

1 An object-based convolutional neural network (OCNN) for urban land use 2 classification

3 Ce Zhang ^{a,*}, Isabel Sargent ^b, Xin Pan ^{c,d}, Huapeng Li ^d, Andy Gardiner ^b, Jonathon Hare ^e,
4 Peter M. Atkinson ^{a,*}

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

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

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

37

38 **1. Introduction**

39 Urban land use information, reflecting socio-economic functions or activities, is essential for
40 urban planning and management. It also provides a key input to urban and transportation
41 models, and is essential to understanding the complex interactions between human activities
42 and environmental change (Patino and Duque, 2013). With the rapid development of modern
43 remote sensing technologies, a huge amount of very fine spatial resolution (VFSR) remotely
44 sensed imagery is now commercially available, opening new opportunities to extract urban
45 land use information at a very detailed level (Pesaresi et al., 2013). However, urban land
46 features captured by these VFSR images are highly complex and heterogeneous, comprising
47 the juxtaposition of a mixture of anthropogenic urban and semi-natural surfaces. Often, the
48 same urban land use types (e.g. residential areas) are characterized by distinctive physical
49 properties or land cover materials (e.g. composed of different roof tiles), and different land use
50 categories may exhibit the same or similar reflectance spectra and textures (e.g. asphalt roads
51 and parking lots) (Pan et al., 2013). Meanwhile, information on urban land use within VFSR
52 imagery is presented implicitly as patterns or high-level semantic functions, in which some
53 identical low-level ground features or object classes are frequently shared amongst different
54 land use categories. This complexity and diversity of spatial and structural patterns in urban
55 areas makes its classification into land use classes a challenging task (Hu et al., 2015).
56 Therefore, it is important to develop robust and accurate urban land use classification
57 techniques by effectively representing the spatial patterns or structures lying in VFSR remotely
58 sensed data.

59 Over the past few decades, tremendous effort has been made in developing automatic urban
60 land use classification methods. These methods can be categorized broadly into four classes
61 based on the spatial unit of representation (i.e. pixels, moving windows, objects and scenes)
62 (Liu et al., 2016). The pixel-level approaches that rely purely upon spectral characteristics are
63 able to classify land cover, but are insufficient to distinguish land uses that are typically
64 composed of multiple land covers, and such problems are particularly significant in urban
65 settings (Zhao et al., 2016). Spatial information, that is, texture (Herold et al., 2003; Myint,
66 2001) or context (Wu et al., 2009), was incorporated to analyse urban land use patterns through
67 moving kernel windows (Niemeyer et al., 2014). However, it could be argued that both pixel-

68 based and moving window-based methods require to predefine arbitrary image structures,
69 whereas actual objects and regions might be irregularly shaped in the real world (Herold et al.,
70 2003). Therefore, object-based image analysis (OBIA) that is built upon automatically
71 segmented objects from remotely sensed imagery is preferable (Blaschke, 2010), and has been
72 considered as the dominant paradigm over the last decade (Blaschke et al., 2014). Those image
73 objects, as the base units of OBIA, offer two kinds of information with a spatial partition,
74 specifically; within-object information (e.g. spectral, texture, shape) and between-object
75 information (e.g. connectivity, contiguity, distances, and direction amongst adjacent objects).
76 Many studies applied OBIA for urban land use classification using within-object information
77 with a set of low-level features (such as spectra, texture, shape) of the ground features (e.g.
78 Blaschke, 2010; Blaschke et al., 2014; Hu and Wang, 2013). These OBIA approaches, however,
79 might overlook semantic functions or spatial configurations due to the inability to use low-
80 level features in semantic feature representation. In this context, researchers have attempted to
81 incorporate between-object information by aggregating objects using spatial contextual
82 descriptive indicators on well-defined land use units, such as cadastral fields or street blocks.
83 Those descriptive indicators were commonly derived by means of spatial metrics to quantify
84 their morphological properties (Yoshida and Omae, 2005) or graph-based methods that model
85 the spatial relationships (Barr and Barnsley, 1997; Walde et al., 2014). However, the ancillary
86 geographic data for specifying the land use units might not be available for some regions, and
87 the spatial contexts are often hard to describe and characterize as a set of “rules”, even though
88 the complex structures or patterns might be recognizable and distinguishable by human experts
89 (Oliva-Santos et al., 2014). Thus, advanced data-driven approaches are highly desirable to learn
90 land use semantics automatically through high-level feature representations.

91 Recently, deep learning has become the new hot topic in machine learning and pattern
92 recognition, where the most representative and discriminative features are learnt end-to-end,
93 hierarchically (Chen et al., 2016a). This breakthrough was triggered by a revival of interest in
94 the use of multi-layer neural networks to model higher-level feature representations without
95 human-designed features or rules. Convolutional neural networks (CNNs), as a well-
96 established and popular deep learning method, has produced state-of-the-art results for multiple
97 domains, such as visual recognition (Krizhevsky et al., 2012), image retrieval (Yang et al.,
98 2015) and scene annotation (Othman et al., 2016). Owing to its superiority in higher-level
99 feature representation and scene understanding, the CNN has demonstrated great potential in
100 many remote sensing tasks such as vehicle detection (Chen et al., 2014; Dong et al., 2015),

101 road network extraction (Cheng et al., 2017), remotely sensed scene classification (Othman et
102 al., 2016; Sargent et al., 2017), and semantic segmentation (Zhao et al., 2017b). Interested
103 readers are referred to a comprehensive review of deep learning in remote sensing (Zhu et al.,
104 2017).

105 Land use information extraction from remotely sensed data using CNN models has been
106 undertaken in the form of land-use scene classification, which aims to assign a semantic label
107 (e.g. tennis court, parking lot, etc.) to an image according to its content (Chen et al., 2016b;
108 Nogueira et al., 2017). There are broadly two strategies to exploit the CNN models for scene-
109 level land use classification, namely; *i*) pre-trained or fine-tuned CNN, and *ii*) fully-trained
110 CNN from scratch. The first strategy relies on pre-trained CNN networks transferred from an
111 auxiliary domain with natural images, which has been demonstrated empirically to be useful
112 for land-use scene classification (Hu et al., 2015; Nogueira et al., 2017). However, it requires
113 three input channels derived from natural images with RGB only, whereas the multispectral
114 remotely sensed imagery often involves the near infrared band, and such a distinction restricts
115 the utility of pre-trained CNN networks. Alternatively, the (ii) fully-trained CNN strategy gives
116 full control over the network architecture and parameters, which brings greater flexibility and
117 expandability (Chen et al., 2016). Previous researchers have explored the feasibility of the
118 fully-trained strategy in building CNN models for scene level land-use classification. For
119 example, Luus et al. (2015) proposed a multi-view CNN with multi-scale input strategies to
120 address the issue of land use scene classification and its scale-dependent characteristics.
121 Othman et al. (2016) used convolutional features and a sparse auto-encoder for scene-level
122 land-use image classification, which further demonstrated the superiority of CNNs in feature
123 learning and representation. Xia et al., (2017) even constructed a large-scale aerial scene
124 classification dataset (AID) for performance evaluation among various CNN models and
125 architectures developed by both strategies. However, the goal of these land use scene
126 classifications is essentially *image* categorization, where a small patch extracted from the
127 original remote sensing image is labelled into a semantic category, such as ‘airport’, ‘residential’
128 or ‘commercial’ (Maggiori et al., 2017). Land-use scene classification, therefore, does not meet
129 the actual requirement of remotely sensed land use image classification, which requires all
130 pixels in an entire image to be identified and labelled into land use categories (i.e., producing
131 a thematic map).

132 With the intrinsic advantages of hierarchical feature representation, the patch-based CNN
133 models provide great potential to extract higher-level land use semantic information. However,

134 this patch-wise procedure introduces artefacts on the border of the classified patches and often
135 produces blurred boundaries between ground surface objects (Zhang et al., 2018a, 2018b), thus,
136 introducing uncertainty in the classification. In addition, to obtain a full resolution
137 classification map, pixel-wise densely overlapped patches were used at the model inference
138 phase, which inevitably led to extremely redundant computation. As an alternative, Fully
139 Convolutional Networks (FCN) and its extensions have been introduced into remotely sensed
140 sematic segmentation to address the pixel-level classification problem (e.g. Liu et al., 2017;
141 Paisitkriangkrai et al., 2016; Volpi and Tuia, 2017). These FCN-based methods are, however,
142 mostly developed to solve low-level semantic (i.e. land cover) classification tasks, due to the
143 insufficient spatial information in the inference phase and the lack of contextual information at
144 up-sampling layers (Liu et al., 2017). In short, we argue that the existing CNN models,
145 including both patch-based and pixel-level approaches, are not well designed in terms of
146 accuracy and/or computational efficiency to cope with the complicated problem of urban land
147 use classification using VFSR remotely sensed imagery.

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

163 The remainder of this paper is organized as follows: Section 2 introduces the general workflow
164 and the key components of the proposed methods. Section 3 describes the study area and data
165 sources. The results are presented in section 4, followed by a discussion in section 5. The
166 conclusions are drawn in the last section.

167

168 **2. Method**

169 **2.1 Convolutional Neural Networks (CNN)**

170 A Convolutional Neural Network (CNN) is a multi-layer feed-forward neural network that is
171 designed specifically to process large scale images or sensory data in the form of multiple
172 arrays by considering local and global stationary properties (LeCun et al., 2015). The main
173 building block of a CNN is typically composed of multiple layers interconnected to each other
174 through a set of learnable weights and biases (Romero et al., 2016). Each of the layers is fed
175 by small patches of the image that scan across the entire image to capture different
176 characteristics of features at local and global scales. Those image patches are generalized
177 through alternative convolutional and pooling/subsampling layers within the CNN framework,
178 until the high-level features are obtained on which a fully connected classification is performed
179 (Schmidhuber, 2015). Additionally, several feature maps may exist in each convolutional layer
180 and the weights of the convolutional nodes in the same map are shared. This setting enables
181 the network to learn different features while keeping the number of parameters tractable.
182 Moreover, a nonlinear activation (e.g. sigmoid, hyperbolic tangent, rectified linear units)
183 function is taken outside the convolutional layer to strengthen the non-linearity (Strigl et al.,
184 2010). Specifically, the major operations performed in the CNN can be summarized as:

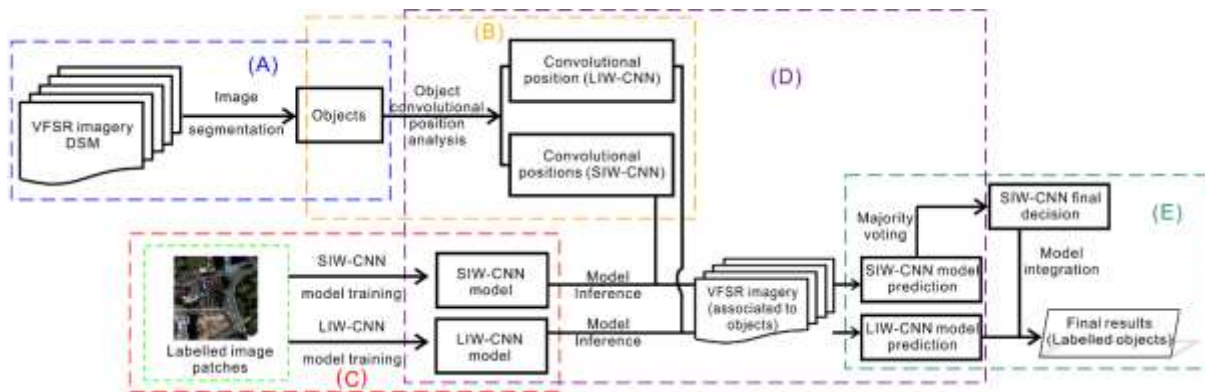
$$185 \quad O^l = pool_p(\sigma(O^{l-1} * W^l + b^l)) \quad (1)$$

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

192 **2.2 Object-based CNN (OCNN)**

193 An object-based CNN (OCNN) is proposed for the urban land use classification using VFSR
194 remotely sensed imagery. The OCNN is trained as the standard CNN models with labelled
195 image patches, whereas the model prediction is to label each segmented object derived from
196 image segmentation. The segmented objects are generally composed of two distinctive objects
197 in geometry, including linearly shaped objects (LS-objects) (e.g. Highway, Railway and Canal)

198 and other (non-linearly shaped) general objects (G-objects). To accurately predict the land use
 199 membership association of a G-object, a large spatial context (i.e. a large image patch) is
 200 required when using the CNN model. Such a large image patch, however, often may lead to a
 201 large uncertainty in the prediction of LS-objects due to narrow linear features being ignored
 202 throughout the convolutional process. Thus, a large input window CNN (LIW-CNN) and a
 203 range of small input window CNNs (SIW-CNN) were thereafter trained to predict the G-object
 204 and the LS-object, respectively, where the appropriate convolutional positions of both models
 205 were derived from a novel object convolutional position analysis (OCPA). The final
 206 classification results were determined by the decision fusion of the LIW-CNN and the SIW-
 207 CNN. As illustrated by Figure 1, the general workflow of the proposed OCNN consists of five
 208 major steps, including (A) image segmentation, (B) OCPA, (C) LIW-CNN and SIW-CNN
 209 model training, (D) LIW-CNN and SIW-CNN model inference, and (E) Decision fusion of
 210 LIW-CNN and SIW-CNN. Each of these steps is elaborated in the following section.



211
 212 Figure 1 Flowchart of the proposed object-based CNN (OCNN) method with five major steps: (A) image
 213 segmentation, (B) object convolutional position analysis (OCPA), (C) LIW-CNN and SIW-CNN model
 214 training, (D) LIW-CNN and SIW-CNN model inference, and (E) fusion decision of LIW-CNN and SIW-CNN.

215 2.2.1 Image segmentation

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

225 **2.2.2 Object convolutional position analysis (OCPA)**

226 The object convolutional position analysis (OCPA) is employed based on the **moment**
 227 **bounding (MB) box** of each object to identify the position of LIW-CNN and those of SIW-
 228 CNNs. The MB box, proposed by Zhang and Atkinson, (2016), refers to the minimum
 229 bounding rectangle built upon the moment orientation (the orientation of the major axis) of a
 230 polygon (i.e. an object), derived from planar characteristics defined by mechanics (Zhang and
 231 Atkinson, 2016; Zhang et al., 2006). The MB box theory is briefly described hereafter.

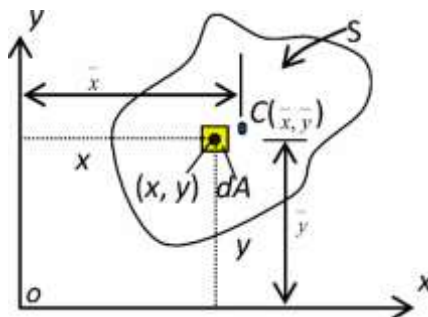
232 Suppose that (x, y) is a point within a planar polygon (S) (Figure 2), whose centroid is $C(\bar{x}, \bar{y})$.
 233 The moment of inertia about the x-axis (I_{xx}) and y-axis (I_{yy}), and the product of inertia (I_{xy})
 234 are expressed by Equations 2, 3 and 4, respectively.

235
$$I_{xx} = \int y^2 dA \quad (2)$$

236
$$I_{yy} = \int x^2 dA \quad (3)$$

237
$$I_{xy} = \int xy dA \quad (4)$$

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



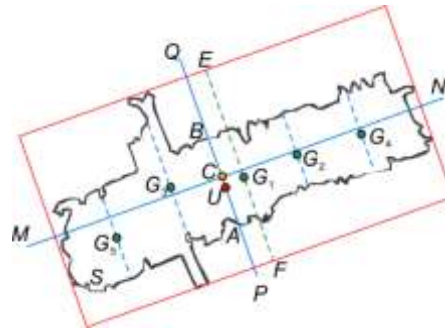
239
 240 Figure 2 A patch (S) with centroid $C(\bar{x}, \bar{y})$, dA is the differential area of point (x, y) , Oxy is the geographic
 241 coordinate system.

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

246
$$\tan 2\theta_{MB} = \frac{2I_{xy}}{I_{yy} - I_{xx}} \quad (5)$$

247
$$\theta_{MB} = \frac{1}{2} \tan^{-1} \left(\frac{2I_{xy}}{I_{yy} - I_{xx}} \right) \quad (6)$$

248 The moment bounding (MB) box (the rectangle in red shown in Figure 3) that minimally
 249 encloses the polygon, S , is then constructed by taking θ_{MB} as the orientation of the long side
 250 of the box, and EF is the perpendicular bisector of the MB box with respect to its long side.
 251 The discrete forms of Equations 2-6 suitable for patch computation, are further deduced by
 252 associating the value of a line integral to that of a double integral using Green's theorem (see
 253 Zhang et al. (2006) for theoretical details).



254
 255 Figure 3 Moment bounding (MB) box and the CNN convolutional positions of a polygon S .

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

262
$$n = \frac{l - d}{d} \quad (7)$$

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

266 2.2.3 LIW-CNN and SIW-CNN model training

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

271 2.2.4 LIW-CNN and SIW-CNN model inference

272 After the above steps, the trained LIW-CNN and SIW-CNN models, and the convolutional
273 position of LIW-CNN and those of SIW-CNN for each object are available. For a specific
274 object, its land use category can be predicted by the LIW-CNN at the derived convolutional
275 position within the VFSR imagery; at the same time, the predictions on the land use
276 membership associations of the object can also be obtained by employing SIW-CNN models
277 at the corresponding convolutional positions. Thus each object is predicted by both LIW-CNN
278 and SIW-CNN models.

279 2.2.5 Fusion decision of LIW-CNN and SIW-CNN

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

286 2.3 Accuracy assessment

287 Both pixel-based and object-based methods were adopted to comprehensively test the
288 classification performance using the testing sample set through five-fold cross validation. The
289 pixel-based approach was assessed based on the overall accuracy and Kappa coefficient as well
290 as per-class mapping accuracy computed from a confusion matrix. The object-based
291 assessment was based on geometry (Clinton et al., 2010; Li et al., 2015; Radoux and Bogaert,
292 2017). Specifically, suppose that a classified object M_i overlaps a set of reference objects O_{ij} ,
293 where $j = 1, 2, \dots, r$, r refers to the total number of reference objects overlapped by M_i . For each
294 pair of objects (M_i, O_{ij}), a weight parameter deduced by the ratio between the area of a reference
295 object (area (O_{ij})) and the total area of reference objects $\sum_{j=1}^r \text{area}(O_{ij})$ was introduced to
296 calculate over-classification $OC(M_i)$ and under-classification $UC(M_i)$ error indices as:

$$297 \quad OC(M_i) = \sum_{i=1}^r (w \cdot (1 - \frac{\text{area}(M_i \cap O_{ij})}{\text{area}(O_{ij})})), \quad w = \frac{\text{area}(O_{ij})}{\sum_{j=1}^r \text{area}(O_{ij})} \quad (8)$$

$$298 \quad UC(M_i) = 1 - \frac{\sum_{j=1}^r \text{area}(M_i \cap O_{ij})}{\text{area}(M_i)} \quad (9)$$

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

$$301 \quad TCE(M_i) = \sqrt{\frac{OC(M_i)^2 + UC(M_i)^2}{2}} \quad (10)$$

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

304 **3. Experimental Results and Analysis**

305 *3.1 Study area and data sources*

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

311 Aerial photos of S1 and S2 were captured using Vexcel UltraCam Xp digital aerial cameras on
312 22/07/2012 and 20/04/2016, respectively. The images have four multispectral bands (Red,
313 Green, Blue and Near Infrared) with a spatial resolution of 50 cm. The study sites were subset
314 into the city centres and their surrounding regions with spatial extents of 5802×4850 pixels for
315 S1 and 5875×4500 pixels for S2, respectively. Land use categories of the study areas were
316 defined according to the official land use classification system provided by the UK government
317 Department for Communities and Local Government (DCLG). Detailed descriptions of each
318 land use class and its corresponding sub-classes in S1 and S2 are listed in Tables 1 and 2,
319 respectively. 10 dominant land use classes were identified within S1, including *high-density*
320 *residential*, *commercial*, *industrial*, *medium-density residential*, *highway*, *railway*, *park and*
321 *recreational area*, *parking lot*, *redeveloped area*, and *harbour and sea water*. In S2, nine land
322 use categories were found, including *residential*, *commercial*, *industrial*, *highway*, *railway*,
323 *park and recreational area*, *parking lot*, *redeveloped area*, and *canal*.

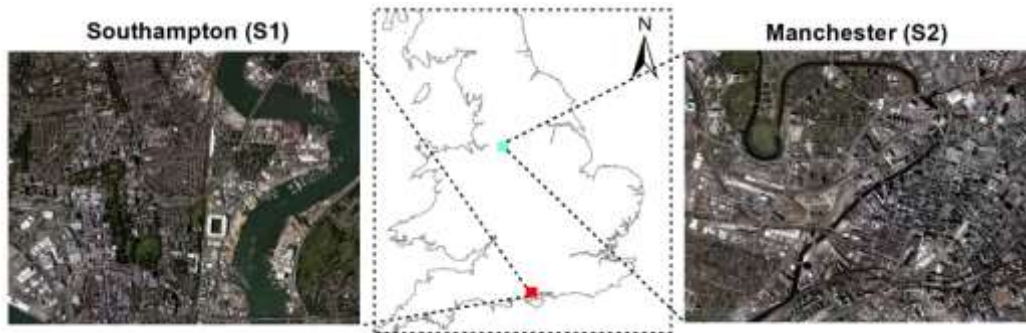


Figure 4 The two study areas of urban scenes: S1 (Southampton) and S2 (Manchester).

324
325
326
327

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

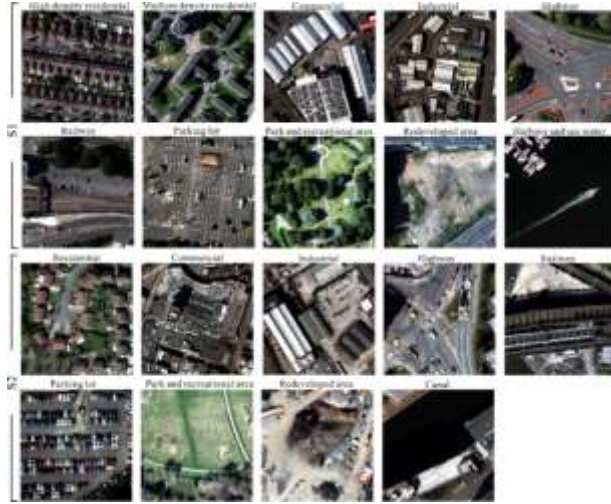
Land Use Class	Train	Test	Sub-class Components
High-density residential	1026	684	Residential houses, terraces, a small coverage of green space
Medium-density residential	984	656	Residential flats with a large green space and parking lots
Commercial	972	648	Commercial services with complex buildings, 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

328
329

Table 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 and vegetation
Commercial	1028	685	Shopping centre, retail parks 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 and vegetation, bare soil, lake
Redeveloped area	1032	688	Bare soil, scattered vegetation, reconstructions
Canal	994	662	Canal water

330



331

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

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

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

350 **3.2 Model structure and parameter settings**

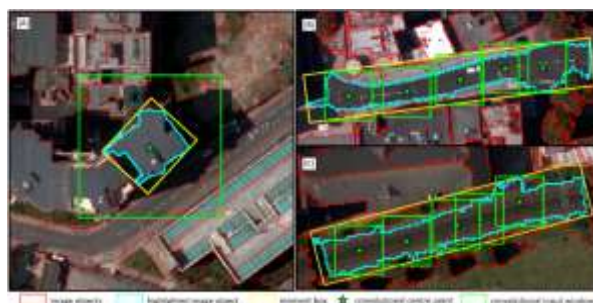
351 The proposed method was implemented based on vector objects extracted by means of image
 352 segmentation. The objects were further classified through object-based CNN networks
 353 (OCNN). Detailed parameters and model structures optimised by S1 and directly generalised
 354 in S2 were clarified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

355 **3.2.1 Segmentation parameter settings**

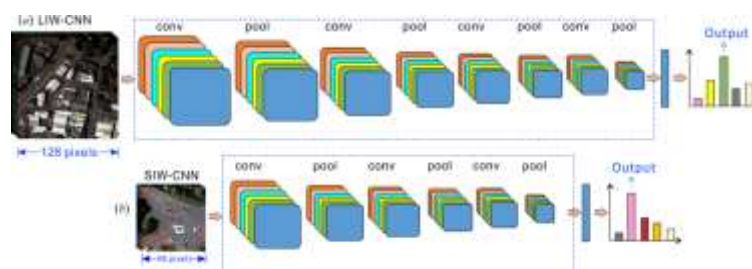
356 The initial mean-shift segmentation algorithm was implemented using the Orfeo Toolbox
357 open-source software. Two spatial and spectral bandwidth parameters, namely the spatial
358 radius and the range (spectral) radius, were optimized as 15.5 and 20 through cross-validation
359 coupled with a small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

362 **3.2.2 LIW-CNN and SIW-CNN model structures and parameters**

363 Within the two study sites, the highway, railway in S1 and the highway, railway, and canal in
364 S2 belong to linearly shaped objects (LS-objects) in consideration of the elongated geometric
365 characteristics (e.g. Figure 6(B), (C)), while all the other objects belong to general objects (G-
366 objects) (e.g. Figure 6(A)). The LIW-CNN with a large input window (Figure 6(A)), and SIW-
367 CNNs with small input windows (Figure 6(B), (C)) that are suitable for the prediction of G-
368 objects and LS-objects, respectively, were designed here. Note, the other type of CNN models
369 employed on each object, namely, the SIW-CNNs in Figure 6(A) and the LIW-CNN in both
370 Figure 6(B) and 6(C) were not presented in the figure to gain a better visual effect. The model
371 structures and parameters of LIW-CNN and SIW-CNN are illustrated by Figure 7(a) and 7(b)
372 and are detailed hereafter.



373
374 Figure 6 An illustration of object convolutional position analysis with the moment box (yellow rectangle), the
375 convolutional centre point (green star), and the convolutional input window (green rectangle), as well as the
376 highlighted image object (in cyan). All the other segmented objects are demonstrated as red polygons. (A)
377 demonstrates the large input window for a general object, and (B), (C) illustrate the small input windows for
378 linearly shaped objects (highway and railway, respectively, in these exemplars).



379

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

383 The model structure of the LIW-CNN was designed similar to the AlexNet (Krizhevsky et al.,
384 2012) with eight layers (Figure 7(a)) using a large input window size (128×128), but with small
385 convolutional filters (3×3) for the majority of layers except for the first one (which was 5×5).
386 The input window size was determined through cross-validation on a range of window sizes,
387 including $\{48 \times 48, 64 \times 64, 80 \times 80, 96 \times 96, 112 \times 112, 128 \times 128, 144 \times 144, 160 \times 160\}$ to
388 sufficiently cover the contextual information of general objects relevant to land use semantics.
389 The number of filters was tuned to 64 to extract deep convolutional features effectively at each
390 level. The CNN network involved alternating convolutional (conv) and pooling layers (pool)
391 as shown in Figure 7(a), where the maximum pooling within a 2×2 window was used to
392 generalize the feature and keep the parameters tractable.

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

399 All the other parameters for both CNN networks were optimized empirically based on standard
400 computer vision. For example, the number of neurons for the fully connected layers was set as
401 24, and the output labels were predicted through softmax estimation with the same number of
402 land use categories. The learning rate and the epoch were set as 0.01 and 600 to learn the deep
403 features through backpropagation.

404 **3.2.3 OCNN parameter settings**

405 In the proposed OCNN method, the LIW-CNN and the SIW-CNN networks were integrated to
406 predict the land use classes of general objects and linearly shaped objects at the model inference
407 phase. Based on object convolutional position analysis (OCPA), the LIW-CNN with a 128×128
408 input window (denoted as OCNN_{128}) was employed only once per object, and the SIW-CNNs
409 with a 48×48 input window (denoted as OCNN_{48^*} , the 48^* here represents multiple image
410 patches sized 48×48) were used at multiple positions to predict the land use label of an object
411 through majority voting (see section 2.2.2 for theoretical details). The parallel distance

412 parameter d in OCPA that controls the convolutional locations and the number of small window
413 size CNNs, was estimated by the length distribution of the moment box together with a trial-
414 and-error procedure in a wide search space (0.5 m – 20 m) with a step of 0.5 m. The d was
415 optimized as 5 m for the objects with moment box length (l) larger than or equal to 20 m, and
416 was estimated by $l/4$ for those objects with l less than 20 m (i.e. the minimum number of small
417 window size CNNs was 3) to perform a statistical majority voting. The proposed method
418 (OCNN_{128+48*}) integrates both OCNN₁₂₈ and OCNN_{48*}, which is suitable for the prediction of
419 urban land use semantics for any shaped objects.

420 **3.2.4 Other benchmark methods and their parameters**

421 To evaluate the classification performance of the proposed method, three existing benchmark
422 methods (i.e. Markov Random Field (MRF), object-based image analysis with support vector
423 machine (OBIA-SVM), and the pixel-wise CNN) that each incorporate spatial context were
424 compared comprehensively, as follows:

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

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

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

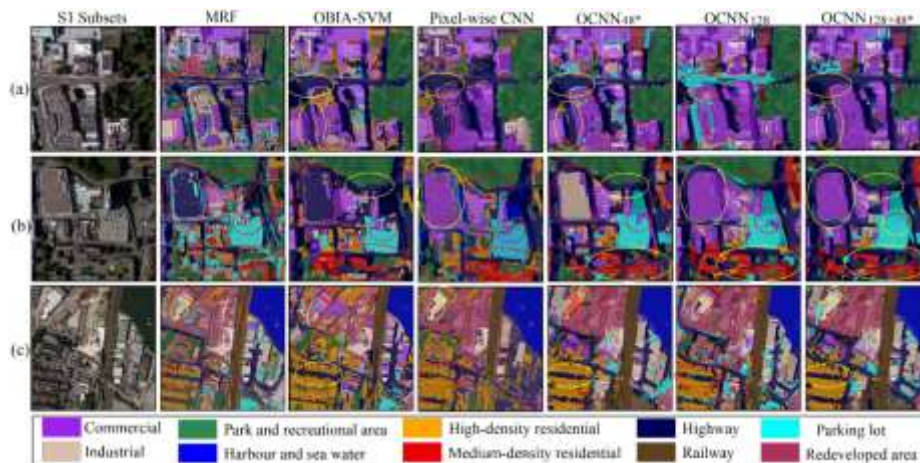
444 input image size was chosen from $\{28 \times 28, 32 \times 32, 36 \times 36, 40 \times 40, 44 \times 44, 48 \times 48, 52 \times 52$ and
445 $56 \times 56\}$ to evaluate the influence of contextual area on classification performance. The optimal
446 input image patch size for the pixel-wise CNN was found to be 48×48 to leverage the training
447 sample size and the computational resources (e.g. GPU memory). The depth configuration of
448 the CNN network plays a key role in classification accuracy because the quality of the learnt
449 features is highly influenced by the level of abstraction and representation. As suggested by
450 Chen et al., (2016a), the number of CNN layers was chosen as six to balance the network
451 complexity and robustness. Other CNN parameters were tuned empirically through cross-
452 validation. For example, the filter size was set to 3×3 for the convolutional layer with a stride
453 of 1, and the number of filters was set to 24 to extract multiple convolutional features at each
454 level. The learning rate was set as 0.01 and the number of epochs was chosen as 600 to fully
455 learn the features through backpropagation.

456 *3.3 Classification results and analysis*

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

464 ***Experiment 1:*** A desirable classification result was obtained in S1 by using the proposed
465 OCNN_{128+48^*} . To provide a useful visualization, three subsets of S1 classified by different
466 approaches were presented in Figure 8, with the correct or incorrect classification results
467 marked in yellow or red circles, respectively. In general, the proposed method achieved the
468 smoothest visual results with precise boundary information compared with other benchmark
469 methods. Most importantly, the semantic contents of complex urban land uses (e.g. commercial,
470 industrial etc.) were effectively characterized, and the linearly shaped features including
471 highway and railway were identified with high geometric fidelity. As shown by Figure 8(a)
472 and 8(c), the highway (a linear feature) was misclassified as a parking lot (red circles) by
473 OCNN_{128} , whereas the highway feature was accurately identified by the OCNN_{48^*} (yellow
474 circles). However, OCNN_{48^*} was inferior to OCNN_{128} when identifying general objects, as
475 demonstrated by Figure 8(b). Fortunately, these complementary behaviours of the two sub-

476 modules were captured by the proposed OCNN_{128+48^*} , which was able to label the highway
 477 accurately (yellow circles in Figure 8(b)). The pixel-wise CNN demonstrated some capacity
 478 for extracting semantic functions for complex objects; for example, the commercial area in
 479 Figure 8(b) was correctly distinguished (yellow circle). However, classification errors along
 480 the edges or boundaries between objects were found. For example, the edges of the highway
 481 were misclassified as high-density residential as shown by Figure 8(a). For the OBIA-SVM,
 482 the simple land uses with less within-object variation (e.g. highway) were more accurately
 483 classified (yellow circle in Figure 8(a) and 8(c)), whereas, those highly complex land uses with
 484 great within-object variation (e.g. commercial, industrial etc.) were more likely to be
 485 misclassified (red circle in Figure 8(b)). In addition, the OBIA-SVM could also discover some
 486 sub-objects (e.g. balcony on the residential house) through the information context. The results
 487 of the MRF, in contrast to the other object-based approaches, were the least smooth even
 488 though local neighbourhood information was used. Nevertheless, there were still some benefits
 489 of the MRF: spectrally distinctive land uses, such as highway, park and recreational area, were
 490 classified with a relatively high accuracy.



491
 492 Figure 8 Three typical image subsets (a, b and c) in study site S1 with their classification results. Columns from
 493 left to right represent the original images (R G B bands only), and the MRF, OBIA-SVM, Pixel-wise CNN,
 494 OCNN_{48^*} , OCNN_{128} , and the proposed OCNN_{128+48^*} results. The red and yellow circles denote incorrect and
 495 correct classification, respectively.

496

497 The effectiveness of the OCNN_{128+48^*} was also demonstrated by quantitative classification
 498 accuracy assessment. As shown in Table 2, the OCNN_{128+48^*} achieved the largest overall
 499 accuracy of 89.52% with a Kappa coefficient (κ) of 0.88, consistently larger than its sub-
 500 module OCNN_{128} (87.31% OA and κ of 0.86) and the OCNN_{48^*} (OA of 84.23% and κ of 0.82),
 501 respectively. The accuracy increase was much more dramatic in comparison with other

502 benchmark methods, including the pixel-wise CNN (81.62% OA and κ of 0.80), the OBIA-
503 SVM (79.54% OA and κ of 0.78), as well as the MRF (OA of 78.67% and κ of 0.76). The
504 superiority of the proposed OCNN_{128+48*} was further demonstrated by the per-class mapping
505 accuracy (Table 3). From the table, it can be seen that the accuracies of highway and railway
506 were increased significantly by 5.34% and 4.64% respectively, compared with the OCNN₁₂₈.
507 This was followed by a moderate increase of 3.24% for the parking lot class. Other land use
508 classes (e.g. commercial, industrial, etc.) were slightly increased in terms of classification
509 accuracy (less than 1.5%) without statistical significance in comparison with OCNN₁₂₈. When
510 comparing with the OCNN_{48*}, the accuracy increase of the proposed OCNN_{128+48*} was
511 remarkable for the majority of general object classes, with increases of up to 6.06%, 6.51%,
512 4.98%, 4.7% and 4.68%, for the classes of commercial, industrial, redeveloped area, park and
513 recreational area, and high-density residential, respectively; whereas the accuracies of the
514 medium-density residential and the parking lot increased moderately, by 3.31% and 3.81%,
515 respectively. For linearly shaped objects, however, the OCNN_{128+48*} was not substantially
516 superior to the OCNN_{48*}, with just a slight accuracy increase of 1.52% for highway and 2.41%
517 for railway, respectively. For general objects with complex semantic functions, including
518 commercial, industrial, redeveloped area, park and recreational area, and high-density
519 residential, the increase in accuracy of the OCNN_{128+48*} was much more significant, by up to
520 6.06%, 6.51%, 4.98%, 4.7% and 4.68%, respectively.

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

530

531 Table 3. Classification accuracy comparison amongst MRF, OBIA-SVM, Pixel-wise CNN, OCNN_{48*}, OCNN₁₂₈,
532 and the proposed OCNN_{128+48*} method for Southampton using the per-class mapping accuracy, overall accuracy
533 (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

534

535 An object-based accuracy assessment was implemented in S1 to validate the classification
536 performance in terms of over-classification (*OC*), under-classification (*UC*), and total
537 classification error (*TCE*). Three typical methods, including OBIA-SVM (denoted as OBIA),
538 pixel-wise CNN (denoted as CNN), and the proposed OCNN_{128+48*} method (denoted as OCNN),
539 were evaluated, with accuracy comparisons of each land use class listed in Table 4. Clearly,
540 the proposed OCNN method produced the smallest *OC*, *UC*, and *TCE* errors, respectively
541 (highlighted by bold font), constantly smaller than those of the CNN and OBIA. Generally, the
542 *UC* errors are smaller than *OC* errors, demonstrating that a slight over-segmentation was
543 produced. Specifically, the OCNN demonstrates excellent object-level classification, with the
544 majority of classes less than 0.2 in *TCE*. Those complex land use classes, including commercial
545 and industrial, can be segmented precisely and classified with small *TCE* of 0.22 and 0.20, less
546 than those of CNN (0.29 and 0.27) and OBIA (0.39 and 0.38). The parking lot objects with
547 complex land use patterns, were also recognised accurately with high fidelity (*OC* of 0.22, *UC*
548 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,
549 and 0.37). For those LS-objects, the OCNN achieved promising accuracy in comparison with
550 the other two benchmarks. For example, the *TCE*s of highway and railway produced by the
551 OCNN were 0.17 and 0.09, smaller than those of the CNN (0.25 and 0.22) and OBIA (0.20 and
552 0.18). All the other land use categories demonstrate increased segmentation accuracy. For
553 instance, the *TCE* of park and recreational area was 0.18 with the OCNN, less than for the CNN
554 of 0.24 and OBIA of 0.32.

555

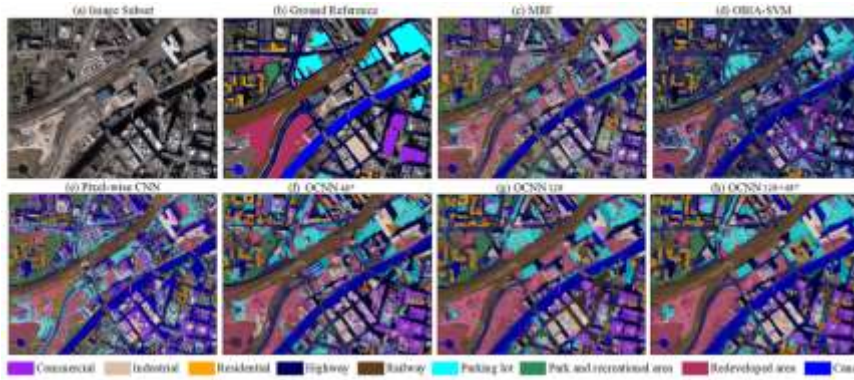
556 Table 4 Object-based accuracy assessment among OBIA-SVM (OBIA), Pixel-wise CNN (CNN), and the
557 proposed OGC-CNN_{128+48*} method (OCNN) for Southampton using error indices of *OC*, *UC*, and *TCE*. The bold
558 font highlights the smallest classification error of a specific index per row.

Class	<i>OC</i>			<i>UC</i>			<i>TCE</i>		
	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

559

560 **Experiment 2:** The most accurate classification performance was also achieved in S2 by the
561 proposed method, as illustrated by the quantitative accuracy results in Table 5. From the table,
562 it can be seen that OCNN_{128+48*} obtained the greatest overall accuracy (OA) of 90.87% with a
563 Kappa coefficient (κ) of 0.88, significantly larger than the OCNN₁₂₈ (OA of 88.74% and κ of
564 0.86), the OCNN_{48*} (OA of 85.06% with κ of 0.83), the Pixel-wise CNN (OA of 82.39% and
565 κ of 0.81), the OBIA-SVM (OA of 80.37% with κ of 0.79), and the MRF (OA of 78.52% with
566 κ of 0.76). The effectiveness of the OCNN_{128+48*} was also demonstrated by the per-class
567 mapping accuracy. Compared with the OCNN₁₂₈, the classes formed by linearly shaped objects,
568 including the highway, railway and canal, had significantly increased accuracies of up to 5.36%,
569 3.06% and 3.48%, respectively (Table 5). Such increases can also be noticed in Figure 9 (a
570 subset of S2), where the misclassifications of railway and highway shown in Figure 9(g) were
571 rectified in Figure 9(h) classified by the OCNN_{128+48*}. At the same time, the parking lot land
572 use class was moderately increased by 2.28%. Whereas, other land use classes had slightly
573 increases in accuracy of less than 1% on average. In contrast, the OCNN_{128+48*} led to no
574 significant increases over the OCNN_{48*} for the linear object classes, with accuracy increases
575 for highway, railway and canal of 1.8%, 0.42% and 1.22%, respectively. For the general classes,
576 especially the complex land uses (e.g. commercial, industrial etc.), remarkable accuracy
577 increases were achieved with an average up to 6.75%. Figure 9(f) (classified by OCNN_{48*}) also
578 showed the confusion between the commercial and industrial land use classes, which was
579 revised in Figure 9(h). With respect to the benchmark comparators, the accuracy increase of
580 OCNN_{128+48*} was much more obvious for most of the land use classes, with the largest accuracy
581 increase up to 12.39% for parking lot, 11.21% for industrial, and 8.56% for commercial,
582 compared with the MRF, OBIA-SVM and Pixel-wise CNN, respectively. The undesirable
583 visual effects and misclassifications can also be seen in Figure 9(c-e), which were corrected in
584 Figure 9(h).



585

586

587

588

589

Figure 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) OCNN₁₂₈ classification, and (h) OCNN_{128+48*} classification.

590

591

592

Table 5 Classification accuracy comparison amongst MRF, OBIA-SVM, Pixel-wise CNN, OCNN_{48*}, OCNN₁₂₈, 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

593

594

595

596

597

598

599

600

601

602

603

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 6). The error indices in S2 (Table 6) present a similar trend with those in S1 (Table 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 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

604 highway, railway, and canal, were precisely characterised by the OCNN method, with a *TCE*
605 of 0.16, 0.09, and 0.08, significantly smaller than for the CNN (0.22, 0.21, and 0.14) and OBIA
606 (0.18, 0.19, and 0.12). The residential land use was also clearly improved with a very small
607 *TCE* of 0.10, smaller than for the CNN (0.22) and OBIA (0.26). Other land use classes, such
608 as the park and recreational area and the redeveloped area, were also better distinguished by
609 the OCNN (0.16 and 0.15 in terms of *TCE*), smaller than for the CNN (0.21 and 0.25) and
610 OBIA (0.28 and 0.30).

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

Class	<i>OC</i>			<i>UC</i>			<i>TCE</i>		
	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

614
615 A sensitivity analysis was conducted to further investigate the effect of different input window
616 sizes on the overall accuracy of urban land use classification (see Figure 10). The window sizes
617 varied from 16×16 to 144×144 with a step size of 16. From Figure 10, it can be seen that both
618 S1 and S2 demonstrated similar trends for the proposed OCNN and the pixel-wise CNN (CNN).
619 With window sizes smaller than 48×48 (i.e. relatively small windows), the classification
620 accuracy of OCNN is lower than that of CNN, but the accuracy difference decreases with an
621 increase of window size. Once the window size is larger than 48×48 (i.e. relatively large
622 windows), the overall accuracy of the OCNN increases steadily until the window is as large as
623 128×128 (up to around 90%), and outperforms the CNN which has a generally decreasing trend
624 in both study sites. However, an even larger window size (e.g. 144×144) in OCNN could result
625 in over-smooth results, thus reducing the classification accuracy.

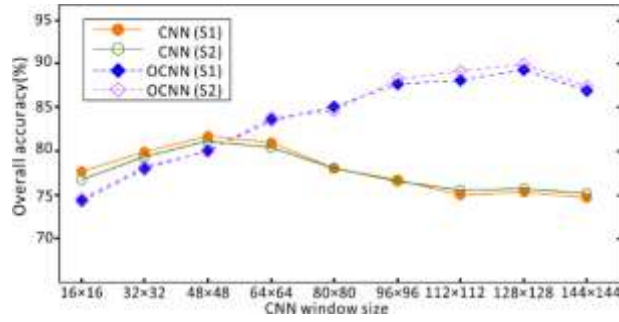


Figure 10 The influence of CNN window size on the overall accuracy of pixel-wise CNN and the proposed OCNN method for both study sites S1 and S2.

3.4 Computational efficiency

The computational efficiency of the proposed method was evaluated and compared with the other methods listed in Table 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 7, the training time of the Pixel-wise CNN, OCNN_{48*}, OCNN₁₂₈ 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 OCNN₁₂₈ 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 7. Comparison of computational times amongst MRF, OBIA-SVM, Pixel-wise CNN, OCNN_{48*}, OCNN₁₂₈, 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

650 **4. Discussion**

651 Urban land use captured in VFSR remotely sensed imagery is highly complex and
652 heterogeneous, with spatial patterns presented that imply a hierarchical or nested class structure.
653 Classifying urban land use requires not only a precise characterisation of image objects as
654 functional units, but also an accurate and robust representation of spatial context. A novel
655 object-based CNN method for urban land use classification using VFSR remotely sensed
656 imagery was, therefore, proposed, in which the functional units are derived at object levels and
657 the spatial patterns are learned through CNN networks with hierarchical feature representation.
658 The OCNN method is fundamentally different from the work proposed by Zhao et al. (2017a)
659 in multiple aspects, including: (1) the realisation of an object-based CNN for land use
660 classification under the OBIA framework using geometric characterisations to guide the choice
661 of sizes and locations of image patches; (2) the use of within-object and between-object
662 information learnt by the OCNN model to represent the spatial and hierarchical relationships;
663 (3) the high computational efficiency achieved with targeted sampling at the object level to
664 avoid a pixel-wise (i.e., densely overlapping) convolutional process.

665 ***4.1 Convolutional neural networks for urban land use feature representation***

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

683 innovation through adaptation of the CNNs for urban land use classification using appropriate
684 functional units and convolutional processes.

685 ***4.2 Object-based CNN (OCNN) for urban land use classification***

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

700 Each segmented object has its distinctive geometric characteristics with respect to the specific
701 land use category. Representations of objects using OCNN should be scale-dependent with
702 appropriate window sizes and convolutional positions to match the geometric distributions,
703 especially when dealing with the two types of objects with geometrically distinctive
704 characteristics, namely, general objects (G-objects) and linearly-shaped objects (LS-objects).
705 For those G-objects with complex urban land use, a deep CNN network (eight-layers) with a
706 large input image patch (128×128) was used to accurately identify an object with a large extent
707 of contextual information. Such an image patch could reflect the real dimension of G-objects
708 and their wide context (64m×64m in geographical space). The convolutional position of the
709 CNN network was theoretically derived close to the central region of a moment box, where
710 both object geometry and spatial anisotropy were characterised. In this way, the within-object
711 (at the centre of the image patch) and between-object (surrounding context within the image
712 patch) information are used simultaneously to learn the objects and the surrounding complex
713 spatial structures or patterns, with the largest overall accuracy at large context (Figure 10). The
714 LS-objects, such as Highway, Railway and Canal, were sampled along the objects using a range
715 of less deep CNNs (six-layers) with small window size (48×48) (or 24m×24m geographically)

716 and were classified through majority voting. These small window size CNNs focus on the
717 within-object information, which often includes homogeneous characteristics within objects
718 (e.g. rail tracks, asphalt road), and avoid the great variation between adjacent objects (e.g. trees,
719 residential buildings, bare land etc. alongside the Highway). Moreover, the small contextual
720 image patches with less deep networks cover the elongated objects sufficiently, without losing
721 useful within-object information through the convolutional process. To integrate the two
722 classification models for G-objects and LS-objects, a simple rule-based classification
723 integration was employed conditional upon model predictions, in which the majority of the
724 classification results were derived from the CNNs with large window size, whereas the
725 predictions of Highway, Railway and Canal were trusted by the voting results of small window
726 CNNs alone. Thus, the type of object (either as a G-object or a LS-object) is determined through
727 CNN model predictions and rule-based classification integration. Such a decision fusion
728 approach provides a pragmatic and effective manner to combine the two models by considering
729 the object geometry and class-specific adaptations. Overall, the proposed OCNN method with
730 large and small window size feature representations is a feasible solution for the complex urban
731 land use classification problem using VFSR remotely sensed imagery, with massive
732 generalisation capability for a broad range of applications.

733 ***4.3 Computational complexity and efficiency***

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

747 **4.4 Future research**

748 The proposed OCNN method provides a very high accuracy and efficiency for urban land use
749 classification using VFSR remotely sensed imagery. The image objects are identified through
750 decision fusion between a large input window CNN with a deep network and several small
751 input window CNNs with less deep networks, to account for typical distinctive object sizes and
752 geometries. However, such two-scale feature representation might be insufficient to
753 characterise some complex geometric characteristics. Therefore, a range of CNNs with
754 different input patch sizes will be adopted in the future to adapt to the diverse sizes and shapes
755 of the urban objects through weighted decision fusion. In addition, urban land use classification
756 was undertaken at a generalized spatial and semantic level (e.g., residential area, commercial
757 area and industrial area), without identifying smaller functional sites (e.g., supermarkets,
758 hospitals and playgrounds etc.). This issue might be addressed by incorporating multi-source
759 geospatial data, for example, those classified commercial areas might be further differentiated
760 as supermarkets, retail outlets, and café areas through indoor human activities. Future research
761 will, therefore, mine the semantic information from GPS trajectories, transportation networks
762 and social media data to characterise these smaller functional units in a hierarchical way, as
763 well as socioeconomic activities and population dynamics.

764 **5. Conclusions**

765 Urban land use classification using VFSR remotely sensed imagery remains a challenging task,
766 due to the indirect relationship between the desired high-level land use categories and the
767 recorded spectral reflectance. A precise partition of functional units as image objects together
768 with an accurate and robust representation of spatial context are, therefore, needed to
769 characterise urban land use structures and patterns into high-level feature thematic maps. This
770 paper proposed a novel object-based CNN (OCNN) method for urban land use classification
771 from VFSR imagery. In the OCNN, segmented objects consisting of linearly shaped objects
772 (LS-objects) and other general objects (G-objects), were utilized as functional units. The G-
773 objects were precisely identified and labelled through a single large input window (128×128)
774 CNN with a deep (eight-layer) network to perform a contextual object-based classification.
775 Whereas the LS-objects were each distinguished accurately using a range of small input
776 window (48×48) CNNs with less deep (six-layer) networks along the objects' lengths through
777 majority voting. The locations of the input image patches for both CNN networks were
778 determined by considering both object geometry and its spatial anisotropy, such as to
779 accurately classify the objects into urban land use classes. Experimental results on two

780 distinctive urban scenes demonstrated that the proposed OCNN method significantly increased
781 the urban land use classification accuracy for all land use categories. The proposed OCNN
782 method with large and small window size CNNs produced the most accurate classification
783 results in comparison with the sub-modules and other contextual-based and object-based
784 benchmark methods. Moreover, the OCNN method demonstrated a high computational
785 efficiency with much more acceptable time requirements than the standard pixel-wise CNN
786 method in the process of model inference. We conclude that the proposed OCNN is an effective
787 and efficient method for urban land use classification from VFSR imagery. Meanwhile, the
788 OCNN method exhibited an excellent generalisation capability on distinctive urban land use
789 settings with great potential for a broad range of applications.

790 **Acknowledgements**

791 This research was funded by PhD studentship “Deep Learning in massive area, multi-scale
792 resolution remotely sensed imagery” (NO. EAA7369), sponsored by Ordnance Survey and
793 Lancaster University. The authors thank the staff of the Ordnance Survey for supplying the
794 aerial imagery and the supporting ground data.

795

796 **Reference**

- 797 Barr, S.L., Barnsley, M.J., 1997. A region-based, graph- theoretic data model for the
798 inference of second-order thematic information from remotely-sensed images. *Int. J.*
799 *Geogr. Inf. Sci.* 11, 555–576. <https://doi.org/10.1080/136588197242194>
- 800 Berthod, M., Kato, Z., Yu, S., Zerubia, J., 1996. Bayesian image classification using Markov
801 random fields. *Image Vis. Comput.* 14, 285–295. [https://doi.org/10.1016/0262-](https://doi.org/10.1016/0262-8856(95)01072-6)
802 [8856\(95\)01072-6](https://doi.org/10.1016/0262-8856(95)01072-6)
- 803 Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS J. Photogramm.*
804 *Remote Sens.* 65, 2–16. <https://doi.org/10.1016/j.isprsjprs.2009.06.004>
- 805 Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R.,
806 van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-
807 based image analysis - towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.*
808 87, 180–191. <https://doi.org/10.1016/j.isprsjprs.2013.09.014>
- 809 Chen, C., Zhang, B., Su, H., Li, W., Wang, L., 2016. Land-use scene classification using
810 multi-scale completed local binary patterns. *Signal, Image Video Process.* 10, 745–752.
811 <https://doi.org/10.1007/s11760-015-0804-2>

812 Chen, X., Xiang, S., Liu, C.-L., Pan, C.-H., 2014. Vehicle detection in satellite images by
813 hybrid deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* 11,
814 1797–1801. <https://doi.org/10.1109/LGRS.2014.2309695>

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

819 Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C., 2017. Automatic road detection
820 and centerline extraction via cascaded end-to-end Convolutional Neural Network. *IEEE*
821 *Trans. Geosci. Remote Sens.* 55, 3322–3337.
822 <https://doi.org/10.1109/TGRS.2017.2669341>

823 Clinton, N., Holt, A., Scarborough, J., Yan, L., Gong, P., 2010. Accuracy Assessment
824 Measures for Object-based Image Segmentation Goodness. *Photogramm. Eng. Remote*
825 *Sens.* 76, 289–299. <https://doi.org/10.14358/PERS.76.3.289>

826 Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis.
827 *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 1–37. <https://doi.org/10.1109/34.1000236>

828 Dong, Z., Pei, M., He, Y., Liu, T., Dong, Y., Jia, Y., 2015. Vehicle type classification using
829 unsupervised Convolutional Neural Network. *IEEE Trans. Intell. Transp. Syst.* 16,
830 2247–2256. <https://doi.org/10.1109/ICPR.2014.39>

831 Herold, M., Liu, X., Clarke, K.C., 2003. Spatial metrics and image texture for mapping urban
832 land use. *Photogramm. Eng. Remote Sens.* 69, 991–1001.
833 <https://doi.org/doi:10.14358/PERS.69.9.991>

834 Hofmann, P., Blaschke, T., Strobl, J., 2011. Quantifying the robustness of fuzzy rule sets in
835 object-based image analysis. *Int. J. Remote Sens.* 32, 7359–7381.
836 <https://doi.org/10.1080/01431161.2010.523727>

837 Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep Convolutional Neural Networks
838 for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* 7,
839 14680–14707. <https://doi.org/10.3390/rs71114680>

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

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

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

847 Li, M., Bijker, W., Stein, A., 2015. Use of Binary Partition Tree and energy minimization for
848 object-based classification of urban land cover. *ISPRS J. Photogramm. Remote Sens.*
849 102, 48–61. <https://doi.org/10.1016/j.isprsjprs.2014.12.023>

850 Liu, X., Kang, C., Gong, L., Liu, Y., 2016. Incorporating spatial interaction patterns in
851 classifying and understanding urban land use. *Int. J. Geogr. Inf. Sci.* 30, 334–350.
852 <https://doi.org/10.1080/13658816.2015.1086923>

853 Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W., Munteanu, A., 2017. Hourglass-shape
854 network based semantic segmentation for high resolution aerial imagery. *Remote Sens.*
855 9, 522. <https://doi.org/10.3390/rs9060522>

856 Luus, F.P.S., Salmon, B.P., Bergh, F. Van Den, Maharaj, B.T.J., 2015. Multiview deep
857 learning for land-use classification. *IEEE Geosci. Remote Sens. Lett.* 12, 2448–2452.

858 Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional Neural Networks
859 for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.*
860 55, 645–657. <https://doi.org/10.1109/TGRS.2016.2612821>

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

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

867 Nogueira, K., Penatti, O.A.B., dos Santos, J.A., 2017. Towards better exploiting
868 convolutional neural networks for remote sensing scene classification. *Pattern Recognit.*
869 61, 539–556. <https://doi.org/10.1016/j.patcog.2016.07.001>

870 Oliva-Santos, R., Maciá-Pérez, F., Garea-Llano, E., 2014. Ontology-based topological
871 representation of remote-sensing images. *Int. J. Remote Sens.* 35, 16–28.

872 <https://doi.org/10.1080/01431161.2013.858847>

873 Othman, E., Bazi, Y., Alajlan, N., Alhichri, H., Melgani, F., 2016. Using convolutional
874 features and a sparse autoencoder for land-use scene classification. *Int. J. Remote Sens.*
875 37, 2149–2167. <https://doi.org/10.1080/01431161.2016.1171928>

876 Paisitkriangkrai, S., Sherrah, J., Janney, P., Van Den Hengel, A., 2016. Semantic labeling of
877 aerial and satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 9, 2868–
878 2881. <https://doi.org/10.1109/JSTARS.2016.2582921>

879 Pan, G., Qi, G., Wu, Z., Zhang, D., Li, S., 2013. Land-use classification using taxi GPS
880 traces. *IEEE Trans. Intell. Transp. Syst.* 14, 113–123.
881 <https://doi.org/10.1109/TITS.2012.2209201>

882 Patino, J.E., Duque, J.C., 2013. A review of regional science applications of satellite remote
883 sensing in urban settings. *Comput. Environ. Urban Syst.*
884 <https://doi.org/10.1016/j.compenvurbsys.2012.06.003>

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

890 Pingel, T.J., Clarke, K.C., McBride, W.A., 2013. An improved simple morphological filter
891 for the terrain classification of airborne LIDAR data. *ISPRS J. Photogramm. Remote*
892 *Sens.* 77, 21–30. <https://doi.org/10.1016/j.isprsjprs.2012.12.002>

893 Radoux, J., Bogaert, P., 2017. Good practices for object-based accuracy assessment. *Remote*
894 *Sens.* 9, 1–23. <https://doi.org/10.3390/rs9070646>

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

898 Reis, S., Tasdemir, K., 2011. Identification of hazelnut fields using spectral and gabor
899 textural features. *ISPRS J. Photogramm. Remote Sens.* 66, 652–661.
900 <https://doi.org/10.1016/j.isprsjprs.2011.04.006>

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

902 extraction for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.*
903 54, 1349–1362. <https://doi.org/10.1109/TGRS.2015.2478379>.

904 Sargent, I., Hare, J., Young, D., Wilson, O., Doidge, C., Holland, D., Atkinson, P.M., 2017.
905 Inference and discovery in remote sensing data with features extracted using deep
906 networks, in: Bramer, M., Petridis, M. (Eds.), *AI-2017 Thirty-Seventh SGAI*
907 *International Conference on Artificial Intelligence*. Cambridge, United Kingdom, pp.
908 131–136. https://doi.org/10.1007/978-3-319-71078-5_10

909 Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural Networks*.
910 <https://doi.org/10.1016/j.neunet.2014.09.003>

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

915 Timoshenko, S., Gere, J.M., 1972. *Mechanics of materials*. Van Nostrand Reinhold Co., New
916 York, NY, USA.

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

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

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

926 Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A
927 benchmark data set for performance evaluation of aerial scene classification. *IEEE*
928 *Trans. Geosci. Remote Sens.* 55, 3965–3981.
929 <https://doi.org/10.1109/TGRS.2017.2685945>

930 Yang, X., Qian, X., Mei, T., 2015. Learning salient visual word for scalable mobile image
931 retrieval. *Pattern Recognit.* 48, 3093–3101. <https://doi.org/10.1016/j.patcog.2014.12.017>

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

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

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

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

946 Zhang, S., Zhang, J., Li, F., Cropp, R., 2006. Vector analysis theory on landscape pattern
947 (VATLP). *Ecol. Modell.* 193, 492–502. <https://doi.org/10.1016/j.ecolmodel.2005.08.022>

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

951 Zhao, W., Du, S., Emery, W.J., 2017a. Object-Based Convolutional Neural Network for
952 High-Resolution Imagery Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote*
953 *Sens.* 10, 3386–3396. <https://doi.org/10.1109/JSTARS.2017.2680324>

954 Zhao, W., Du, S., Wang, Q., Emery, W.J., 2017b. Contextually guided very-high-resolution
955 imagery classification with semantic segments. *ISPRS J. Photogramm. Remote Sens.*
956 132, 48–60. <https://doi.org/10.1016/j.isprsjprs.2017.08.011>

957 Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep
958 Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE*
959 *Geosci. Remote Sens. Mag.* <https://doi.org/10.1109/MGRS.2017.2762307>

960

961