecture. They typically involve two separate but complementary steps for feature extraction and classification (Volpi and Tuia 2017). Feature extraction is implemented by specific operators on local portions of the image (e.g. image patches, super-pixels or regions, objects etc.), to transform the original spectral feature space into compact and/or abstract representations that are amenable to being readily separated by a classifier (Sun et al. 2014). Such transformed spatial features are thereafter used together with original spectra to train some sort of supervised classifiers (e.g. Support Vector Machine), in order to recognise the semantic content of the input imagery (Chen et al. 2016). The performance of any classifier used is heavily affected by the used transformations and the consequent spatial features. Common examples of such operators include texture

descriptors (Reis and Tasdemir 2011), mathematical morphology (Pingel et al. 2013), and oriented gradients (Cheng et al. 2013). The process of hand-engineering features, however, often involves a tedious trial-and-error procedure for feature extraction and selection (Volpi and Tuia 2017). Those hand-coded features are often task-specific and might be useful for a particular region and/or problem. Moreover, the used low-level features followed by shallow architectures of the classifiers, are insufficient to mine the underlying semantics or functions due to the lack of higher-level feature representations (Liu et al. 2017). Thus, limited classification performance has been achieved to-date using VFSR images that are spectrally and structurally complicated.

2.3 An Overview of Deep Learning in Remote Sensing

Deep learning offers a different outlook on feature learning and representations, where robust, abstract and invariant features are learnt end-to-end, hierarchically, from raw data (e.g. image pixels) to semantic labels, which is a key advantage in comparison with previous state-of-the-art methods (Nogueira et al. 2017). Many deep learning-based methods have been proposed, including deep belief networks (DBNs) (Chen et al. 2015), deep Boltzmann machines (DBMs) (Qin et al. 2017), stacked autoencoder (SDE) (Yao et al. 2016), and deep convolutional neural networks (CNNs) (Maggiori et al. 2017). Amongst them, the CNN model represents the most well-established method, with impressive performance and great success in the field of computer vision and pattern recognition (Schmidhuber 2015), such as for visual recognition (Krizhevsky et al. 2012, Farabet et al. 2013), image retrieval (Yang et al. 2015) and scene annotation (Othman et al. 2016).

Deep learning is taking off in the field of remote sensing (Zhu et al. 2017). Figure 2-1 illustrates the number of papers published on the related topic since 2014. Clearly, the

exponential increase from only 3 in 2014 up to projected 200+ in 2018 demonstrates the rapid surge of interest in deep learning within the remote sensing community. These publications show huge potential and practical utility in several remote sensing tasks, such as object detection (Zhao et al. 2017), semantic segmentation (Fu et al. 2017), road extraction (Wei, et al. 2017), and land use scene classification (Liu, et al. 2018).

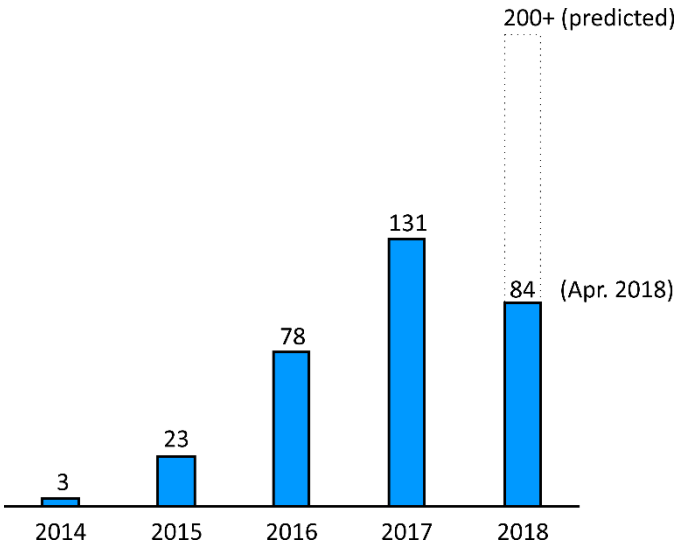


Figure 2-1: The statistics for published papers related to deep learning in remote sensing.

Table 2-1 - Summary of the deep learning for solving remote sensing tasks with the number of papers and example sources.

Remote sensing tasks	No of papers	Source (examples)
Target detection	**22**	Chen et al. 2014, Zhang et al. 2016, Long et al. 2017, Pei et al. 2018
Road extraction	6	Wang et al. 2015, Panboonyuen et al. 2017, Wei et al. 2017
Image processing	4	(Masi et al. 2016, Wei, Yuan, et al. 2017, Wang et al. 2018)
Semantic segmentation	**49**	Paisitkriangkrai et al. 2016, Maggiori et al. 2017, Wu et al. 2018
Land cover image classification	8	Kussul et al. 2017, Pan and Zhao 2017, Zhang et al. 2018
Land use scene classification	**72**	Hu et al. 2015, Othman et al. 2016, Nogueira et al. 2017, Liu et al. 2018
Change detection	5	Zhang et al. 2016, Gong et al. 2017, Su et al. 2017

Table 2-1 summarises various remotely sensed applications using deep learning based methods. From the table, it can be seen that, there are seven types of major applications

in the remote sensing domain: target detection, road extraction, image processing, semantic segmentation, land cover image classification, land use scene classification and change detection. Among the seven applications, the land use scene classification (72 papers), semantic segmentation (49 papers) and target detection (22 papers) constitute the majority cases, whereas others are less researched so far. These previous works represent the research focus and hot topics of deep learning in the remote sensing domain. Note, while this section covers the most important contributions in the literature; it will not provide a comprehensive review of deep learning in remote sensing (Zhang et al. 2016, Zhu et al. 2017). Instead, the purpose is to provide a concise overview of deep learning methods for classifying LC and LU using VFSR remotely sensed imagery. We focus on deep convolutional neural networks (CNN), as they are the most typical and well-established deep learning method that has been adopted in the remote sensing domain.

Deep CNNs are a variant of multilayer neural networks that are specifically designed to process large-scale images or sensory data in the form of multiple arrays by considering local and global stationary properties (LeCun et al. 2015). The essential characteristic of CNNs is their translational invariance through a patch-based procedure, in which a higher-level object within an image patch can be recognised even if the pixels comprising the object are shifted or distorted. Deep CNNs were originally designed to solve the *image* categorisation task, i.e. to assign the entire image into a semantic class such as a digit (LeCun et al. 1998) or an object category (Krizhevsky et al. 2012). In the remote sensing domain, the equivalent problem is to solve the remotely sensed scene classification task, in which an image patch is assigned to a specific category, such as ‘airport’, ‘residential’, ‘commercial’. These sorts of land use scene classification tasks are closely related to object detection (Zhang et al. 2016) and

localisation (Long et al. 2017), where the translational invariance is the key advantage of the CNN to *detect* the object with higher order features, such as land use or functional sites. However, this characteristic becomes a major weakness in LC and LU classification for pixel-level differentiation, from which blurred boundaries are produced between ground surface objects. Here, we review the classification of both LC and LU using CNNs to elaborate these challenges in detail and identify the research gaps.

2.4 Deep CNN for LC Classification

Land cover (LC) classification using CNN models can be divided broadly into two categories based on processing units, including patch-based and pixel-based procedures. The patch-based processes for LC classification involve an image patch passing through the entire image pixel-by-pixel, with densely overlapping patches used for land cover predictions (Fu et al. 2017). In this context, researchers have made some progress using patch-based CNN models. For example, Mnih (2013) proposed a patch-based CNN architecture to learn large-scale contextual features for aerial image labelling. The model produced a dense classification patch, instead of outputting a single value image category, in which spatial contextual features were learnt to better distinguish the land cover classes. Längkvist et al. (2016) used the standard pixel-wise CNN with densely overlapping patches to classify the image into five land cover classes (vegetation, ground, road, building, and water), outperforming the existing classification approaches. Sharma et al. (2017) extracted image patches for all possible locations within medium-resolution remotely sensed data, and classified them into LC categories respectively. However, such a patch-wise procedure has the disadvantage of introducing artefacts at the border of the classified patches, and the use of the

overlapped patches introduces too much redundant computations, thus, severely restricting the actual utility of the method for large-scale land cover classification (Fu et al. 2017, Maggiori et al. 2017). Recent research has shifted the focus on patch-based CNN for land cover classification towards designing pixel-level architectures for pixel labelling using VFSR remotely sensed imagery (Volpi and Tuia 2017). Particularly, the fully convolutional networks (FCN) and their extensions (Paisitkriangkrai et al. 2016, Wang et al. 2017, Zhao et al. 2017) were proposed for the task of semantic segmentation to classify a set of low-level land cover semantics, such as building, grassland and cars (Liu et al. 2017). These FCN-based methods involve convolution and down-sampling together with subsequent up-sampling to maintain the resolution of output map to be the same as the original input image, where the class likelihoods for an entire image were produced for pixel-wise semantic segmentation (Chen et al. 2016). However, the convolution utilises the neighbourhood information as context, and there is a trade-off between strong down-sampling, which allows the network to see a large context, but loses fine spatial details for precise boundary delineation (Marmanis et al. 2018). Besides, the up-sampling layers are performed in a sense of interpolation at the pixel level that tends to over-smooth the object with insufficient spatial information during the inference stage (Liu et al. 2017). As a consequence, the FCN models still face challenges in pixel-wise dense labelling.

Some other research has attempted to mitigate the blurring of boundaries due to down-sampling and up-sampling, either by using the "atrous" convolution (dilated convolution) to increase the density of the predicted class labels, or by adding skip connections within the network architectures, so that the fine resolution details were re-introduced after up-sampling (Marmanis et al. 2018). Still, these extension methods resulted in blurred boundary delineations when applied to VFSR remotely sensed

imagery with many small objects nested within each other. Others took the CNN as a rough classifier for object localisation, and further rectified the edges during the post-classification process by using the original image as the guidance for precise segmentation (Maggiori et al. 2017). For example, Längkvist et al. (2016) merged the standard pixel-wise CNN with segmented regions to smooth the classification results through average post-processing. Zhao et al. (2017) proposed a contour-preserving CNN method for semantic segmentation, and smoothed the classification results through post-processing using a conditional random field (CRF). Similarly, Fu et al. (2017) used FCN-based approaches for dense classification, and then performed the CRF method as a post-processing to refine the region boundaries. Marmanis et al. (2018) applied a special structure of FCN (SegNet), and smoothed the results using CRF for semantic segmentation. However, these post-processing procedures (either by averaging over segmented regions or using a CRF approach) can only partially address the boundary issues caused by CNN models by smoothing the outputs at the price of losing fine spatial detail. Often, some small features with linearly shaped objects, such as canal, railway, were easily eliminated through post-processing processes, which is undesirable in the case of VFSR remotely sensed image classification.

2.5 Deep CNN for LU Classification

Land use (LU) classification from VFSR remotely sensed data using CNN models has been undertaken in the form of land-use scene classification, which aims to assign a semantic label (e.g. tennis court, parking lot, etc.) to an image according to its content (Chen et al., 2016; Nogueira et al., 2017). There are broadly two strategies to exploit the CNN models for remotely sensed scene classification, namely; i) pre-trained or fine-tuned CNN, and ii) fully-trained CNN from scratch. Many researchers used the first

strategy primarily because the sample size in remote sensing tends to be small (up to several thousands), and sometimes cannot support the parameterisation of a deep network, such as AlexNet with 8 layers (Krizhevsky et al. 2012), visual geometry group network (VGG-Net) with 16 layers (Simonyan and Zisserman 2015), GoogLeNet with 22 layers (Szegedy et al. 2015), deep residual network with 34 or 50 layers (He et al. 2016). They normally pre-trained a deep CNN network on a natural image dataset such as ImageNet (Krizhevsky et al. 2012), and transferred this to the scene classification problem in the remote sensing domain, which was demonstrated to be empirically useful for the classification of land use scenes. For example, Hu et al. (2015) investigated the problem of transferring features from CNN models that are pre-trained on a large auxiliary labelled dataset. Marmanis et al. (2016) used the pre-trained AlexNet model as a feature extractor, and then transferred this into a supervised CNN for scene classification. Chaib et al. (2017) extracted deep features from a pre-trained VGG-Net, and fine-tuned on remotely sensed scene datasets, including the UC Merced (Yang and Newsam 2010), WHU-RS (Shao et al. 2013) and AID datasets (Xia et al. 2017). Nogueira et al. (2017) thoroughly discussed three strategies for exploiting the power of CNNs, including fully trained, pre-trained with fine-tuning, and using CNN directly as feature extractor. The empirical results demonstrate that the "pre-trained with fine-tuning" strategy tends to be more accurate than others given the limited training sample size. However, it requires three input channels derived from natural images with RGB only, whereas the multispectral remotely sensed imagery often involves the near infrared band, and such a distinction restricts the utility of pre-trained CNN networks. Alternatively, the (ii) fully-trained CNN strategy gives full control over the network architecture and parameters, which brings greater flexibility and expandability (Chen et al. 2016). Previous researchers have explored the feasibility of

the fully-trained approaches in building CNN models for scene level land-use classification. For example, Luus et al. (2015) proposed a multi-view CNN with multi-scale input strategies to address the issue of land use scene classification and its scale-dependent characteristics. Othman et al. (2016) used convolutional features and a sparse auto-encoder for scene-level land-use image classification, which further demonstrated the superiority of CNNs in feature learning and representation. Zhang et al. (2016) proposed a gradient boosting CNN model that outperformed other classical machine learning methods for remotely sensed scene classification. Liu et al. (2018) developed a deep random-scale stretched CNN for fine resolution remotely sensed scene classification, with patches of random scales used as inputs to strengthen the robustness of the CNN to scale variation. Experimental results demonstrated that the proposed CNN with random-scale stretched patches outperformed both classical machine learning methods and other off-the-shelf deep learning methods.

Although great success has been achieved in land use scene classification based on CNN models, they are essentially different from remotely sensed image classification, which requires all pixels in an entire image to be identified and labelled into land use categories (i.e., producing a LU thematic map). In such a context, a dense pixel-wise semantic labelling is required, and the spatial resolution should be preserved to precisely locate the object boundaries (Maggiori et al. 2017). This is not straightforward to implement, because the well-known trade-off between recognition with translation invariance and localisation without translation invariance. Due to the characteristic of translation invariance, CNNs demonstrate a structural limitation to perform fine-grained classification at the pixel level. Indeed, in VFSR remotely sensed imagery, even if few pixels were moved, the label would change abruptly when passing through the object boundaries. Therefore, it is of paramount importance to design specific deep

architectures to overcome the structural issues in terms of losing resolution and to solve the challenging problem of land use image classification.

Let us remark that the existing CNN models, either patch-based or pixel-based architectures, are not well designed and cannot be directly transferred to the problem of dense LC and LU image classification, which forms the research *gap* to be solved in this thesis. Based on the classification hierarchies, the previous work for pixel-wise classification using CNNs were all designed to address land cover classification tasks. For example, the well-known ISPRS semantic labelling dataset considers six land covers to be classified, including buildings, impervious surface, low vegetation, trees, cars, and clutters (Wang et al. 2017). Whereas land use image classification, which expects to classify each pixel into a specific land use, such as residential, commercial, industrial etc., has not been explored through the CNN based methods. The key thing is the context in a sense that a building cannot be identified as a commercial area without understanding the spatial and hierarchical relationships in a wide context (e.g. supermarkets, parking lots, and surrounding residential areas). However, as discussed beforehand, there is a strong trade-off between contextual information and the spatial precision. Essentially, a fine resolution thematic map is required for land use image classification that characterises the higher level functional representations and semantic meanings.

Table 2-2 - Summary of the state-of-the-art LC/LU classification methods with their typical input data, LC/LU classes as well as accuracy ranges.

LC/LU Classification Method	Source (Examples)	Typical Input Data	LC/LU Classes	Accuracy Range
Pixel-based MLP	Atkinson and Tatnall 1997, Del Frate et al. 2007, Hu and Weng 2009, Pacifici et al. 2009, Jiang et al. 2018	Landsat 5, SPOT 5, IKNOS, QuickBird satellite imagery	Land cover/use depends on specific applications (urban fabric extraction, forest canopy mapping, crop classification, ecological habitat mapping, et al.)	75.62% - 81.04%
Pixel-based SVM	Foody and Mathur 2004, Mountrakis et al. 2011, Ok 2013, Dragozi et al. 2014, Ozdarici-Ok et al. 2015, Zhang et al. 2015	Landsat 5, SPOT 5, IKNOS, QuickBird satellite imagery		74.85% - 81.96%
Pixel-based RF	Sesnie et al. 2010, Guo et al. 2011, Bechtel and Daneke 2012, Naidoo et al. 2012, Ahmed et al. 2015	Landsat 5, SPOT 5, IKNOS, QuickBird satellite imagery		74.56% - 82.42%
Object-based OBIA SVM	Conchedda et al. 2008, Li et al. 2010, 2015, Duro et al. 2012, Vetrivel et al. 2015	SPOT 5, IKNOS, QuickBird, WorldView-2, WorldView-3, GeoEye satellite imagery		80.28% - 83.86%
Patch-based CNN	Romero et al. 2016, Zhao and Du 2016, Pan and Zhao 2017, Sharma et al. 2017	WorldView-2 satellite imagery (Beijing Dataset), Landsat 8 imagery	Vegetation, ground, road, building, water	80.12% - 84.53%
CNN with segmented object averaging	Längkvist et al. 2016, Paisitkriangkrai et al. 2016, Zhao et al. 2017	Aerial imagery (Vaihingen Dataset), WorldView-2 satellite image (Beijing Dataset)	Building, Impervious Surface, Tree, Low Vegetation, Car	82.03% - 85.81%
CNN + Conditional Random Field (CRF)	Panboonyuen et al. 2017, Zhao et al. 2017, Pan and Zhao 2018, Wei et al. 2018	Aerial imagery (Vaihingen Dataset and Potsdam Dataset)	Building, Impervious Surface, Tree, Low Vegetation, Car	84.06% - 86.64%
SegNet	Cheng et al. 2017, Volpi and Tuia 2017, Arief et al. 2018, Audebert et al. 2018	Aerial imagery (Vaihingen Dataset and Potsdam Dataset) and DSM	Building, Impervious Surface, Tree, Low Vegetation, Car	83.79% - 86.75%
Fully convolutional network (FCN) + CRF	Fu et al. 2017, Maggiori et al. 2017, Wang et al. 2017, Marmanis et al. 2018	Aerial imagery (Vaihingen Dataset and Potsdam Dataset) and DSM	Building, Impervious Surface, Tree, Low Vegetation, Car	84.92% - 86.80%

2.6 Summary of LC and LU Classification methods

To synthesis the literature review in a systematic fashion, the LC and LU classification methods developed so far are summarised in Table 2-2. The reviewed classification methods start with pixel-based machine learning methods (MLP, SVM, RF), follow by object-based SVM as the OBIA paradigm, and reach the existing deep learning methods (patch-based CNN, CNN with segmented object averaging, CNN with CRF, and FCN with CRF). These classification methods take satellite and aerial imagery as their input data sources, which are classified into several land cover/use classes (primarily simple land covers), such as buildings, roads, trees etc. The pixel-based MLP has an accuracy range of (75.62% - 81.04%), similar to the pixel-based SVM (74.85% - 81.96%), and the pixel-based RF (74.56% - 82.42%). Compared with pixel-based approaches, Object-based SVM achieves a slightly higher accuracy (80.28% - 83.86%), owing to its consideration of spatial context. However, these traditional methods rely on hand coded features or rules that are hard to be designed and subject to user knowledge and expertise. Deep learning based methods have been actively studied as they can automatically learn the feature representations other than human designed features, with promising performance being demonstrated over the past few years. The patch-based CNN has an increased accuracy range compared with other traditional methods (80.12% - 84.53%), but at the price of losing spatial resolution and blurring the edges of ground features. Recent development in deep learning methods have made some progresses in addressing these problems through post-processing. For example, the CNN integrates with segmented object averaging at the post-classification process acquires an accuracy range of (82.03% - 85.81%). The patch-based CNNs together with conditional random field (CRF) are proposed to further enhance the accuracy range

using the class specific probability distribution along the boundaries (84.06% - 86.64%). A special deep architecture (SegNet) gains popular interest in remote sensing community for pixel-wise semantic labelling, leading to an accuracy range of (83.79% - 86.75%). Such networks are generalised as fully convolutional networks (FCN) with various extensions (e.g. via dilated convolution or through CRF post-processing) for semantic segmentation. The typical accuracy range for FCN with CRF is up to 84.92% - 86.80%, with an accuracy of 85% in average.

To summarise, the traditional pixel-based machine learning methods and the object-based methods have an average classification accuracy of 80%, whereas most of the reviewed deep learning methods achieve around 85% classification accuracy. However, all the existing methods have different kinds of problems and challenges outlined in 2.3 – 2.5. The aim of this thesis is to develop novel deep learning methods that can push the boundary of classification accuracy towards the most accurate LC and LU classification approaches.

Chapter 3 A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification[1]

[1] This chapter is based on the published paper: 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 Journal of Photogrammetry and Remote Sensing*, 140: 133-144. https://doi.org/10.1016/j.isprsjprs.2017.07.014.

Abstract

The contextual-based convolutional neural network (CNN) with deep architecture and pixel-based multilayer perceptron (MLP) with shallow structure are well-recognized neural network algorithms, representing the state-of-the-art deep learning method and the classical non-parametric machine learning approach, respectively. The two algorithms, which have very different behaviours, were integrated in a concise and effective way using a rule-based decision fusion approach for the classification of very fine spatial resolution (VFSR) remotely sensed imagery. The decision fusion rules, designed primarily based on the classification confidence of the CNN, reflect the generally complementary patterns of the individual classifiers. In consequence, the proposed ensemble classifier MLP-CNN harvests the complementary results acquired from the CNN based on deep spatial feature representation and from the MLP based on spectral discrimination. Meanwhile, limitations of the CNN due to the adoption of convolutional filters such as the uncertainty in object boundary partition and loss of useful fine spatial resolution detail were compensated. The effectiveness of the ensemble MLP-CNN classifier was tested in both urban and rural areas using aerial photography together with an additional satellite sensor dataset. The MLP-CNN classifier achieved promising performance, consistently outperforming the pixel-based MLP, spectral and textural-based MLP, and the contextual-based CNN in terms of classification accuracy. This research paves the way to effectively address the complicated problem of VFSR image classification.

Keywords: convolutional neural network; multilayer perceptron; VFSR remotely sensed imagery; fusion decision; feature representation

3.1 Introduction

With the rapid development of modern remote sensing technologies, a large quantity of very fine spatial resolution (VFSR) images is now commercially available. These VFSR images, typically acquired at sub-metre spatial resolution, have opened up many opportunities for new applications (Zhong et al. 2014), for example, urban land use retrieval , precision agriculture (Zhang and Kovacs 2012, Ozdarici-Ok et al. 2015), and tree crown delineation (Ardila et al. 2011, Yin et al. 2015). However, despite the presence of a rich spatial data content (Huang et al. 2014), the information conveyed by the imagery is conditional upon the quality of the processing (Längkvist et al. 2016). With fewer spectral channels in comparison with coarse or medium spatial resolution remotely sensed data, it can be challenging to differentiate subtle differences amongst similar land cover types (Powers et al. 2015). Meanwhile, objects of the same class may exhibit strong spectral heterogeneity due to differences in age, level of maintenance and composition as well as illumination conditions (Demarchi et al. 2014), which might be further complicated by the scattering of peripheral ground objects (Chen et al. 2014). As a consequence, such high intra-class variability and low inter-class disparity make automatic classification of VFSR images a challenging task.

Ever since the advent of VFSR imagery, tremendous efforts have been made to develop robust and accurate, automatic image classification methods. Among these, machine learning is currently considered as the most promising and evolving approach (Zhang et al. 2015). Popular pixel-based machine learning algorithms, such as Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Random Forest (RF), have drawn considerable attention in the remote sensing community (Yang et al. 2012, Attarchi and Gloaguen 2014, Zhang et al. 2015). The MLP, as a typical non-parametric

neural network classifier, is designed to learn the nonlinear spectral feature space at the pixel level irrespective of its statistical properties (Atkinson and Tatnall 1997, Foody and Arora 1997, Mas and Flores 2008). The MLP has been used widely in remote sensing applications, including VFSR-based land cover classification (e.g. Del Frate et al. (2007), Pacifici et al. (2009)). The MLP algorithm is mathematically complicated yet can be simple in model architecture (e.g., a shallow classifier with one or two feature representation levels). At the same time, a pixel-based MLP classifier does not consider, or make use of, the spatial patterns implicit in images, especially for VFSR imagery with unprecedented spatial detail. In essence, the MLP (and related algorithms, e.g. SVM, RF, etc.) is a pixel-based classifier with shallow structure (one or two layers) (Chen et al. 2016), where the membership association of a pixel for each class is predicted.

Recent advances in neuroscience have shown that deep feature representations can be learned hierarchically from simple concepts such as oriented edges to higher-level complex patterns such as textures, segments, parts and objects (Arel et al. 2010). This discovery motivated the breakthrough of the so-called "deep learning" methods that represent the state-of-the-art in a variety of domains, including target detection (Chen et al. 2016, Tang et al. 2015), image recognition (Krizhevsky et al. 2012, Farabet et al. 2013) and robotics (Yu et al. 2013, Bezak et al. 2014, Lenz et al. 2015), amongst others. The convolutional neural network (CNN), a well-established deep learning approach, has produced excellent results in the field of computer vision and pattern recognition (Schmidhuber 2015), such as for visual recognition (Krizhevsky et al. 2012, Farabet et al. 2013), image retrieval (Yang et al. 2015) and scene annotation (Othman et al. 2016).

In the remote sensing domain, CNNs have been studied actively and shown to produce state-of-the-art results over the past few years, focusing primarily on object detection (Dong et al. 2015) or scene classification (Hu et al. 2015, Zhang et al. 2016). Recent studies further explored the feasibility of CNNs for the task of remotely sensed image classification. For example, Yue et al. (2016) utilized spatial pyramid pooling to learn multi-scale spatial features from hyperspectral data, Chen et al. (2016) introduced a 3D CNN to jointly extract spectral–spatial features, thus, making full use of the continuous hyperspectral and spatial spaces. In terms of the classification of multi- and hyperspectral imagery, a deep CNN model was formulated through a greedy layer-wise unsupervised pre-training strategy (Romero et al. 2016), whereas an image pyramid was built up through upscaling the original image to capture the contextual information at multiple scales (Zhao and Du 2016). For VFSR image classification, CNN models with varying contextual input size were constructed to learn multi-scale features while preserving the original fine resolution information (Längkvist et al. 2016). All of the above-mentioned work applied CNNs with contextual patches as their inputs, and demonstrated the robustness and effectiveness in spatial feature representations with excellent classification performance. However, the benefits and shortcomings of the CNN as a classifier itself have not been studied thoroughly. In particular, the CNN, as a contextual classifier with deep structures (Szegedy et al. 2015), explores the complex spatial patterns hidden in the image that are not seen by representation in its shallow counterparts, whereas it may overlook certain information in spectral space observed by pixel-based classifiers. Moreover, uncertainties may appear in object boundaries due to the usage of convolutional filters of the CNN. These issues deserve further investigation.

Any single set of features (e.g., spectral only) or a specific classifier (e.g., pixel-based only) is unlikely to achieve the highest classification accuracies for VFSR imagery because the result is conditional upon both spectral and spatial information. In this context, two categories of spectral and spatial information were fused for classification or handled with a classifier ensemble. Information fusion can be realized by stacking the spatial and spectral information as feature bands. However, this does not allow the specification of the relative influence of the extracted features (Wang et al. 2016). Others proposed integrative algorithms considering the spatial and spectral features at the same time. For example, Fauvel et al. (2012) proposed a composite kernel-based SVM with spectral and spatial kernels applied simultaneously. However, the spatial kernel summarizes only basic information (e.g. median) of the spatial neighbourhood (Wang et al. 2016).

In terms of classifier ensemble technology, two strategies, namely "multiple classifier systems" (Benediktsson 2009) and "decision fusion" (Fauvel et al. 2006) are employed. Multiple classifier systems are based on the manipulation of training sample sets, including boosting (Freund et al. 2003) and bagging (Breiman 1996). This ensemble approach, however, usually requires a relatively large sample size and the computational complexity tends to be high. An alternative classifier ensemble is derived from decision fusion of the outputs of different classification algorithms according to a certain combination of approaches (Du et al. 2012, Löw et al. 2015). This decision fusion-based ensemble approach is preferable where the individual classifiers demonstrate complementary behaviour. For instance, different non-parametric classifiers are sometimes accurate in different locations in a classification map, thus, producing complementary results from the ensemble (Clinton et al. 2015, Löw et al. 2015). However, all the aforementioned fusion strategies are conducted using

pixel-based classifiers with shallow structures, whose complementary behaviours are insufficient to address the challenges of VFSR image classification.

In this chapter, a hybrid classification system was proposed that combines the CNN (a contextual-based classifier with deep architectures) and MLP (a pixel-based classifier with shallow structures) using a rule-based decision fusion strategy. The hypothesis is that both MLP and CNN classifiers can provide different views or feature representations with strong complementarity. Thus, the classifier ensemble has the potential to enhance the final classification performance. The decision fusion rules were built up at the post-classification stage, primarily based on the confidence distribution of the contextual-based CNN classifier, such that the classified pixels with low confidence can be rectified by the MLP at the pixel level. The effectiveness of the proposed method was tested on images of both an urban scene and a rural area. A benchmark comparison was provided by the standard pixel-based MLP, spectral-texture based MLP as well as contextual-based CNN classifiers.

3.2 Methodology

3.2.1 Brief review of multilayer perceptron (MLP)

A multilayer perceptron (MLP) is a network that maps sets of input data onto a set of outputs in a feedforward manner (Atkinson and Tatnall 1997). The typical structure is that the MLP is composed of interconnected nodes in multiple layers (input, hidden and output layers), with each layer fully connected to the preceding layer as well as the succeeding layer (Del Frate et al. 2007). The outputs of each node are weighted units followed by a nonlinear activation function to distinguish the data that are not linearly separable (Pacifici et al. 2009). Formally, the output activation $a^{(l+1)}$ at layer l+1 is derived by the input activation $a^{(l)}$:

$$a^{(l+1)} = \sigma(w^{(l)}a^{(l)} + b^{(l)}) \tag{3-1}$$

where *l* corresponds to a specific layer, $w^{(l)}$ and $b^{(l)}$ denote the weight and bias at layer *l*, and σ represents the nonlinear activation operation (e.g. sigmoid, hyperbolic tangent, rectified linear units) function. For an *m* layer multilayer perceptron, the first input layer is $a^{(1)} = x$ while the last output layer is:

$$h_{w,b}(x) = a^{(m)} \tag{3-2}$$

The weights w and bias b in equation (3-2) are learned by supervised training using a backpropagation algorithm to approximate an unknown input-output relation (Del Frate et al. 2007). The objective function is to minimize the difference between the predicted outputs and the desired outputs:

$$J(W,b;x,y) = \frac{1}{2}\left\|h_{w,b}(x) - y\right\|^2 \tag{3-3}$$

3.2.2 Brief review of Convolutional Neural Networks (CNN)

A Convolutional Neural Network (CNN), is a variant of the multilayer feed forward neural networks, and is designed specifically to process large scale images or sensory data in the form of multiple arrays by considering local and global stationary properties (LeCun et al. 2015). Similar to the MLP, the CNN is a network stacked into a number of *layers*, where the output of the previous layer is connected sequentially to the input of the next one by a set of learnable weights and biases (Romero et al. 2016). The major difference is that each layer is represented as input and output feature maps by capturing different perspectives on features through the operation of convolution.

The CNN basically consists of three major operations: convolution, nonlinearity and pooling/subsampling (Schmidhuber 2015). The convolutional and pooling layers are stacked together alternatively in the CNN framework, until obtaining the high-level features on which a fully connected classification is performed (LeCun et al. 2015). In addition, several feature maps may exist in each convolutional layer and the weights of convolutional nodes in the same map are shared. This setting enables the network to learn different features while keeping the number of parameters tractable. Mathematically, the output feature map ${y_{i,j}}^{(l)}$ at convolutional layer *l* is calculated as:

$$ {y_{i,j}}^{(l)} = \sigma^{(l)}\left(\sum_{n=1}^{k}\sum_{m=1}^{k} {w_{n,m}}^{(l)} \cdot {x_{i+n,j+m}}^{(l-1)} + b^{(l)}\right) \tag{3-4} $$

where the ${w_{n,m}}^{(l)}$ denotes the convolutional filter with size $k \times k$ at layer *l*, and the ${x_{i+n,j+m}}^{(l-1)}$ represents the spatial position of the corresponding feature map at the preceding layer *l*-1. The algorithm passes the convolutional filter throughout the input feature map using the dot product $(\cdot)$ between them with an addition of a bias unit $b^{(l)}$ (Arel et al. 2010). Moreover, a nonlinear activation function $\sigma^{(l)}$ at layer *l* is taken outside the dot product to strengthen the nonlinearity (Strigl et al. 2010).

The pooling/subsampling layer can generalize the convolved features through down-sampling and thereby reduce the computational complexity during the training process (Zhao and Du 2016). Given a pooling/subsampling layer *q*, the feature output F^q can be derived from the preceding layer $f^{(q-1)}$ through

$$ F_{i,j}^{q} = \max\left(f_{1+p(i-1),1+p(j-1)}^{q-1}, \ldots, f_{pi,1+p(j-1)}^{q-1}, \ldots, f_{1+p(i-1),pj}^{q-1}, \ldots f_{pi,pj}^{q-1} \right) \tag{3-5} $$

where $p \times p$ is the size of the local spatial region, and $1 \le i, j \le (m-n+1)/p$, here the *m* refers to the size of input feature map, while *n* corresponds to the size of filter

(Längkvist et al. 2016). The *max* simply summarizes the input features within local spatial region using the maximum value (Figure 3-1: Pooling). By doing this, the learnt features become robust and abstract with certain sparseness and translation invariance.

Once the higher level features are extracted, the output feature maps are flattened into a one-dimensional vector, followed by a fully connected output layer (Figure 3-1: fully connect). This operation is exactly a simple logistic regression, which is equivalent to the standard MLP discussed in section 2.1, but without any hidden layer.

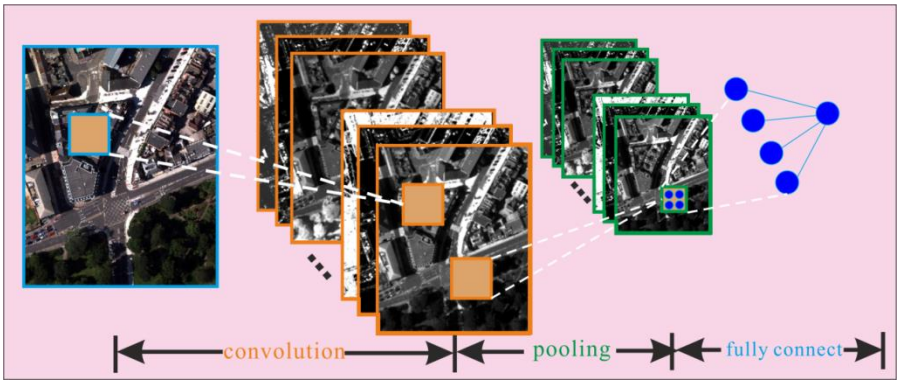


Figure 3-1: A schematic illustration of the three core layers within the CNN architecture, including the convolutional layer (convolution), pooling layer (pooling) and fully connected layer (fully connect).

3.2.3 Hybrid MLP-CNN Classification Approach

Suppose the predictive outputs of the MLP and CNN at each pixel are *n*-dimensional vectors $C = (c_1, c_2, ..., c_n)$, where *n* represents the number of classes and each dimension $i \in [1, n]$ corresponds to the probability of a specific class (*i*-th class) with certain membership association. Ideally, the probability of the classification prediction would be 1 for the target class and 0 for the others. However, due to the uncertainty in the process of remotely sensed image classification, the probability value *c* is denoted as $f(x) = \{c_x \mid x \in (1,2,...,n)\}$, where $c_x \in [0,1]$ and $\sum_1^n c_x = 1$. The classification model

simply takes the maximum membership association as the predicted output label (denoted as class(*C*)):

$$class(C) = \arg\max(\{ f(x) = c_x \mid x \in (1,2,...,n)\}) \tag{3-6}$$

The confidence $conf$ of such membership association is defined here as:

$$conf = Max(C) - Mean(C) \tag{3-7}$$

In equation (3-7), *Max*(*C*) represents the maximum value of vector *C*, while *Mean*(*C*) denotes the average of all the values of *C*. The *conf*, quantified by the difference between *Max*(*C*) and *Mean*(*C*), measures the confidence or reliability of the class membership allocation (i.e. classification confidence map). Since the CNN takes contextual image patches as its inputs instead of image pixels, it has the following properties:

(1). If the input image patch is located at the central homogeneous region, its class purity is relatively high with large difference between the membership association of the predicted class and those of the other classes, and the *conf* tends to be large (White regions in Figure 3-2(*c*)).

(2). If the image patch contains other land cover classes as contextual information, the resulting distinction between the membership association of prediction and those of the others is relatively low, and the *conf* tends to be small (Dark regions in Figure 3-2(*c*)).

However, the MLP (spectral feature only) is based on per-pixel spectral information, thereby ruling out the difference of membership association between central and boundary regions of the classified objects (Figure 3-1(*b*)). According to the aforementioned properties, the fusion decision rules are constructed primarily based on

CNN confidence. To be more specific, the fusion output gives credit to the CNN when its confidence is larger than a predefined threshold (α_1), while the MLP is trusted given that the CNN confidence is lower than another threshold (α_2); once the confidence of the CNN lies in-between the two thresholds ($\in(\alpha_1,\alpha_2)$), the fusion output chooses the CNN or MLP classification result with a larger confidence. Therefore, for a given image pixel at location (h,v), a rule-based decision fusion approach to determining the class label ($class(h,v)$) of the corresponding pixel is formulated as follows:

$$class(h,v)=\begin{cases} class_{mlp} & conf_{cnn}<\alpha_1 \\ class_{mlp} & (\alpha_1 \le conf_{cnn}<\alpha_2 \,\&\, conf_{cnn}<conf_{mlp}) \\ class_{cnn} & (\alpha_1 \le conf_{cnn}<\alpha_2 \,\&\, conf_{cnn}>conf_{mlp}) \\ class_{cnn} & conf_{cnn}\ge\alpha_2 \end{cases} \quad (3\text{-}8)$$

where the $class_{mlp}$ and $class_{cnn}$ represent the classification results of the MLP and CNN respectively; the $conf_{mlp}$ and $conf_{cnn}$ denote the classification confidence of the MLP and CNN accordingly.

Estimation of the two thresholds (α_1,α_2) is conducted using grid search with cross-validation (Min and Lee 2005, Zhang et al. 2015) based on the CNN classification confidence map (as illustrated by Figure 3-2(*c*)). Specifically, the α_1 was searched from 0.1 to 0.5 to detect those regions with low confidence as predicted by the CNN, while the α_2 was chosen from 0.5 to 0.9 to discover the high confidence regions. By initially fixing α_1 as 0.1, α_2 was tuned with step size of 0.05 (i.e. α_2=0.5, 0.55, 0.6, ..., 0.9) to cross-validate the classification accuracy influenced by the selected thresholds; α_1 was then increased to further tune α_2 in a similar way until the optimal α_1 and α_2 were found with the best classification accuracy.

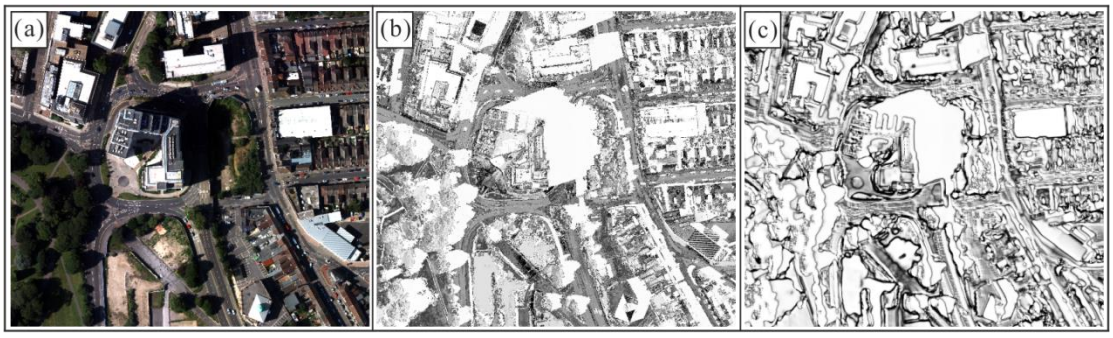


Figure 3-2: A subset of the original imagery (a) with RGB spectral bands, (b) the classification confidence of the MLP and (c) the classification confidence of the CNN. The dark pixels represent low confidence, while white pixels signify high confidence.

3.3 Experiment

3.3.1 Study area and data source

For this study, the city of Southampton, UK and its surrounding environment, which lies on the south coast of England, was chosen as a case study area (Figure 3-3). The urban and suburban areas in Southampton are strongly heterogeneous with a mixture of anthropogenic urban surface (e.g. roof materials, asphalt, concrete) and semi-natural environment (e.g. vegetation, bare soil), thereby representing a good test for classification algorithms.

A scene of aerial imagery of Southampton was captured on 22 July 2012 using a Vexcel UltraCam Xp digital aerial camera with 50 cm spatial resolution and four multispectral bands (Red, Green, Blue and Near Infrared). Two study sites S1 (3087×2750 pixels) and S2 (2022×1672 pixels) were selected to investigate the effectiveness of the proposed algorithm. S1 is located in the city centre of Southampton, which consists of eight dominant land cover classes, including Clay roof, Concrete roof, Metal roof, Asphalt, Grassland, Trees, Bare soil and Shadow, with detailed descriptions listed in Table 3-1. S2, on the other hand, is situated in a suburban and rural area of Southampton

comprised of large patches of forest, grassland and bare soil speckled with small buildings and roads. There are six land cover categories in this study site, namely, Buildings, Road-or-track, Grassland, Trees, Bare soil and Shadow (Table 3-1).

Figure 3-3: Southampton, UK Location of study area and aerial imagery with two study sites S1 and S2.

Sample points were collected using a stratified random scheme from ground data provided by local surveyors at Southampton, and split into 50% training samples and 50% testing samples for each class (Table 3-1). Field land cover survey was conducted throughout the study area on July 2012 to further check the validity and precision of the selected samples. In addition, a highly detailed vector map from Ordnance Survey, namely the MasterMap Topographic Layer (Regnauld and Mackaness 2006), was fully consulted and cross-referenced to gain a comprehensive appreciation of the land cover and land use within the study area.

To further test the applicability of the proposed method, another scene of Worldview-2 satellite sensor imagery was acquired on 24 July 2013 in the same region of Southampton with urban (S1') and rural (S2') study sites close to the Northwest of S1 and S2. The Worldview-2 image was geometrically and atmospherically corrected, and pan-sharpened at 50 cm spatial resolution to be consistent with the aerial imagery.

Figure 3-4 demonstrates the WorldView-2 satellite sensor image together with two subsets S1' and S2'.

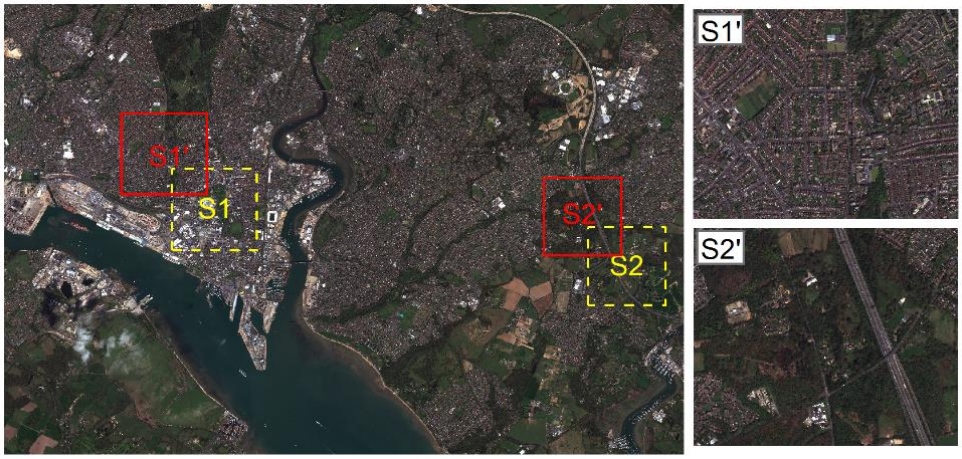


Figure 3-4: Additional WorldView-2 satellite sensor image covering the same region of Southampton with the S1' and S2' study sites to the northwest of S1 and S2, respectively.

Table 3-1 - Detailed description of land covers at two study sites with training and testing sample size per class.

Study Sites	Class	Train	Test	Description
S1	Clay roof	144	144	Predominantly residential buildings in red clay tiles
	Concrete roof	132	132	Predominantly residential buildings in grey clay tiles
	Metal roof	134	134	Predominantly industrial buildings in white metal panels
	Asphalt	136	136	Urban road or cark parks covered by asphalt
	Grassland	126	126	Areas of grass covering the urban park or lawn
	Trees	137	137	Patches of tree species
	Bare soil	118	118	Open areas covered by bare soil
	Shadow	123	123	Areas of shadow cast from buildings and trees
S2	Building	82	82	Predominantly small buildings at rural areas
	Road-or-track	85	85	Asphalt road or small path
	Grassland	86	86	Large areas of wild grass or lawn
	Trees	98	98	Large patches of deciduous trees
	Bare soil	84	84	Open areas covered by bare soil
	Shadow	86	86	Areas of shadow cast from buildings and trees

3.3.2 Model input variables and parameters

Model inputs: the standard pixel-based MLP (hereafter, MLP) and CNN take only the four spectral bands as their input variables, whereas the pixel-based texture MLP based on the standard Grey Level Co-occurrence Matrix (hereafter, GLCM-MLP) simultaneously makes use of both the four spectral bands and the texture features derived from GLCM textural features including the Mean, Variance, Homogeneity, Contrast, Dis-similarity, Entropy, Second moment and Correlation (Haralick et al. 1973, Zhang et al. 2003, Xia et al. 2010, Rodriguez-Galiano et al. 2012). Three window sizes for each spectral band, including 3×3 (1.5×1.5 m), 5×5 (2.5×2.5 m), and 7×7 (3.5×3.5 m), were optimally chosen to perform multi-scale texture feature representation, thus generating 96 GLCM texture features in total. It should be noted that both the MLP and the CNN as well as the GLCM-MLP were trained to predict all pixels within the images. Although the CNN was designed to predict a single label from a small image patch, the sliding window was densely overlapping to cover the entire image at the inference phase.

Both the MLP (also including GLCM-MLP) and CNN models require a series of predefined parameters to optimize the learning accuracy and generalization capability. Following the recommendations of Mas and Flores (2008), the MLPs with one, two and three hidden layers were tested, using a varying number of {4, 8, 12, 16, 20, and 24} nodes in each layer. The learning rate was chosen optimally as 0.2 and the momentum factor was set as 0.7. In addition, the number of iterations was set as 1000 to fully converge to a stable state. Through cross-validation with different numbers of nodes and hidden layers, the best predicting MLP was found using two hidden layers with 8 nodes in each layer. Similar parameters were also set for the GLCM-MLP, except that

two hidden layers with 20 nodes in each layer were found to be the optimal solution in this case.

For the CNN, a range of parameters including the number of layers, the input image patch size, the number and size of convolutional filter, as well as other predefined parameters, such as the learning rate and number of epochs (iterations), need to be tuned (Romero et al. 2016). Following the discussion by Längkvist et al. (2016), the input image size was chosen from {8×8, 10×10, 12×12, 14×14, 16×16, 18×18, 20×20, 22×22 and 24×24} to evaluate the influence of context area on classification performance. In general, a small-sized contextual area results in overfitting of the model, whereas a large one often leads to under-segmentation. In consideration of the image object size and contextual relationship coupled with a small amount of trial and error, the optimal input image patch size was set to 16×16 in this research. Besides, as discussed by Chen et al., (2014) and Längkvist et al. (2016), the depth plays a key role in classification accuracy because the quality of learnt feature is highly influenced by the level of abstraction and representation. As suggested by Chen et al. (2016), the number of CNN layers was chosen as four to balance the network complexity and robustness. Other parameters were set based on standard practice in the field of computer vision. For example, the filter size was set to 5×5 for the first convolution layer and 3×3 for the rest with stride of 1, and the number of the filters was set to 24 to extract multiple convolutional features at each level. The fully connected layer was tuned as 12 nodes followed by a softmax classification. The learning rate was set to 0.01 and the number of epochs (iterations) was chosen as 600 to fully learn the features through backpropagation. The detailed architecture of the CNN and its parameter configurations is illustrated in Figure 3-5.

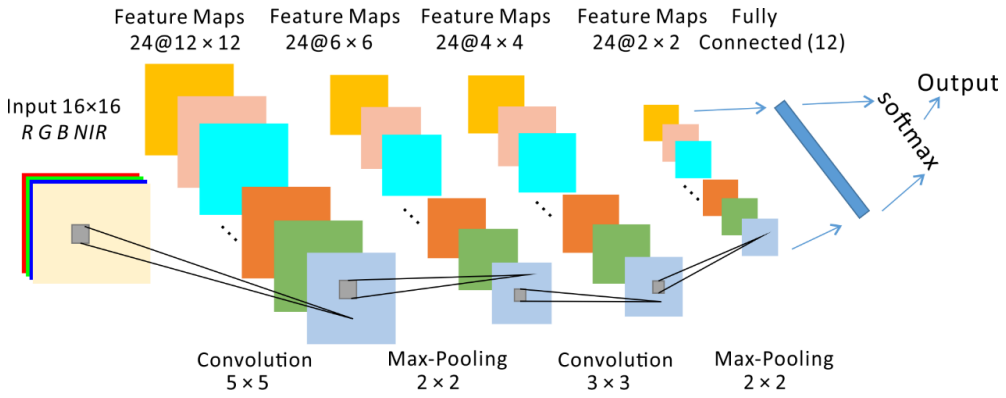


Figure 3-5: The architecture of the CNN and its configurations.

3.3.3 Decision Fusion Parameter Setting and analysis

A rule-based decision fusion approach was implemented based on the classification confidence maps of the CNN (e.g. Figure 3-2(*b*)) and MLP (e.g. Figure 3-2(*c*)). The parameters of decision fusion, including two thresholds α_1 and α_2, were determined by grid search with cross-validation using 10% of the randomly chosen samples. In this study, the optimal thresholds α_1=0.4 and α_2=0.6 were found that reported the greatest classification accuracy.

For the sake of visual interpretation, the confidence distribution of the CNN and MLP influenced by the chosen thresholds is shown in Figure 3-6. Clearly, the CNN and MLP demonstrated individually consistent, but mutually converse distribution patterns in the two study sites: along with the increase in the CNN's confidence, the MLP inversely exhibited a decreasing trend. Specifically, for low CNN confidence (<0.4), the MLP confidence was around 0.75, significantly higher than that of the CNN, thus outputting the results of MLP in the final decision; once the CNN confidence ranged from 0.4 to 0.6, no significant difference was shown between the two classifiers, thereby, optimally choosing the classification results based on the competitive "winner-takes-all" approach; while for large CNN confidence (>0.6), the MLP was, in contrast, much less reliable (<0.45), thus, taking the classification results of the CNN only in this situation.

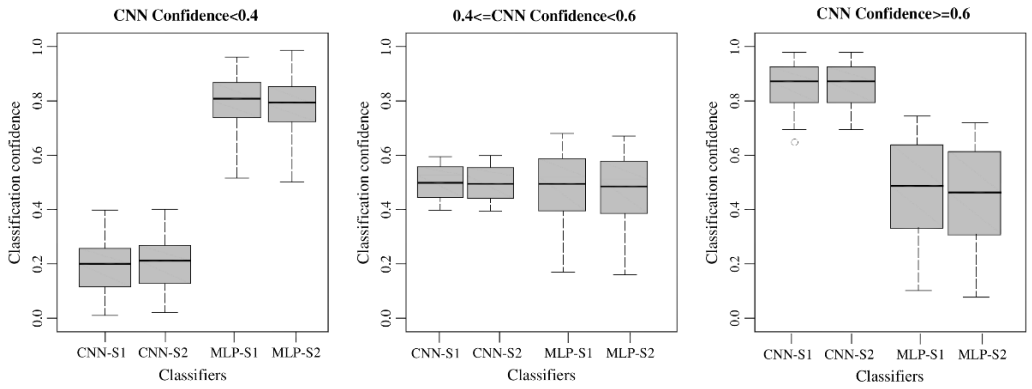


Figure 3-6: Classification confidence distributions of the CNN and MLP at two study sites (S1 and S2) under different fusion thresholds.

3.3.4 Classification results and analysis

3.3.4.1 Classification results and visual assessment

By integrating the classification results of the MLP and CNN using the above-mentioned fusion parameters, the final classification of the proposed MLP-CNN was obtained at both study sites, S1 (city centre with complex urban scene) and S2 (rural areas with natural phenomena). To provide a better visualization, Figure 3-7 (three subsets of S1) and Figure 3-8 (three subsets of S2) highlights the correct or incorrect classification results of different classifiers marked in yellow or red circles, respectively.

From Figure 3-7, it can be seen that the MLP classification results consist of undesirable noise (marked in red circle), such as a severe salt-and-pepper effect in Figure 3-7(*a*) and 3-7(*b*), and linear noisy textures in Figure 3-8(*c*). Besides, Trees and Grassland are seriously confused with each other as illustrated by Figure 3-7(*c*) and Figure 3-8(*a*) and 3-8(*b*). However, as shown by Figure 3-7(*b*), the MLP has certain advantages over CNN in identifying the Clay roof class with spectrally distinctive features (marked in yellow circle). With the addition of the GLCM textures, the GLCM-MLP achieved certain improvements in both spectral and spatial pattern differentiation. For example, Trees

and Grassland are better distinguished to some extent compared with the pixel-based MLP results, as illustrated in Figure 3-7(*c*) and Figure 3-8(*b*). Besides, the clear linear noisy textures in Figure 3-8(*c*) are much reduced, and primarily turned into small speckles due to the introduction of texture features. Yet, the GLCM-MLP falsely identifies some edges or boundaries as Clay Roof, as shown in Figure 3-7(*c*) and Figure 3-8(*a*) and 3-8(*b*) (marked in red circle). Additionally, some geometrical distortions of building roof tops, e.g. the Metal Roof and Concrete Roof in Figure 3-7(*b*), are shown in the GLCM-MLP classification results caused by the GLCM texture filters.

In contrast to the pixel-based MLP and the GLCM-MLP, the classification results of the CNN in both study sites exhibit smoothed visual effects with the least speckle noise as shown by Figure 3-7 and 3-8. Additionally, the classes of green vegetation including Grassland and Trees are accurately distinguished as demonstrated by the yellow circles in Figure 3-7(*c*) and Figure 3-8(*a*) and 3-8(*b*) in spite of their spectral similarity. Moreover, the CNN is able to discriminate the Concrete roof from Asphalt with a moderate accuracy, as highlighted by the yellow circle in Figure 3-7(*a*). Nevertheless, the CNN delivers some uncertainties in partitioning object boundaries. For example, the regular shapes of some buildings (e.g. the geometries of some Clay roof and Concrete roof areas) are distorted with false boundary partitions, as marked by the red circle in Figure 3-7(*b*). In addition, small or linear features are either merged into a large object or discarded by over-smoothness. For instance, some Clay roof buildings (small objects) are falsely connected together, while Asphalt is sometimes misclassified as Clay roof (Figure 3-7(*c*)) and the small paths covered by Bare soil are discarded (Figure 3-8(*b*)).

With respect to the results of the MLP-CNN, all of the aforementioned misclassifications produced by MLP or CNN are resolved with a higher resulting

accuracy. Thus, the incorrect classifications (marked by red circles) which appeared in Figure 3-7 and 3-8 are revised accordingly, with no red circles appearing in the classification results of MLP-CNN. The MLP-CNN modifies the classification errors of the CNN for Asphalt, as illustrated by the red circles in Figure 3-7(*c*) and Figure 3-8(*b*), thanks to the correct classification results of the MLP. Moreover, the linear-shaped Bare Soil area missed by the CNN in Figure 3-8(*a*) is brought back correctly without losing useful information. In addition, the original shapes of the Clay roof and Concrete roof areas shown in Figure 3-7(*b*) are accurately restored. Most importantly, some mutual misclassifications between the MLP and CNN are successfully rectified. For example, the MLP-CNN correctly differentiates some Asphalt (with spectrally distinctive but spatially confusing characteristics) and Concrete roof (distinctive in texture and geometry but vague in spectrum) areas that are mutually misclassified by the MLP and CNN respectively (see the regions marked by red circles in Figure 3-7(*a*)).

3.3.4.2 Classification accuracy assessment

The classification performance of the proposed MLP-CNN approach was further investigated through benchmark comparison with the MLP, GLCM-MLP and the CNN. Table 3-2 lists the classification accuracy assessment, including the overall accuracy (OA), Kappa coefficient (κ), and the class-wise mapping accuracy. From the table, it can be seen that the decision fusion approach (MLP-CNN) consistently reports the best classification OA with up to 90.93% for S1 and 89.64% for S2, higher than that of the CNN (85.39% and 86.56%, respectively) and GLCM-MLP (83.12% and 82.63%, respectively) as well as MLP (81.62% and 80.73%, respectively) (Table 3-2). Moreover, a Kappa *z*-test for pair-wise comparison also shows that a significant increase in classification accuracy has been achieved by the proposed MLP-CNN classifier over the MLP, GLCM-MLP and CNN in S1, with *z*-value=3.68, 3.12 and

2.25, respectively. For S2, the MLP-CNN also revealed a significant increase over the MLP with *z*-value=3.71 as well as GLCM-MLP with *z*-value=3.18, but no significant difference in comparison with the CNN ($z = 1.59$, smaller than 1.96 at 95% confidence level) (Congalton 1991), despite the obvious improvement shown in Table 3-2.

The increase in classification accuracy was also checked by class-wise accuracy assessment (Table 3-3). As illustrated by the table, MLP-CNN outperforms CNN for all classes at both study sites in terms of classification accuracy. The largest increase is up to 9.77% for the class of Concrete roof in S1 and 7.16% for the class of Road-or-track in S2. Similar patterns were found such that the MLP-CNN was constantly superior to GLCM-MLP at the class-wise level, where the greatest increase in accuracy was shown up to 11.56% for the class of Concrete Roof in S1 and 11.74% for the class of Grassland in S2. When compared with the MLP, most classes in the two sites except for Asphalt and Shadow in S1 are classified with higher accuracy by the MLP-CNN. Here, Grassland exhibits the highest increase in classification accuracy, up to 33.51% and 18.83% for S1 and S2, respectively. For the classes of Asphalt and Shadow, the accuracy of the MLP is slightly larger than that of the MLP-CNN, but without a statistically significant difference. Thus, they can be regarded as similar to each other.

With respect to the three benchmark classifiers themselves (i.e. MLP, GLCM-MLP and CNN), it can be seen from Table 3-2 that their classification accuracies are ordered as: MLP <GLCM-MLP < CNN. While the accuracy of CNN is remarkably higher (3%-5%) than that of the MLP and GLCM-MLP, the GLCM-MLP is just slightly higher (<2%) than the MLP. The Kappa *z*-tests (Table 3-3) further demonstrate that the CNN is statistically significantly more accurate than MLP and GLCM-MLP in both urban and rural areas, whereas a significant increase in accuracy of the GLCM-MLP over the MLP appears only in the rural area rather than the urban area.

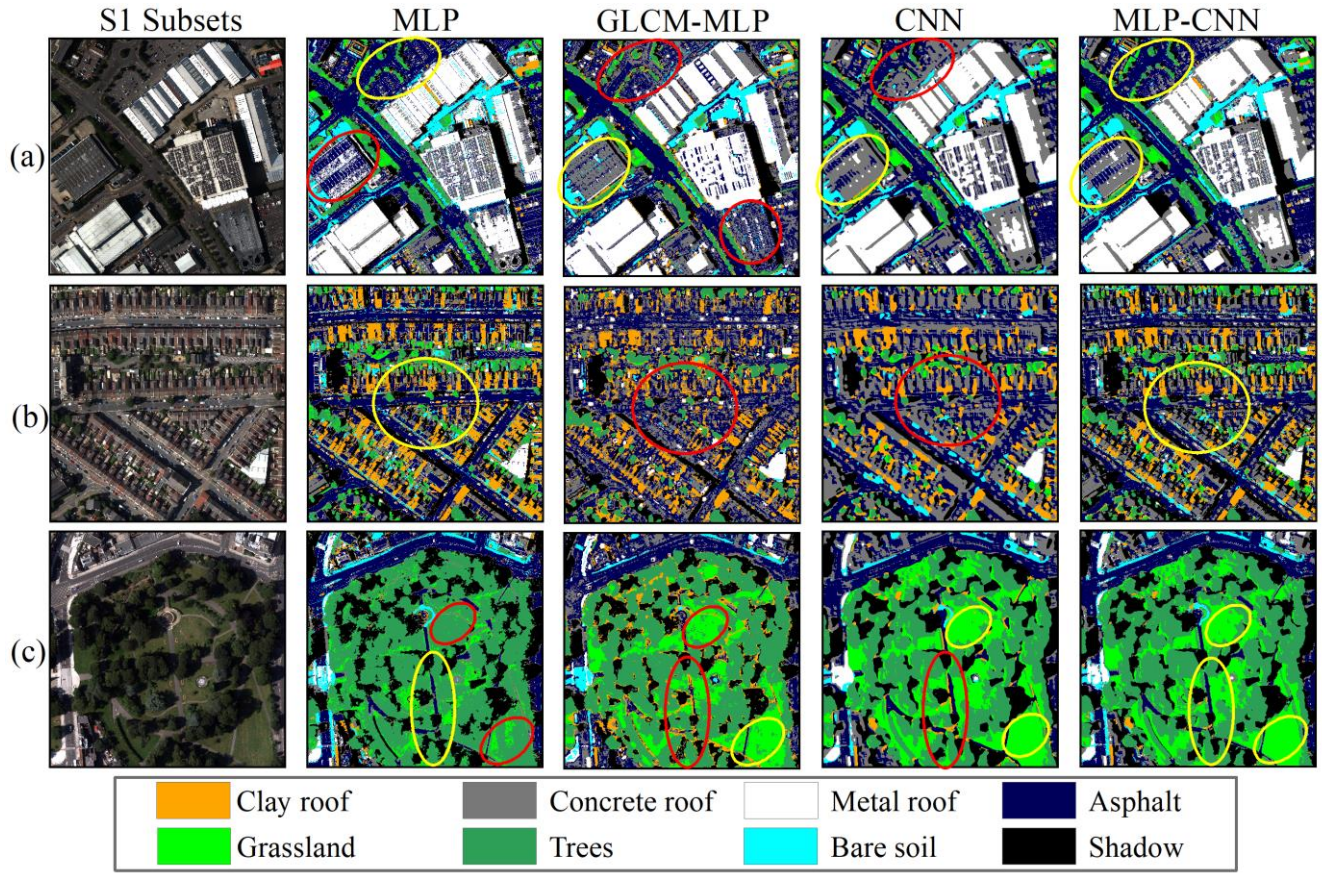


Figure 3-7: Three typical image subsets (a, b and c) in study site S1 with their classification results. Columns from left to right represent the original images (R G B bands), the MLP classification, the GLCM-MLP classification, the CNN classification and the MLP-CNN classification correspondingly. The red and yellow circles denote incorrect and correct classification, respectively.

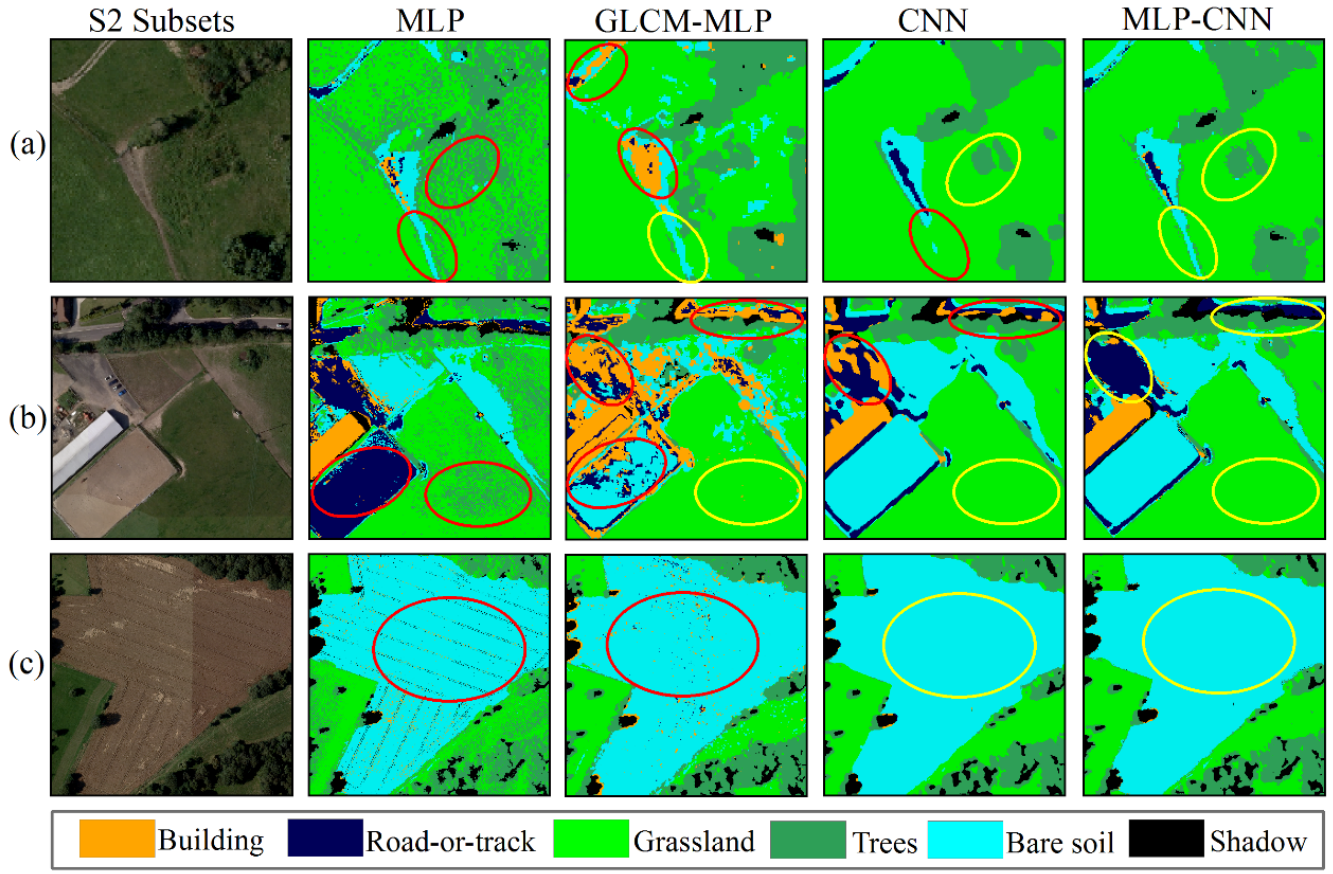


Figure 3-8: Three typical image subsets (a, b and c) in study site S2 with their classification results. Columns from left to right represent the original images (R G B bands), the MLP classification, the GLCM-MLP classification, the CNN classification and the MLP-CNN classification correspondingly. The red and yellow circles denote incorrect and correct classification, respectively.

Table 3-2 - Classification accuracy comparison amongst MLP, GLCM-MLP, CNN and the proposed MLP-CNN approach for study sites S1 and S2 using the per-class mapping accuracy, overall accuracy (OA) and Kappa coefficient (κ). The bold font highlights the greatest classification accuracy per row.

Study Sites	Class	MLP	GLCM-MLP (Benchmark)	CNN	MLP-CNN
S1	Clay roof	92.26%	91.43%	90.11%	**95.03%**
	Concrete roof	67.06%	62.44%	64.23%	**74.00%**
	Metal roof	91.13%	90.36%	94.19%	**94.63%**
	Asphalt	**92.72%**	88.67%	85.98%	91.26%
	Grassland	60.51%	82.58%	90.73%	**94.02%**
	Trees	63.88%	78.46%	82.28%	**88.83%**
	Bare soil	79.63%	83.05%	86.16%	**92.49%**
	Shadow	**92.33%**	91.06%	91.14%	91.52%
	Overall Accuracy (OA)	81.62%	83.12%	85.39%	**90.93%**
	Kappa Coefficient (κ)	0.78	0.81	0.84	**0.89**
S2	Building	82.83%	80.79%	83.08%	**88.48%**
	Road or track	83.02%	80.14%	82.42%	**89.58%**
	Grassland	71.11%	78.20%	88.34%	**89.94%**
	Trees	79.31%	84.55%	90.70%	**92.86%**
	Bare soil	74.07%	76.32%	81.36%	**86.86%**
	Shadow	89.41%	88.25%	88.37%	**90.17%**
	Overall Accuracy (OA)	80.73%	82.63%	86.56%	**89.64%**
	Kappa Coefficient (κ)	0.78	0.79	0.84	**0.87**

The proposed MLP-CNN method and the other three benchmarks (MLP, GLCM-MLP and the CNN) were also validated using an additional WorldView-2 satellite sensor dataset at the S1' and S2' study sites. The OA and κ of both study sites are in accordance with the results of aerial photo classification, where the decision fusion approach (MLP-CNN) acquires the largest OA of 90.56% at S1' and 89.77% at S2', consistently higher than the CNN (86.15% and 86.39%), the GLCM-MLP (83.26% and 82.52%) and the MLP (81.42% and 80.32%) (Table 3-4). Such coherency of classification results further demonstrates the wide applicability of the proposed method with different datasets.

Table 3-3 - Kappa *z*-test (*p*-value) comparing the performance of the three classifiers for two study sites S1 and S2. Significantly different accuracies with confidence of 95% (*z*-value > 1.96 with *p*-value < 0.05) are indicated by *.

Study sites	Classifiers	Kappa Z-test (p-value)			
		MLP	GLCM-MLP (Benchmark)	CNN	MLP-CNN
S1	MLP	—			
	GLCM-MLP	1.56 (0.1188)	—		
	CNN	2.64* (0.0083)	2.44* (0.0147)	—	
	MLP-CNN	3.68* (0.0002)	3.12* (0.0018)	2.25* (0.0244)	—
S2	MLP	—			
	GLCM-MLP	2.05* (0.0404)	—		
	CNN	2.51* (0.0121)	2.36* (0.0183)	—	
	MLP-CNN	3.71* (0.0002)	3.18* (0.0015)	1.59 (0.1118)	—

Table 3-4 - Classification accuracy comparison amongst MLP, GLCM-MLP (Benchmark), CNN and the proposed MLP-CNN approach for study sites S1' and S2' from the WorldView-2 satellite sensor image using overall accuracy (OA) and Kappa coefficient (κ). The bold font highlights the greatest classification accuracy per row.

WorldView-2	Classification	MLP	GLCM-MLP (Benchmark)	CNN	MLP-CNN
S1'	OA	81.42%	83.26%	86.15%	**90.56%**
	κ	0.77	0.80	0.82	**0.89**
S2'	OA	80.32%	82.52%	86.39%	**89.77%**
	κ	0.77	0.79	0.83	**0.87**

3.4 Discussion

In this research, a rule-based decision fusion approach (MLP-CNN) was proposed to integrate classifiers of the pixel-based MLP with shallow structures and the contextual-based CNN with deep architectures for the classification of VFSR remotely sensed

imagery. The MLP-CNN takes advantage of the merits of the two classifiers and overcomes their individual shortcomings as discussed below.

3.4.1 Characteristics of MLP and GLCM-MLP classification

In principle, the MLP builds the decision boundaries among classes in feature space based on per-pixel spectral information (Mokhtarzade and Zoej 2007). Such classification boundaries are very sensitive to the class with salient spectral properties that are spectrally distinctive from other classes (Berberoglu et al. 2000). For example, classes like Clay roof, Asphalt and Shadow in Site 1 are spectrally exclusive to other classes, leading to high classification accuracies, up to 92.26%, 92.72% and 92.33%, respectively (Table 3-2). However, the MLP relies on the pixel-based spectral information in the classification process without exploiting the abundant spatial information appearing in the VFSR imagery (e.g. texture, geometry or contextual relationship) (Wang et al. 2016). These limitations often result in unsatisfactory classification performance; for example, confusion and misclassification between the Trees and Grassland classes that are spectrally similar. Even for those correctly identified objects, severe salt and pepper effects still exist (Dark and Bram 2007), for example, the linear texture noise appearing for Bare soil in Figure 3-8(*c*). For these reasons, the classification accuracy of MLP is generally statistically significantly lower than that of the CNN and the proposed MLP-CNN. However, objects in VFSR imagery are mostly depicted by pure pixels, especially for human-made features with crisp boundaries, such as buildings, residential houses and cultivated land. The membership association of a pixel deduced by MLP is, therefore, not affected by its relative position (e.g. lying on or close to boundaries), as long as the corresponding spectral space is separable.

The inclusion of GLCM texture features in the GLCM-MLP classifier enables the model to process spectral and spatial information simultaneously. Those GLCM texture descriptors are handcrafted features that are designed to capture statistical co-occurrence information (Xia et al. 2010). However, the GLCM textures are essentially first or second order feature transformations instead of feature learning. Such hand-coded features might be effective for a particular region and/or season, but are often challenging to generalize to other domains and datasets. Besides, the addition of 96 GLCM textures results in a dramatically increased number of input variables, which leads to a relatively high dimensional feature space. The so-called "curse of dimensionality" (Hughes 1968) and collinearity make the GLCM-MLP hard to parameterize and potentially leads to texture overfitting. That is why the GLCM-MLP cannot substantially increase the classification accuracy compared to the MLP. That is, the spectral and spatial information cannot be effectively exploited by the GLCM-MLP. For example, some spectrally different classes but with similar textures such as Clay Roof, Concrete Roof and Asphalt are confused to some degree.

3.4.2 Characteristics of CNN classification

Spatial features in remotely sensed data like VFSR imagery are intrinsically local (especially in lower layers) and spatially invariant (Masi et al. 2016). The MLP, however, assumes that the location of the data in the input is irrelevant to the model construction and it is, thus, incapable of learning spatial features of remote sensing data. In contrast, by using multiple convolution and pooling operations, CNN models the way that the human visual cortex works and enforces weight sharing with translation invariance that enables the extraction of high-level spatial features from image patches. It should be mentioned that the pooling operations play an important role in dimension reduction, thus, avoiding "the curse of dimensionality" present in the GLCM-MLP

classifier. Thanks to these superior characteristics, the CNN classifier outperforms the MLP and GLCM-MLP classifiers in both the urban scene and rural areas. Especially, classes like Concrete roof and Road-or-track that are difficult to distinguish from their backgrounds with only spectral or low-level features (e.g. distance between the prediction and the target class at spectral space), are identified with relatively high accuracies. In addition, classes with heavy spectral confusion in both study sites (e.g. Trees and Grassland), are accurately differentiated due to their obvious spatial pattern differences; for example, the texture of tree canopies is generally much rougher than for grassland. As a contextual classifier with deep architectures, the CNN could reveal the spatial patterns hidden in the image data that cannot be perceived by its shallow counterparts (e.g. MLP classifier or even the GLCM-MLP classifier). The higher layers in CNN models provide more semantically meaningful information concentrating on global semantics rather than local or pixel-level information, making the CNN classification work well for classes with spectral confusion (Hu et al. 2015a, Hu et al. 2015b, Yang et al. 2015). Therefore, the CNN shows an impressive stability and effectiveness in spatial feature representation, which is crucial for VFSR image classification (Zhao and Du 2016).

However, according to the "no free lunch" theorem (Wolpert and Macready 1997), any elevated performance in one aspect of a problem will be paid for through others, and the CNN is no exception. Using contextual image patches as inputs and learning deep spatial features, the CNN demonstrates power in spatial pattern recognition but also weakness in spatial partition. Boundary uncertainties (over-smoothness) often appear in the classified object and small useful features are erased, somewhat similar to morphological or Gabor filter methods (Reis and Tasdemir 2011, Pingel et al. 2013). For example, the human-made objects in urban scenes like buildings and asphalt are

often geometrically enlarged with distortion to some degree (See Figure 3-7(*b*)). As for natural objects in rural areas (S2), edges or porosities of a landscape patch are simplified or ignored, and even worse, linear features like river channels or dams that are of ecological importance, are erroneously erased. One may argue that the reduction of image patch size might be able to detect small features by multiple CNNs by varying the contextual filter size as adopted in Längkvist et al. (2016). However, objects, whether large or small in size, all have boundaries, thus, retaining the problem of smoothing edges. In addition, the adoption of convolution and pooling operations intrinsically reduces the image contextual size but strengthens the spatial feature representation. Thus, a far too small initial image patch size can limit the network depth of a CNN model. In fact, the currently used 16×16 window size is close to the minimum requirements for a deep CNN with four hidden layers in total. Moreover, certain spectrally distinctive features without obvious spatial patterns are poorly differentiated. For example, some Asphalt pixels are wrongly identified as Concrete roofs as illustrated in Figure 3-7(*a*). This further demonstrates the necessity of introducing spectral features for VFSR image classification.

3.4.3 fusion decision of MLP-CNN classification

Huge uncertainty and inconsistency exists inherently in any remotely sensed data (including VFSR imagery), and this runs through the training and the testing samples. In fact, different classification algorithms vary in terms of remote sensing data processing strategies. Thus there is no 'one-algorithm-fits-all' solution (Löw et al. 2015) to various applications of VFSR image classification, even for the powerful CNN classifier with deep spatial feature representations. It is therefore especially important to make use of the complementarities of different classifiers. It should be mentioned that, the more heterogeneous the classification algorithms' behaviours, the more that

different places might be accurately classified by each individual classifier, and the more accurate the ensemble classifier might be (Löw et al. 2015). An ideal ensemble classifier, thereby, should be established using individual classifiers that are very differently behaved.

The experimental results show that the pixel-based MLP classifier with shallow structures and the contextual-based CNN classifier with deep architectures can provide complementary information, leading to a more accurate classification result than either classifier alone. In addition to the elimination of heavy noise, the CNN can accurately identify classes with rich spatial information implicit in VFSR data. Such characteristics of the CNN emphasize the limitations of the MLP classifier for VFSR image classification. At the same time, the CNN might lose some useful details, and it has difficulties in utilizing spectral information and delineating object boundaries and is, thus, incapable of maintaining geometric fidelity. The MLP classifier, however, compensates directly with regard to the limitations of the CNN. The aforementioned complementary properties between the CNN and MLP are well reflected from the inverse confidence trends of the two classifiers (Figure 3-2). Specifically, in the case of the CNN with the highest confidence, the MLP has the least confidence and *vice versa*, which further indicates that the proposed MLP-CNN ensemble classifier can take advantage of the MLP and CNN.

The proposed fusion decision rules were derived primarily on the basis of the CNN's confidence distribution, in consideration of the superiority of CNN classification performance and the regularity of its confidence distribution. Such a decision fusion strategy captures the patterns of the complementarities between the two individual classifiers in general, thus, achieving a desirable classification result. At the same time,

the MLP-CNN classifier demonstrates great utility and wide applicability for both aerial photography and WorldView-2 satellite sensor imagery with consistent and competitive classification performance. However, in comparison with MLP, the classification accuracies of Asphalt and Shadow were slightly higher than for the proposed MLP-CNN. This means that there is still room for improvement of the decision fusion rules at the class-wise level for VFSR image classification. It might be better to incorporate the spectral separability differentiated by MLP to achieve the best classification performance at class level. Besides, no significant improvement was acquired for rural areas (S2) by the MLP-CNN compared with the CNN. This is mainly due to the ineffectiveness of the MLP in classifying natural features that dominate in the rural environment. This shortcoming might be overcome by the replacement of the MLP by other non-parametric machine learning classifiers (e.g. SVM, RF, etc.). Moreover, incorporating other data sources (e.g. digital surface model) might be needed to increase the accuracy of the MLP-CNN for both the CNN and MLP with very low confidence simultaneously. These aforementioned issues will be investigated in future research.

3.5 Conclusion

Due to its high intra-class variability and low inter-class disparity, VFSR image classification poses great challenges to any single machine learning algorithm, even for the powerful deep learning convolutional neural network (CNN). In this chapter, two neural network classifiers with strong heterogeneous behaviours (i.e. pixel-based MLP with shallow structures and contextual-based CNN with deep architectures), were integrated in a concise and effective way using a rule-based decision fusion strategy. The decision fusion rules, designed primarily on the basis of the classification

confidence of the CNN, reflect the general complementary patterns of both the MLP and CNN. In consequence, the proposed ensemble classifier MLP-CNN harvests the complementary results acquired from the CNN with deep spatial feature representations (CNN) and from the MLP based on spectral discrimination. Meanwhile, limitations of the CNN such as uncertainty in object boundary partition and loss of useful fine resolution detail were compensated. The effectiveness of the new MLP-CNN algorithm was tested in both urban and rural areas using aerial and satellite sensor images. The MLP-CNN algorithm consistently outperformed both of the individual classifiers (MLP and CNN) as well as the GLCM-MLP that includes the GLCM texture features, with a statistically significant difference in the majority of cases. This research paves the way to an effective solution to the complicated problem of automatic VFSR image classification.

Chapter 4 VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images[2]

[2] This chapter is based on the published paper: 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 Transactions on Geoscience and Remote Sensing*, 56(8): 4507-4521. https://doi.org/10.1109/TGRS.2018.2822783.

Abstract

Recent advances in computer vision and pattern recognition have demonstrated the superiority of deep neural networks using spatial feature representation, such as convolutional neural networks (CNN), for image classification. However, any classifier, regardless of its model structure (deep or shallow), involves prediction uncertainty when classifying spatially and spectrally complicated very fine spatial resolution (VFSR) imagery. We propose here to characterise the uncertainty distribution of CNN classification and integrate it into a regional decision fusion to increase classification accuracy. Specifically, a variable precision rough set (VPRS) model is proposed to quantify the uncertainty within CNN classifications of VFSR imagery, and partition this uncertainty into positive regions (correct classifications) and non-positive regions (uncertain or incorrect classifications). Those "more correct" areas were trusted by the CNN, whereas the uncertain areas were rectified by a Multi-Layer Perceptron (MLP)-based Markov random field (MLP-MRF) classifier to provide crisp and accurate boundary delineation. The proposed MRF-CNN fusion decision strategy exploited the complementary characteristics of the two classifiers based on VPRS uncertainty description and classification integration. The effectiveness of the MRF-CNN method was tested in both urban and rural areas of southern England as well as Semantic Labelling datasets. The MRF-CNN consistently outperformed the benchmark MLP, SVM, MLP-MRF and CNN and the baseline methods. This research provides a regional decision fusion framework within which to gain the advantages of model-based CNN, while overcoming the problem of losing effective resolution and uncertain prediction at object boundaries, which is especially pertinent for complex VFSR image classification.

Keywords: rough set, convolutional neural network, Markov random field, uncertainty, regional fusion decision.

4.1 Introduction

Remote sensing technologies have evolved greatly since the launch of the first satellite sensors, with a significant change being the wide suite of very fine spatial resolution (VFSR) sensors borne by diverse platforms (satellite, manned aircraft or unmanned aerial vehicles UAV) (Benediktsson et al. 2012). These technical advances have resulted in immense growth in the available VFSR remotely sensed imagery typically acquired at sub-metre spatial resolution (Yao et al. 2016), such as QuickBird, GeoEye-1, Pleiades-1, and WorldView-2, 3, and 4. The fine spatial detail presented in VFSR images offer huge opportunities for extracting a higher quality and larger quantity of information, which may underpin a wide array of geospatial applications, including urban land use change monitoring (Shi et al. 2015), precision agriculture (Ozdarici-Ok et al. 2015), and tree crown delineation (Ardila et al. 2011), to name but a few. One of the bases of these applications is image classification where information embedded at the pixel level is captured, processed and classified into different land cover classes (Zhang et al. 2016). Image classification applied to VFSR imagery, however, can be a very complicated task due to the large spectral variation that the same land cover class can produce, which increases the difficulty of discriminating complex and ambiguous image features (Lei et al. 2011). The increased spatial resolution, often in conjunction with a limited number of wavebands, can lead to reduced spectral separability amongst different classes. As a consequence, it is of prime concern to develop robust and accurate image classification methods to fully exploit and analyse such data effectively and to keep pace with the technological advances in remote sensors.

Over the last few decades, a vast array of computer-based image classification methods have been developed (Zhang et al. 2015), ranging from unsupervised methods such as K-means clustering, supervised statistical approaches such as maximum likelihood classification, and non-parametric machine learning algorithms, such as the multilayer perceptron (MLP), support vector machine (SVM) and random forest (RF), amongst others. Non-parametric machine learning is currently considered as the most promising and evolving approach (Pacifici et al. 2009). The MLP, as a typical non-parametric neural network classifier, is designed to learn the non-linear spectral feature space at the pixel level irrespective of its statistical properties. The MLP has been used widely in remote sensing applications, including VFSR-based land cover classification (e.g. Del Frate et al. 2007, Pacifici et al. 2009). However, a pixel-based MLP classifier does not make use of the spatial patterns implicit in images, especially for VFSR imagery with unprecedented spatial detail. Thus, limited classification performance can be obtained by the pixel-based MLP classifier (and related algorithms, e.g. SVM, RF, etc.) that purely relies on spectral differentiation.

To better exploit the potential in VFSR remotely sensed imagery, many researchers proposed to incorporate spatial information to distinguish spatial features through context. These spatial features may be associated with a regular spatial organization specific to particular types of land cover (Regniers et al. 2016). For example, the juxtaposition of buildings and roads can create a specific spatial pattern. Similarly, the periodic row structure in cereals can be a useful cue in classifying VFSR image data. These spatial patterns can be captured directly through spatial contextual information in the classification process. A typical example of such is the Markov Random field (MRF) (Nishii 2003), that has been used widely in the field of remote sensing. The MRF models the conditional spatial dependencies within a pixel neighbourhood to

support prediction for the central pixel, to increase classification accuracy (Wang and Liu 1999). However, the contextual MRF often uses small neighbourhood windows to achieve the robustness as well as to balance the computational complexity, which might downgrade the performance for the classification of VFSR imagery that requires wider contexts to handle the rich spatial details.

Recent advances in computer vision and machine learning have suggested that spatial feature representation can be learnt hierarchically at multiple levels through deep learning algorithms (Arel et al. 2010). These deep learning approaches learn the spatial contexts at higher levels through the models themselves to achieve enhanced generalization capabilities. The convolutional neural network (CNN), as a well-established deep learning method, has produced state-of-the-art results for multiple domains, such as visual recognition (Krizhevsky et al. 2012), image retrieval (Yang et al. 2015) and scene annotation (Othman et al. 2016). CNNs have been introduced and actively investigated in the field of remote sensing over the past few years, focusing primarily on object detection (Dong et al. 2015) and scene classification (Zhang et al. 2016). Recent work has demonstrated the feasibility of CNNs for remote sensing image classification, as here. For example, Zhao and Du (2016) used an image pyramid of hyperspectral imagery to learn deep features through the CNN at multiple scales. Chen et al. (2016) introduced a 3D CNN to jointly extract spectral–spatial features, thus, making full use of the continuous hyperspectral and spatial spaces. Längkvist et al. (2016) used a CNN model with different contextual sizes to classify and segment VFSR satellite images. Volpi and Tuia (2017) used deep CNNs to perform a patch-based semantic labelling of VFSR aerial imagery together with normalized DSMs. All of these works demonstrated the superiority of CNNs by using contextual patches as their inputs and the convolutional operations for spatial feature representation.

The contextual-based CNN classifiers, however, might introduce uncertainties along object boundaries, leading to over-smoothness to some degree (Zhang et al. 2018). Besides, objects with little spatial information are likely to be misclassified, even for those with distinctive spectral characteristics (Zhang et al. 2018). In fact, any classifier, regardless of its model structure, predicts with uncertainty when handling spatially and spectrally complex VFSR imagery. A key problem to be addressed is, thus, for a given classification map, which areas are correctly classified and which are not? This information is important for classification map *producers* who need to further increase classification accuracy. Information on uncertainty is also very useful for classification map *users*, because if it is available, at least in some generalised form, users can better target their attention and effort. Currently, classification model uncertainty is assessed mainly using measures such as the difference between the first and second largest class membership value (Olofsson et al. 2014), Shannon's entropy (Wang and Shi 2013), α-quadratic entropy (Giacco et al. 2010), and so on, but there is generally a lack of objective and automatic approaches to partition and label the correct and incorrect classification regions.

The real problem with image classification, using a CNN or any other classifier, is, thus, how to reasonably describe and partition the geometric space given the inherent prediction uncertainties in a classification map. We previously proposed to create rules to threshold the classification results and deal with uncertainties through decision fusion (Zhang et al. 2018). This method, although having potential to achieve desirable classification results, involves a large amount of trial and error and prior knowledge of feature characteristics, thus was hard to be generalized and applied in an automatic fashion. As a well-established mathematical tool, rough set theory is proposed here as a means of providing an uncertainty description with no need for prior knowledge, and

this can be applied to model uncertainties of classification results.

Rough set theory, as proposed by Pawlak (1982), is an extension of conventional set theory that describes and models the vagueness and uncertainty in decision making (Pan et al. 2010). It has been applied in diverse domains such as pattern recognition (Swiniarski and Skowron 2003), machine learning (Chen et al. 2010), knowledge acquisition (Yu et al. 2014), and decision support systems (Zhan and Zhu 2017). Unlike other approaches that deal with vague concepts such as fuzzy set theory, rough set theory provides an objective form of analysis without any preliminary assumptions on membership association, thus, demonstrating power in information granulation (Qian et al. 2017) and uncertainty analysis (Chen et al. 2017). In the field of remote sensing and GIS, rough set theory has been applied in rule-based feature reduction and knowledge induction (Leung et al. 2008, Pan et al. 2010), land use spatial relationship extraction (Ge et al. 2011), spatio-temporal outlier detection (Albanese et al. 2014), and land cover classification and knowledge discovery (Sikder 2016). However, description of the uncertainty in remote sensing image classification results, as identified as a need and proposed here, has not been addressed through rough set theory, except for the pioneering work of Ge et al. (2009) on classification accuracy assessment. In fact, as one of the basic theories of granular computation, the predominant role of rough sets is to transform an original target granularity (i.e., continuous and intricate) into a simpler and more easily analysable variable. Thus, by using rough sets, the uncertainty of remote sensing classification can be simplified and the resulting data is more readily used to support decision-making.

In this chapter, a variant of rough set theory, variable precision rough set (VPRS) (Pan et al. 2010), is introduced for the first time to model and quantify the uncertainties in

CNN classification of VFSR imagery with a certain level of error tolerance, which is more suitable for the remote sensing domain than standard rough set theory due to its complexity. Through the VPRS theory, these classification uncertainties are partitioned and labelled automatically into positive regions (correct classifications), negative regions (misclassifications) and boundary regions (uncertain areas), respectively. These labelled regions are then used to guide the regional decision fusion for final classification. Specifically, the positive regions are trusted directly by the CNN, whereas the non-positive regions (negative and boundary regions) with high uncertainty (often occurring along object edges) are rectified by the results of an MLP-based MRF (MLP-MRF). Such a region-based fusion decision strategy performs classification integration at the regional level, as distinct from the commonly used pixel-based strategies. The proposed VPRS-based MRF-CNN regional decision fusion aims to capture the mutual complementarity between the CNN in spatial feature representation and the MLP-MRF in spectral differentiation and boundary segmentation.

The key innovations of this research can be summarized as: 1) a novel variable precision rough set model is proposed to quantify the uncertainties in CNN classification of VFSR imagery, and 2) a spatially explicit regional decision fusion strategy is introduced for the first time to improve the classification in uncertain regions using the distribution characteristics of the CNN classification map.

The effectiveness of the proposed method was tested on images of both an urban scene and a rural area as well as semantic labelling datasets. A benchmark comparison was provided by pixel-based MLP and SVM, spectral-contextual based MLP-MRF as well as contextual-based CNN classifiers, together with mainstream baseline methods.

4.2 Methodology

A novel VPRS-based method for regional decision fusion of CNN and MRF (MRF-CNN) is proposed for the classification of VFSR remotely sensed imagery. The methodology consists of the following steps:

1. perform CNN and MLP classification using a training sample set (*T*1) and validate them using a testing sample set (*T*3),

2. estimate the uncertainty of the CNN classification result to achieve a CNN classification confidence map (CCM), and perform MLP-based MRF (MLP-MRF) classification,

3. construct a VPRS fusion decision model to partition the CCM into positive regions and non-positive (i.e. boundary and negative) regions using a test sample set (denoted as *T*2), and

4. obtain the final classification result by taking the classification results of the CNN for the positive regions and those of MLP-MRF for the non-positive regions.

Principles and major workflows are detailed hereafter.

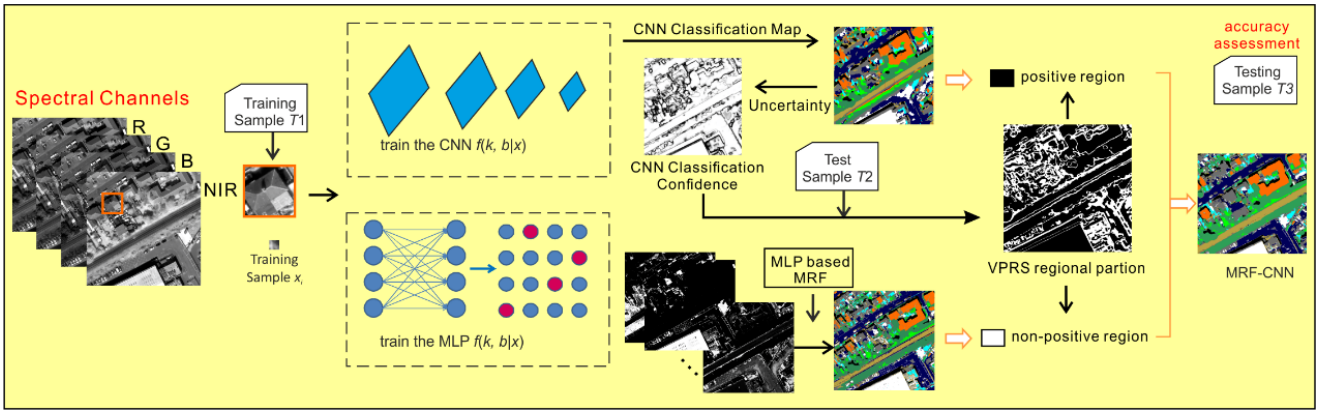


Figure 4-1: A workflow illustrating the proposed MRF-CNN methodology.

4.2.1 Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a multi-layer feed-forward neural network that is designed specifically to process large scale images or sensory data in the form of multiple arrays by considering local and global stationary properties (LeCun et al. 2015). The main building block of a CNN is typically composed of multiple layers interconnected to each other through a set of learnable weights and biases (Romero et al. 2016). Each of the layers is fed by small patches of the image that scan across the entire image to capture different perspectives of features at local and global scales. Those image patches are generalized through a convolutional layer and a pooling/subsampling layer alternatively within the CNN framework, until the high-level features are obtained on which a fully connected classification is performed (LeCun et al. 2015). Additionally, several feature maps may exist in each convolutional layer and the weights of the convolutional nodes in the same map are shared. This setting enables the network to learn different features while keeping the number of parameters tractable. Moreover, a nonlinear activation (e.g. sigmoid, hyperbolic tangent, rectified linear units) function is taken outside the convolutional layer to strengthen the non-linearity (Strigl et al. 2010). Specifically, the major operations performed in the CNN can be summarized as:

$$O^l = pool_p(\sigma(O^{l-1} * W^l + b^l)) \tag{4-1}$$

where the O^{l-1} denotes the input feature map to the lth layer, the W^l and the b^l represent the weights and biases of the layer, respectively, that convolve the input feature map through linear convolution*, and the $\sigma(\cdot)$ indicates the non-linearity function outside the convolutional layer. These are often followed by a max-pooling operation with $p \times p$

window size ($pool_p$) to aggregate the statistics of the features within specific regions, which forms the output feature map O^l at the lth layer (Romero et al. 2016).

4.2.2 Multilayer perceptron based Markov random field (MLP-MRF)

A multilayer perceptron (MLP) is a classical neural network model that maps sets of input data onto a set of outputs in a feed-forward manner (Atkinson and Tatnall 1997). The typical structure of a MLP is cascaded by interconnected nodes at multiple layers (input, hidden and output layers), with each layer fully connected to the preceding layer as well as the succeeding layer (Del Frate et al. 2007). The outputs of each node are weighted units and biases followed by a non-linear activation function to distinguish the data that are not linearly separable (Pacifici et al. 2009). The weights and biases at each layer are learned by supervised training using a back-propagation algorithm to approximate an unknown input-output relation between the input feature vectors and the desired outputs (Del Frate et al. 2007).

The predictive output of the MLP is the membership probability/likelihood to each class at the pixel level, which forms the conditional probability distribution function according to the Bayesian theorem (Foody 2000). The objective of Bayesian prediction is to achieve the maximum posterior probability by combining the prior and conditional probability distribution functions, so as to solve the classification problem effectively. The MRF classifier provides a convenient way to model the local properties of an image into positivity, Marknovianity and Homogeneity as its prior probability, together with the learnt likelihood from the MLP, which constitutes the MLP-MRF (Dunne and Campbell 1995, Tso and Mather 2009). Such local neighbourhood information can further be converted into its global equivalence of the Gibbs random field as an energy function based on the Hamersley-Clifford theorem (Wang and Liu 1999). The MLP-

MRF is hence iteratively solved by minimizing the energy function to search for the global minima. See Tso and Mather (2009) and Li (2009) for more theoretical concepts on MLP-based MRF and its application to image classification.

4.2.3 VPRS based decision fusion between CNN and MRF

4.2.3.1 Introduction to variable precision rough set theory

In rough set theory (Pawlak 1982), a dataset is represented as a table, which is called an information system, denoted as $S = (U, A)$, where U is a non-empty finite set of objects known as the universe of discourse, and A is a non-empty finite set of attributes, such that $U \rightarrow Va$ exists for each $a \in A$. The set Va denotes the set of attribute values that a may take. A decision table is an information system in the form of $S = (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. For any attribute set $P \subseteq A$, there is an indiscernible relation R between two objects x and y:

$$R = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\} \tag{4-2}$$

where R explains that the x and y are indiscernible by the attributes from P (i.e. both x and y share the same attribute values).

The equivalence classes of the indiscernible relation based on R can be defined as:

$$[x]_R = \{y \in U \mid (x, y) \in R\} \tag{4-3}$$

Given a target set $X \subseteq U$, X can then be approximated by using the equivalence classes of the indiscernible relation R, including a R-lower approximation: $\underline{R}X = \{x \mid [x]_R \subseteq X\}$ and a R-upper approximation: $\overline{R}X = \{x \mid [x]_R \cap X \neq \phi\}$. If $\underline{R}X \neq \overline{R}X$, then the tuple $(\underline{R}X, \overline{R}X)$ forms a rough set. The positive ($POS_R(X)$), negative ($NEG_R(X)$) and boundary

($BND_R(X)$) regions can be defined as:

$$POS_R(X) = \underline{R}X \tag{4-4}$$

$$NEG_R(X) = U - \overline{R}X \tag{4-5}$$

$$BND_R(X) = \overline{R}X - POS_R(X) \tag{4-6}$$

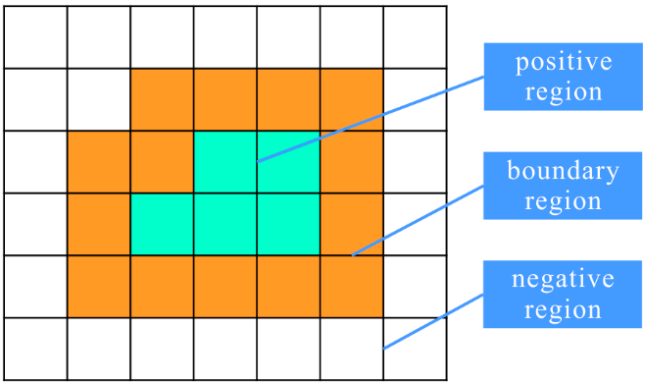


Figure 4-2: An illustration of the standard rough set with positive, boundary and negative regions.

However, the above standard definition of the set inclusion relation is too rigorous to represent any "almost" complete set inclusion (Ziarko 1993) (i.e., equation (4-4) is difficult to be satisfied strictly). Thus, a variable precision rough set (VPRS) model was proposed to allow a certain number of inclusion errors. Let X and Y be two non-empty subsets of a finite universe U, the degree (or level) of inclusion error of Y within X can be defined as (Chen et al. 2017):

$$e(Y,X) = 1 - \frac{Card(Y \cap X)}{Card(Y)}, \ (Y \neq \phi) \tag{4-7}$$

where the $Card(*)$ denotes the cardinality of a set. The $e(Y,X) = 0$ if and only if $Y \subseteq X$, that is, the case of standard rough set theory (Figure 4-2). Suppose $e(Y,X) \neq 0$, then a level of inclusion error β is introduced to tolerate a certain level of inclusion. Given a

level of inclusion error β, Y being included by X can be defined as:

$$Y \subseteq_{\beta} X \quad iff \quad e(Y,X) \le \beta, \quad 0 \le \beta \le 1 \tag{4-8}$$

Having defined the relative inclusion error β, the β-lower approximation and the β-upper approximation can be characterized as:

$$\underline{R}_{\beta}X = \{x \in U \mid e([x]_R, X) \le \beta\} \tag{4-9}$$

$$\overline{R}_{\beta}X = \{x \in U \mid e([x]_R, X) \le 1-\beta\} \tag{4-10}$$

Given equations (4-9) and (4-10), the positive ($POS_{R,\beta}(X)$), negative ($NEG_{R,\beta}(X)$) and boundary ($BND_{R,\beta}(X)$) regions with a level of inclusion error β can be inferred as:

$$POS_{R,\beta}(X) = \underline{R}_{\beta}X \tag{4-11}$$

$$NEG_{R,\beta}(X) = U - \overline{R}_{\beta}X \tag{4-12}$$

$$BND_{R,\beta}(X) = \overline{R}_{\beta}X - POS_{R,\beta}(X) \tag{4-13}$$

4.2.3.2 VPRS-based MRF-CNN fusion decision

Suppose the membership prediction of the CNN at each pixel is an n-dimensional vector $C = (c_1, c_2, ..., c_n)$, where n represents the number of classes, while each dimension $c_i (i \in [1, n])$ corresponds to the pixel's probability of a specific (i-th) class with certain membership association. Ideally, the probability value of the classification prediction is 1 for the target class but 0 for the other classes, which is usually unobtainable due to the extensive uncertainty in the process of remotely sensed image classification. The probability value C is, therefore, denoted as:

$$f(z) = \{c_z \mid z \in (1,2,...,n)\} \; c_z \in [0,1], \sum_1^n c_z = 1 \tag{4-14}$$

By default, the classification model simply takes the maximum membership association as the predicted output label (denoted as class(*C*)):

$$class(C) = \arg\max_{z}(\{ f(z) = c_z \mid z \in (1,2,...,n)\}) \tag{4-15}$$

The confidence of being determined as *class*(*C*) is derived from one minus the normalized Shannon Entropy (Ge et al. 2009):

$$conf = 1 - \frac{E_i}{E_{max} - E_{min}} = 1 - \frac{-\sum_{z=1}^{n} f_i(z)\log_2(f_i(z))}{E_{max} - E_{min}} \tag{4-16}$$

where, $E_i = -\sum_{z=1}^{n} f_i(z)\log_2(f_i(z))$ denotes the entropy value of the *i*th pixel, whereas the E_{max} and the E_{min} refer to the maximum and minimum entropy values, respectively, of the entire classification map. When the entropy of a pixel is maximized (i.e., E_{max} in equation (4-16)), $f(z)$ approximates a uniform probability distribution, representing that there is a strong possibility that the pixel is wrongly classified, and therefore the confidence value *conf* tends to be small (i.e., the level of the corresponding uncertainty tends to be higher) and *vice versa*. Therefore, the *conf* $(\in [0,1])$ is inversely correlated with the normalized entropy.

Given a CNN classification map, the confidence value of an object is spatially heterogeneous: the central region is often accurately classified, but the boundary region is likely to be misclassified (Zhang et al. 2018). The two regions (i.e., patch centre and patch boundary) can then be described theoretically by using rough set theory (Pan et al. 2010). That is, the correctness, incorrectness and uncertainty of image classification can be modelled via the positive (equation (4-4)), negative (equation (4-5)) and

boundary (equation (4-6)) regions, respectively.

The decision attribute $\{d\}$ of the rough set model, commonly referred to as the attribute for the identification of a specific land cover class, is used here to describe whether a test sample is correctly classified (i.e., a strength and weakness analysis on the classification results of the region corresponding to the sample). The confidence value (*conf*) of any two samples within this region should belong to the same indiscernible relation, of which they should be treated simultaneously. For the confidence map of CNN classification (i.e., the image with a *conf* value at each pixel), it can, therefore, be partitioned into a series of intervals, each of which represents a particular indiscernible relation:

$$[0, step), [step, step \times 2), \ldots, [step \times floor(conf / step), 1] \tag{4-17}$$

where, *step* is the atomic granule representing the least unit of indiscernible relation. Each interval forms an indiscernible region (denoted as IND_{Area}) on the CNN classification map. By checking the consistency of the classification results with respect to the test samples ($T2$), the partitions can then be characterized as: the positive region (the negative region, respectively) where the entirety of $T2$ lying in the region are correctly (incorrectly, respectively) classified, and the boundary region in which the $T2$ are partially correctly classified.

There exists extensive uncertainty and inconsistency in remotely sensed image classification, especially for VFSR imagery. A small amount of error (even with only one misclassified sample) could inevitably turn a positive region into a boundary region. Thus, equation (4-4) is too restrictive and might not be sufficiently satisfied. Therefore, the introduction of the VPRS model with a relative classification error β is

necessary to allow for some degree of misclassification in the largely correct classification. Based on the VPRS model, the CNN classification confidence map can be partitioned into indiscernible regions (i.e. IND_{Area}). The accuracy of each region is evaluated further using the test sample sets ($T2$) to quantify the ratio of the labelled samples that are consistent or inconsistent to the categories of the classification results. Those indiscernible regions that meet the accuracy requirements of (equation (4-11)) are labelled as positive regions, whereas those fitting equations (4-12) and (4-13) are characterised as non-positive regions.

As illustrated by equation (4-7), the real level of inclusion error (denoted as *error*) in a specific IND_{Area} is essentially the classification error of the test sample ($T2$), that is, the ratio between the number of misclassified samples and the total number of the samples within the region. The IND_{Area} can then be identified either as a *positive region* or a *non-positive region* based on the relative inclusion error β:

$$IND_{Area} = \begin{cases} positive\ region & error \le \beta \\ non\text{-}positive\ region & error > \beta \end{cases} \tag{4-18}$$

The final classification results of all pixels within the region can then be determined by using either the results ($class_{cnn}$) of CNN (in the case of *positive region*), or the results ($class_{mlp\text{-}mrf}$) of MLP-MRF (in the case of *non-positive region*). The positive region and the non-positive region are, therefore, allocating priority to the CNN and the MLP-MRF accordingly.

Following the strategy mentioned above, the VPRS-based decision fusion algorithm for remotely sensed image classification is illustrated using pseudo-code in Table 4-1:

Table 4-1 - Detailed description of the VPRS-based regional decision fusion algorithm for remotely sensed image classification

VPRS-Based Regional Decision Fusion Algorithm	
Input: remotely sensed (RS) image, level of inclusion error β, training sample set $T1$, rough set test sample set $T2$, atomic granule *step*	
Output: classification result result*Img*	
1.	$Model_{cnn}$ = The CNN model trained by sample set $T1$
2.	$Model_{mlp\text{-}mrf}$ = The MLP-MRF model trained by sample set $T1$
3.	fuzzyMatrix = The RS image classified by using $Model_{cnn}$ to obtain decision vector
4.	*conf* = The uncertain level within fuzzyMatrix (1 – Normalized Entropy)
6.	For each region IND_{Area} partitioned from *conf* using an atomic granule *step*
7.	using *error* (derived from $T2$) and β to determine IND_{Area} (18)
8.	If *error* $\leq \beta$ then IND_{Area} belongs to *positive region*
9.	result*Pixels* = each pixel within IND_{Area} is classified by CNN ($class_{cnn}$)
10.	Else IND_{Area} belongs to *non-positive region*
11.	result*Pixels* = each pixel within IND_{Area} is classified by MLP-MRF ($class_{mlp\text{-}mrf}$)
12.	End if
13.	result*Img* = result*Img* ∪ result*Pixels*
14.	End for
15.	Return result*Img*
16.	End

4.3 Experimental Results and Analysis

4.3.1 Data description and experimental design

Experiment 1: The city of Bournemouth, UK and its surrounding environment, located on the southern coast of England, was selected as a case study area (Figure 4-3). The urban area of Bournemouth city is very developed with a high density of anthropogenic structures such as residential houses, commercial buildings, roads and railways. In the contrast, the suburban and rural areas near Bournemouth are less densely populated, predominantly covered by natural and semi-natural environments.

An aerial image was captured on 20 April 2015 using a Vexcel UltraCam Xp digital

aerial camera with 25 cm spatial resolution and four multispectral bands (Red, Green, Blue and Near Infrared), referenced to the British National Grid coordinate system (Figure 4-3). Two subsets of the imagery with different environmental settings, including S1 (2772×2515 pixels) within Bournemouth city centre and S2 (2639×2407 pixels) in the rural and suburban area were chosen to test the classification algorithms. S1 consists mainly of nine dominant land cover classes, including Clay roof, Concrete roof, Metal roof, Asphalt, Railway, Grassland, Trees, Bare soil and Shadow, listed in Table 4-2. S2 includes Queen's Park Golf Course and is comprised of large patches of woodland, grassland and bare soil speckled with small buildings and roads. There are seven land cover categories in this study site, namely, Clay roof, Concrete roof, Road-or-track, Grassland, Trees, Bare soil and Shadow (Table 4-2).

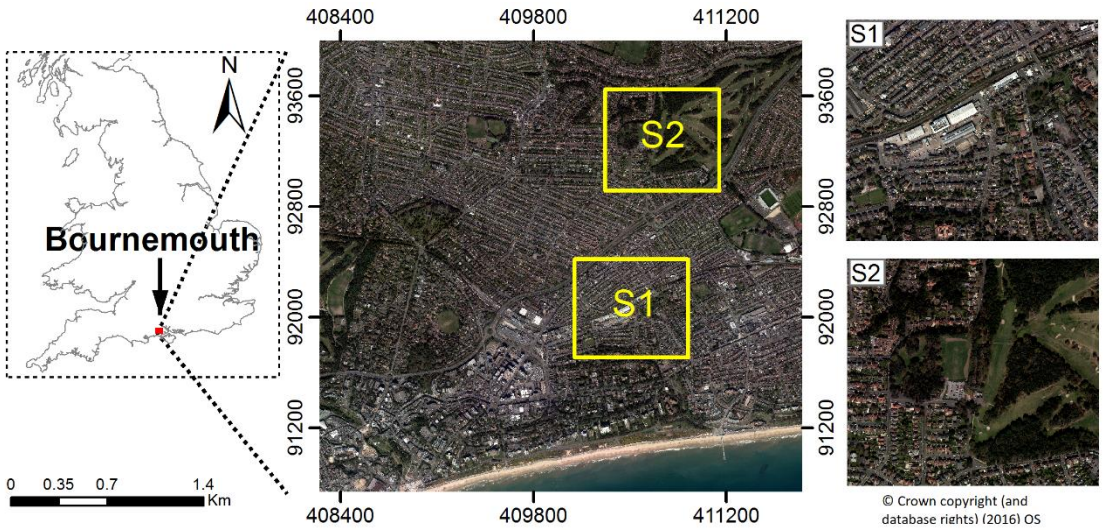


Figure 4-3: Location of study area at Bournemouth within the UK, and aerial imagery showing zooms of the two study sites S1 and S2.

Sample points were collected using a stratified random scheme from ground data provided by local surveyors in Bournemouth, and split into 50% training samples (Training Sample *T*1 at Table 4-2) and 50% testing samples (Testing Sample *T*3 at Table 4-2) for each class. In addition, a set of test samples (Test Sample *T*2, see Table

4-2) with which to construct the variable precision rough set (VPRS) model were stratified randomly collected throughout the imagery and manually labelled into different land cover classes. The sample labelling was based on expert knowledge and historical references provided by local surveyors and photogrammetrists. Field survey was conducted on April 2015 to further check the validity and precision of the selected samples. Moreover, a highly detailed vector map from the Ordnance Survey, namely the MasterMap Topography Layer (Regnauld and Mackaness 2006), was fully consulted and cross-referenced to gain a comprehensive appreciation of the land cover and land use within the study area.

Table 4-2 - Land cover classes at two study sites with training and testing sample size per class. training sample *T*1 and testing sample *T*3 were used for model construction and accuracy validation, while test sample *T*2 was used for building the variable precision rough set.

Study Sites	Class	Training Sample *T*1	Test Sample *T*2	Testing Sample *T*3
S1	Clay roof	110	156	110
	Concrete roof	107	148	107
	Metal roof	103	139	103
	Asphalt	107	148	107
	Grassland	114	162	114
	Trees	104	141	104
	Bare soil	103	139	103
	Shadow	103	139	103
	Railway	102	137	102
S2	Clay roof	82	104	82
	Concrete roof	90	115	90
	Road-or-track	85	108	85
	Grassland	86	110	86
	Trees	98	124	98
	Bare soil	84	106	84
	Shadow	86	110	86

Experiment 2: Two well-known semantic labelling datasets, the Vaihingen dataset and the Potsdam dataset, were used to further evaluate the effectiveness of the proposed method.

The Vaihingen dataset contains 33 true orthophoto tiles with a spatial resolution of 9 cm. For each tile, four channels are provided, namely near-infrared (NIR), red (R) and green (G), together with digital surface models (DSMs). Six semantic categories were manually classified by ISPRS, including impervious surfaces, building, low vegetation, tree, car, and clutter/background. As previously with other authors (e.g. Kampffmeyer et al. 2016, Volpi and Tuia 2017), the clutter/background class (mainly involving water bodies, background and others) was not considered in the experiments since it accounts only for 0.88% of the total number of pixels.

Following the same training and testing procedures set by FCN (Kampffmeyer et al. 2016) and SegNet (Volpi and Tuia 2017), we used the sixteen annotated tiles in our experiments. Eleven tiles (areas: 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37) were selected for training, while the other five tiles (areas: 11, 15, 28, 30, 34) were reserved for testing.

The Potsdam 2D segmentation dataset includes 38 tiles of fine spatial resolution remote sensing images. All of them feature a spatial resolution of 5 cm and have a uniform resolution of 6000×6000 pixels. Twenty-four tiles are provided with Ground Reference pixel labels, using the same five classes as in the Vaihingen dataset without the clutter/background class. In the experiments, Following the practice in (Kampffmeyer et al. 2016), six tiles (02_12, 03_12, 04_12, 05_12, 06_12, 07_12) were selected as the testing set, while the other eighteen among the annotated tiles were used for training.

Sample points for both datasets were acquired using a stratified random scheme from the Ground Reference with a stride of 300 pixels to ensure the adequacy of GPU

memory, and these were partitioned into 30%, 40% and 30% sets for Training Sample *T*1, Test Sample *T*2 and the Testing Sample *T*3. SVM and other mainstream methods, such as FCN (Kampffmeyer et al. 2016), SegNet (Volpi and Tuia 2017) and Deeplab-v2 (Chen et al. 2016), were applied as benchmarks.

4.3.2 Model Architectures and Parameter Settings

Since the MRF used in this research was based on the probabilistic output from a pixel-based MLP, good choices for the model architectures and parameter settings of the MLP and CNN are essential for the proposed MRF-CNN approach. To make a fair comparison, both CNN and MLP models were assigned the same parameters for the learning rate as 0.1, the momentum factor as 0.7, the logistic non-linearity function, and the maximum iteration number of 1000 to allow the networks to fully converge to a stable state through back-propagation. In the MLP, the numbers of nodes and hidden layers were tuned with 1-, 2-, and 3-hidden layers through cross-validation, and the best predicting MLP was found using two hidden layers with 20 nodes in each layer. For the CNN, a range of parameters including the number of hidden layers, the input image patch size, the number and size of convolutional filter, need to be tuned (Romero et al. 2016). Following the discussion by Längkvist et al. (2016), the input patch size was chosen from {12×12, 14×14, 16×16, 18×18, 20×20, 22×22, 24×24} to evaluate the influence of context area on classification performance. In general, a small-sized contextual area results in overfitting of the model, whereas a large one often leads to under-segmentation. In consideration of the image object size and contextual relationship coupled with a small amount of trial and error, the optimal input image patch size was set to 16×16 in this research. Besides, as discussed by Chen et al. (2014) and Längkvist et al. (2016), the depth plays a key role in classification accuracy because the quality of learnt feature is highly influenced by the level of abstraction and

representation. As suggested by Längkvist et al. (2016), the number of CNN hidden layers was chosen as four to balance the network complexity and robustness. Other parameters were tuned empirically based on cross-validation accuracy, for example, the kernel size of the convolutional filters within the CNN was set as 3×3 and the number of filters was tuned as 24 at each convolutional layer.

The MLP-MRF requires to predefine a fixed size of neighbourhood and a parameter γ that controls the smoothness level. The window size of the neighbourhood in the MLP-MRF model was chosen optimally as 7×7 in consideration of the spatial context and the fidelity maintained in the classification output. Due to the fine spatial detail contained in the VFSR imagery, the parameter γ controlling the level of smoothness was set as 0.7 to achieve an increasing level of smoothness in terms of the MRF. The simulated annealing optimization using a Gibbs sampler (Berthod et al. 1996) was employed in MLP-MRF to maximize the posterior probability through iteration.

An SVM classifier was further used as a benchmark comparator to test the classification performance. The SVM model involves a penalty value C and a kernel width σ that needs to be parameterised. Following the recommendation by Zhang et al. (2015), a grid search with 5-fold cross-validation was implemented to exhaustively search within a wide parameter space (C and σ within $[2^{-10}, 2^{10}]$). Such parameter settings would lead to high validation accuracy using support vectors to formulate an optimal classification hyperplane.

4.3.3 Decision Fusion Parameter Setting and Analysis

The decision fusion between the MLP-MRF and the CNN, namely, the MRF-CNN, based on the VPRS model, involves parameters β (the level of inclusion error) and *step* (the atomic granule). The two parameters were optimized through grid search with

cross-validation using Training Sample 2 (Listed in Table 4-2). Specifically, β was varied from 0 to 1 with incremental steps of 0.01, while the *step* was tuned between 0 to 0.5 through a small step of 0.025 (i.e. with a wider parameter searching space) to obtain a higher validation accuracy. By doing so, β and *step* were chosen optimally as 0.1 and 0.075, respectively.

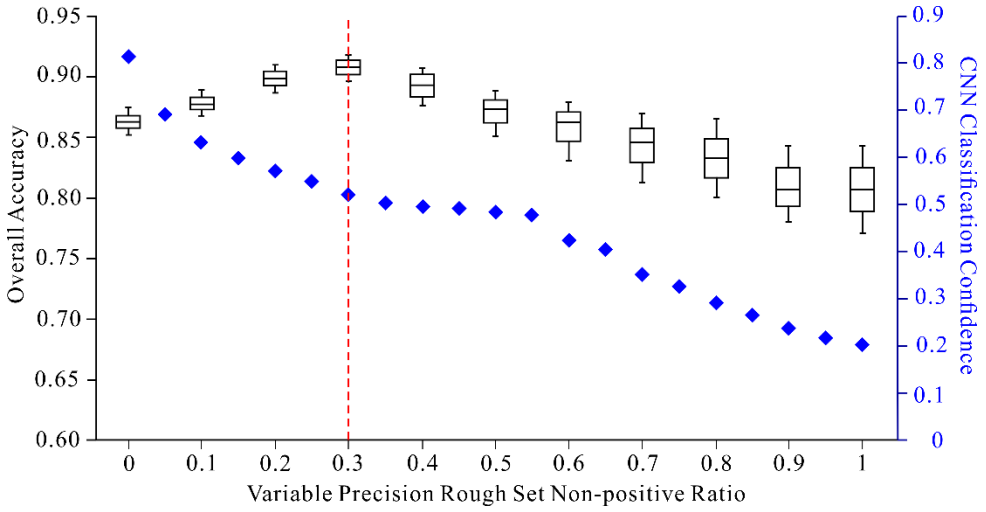


Figure 4-4: The CNN classification confidence value and the overall accuracy influenced by the fusion decision parameter setting (in the form of the non-positive to positive ratio).

Both of the fusion decision parameters (β and *step*) jointly determined the partition of the positive and non-positive regions. As shown in Figure 4-4, these parameter settings, reflected by variation between the ratios of VPRS non-positive and positive regions (horizontal axis coordinates ranging from 0 to 1), have an impact on the CNN classification confidence values (blue dots) and the overall accuracies (boxplots). From the figure, it can be seen that along with the increase of the non-positive ratio, the CNN classification confidence decreases constantly, except for the non-positive ratio from 0.3 to 0.55; whereas the overall accuracy initially increases from around 0.86 to around 0.9 and then decreases constantly until around 0.81. Another observation is that the boxplot tends to be wider as the ratio of non-positive to positive region becomes larger,

with more credits being given from the CNN to the MLP-MRF. The optimal non-positive ratio (determined by decision fusion parameter setting) was found to be 0.3 (marked by the red dotted line in Figure 4-4).

4.3.4 Classification Results and Analysis

Experiment 1: The classification performance of the MRF-CNN and the other benchmark methods, including the MLP, SVM, MLP-MRF and the CNN, were compared using the Testing samples of Bournemouth dataset. Table 4-3 lists the detailed accuracy assessment of both S1 for Bournemouth city centre and S2 for the rural and suburban areas with overall accuracy (OA), Kappa coefficient (κ) as well as per-class mapping accuracy. Clearly, the MRF-CNN achieved the best overall accuracy of 90.96% for S1 and 89.76% for S2 with Kappa coefficients of 0.89 and 0.88 respectively, consistently higher than the CNN (85.37% and 86.39% OA with κ of 0.84 and 0.83, respectively), the MLP-MRF (83.76% and 84.52% with corresponding κ of 0.79 and 0.80), the SVM (81.65% and 81.24% with corresponding κ of 0.77 and 0.78), and the MLP (81.52% and 80.32% with the same κ of 0.77) (Table 4-3). In addition, a McNemar *z*-test that accounts for the pair-wise classification comparison further demonstrates that a statistically significant increase has been achieved by the MRF-CNN over the MLP, SVM, MLP-MRF and the CNN, with *z*-value = 3.27, 3.02, 2.74 and 2.02 in S1 and *z*-value = 3.89, 3.51, 3.06 and 2.05 in S2 respectively, greater than 1.96 at 95% confidence level (Table 4-4). Moreover, the class-wise classification accuracy of MRF-CNN constantly reports the most accurate results highlighted by the bold font in Table 4-3, except for the trees in S2 (89.32%) for which accuracy is slightly lower than for the CNN (90.42%). In particular, the mapping accuracies of most land covers classified by the MRF-CNN were higher than 90%, with the greatest accuracy achieved in grassland at both study sites S1 and S2, up to 93.57% and 92.94%,

respectively.

With respect to the four benchmark classifiers themselves (i.e., MLP, SVM, MLP-MRF and CNN), it can be seen from Table 4-3 that their classification accuracies are ordered as: MLP < SVM < MLP-MRF < CNN. For the urban area at S1, the accuracy of the MLP-MRF and the SVM is closer to the MLP (<2%), but with larger difference (>3%) from the CNN. This is further demonstrated by the McNemar *z*-test in Table 4-4 where the CNN is significantly different from the MLP, the SVM and the MLP-MRF (z = 3.12, 2.85 and 2.14, respectively), but the increase of the MLP-MRF is not significant compared with the MLP (z = 1.57) and the SVM (z = 1.68). In the rural area at S2, on the contrary, the accuracy of the MLP-MRF is remarkably higher (>4%) than that of the MLP and SVM with statistical significance (z = 2.12 and 2.04), and only slightly lower than that of the CNN (<2%) without significant difference (z = 1.59).

Figure 4-5 and 4-6 demonstrate visual comparisons of the five classification results using three subset images at each study site (S1 and S2). For the Concrete roof class, from the upper right of Figure 4-5(*a*), it is clear that the MLP and SVM classification results maintain the rectangular geometry of the building, but at the same time present very noisy information with salt-and-pepper effects in white throughout the Concrete roof (see the red circles at the figure). Such noise has been largely reduced by the MLP-MRF but still not yet completely eliminated (shown by red circle). The noise has been erased thoroughly by the CNN. However, some serious mistakes have been introduced by misclassifying the asphalt on top of the Concrete roof (highlighted by red circle). Fortunately, the MRF-CNN removed all of the noise while keeping the correctness of the semantic segmentation (yellow circle). A similar pattern was found in the middle of Figure 4-5(*b*), where the MLP-MRF is less noisy than the MLP and the SVM (red

circles), and the CNN obtains the smoothest classification result, but tends to be under-segmented along the object boundaries (highlighted by red circle). The MRF-CNN, in contrast, keeps the central regions smooth while preserving the precise boundary information (e.g. the rectangularity of the concrete roofs and the shadow next to them; shown in yellow circle). Similar situations are found in the Clay roof, as shown in Figure 4-6(*a*) and 4-6(*c*), where the MLP, SVM and MLP-MRF introduced some noise in the central region, whereas the CNN eradicated them but with obvious geometric distortions. The MRF-CNN, surprisingly, removes all the noise while keeping the crisp boundaries with accuracy. In terms of the railway class illustrated in the middle of Figure 4-5(*a*), it was noisily classified by the MLP, the SVM and the MLP-MRF (red circles). This noise was eliminated by the CNN as well as the MRF-CNN (yellow circles). Moreover, some small Road-or-tracks exemplified by Figure 4-6(*a*) and 4-6(*b*) were successfully maintained by the MLP, SVM, MLP-MRF as well as MRF-CNN, yet omitted by CNN due to the convolutional operations.

For the natural land cover classes, the grassland patch shown in Figure 4-5(*b*) is shaped approximately square (see the original image in Figure 4-5(*b*)). The MLP and SVM produced noisy results confused with the surrounding tree species (shown in red circles). A similar pattern was found in the result of the MLP-MRF but with less noise (marked by red circle). The CNN and the MRF-CNN did not show any noise in the classification map. However, the CNN did not maintain the squared shape of the grassland (shown in red circle), whereas the MRF-CNN successfully kept the geometric fidelity as a square shaped object (highlighted by yellow circle). With regard to the Trees indicated in Figure 4-6(*a*) and 4-6(*b*), the MLP, SVM and MLP-MRF produce different noise: the MLP tends to misclassify the trees as grassland (shown in red circle), whereas the SVM and MLP-MRF sometimes falsely considers the leaf-off trees

or the shade of trees as the shaded Clay roof (marked by red circle). All these misclassifications are rectified by the CNN and the MRF-CNN (in yellow circle).

Table 4-3 - Classification accuracy comparison amongst MLP, SVM, MLP-MRF, CNN and the proposed MRF-CNN approach for Bournemouth city centre (S1) and the suburban area (S2) using the per-class mapping accuracy, overall accuracy (OA) and kappa coefficient (κ). the bold font highlights the greatest classification accuracy per row.

Study Sites	Class	MLP	SVM	MLP-MRF	CNN	MRF-CNN
S1	Clay roof	91.37%	91.45%	90.58%	88.56%	**92.68%**
	Concrete roof	68.52%	68.74%	72.23%	74.37%	**78.25%**
	Metal roof	89.75%	89.52%	90.12%	91.42%	**92.23%**
	Asphalt	88.59%	88.55%	88.67%	85.98%	**91.26%**
	Grassland	73.51%	74.28%	76.42%	88.63%	**93.57%**
	Trees	65.68%	65.79%	72.28%	82.28%	**88.53%**
	Bare soil	80.46%	80.51%	80.82%	85.23%	**90.24%**
	Shadow	91.56%	91.23%	91.23%	90.14%	**92.16%**
	Railway	82.14%	82.35%	83.57%	90.23%	**91.56%**
	OA	81.52%	81.65%	83.26%	86.37%	**90.96%**
	κ	0.77	0.77	0.79	0.84	**0.89**
S2	Clay roof	88.56%	88.27%	86.75%	82.37%	**90.16%**
	Concrete roof	79.84%	79.62%	81.26%	84.17%	**88.27%**
	Road-or-track	83.02%	83.36%	83.17%	86.54%	**92.38%**
	Grassland	72.11%	73.64%	80.57%	88.58%	**92.94%**
	Trees	79.31%	79.24%	85.26%	90.42%	**89.32%**
	Bare soil	76.18%	76.42%	78.25%	81.36%	**88.75%**
	Shadow	89.42%	89.56%	89.42%	88.25%	**89.58%**
	OA	80.32%	81.24%	84.52%	86.39%	**89.76%**
	κ	0.77	0.78	0.80	0.83	**0.88**

As for the other land cover classes (e.g., bare soil and shadow) the four classification methods do not show significant differences, although some increases in classification accuracy were still obtained by the MRF-CNN. For example, the bare soil shown in Figure 4-6(*c*) is highly influenced by the cars and other small objects, which results in over-segmented noise by the MLP and the SVM (shown in red circles) or false identification into Clay roof by the CNN (marked in red circle). The MLP-MRF and the proposed MRF-CNN, fortunately, addressed those challenges with smooth yet semantically accurate geometric results (in yellow circle).

Table 4-4 - McNemar *Z*-test comparing the performance of the four classifiers for two study sites s1 and s2. significantly different accuracies with confidence of 95% (*z*-value > 1.96) are indicated by *.

Study sites	Classifiers	McNemar *Z*-test				
		MLP	SVM	MLP-MRF	CNN	MRF-CNN
S1	MLP	—				
	SVM	1.32	—			
	MLP-MRF	1.57	1.68	—		
	CNN	3.12*	2.85*	2.14*	—	
	MRF-CNN	3.27*	3.02*	2.74*	2.02*	—
S2	MLP	—				
	SVM	1.66	—			
	MLP-MRF	2.12*	2.04*	—		
	CNN	2.42*	2.15*	1.59	—	
	MRF-CNN	3.89*	3.51*	3.06*	2.05*	—

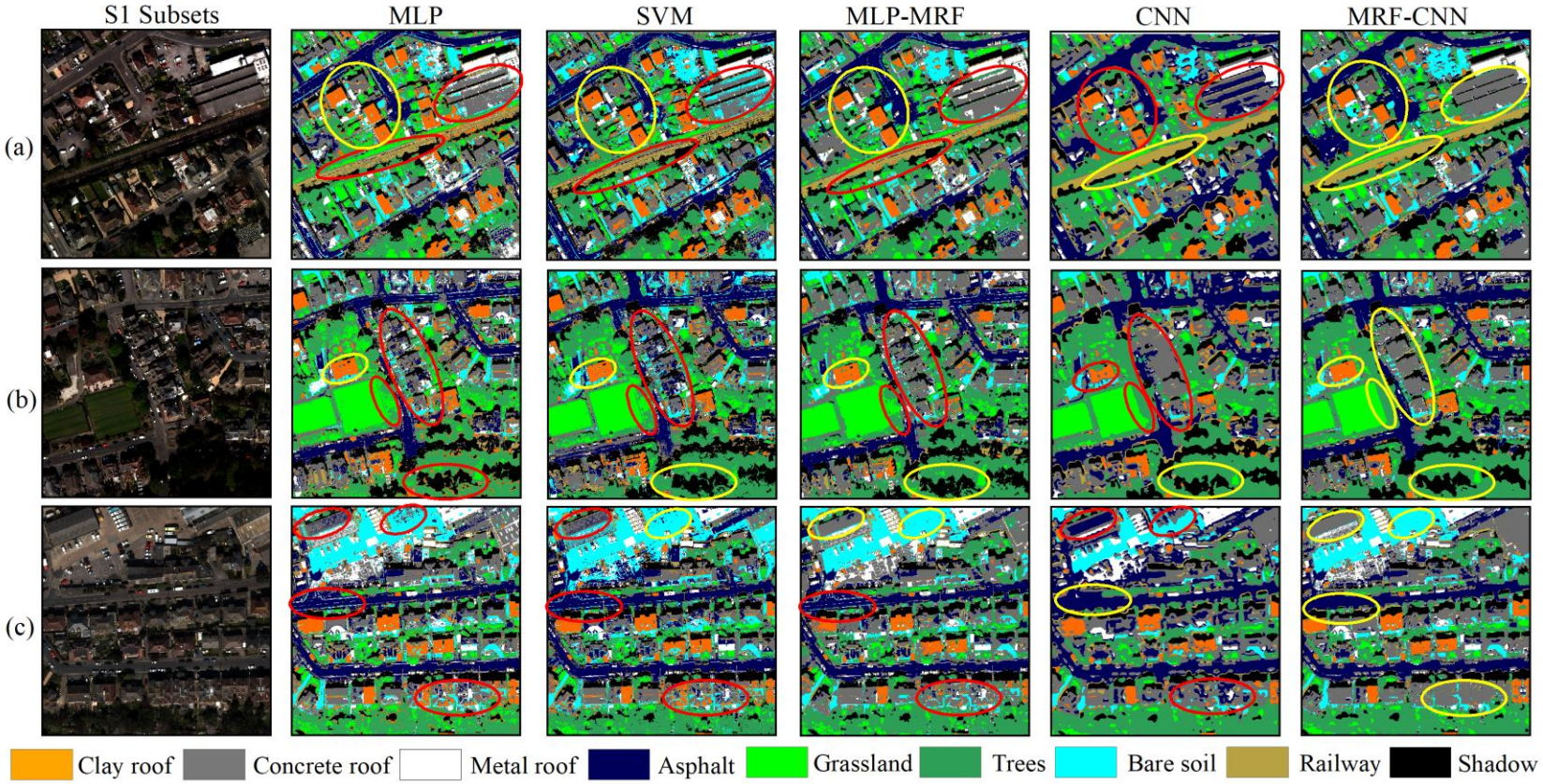


Figure 4-5: Three typical image subsets (a, b and c) in study site S1 with their classification results. Columns from left to right represent the original images (R G B bands), the MLP, the SVM, the MLP-MRF, the CNN, and the MRF-CNN classification results. The red and yellow circles denote incorrect and correct classification, respectively.

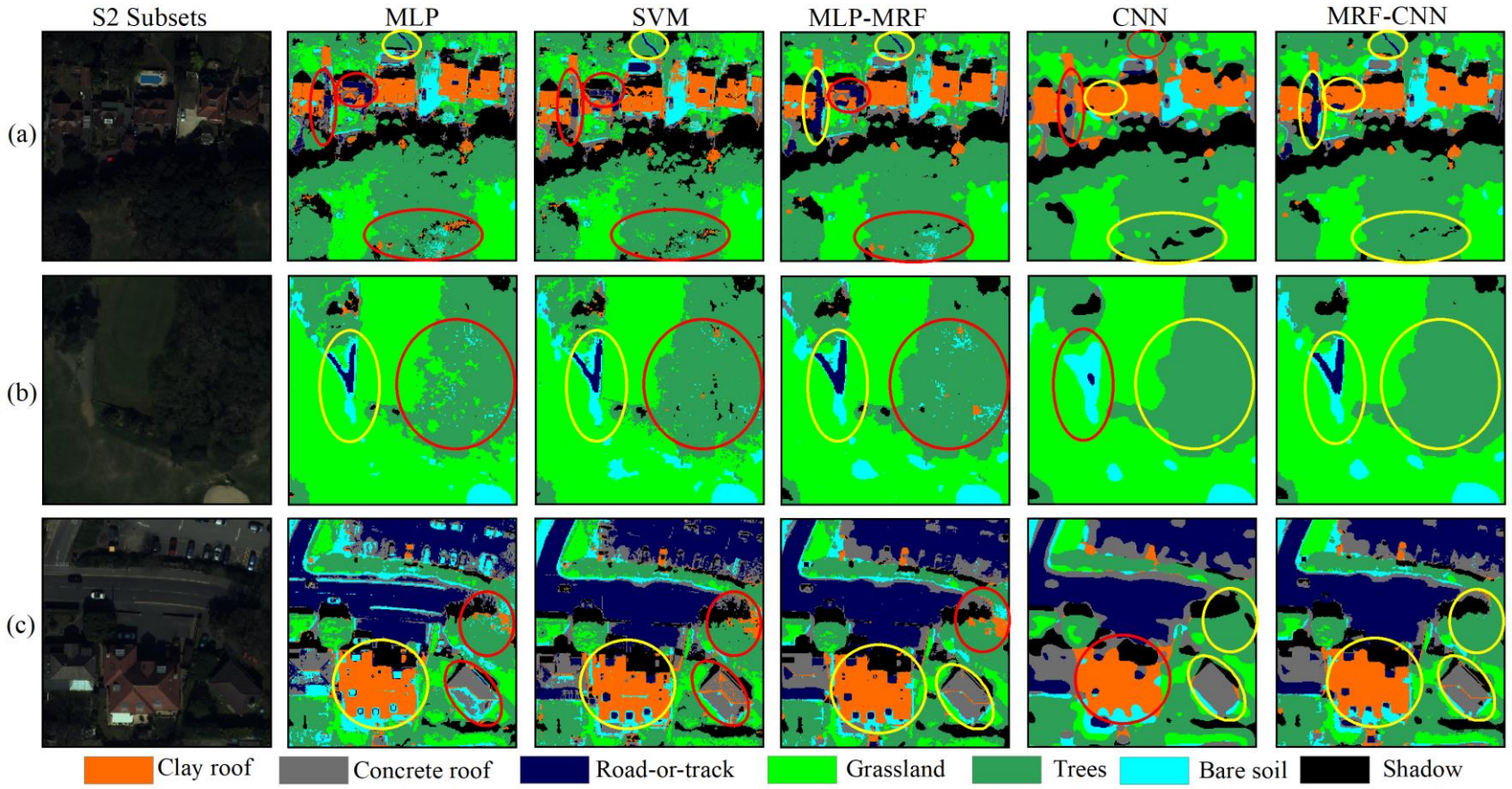


Figure 4-6: Three typical image subsets (a, b and c) in study site S2 with their classification results. Columns from left to right represent the original images (R G B bands), the MLP, the SVM, the MLP-MRF, the CNN, and the MRF-CNN classification results. The red and yellow circles denote incorrect and correct classification, respectively.

Experiment 2: The proposed MRF-CNN and its sub-modules (MLP, MLP-MRF and CNN) as well as other benchmark methods were validated on the Vaihingen and Potsdam semantic segmentation datasets. Table 4-5 and 4-6 present the classification accuracies of all four methods together with the four benchmark methods (SVM, FCN, SegNet and Deeplab-v2). The MRF-CNN achieved the largest OA of 88.4% and 89.4% for the two datasets, larger than its sub-modules (86.2% and 86.5%, 82.1% and 83.7%, and 81.4% and 82.1% OA of CNN, MLP-MRF and the MLP, respectively). The MRF-CNN also demonstrates greater accuracy than the benchmarks, including the Deeplab-v2 with an OA of 86.7% and 88.2%, the FCN with an OA of 85.9% and 86.2% (Kampffmeyer et al. 2016), the SegNet with an OA of 82.8% and 83.6% (Volpi and Tuia 2017), and the SVM with an OA of 81.7% and 82.4%.

The per-class mapping accuracy (Table 4-5 and 4-6) shows the effectiveness of the proposed MRF-CNN for the majority of classes. Significant increases in accuracy are realized for the classes of Impervious surfaces, Low vegetation, Building and Car relative to the individual classifier CNN and MLP-MRF, with an average large margin of 3.9%, 4%, 5.55% and 8.75%, respectively. The Tree class accuracy, however, was less significantly increased compared to the CNN, with small margins of 0.8% and 0.6%. In terms of benchmark methods, the MRF-CNN demonstrates higher accuracy for the majority of classes, except for the Car class (79.6% and 80.3%), for which the accuracy is less than for the state-of-the-art Deeplab-v2 (84.7% and 83.9%).

Figure 4-7 and 4-8 illustrate full tile predictions of Vaihingen dataset (No. 30) and Potsdam dataset (No. 05_12), with red and dashed circles highlighting broadly incorrect and correct classifications, respectively. Both MLP and SVM classifications result in salt-and-pepper effects due to pixel-level differentiation with subtle differences

between them (e.g. red circles shown in Figure 4-7(*d*) and 4-7(*e*)). The MLP-MRF (Figure 4-7(*f*) and 4-8(*f*)) improves on the MLP (Figure 4-7(*d*) and 4-8(*d*)) with homogeneous blocks and crisp boundary differentiation. This can be seen at the lower right side of the Building that has reduced salt-and-pepper effect (dashed circle in Figure 4-7(*d*) and 4-8(*d*)). The CNN acquires the greatest smoothness (Figure 4-7(*g*) and 4-8(*g*)) thanks to higher-level spatial feature representation. However, it makes some blunders by misclassifying Building as Car (red circles in Figure 4-7(*g*) or falsely producing some building edge artefacts as Impervious Surface (the red circle in Figure 4-8(*g*)). The MRF-CNN (Figure 4-7(*h*) and 4-8(*h*)), solved the aforementioned problems (all dashed circles) by taking advantage of the rough set uncertainty partition as well as the subsequent decision fusion.

Table 4-5 - Per-class accuracy and overall accuracy (OA) for the MLP, SVM, MLP-MRF, CNN and the proposed MRF-CNN approach, as well as baseline methods, for the Vaihingen dataset. the bold font highlights the largest classification accuracy per row.

Method	**Imp Surf**	**Building**	**Low Veg**	**Tree**	**Car**	**OA**
MLP	83.5%	82.1%	68.3%	86.1%	64.2%	81.4%
SVM	82.7%	82.4%	69.2%	84.3%	66.5%	81.7%
MLP-MRF	84.3%	83.6%	72.7%	83.9%	71.7%	82.1%
CNN	86.2%	89.2%	76.9%	86.9%	69.7%	86.2%
FCN	87.1%	91.8%	75.2%	86.1%	63.8%	85.9%
SegNet	82.7%	89.1%	66.3%	83.9%	55.7%	82.8%
Deeplab-v2	88.5%	93.3%	73.9%	86.9%	**84.7%**	86.7%
MRF-CNN	**89.7%**	**93.8%**	**80.1%**	**87.7%**	79.6%	**88.4%**

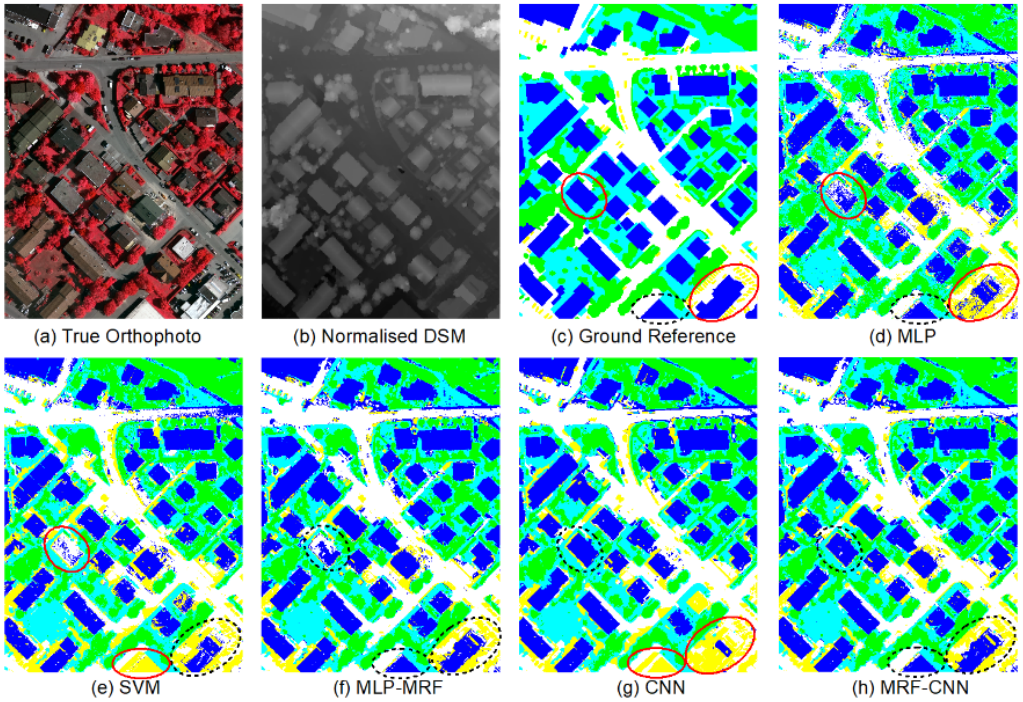


Figure 4-7: Full tile prediction for No. 30. Legend on the Vaihingen dataset: white=impervious surface; blue=buildings; cyan=low vegetation; green=trees; yellow=cars. (a) True Orthophoto; (b) Normalised DSM; (c) Ground Reference, ground reference labelling; (d, e, f, g) the inference result from MLP, SVM, MLP-MRF, CNN, respectively; (f) the proposed MRF-CNN classification result. The red and dashed circles denote incorrect and correct classification, respectively.

Table 4-6 - Per-class accuracy and overall accuracy (OA) for the MLP, SVM, MLP-MRF, CNN and the proposed MRF-CNN approach, as well as baseline methods, for the Potsdam dataset. the bold font highlights the largest classification accuracy per row.

Method	**Imp Surf**	**Building**	**Low Veg**	**Tree**	**Car**	**OA**
MLP	84.3%	81.3%	71.5%	85.6%	70.4%	82.1%
SVM	83.6%	81.8%	72.2%	84.3%	70.9%	82.4%
MLP-MRF	85.8%	83.6%	73.4%	84.8%	72.3%	83.7%
CNN	86.5%	88.7%	76.7%	87.6%	72.7%	86.5%
FCN	85.5%	90.6%	75.8%	86.1%	69.8%	86.2%
SegNet	82.9%	89.5%	73.1%	84.3%	70.5%	83.6%
Deeplab-v2	88.7%	93.6%	77.2%	86.5%	**83.9%**	88.2%
MRF-CNN	**90.8%**	**95.2%**	**81.5%**	**88.2%**	80.3%	**89.4%**

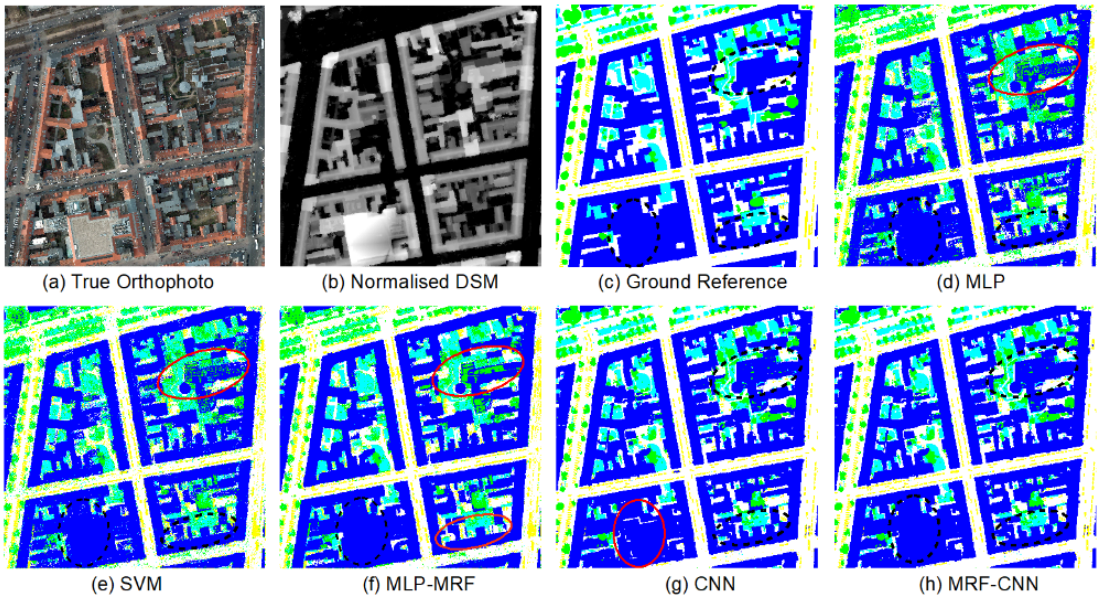


Figure 4-8: Full tile prediction for No. 05_12. Legend on the Potsdam dataset: white=impervious surface; blue=buildings; cyan=low vegetation; green=trees; yellow=cars. (a) True Orthophoto; (b) Normalised DSM; (c) Ground Reference, ground reference labelling; (d, e, f, g) the inference result from MLP, SVM, MLP-MRF, CNN, respectively; (f) the proposed MRF-CNN classification result. The red and dashed circles denote incorrect and correct classification, respectively.

4.3.5 Function of the VPRS fusion decision parameter β and *step*

The VPRS fusion decision parameters (β and *step*) were analysed separately to investigate each of their contributions in describing and integrating the classification results. As illustrated by Figure 4-9(*a*) and 4-9(*b*), relations between the fused classification accuracy and each of the parameters (while fixing the other) can be plotted. Generally, there are similar trends in terms of the influence of two parameters on classification accuracy: the accuracy increases initially until reaching the maximum accuracy at β = 0.1 and *step* around 0.075-0.1, and then decreases constantly, along with further increases of the inclusion error β (Figure 4-9(*a*)) and the atomic granule *step* (Figure 4-9(*b*)) respectively. This means that both β and *step* can impact the

accuracy. However, compared with the step, the change in accuracy caused by β is greater accompanied by greater accuracy variation, indicating that β is the crucial factor for VPRS parameter setting. It can be imagined that a large value of β can wrongly take the CNN's problematic boundary information as positive regions, whereas the "should-be" positive regions can be eliminated by too small a value of β. In terms of *step*, the smaller its value (i.e. a finer information granularity), the larger the test samples for the VPRS will be required, to provide enough samples within each information granularity level. An atomic granularity should, therefore, ideally match with the sampling density level; otherwise, it will reduce the classification accuracy (Figure 4-9(*b*)).

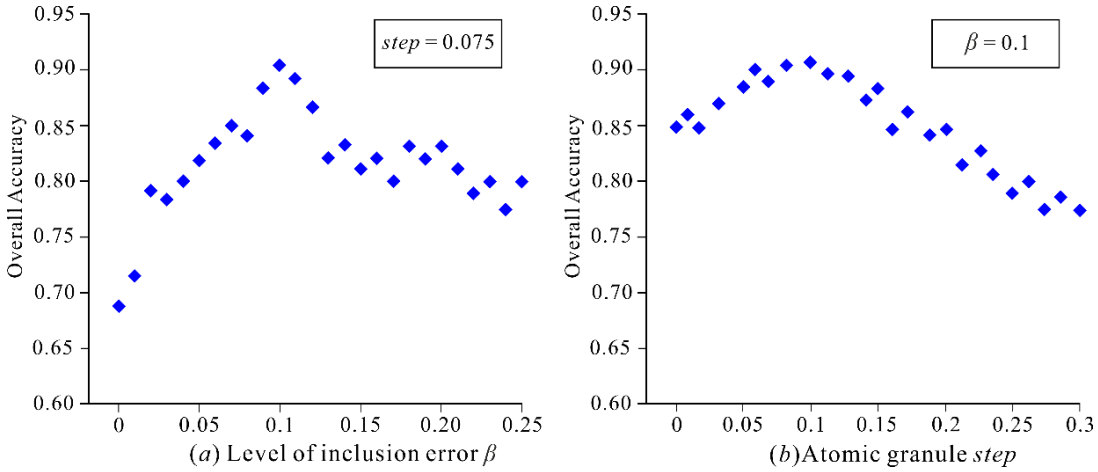


Figure 4-9: Accuracies of VPRS (a) influenced by β when fixing the step as 0.075, (b) influenced by step when fixing the β as 0.1.

4.4 Discussion

Due to the spatial and spectral complexity within VFSR imagery, any classification model prediction is inherently uncertain, including the advanced CNN classifier. Thus, for the integration of classifiers, it would be of paramount importance to discriminate the less uncertain and more uncertain results of each individual classification. A VPRS based regional fusion decision strategy was, thus, proposed to integrate the spectral-contextual-based MLP-MRF classifier with precise boundary partitions and the CNN

classifier with spatial feature representations for high accuracy classification of VFSR remotely sensed imagery. The proposed MRF-CNN regional decision fusion method takes advantage of the merits of the two individual classifiers and overcomes their respective shortcomings as discussed below.

4.4.1 Characteristics of MLP-MRF classification

The MLP-MRF classifier is constructed based on the pixel-based MLP as its conditional probability and models the prior probability using its contextual neighbourhood information to achieve a certain amount of smoothness (Wang and Liu 1999). That is, the MLP-MRF depends primarily on the spectral feature differentiation from the MLP with consideration of its spatial connectivity/smoothness (Wang et al. 2013). Such characteristics result in similar classification performance to the result of MLP but with less salt and pepper effect. One positive attribute of the MLP-MRF, inherited from the non-parametric learning classifier MLP, is the ability to maintain precise boundaries of some objects with high accuracy and fidelity. In particular, the classification accuracy of a pixel in the MLP model is not affected by the relative position (e.g. lying on or close to boundaries) of the object it belongs to, as long as the corresponding spectral space is separable. Some land cover classes (e.g. Clay roof, Metal roof and Shadow), with salient spectral properties that are spectrally exclusive to other classes, are therefore not only accurately classified with high classification accuracies (>90% overall accuracy), but also with less noise in comparison with the standard MLP and SVM classification results. At the same time, the MLP-MRF can elaborately identify some components of an object, for example, the Velux^{TM} windows of a building (shown by yellow circle in Figure 4-6(*c*)), indicating that the object and its sub-objects might be possibly mapped accurately in future. However, the classification accuracy increase of the MLP-MRF over the MLP is not substantial or

less remarkable, with just a 2-3% accuracy increase (see Table 4-3 in experiment 1 and Table 4-5 in experiment 2). In comparison with the CNN, the MLP-MRF usually demonstrates a much larger intra-class variation, which can be demonstrated by the fact that the boxplots of confidence values are larger when gradually trusting the MLP-MRF (Figure 4-4). This is mainly because the MLP-MRF utilizes the spectral information in the classification process without fully exploiting the abundant spatial information appearing in the VFSR imagery (e.g. texture, geometry or spatial arrangement) (Wang et al. 2016). Such deficiencies often lead to unsatisfactory classification performance in classes with spectrally mixed but spatially distinctive characteristics (e.g., the confusion and misclassification between Trees and Grassland or Low Vegetation that are spectrally similar, the severe salt and pepper effects on railway with linear textures, etc.).

4.4.2 Characteristics of CNN classification

Spatial features in remotely sensed data like VFSR imagery are intrinsically local and stationary that represent a coherent spatial pattern (Masi et al. 2016). The presence of such spatial features are detected by the convolutional filters within the CNN, and well generalized into increasingly abstract and robust features through hierarchical feature representations. Therefore, the CNN shows an impressive stability and effectiveness in VFSR image classification (Zhao and Du 2016). Especially, classes like Concrete roof and Road-or-track that are difficult to distinguish from their backgrounds with only spectral features at pixel level, are identified with relatively high accuracies. In addition, classes with heavy spectral confusion in both study sites (e.g. Trees and Grassland), are accurately differentiated due to their obvious spatial pattern differences; for example, the texture of tree canopies is generally rougher than that of grassland, which is captured by the CNN through spatial feature representations. Moreover, the convolutional filters

applied at each layer within the CNN framework remove all of the noise that is smaller than the size of the image patch, which leads to the smoothest classification results compared with the MLP, the SVM and the MLP-MRF (see Figure 4-5 - Figure 4-8). This is also demonstrated by Figure 4-4, where the boxplots of the CNN are much narrower than those of the MLP-MRF.

As discussed above, the CNN classifier demonstrates obvious superiority over the spectral-contextual based MLP-MRF (and the pixel-based MLP and SVM classifiers) for the classification of the spatially and spectrally complex VFSR remotely sensed imagery. However, according to the "no free lunch" theorem (Wolpert and Macready 1997), any elevated performance in one aspect of a problem will be paid for through others, and the CNN is no exception. the CNN also demonstrates some deficiencies for boundary partition and small feature identification, which is essential for VFSR image classification with unprecedented spatial detail. Such a weakness occurs mainly because of over-smoothness that leads to boundary uncertainties with small useful features being falsely erased, somehow similar to morphological or Gabor filter methods (Reis and Tasdemir 2011, Pingel et al. 2013). For example, the human-made objects in urban scenes like buildings and asphalt are often geometrically enlarged with distortion to some degree (See Figure 4-5(*b*) and 4-6(*c*)), and the impervious surfaces and the building are confused with cars being enlarged or misclassified (Figure 4-7(*e*)). As for natural objects in rural areas (S2), edges or porosities of a landscape patch are simplified or ignored, and even worse, linear features like river channels or dams that are of ecological importance, are erroneously erased (e.g. Figure 4-5(*b*)). Besides, certain spectrally distinctive features without obvious spatial patterns are poorly differentiated. For example, some Concrete roofs are wrongly identified as Asphalt as illustrated in Figure 4-5(*c*). Previous work also found that the CNN was inferior to some

global low level feature descriptors like Border/ Interior Pixel Classification when dealing with a remote sensing image that has abundant spectral but lacks spatial information (Nogueira et al. 2017). However, the uncertainties in the CNN classification demonstrate regional distribution characteristics, either along the object boundaries (e.g. Figure 4-5(*b*)) or entire objects (e.g. Figure 4-5(*c*)). These provide the justification of regional decision fusion to further improve the CNN for VFSR image classification.

4.4.3 The VPRS based MRF-CNN fusion decision

This chapter proposed to explore rough set theory for region-based uncertainty description and classification decision fusion using VFSR remotely sensed imagery. The classification uncertainties in the CNN results were quantified at a regional level, with each region determined as positive or non-positive (boundary and negative) regions by matching the correctness of a group of samples in the Test Sample *T*2. Nevertheless, in the standard rough set, most of the actual positive regions are occupied by boundary (i.e. non-positive) regions due to the huge uncertainty and inconsistency in VFSR image classification results. Such issues limit the practical application of the standard rough set because of its ignorance of the desired positive regions. A variable precision rough set (VPRS) is proposed for uncertainty description and classification integration by incorporating a small level of inclusion error (i.e. parameter β). The VPRS theory is used here as a spatially explicit framework for regional decision fusion, where the non-positive regions in this research represent the spatial uncertainties in the CNN classification result. For those positive regions of CNN classifications, including the very close to 100% correct classifications, are identified and utilized; whereas the rest (i.e. the non-positive) regions are replaced by the MLP-MRF results with crisp and accurate boundary delineation.

To integrate the CNN and the MLP-MRF classifier, the CNN was served as the base classifier to derive the classification confidence, considering its superiority in terms of classification accuracy and the regional homogeneity of classification results. Therefore, the regional decision fusion process is based on the CNN classification results, and the MLP-MRF is only trusted at the regions where the CNN is less believable (i.e. the non-positive regions). Such a fusion decision strategy achieves an accurate and stable result with the least variation in accuracy, as illustrated by the narrow box in Figure 4-4. The complete correctness of the MLP-MRF results at the non-positive regions are not guaranteed, but one thing is certain: the corresponding MLP-MRF results are much more accurate than those of the CNN. In fact, while the CNN accurately classifies the interiors of objects with spatial feature representations, the MLP-MRF could provide a smooth, but also crisp boundary segmentation with high fidelity (Wang et al. 2013). These supplementary characteristics inherent in the MLP-MRF and CNN, are captured well by the proposed VPRS-based MRF-CNN regional decision fusion approach. As shown by Figure 4-4, although the values of the CNN confidence map decrease gradually from the centre to its boundary (i.e. the edge between the positive and non-positive regions, at 0.3 marked by the red vertical line), the classification accuracies rise constantly until reaching the maximum accuracy. For these MLP-MRF results in the non-positive regions, the corresponding non-positive regions (i.e. the problematic areas of the final fusion decision results) can be further clarified. Moreover, additional improvement might be obtained by means of imposing extra expert knowledge and/or combining other advanced classifiers (e.g. SVM, Random Forest, etc.).

In summary, the proposed method for classification data description and integration is, in fact, a general framework extensively applicable to any classification algorithms (not

just for the mentioned individual classifiers), and to any remote sensing images (not just for the VFSR remotely sensed imagery). The general approach, thus, addresses the complex problem of remote sensing image classification in a flexible, automatic and active manner.

The proposed MRF-CNN relies on an efficient and relatively limited CNN network with just four layers (c.f. state-of-the-art networks, such as Deeplab-v2, built on extremely deep ResNet-101). Nevertheless, it still achieves comparable and promising classification performance with the largest accuracy overall. This demonstrates that the proposed method has practical utility, especially when facing the problems of limited computational power with insufficient training data, which are commonly encountered in the remote sensing domain when building a deep CNN network.

4.5 Conclusion

Spatial uncertainty is always a key concern in remote sensing image classification, which is essential when facing the spatially and spectrally complex VFSR remotely sensed imagery. Characterising the spatial distribution of uncertainties has great potential for practical application of the data. In this chapter, a novel variable precision rough set (VPRS) based regional fusion decision between CNN and MRF was presented for the classification of VFSR remotely sensed imagery. The VPRS model quantified the uncertainties in CNN classification of VFSR imagery by partitioning the result into spatially explicit granularities that represent positive regions (correct classifications) and non-positive regions (uncertain or incorrect classifications). Such a region-based fusion decision approach reflects the regional homogeneity of the CNN classification map. The positive regions were directly trusted by the CNN, whereas non-positive regions were rectified by the MLP-MRF in consideration of their

complementary behaviour in spatial representation. The proposed regional fusion of MRF-CNN classifiers consistently outperformed the standard pixel-based MLP and SVM, spectral-contextual based MLP-MRF as well as contextual-based CNN classifiers, and increased classification accuracy above state-of-the-art methods when applied to the ISPRS Semantic Labelling datasets. Therefore, this VPRS-based regional classification integration of CNN and MRF classification results provides a framework to achieve fully automatic and effective VFSR image classification.

Chapter 5 An object-based convolutional neural network (OCNN) for urban land use classification [3]

[3] This chapter is based on the published paper: Ce Zhang, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, Peter M. Atkinson, 2018c, An object-based convolutional neural networks (OCNN) for urban land use classification. *Remote Sensing of Environment*, 216:57-70.

Abstract

Urban land use information is essential for a variety of urban-related applications such as urban planning and regional administration. The extraction of urban land use from very fine spatial resolution (VFSR) remotely sensed imagery has, therefore, drawn much attention in the remote sensing community. Nevertheless, classifying urban land use from VFSR images remains a challenging task, due to the extreme difficulties in differentiating complex spatial patterns to derive high-level semantic labels. Deep convolutional neural networks (CNNs) offer great potential to extract high-level spatial features, thanks to its hierarchical nature with multiple levels of abstraction. However, blurred object boundaries and geometric distortion, as well as huge computational redundancy, severely restrict the potential application of CNN for the classification of urban land use. In this chapter, a novel object-based convolutional neural network (OCNN) is proposed for urban land use classification using VFSR images. Rather than pixel-wise convolutional processes, the OCNN relies on segmented objects as its functional units, and CNN networks are used to analyse and label objects such as to partition within-object and between-object variation. Two CNN networks with different model structures and window sizes are developed to predict linearly shaped objects (e.g. Highway, Canal) and general (other non-linearly shaped) objects. Then a class-specific decision fusion is performed to integrate the classification results. The effectiveness of the proposed OCNN method was tested on aerial photography of two large urban scenes in Southampton and Manchester in Great Britain. The OCNN combined with large and small window sizes achieved excellent classification accuracy and computational efficiency, consistently outperforming its sub-modules, as well as other benchmark comparators, including the pixel-wise CNN, contextual-based MRF and object-based

OBIA-SVM methods. The proposed method provides the first object-based CNN framework to effectively and efficiently address the complicated problem of urban land use classification from VFSR images.

Keywords: convolutional neural network; OBIA; urban land use classification; VFSR remotely sensed imagery; high-level feature representations

5.1 Introduction

Urban land use information, reflecting socio-economic functions or activities, is essential for urban planning and management. It also provides a key input to urban and transportation models, and is essential to understanding the complex interactions between human activities and environmental change (Patino and Duque 2013). With the rapid development of modern remote sensing technologies, a huge amount of very fine spatial resolution (VFSR) remotely sensed imagery is now commercially available, opening new opportunities to extract urban land use information at a very detailed level (Pesaresi et al. 2013). However, urban land features captured by these VFSR images are highly complex and heterogeneous, comprising the juxtaposition of a mixture of anthropogenic urban and semi-natural surfaces. Often, the same urban land use types (e.g. residential areas) are characterized by distinctive physical properties or land cover materials (e.g. composed of different roof tiles), and different land use categories may exhibit the same or similar reflectance spectra and textures (e.g. asphalt roads and parking lots) (Pan et al. 2013). Meanwhile, information on urban land use within VFSR imagery is presented implicitly as patterns or high-level semantic functions, in which some identical low-level ground features or object classes are frequently shared amongst different land use categories. This complexity and diversity of spatial and structural patterns in urban areas makes its classification into land use classes a

challenging task (Hu et al. 2015a). Therefore, it is important to develop robust and accurate urban land use classification techniques by effectively representing the spatial patterns or structures lying in VFSR remotely sensed data.

Over the past few decades, tremendous effort has been made in developing automatic urban land use classification methods. These methods can be categorized broadly into four classes based on the spatial unit of representation (i.e. pixels, moving windows, objects and scenes) (Liu et al. 2016). The pixel-level approaches that rely purely upon spectral characteristics are able to classify land cover, but are insufficient to distinguish land uses that are typically composed of multiple land covers, and such problems are particularly significant in urban settings (Zhao et al. 2016). Spatial information, that is, texture (Myint 2001, Herold et al. 2003) or context (Wu et al. 2009), was incorporated to analyse urban land use patterns through moving kernel windows (Niemeyer et al. 2014). However, it could be argued that both pixel-based and moving window-based methods require to predefine arbitrary image structures, whereas actual objects and regions might be irregularly shaped in the real world (Herold et al. 2003). Therefore, object-based image analysis (OBIA) that is built upon automatically segmented objects from remotely sensed imagery is preferable (Blaschke 2010), and has been considered as the dominant paradigm over the last decade (Blaschke et al. 2014). Those image objects, as the base units of OBIA, offer two kinds of information with a spatial partition, specifically; within-object information (e.g. spectral, texture, shape) and between-object information (e.g. connectivity, contiguity, distances, and direction amongst adjacent objects). Many studies applied OBIA for urban land use classification using within-object information with a set of low-level features (such as spectra, texture, shape) of the ground features (e.g. Blaschke 2010, Blaschke et al. 2014, Hu and Wang, 2013). These OBIA approaches, 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 attempted to incorporate between-object information by aggregating objects using spatial contextual descriptive indicators on well-defined land use units, such as cadastral fields or street blocks. Those descriptive indicators were 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). However, the ancillary geographic data for specifying the land use units might not be available for some regions, and the spatial contexts are often hard to describe and characterise as a set of "rules", even though the complex structures or patterns might be recognizable and distinguishable by human experts (Oliva-Santos et al. 2014). Thus, advanced data-driven approaches are highly desirable to learn land use semantics automatically through high-level feature representations.

Recently, deep learning has become the new hot topic in machine learning and pattern recognition, where the most representative and discriminative features are learnt end-to-end, hierarchically (Chen et al. 2016). This breakthrough was triggered by a revival of interest in the use of multi-layer neural networks to model higher-level feature representations without human-designed features or rules. Convolutional neural networks (CNNs), as a well-established and popular deep learning method, has produced state-of-the-art results for multiple domains, such as visual recognition (Krizhevsky et al. 2012), image retrieval (Yang et al. 2015) and scene annotation (Othman et al. 2016). Owing to its superiority in higher-level feature representation and scene understanding, the CNN has demonstrated great potential in many remote sensing tasks such as vehicle detection (Chen et al. 2014, Dong et al. 2015), road network extraction (Cheng, Wang, et al. 2017), remotely sensed scene classification (Othman et

al. 2016, Sargent et al. 2017), and semantic segmentation (Zhao et al. 2017). Interested readers are referred to a comprehensive review of deep learning in remote sensing (Zhu et al. 2017).

Land use information extraction from remotely sensed data using CNN models has been undertaken in the form of land-use scene classification, which aims to assign a semantic label (e.g. tennis court, parking lot, etc.) to an image according to its content (Chen et al. 2016, Nogueira et al. 2017). There are broadly two strategies to exploit the CNN models for scene-level land use classification, namely; *i*) pre-trained or fine-tuned CNN, and *ii*) fully-trained CNN from scratch. The first strategy relies on pre-trained CNN networks transferred from an auxiliary domain with natural images, which has been demonstrated empirically to be useful for land-use scene classification (Hu et al. 2015b, Nogueira et al. 2017). However, it requires three input channels derived from natural images with RGB only, whereas the multispectral remotely sensed imagery often involves the near infrared band, and such a distinction restricts the utility of pre-trained CNN networks. Alternatively, the (ii) fully-trained CNN strategy gives full control over the network architecture and parameters, which brings greater flexibility and expandability (Chen et al. 2016). Previous researchers have explored the feasibility of the fully-trained strategy in building CNN models for scene level land-use classification. For example, Luus et al. (2015) proposed a multi-view CNN with multi-scale input strategies to address the issue of land use scene classification and its scale-dependent characteristics. Othman et al. (2016) used convolutional features and a sparse auto-encoder for scene-level land-use image classification, which further demonstrated the superiority of CNNs in feature learning and representation. Xia et al. (2017) even constructed a large-scale aerial scene classification dataset (AID) for performance evaluation among various CNN models and architectures developed by both strategies.

However, the goal of these land use scene classifications is essentially *image* categorization, where a small patch extracted from the original remote sensing image is labelled into a semantic category, such as 'airport', 'residential' or 'commercial' (Maggiori et al. 2017). Land-use scene classification, therefore, does not meet the actual requirement of remotely sensed land use image classification, which requires all pixels in an entire image to be identified and labelled into land use categories (i.e., producing a thematic map).

With the intrinsic advantages of hierarchical feature representation, the patch-based CNN models provide great potential to extract higher-level land use semantic information. However, this patch-wise procedure 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 in the classification. In addition, to obtain a full resolution classification map, pixel-wise densely overlapped patches were used at the model inference phase, which inevitably led to extremely redundant computation. As an alternative, Fully Convolutional Networks (FCN) and its extensions have been introduced into remotely sensed sematic segmentation to address the pixel-level classification problem (e.g. Liu et al. 2017; Paisitkriangkrai et al. 2016; Volpi and Tuia, 2017). These FCN-based methods are, however, mostly developed to solve low-level semantic (i.e. land cover) classification tasks, due to the insufficient spatial information in the inference phase and the lack of contextual information at up-sampling layers (Liu et al. 2017). In short, we argue that the existing CNN models, including both patch-based and pixel-level approaches, are not well designed in terms of accuracy and/or computational efficiency to cope with the complicated problem of urban land use classification using VFSR remotely sensed imagery.

In this chapter, we propose an innovative object-based CNN (OCNN) method to address the complex urban land-use classification task using VFSR imagery. Specifically, object-based segmentation was initially employed to characterize the urban landscape into functional units, which consist of two geometrically different objects, namely linearly shaped objects (e.g. Highway, Railway, Canal) and other (non-linearly shaped) general objects. Two CNNs with different model structures and window sizes were applied to analyse and label these two kinds of objects, and a rule-based decision fusion was undertaken to integrate the models for urban land use classification. The innovations of this research can be summarised as 1) to develop and exploit the role of CNNs under the framework of OBIA, where both within-object information and between-object information is used jointly to fully characterise objects and their spatial context. 2) to design the CNN networks and position them appropriately with respect to object size and geometry, and integrate the models in a class-specific manner to obtain an effective and efficient urban land use classification output (i.e., a thematic map). The effectiveness and the computational efficiency of the proposed method were tested on two complex urban scenes in Great Britain.

The remainder of this chapter is organized as follows: Section 5.2 introduces the general workflow and the key components of the proposed methods. Section 5.3 describes the study area and data sources. The results are presented in section 5.4, followed by a discussion in section 5.5. The conclusions are drawn in the last section.

5.2 Methodology

5.2.1 Convolutional Neural Networks (CNN)

A Convolutional Neural Network (CNN) is a multi-layer feed-forward neural network that is designed specifically to process large scale images or sensory data in the form

of multiple arrays by considering local and global stationary properties (LeCun et al. 2015). The main building block of a CNN is typically composed of multiple layers interconnected to each other through a set of learnable weights and biases (Romero et al. 2016). Each of the layers is fed by small patches of the image that scan across the entire image to capture different characteristics of features at local and global scales. Those image patches are generalized through alternative convolutional and pooling/subsampling layers within the CNN framework, until the high-level features are obtained on which a fully connected classification is performed (Schmidhuber 2015). Additionally, several feature maps may exist in each convolutional layer and the weights of the convolutional nodes in the same map are shared. This setting enables the network to learn different features while keeping the number of parameters tractable. Moreover, a nonlinear activation (e.g. sigmoid, hyperbolic tangent, rectified linear units) function is taken outside the convolutional layer to strengthen the non-linearity (Strigl et al. 2010). Specifically, the major operations performed in the CNN can be summarized as:

$$O^l = pool_p(\sigma(O^{l-1} * W^l + b^l)) \tag{5-1}$$

where the O^{l-1} denotes the input feature map to the lth layer, the W^l and the b^l represent the weights and biases of the layer, respectively, that convolve the input feature map through linear convolution*, and the $\sigma(\cdot)$ indicates the non-linearity function outside the convolutional layer. These are often followed by a max-pooling operation with $p \times p$ window size ($pool_p$) to aggregate the statistics of the features within specific regions, which forms the output feature map O^l at the lth layer (Romero et al. 2016).

5.2.2 Object-based CNN (OCNN)

An object-based CNN (OCNN) is proposed for the urban land use classification using VFSR remotely sensed imagery. The OCNN is trained as the standard CNN models with labelled image patches, whereas the model prediction is to label each segmented object derived from image segmentation. The segmented objects are generally composed of two distinctive objects in geometry, including linearly shaped objects (LS-objects) (e.g. Highway, Railway and Canal) and other (non-linearly shaped) general objects (G-objects). To accurately predict the land use membership association of a G-object, a large spatial context (i.e. a large image patch) is required when using the CNN model. Such a large image patch, however, often may lead to a large uncertainty in the prediction of LS-objects due to narrow linear features being ignored throughout the convolutional process. Thus, a large input window CNN (LIW-CNN) and a range of small input window CNNs (SIW-CNN) were thereafter trained to predict the G-object and the LS-object, respectively, where the appropriate convolutional positions of both models were derived from a novel object convolutional position analysis (OCPA). The final classification results were determined by the decision fusion of the LIW-CNN and the SIW-CNN. As illustrated by Figure 5-1, the general workflow of the proposed OCNN consists of five major steps, including (*A*) image segmentation, (*B*) OCPA, (*C*) LIW-CNN and SIW-CNN model training, (*D*) LIW-CNN and SIW-CNN model inference, and (*E*) Decision fusion of LIW-CNN and SIW-CNN. Each of these steps is elaborated in the following section.

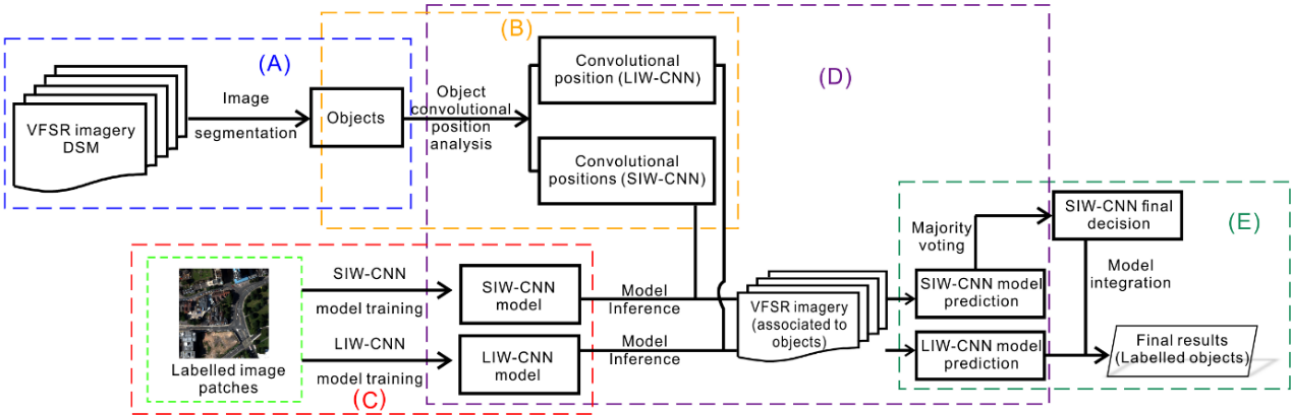


Figure 5-1: Flowchart of the proposed object-based CNN (OCNN) method with five major steps: (A) image segmentation, (B) object convolutional position analysis (OCPA), (C) LIW-CNN and SIW-CNN model training, (D) LIW-CNN and SIW-CNN model inference, and (E) fusion decision of LIW-CNN and SIW-CNN.

5.2.2.1 Image segmentation

The proposed method starts with an initial image segmentation to achieve an object-based image representation. Mean-shift segmentation (Comaniciu and Meer 2002), as a nonparametric clustering approach, was used to partition the image into objects with homogeneous spectral and spatial information. Four multispectral bands (Red, Green, Blue, and Near Infrared) together with a digital surface model (DSM), useful for differentiating urban objects with height information (Niemeyer et al. 2014), were incorporated as multiple input data sources for the image segmentation (Figure 5-1(*A*)). A slight over-segmentation rather than under-segmentation was produced to highlight the importance of spectral similarity, and all the image objects were transformed into GIS vector polygons with distinctive geometric shapes.

5.2.2.2 Object convolutional position analysis (OCPA)

The object convolutional position analysis (OCPA) is employed based on the **moment bounding (MB) box** of each object to identify the position of LIW-CNN and those of SIW-CNNs. The MB box, proposed by Zhang and Atkinson (2016), refers to the

minimum bounding rectangle built upon the moment orientation (the orientation of the major axis) of a polygon (i.e. an object), derived from planar characteristics defined by mechanics (Zhang et al. 2006, Zhang and Atkinson 2016). The MB box theory is briefly described hereafter.

Suppose that (*x*, *y*) is a point within a planar polygon (*S*) (Figure 5-2), whose centroid is $C(\bar{x}, \bar{y})$. The moment of inertia about the x-axis (I_{xx}) and y-axis (I_{yy}), and the product of inertia (I_{xy}) are expressed by equations (5-2), (5-3) and (5-4), respectively.

$$I_{xx} = \int y^2 dA \tag{5-2}$$

$$I_{yy} = \int x^2 dA \tag{5-3}$$

$$I_{xy} = \int xy dA \tag{5-4}$$

Note, dA(= $dx \cdot dy$) refers to the differential area of point (*x*, *y*) (Timoshenko and Gere 1972).

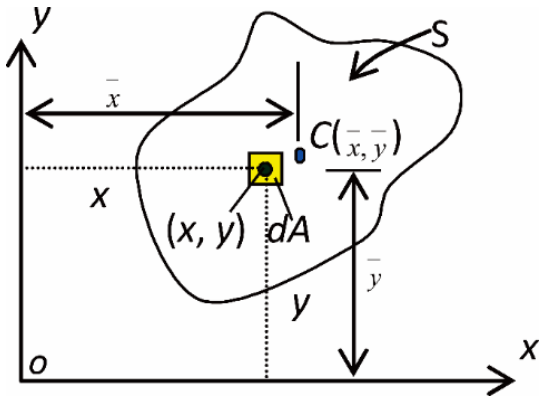


Figure 5-2: A patch (S) with centroid C ($\bar{x}, \bar{y}$), dA is the differential area of point (x, y), Oxy is the geographic coordinate system.

As illustrated by Figure 5-3, two orthogonal axes (*MN* and *PQ*), the major and minor axes, pass through the centroid (*C*), with the minimum and maximum moment of inertia about the major and minor axes, respectively. The moment orientation θ_{MB} (i.e. the orientation of the major axis) is calculated by equations (5-5) and (5-6) (Gere and Timoshenko 1972).

$$\tan 2\theta_{MB} = \frac{2I_{xy}}{I_{yy} - I_{xx}} \tag{5-5}$$

$$\theta_{MB} = \frac{1}{2}\tan^{-1}\left(\frac{2I_{xy}}{I_{yy} - I_{xx}}\right) \tag{5-6}$$

The moment bounding (MB) box (the rectangle in red shown in Figure 5-3) that minimally encloses the polygon, *S*, is then constructed by taking θ_{MB} as the orientation of the long side of the box, and *EF* is the perpendicular bisector of the MB box with respect to its long side.

The discrete forms of equations (5-2) - (5-6) suitable for patch computation, are further deduced by associating the value of a line integral to that of a double integral using Green's theorem (see Zhang et al. (2006) for theoretical details).

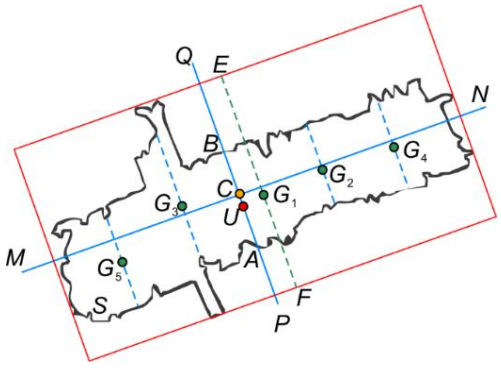


Figure 5-3: Moment bounding (MB) box and the CNN convolutional positions of a polygon S.

The **CNN convolutional positions** are determined by the minor axis (*PQ*) and the bisector of the MB box (*EF*) to approximate the central region of the polygon (*S*). For the LIW-CNN, the central point (the red point *U*) of the line segment (*AB*) intersected by *PQ* and polygon *S* is assigned as the convolutional position. As for the SIW-CNN, a distance parameter (*d*) (a user defined constant) is used to determine the number of SIW-CNN sampled along the polygon. Given the length of a MB box as *l*, the number (*n*) of SIW-CNNs is derived as:

$$n = \frac{l - d}{d} \tag{5-7}$$

The convolutional positions of the SIW-CNN are assigned to the intersection between the centre of the bisector (*EF*) as well as its parallel lines and the polygon *S*. The points (G_1, G_2, …, G_5) in Figure 5-3 illustrate the convolutional positions of SIW-CNN for the case of $n = 5$.

5.2.2.3 LIW-CNN and SIW-CNN model training

Both the LIW-CNN and SIW-CNN models are trained using image patches with labels as input feature maps. The parameters and model structures of these two models are empirically tuned as demonstrated in the Experimental Results and Analysis sections. Those trained CNN models are used for model inference in the next stage.

5.2.2.4 LIW-CNN and SIW-CNN model inference

After the above steps, the trained LIW-CNN and SIW-CNN models, and the convolutional position of LIW-CNN and those of SIW-CNN for each object are available. For a specific object, its land use category can be predicted by the LIW-CNN at the derived convolutional position within the VFSR imagery; at the same time, the predictions on the land use membership associations of the object can also be obtained

by employing SIW-CNN models at the corresponding convolutional positions. Thus each object is predicted by both LIW-CNN and SIW-CNN models.

5.2.2.5 Fusion decision of LIW-CNN and SIW-CNN

Given an object, the two LIW-CNN and SIW-CNN model predictions might be inconsistent between each other, and the distinction might also occur within those of the SIW-CNN models. Therefore, a simple majority voting strategy is applied to achieve the final decision of the SIW-CNN model. A fusion decision between the LIW-CNN and the SIW-CNN is then conducted to give priority to the SIW-CNN model for LS-objects, such as roads, railways etc.; otherwise, the prediction of the LIW-CNN is chosen as the final result.

5.2.3 Accuracy assessment

Both pixel-based and object-based methods were adopted to comprehensively test the classification performance using the testing sample set through five-fold cross validation. The pixel-based approach was assessed based on the overall accuracy and Kappa coefficient as well as per-class mapping accuracy computed from a confusion matrix. The object-based assessment was based on geometry (Clinton et al. 2010, Li et al. 2015, Radoux and Bogaert 2017). Specifically, suppose that a classified object M_i overlaps a set of reference objects O_{ij}, where $j = 1, 2, \cdots r$, r refers to the total number of reference objects overlapped by M_i. For each pair of objects (M_i, O_{ij}), a weight parameter deduced by the ratio between the area of a reference object (area (O_{ij})) and the total area of reference objects $\sum_{j=1}^{r}\text{area}(O_{ij})$ was introduced to calculate over-classification $OC(M_i)$ and under-classification $UC(M_i)$ error indices as:

$$OC(M_i)=\sum_{i=1}^{r}(w\cdot(1-\frac{\text{area}(M_i\bigcap O_{ij})}{\text{area}(O_{ij})})),\ w=\frac{\text{area}(O_{ij})}{\sum_{j=1}^{r}\text{area}(O_{ij})} \tag{5-8}$$

$$UC(M_i) = 1 - \frac{\sum_{j=1}^{r} \text{area}(M_i \cap O_{ij})}{\text{area}(M_i)} \tag{5-9}$$

The total classification error (*TCE*) of M_i is designed to integrate the over-classification and under-classification error as:

$$TCE(M_i) = \sqrt{\frac{OC(M_i)^2 + UC(M_i)^2}{2}} \tag{5-10}$$

All three indices (i.e. *OC*, *UC*, and *TCE*) represent the average of all the classified objects for each land use category in the classification map to formulate the final validation results.

5.3 Experimental Results and Analysis

5.3.1 Study area and data sources

In this research, two UK cities, Southampton (S1) and Manchester (S2), lying on the Southern coast and in North West England, respectively, were chosen as our case study sites (Figure 5-4). Both of the study areas are highly heterogeneous and distinctive from each other in land use characteristics, and are thereby suitable for testing the generalization capability of the proposed land use classification algorithm.

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 5802×4850 pixels for S1 and 5875×4500 pixels for S2, respectively. Land use categories of the study areas were defined according to the official land use classification system provided by the UK government Department for

Communities and Local Government (DCLG). Detailed descriptions of each land use class and its corresponding sub-classes in S1 and S2 are listed in Tables 5-1 and 5-2, respectively. 10 dominant land use classes were identified within S1, including *high-density residential*, *commercial*, *industrial*, *medium-density residential*, *highway*, *railway*, *park and recreational area*, *parking lot*, *redeveloped area*, and *harbour and sea water*. In S2, nine land use categories were found, including *residential*, *commercial*, *industrial*, *highway*, *railway*, *park and recreational area*, *parking lot*, *redeveloped area*, and *canal*.

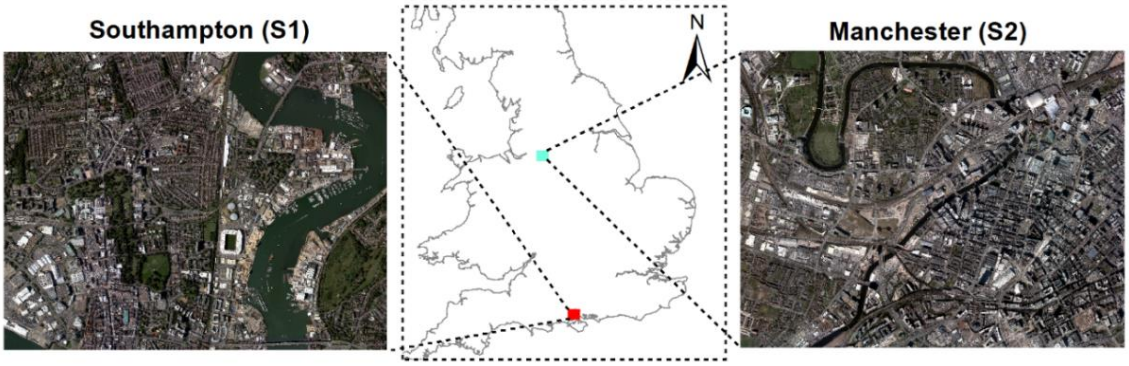


Figure 5-4: The two study areas of urban scenes: S1 (Southampton) and S2 (Manchester).

Table 5-1 - The land use classes in S1 (Southampton) and the corresponding sub-class components.

Land Use Class	Train	Test	Sub-class Components
High-density residential	1026	684	Residential houses, terraces, a small coverage of green space
Medium-density residential	984	656	Residential flats with a large green space and parking lots
Commercial	972	648	Commercial services with complex buildings, and parking lots
Industrial	986	657	Marine transportation, car factories
Highway	1054	703	Asphalt road, lane, cars
Railway	1008	672	Rail tracks, gravel, sometimes covered by trains
Parking lot	982	655	Asphalt road, parking line, cars
Park and recreational area	996	664	A large coverage of green space and vegetation, bare soil, lake
Redeveloped area	1024	683	Bare soil, scattered vegetation, reconstructions
Harbour and sea water	1048	698	Sea shore, ship, sea water

Table 5-2 - The land use classes in S2 (Manchester) and the corresponding sub-class components.

Land Use Class	Train	Test	Sub-class Components
Residential	1009	673	Residential buildings, a small coverage of green space
Commercial	1028	685	Shopping centre and commercial services with parking lots
Industrial	1004	669	Digital services, science and technology, gas industry
Highway	997	665	Asphalt road, lane, cars
Railway	1024	683	Rail tracks, gravel, sometimes covered by trains
Parking lot	1015	677	Asphalt road, parking line, cars
Park and recreational area	993	662	A large coverage of green space, bare soil, lake
Redeveloped area	1032	688	Bare soil, scattered vegetation, reconstructions
Canal	994	662	Canal water

Figure 5-5: Representative exemplars (image patches) of each land use category at the two study sites (S1 and S2).

In addition to the above-mentioned aerial photographs, Digital Surface Models (DSM) of the study sites with 50 cm spatial resolution were incorporated into the process of image segmentation. Moreover, other data sources, including Google Maps, Microsoft Bing Maps, and the MasterMap Topographic Layer (a highly detailed vector map from Ordnance Survey) (Regnauld and Mackaness 2006), were fully consulted and cross-referenced to gain a comprehensive appreciation of the land cover and land use within the study sites.

Sample points were collected using a stratified random scheme from ground data provided by local surveyors and photogrammetrists, and split into 60% training samples and 40% testing samples for each class. The training sample size was guaranteed above an average of 1,000 per class, which is sufficient for CNN networks, as recommended by Chen et al. (2016). In S1, a total of 10,080 training samples and 6,720 testing samples were obtained, and each category's sample size together with its sub-class components are listed in Table 5-1. In S2, 9,096 training samples and 6,064 testing samples were acquired (see Table 5-2 for the detailed sample size per class and the corresponding sub-classes). Figure 5-5 demonstrates typical examples of the land use categories: note that they are highly heterogeneous and spectrally overlapping. Field survey was conducted throughout the study areas in July 2016 to further check the validity and precision of the selected samples.

5.3.2 Model structure and parameter settings

The proposed method was implemented based on vector objects extracted by means of image segmentation. The objects were further classified through object-based CNN networks (OCNN). Detailed parameters and model structures optimised by S1 and directly generalised in S2 were clarified as follows.

5.3.2.1 Segmentation parameter settings

The initial mean-shift segmentation algorithm was implemented using the Orfeo Toolbox open-source software. Two spatial and spectral bandwidth parameters, namely the spatial radius and the range (spectral) radius, were optimized as 15.5 and 20 through cross-validation coupled with a small amount of trial-and-error. In addition, the minimum region size (the scale parameter) was chosen as 80 to produce a small amount of over-segmentation and, thereby, mitigate salt and pepper effects simultaneously.

Figure 5-6: An illustration of object convolutional position analysis with the moment box (yellow rectangle), the convolutional centre point (green star), and the convolutional input window (green rectangle), as well as the highlighted image object (in cyan). All the other segmented objects are demonstrated as red polygons. (A) demonstrates the large input window for a general object, and (B), (C) illustrate the small input windows for linearly shaped objects (highway and railway, respectively, in these exemplars).

5.3.2.2 LIW-CNN and SIW-CNN model structures and parameters

Within the two study sites, the highway, railway in S1 and the highway, railway, and canal in S2 belong to linearly shaped objects (LS-objects) in consideration of the elongated geometric characteristics (e.g. Figure 5-6(*B*), (*C*)), while all the other objects

belong to general objects (G-objects) (e.g. Figure 5-6(*A*)). The LIW-CNN with a large input window (Figure 5-6(*A*)), and SIW-CNNs with small input windows (Figure 5-6(*B*), (*C*)) that are suitable for the prediction of G-objects and LS-objects, respectively, were designed here. Note, the other type of CNN models employed on each object, namely, the SIW-CNNs in Figure 5-6(*A*) and the LIW-CNN in both Figure 5-6(*B*) and 5-6(*C*) were not presented in the figure to gain a better visual effect. The model structures and parameters of LIW-CNN and SIW-CNN are illustrated by Figure 5-7(*a*) and 5-7(*b*) and are detailed hereafter.

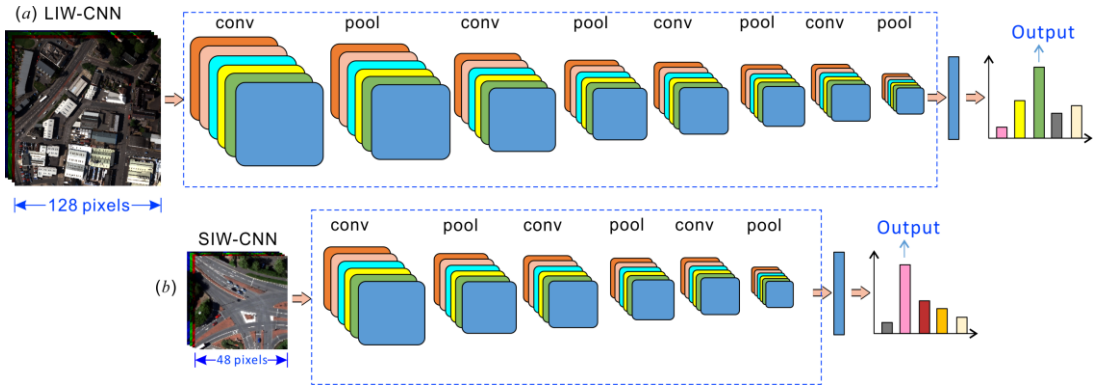


Figure 5-7: The model architectures and structures of the large input window CNN (LIW-CNN) with 128×128 input window size and eight-layer depth and small input window CNN (SIW-CNN) with 48×48 input window size and six-layer depth.

The SIW-CNN (Figure 5-7(*b*)) with a small input window size (48×48) and six-layer depth is a simplified structure with similar parameters to the LIW-CNN network, except for the number of convolutional filters at each layer, which was reduced to 32 in order to avoid over-fitting the model. The input window size was cross-validated on linear objects with a range of small window sizes, including {24×24, 32×32, 40×40, 48×48, 56×56, 64×64, 72×72}, and 48×48 was found to be optimal to capture the contextual information about land use for linear objects.

All the other parameters for both CNN networks were optimized empirically based on standard computer vision. 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 land use categories. The learning rate and the epoch were set as 0.01 and 600 to learn the deep features through backpropagation.

5.3.2.3 OCNN parameter settings

In the proposed OCNN method, the LIW-CNN and the SIW-CNN networks were integrated to predict the land use classes of general objects and linearly shaped objects at the model inference phase. Based on object convolutional position analysis (OCPA), the LIW-CNN with a 128×128 input window (denoted as $OCNN_{128}$) was employed only once per object, and the SIW-CNNs with a 48×48 input window (denoted as $OCNN_{48*}$, the 48* here represents multiple image patches sized 48×48) were used at multiple positions to predict the land use label of an object through majority voting (see section 2.2.2 for theoretical details). The parallel distance parameter *d* in OCPA that controls the convolutional locations and the number of small window size CNNs, was estimated by the length distribution of the moment box together with a trial-and-error procedure in a wide search space (0.5 m – 20 m) with a step of 0.5 m. The *d* was optimized as 5 m for the objects with moment box length (*l*) larger than or equal to 20 m, and was estimated by *l*/4 for those objects with *l* less than 20 m (i.e. the minimum number of small window size CNNs was 3) to perform a statistical majority voting. The proposed method ($OCNN_{128+48*}$) integrates both $OCNN_{128}$ and $OCNN_{48*}$, which is suitable for the prediction of urban land use semantics for any shaped objects.

5.3.2.4 Other benchmark methods and their parameters

To evaluate the classification performance of the proposed method, three existing benchmark methods (i.e. Markov Random Field (MRF), object-based image analysis

with support vector machine (OBIA-SVM), and the pixel-wise CNN) that each incorporate spatial context were compared comprehensively, as follows:

MRF: The Markov Random Field, a spatial contextual classifier, was used as a benchmark comparator. The MRF was constructed by the conditional probability formulated by a support vector machine (SVM) at pixel level, which was parameterized through grid search with a 5-fold cross-validation. The spatial context was incorporated by a fixed size of neighbourhood window (7×7) and a parameter γ that controls the smoothness level, set as 0.7, to achieve an appropriate level of smoothness in the MRF. The simulated annealing optimization approach with a Gibbs sampler (Berthod et al. 1996) was employed in the MRF to maximize the posterior probability through iteration.

OBIA-SVM: The multi-resolution segmentation was implemented initially to segment objects through the image. A range of features was 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 degree of similarity 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 parameterized SVM for object-based classification.

Pixel-wise CNN: The standard pixel-wise CNN was trained to predict all pixels within the images using densely overlapping image patches. The most important parameters that influence directly the classification performance of the pixel-wise CNN are the input image patch size and the number of layers (depth). Following the discussion by Längkvist et al., (2016), the input image size was chosen from {28×28, 32×32, 36×36,

40×40, 44×44, 48×48, 52×52 and 56×56} to evaluate the influence of contextual area on classification performance. 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 plays a key role in classification accuracy because the quality of the learnt features is highly influenced by the level of abstraction and representation. As suggested by Chen et al., (2016a), the number of CNN layers was chosen as six to balance the network complexity and robustness. Other CNN parameters were tuned empirically through cross-validation. For example, the filter size was set to 3×3 for the convolutional layer with a stride of 1, and the number of filters was set to 24 to extract multiple convolutional features at each level. The learning rate was set as 0.01 and the number of epochs was chosen as 600 to fully learn the features through backpropagation.

5.3.3 Classification results and analysis

The classification performance of the proposed $OCNN_{128+48*}$ method using the above-mentioned parameters was investigated on both S1 (experiment 1) and S2 (experiment 2). The proposed method was compared with $OCNN_{128}$ and $OCNN_{48*}$ as well as the benchmark MRF, OBIA-SVM and the pixel-wise CNN. Visual inspection and quantitative accuracy assessment, including pixel-based overall accuracy (OA), Kappa coefficient (κ) and the per-class mapping accuracy as well as object-based accuracy assessment, were adopted to evaluate the classification results hereafter.

Experiment 1: A desirable classification result was obtained in S1 by using the proposed $OCNN_{128+48*}$. To provide a useful visualization, three subsets of S1 classified by different approaches were presented in Figure 5-8, with the correct or incorrect classification results marked in yellow or red circles, respectively. In general, the proposed method achieved the smoothest visual results with precise boundary

information compared with other benchmark methods. Most importantly, the semantic contents of complex urban land uses (e.g. commercial, industrial etc.) were effectively characterized, and the linearly shaped features including highway and railway were identified with high geometric fidelity. As shown by Figure 5-8(*a*) and 5-8(*c*), the highway (a linear feature) was misclassified as a parking lot (red circles) by $OCNN_{128}$, whereas the highway feature was accurately identified by the $OCNN_{48*}$ (yellow circles). However, $OCNN_{48*}$ was inferior to $OCNN_{128}$ when identifying general objects, as demonstrated by Figure 5-8(*b*). Fortunately, these complementary behaviours of the two sub-modules were captured by the proposed $OCNN_{128+48*}$, which was able to label the highway accurately (yellow circles in Figure 5-8(*b*)). The pixel-wise CNN demonstrated some capacity for extracting semantic functions for complex objects; for example, the commercial area in Figure 5-8(*b*) was correctly distinguished (yellow circle). However, classification errors along the edges or boundaries between objects were found. For example, the edges of the highway were misclassified as high-density residential as shown by Figure 5-8(*a*). For the OBIA-SVM, the simple land uses with less within-object variation (e.g. highway) were more accurately classified (yellow circle in Figure 5-8(*a*) and 5-8(*c*)), whereas, those highly complex land uses with great within-object variation (e.g. commercial, industrial etc.) were more likely to be misclassified (red circle in Figure 5-8(*b*)). In addition, the OBIA-SVM could also discover some sub-objects (e.g. balcony on the residential house) through the information context. The results of the MRF, in contrast to the other object-based approaches, were the least smooth even though local neighbourhood information was used. Nevertheless, there were still some benefits of the MRF: spectrally distinctive land uses, such as highway, park and recreational area, were classified with a relatively high accuracy.

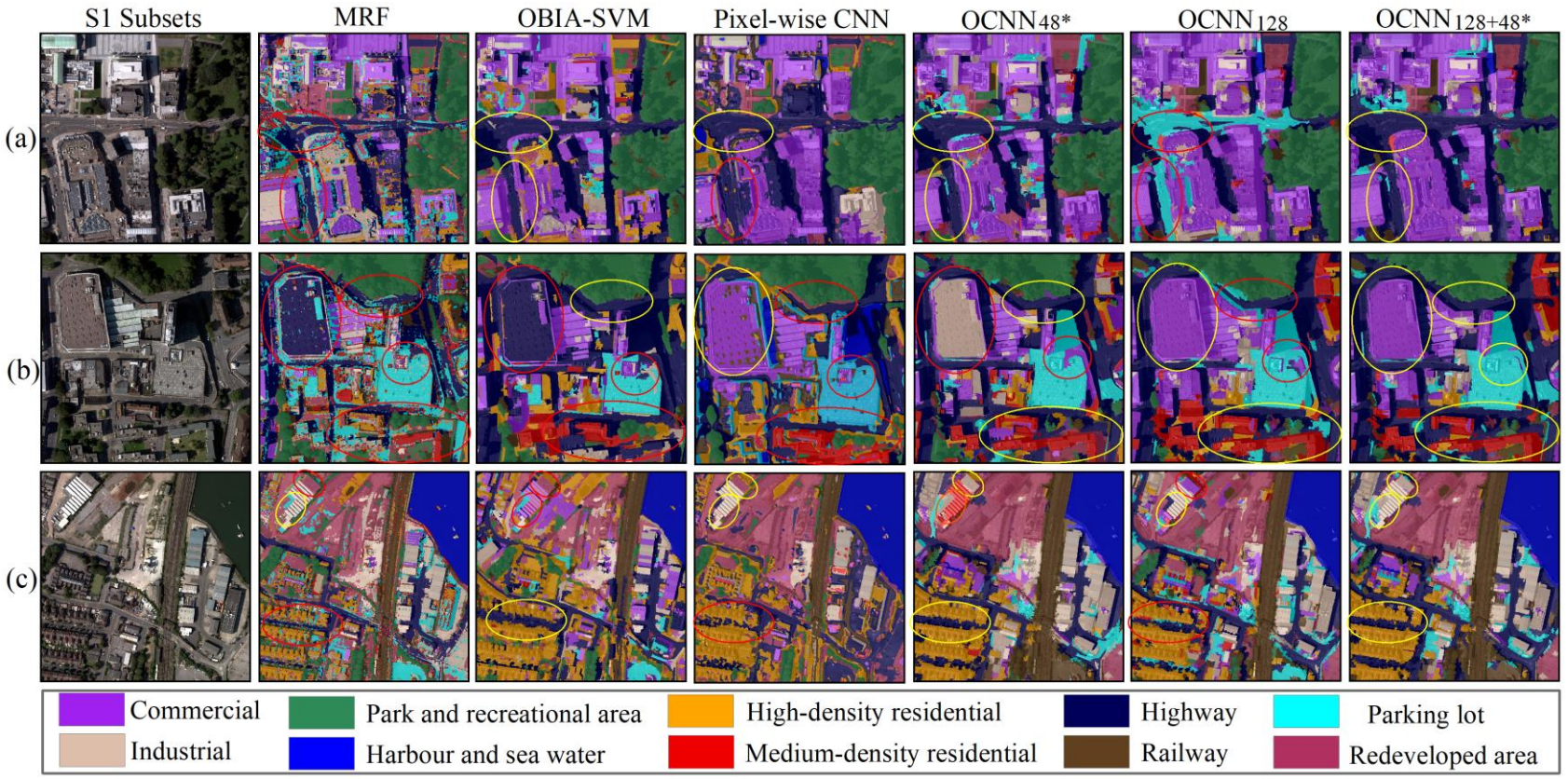


Figure 5-8: Three typical image subsets (a, b and c) in study site S1 with their classification results. Columns from left to right represent the original images (R G B bands only), and the MRF, OBIA-SVM, Pixel-wise CNN, $OCNN_{48*}$, $OCNN_{128}$, and the proposed $OCNN_{128+48*}$ results. The red and yellow circles denote incorrect and correct classification, respectively.

The effectiveness of the $OCNN_{128+48*}$ was also demonstrated by quantitative classification accuracy assessment. As shown in Table 5-2, the $OCNN_{128+48*}$ achieved the largest overall accuracy of 89.52% with a Kappa coefficient (κ) of 0.88, consistently larger than its sub-module $OCNN_{128}$ (87.31% OA and κ of 0.86) and the $OCNN_{48*}$ (OA of 84.23% and κ of 0.82), respectively. The accuracy increase was much more dramatic in comparison with other benchmark methods, including the pixel-wise CNN (81.62% OA and κ of 0.80), the OBIA-SVM (79.54% OA and κ of 0.78), as well as the MRF (OA of 78.67% and κ of 0.76). The superiority of the proposed $OCNN_{128+48*}$ was further demonstrated by the per-class mapping accuracy (Table 5-3). From the table, it can be seen that the accuracies of highway and railway were increased significantly by 5.34% and 4.64% respectively, compared with the $OCNN_{128}$. This was followed by a moderate increase of 3.24% for the parking lot class. Other land use classes (e.g. commercial, industrial, etc.) were slightly increased in terms of classification accuracy (less than 1.5%) without statistical significance in comparison with $OCNN_{128}$. When comparing with the $OCNN_{48*}$, the accuracy increase of the proposed $OCNN_{128+48*}$ was remarkable for the majority of general object classes, with increases of up to 6.06%, 6.51%, 4.98%, 4.7% and 4.68%, for the classes of commercial, industrial, redeveloped area, park and recreational area, and high-density residential, respectively; whereas the accuracies of the medium-density residential and the parking lot increased moderately, by 3.31% and 3.81%, respectively. For linearly shaped objects, however, the $OCNN_{128+48*}$ was not substantially superior to the $OCNN_{48*}$, with just a slight accuracy increase of 1.52% for highway and 2.41% for railway, respectively. For general objects with complex semantic functions, including commercial, industrial, redeveloped area, park and recreational area, and high-density residential, the increase in accuracy of the

$OCNN_{128+48*}$ was much more significant, by up to 6.06%, 6.51%, 4.98%, 4.7% and 4.68%, respectively.

In terms of the pixel-wise CNN, effectiveness was observed for certain complex objects (e.g. the accuracy for the industrial land use was up to 80.23%). However, the simple and geometrically distinctive land use classes were not accurately mapped, with the largest accuracy difference up to 6.57% for the class highway compared with the $OCNN_{128+48*}$. By contrast, the OBIA-SVM demonstrated some advantages on simple land use classes (e.g. the accuracy of railway up to 90.65%), but it failed to accurately identify more complex general objects (e.g. an accuracy as low as 71.87% for commercial land use). The MRF presented the smallest classification accuracy for most land use classes, especially the complex general land uses (e.g. 12.37% accuracy lower than the $OCNN_{128+48*}$ for commercial land use).

Table 5-3 - Classification accuracy comparison amongst MRF, OBIA-SVM, Pixel-wise CNN, $OCNN_{48*}$, $OCNN_{128}$, and the proposed $OCNN_{128+48*}$ method for Southampton using the per-class mapping accuracy, overall accuracy (OA) and Kappa coefficient (κ). The bold font highlights the greatest classification accuracy per row.

Class	MRF	OBIA-SVM	Pixel-wise CNN	$OCNN_{48*}$	$OCNN_{128}$	$OCNN_{128+48*}$
commercial	70.09	72.87	73.26	76.4	81.13	**82.46**
highway	77.23	78.04	76.12	78.17	74.35	**79.69**
industrial	67.28	69.01	71.23	78.24	83.87	**84.75**
high-density residential	81.52	80.59	80.05	81.75	85.35	**86.43**
medium-density residential	82.74	84.42	85.27	87.28	90.34	**90.59**
park and recreational area	91.05	93.14	92.34	92.59	96.41	**97.09**
parking lot	80.09	83.17	84.76	86.02	85.59	**88.83**
railway	88.07	90.65	86.57	89.51	87.28	**91.92**
redeveloped area	89.13	90.02	89.26	89.71	94.57	**94.69**
harbour and sea water	97.39	98.43	98.54	98.62	98.75	**98.95**
Overall Accuracy (OA)	78.67%	79.54%	81.62%	84.23%	87.31%	**89.52%**
Kappa Coefficient (κ)	0.76	0.78	0.8	0.82	0.86	**0.88**

An object-based accuracy assessment was implemented in S1 to validate the classification performance in terms of over-classification (*OC*), under-classification (*UC*), and total classification error (*TCE*). Three typical methods, including OBIA-SVM (denoted as OBIA), pixel-wise CNN (denoted as CNN), and the proposed $OCNN_{128+48*}$ method (denoted as OCNN), were evaluated, with accuracy comparisons of each land use class listed in Table 5-4. Clearly, the proposed OCNN method produced the smallest *OC*, *UC*, and *TCE* errors, respectively (highlighted by bold font), constantly smaller than those of the CNN and OBIA. Generally, the *UC* errors are smaller than *OC* errors, demonstrating that a slight over-segmentation was produced. Specifically, the OCNN demonstrates excellent object-level classification, with the majority of classes less than 0.2 in *TCE*. Those complex land use classes, including commercial and industrial, can be segmented precisely and classified with small *TCE* of 0.22 and 0.20, less than those of CNN (0.29 and 0.27) and OBIA (0.39 and 0.38). The parking lot objects with complex land use patterns, were also recognised accurately with high fidelity (*OC* of 0.22, *UC* of 0.13, and *TCE* of 0.17), less than CNN (0.28, 0.17, and 0.22) as well as OBIA (0.41, 0.32, and 0.37). For those LS-objects, the OCNN achieved promising accuracy in comparison with the other two benchmarks. For example, the *TCEs* of highway and railway produced by the OCNN were 0.17 and 0.09, smaller than those of the CNN (0.25 and 0.22) and OBIA (0.20 and 0.18). All the other land use categories demonstrate increased segmentation accuracy. For instance, the *TCE* of park and recreational area was 0.18 with the OCNN, less than for the CNN of 0.24 and OBIA of 0.32.

Table 5-4 - Object-based accuracy assessment among OBIA-SVM (OBIA), Pixel-wise CNN (CNN), and the proposed OGC-CNN$_{128+48*}$ method (OCNN) for Southampton using error indices of *OC*, *UC*, and *TCE*. The bold font highlights the lowest classification error of a specific index per row.

Class	*OC*			*UC*			*TCE*		
	OBIA	CNN	OCNN	OBIA	CNN	OCNN	OBIA	CNN	OCNN
commercial	0.45	0.33	**0.26**	0.34	0.26	**0.18**	0.39	0.29	**0.22**
highway	0.23	0.29	**0.19**	0.17	0.21	**0.16**	0.20	0.25	**0.17**
industrial	0.42	0.31	**0.23**	0.36	0.24	**0.17**	0.38	0.27	**0.20**
high-density residential	0.34	0.28	**0.14**	0.26	0.19	**0.08**	0.30	0.23	**0.11**
medium-density residential	0.29	0.21	**0.16**	0.21	0.14	**0.09**	0.25	0.17	**0.12**
park and recreational area	0.36	0.29	**0.24**	0.28	0.19	**0.12**	0.30	0.24	**0.18**
parking lot	0.41	0.28	**0.22**	0.32	0.17	**0.13**	0.37	0.22	**0.17**
railway	0.25	0.27	**0.12**	0.11	0.18	**0.06**	0.19	0.21	**0.09**
redeveloped area	0.37	0.32	**0.21**	0.29	0.25	**0.13**	0.33	0.28	**0.17**
harbour and sea water	0.18	0.19	**0.14**	0.07	0.11	**0.06**	0.12	0.15	**0.09**

Experiment 2: The most accurate classification performance was also achieved in S2 by the proposed method, as illustrated by the quantitative accuracy results in Table 5-5. From the table, it can be seen that OCNN$_{128+48*}$ obtained the greatest overall accuracy (OA) of 90.87% with a Kappa coefficient (κ) of 0.88, significantly larger than the OCNN$_{128}$ (OA of 88.74% and κ of 0.86), the OCNN$_{48*}$ (OA of 85.06% with κ of 0.83), the Pixel-wise CNN (OA of 82.39% and κ of 0.81), the OBIA-SVM (OA of 80.37% with κ of 0.79), and the MRF (OA of 78.52% with κ of 0.76). The effectiveness of the OCNN$_{128+48*}$ was also demonstrated by the per-class mapping accuracy. Compared with the OCNN$_{128}$, the classes formed by linearly shaped objects, including the highway, railway and canal, had significantly increased accuracies of up to 5.36%, 3.06% and 3.48%, respectively (Table 5-5). Such increases can also be noticed in Figure 5-9 (a subset of S2), where the misclassifications of railway and highway shown in

Figure 5-9(*g*) were rectified in Figure 5-9(*h*) classified by the $OCNN_{128+48*}$. At the same time, the parking lot land use class was moderately increased by 2.28%. Whereas, other land use classes had slightly increases in accuracy of less than 1% on average. In contrast, the $OCNN_{128+48*}$ led to no significant increases over the $OCNN_{48*}$ for the linear object classes, with accuracy increases for highway, railway and canal of 1.8%, 0.42% and 1.22%, respectively. For the general classes, especially the complex land uses (e.g. commercial, industrial etc.), remarkable accuracy increases were achieved with an average up to 6.75%. Figure 5-9(*f*) (classified by $OCNN_{48*}$) also showed the confusion between the commercial and industrial land use classes, which was revised in Figure 5-9(*h*). With respect to the benchmark comparators, the accuracy increase of $OCNN_{128+48*}$ was much more obvious for most of the land use classes, with the largest accuracy increase up to 12.39% for parking lot, 11.21% for industrial, and 8.56% for commercial, compared with the MRF, OBIA-SVM and Pixel-wise CNN, respectively. The undesirable visual effects and misclassifications can also be seen in Figure 5-9(*c-e*), which were corrected in Figure 5-9(*h*).

Similar to S1, the object-based accuracy assessment was conducted in S2 to investigate the over-, under-, and total classification errors of each class using the OCNN, CNN and OBIA methods (Table 5-6). The error indices in S2 (Table 5-6) present a similar trend with those in S1 (Table 5-4), although the geometric errors for S2 are smaller than for S1 due to the relatively regular land use structures and configurations in Manchester city centre. The proposed OCNN yielded the greatest classification accuracy with the smallest error indices (highlighted by bold font), smaller than those of the CNN and OBIA. The OCNN accurately differentiated the complex land use classes, with a *TCE* of 0.20, 0.17, and 0.15 for the classes of commercial, industrial and parking lot, respectively (Table 5-6), significantly smaller than for the CNN (0.27, 0.26, and 0.24),

and OBIA (0.37, 0.35, and 0.32). Those linearly shaped objects, including highway, railway, and canal, were precisely characterised by the OCNN method, with a *TCE* of 0.16, 0.09, and 0.08, significantly smaller than for the CNN (0.22, 0.21, and 0.14) and OBIA (0.18, 0.19, and 0.12). The residential land use was also clearly improved with a very small *TCE* of 0.10, smaller than for the CNN (0.22) and OBIA (0.26). Other land use classes, such as the park and recreational area and the redeveloped area, were also better distinguished by the OCNN (0.16 and 0.15 in terms of *TCE*), smaller than for the CNN (0.21 and 0.25) and OBIA (0.28 and 0.30).

Table 5-5 - Classification accuracy comparison amongst MRF, OBIA-SVM, Pixel-wise CNN, $OCNN_{48*}$, $OCNN_{128}$, and the proposed $OCNN_{128+48*}$ method for Manchester, using the per-class mapping accuracy, overall accuracy (OA) and Kappa coefficient (κ). The bold font highlights the greatest classification accuracy per row.

Class	MRF	OBIA-SVM	Pixel-wise CNN	$OCNN_{48*}$	$OCNN_{128}$	$OCNN_{128+48*}$
commercial	71.11	72.47	74.16	76. 27	82.43	**82.72**
highway	80.43	79.26	80.59	82.57	79.01	**84.37**
industrial	73.52	72.05	74.84	76.22	82.19	**83.26**
residential	78.41	80.45	80.56	83.09	84.75	**84.99**
parking lot	79.63	82.06	84.37	87.86	89.74	**92.02**
railway	85.94	88.14	88.32	91.06	88.42	**91.48**
park and recreational area	88.42	89.54	90.76	91.34	94.38	**94.59**
redeveloped area	82.07	84.15	87.04	88.83	93.16	**93.75**
canal	90.02	92.28	94.18	97.52	95.26	**98.74**
Overall Accuracy (OA)	78.52%	80.37%	82.39%	85.06%	88.74%	**90.87%**
Kappa Coefficient (κ)	0.76	0.79	0.81	0.83	0.86	**0.88**

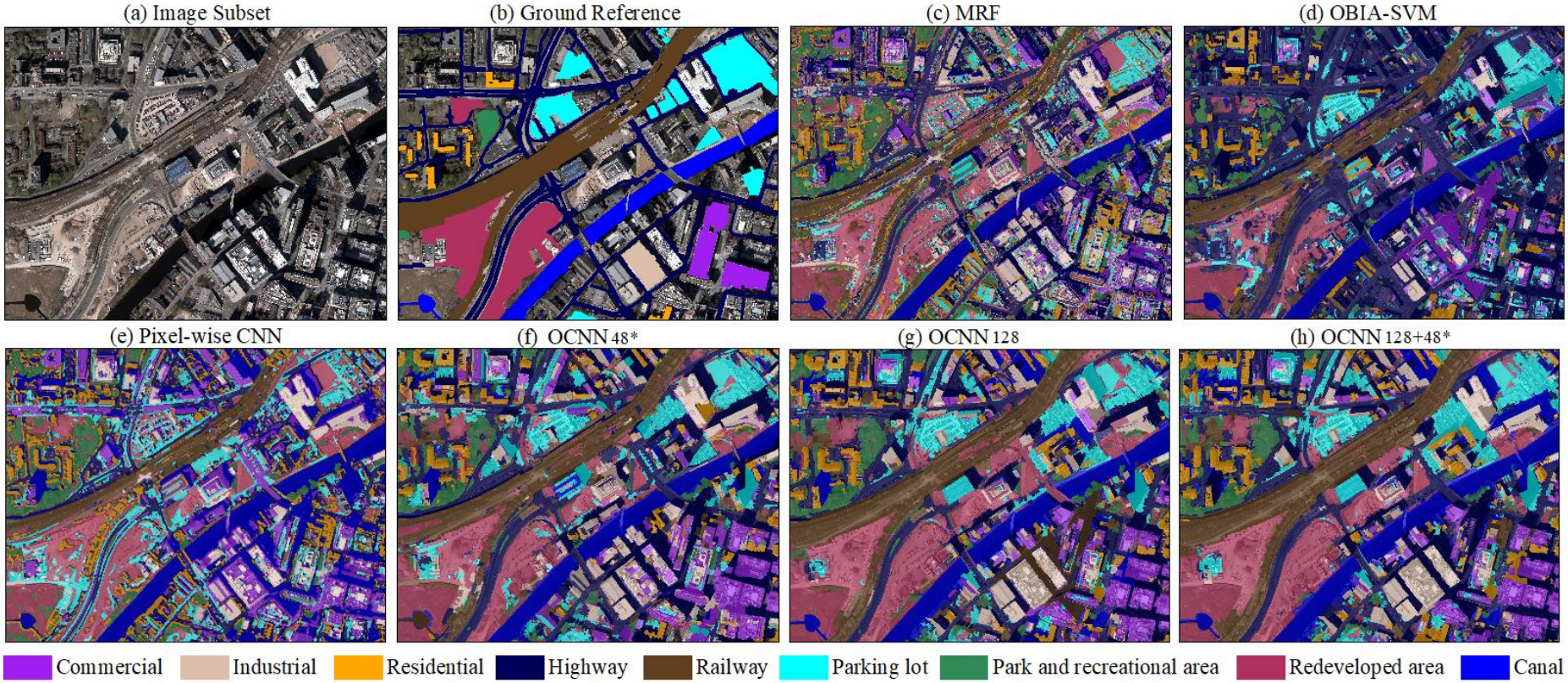


Figure 5-9: Classification results in study site S2, with (a) an image subset (R G B bands only), (b) the ground reference, (c) MRF classification, (d) OBIA-SVM classification, (e) Pixel-wise CNN classification, (f) $OCNN_{48*}$ classification, (g) $OCNN_{128}$ classification, and (h) $OCNN_{128+48*}$ classification

Table 5-6 - Object-based accuracy assessment among OBIA-SVM (OBIA), Pixel-wise CNN (CNN), and the proposed OGC-CNN$_{128+48*}$ method (OCNN) for Manchester using error indices of *OC*, *UC*, and *TCE*. The bold font highlights the lowest classification error of a specific index per row.

Class	*OC*			*UC*			*TCE*		
	OBIA	CNN	OCNN	OBIA	CNN	OCNN	OBIA	CNN	OCNN
commercial	0.41	0.32	**0.24**	0.32	0.23	**0.16**	0.37	0.27	**0.20**
highway	0.22	0.27	**0.18**	0.15	0.19	**0.15**	0.18	0.23	**0.16**
industrial	0.39	0.31	**0.20**	0.31	0.22	**0.14**	0.35	0.26	**0.17**
residential	0.30	0.24	**0.12**	0.22	0.20	**0.09**	0.26	0.22	**0.10**
parking lot	0.37	0.26	**0.19**	0.28	0.22	**0.12**	0.32	0.24	**0.15**
railway	0.22	0.25	**0.10**	0.14	0.19	**0.07**	0.18	0.22	**0.09**
park and recreational area	0.31	0.25	**0.21**	0.26	0.17	**0.10**	0.28	0.21	**0.16**
redeveloped area	0.34	0.29	**0.18**	0.26	0.22	**0.12**	0.30	0.25	**0.15**
canal	0.16	0.17	**0.12**	0.08	0.12	**0.05**	0.12	0.14	**0.08**

A sensitivity analysis was conducted to further investigate the effect of different input window sizes on the overall accuracy of urban land use classification (see Figure 5-10). The window sizes varied from 16×16 to 144×144 with a step size of 16. From Figure 5-10, it can be seen that both S1 and S2 demonstrated similar trends for the proposed OCNN and the pixel-wise CNN (CNN). With window sizes smaller than 48×48 (i.e. relatively small windows), the classification accuracy of OCNN is lower than that of CNN, but the accuracy difference decreases with an increase of window size. Once the window size is larger than 48×48 (i.e. relatively large windows), the overall accuracy of the OCNN increases steadily until the window is as large as 128×128 (up to around 90%), and outperforms the CNN which has a generally decreasing trend in both study sites. However, an even larger window size (e.g. 144×144) in OCNN could result in over-smooth results, thus reducing the classification accuracy.

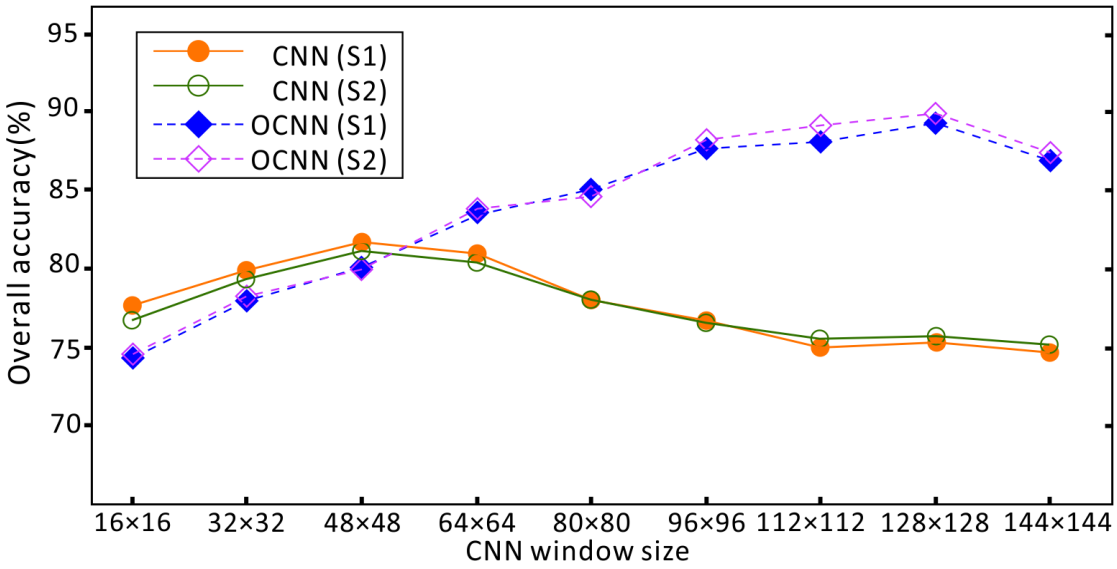


Figure 5-10: The influence of CNN window size on the overall accuracy of pixel-wise CNN and the proposed OCNN method for both study sites S1 and S2.

5.3.4 Computational efficiency

The computational efficiency of the proposed method was evaluated and compared with the other methods listed in Table 5-7. The classification experiments were implemented using Keras/Tensorflow under a Python environment with a laptop of NVIDIA 940M GPU and 12.0 GB memory. As shown in Table 5-7, the training time of the Pixel-wise CNN, $OCNN_{48*}$, $OCNN_{128}$ and the proposed $OCNN_{128+48*}$ were similar in both experiments, with an average time of 4.27 h, 4.36 h, 4.74 h, and 4.78 h, respectively. The prediction time for the Pixel-wise CNN was the longest compared with other OCNN-based approaches with 321.07 h on average, about 100 times longer than those of the OCNN-based approaches. Among the three OCNN methods, the $OCNN_{128}$ and the $OCNN_{128+48*}$ were similar in computational efficiency with average of 2.81 h and 2.9 h, respectively, longer than that of the $OCNN_{48*}$ (1.78 h on average) for the two experiments. The benchmark methods, the MRF and OBIA-SVM, spent much less time on the training and prediction phases than the CNN-based methods, with an average of

1.4 h and 1.2 h for the two experiments, about 20 times and 3 times less than the pixel-wise CNN and the OCNN-based approaches, respectively.

Table 5-7 - Comparison of computational times amongst MRF, OBIA-SVM, Pixel-wise CNN, $OCNN_{48*}$, $OCNN_{128}$, and the proposed $OCNN_{128+48*}$ approach in S1 and S2.

	Study area	No. of object	Mean Area (m^2)	Computation time (h)					
				MRF	OBIA-SVM	Pixel-wise CNN	$OCNN_{48*}$	$OCNN_{128}$	$OCNN_{128+48*}$
Train	S1	6328	25.37	1.42	0.58	4.45	4.45	4.88	4.92
	S2	6145	25.92	1.37	0.44	4.08	4.27	4.59	4.64
Predict	S1	61 921	26.61	1.52	1.76	326.78	1.82	2.83	2.94
	S2	58 408	25.75	1.33	1.55	315.36	1.74	2.78	2.86

5.4 Discussion

Urban land use captured in VFSR remotely sensed imagery is highly complex and heterogeneous, with spatial patterns presented that imply a hierarchical or nested class structure. Classifying urban land use requires not only a precise characterisation of image objects as functional units, but also an accurate and robust representation of spatial context. A novel object-based CNN method for urban land use classification using VFSR remotely sensed imagery was, therefore, proposed, in which the functional units are derived at object levels and the spatial patterns are learned through CNN networks with hierarchical feature representation. The OCNN method is fundamentally different from the work proposed by Zhao et al. (2017) in multiple aspects, including: (1) the realisation of an object-based CNN for land use classification under the OBIA framework using geometric characterisations to guide the choice of sizes and locations of image patches; (2) the use of within-object and between-object information learnt by the OCNN model to represent the spatial and hierarchical relationships; (3) the high

computational efficiency achieved with targeted sampling at the object level to avoid a pixel-wise (i.e., densely overlapping) convolutional process.

5.4.1 CNN for urban land use feature representation

Urban land use information is characterised as high-level spatial features in VFSR remotely sensed data, which are an abstraction of the observed spatial structures or patterns. Convolutional neural networks (CNN) are designed to learn such complex feature representations effectively from raw imagery, end-to-end, by cascading multiple layers of nonlinear processing units. As shown in Table 5-3, the pixel-wise CNN achieved greater classification accuracy than the traditional MRF and OBIA-SVM methods on complex land use categories, such as Commercial, Industrial, and Parking lot, owing to its capacity for complex spatial contextual feature representation. Nevertheless, the pixel-wise CNN is essentially designed to predict image patches, whereas urban land use classification requires each pixel of the remotely sensed imagery to be labelled as a particular land use class to create a thematic map. The boundary information of the land use is often weakened by the pixel-wise convolutional process with image patches, where blurred boundaries occur between the classified objects with a loss of small useful land features, somewhat similar to morphological or Gabor filter methods (Reis and Tasdemir 2011, Pingel et al. 2013). This problem is exacerbated when trying to extract high-level land use semantics using deep CNN networks with large input window sizes (see the declining trend of overall accuracy for large window sizes as illustrated by Figure 5-10 due to the over-smoothness). These demonstrate the need for innovation through adaptation of the CNNs for urban land use classification using appropriate functional units and convolutional processes.

5.4.2 Object-based CNN (OCNN) for urban land use classification

The proposed object-based CNN (OCNN) is built upon segmented objects with spectrally homogeneous characteristics as the functional units, in which the precise boundary information is characterised at the object level. Unlike the standard pixel-wise CNN with image patches that are densely overlapping throughout the image, the OCNN method analyses and labels objects using CNN networks by incorporating the objects and their spatial context within image patches. This provides a new perspective for object description and feature characterisation, where both within-object information and between-object information are jointly learned inside the model. Since each segmented object is labelled with a single land use as a whole, the homogeneity of each object is crucial to achieving high land use classification accuracy. To produce a set of such objects with local homogeneity, a slight over-segmentation was adopted in this research, as suggested by previous studies (e.g. Hofmann et al. 2011, Li et al. 2015). In short, the OCNN method, as a combination of CNN and OBIA, demonstrates strong capacity for classifying complex urban land uses through deep feature representations, while maintaining the fine spatial details using regional partition and boundary delineation.

Each segmented object has its distinctive geometric characteristics with respect to the specific land use category. Representations of objects using OCNN should be scale-dependent with appropriate window sizes and convolutional positions to match the geometric distributions, especially when dealing with the two types of objects with geometrically distinctive characteristics, namely, general objects (G-objects) and linearly-shaped objects (LS-objects). For those G-objects with complex urban land use, a deep CNN network (eight-layers) with a large input image patch (128×128) was used to accurately identify an object with a large extent of contextual information. Such an

image patch could reflect the real dimension of G-objects and their wide context (64m×64m in geographical space). The convolutional position of the CNN network was theoretically derived close to the central region of a moment box, where both object geometry and spatial anisotropy were characterised. In this way, the within-object (at the centre of the image patch) and between-object (surrounding context within the image patch) information are used simultaneously to learn the objects and the surrounding complex spatial structures or patterns, with the largest overall accuracy at large context (Figure 5-10). The LS-objects, such as Highway, Railway and Canal, were sampled along the objects using a range of less deep CNNs (six-layers) with small window size (48×48) (or 24m×24m geographically) and were classified through majority voting. These small window size CNNs focus on the within-object information, which often includes homogeneous characteristics within objects (e.g. rail tracks, asphalt road), and avoid the great variation between adjacent objects (e.g. trees, residential buildings, bare land etc. alongside the Highway). Moreover, the small contextual image patches with less deep networks cover the elongated objects sufficiently, without losing useful within-object information through the convolutional process. To integrate the two classification models for G-objects and LS-objects, a simple rule-based classification integration was employed conditional upon model predictions, in which the majority of the classification results were derived from the CNNs with large window size, whereas the predictions of Highway, Railway and Canal were trusted by the voting results of small window CNNs alone. Thus, the type of object (either as a G-object or a LS-object) is determined through CNN model predictions and rule-based classification integration. Such a decision fusion approach provides a pragmatic and effective manner to combine the two models by considering the object geometry and class-specific adaptations. Overall, the proposed OCNN method with

large and small window size feature representations is a feasible solution for the complex urban land use classification problem using VFSR remotely sensed imagery, with massive generalisation capability for a broad range of applications.

5.4.3 Computational complexity and efficiency

Throughout the computational process, the model inference of the pixel-wise CNN is the most time-consuming stage for urban land use classification using VFSR remotely sensed imagery. The prediction of the CNN model over the entire image with densely overlapping image patches gives rise to a time complexity of O(N), where N represents the total number of pixels of the image. Such a time complexity could be huge when classifying a large image coupled with relatively large image patches as input feature maps. In contrast, the time complexity of the proposed OCNN method is remarkably reduced from O(N) at pixel level to O(M) at object level with M segmented objects, where a significant time decrease of up to N/M times (N/M here denotes the average object size in pixels) can be achieved. The time reductions for both S1 and S2 are around 100 times, approximating to those of the mean object sizes (Table 5-7), thus, being more acceptable than the standard pixel-wise CNN. Such a high computational efficiency demonstrates the practical utility of the proposed OCNN method to general users with limited computational resources.

5.4.4 Future research

The proposed OCNN method provides a very high accuracy and efficiency for urban land use classification using VFSR remotely sensed imagery. The image objects are identified through decision fusion between a large input window CNN with a deep network and several small input window CNNs with less deep networks, to account for typical distinctive object sizes and geometries. However, such two-scale feature

representation might be insufficient to characterise some complex geometric characteristics. Therefore, a range of CNNs with different input patch sizes will be adopted in the future to adapt to the diverse sizes and shapes of the urban objects through weighted decision fusion. In addition, urban land use classification was undertaken at a generalized spatial and semantic level (e.g., residential area, commercial area and industrial area), without identifying smaller functional sites (e.g., supermarkets, hospitals and playgrounds etc.). This issue might be addressed by incorporating multi-source geospatial data, for example, those classified commercial areas might be further differentiated as supermarkets, retail outlets, and café areas through indoor human activities. Future research will, therefore, mine the semantic information from GPS trajectories, transportation networks and social media data to characterise these smaller functional units in a hierarchical way, as well as socioeconomic activities and population dynamics.

5.5 Conclusion

Urban land use classification using VFSR remotely sensed imagery remains a challenging task, due to the indirect relationship between the desired high-level land use categories and the recorded spectral reflectance. A precise partition of functional units as image objects together with an accurate and robust representation of spatial context are, therefore, needed to characterise urban land use structures and patterns into high-level feature thematic maps. This chapter proposed a novel object-based CNN (OCNN) method for urban land use classification from VFSR imagery. In the OCNN, segmented objects consisting of linearly shaped objects (LS-objects) and other general objects (G-objects), were utilized as functional units. The G-objects were precisely identified and labelled through a single large input window (128×128) CNN with a

deep (eight-layer) network to perform a contextual object-based classification. Whereas the LS-objects were each distinguished accurately using a range of small input window (48×48) CNNs with less deep (six-layer) networks along the objects' lengths through majority voting. The locations of the input image patches for both CNN networks were determined by considering both object geometry and its spatial anisotropy, such as to accurately classify the objects into urban land use classes. Experimental results on two distinctive urban scenes demonstrated that the proposed OCNN method significantly increased the urban land use classification accuracy for all land use categories. The proposed OCNN method with large and small window size CNNs produced the most accurate classification results in comparison with the sub-modules and other contextual-based and object-based benchmark methods. Moreover, the OCNN method demonstrated a high computational efficiency with much more acceptable time requirements than the standard pixel-wise CNN method in the process of model inference. We conclude that the proposed OCNN is an effective and efficient method for urban land use classification from VFSR imagery. Meanwhile, the OCNN method exhibited an excellent generalisation capability on distinctive urban land use settings with great potential for a broad range of applications.

Chapter 6 Joint Deep Learning for land cover and land use classification[4]

[4] This chapter is based on the paper under 2nd round review: Zhang, C., Sargent, I., Pan, X., Li, Andy Gardiner, Jonathon Hare, Peter M. Atkinson *, 2018d, Joint Deep Learning for land cover and land use classification. *Remote Sensing of Environment*. (Under 2nd Round Review)

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 chapter, 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 increasing 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 90.24% for LC and 88.01% 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, proposes 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

6.1 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 (Zhang et al. 2018a). An example of the latter is the ability to identify a railway station from the character of the objects that comprise it (e.g. long thin platforms, long thin roofs) and the set of objects that surround it (e.g. railway lines, car park and multiple roads) (Tang et al. 2016). 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 in comparison with coarse or medium spatial resolution remotely sensed data. 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-segmentation and under-segmentation commonly exist within the same image (Ming et al. 2015). To date, no effective solution has been proposed for LC classification using VFSR remotely sensed imagery, where objects of the same LC may exhibit strong spectral heterogeneity due to differences in age, level of maintenance and composition as well as illumination conditions (Demarchi et al. 2014).

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 (Myint 2001, Herold et al. 2003) 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 for some regions, and the spatial contexts are often hard to describe and characterise 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).

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 machine learning and pattern recognition 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. 2015b, 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, Yang et al. 2015). Owing to its superiority in higher-level feature representation, the CNN has demonstrated great

potential in many remote sensing tasks such as vehicle detection (Chen et al. 2014, Dong et al. 2015), road network extraction (Cheng, Wang, et al. 2017), remotely sensed scene classification (Othman et al. 2016, Sargent et al. 2017), and semantic segmentation (Zhao et al. 2017).

The essential characteristic of CNNs is their translational invariance through a patch-wise procedure, in which a higher-level object within an image patch can be recognised even if the pixels comprising the object are shifted or 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, Zhang et al. 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 chapter 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 chapter, 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 chapter is organised as follows: Section 6.2 introduces the general workflow and the key components of the proposed methods. Section 6.3 describes the study area and data sources. The results are presented in section 6.4, followed by a discussion in section 6.5. The conclusions are drawn in the last section.

6.2 Methodology

6.2.1 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.

6.2.2 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.

6.2.3 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. 2018). 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 (Zhang et al. 2018). 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.

6.2.4 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

$$p(x) = \prod_{v=1}^{k} p(x_v \mid x_{pa(v)}) \tag{6-1}$$

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 $C_{LC} = \{C_{LC1}, C_{LC2}, \ldots, C_{LCi} \ldots, C_{LCm}\}$ ($i \in [1, m]$), where C_{LCi} denotes the set of LC samples of the *i*th class, and *m* represents the number of LC classes; $C_{LU} = \{ C_{LU1}, C_{LU2}, \ldots, C_{LUj} \ldots, C_{LCn}\}$($j \in [1, n]$), 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 the 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 is derived by

$$C^* = \arg\max_{C_i} p(C_i \mid F) \tag{6-2}$$

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

Through the Bayes' theorem

$$p(C_i \mid F) = \frac{p(C_i)\,p(F \mid C_i)}{p(F)} \tag{6-3}$$

The classification result C^* is obtained as

$$C^* = \arg\max_{C_i} p(C_i)\,p(F \mid C_i) \tag{6-4}$$

where $p(F)$ is the same for all possible states of C_i.

The $p(C_i)$ models the prior probability distribution of different LC/LU classes. In this research, we do not consider any specific 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 6-1 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.

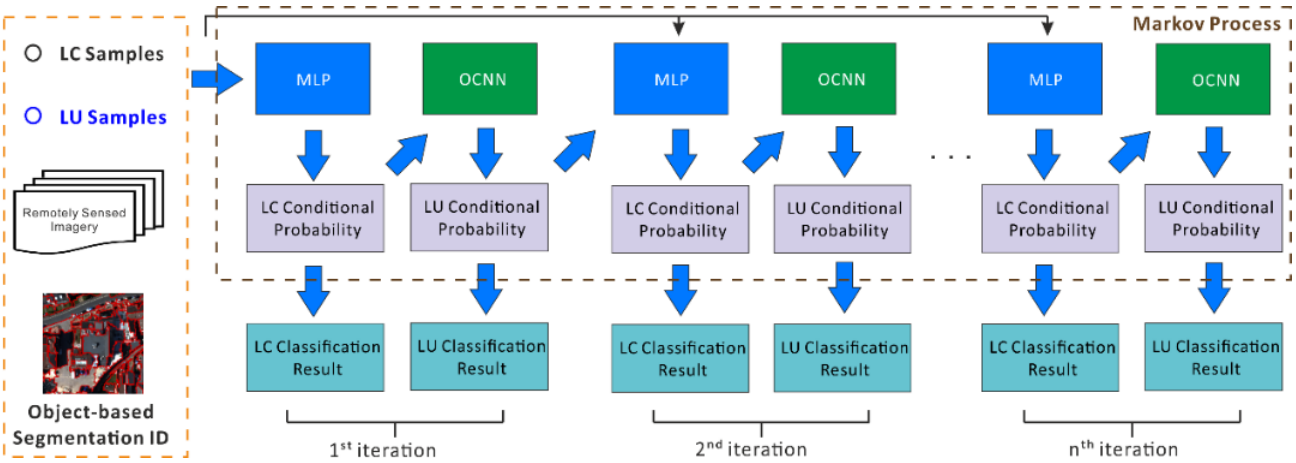


Figure 6-1: The general workflow of the land cover (LC) and land use (LU) joint deep learning (JDL).

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):

$$P(\text{LandCover}^{i}, \text{LandUse}^{i}) = P(\text{LandCover}^{i}, \text{LandUse}^{i} \mid \text{LandCover}^{i-1}, \text{LandUse}^{i-1}) \quad (6\text{-}5)$$

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 (6-9) can be expressed as:

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

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 ith 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:

$$P(\text{LandCover}^{i}) = P(\text{LandCover}^{i} \mid \text{LandUse}^{i-1}) \quad (6\text{-}7)$$

where the MLP model is trained to solve the equation (6-11) as:

$$MLPModel^{i} = TrainMLP(concat(\text{LandUse}^{i-1}, \text{Image}), C_{LC}) \quad (6\text{-}8)$$

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:

$$LandCover^{i} = MLPModel^{i}.predict\ (concate\ (\mathrm{LandUse}^{i-1}, \mathrm{Image}) \tag{6-9}$$

Land use (LU) conditional probabilities are deduced as:

$$P(\mathrm{LandUse}^{i}) = P(\mathrm{LandUse}^{i} \mid \mathrm{LandCover}^{i}) \tag{6-10}$$

where the OCNN model is built to solve equation (6-14) as:

$$OCNNModel^{i} = TrainCNN(\mathrm{LandCover}^{i}, C_{LU}) \tag{6-11}$$

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 CNN model training. The trained CNN model can then be used to predict the LU association as:

$$LandUse^{i} = CNNModel^{i}.predict(\mathrm{cast}(\mathrm{LandCover}^{i}, \mathrm{SegmentImage}) \tag{6-12}$$

where the function *cast* denotes the cropped image patch with LC probabilities derived from $\mathrm{LandCover}^{i}$, 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.

6.3 Experimental Results and Analysis

6.3.1 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 6-2). Both study areas involve urban and rural areas that are highly heterogeneous and distinctive from each other in LC and LU characteristics and are, therefore, suitable for testing the generalisation capability of the joint deep learning approach.

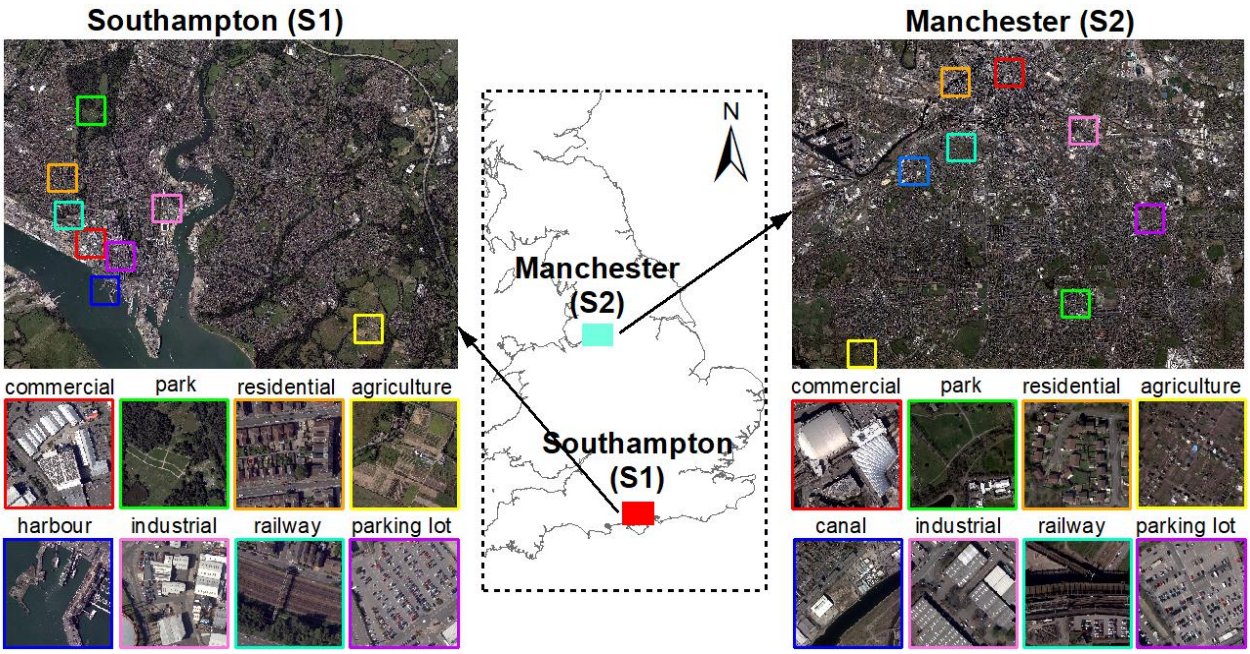


Figure 6-2: The two study areas: S1 (Southampton) and S2 (Manchester) with highlighted regions representing the majority of land use categories.

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, including *clay roof*, *concrete roof*, *metal roof*, *asphalt*, *rail*, *bare soil*, *woodland*, *grassland*, *crops*, and *water* (Table 6-1). 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 area*, *parking lot*, *redeveloped area*, and *canal* (Table 6-1). The majority of LU types for both study sites are highlighted and exemplified in Figure 6-2. 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 6-1.

Table 6-1 - 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 that are collected by local surveyors and digitised manually by photogrammetrists in the UK. These reference polygons with well-defined labelling protocols specified in Table 6-1 served as the basis for probability-based sample design. A stratified random sampling scheme was used to generate unbiased sample points for each class proportional upon the size of every individual reference polygon, the sample points were further split into 60% training samples and 40% testing samples at each class. 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).

6.3.2 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 series of predefined parameters to optimise the learning accuracy and generalisation capability. Detailed model structures and parameters were clarified as below.

6.3.2.1 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. Following the recommendations of Mas and Flores (2008), MLPs with one, two and three hidden layers were tested, using a varying number of {4, 8, 12, 16, 20, and 24} nodes in each layer. The learning rate was chosen optimally as 0.2 and the momentum factor was set as 0.7. In addition, the number of iterations was set as 800 to fully converge to a stable state. Through cross validation with different numbers of nodes and hidden layers, the optimal MLP parameter setting was found using two hidden layers with 16 nodes in each layer.

6.3.2.2 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 coupled with a small amount of 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.

6.3.2.3 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 6-3) 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 number of filters was tuned to 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 6-3, where the maximum pooling within a 2×2 window was used to generalise the feature and keep the parameters tractable.

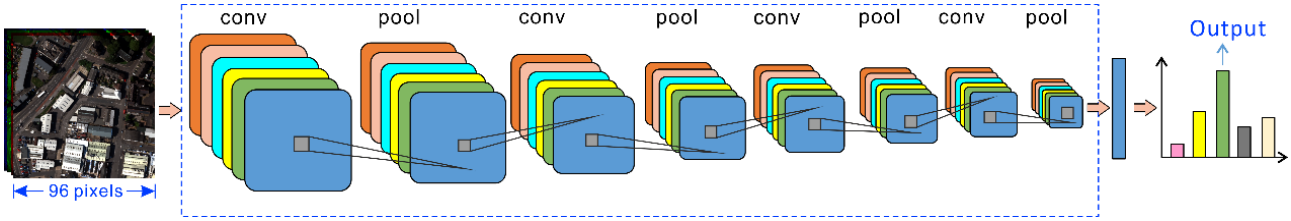


Figure 6-3: Model architectures and structures of the CNN with 96×96 input window size and eight-layer depth.

All the other parameters were optimised empirically based on 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.

6.3.2.4 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 with 16 nodes in each layer. Such consistency in parameter setting makes the baseline results comparable.

SVM: The SVM model involves a penalty value C and a kernel width σ that needs to be parameterised. Following the recommendation by Zhang et al. (2015), a grid search with 5-fold cross validation was implemented to search exhaustively within a wide parameter space (C and σ within $[2^{-10}, 2^{10}]$). Such parameter settings should lead to high validation accuracy using support vectors to formulate an optimal classification hyperplane.

MRF: The Markov Random Field, a spatial contextual classifier, was used 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 fixed size of neighbourhood window (7×7) and a parameter γ that controls the smoothness level was set as 0.7 to achieve an appropriate level of smoothness in the MRF. The simulated annealing

optimization approach with a Gibbs sampler (Berthod et al. 1996) was employed in the MRF to maximise the posterior probability through 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 degree of similarity 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 all pixels within the images using densely overlapping image patches. The most crucial parameters that influence directly the classification performance of the pixel-wise CNN are the input image patch size and the number of layers (depth). Following the discussion by Längkvist et al., (2016), the input image size was chosen from {28×28, 32×32, 36×36, 40×40, 44×44, 48×48, 52×52 and 56×56} to evaluate the influence of contextual area on classification performance. 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 plays a key role in classification accuracy because the quality of the learnt features is highly influenced by the level of abstraction and representation. As suggested by Chen et al. (2016), the number of CNN layers was chosen as six with three convolutional layers and three pooling layers to balance the network complexity and robustness. Other CNN parameters were tuned empirically through cross validation. For example, the filter size

was set to 3×3 for the convolutional layer with a stride of 1, and the number of filters was set to 24 to extract multiple convolutional features at each level. The learning rate was set as 0.01 and the number of epochs was chosen as 600 to fully learn the features through backpropagation.

6.3.3 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, including 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).

6.3.3.1 LC-LU JDL Classification Iteration

The proposed LC-LU JDL was implemented through iteration. For each iteration, the LC and LU classifications were implemented 10 times with 60% training and 40% testing sample sets split randomly using the Monte Carlo method, and the average overall accuracy (OA) was reported for each iteration. Figure 6-4 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% and 78.3% 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.

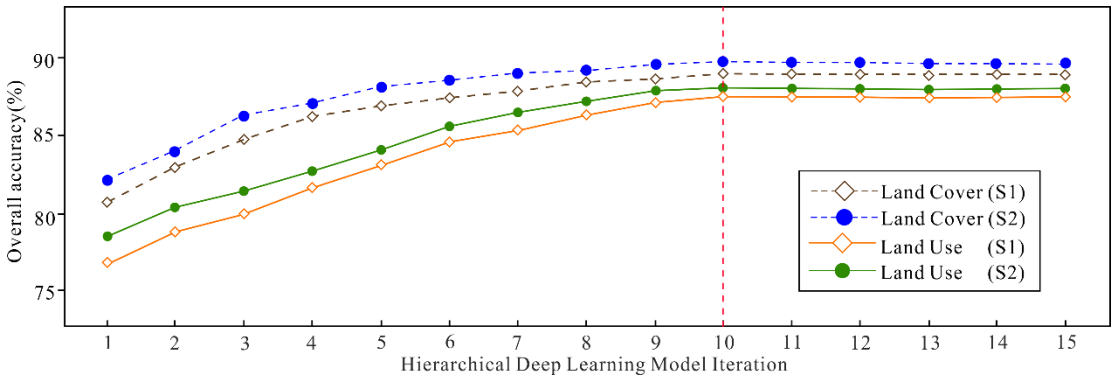


Figure 6-4: The overall accuracy curves for the Joint Deep Learning iteration of land cover (LC) and land use (LU) classification results in S1 and S2. The red dash line indicates the optimal accuracy for the LC and LU classification at iteration 10.

6.3.3.2 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 6-5 and 6-6, respectively, with the optimal classification outcome (at iteration 10) marked by blue boxes. In Figure 6-5, four subsets of S1 at different iterations (1, 2, 4, 6, 8, and 10) are presented to provide better visualisation, with red and yellow circles highlighting incorrect and correct 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 6-5(*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 6-5(*a*) and 6-5(*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 6-5(*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 6-5(*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 6-5(*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 6-5(*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 6-5(*b*)), but the final result appeared to be reasonably smooth with accurate characterisation of asphalt road and clay roof (yellow circles in Figure 6-5(*b*) of iteration 10). In addition, confusion between metal roof and concrete roof (iteration 1 – 8 with red circles in Figure 6-5(*c*)) was rectified step-by-step through iteration, with the entire building successfully classified as metal roof at iteration 10 (yellow circle in Figure 6-5(*c*)). Moreover, the crops within Figure 6-5(*d*) was smoothed gradually from

severe salt-and-pepper effects in iteration 1 (red circles in Figure 6-5(*d*)) to sufficiently smoothed representations in iteration 10 (yellow circle in Figure 6-5(*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 6-5(*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 6-5(*d*)).

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 6-4) and best visual appeal (Figure 6-6). 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 6-6(*c*)). Such problems were tackled with increasing iteration (Figure 6-6(*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 6-6(*c*)). Such shadow effects were gradually reduced in Figure 6-6(*d*-*g*) and completely eliminated in Figure 6-6(*h*) at iteration 10, which was highlighted by blue box as the best classification result in JDL-LC (Figure 6-6(*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.

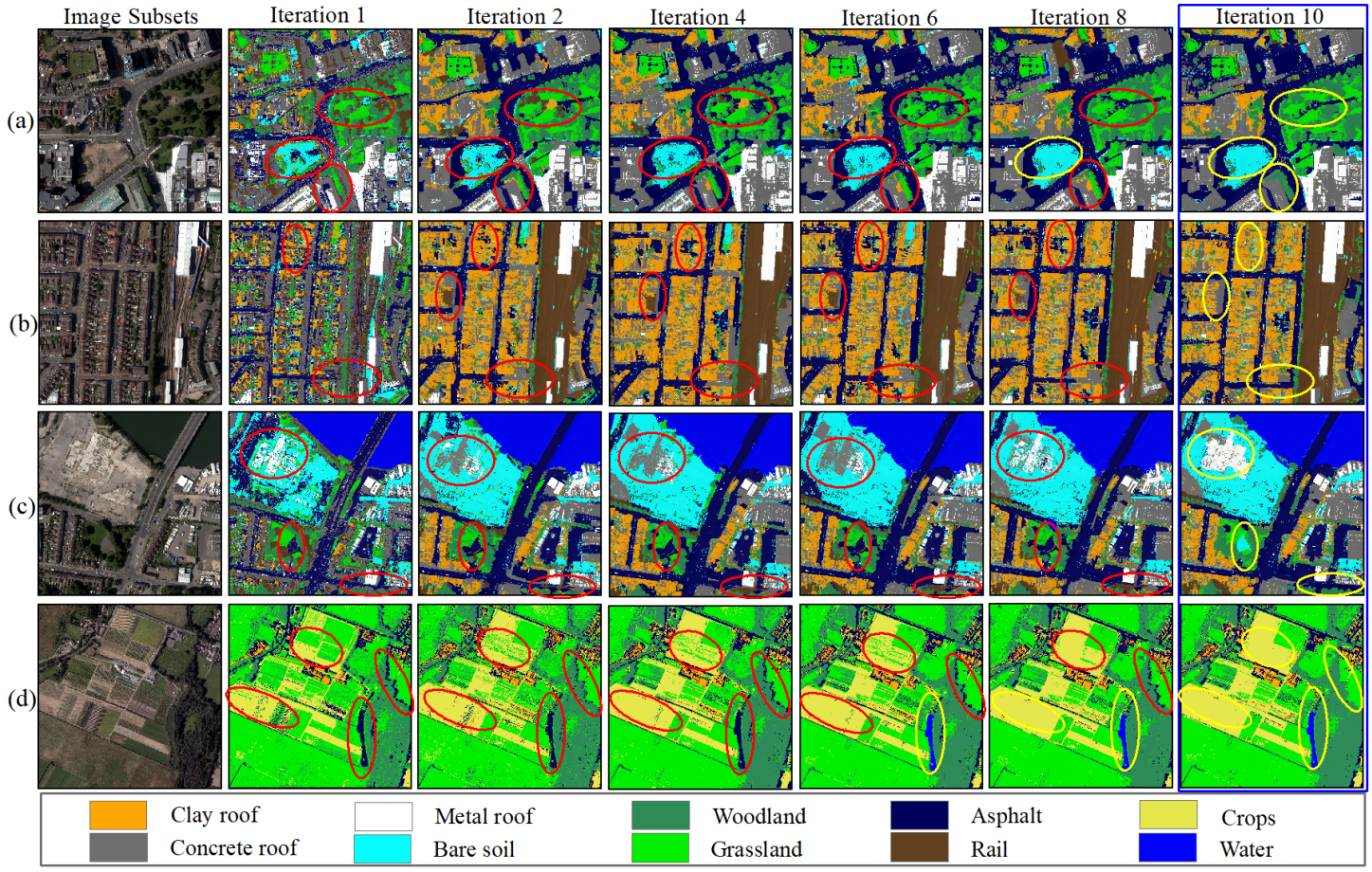


Figure 6-5: Four subset land cover classification results in S1 using Joint Deep Learning – Land cover (JDL-LC), the best results at iteration 10 were highlighted with blue box. The red and yellow circles denote incorrect and correct classification, respectively.

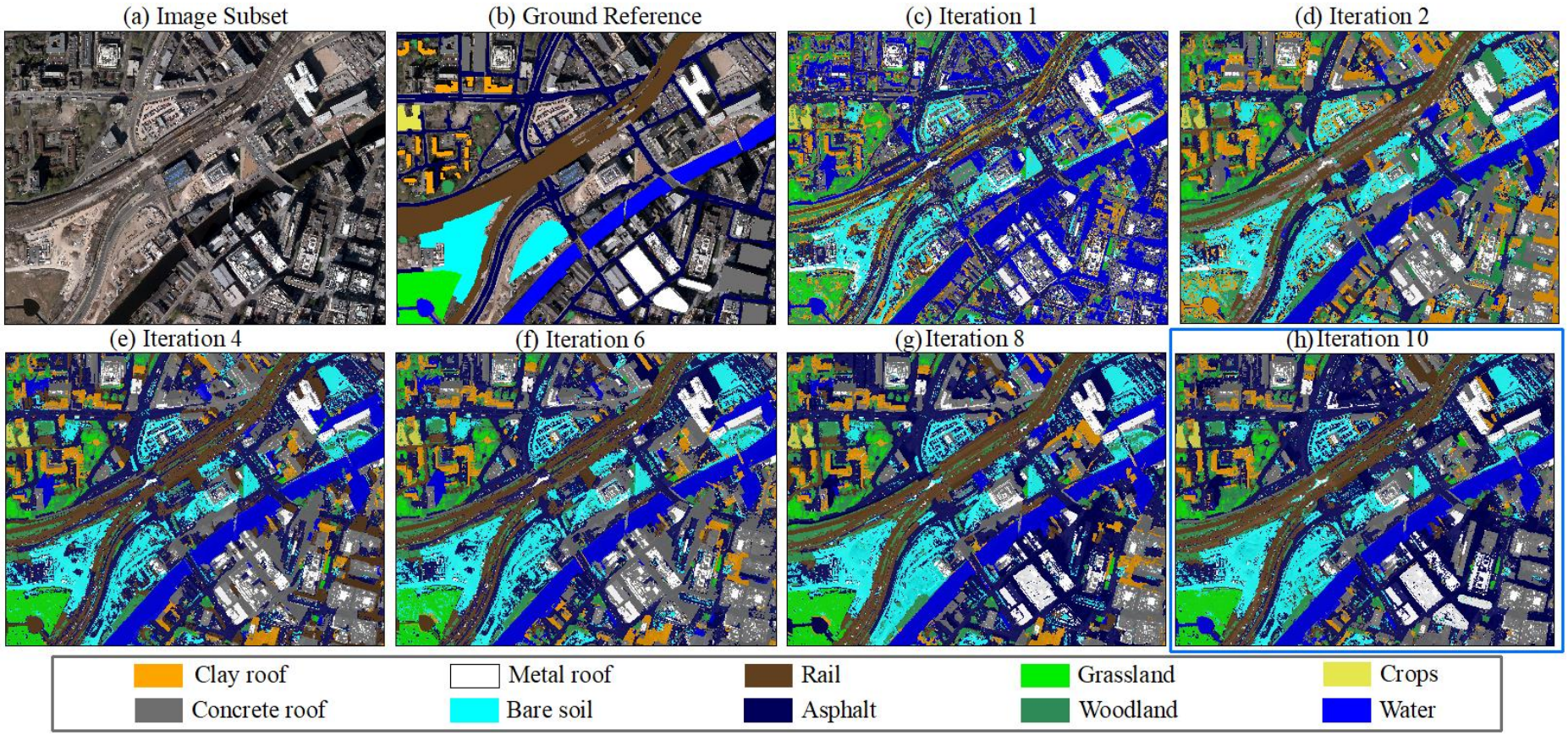


Figure 6-6: The land cover classification results in S2 using Joint Deep Learning – Land cover (JDL-LC), the best results at (h) iteration 10 were highlighted with blue box.

6.3.3.3 JDL – Land use (JDL-LU) classification Iteration

LU classifications from the JDL – Land use (JDL-LU) are demonstrated in Figures 6-7 and 6-8 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 6-4.

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 6-7(*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 6-7(*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 6-7(*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 instance, the misclassifications of residential and highway shown in iteration 1-6 (red circles in Figure 6-7(*b*)) were mostly rectified in iteration 8 and were completely distinguished in iteration 10 with high accuracy ((yellow circles

in Figure 6-7(*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 6-7(*a*)) were rectified in iteration 10 (yellow circle in Figure 6-7(*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 6-7(*a*) and 6-7(*c*)), were accurately removed either at iteration 8 (yellow circle in Figure 6-7(*a*)) or at iteration 10 (yellow circle in Figure 6-7(*c*)).

In S2, the iterative process also exhibits similar improvements with iteration. For example, the mixture of commercial areas and industrial areas in S2 (Figure 6-8(*c*)) was gradually reduced through the process (Figure 6-8(*d-g*)), and was surprisingly resolved at iteration 10 (Figure 6-8(*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 6-8(*c-g*), parts of the highway and redeveloped area were falsely identified as parking lot, but were accurately distinguished at iteration 10 (Figure 6-8(*h*)). Moreover, a narrow highway that was spatially adjacent to the railway, that was not identified at iteration 1 (Figure 6-8(*c*)), was identified at iteration 10 (Figure 6-8(*h*)), demonstrating the ability of the proposed JDL method to differentiate small linear features.

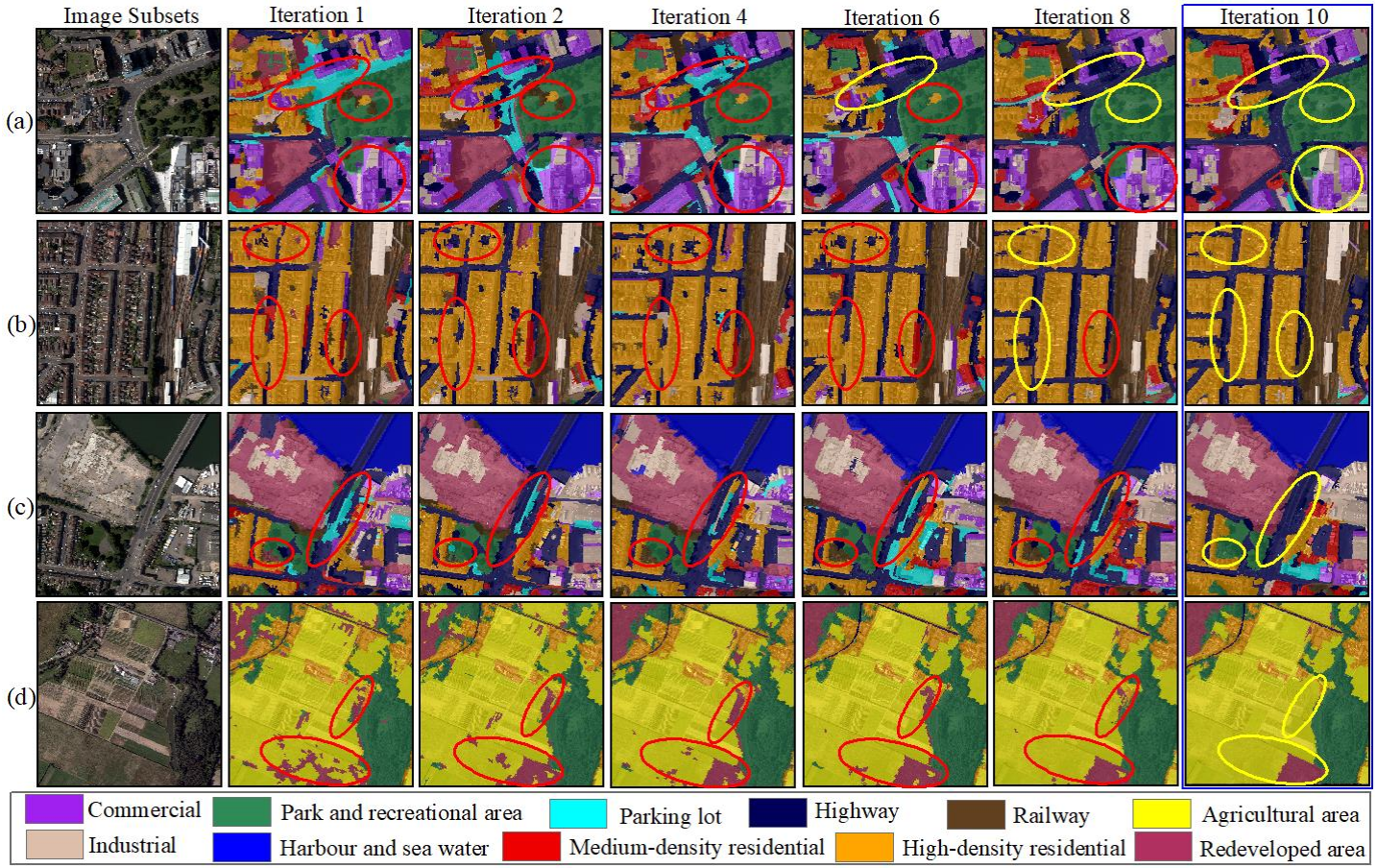


Figure 6-7: Four subset land use classification results in S1 using Joint Deep Learning – Land use (JDL-LU), the best results at iteration 10 were highlighted with blue box. The red and yellow circles denote incorrect and correct classification, respectively.

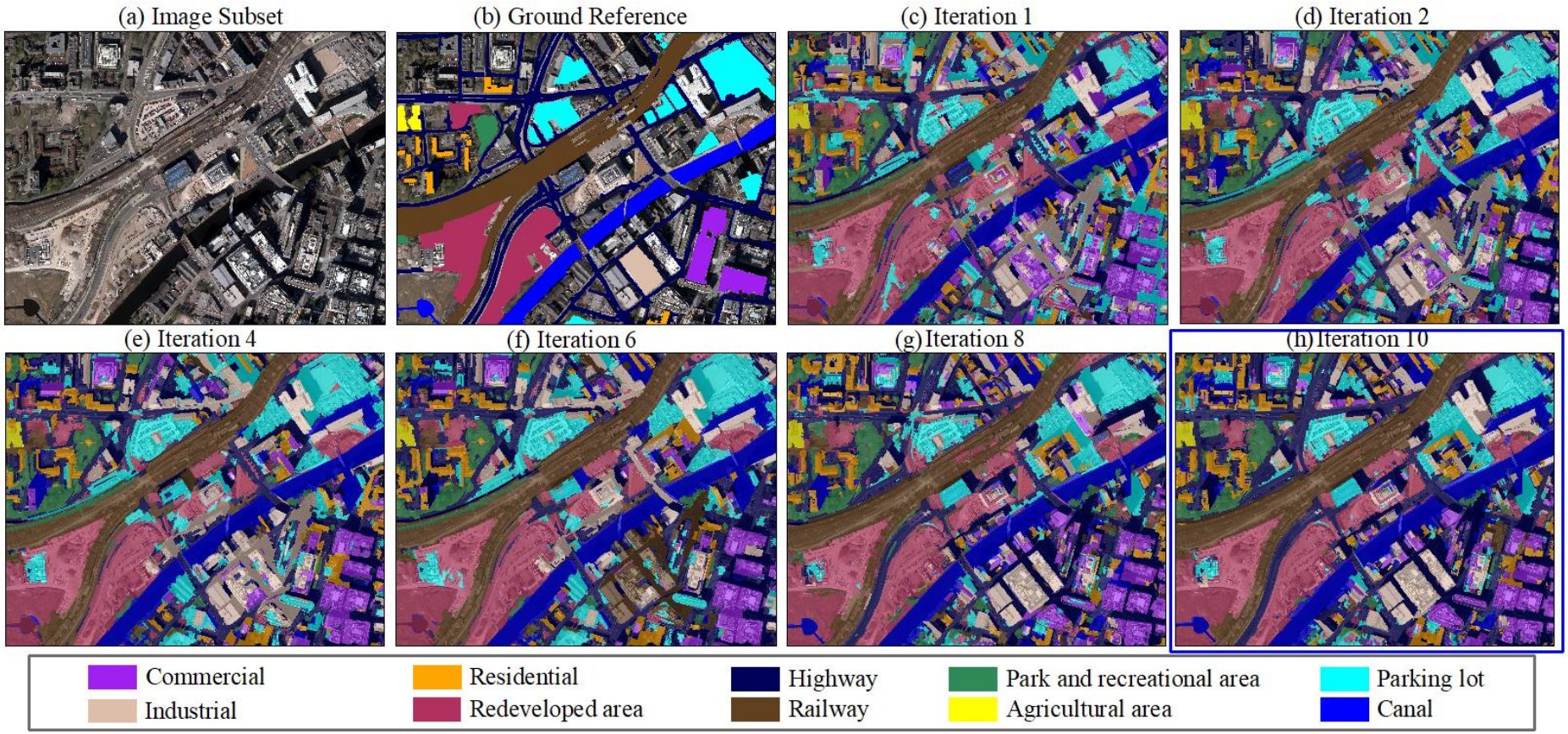


Figure 6-8: The land use classification results in S2 using Joint Deep Learning – Land use (JDL-LU), the best results at (h) iteration 10 were highlighted with blue box.

6.3.3.4 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 6-9(*a*) and Figure 6-9(*b*), respectively. As shown by Figure 6-9(*a*), the HDLJDL-LC achieved the largest OA of up to 89.72% and 90.76% for the S1 and S2, larger than the MRF of 84.88% and 84.46%, the SVM of 82.46% and 82.33%, and the MLP of 81.35% and 82.24%, respectively. For the LU classification in Figure 6-9(*b*), the proposed HDLJDL-LU achieved 87.63% and 88.39% for S1 and S2, higher than those of CNN (84.12% and 83.36%), OBIA-SVM (80.36% and 80.48%), and MRF (79.44% and 79.34%) respectively.

In addition to the OA, the proposed JDL method achieved consistently the smallest values for both Quantity and Allocation Disagreement, respectively. From Table 6-2 and 6-3, the JDL-LC has the smallest disagreement in terms of LC classification, with an average of 6.87% and 6.75% for S1 and S2 accordingly, which is far smaller than for any of the three benchmarks. Similar patterns were found in LU classification (Table 6-4 and 6-5), where the JDL-LU acquired the smallest average disagreement in S1 and S2 (9.94% and 9.14%), much smaller than for the MRF (20.28% and 19.08%), OBIA-SVM (18.55% and 16.77%), and CNN (14.20% and 13.96%).

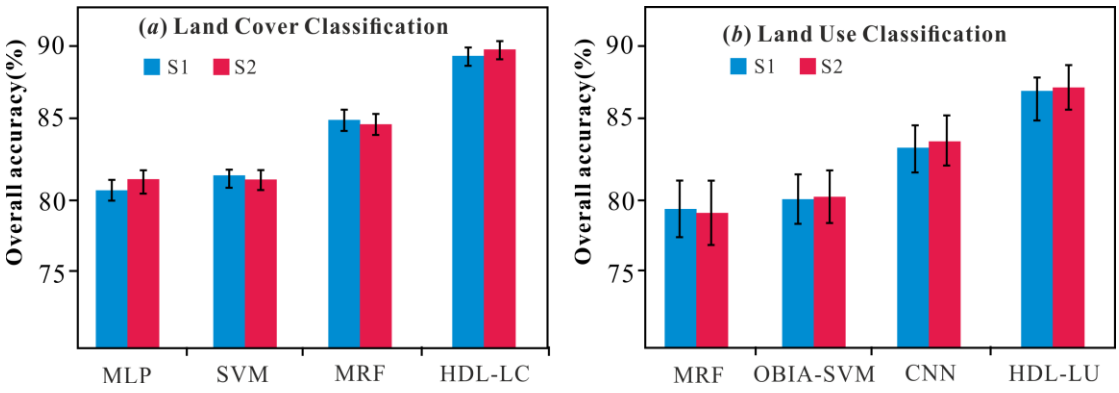


Figure 6-9: Overall accuracy comparisons among the MLP, SVM, MRF, and the proposed JDL-LC for land cover classification, and the MRF, OBIA-SVM, CNN, and the proposed JDL-LU for land use classification.

Per-class mapping accuracies of the two study sites (S1 and S2) were listed to provide detailed comparison of each LC (Table 6-2 and Table 6-3) and LU (Table 6-4 and Table 6-5) 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 6-2 and Table 6-3), the mapping accuracies of Clay roof, Metal roof, Grassland, Asphalt and Water are higher than 90%, with the greatest accuracy achieved in water at both S1 and S2, up to 98.37% and 98.42%, respectively. The most remarkable increase in accuracy can be seen in Grassland with an accuracy of up to 90.12% and 90.65%, respectively, much higher than for the other three benchmarks, including the MRF (75.62% and 75.42%), the SVM (73.23% and 73.59%), and the MLP (71.26% and 70.36%). Another significant increase in accuracy was found in Woodland through JDL-LC with the mapping accuracy of 88.43% (S1) and 88.24% (S2), dramatically higher than for the MRF of 76.09% and 75.39%, SVM of 70.28% and 70.16%, and MLP of 68.59% and 69.45%, respectively. Likewise, the Concrete roof also demonstrated an obvious increase in accuracy from just 69.43% and

70.54% classified by the MLP to 79.52% and 79.25% 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 5.25% and 5.46%, respectively. Other LC classes (e.g. 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.

With respect to the LU classification, the proposed JDL-LU achieved excellent classification accuracy for the majority of LU classes at both S1 (Table 6-4) and S2 (Table 6-5). 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 98.42% for Harbour and sea water, 98.74% for Canal, and an average of 95.84% 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 12.37% and 11.61% for the commercial areas, 17.47% and 10.74% for industrial areas, and 13.74% and 12.39% 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.

Table 6-2 - 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	89.52%	89.45%	89.14%	**92.43%**
Concrete roof	69.43%	69.82%	73.27%	**79.52%**
Metal roof	90.28%	90.93%	90.23%	**91.65%**
Woodland	68.59%	70.28%	76.09%	**88.43%**
Grassland	71.26%	73.23%	75.62%	**90.12%**
Asphalt	88.54%	88.37%	89.46%	**91.24%**
Rail	82.18%	82.35%	83.58%	**87.29%**
Bare soil	80.07%	80.15%	82.57%	**85.64%**
Crops	84.28%	84.75%	86.52%	**89.58%**
Water	97.32%	97.43%	98.48%	**98.62%**
Overall Accuracy (OA)	81.35%	82.46%	84.88%	**89.72%**
Quantity Disagreement	17.15%	16.88%	11.26%	**7.56%**
Allocation Disagreement	16.23%	16.34%	13.42%	**6.18%**

Table 6-3 - 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	90.12%	90.28%	89.58%	**92.87%**
Concrete roof	70.54%	70.43%	74.23%	**79.25%**
Metal roof	90.17%	90.91%	90.02%	**91.34%**
Woodland	69.45%	70.16%	75.39%	**88.24%**
Grassland	72.36%	73.59%	75.42%	**90.65%**
Asphalt	89.42%	89.58%	89.45%	**91.68%**
Rail	83.21%	83.15%	84.26%	**88.54%**
Bare soil	80.23%	80.34%	82.27%	**85.59%**
Crops	85.04%	85.32%	87.86%	**90.74%**
Water	97.58%	97.23%	98.07%	**98.37%**
Overall Accuracy (OA)	82.24%	82.33%	84.46%	**90.76%**
Quantity Disagreement	16.28%	16.37%	11.36%	**7.26%**
Allocation Disagreement	15.76%	15.89%	12.18%	**6.25%**

Table 6-4 - 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	70.09%	72.87%	73.26%	**82.46%**
Highway	77.23%	78.04%	76.12%	**79.69%**
Industrial	67.28%	69.01%	71.23%	**84.75%**
High-density residential	81.52%	80.59%	80.05%	**86.43%**
Medium-density residential	82.74%	84.42%	85.27%	**88.59%**
Park and recreational area	91.05%	93.14%	92.34%	**97.09%**
Agricultural area	85.07%	88.59%	87.42%	**90.96%**
Parking lot	78.09%	80.17%	83.76%	**91.83%**
Railway	88.07%	90.65%	86.57%	**91.92%**
Redeveloped area	89.13%	90.02%	89.26%	**90.69%**
Harbour and sea water	97.39%	98.43%	98.54%	**98.42%**
Overall Accuracy (OA)	79.44%	80.36%	84.12%	**87.63%**
Quantity Disagreement	20.64%	18.32%	14.36%	**10.26%**
Allocation Disagreement	19.92%	18.78%	14.05%	**9.62%**

Table 6-5 - 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	71.11%	72.47%	74.16%	**82.72%**
Highway	81.43%	79.26%	80.59%	**84.37%**
Industrial	72.52%	72.05%	74.84%	**83.26%**
Residential	78.41%	80.45%	80.56%	**84.99%**
Parking lot	79.63%	82.06%	84.37%	**92.02%**
Railway	85.94%	88.14%	88.32%	**91.48%**
Park and recreational area	88.42%	89.54%	90.76%	**94.59%**
Agricultural area	84.64%	87.13%	86.58%	**91.42%**
Redeveloped area	82.57%	84.15%	87.04%	**93.75%**
Canal	90.63%	92.28%	94.18%	**98.74%**
Overall Accuracy (OA)	79.34%	80.48%	83.36%	**88.39%**
Quantity Disagreement	19.42%	17.03%	14.28%	**9.82%**
Allocation Disagreement	18.74%	16.52%	13.65%	**8.46%**

6.3.3.5 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 sample sizes of 10%, 30%, and 50% (Figure 6-10). Similar patterns in reduction in accuracy as a function of sample size reduction were observed for S1 and S2. From Figure 6-10, 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 6-10).

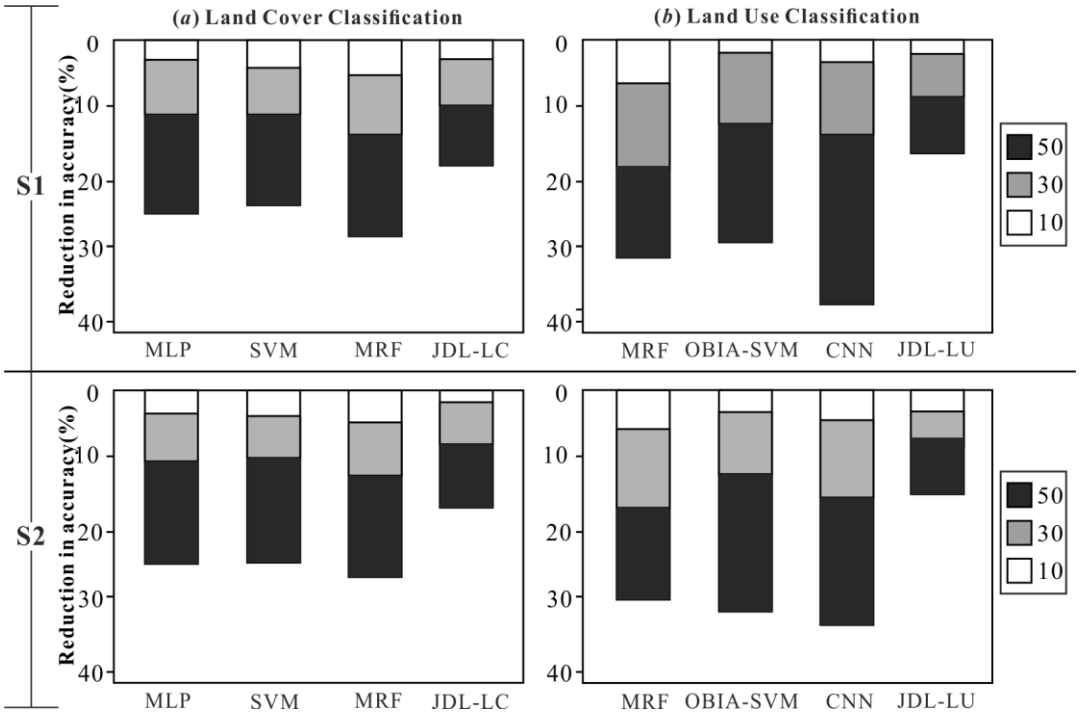


Figure 6-10: The effect of reducing sample size (50%, 30%, and 10% of the total amount of samples) on the accuracy of (a) land cover classification (JDL-LC) and (b) land use classification (JDL-LU), and their respective benchmark comparators at study sites S1 and S2.

For the LC classification (Figure 6-10(*a*)), the accuracy distributions of the MLP and SVM were similar, although the SVM was slightly less sensitive to sample size reduction, with about 2% higher accuracy with a 50% reduction than for the MLP. The MRF was the most sensitive method to LC sample reduction, with decreases of up to 30% and 28% in accuracy for S1 and S2, respectively. The JDL-LC was the least sensitive to a reduction in training sample size, with less than 10% accuracy reduction for 30% reduced sample size and less than 20% decreased accuracy for 50% sample size reduction, outperforming the benchmarks in terms of model robustness.

In terms of the LU classification (Figure 6-10(*b*)), the CNN was most sensitive to sample size reduction, particularly the 50% sample size reduction, where significantly decreased accuracy was observed (with 40% and 32% decreases in accuracy in S1 and S2, respectively). MRF and OBIA-SVM were less sensitive to sample size reduction than the CNN, with around a 30% decrease in accuracy while reducing the sample size to 50%. The JDL-LU, however, demonstrated the most stable performance with respect to sample size reduction, with less than a 20% decrease in accuracy when 50% of the training samples were used.

6.4 Discussion

This chapter 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.

6.4.1 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 6-11(*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 6-4), 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).

6.4.2 Mutual Benefit for 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 efficiency to separate different types of LCs (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.

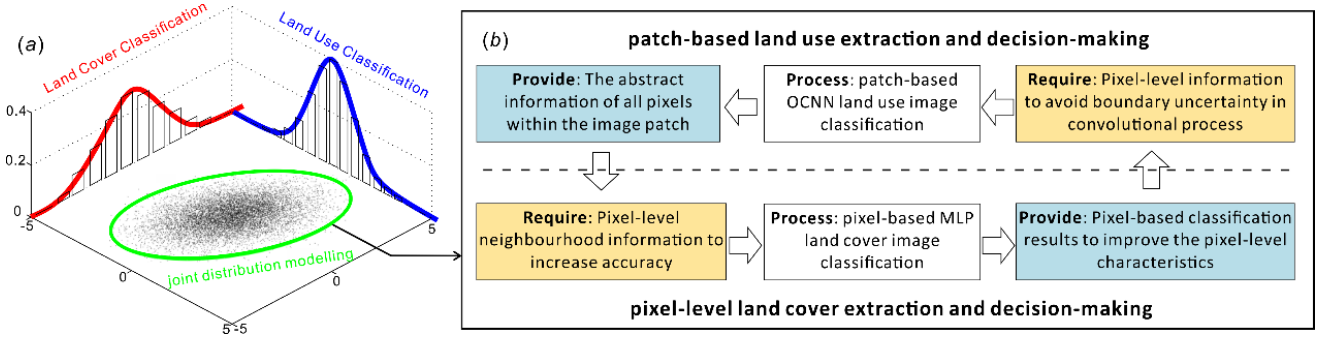


Figure 6-11: Joint deep learning with joint distribution modelling (a) through iterative process for pixel-level land cover (LC) and patch-based land use (LU) extraction and decision-making (b).

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 6-11(*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 6-8).

From an AI 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 6-10), 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.

6.5 Conclusion

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 chapter, 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. This research paves the way to effectively address 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. 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.

Chapter 7 Discussion and Conclusion

Fully automated land cover (LC) and land use (LU) classification of remotely sensed imagery, especially VFSR imagery, is an extremely challenging task that is constantly pushing the envelope of AI and machine learning. The major challenges in classifying these VFSR remotely sensed images are their spectral and spatial complexity due to the increased intra-class variation and the reduced inter-class disparity, relative to fixed objects of interest, coupled with varied illumination conditions and shadow that interact between adjacent ground objects. Indeed, existing techniques remain inadequate to analyse VFSR data effectively and efficiently, which calls for the development of advanced methodologies to accelerate innovation in fully automatic image classification.

Deep learning, as a new frontier of AI, holds great promise to fulfil the challenging needs of VFSR remotely sensed image classification. The idea of deep learning methods is to perform human-like reasoning and to extract high-level and abstract features that represent the semantics of input images. The process is greatly inspired by human visual cognition, in which hierarchical structures are learnt through multiple levels of feature representations. Typically, the high-level features (e.g. patterns and associations) are composed of a set of low-level characteristics (e.g. edges, textures, and shape), and are often indirectly correlated to the recorded spectral reflectance. Deep learning based methods have strong capabilities to characterise and differentiate the semantics of LC and LU by extracting the most expressive and discriminative features end-to-end, hierarchically. Therefore, introducing deep learning methods has great potential to obtain better feature representations in identifying unique characteristics of

VFSR imagery, and to keep pace with the advances in sensor technologies through automation.

This thesis developed a set of novel deep learning methods for land cover and land use classification using VFSR remotely sensed imagery. The mapping objective started with land cover classification (e.g. buildings, asphalt, grassland, woodland etc.) and reached complex land use classification (e.g. residential, commercial, industrial etc.). These challenging land features were extracted and classified by innovating deep learning methods, in particular, convolutional neural networks (CNN) as an example (i.e., developing entirely new methods based on the CNN). Experimental results demonstrated several difficulties in applying the standard pixel-wise CNN for remotely sensed image classification, including geometric distortions, boundary uncertainties and huge computational redundancy. Essentially, the problem of producing LC/LU thematic maps using patch-based CNN networks is an inherent tension between "what" from deep layers as semantics and "where" from shallow layer to provide the details, and there is a strong trade-off between high-level semantic recognition and low-level boundary delineation (Sun and Wang 2018). These technological and technical challenges were addressed systematically in this thesis through methodological innovations. In addition, the learning process was further extended into a hierarchical and iterative procedure rather than finding a one-off solution. This is similar to human interpretations on aerial imagery using a wide range of real-world knowledge and expertise through repetition. In summary, the major contributions of this thesis involve:

(1) Developing novel deep learning methods or architectures as a learning process towards human interpretation on VFSR remotely sensed imagery;

(2) Solving complex LC and LU image classification with a set of innovations through technological integrations and model iterations;

(3) Understanding the classification hierarchies and their intrinsic relationships across different mapping objectives, from simple land covers to complex land use feature representations.

This chapter discusses the key research findings from the four published papers in chapter 3 – 6 by answering the raised research questions, followed by future recommendations.

7.1 Research Findings and Conclusions

This section presents the research findings and conclusions with respect to each research objective as described in section 1.5.

- **Objective 1: Develop a deep learning method for land cover classification using VFSR remotely sensed images.**

  A hybrid MLP-CNN method was proposed for land cover classification using VFSR images. The MLP-CNN was designed to integrate the contextual-based convolutional neural network (CNN) with deep architecture and pixel-based multilayer perceptron (MLP) with shallow structure using a rule-based fusion decision strategy. The decision fusion rules, designed primarily based on the classification confidence of the CNN, reflect the generally complementary patterns of the individual classifiers with different behaviours, in which the CNN based on deep spatial feature representation and the MLP based on spectral discrimination were integrated to differentiate land cover objects from VFSR remotely sensed imagery. The effectiveness of the ensemble MLP-CNN classifier was tested in both

urban and rural areas using aerial photography and the WorldView-2 satellite imagery with sub-metre spatial resolution. The MLP-CNN classifier achieved promising performance, consistently outperforming the pixel-based MLP, spectral and textural-based MLP, and the contextual-based CNN in terms of classification accuracy.

This study concluded that the fusion of the patch-based CNN and the pixel-based MLP is an effective solution for land cover classification using VFSR remotely sensed imagery. The complementarity acquired by the CNN for deep spatial feature representation and MLP in spectral discrimination can rectify the blurred boundaries and the loss of fine spatial details reduced through the convolutional process using the fusion decision strategy. This research provides an effective solution to well balance the trade-off between feature recognition and localisation, and paves the way to address complex land cover image classification tasks using VFSR imagery.

- **Objective 2: Model the uncertainty in deep learning for VFSR land cover image classification.**

  A variable precision rough set (VPRS) model was proposed to quantify the uncertainties in the deep learning based method, and a spatially explicit regional decision fusion strategy was introduced to further improve the CNN-based VFSR image classification through the fusion of a Markov random field (MRF). The VPRS model, based on rough set theory, was developed to partition the classification confidence map derived from CNN-based classifications into the positive regions (correct classifications) and the non-positive regions (uncertain or incorrect classifications). Those "more correct" areas were trusted by the CNN, whereas the uncertain areas were rectified by a multi-layer perceptron (MLP)-based

Markov random field (MLP-MRF) classifier to provide crisp and accurate boundary delineation. The proposed MRF-CNN fusion decision strategy exploited the complementary characteristics of the two classifiers based on VPRS uncertainty description and classification integration. The effectiveness of the MRF-CNN method was tested in both urban and rural areas of southern England as well as with Semantic Labelling datasets. The MRF-CNN consistently outperformed the standard pixel-based MLP and SVM, spectral-contextual based MLP-MRF as well as contextual-based CNN classifiers, and state-of-the-art baseline methods for semantic segmentation.

This study concluded that the proposed VPRS-based regional fusion decision between CNN and MRF was an effective framework for land cover classification using VFSR remotely sensed imagery. The VPRS model quantified the uncertainties in CNN classification of VFSR imagery by partitioning the result into spatially explicit granularities that represent positive regions and non-positive regions, respectively. The positive regions were trusted directly by the CNN, whereas non-positive regions were rectified by the MLP-MRF in consideration of their complementary behaviour in spatial representations. The proposed regional fusion of MRF-CNN classifiers achieved the highest classification accuracy compared with the benchmark approaches. Therefore, this VPRS-based uncertainty description and classification integration between CNN and MRF provides a general framework to achieve fully automatic and effective VFSR land cover image classification.

- **Objective 3: Develop a deep learning method to solve the complex land use classification using VFSR remotely sensed imagery.**

A novel object-based convolutional neural network (OCNN) was proposed for urban land use classification using VFSR images. Rather than pixel-wise convolutional processes, the OCNN relies on segmented objects as its functional units, and CNN networks are applied to characterise objects and their spatial context by using within-object and between-object information. Specifically, two CNN networks with different model structures (six layers and eight layers) and different window sizes (48×48 and 128×128) were developed to predict linearly shaped objects (e.g. Highway, Railway, Canal) and general objects (other non-linearly shaped). Multiple small window size CNNs were sampled along each object based on geometric characteristics, and integrated through statistical majority voting, whereas the large window size CNN was used only once per object for prediction using a wide spatial context. A rule-based decision fusion was designed to integrate the class-specific classification results conditional upon these two CNN models, in which the prediction of a linearly shaped object from the small window size CNNs was given priority, whereas the large window size CNN was trusted in any other cases. The effectiveness of the proposed OCNN method was tested on aerial photography of two large urban scenes in Southampton and Manchester of Great Britain for land use image classification. The OCNN combined with large and small window sizes achieved excellent classification accuracy and computational efficiency, constantly outperforming its sub-modules, as well as other benchmark comparators, including the pixel-wise CNN, contextual-based MRF and object-based OBIA-SVM.

This study concluded that the object-based CNN (OCNN) method was an effective solution for complex land use classification using VFSR imagery. The proposed OCNN method with two CNN networks is designed to sample specific locations

that are defined by size and geometry of image objects, and integrate them in a class-specific manner to obtain an effective and efficient urban land use classification output (i.e., a thematic map). This OCNN method with large and small window size CNNs produced the most accurate classification results in comparison with the sub-modules and other contextual-based and object-based benchmark methods. Moreover, a high computational efficiency was achieved with much more acceptable time requirements than the standard pixel-wise CNN method in the process of model inference. Therefore, the proposed OCNN method is effective *and* efficient in urban land use classification using VFSR imagery with great potential for a broad range of applications.

- **Objective 4: Develop a novel method for joint land cover and land use classification using VFSR remotely sensed imagery.**

  A novel MLP-CNN based Joint Deep Learning (JDL) method that incorporates multilayer perceptron (MLP) and deep convolutional neutral networks (CNN), was proposed for joint land cover (LC) and land use (LU) classifications through iteration. Specifically, LU classifications conducted by the patch-based CNNs for object characterisations were made conditional upon the land cover probabilities derived from the pixel-based MLP using the original imagery. Then the land use probabilities inferred by the CNN together with the original image were re-used as inputs to the MLP to strengthen the spatial and spectral feature representations. Such an iterative process between the MLP and CNN is formulated as a Markov process through joint distribution modelling, in which both the LC and LU are classified simultaneously through an iterative procedure. The effectiveness of the proposed MLP and CNN JDL method was tested on aerial photography of two large urban and suburban scenes (Southampton and Manchester) in Great Britain. The

JDL demonstrated a consistently increasing trend in accuracies for both LC and LU classifications, achieving the best classification accuracies at iteration 10, with the average overall accuracies up to 90.24% for LC and 88.01% for LU for the two study sites, constantly outperforming various benchmark comparators for LC and LU classifications. In particular, complex land cover classes cast by shadows that are extremely difficult to be handled were distinguished precisely, and the complex land use patterns (e.g. parking lots) were recognised accurately.

This study concluded that the MLP-CNN JDL is an effective method to address the complex land cover and land use classification tasks using VFSR remotely sensed imagery in a joint and automatic manner. The proposed method provides a general framework, within which the pixel-based MLP and the patch-based CNN models were mutually complementary between the pixel-level and neighbourhood characteristics, refining each other throughout the classification process with iteration. To the best of our knowledge, this joint LC and LU classification is the first research to jointly model the spatial and hierarchical relationships between land cover and land use, in which pixel-level and neighbourhood information was interacted and updated iteratively, enabling the accurate classifications of LC and LU at both levels simultaneously.

7.2 Reflections

This paper-based thesis is composed of four peer-reviewed journal papers in Chapters 3 – 6. Although the research objectives of these four chapters are completely different, they exhibit strong links among each other and form a logical story. Essentially, the logic in this thesis is from simple land cover classification to complex land use

classification, from pixel-based methods to object-based methods, and from single end-to-end model to joint models through iteration. Detailed links are presented as follows:

Chapter 3 and 4 are strongly linked for solving the land cover classification tasks using VFSR remotely sensed imagery. The fundamental challenge for adapting deep learning based methods for land cover classification is the dilemma between pixel-level spectral differentiation and the patch-based spatial feature representations, where the standard pixel-wise CNNs pose several challenges in land cover classification, including geometric distortions, boundary uncertainties and the loss of fine spatial details through convolutional processes. Chapter 3 proposed to create rules to threshold the classification results and deal with uncertainties through a fusion decision strategy, in which the MLP and CNN were fused to harvest their complementarity between spectral differentiation and spatial feature characterisation. This method, although having potential to achieve desirable classification results, involves a large amount of trial and error and prior knowledge of feature characteristics and, thus, was hard to generalise and apply in an automatic fashion. Chapter 4 proposed a VPRS-based approach to model and partition the uncertainty within the CNN-based land cover classification map automatically. Those uncertain regions are further improved by MLP-MRF for spectral differentiation and boundary segmentation. Such regional fusion decision approaches provide a general framework within which to gain the advantages of the model-based CNN, while overcoming the problem of losing effective resolution and uncertain prediction at object boundaries, which is especially pertinent for complex VFSR image classification.

While Chapters 3 and 4 extracted ground objects with physical properties (i.e. land cover), Chapter 5 focused on predicting land use using higher-order feature

representations, which is closely related to Chapter 6 for joint land cover and land use classification. In Chapter 5, an object-based CNN method was developed to learn both within-object and between-object information, such that the spatial and hierarchical relationships are used to characterise urban land use. A range of small window size CNNs were designed to recognise linearly shaped objects that are often misclassified or ignored by large window size CNNs, and the two contextual windows are used at different scales to represent the classes that are linearly shaped and other general land use classes with complex patterns. Chapter 6, on the other hand, aimed to learn the spatial hierarchies between land cover and land use through iteration, and to jointly classify LC and LU simultaneously. The conditional dependencies between the two classification levels were learnt through the joint deep learning model, and refined each other with mutual complementarity and joint improvement. Such iterative processes are able to use existing knowledge, recall past experience, and consider context and physical phenomena, which are similar to the learning processes for human visual interpretation. Essentially deep learning aims to build a "machine" that can successfully perform (virtually) any tasks a human can. Although we are arguably far from creating such a machine, the iterative process proposed here could potentially mimic the human operators to perform visual interpretation through repetition.

There are clear linkages between each of the analytical chapter and they have progressed from each other significantly. One of the key advances along the chapters is the mitigation of shadow artefacts within aerial images. The shadow appears extensively on images, directly affecting the performance of LC and LU classification. In Chapter 3 (Figure 3-7 and 3-8) and Chapter 4 (Figure 4-5 and 4-6), the shadow was included as a specific land cover class, which was completely eliminated in Chapter 5 (Figure 5-8 and 5-9) and Chapter 6 (Figure 6-5, 6-6, 6-7, and 6-8). Such improvement

is non-trivial for both scientific contribution as the shadow artefacts are huge challenges for LC/LU classification to be dealt with, and for practical applications within Ordnance Survey, in which shadow is of less interest for practitioners and map users. Besides, the study extents in Chapter 3 (pilot study sites in Southampton, UK) and Chapter 4 (case studies in Bournemouth, UK) were relatively small, which were designed to proof the concept for adapting the deep learning based methods for LC classification. While in Chapter 5, large urban scenes of Great Britain, including City of Southampton and Great Manchester, are used for testing the proposed OCNN method for LU classification. These study areas are further expanded to include rural areas (agricultural and crop areas) in Chapter 6 for comprehensive LC and LU classification through the novel JDL method. Moreover, the technical skills (e.g. programming) progress throughout the chapters as a learning process. An old Deep Learning Toolbox based on MATLAB platform was used in Chapter 3 and 4 for LC classifications, which is currently outdated and no longer maintained. The deep learning methods in Chapter 5 and 6 were developed by using the state-of-the-art Keras backend with Tensorflow under Python platform, and coupled with GPU, the methods have the capabilities to process large scenes of imagery through parallel computing. And most importantly, the research focus was shifted from pixel-based LC classification (Chapter 3 and 4) to higher-order LU representations (Chapter 5), where the CNN is considered as the most appropriate approach for higher-order LU classification (up to 90% accuracy), and such hierarchical representations between LC and LU were further modelled in JDL through joint distribution modelling (Chapter 6).

In summary, the developed deep learning based methods represent by far the most accurate LC and LU classification, with approximately 90% accuracy, ~5% improvement over the existing deep learning methods, and ~10% higher than the

traditional methods as listed in Table 2-2. Such highly accurate classification techniques make a significant contribution to the field of LC/LU classification, and have great potential for a wide range of geospatial applications.

7.3 Recommendations

This thesis developed deep learning methods for land cover and land use classification using VFSR remotely sensed imagery. Many aspects of the proposed methods have not yet been fully addressed and explored, and further investigations are needed. In particular, the major limitation of this research is the lack of testing on transferability of the developed approaches. Within the thesis, only Chapter 4 tested the transferability using the Vaihingen and Potsdam semantic segmentation datasets, where the training was conducted at some annotated tiles and the testing was done on the rest of the image tiles that were not used for training purpose. However, the majority of research here took the training and testing samples from the same study area, and the methods were only validated at specific regions without transferring to other unseen regions. Future work would, therefore, devote to testing the transferability of the methods at wider geographical areas, in order to ensure the developed techniques highly robust, transferable and scalable. These techniques will be prototyped and integrated into the commercial image processing pipeline, which will set out the route towards developing an operational system in collaboration with Ordnance Survey.

Apart from developing transferable systems discussed beforehand, the proposed deep learning methods could be developed further at multiple perspectives, including data sources, techniques and applications. Detailed recommendations are made as follows:

1. *Data Sources*

The thesis focused on land cover and land use classification using VFSR imagery. However, many other data sources exist in the field of remote sensing, such as hyperspectral, synthetic aperture radar (SAR), and LiDAR, which have not been used in this research, and the fusion of these multiple data sources by designing novel deep architectures would be a potential future direction. In addition, the urban land use classification in this thesis was undertaken at a generalised spatial and semantic level (e.g., residential, commercial and industrial area), without identifying smaller functional sites (e.g., supermarkets, hospitals and playgrounds etc.). This issue might be addressed by incorporating multi-source geospatial data, for example, those classified commercial areas might be further differentiated as supermarkets, retail outlets, and café areas through indoor human activities. Future research will, therefore, mine the semantic information from GPS trajectories, transportation networks and social media data to characterise these smaller functional units in a hierarchical way, as well as socioeconomic activities and population dynamics.

2. *Techniques*

Many techniques proposed in this thesis could be developed further into real applications. For example, Chapter 5 proposed two-scale representations for urban land use, which might be insufficient to characterise some complex geometric characteristics. Land use features extracted in urban areas are essentially scale-dependent, and manifest across different scales. Therefore, a range of CNNs with different input patch sizes will be adopted in the future to adapt to the diverse sizes and shapes of urban objects through weighted decision fusion. In addition, Chapter 6 considered only the classification of land cover and land use. Such a two-level classification system could be extended to a structured classification hierarchy, from

land cover to land use and to higher order functions. Such hierarchical representations require to develop the Joint Deep Learning framework further to incorporate more deep learning models within the iterative process, and might strengthen each other through active learning and optimisation. Moreover, deep learning methods developed in this research provide point estimates (fixed values) for land cover and land use predictions. Hence, future work will refer to probabilistic methods (in a Bayesian interpretation) to estimate uncertainty associated with the predictions of these models. Chapter 4 explored the uncertainty modelling based on rough set theory, which could be developed further to incorporate prior probabilities under a Bayesian inference framework for uncertainty quantification.

3. *Applications*

There are many potential applications to demonstrate the developed techniques with real-world impact. One possible direction is to explore land cover and land use change detections in fast-growing urban environments, in which many stakeholders, including government, local authorities, and small and medium enterprises need to support their decision-making. Other deep learning techniques, such as recurrent neural networks (RNN) and long short term memory (LSTM), can be developed to model the land cover and land use dynamics through time-series analysis. Additionally, the land use in this research focused on urban areas such as residential, commercial, and industrial etc.. Future work will develop the deep learning models for rural areas and agricultural land uses, to differentiate functional uses between large-scale and small-holder agriculture through higher-order feature representations.

7.4 Conclusions

This thesis developed a set of novel deep learning methods for land cover and land use classification. Both land cover (e.g. buildings, asphalt, grassland, woodland etc.) and land use objects (e.g. residential, commercial, industrial etc.) were classified and mapped with high accuracy and computational efficiency. Below are the main conclusions emerging from this research.

- Land cover classification was solved by deep learning methods through a rule-based decision fusion and a rough set based uncertainty modelling, in which technical challenges in geometric distortions and boundary uncertainties were successfully tackled with innovations.

- Land use classification was addressed by developing a novel object-based convolutional neural networks (OCNN), where a group of pixels was used as an object for target sampling, and further incorporated the spatial context with high accuracy and computational efficiency. Substantial progress in developing a prototype method for extracting urban land use information from remotely sensed data has been made to-date.

- Joint Deep Learning (JDL) was proposed for joint land cover and land use classification to learn a hierarchical representation through iteration. This JDL framework is the major solution of this research, and has great potential to be the most important contribution to land use and land cover classification over the last decade.

References

Ahmed, O. S., Franklin, S. E., Wulder, M. A., and White, J. C., 2015. Characterizing stand-level forest canopy cover and height using Landsat time series, samples of airborne LiDAR, and the Random Forest algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 101, 89–101.

Albanese, A., Pal, S. K., and Petrosino, A., 2014. Rough sets, kernel set, and spatiotemporal outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 26 (1), 194–207.

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., and Friedl, L., 2017. Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-Spatial Information Science*, 20 (2), 77–96.

Ardila, J. P., Tolpekin, V. A., Bijker, W., and Stein, A., 2011. Markov-random-field-based super-resolution mapping for identification of urban trees in VHR images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (6), 762–775.

Arel, I., Rose, D. C., and Karnowski, T. P., 2010. Deep machine learning - A new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine*, 5 (4), 13–18.

Arief, H., Strand, G.-H., Tveite, H., and Indahl, U., 2018. Land Cover Segmentation of Airborne LiDAR Data Using Stochastic Atrous Network. *Remote Sensing* [online], 10 (6), 973. Available from: http://www.mdpi.com/2072-4292/10/6/973.

Atkinson, P. M. and Tatnall, A. R. L., 1997. Introduction Neural networks in remote sensing. *International Journal of Remote Sensing* [online], 18 (4), 699–709.

Available from: http://dx.doi.org/10.1080/014311697218700%5Cnhttp://www.tandfonline.com.www.sndl1.arn.dz/doi/abs/10.1080/014311697218700.

Attarchi, S. and Gloaguen, R., 2014. Classifying complex mountainous forests with L-Band SAR and landsat data integration: A comparison among different machine learning methods in the Hyrcanian forest. *Remote Sensing*, 6 (5), 3624–3647.

Audebert, N., Le Saux, B., and Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 20–32.

Barr, S. L. and Barnsley, M. J., 1997. A region-based, graph- theoretic data model for the inference of second-order thematic information from remotely-sensed images. *International Journal of Geographical Information Science* [online], 11 (6), 555–576. Available from: http://pdfserve.informaworld.com/327980_777306414_713811360.pdf.

Bechtel, B. and Daneke, C., 2012. Classification of local climate zones based on multiple earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5 (4), 1191–1202.

Benediktsson, J. A., 2009. Ensemble classification algorithm for hyperspectral remote sensing data. *IEEE Geoscience and Remote Sensing Letters* [online], 6 (4), 762–766. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5175399.

Benediktsson, J. A., Chanussot, J., and Moon, W. M., 2012. Very High-resolution remote sensing: Challenges and opportunities [point of view]. *In*: *Proceedings of*

the IEEE. 1907–1910.

Berberoglu, S., Lloyd, C. D., Atkinson, P. M., and Curran, P. J., 2000. The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. *Computers & Geosciences* [online], 26 (4), 385–396. Available from: http://www.sciencedirect.com/science/article/pii/S0098300499001193.

Berthod, M., Kato, Z., Yu, S., and Zerubia, J., 1996. Bayesian image classification using Markov random fields. *Image and Vision Computing*, 14 (4), 285–295.

Bezak, P., Bozek, P., and Nikitin, Y., 2014. Advanced robotic grasping system using deep learning. *Procedia Engineering* [online], 96, 10–20. Available from: http://www.sciencedirect.com/science/article/pii/S1877705814031439.

Bibby, P., 2009. Land use change in Britain. *Land Use Policy* [online], 26 (SUPPL. 1), S2–S13. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0264837709001446.

Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 65 (1), 2–16. Available from: http://dx.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., and Tiede, D., 2014. Geographic object-based image analysis - towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 87, 180–191. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2013.09.014.

Bratasanu, D., Nedelcu, I., and Datcu, M., 2011. Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications. *IEEE Journal*

of Selected Topics in Applied Earth Observations and Remote Sensing, 4 (1), 193–204.

Breiman, L., 1996. Bagging Predictors. *Machine Learning* [online], 24 (2), 123–140. Available from: http://link.springer.com/article/10.1023/A%3A1018054314350 [Accessed 11 Jun 2014].

Caccetta, P., Collings, S., Devereux, A., Hingee, K., McFarlane, D., Traylen, A., Wu, X., and Zhou, Z. S., 2016. Monitoring land surface and cover in urban and peri-urban environments using digital aerial photography. *International Journal of Digital Earth* [online], 9 (5), 457–476. Available from: http://www.tandfonline.com/eprint/rNsUJ4vf7MiaDDbKaSIb/full.

Cassidy, L., Binford, M., Southworth, J., and Barnes, G., 2010. Social and ecological factors and land-use land-cover diversity in two provinces in Southeast Asia. *Journal of Land Use Science*, 5 (4), 277–306.

Chaib, S., Liu, H., Gu, Y., and Yao, H., 2017. Deep Feature Fusion for VHR Remote Sensing Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (8), 4775–4784.

Chen, C., Zhang, B., Su, H., Li, W., and Wang, L., 2016. Land-use scene classification using multi-scale completed local binary patterns. *Signal, Image and Video Processing*, 10 (4), 745–752.

Chen, D. G., He, Q., and Wang, X. Z., 2010. FRSVMs: Fuzzy rough set based support vector machines. *Fuzzy Sets and Systems*, 161 (4), 596–607.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., 2016. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs* [online]. arXiv. Available from:

http://arxiv.org/abs/1606.00915.

Chen, S., Member, S., Wang, H., Xu, F., and Member, S., 2016. Target classification using the deep Convolutional Networks for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (8), 4806–4817.

Chen, X., Xiang, S., Liu, C.-L., and Pan, C.-H., 2014. Vehicle detection in satellite images by hybrid deep Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 11 (10), 1797–1801.

Chen, Y., Jiang, H., Li, C., Jia, X., and Ghamisi, P., 2016. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (10), 6232–6251.

Chen, Y., Lin, Z., Zhao, X., Wang, G., and Gu, Y., 2014. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7 (6), 2094–2107.

Chen, Y., Xue, Y., Ma, Y., and Xu, F., 2017. Measures of uncertainty for neighborhood rough sets. *Knowledge-Based Systems* [online], 120, 226–235. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0950705117300096.

Chen, Y., Zhao, X., and Jia, X., 2015. Spectral-Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8 (6), 2381–2392.

Cheng, D., Meng, G., Xiang, S., and Pan, C., 2017. FusionNet: Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10 (12), 5769–5783.

Cheng, G., Han, J., Guo, L., Qian, X., Zhou, P., Yao, X., and Hu, X., 2013. Object detection in remote sensing imagery using a discriminatively trained mixture model. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 85, 32–43. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2013.08.001.

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., and Pan, C., 2017. Automatic road detection and centerline extraction via cascaded end-to-end Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (6), 3322–3337.

Clinton, N., Holt, A., Scarborough, J., Yan, L., and Gong, P., 2010. Accuracy Assessment Measures for Object-based Image Segmentation Goodness. *Photogrammetric Engineering & Remote Sensing* [online], 76 (3), 289–299. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=0099-1112&volume=76&issue=3&spage=289.

Clinton, N., Yu, L., and Gong, P., 2015. Geographic stacking: Decision fusion to increase global land cover map accuracy. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 103, 57–65. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0924271615000556.

Comaniciu, D. and Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (5), 1–37.

Conchedda, G., Durieux, L., and Mayaux, P., 2008. An object-based method for mapping and change analysis in mangrove ecosystems. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 63 (5), 578–589. Available from:

http://linkinghub.elsevier.com/retrieve/pii/S0924271608000294.

Congalton, R. G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37 (1), 35–46.

Dark, S. J. and Bram, D., 2007. The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, 31 (5), 471–479.

Demarchi, L., Canters, F., Cariou, C., Licciardi, G., and Chan, J. C. W., 2014. Assessing the performance of two unsupervised dimensionality reduction techniques on hyperspectral APEX data for high resolution urban land-cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 87 (July), 166–179. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2013.10.012.

Dong, Z., Pei, M., He, Y., Liu, T., Dong, Y., and Jia, Y., 2015. Vehicle type classification using unsupervised Convolutional Neural Network. *IEEE Transactions on Intelligent Transportation Systems* [online], 16 (4), 2247–2256. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6976750.

Dragozi, E., Gitas, I., Stavrakoudis, D., and Theocharis, J., 2014. Burned Area Mapping Using Support Vector Machines and the FuzCoC Feature Selection Method on VHR IKONOS Imagery. *Remote Sensing* [online], 6 (12), 12005–12036. Available from: http://www.mdpi.com/2072-4292/6/12/12005/.

Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y., and Liu, S., 2012. Multiple classifier system for remote sensing image classification: A review. *Sensors* [online], 12 (4), 4764–4792. Available from: https://www.scopus.com/inward/record.url?eid=2-s2.0-

84860267020&partnerID=40&md5=a594deaa7c42472694f0127d4777efd1.

Duro, D. C., Franklin, S. E., and Dubé, M. G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment* [online], 118, 259–272. Available from: http://dx.doi.org/10.1016/j.rse.2011.11.020.

Dwyer, J., 2011. UK Land Use Futures: Policy influence and challenges for the coming decades. *Land Use Policy* [online], 28 (4), 674–683. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0264837710001225.

Farabet, C., Couprie, C., Najman, L., and Lecun, Y., 2013. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8), 1915–1929.

Fauvel, M., Chanussot, J., and Benediktsson, J. A., 2006. Decision fusion for the classification of urban remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* [online], 44 (10), 2828–2838. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1704969.

Fauvel, M., Chanussot, J., and Benediktsson, J. A., 2012. A spatial-spectral kernel-based approach for the classification of remote-sensing images. *Pattern Recognition*, 45 (1), 381–392.

Foody, G. M., 2000. Mapping land cover from remotely sensed data with a softened feedforward neural network classification. *Journal of Intelligent & Robotic Systems* [online], 29 (4), 433–449. Available from: http://www.springerlink.com/index/M1L201W52002V563.pdf.

Foody, G. M. and Arora, M. K., 1997. An evaluation of some factors affecting the

accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, 18, 799–810.

Foody, G. M. and Mathur, A., 2004. A relative evaluation of multi-class image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42, 1335–1343.

Del Frate, F., Pacifici, F., Schiavon, G., and Solimini, C., 2007. Use of neural networks for automatic classification from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 45 (4), 800–809.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y., 2003. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4, 933–969.

Fu, G., Liu, C., Zhou, R., Sun, T., and Zhang, Q., 2017. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sensing* [online], 9 (5), 498. Available from: http://www.mdpi.com/2072-4292/9/5/498.

Ge, Y., Bai, H., Cao, F., Li, S., Feng, X., and Li, D., 2009. Rough set-derived measures in image classification accuracy assessment. *International Journal of Remote Sensing* [online], 30 (20), 5323–5344. Available from: http://www.tandfonline.com/doi/abs/10.1080/01431160903131026.

Ge, Y., Cao, F., Du, Y., Lakhan, V. C., Wang, Y., and Li, D., 2011. Application of rough set-based analysis to extract spatial relationship indicator rules: An example of land use in Pearl River Delta. *Journal of Geographical Sciences*, 21 (1), 101–117.

Gere, J. M. and Timoshenko, S. P., 1972. *Mechanics of Materials* [online]. New

York, NY, USA: Van Nostrand Reinhold Co. Available from: http://link.springer.com/10.1007/978-1-4899-3124-5.

Giacco, F., Thiel, C., Pugliese, L., Scarpetta, S., and Marinaro, M., 2010. Uncertainty analysis for the classification of multispectral satellite images using SVMs and SOMs. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (10), 3769–3779.

Gong, M., Zhan, T., Zhang, P., and Miao, Q., 2017. Superpixel-Based Difference Representation Learning for Change Detection in Multispectral Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* [online], 55 (5), 2658–2673. Available from: http://ieeexplore.ieee.org/document/7839934/.

Guo, L., Chehata, N., Mallet, C., and Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (1), 56–66.

Haralick, R. M., Shanmugam, K., and Dinstein, I., 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3 (6), 610–621.

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep Residual Learning for Image Recognition. *In*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [online]. 770–778. Available from: http://ieeexplore.ieee.org/document/7780459/.

Hedhli, I., Moser, G., Zerubia, J., and Serpico, S. B., 2016. A New Cascade Model for the Hierarchical Joint Classification of Multitemporal and Multiresolution Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (11), 6333–6348.

Hermosilla, T., Ruiz, L. A., Recio, J. A., and Cambra-López, M., 2012. Assessing contextual descriptive features for plot-based classification of urban areas. *Landscape and Urban Planning* [online], 106 (1), 124–137. Available from: http://dx.doi.org/10.1016/j.landurbplan.2012.02.008.

Herold, M., Liu, X., and Clarke, K. C., 2003. Spatial Metrics and Image Texture for Mapping Urban Land Use. *Photogrammetric Engineering & Remote Sensing* [online], 69 (9), 991–1001. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=0099-1112&volume=69&issue=9&spage=991.

Hester, D. B., Cakir, H. I., Nelson, S. a C., and Khorram, S., 2008. Per-pixel Classification of High Spatial Resolution Satellite Imagery for Urban Land-cover Mapping. *Photogrammetric Engineering & Remote Sensing*, 74 (4), 463–471.

Heydari, S. S. and Mountrakis, G., 2018. Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment*, 204, 648–658.

Hofmann, P., Blaschke, T., and Strobl, J., 2011. Quantifying the robustness of fuzzy rule sets in object-based image analysis. *International Journal of Remote Sensing*, 32 (22), 7359–7381.

Hu, F., Xia, G.-S., Hu, J., and Zhang, L., 2015. Transferring deep Convolutional Neural Networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing* [online], 7 (11), 14680–14707. Available from: http://www.mdpi.com/2072-4292/7/11/14680/.

Hu, F., Xia, G. S., Wang, Z., Huang, X., Zhang, L., and Sun, H., 2015. Unsupervised

Feature Learning Via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8 (5), 2015–2030.

Hu, S. and Wang, L., 2013. Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing*, 34 (3), 790–803.

Hu, X. and Weng, Q., 2009. Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks. *Remote Sensing of Environment* [online], 113 (December), 2089–2102. Available from: http://dx.doi.org/10.1016/j.rse.2009.05.014.

Huang, X., Lu, Q., and Zhang, L., 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 90, 36–48. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2014.01.008.

Hughes, G. F., 1968. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, 14 (1), 55–63.

Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K., and Wang, G., 2018. Multilayer perceptron neural network for surface water extraction in landsat 8 OLI satellite images. *Remote Sensing*.

Kampffmeyer, M., Salberg, A.-B., and Jenssen, R., 2016. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. *In*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Las Vegas, NV, USA, 1–9.

Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. ImageNet classification with

deep Convolutional Neural Networks. *In*: *NIPS2012: Neural Information Processing Systems*. Lake Tahoe, Nevada, 1–9.

Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A., 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 14 (5), 778–782.

Längkvist, M., Kiselev, A., Alirezaie, M., and Loutfi, A., 2016. Classification and segmentation of satellite orthoimagery using Convolutional Neural Networks. *Remote Sensing*, 8 (329), 1–21.

Larsen, S. Ø., Salberg, A., and Eikvil, L., 2013. Automatic system for operational traffic monitoring using very-high- resolution satellite imagery. *International Journal of Remote Sensing* [online], 34 (13), 4850–4870. Available from: http://dx.doi.org/10.1080/01431161.2013.782708.

LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *Nature* [online], 521 (7553), 436–444. Available from: http://dx.doi.org/10.1038/nature14539%5Cn10.1038/nature14539.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11), 2278–2323.

Lei, Z., Fang, T., and Li, D., 2011. Land cover classification for remote sensing imagery using conditional texton forest with historical land cover map. *IEEE Geoscience and Remote Sensing Letters*, 8 (4), 720–724.

Lenz, I., Lee, H., and Saxena, A., 2015. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research* [online], 34 (4–5), 705–724. Available from: http://ijr.sagepub.com/content/34/4-5/705.short.

Leung, Y., Fischer, M. M., Wu, W. Z., and Mi, J. S., 2008. A rough set approach for

the discovery of classification rules in interval-valued information systems. *International Journal of Approximate Reasoning*, 47 (2), 233–246.

Li, H., Gu, H., Han, Y., and Yang, J., 2010. Object-oriented classification of high-resolution remote sensing imagery based on an improved colour structure code and a support vector machine. *International Journal of Remote Sensing*, 31 (6), 1453–1470.

Li, M., Bijker, W., and Stein, A., 2015. Use of Binary Partition Tree and energy minimization for object-based classification of urban land cover. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102, 48–61.

Li, M., Stein, A., Bijker, W., and Zhan, Q., 2016. Urban land use extraction from Very High Resolution remote sensing imagery using a Bayesian network. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 122, 192–205. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2016.10.007.

Li, S. Z., 2009. *Markov Random Field Modeling in Image Analysis*. Third Edit. Advances in Pattern Recognition. London: Springer.

Liu, Q., Hang, R., Song, H., and Li, Z., 2018. Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing* [online], 56 (1), 117–126. Available from: http://ieeexplore.ieee.org/document/8036413/.

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., and Hong, Y., 2017. Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science* [online], 31 (8), 1675–1696. Available from: https://doi.org/10.1080/13658816.2017.1324976.

Liu, X., Kang, C., Gong, L., and Liu, Y., 2016. Incorporating spatial interaction

patterns in classifying and understanding urban land use. *International Journal of Geographical Information Science* [online], 30 (2), 334–350. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84948064862&partnerID=40&md5=6cfd4cbb29876d6551a0692721148768.

Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W., and Munteanu, A., 2017. Hourglass-shape network based semantic segmentation for high resolution aerial imagery. *Remote Sensing* [online], 9 (6), 522. Available from: http://www.mdpi.com/2072-4292/9/6/522.

Liu, Y., Zhong, Y., Fei, F., Zhu, Q., and Qin, Q., 2018. Scene Classification Based on a Deep Random-Scale Stretched Convolutional Neural Network. *Remote Sensing* [online], 10 (3), 444. Available from: http://www.mdpi.com/2072-4292/10/3/444.

Long, Y., Gong, Y., Xiao, Z., and Liu, Q., 2017. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* [online], 55 (5), 2486–2498. Available from: http://ieeexplore.ieee.org/document/7827088/.

Löw, F., Conrad, C., and Michel, U., 2015. Decision fusion and non-parametric classifiers for land use mapping using multi-temporal RapidEye data. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 108, 191–204. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2015.07.001.

Luus, F. P. S., Salmon, B. P., Bergh, F. Van Den, and Maharaj, B. T. J., 2015. Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, 12 (12), 2448–2452.

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P., 2017. Convolutional Neural

Networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (2), 645–657.

Malinverni, E. S., Tassetti, A. N., Mancini, A., Zingaretti, P., Frontoni, E., and Bernardini, A., 2011. Hybrid object-based approach for land use/land cover mapping using high spatial resolution imagery. *International Journal of Geographical Information Science* [online], 25 (6), 1025–1043. Available from: http://www.tandfonline.com/doi/abs/10.1080/13658816.2011.566569.

Marmanis, D., Datcu, M., Esch, T., and Stilla, U., 2016. Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geoscience and Remote Sensing Letters*, 13 (1), 105–109.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., and Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 135, 158–172. Available from: https://doi.org/10.1016/j.isprsjprs.2017.11.009.

Mas, J. F. and Flores, J. J., 2008. The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, 29 (3), 617–663.

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., Wang, L., Zhou, G., and Thenkabail, P. S., 2016. Pansharpening by Convolutional Neural Networks. *Remote Sensing*, 8 (594), 1–22.

McRoberts, R. E., 2013. Post-classification approaches to estimating change in forest area using remotely sensed auxiliary data. *Remote Sensing of Environment* [online], 151, 149–156. Available from:

http://dx.doi.org/10.1016/j.rse.2013.03.036.

Min, J. H. and Lee, Y.-C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28, 603–614.

Ming, D., Li, J., Wang, J., and Zhang, M., 2015. Scale parameter selection by spatial statistics for GeOBIA: Using mean-shift based multi-scale segmentation as an example. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 106, 28–41. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0924271615001203.

Mnih, V., 2013. Machine Learning for Aerial Image Labeling. *PhD Thesis*. University of Toronto.

Mokhtarzade, M. and Zoej, M. J. V., 2007. Road detection from high-resolution satellite images using artificial neural networks. *International Journal of Applied Earth Observation and Geoinformation* [online], 9 (1), 32–40. Available from: http://www.sciencedirect.com/science/article/pii/S0303243406000171.

Mountrakis, G., Im, J., and Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 66 (3), 247–259. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2010.11.001.

Myint, S. W., 2001. A robust texture analysis and classification approach for urban land-use and land-cover feature discrimination. *Geocarto International* [online], 16 (4), 29–40. Available from: http://www.tandfonline.com/doi/abs/10.1080/10106040108542212.

Myint, S. W., Gober, P., Brazel, A., Grossman-Clarke, S., and Weng, Q., 2011. Per-

pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sensing of Environment*, 115 (5), 1145–1161.

Naidoo, L., Cho, M. A., Mathieu, R., and Asner, G., 2012. Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 69, 167–179.

Niemeyer, J., Rottensteiner, F., and Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, 152–165.

Nishii, R., 2003. A Markov random field-based approach to decision-level fusion for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 41 (10), 2316–2319.

Nogueira, K., Penatti, O. A. B., and dos Santos, J. A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition* [online], 61, 539–556. Available from: http://dx.doi.org/10.1016/j.patcog.2016.07.001.

Ok, A. O., 2013. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 86, 21–40. Available from: http://www.sciencedirect.com/science/article/pii/S0924271613002050.

Oliva-Santos, R., Maciá-Pérez, F., and Garea-Llano, E., 2014. Ontology-based topological representation of remote-sensing images. *International Journal of Remote Sensing* [online], 35 (1), 16–28. Available from: http://www.tandfonline.com/doi/abs/10.1080/01431161.2013.858847.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment* [online], 148, 42–57. Available from: http://dx.doi.org/10.1016/j.rse.2014.02.015.

Othman, E., Bazi, Y., Alajlan, N., Alhichri, H., and Melgani, F., 2016. Using convolutional features and a sparse autoencoder for land-use scene classification. *International Journal of Remote Sensing* [online], 37 (10), 2149–2167. Available from: http://www.tandfonline.com/doi/full/10.1080/01431161.2016.1171928.

Othman, E., Bazi, Y., Melgani, F., Alhichri, H., Alajlan, N., and Zuair, M., 2017. Domain Adaptation Network for Cross-Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (8), 4441–4456.

Ozdarici-Ok, A., Ok, A., and Schindler, K., 2015. Mapping of agricultural crops from single high-resolution multispectral images—Data-driven smoothing vs. Parcel-based smoothing. *Remote Sensing* [online], 7 (5), 5611–5638. Available from: http://www.mdpi.com/2072-4292/7/5/5611/.

Pacifici, F., Chini, M., and Emery, W. J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment* [online], 113 (6), 1276–1292. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0034425709000625.

Paisitkriangkrai, S., Sherrah, J., Janney, P., and Van Den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9 (7), 2868–2881.

Pan, G., Qi, G., Wu, Z., Zhang, D., and Li, S., 2013. Land-use classification using taxi

GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, 14 (1), 113–123.

Pan, X., Zhang, S., Zhang, H., Na, X., and Li, X., 2010. A variable precision rough set approach to the remote sensing land use/cover classification. *Computers & Geosciences* [online], 36 (12), 1466–1473. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0098300410001639.

Pan, X. and Zhao, J., 2017. A central-point-enhanced convolutional neural network for high-resolution remote-sensing image classification. *International Journal of Remote Sensing* [online], 38 (23), 6554–6581. Available from: https://www.tandfonline.com/doi/full/10.1080/01431161.2017.1362131.

Pan, X. and Zhao, J., 2018. High-Resolution Remote Sensing Image Classification Method Based on Convolutional Neural Network and Restricted Conditional Random Field. *Remote Sensing*, 10 (6), 1–20.

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., and Vateekul, P., 2017. Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields. *Remote Sensing*, 9 (7).

Patino, J. E. and Duque, J. C., 2013. A review of regional science applications of satellite remote sensing in urban settings. *Computers, Environment and Urban Systems*, 37 (1), 1–17.

Pawlak, Z., 1982. Rough sets. *International Journal of Computer & Information Sciences*, 11 (5), 341–356.

Pei, J., Huang, Y., Huo, W., Zhang, Y., Yang, J., and Yeo, T.-S., 2018. SAR Automatic Target Recognition Based on Multiview Deep Learning Framework.

IEEE Transactions on Geoscience and Remote Sensing [online], 56 (4), 2196–2210. Available from: http://ieeexplore.ieee.org/document/8207785/.

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., and Zanchetta, L., 2013. A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6 (5), 2102–2131.

Pingel, T. J., Clarke, K. C., and McBride, W. A., 2013. An improved simple morphological filter for the terrain classification of airborne LIDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 77, 21–30. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2012.12.002.

Pontius, R. G. and Millones, M., 2011. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32 (15), 4407–4429.

Powers, R. P., Hermosilla, T., Coops, N. C., and Chen, G., 2015. Remote sensing and object-based techniques for mapping fine-scale industrial disturbances. *International Journal of Applied Earth Observation and Geoinformation* [online], 34, 51–57. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0303243414001482.

Qian, Y., Liang, X., Lin, G., Guo, Q., and Liang, J., 2017. Local multigranulation decision-theoretic rough sets. *International Journal of Approximate Reasoning* [online], 82, 119–137. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0888613X16303206.

Qin, F., Guo, J., and Sun, W., 2017. Object-oriented ensemble classification for polarimetric SAR Imagery using restricted Boltzmann machines. *Remote Sensing Letters*, 8 (3), 204–213.

R.A. Dunne and N.A. Campbell, 1995. Neighbour-Based MLPs. *In*: *IEEE International Conference on Neural Networks*. 270–274.

Radoux, J. and Bogaert, P., 2017. Good practices for object-based accuracy assessment. *Remote Sensing*, 9 (7), 1–23.

Regnauld, N. and Mackaness, W. a., 2006. Creating a hydrographic network from its cartographic representation: a case study using Ordnance Survey MasterMap data. *International Journal of Geographical Information Science*, 20 (6), 611–631.

Regniers, O., Bombrun, L., Lafon, V., and Germain, C., 2016. Supervised Classification of Very High Resolution Optical Images Using Wavelet-Based Textural Features. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (6), 3722–3735.

Reis, S. and Tasdemir, K., 2011. Identification of hazelnut fields using spectral and gabor textural features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (5), 652–661.

Rodriguez-Galiano, V. F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P. M., and Jeganathan, C., 2012. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment* [online], 121, 93–107. Available from: http://www.sciencedirect.com/science/article/pii/S0034425711004408.

Romero, A., Gatta, C., Camps-valls, G., and Member, S., 2016. Unsupervised deep

feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (3), 1349–1362.

Salehi, B., Zhang, Y., Zhong, M., and Dey, V., 2012. A review of the effectiveness of spatial information used in urban land cover classification of VHR imagery. *International Journal of Geoinformatics* [online], 8 (2), 35–51. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84863431792&partnerID=tZOtx3y1.

Sargent, I., Hare, J., Young, D., Wilson, O., Doidge, C., Holland, D., and Atkinson, P. M., 2017. Inference and discovery in remote sensing data with features extracted using deep networks. *In*: Bramer, M. and Petridis, M., eds. *AI-2017 Thirty-seventh SGAI International Conference on Artificial Intelligence* [online]. Cambridge, United Kingdom, 131–136. Available from: http://link.springer.com/10.1007/978-3-319-71078-5_10.

Schindler, K., 2012. An Overview and Comparison of Smooth Labeling Methods for Land-Cover Classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 50 (11), 4534–4545.

Schmidhuber, J., 2015. Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61, 85–117.

Sesnie, S. E., Finegan, B., Gessler, P. E., Smith, A. M. S., Zayra, R. B., and Thessler, S., 2010. The multispectral separability of Costa Rican rainforest types with support vector machines and Random Forest decision trees. *International Journal of Remote Sensing*, 31 (11), 2885–2909.

Shao, W., Yang, W., and Xia, G. S., 2013. Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification.

International Journal of Remote Sensing, 34 (23), 8588–8602.

Sharma, A., Liu, X., Yang, X., and Shi, D., 2017. A patch-based convolutional neural network for remote sensing image classification. *Neural Networks* [online], 95, 19–28. Available from: http://dx.doi.org/10.1016/j.neunet.2017.07.017.

Shi, H., Chen, L., Bi, F., Chen, H., and Yu, Y., 2015. Accurate urban area detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 12 (9), 1948–1952.

Sikder, I. U., 2016. A variable precision rough set approach to knowledge discovery in land cover classification. *International Journal of Digital Earth* [online], 9 (12), 1206–1223. Available from: https://www.tandfonline.com/doi/full/10.1080/17538947.2016.1194489.

Simonyan, K. and Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICRL)* [online], 1–14. Available from: http://arxiv.org/abs/1409.1556.

Strigl, D., Kofler, K., and Podlipnig, S., 2010. Performance and scalability of GPU-based Convolutional Neural Networks. *In*: *2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing* [online]. 317–324. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5452452.

Su, L., Gong, M., Zhang, P., Zhang, M., Liu, J., and Yang, H., 2017. Deep learning and mapping based ternary change detection for information unbalanced images. *Pattern Recognition*, 66, 213–228.

Sun, Y., Zhao, L., Huang, S., Yan, L., and Dissanayake, G., 2014. L2-SIFT: SIFT

feature extraction and matching for large images in large-scale aerial photogrammetry. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 91, 1–16. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2014.02.001.

Swiniarski, R. W. and Skowron, A., 2003. Rough set methods in feature selection and recognition. *Pattern Recognition Letters*, 24 (6), 833–849.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., 2015. Going deeper with convolutions. *In*: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1–9.

Tang, J., Deng, C., Huang, G.-B., and Zhao, B., 2015. Compressed-domain ship detection on spaceborne optical image using Deep Neural Network and Extreme Learning Machine. *IEEE Transactions on Geoscience and Remote Sensing*, 53 (3), 1174–1185.

Tang, Y., Jing, L., Li, H., and Atkinson, P. M., 2016. A multiple-point spatially weighted k-NN method for object-based classification. *International Journal of Applied Earth Observation and Geoinformation* [online], 52, 263–274. Available from: http://dx.doi.org/10.1016/j.jag.2016.06.017.

Tso, B. and Mather, P. M., 2009. *Classification methods for remotely sensed data* [online]. 2nd ed. Methods. Boca Raton, FL: CRC Press. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:CLASSIFICATION+METHODS+FOR+REMOTELY+SENSED+DATA#0.

Verburg, P. H., Neumann, K., and Nol, L., 2011. Challenges in using land use and land cover data for global change studies. *Global Change Biology*, 17 (2), 974–

989.

Vetrivel, A., Gerke, M., Kerle, N., and Vosselman, G., 2015. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 105, 61–78. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0924271615000982.

Volpi, M. and Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (2), 881–893.

Walde, I., Hese, S., Berger, C., and Schmullius, C., 2014. From land cover-graphs to urban structure types. *International Journal of Geographical Information Science*, 28 (3), 584–609.

Wang, C., Komodakis, N., and Paragios, N., 2013. Markov Random Field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117 (11), 1610–1627.

Wang, H., Wang, Y., Zhang, Q., Xiang, S., and Pan, C., 2017. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing*, 9 (5), 1–15.

Wang, J., Song, J., Chen, M., and Yang, Z., 2015. Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing* [online], 36 (12), 3144–3169. Available from: http://www.tandfonline.com/doi/full/10.1080/01431161.2015.1054049.

Wang, L. and Liu, J., 1999. Texture classification using multiresolution Markov random field models. *Pattern Recognition Letters*, 20 (2), 171–182.

Wang, L., Shi, C., Diao, C., Ji, W., and Yin, D., 2016. A survey of methods incorporating spatial information in image classification and spectral unmixing. *International Journal of Remote Sensing* [online], 37 (16), 3870–3910. Available from: http://www.tandfonline.com/doi/full/10.1080/01431161.2016.1204032.

Wang, Q. and Shi, W., 2013. Unsupervised classification based on fuzzy c-means with uncertainty analysis. *Remote Sensing Letters* [online], 4 (11), 1087–1096. Available from: http://www.tandfonline.com/doi/abs/10.1080/2150704X.2013.832842.

Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., and Jiao, L., 2018. A deep learning framework for remote sensing image registration. *ISPRS Journal of Photogrammetry and Remote Sensing* [online]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0924271617303891.

Wei, X., Fu, K., Gao, X., Yan, M., Sun, X., Chen, K., and Sun, H., 2018. Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model. *Remote Sensing Letters*, 9 (3), 199–208.

Wei, Y., Wang, Z., and Xu, M., 2017. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geoscience and Remote Sensing Letters*, 14 (5), 709–713.

Wei, Y., Yuan, Q., Shen, H., and Zhang, L., 2017. Boosting the Accuracy of Multispectral Image Pansharpening by Learning a Deep Residual Network. *IEEE Geoscience and Remote Sensing Letters*, 14 (10), 1795–1799.

Wolpert, D. H. and Macready, W. G., 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1 (1), 67–82.

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., and Shibasaki, R.,

2018. Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sensing* [online], 10 (3), 407. Available from: http://www.mdpi.com/2072-4292/10/3/407.

Wu, S. S., Qiu, X., Usery, E. L., and Wang, L., 2009. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. *Annals of the Association of American Geographers*, 99 (1), 76–98.

Xia, G. S., Delon, J., and Gousseau, Y., 2010. Shape-based invariant texture indexing. *International Journal of Computer Vision*, 88 (3), 382–403.

Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., and Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (7), 3965–3981.

Yang, W., Dai, D., Triggs, B., and Xia, G. S., 2012. SAR-based terrain classification using weakly supervised hierarchical Markov aspect models. *IEEE Transactions on Image Processing*, 21 (9), 4232–4243.

Yang, W., Yin, X., and Xia, G. S., 2015. Learning high-level features for satellite image classification with limited labeled samples. *IEEE Transactions on Geoscience and Remote Sensing*, 53 (8), 4472–4482.

Yang, X., Qian, X., and Mei, T., 2015. Learning salient visual word for scalable mobile image retrieval. *Pattern Recognition* [online], 48 (10), 3093–3101. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0031320314005287.

Yang, Y. and Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. *In*: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10* [online].

270. Available from: http://portal.acm.org/citation.cfm?doid=1869790.1869829.

Yao, X., Han, J., Cheng, G., Qian, X., and Guo, L., 2016. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (6), 3660–3671.

Yin, W., Yang, J., Yamamoto, H., and Li, C., 2015. Object-based larch tree-crown delineation using high-resolution satellite imagery. *International Journal of Remote Sensing* [online], 36 (3), 822–844. Available from: http://www.tandfonline.com/doi/abs/10.1080/01431161.2014.999165.

Yoshida, H. and 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. *Computers, Environment and Urban Systems* [online], 29 (2), 223–247. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0198971504000432.

Yu, H., Liu, Z., and Wang, G., 2014. An automatic method to determine the number of clusters using decision-theoretic rough set. *International Journal of Approximate Reasoning*, 55 (1 PART 2), 101–115.

Yu, J., Weng, K., Liang, G., and Xie, G., 2013. A vision-based robotic grasping system using deep learning for 3D object recognition and pose estimation. *In*: *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)* [online]. 1175–1180. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6739623.

Yue, J., Mao, S., and Li, M., 2016. A deep learning framework for hyperspectral image classification using spatial pyramid pooling. *Remote Sensing Letters*, 7 (9), 875–884.

Zhan, J. and Zhu, K., 2017. A novel soft rough fuzzy set: Z-soft rough fuzzy ideals of hemirings and corresponding decision making. *Soft Computing*, 21 (8), 1923–1936.

Zhang, C. and Atkinson, P. M., 2016. Novel shape indices for vector landscape pattern analysis. *International Journal of Geographical Information Science* [online], 30 (12), 2442–2461. Available from: https://www.tandfonline.com/doi/full/10.1080/13658816.2016.1179313.

Zhang, C. and Kovacs, J. M., 2012. The application of small unmanned aerial systems for precision agriculture: A review. *Precision Agriculture*, 13 (6), 693–712.

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., and Atkinson, P. M., 2018. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 140, 133–144. Available from: https://doi.org/10.1016/j.isprsjprs.2017.07.014.

Zhang, C., Sargent, I., Pan, X., Gardiner, A., Hare, J., and Atkinson, P. M., 2018. VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing* [online], 56 (8), 4507–4521. Available from: https://doi.org/10.1109/TGRS.2018.2822783.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., and Atkinson, P. M., 2018. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sensing of Environment* [online], 216, 57–70. Available from: https://www.sciencedirect.com/science/article/pii/S0034425718303122.

Zhang, C., Wang, T., Atkinson, P. M., Pan, X., and Li, H., 2015. A novel multi-

parameter support vector machine for image classification. *International Journal of Remote Sensing* [online], 36 (7), 1890–1906. Available from: http://www.tandfonline.com/doi/full/10.1080/01431161.2015.1029096.

Zhang, F., Du, B., and Zhang, L., 2016. Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (3), 1793–1802.

Zhang, F., Du, B., Zhang, L., and Xu, M., 2016. Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection. *IEEE Transactions on Geoscience and Remote Sensing* [online], 54 (9), 5553–5563. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84973571837&doi=10.1109%2FTGRS.2016.2569141&partnerID=40&md5=89f15dee33757e12c2d4259163c962d1.

Zhang, L., Zhang, L., and Kumar, V., 2016. Deep learning for Remote Sensing Data. *IEEE Geoscience and Remote Sensing Magazine*, 4 (2), 22–40.

Zhang, P., Gong, M., Su, L., Liu, J., and Li, Z., 2016. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, 24–41.

Zhang, Q., Wang, J., Gong, P., and Shi, P., 2003. Study of urban spatial patterns from SPOT panchromatic imagery using textural analysis. *International Journal of Remote Sensing*, 24 (21), 4137–4160.

Zhang, S., Zhang, J., Li, F., and Cropp, R., 2006. Vector analysis theory on landscape pattern (VATLP). *Ecological Modelling*, 193 (3–4), 492–502.

Zhang, T., Yan, W., Li, J., and Chen, J., 2016. Multiclass Labeling of Very High-

Resolution Remote Sensing Imagery by Enforcing Nonlocal Shared Constraints in Multilevel Conditional Random Fields Model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9 (7), 2854–2867.

Zhao, A., Fu, K., Wang, S., Zuo, J., Zhang, Y., Hu, Y., and Wang, H., 2017. Aircraft recognition based on landmark detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 14 (8), 1413–1417.

Zhao, B., Zhong, Y., and Zhang, L., 2016. A spectral-structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 116, 73–85. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2016.03.004.

Zhao, W. and Du, S., 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 113, 155–165. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2016.01.004.

Zhao, W., Du, S., and Emery, W. J., 2017. Object-Based Convolutional Neural Network for High-Resolution Imagery Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10 (7), 3386–3396.

Zhao, W., Du, S., Wang, Q., and Emery, W. J., 2017. Contextually guided very-high-resolution imagery classification with semantic segments. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 132, 48–60. Available from: http://dx.doi.org/10.1016/j.isprsjprs.2017.08.011.

Zheng, C. and Wang, L., 2015. Semantic Segmentation of Remote Sensing Imagery Using Object-Based Markov Random Field Model With Regional Penalties. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote*

Sensing, 8 (5), 1924–1935.

Zhong, Y., Zhao, J., and Zhang, L., 2014. A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52 (11), 7023–7037.

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

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., and Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5 (4), 8–36.

Ziarko, W., 1993. Variable precision rough set model. *Journal of Computer and System Sciences*, 46 (1), 39–59.