Remote Sensing of Environment 234 (2019) 111448

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Deprivation pockets through the lens of convolutional neural networks

Jiong Wang[a], Monika Kuffer[a,*], Debraj Roy[b], Karin Pfeffer[a]

[a] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, P.O. Box 217, 7500 AE, Enschede, the Netherlands
[b] Computational Science, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Deprivation pockets
Slums
Bangalore
Deep learning
Convolutional neural networks

ABSTRACT

Machine learning techniques have been frequently applied to map urban deprivation (commonly referred to as slums) in very high-resolution satellite images. Among these, Deep Convolutional Neural Networks have shown exceptional efficiency in automated deprivation mapping at the local scale. Yet these networks have never been used to map very small heterogeneous deprivation areas (pockets) at large scale. This study proposes and evaluates a U-Net-Compound model to map deprivation pockets in Bangalore, India. The model only relies on RGB satellite images with a resolution of 2 m as these are more commonly accessible to local urban planning departments. The experiment assumes a practical situation where only limited reference data is available for the model to learn the spatial morphology of deprivation pockets. It tests whether an updated map of deprivation pockets can be obtained with limited information. The model performance to map a large number of deprivation pockets is examined by incrementally changing the model architecture and the amount of training data. Results show that the proposed model is sensitive to the amount of spatial information contained in the training data. Once sufficient spatial information is learnt through a few samples, the city scale mapping accuracy outperforms existing models in mapping small deprivation pockets, achieving a Jaccard Index of 54%. This study demonstrated that a well-designed convolutional neural network can map the existence, extent, as well as distribution patterns of deprivation pockets at the city scale with limited training data, which is essential for upscaling research outputs to provide important information for the formulation of pro-poor policies.

1. Introduction

More than half of the world's population is living in cities, and this proportion is expected to be 68% by 2050 (UN, 2018). The rapid growth of urban population, especially in the global south, is often beyond the planning and management capability of local governments in providing housing and basic infrastructure (Hachmann et al., 2018; Martinez et al., 2008), which, among other issues, contributes to the expansion of deprived areas (often referred to as slums). Such areas are inhabited by an increasing number of dwellers deprived of durable housing and basic services (Ezeh et al., 2017; Habitat, 2003) and are significantly underestimated in their number (Hofmann et al., 2015; Taubenböck et al. 2018b, 2018c; Taubenböck and Wurm, 2015). The role of such areas is manifold. On the one hand, they pave the way for their inhabitants to urban functions, yet, on the other hand, restrain them under poor living conditions (Taubenböck et al., 2018a; Turok and Borel-Saladin, 2018). However, data on the morphology of deprived areas such as location, extent and dynamics is often not available, outdated or inconsistent.

The increasing availability of multi-temporal very high resolution (VHR) satellite image data allows earth observation (EO) based monitoring for detailed and frequent observation of urban deprivation dynamics in space and time (Kuffer et al., 2016a; Mahabir et al., 2016), and capturing spatial changes of deprivation over an arbitrary period of time (Kit and Lüdeke, 2013; Veljanovski et al., 2012). In general, EO-based deprivation mapping activities are largely based upon two premises. First, the physical appearance of a human settlement can be a strong indicator of their socio-economic conditions and can be used as a proxy to locate urban deprivation (Arribas-Bel et al., 2017; Jain, 2008; Taubenböck et al., 2009). Second, the physical appearance of deprivation can be encoded as shared image features for classifying and mapping deprivation (Graesser et al., 2012; Kohli et al., 2012; Kuffer et al., 2016b). Hence, an EO-based approach explicitly leverages the spatial information captured in images for either object or feature-based deprivation mapping (Benediktsson et al., 2003; Pesaresi, 2000; Pesaresi et al., 2008). Consequently, EO-based results can complement and help to validate the missing spatial dimension in deprivation modeling (Roy et al., 2014). However, the above premises are weakly supported due to varying deprivation morphology. For example, socio-economically deprived areas can be hidden by their physical

* Corresponding author.
E-mail addresses: j.wang2@uu.nl (J. Wang), m.kuffer@utwente.nl (M. Kuffer), D.Roy@uva.nl (D. Roy), k.pfeffer@utwente.nl (K. Pfeffer).

https://doi.org/10.1016/j.rse.2019.111448
Received 30 January 2019; Received in revised form 14 September 2019; Accepted 23 September 2019
Available online 20 October 2019

morphology while areas morphologically similar to deprived areas can be formal areas (Baud et al., 2010; Kuffer et al., 2016a; Mahabir et al. 2016, 2018). Another unsolved problem is that rule sets and feature sets are not only region specific but also image dependent (Liu et al., 2017). For instance, the variability of deprivation morphology and size has been observed at both city and global scale, indicating limited transferability of object-based rule sets and image feature sets from one case to another (Taubenböck et al., 2018a). Thus it is questionable on how to design features that best represent patterns (LeCun et al., 2015), especially due to limited knowledge of heterogeneous morphology of deprivation (Kuffer et al., 2016a). One alternative is to arbitrarily select and assess the capability of several features to capture deprivation heterogeneity (Graesser et al., 2012). However, in addition to overlooking very important features, such an approach may suffer from overfitting considering the commonly limited availability of training data within a high dimensional feature space (Huang and Zhang, 2013).

Deep learning, as a representation learning, outperforms conventional machine learning in two aspects: (1) it operates directly on raw data inputs and (2) automatically learns discriminative representations for detection and classification (LeCun et al. 2010, 2015). Deep Convolutional Neural Networks (DCNN) are one type of deep learning models that can process multi-dimensional data arrays. They have already been applied to airborne image classification (Albert et al., 2017; Längkvist et al., 2016; Maggiori et al., 2017) and also to map deprivation within cities (Li et al., 2017; Mboga et al., 2017; Persello and Stein, 2017). However, these experiments only focused on small fractions of urban areas with large contiguous patches of deprivation which have rather clear boundaries and are surrounded by distinctively different urban morphologies. In addition, sufficient labeled deprivation data in these areas in conjunction with 4-band pansharpened VHR multi-spectral images allowed to train a model with a complex architecture (Mboga et al., 2017). Some of the experiments used up to 60% of available data for training to predict the other 40% (Jenerette et al., 2016), which assumed most of the deprived information is known and set the experiment far from being realistic, where such data is commonly limited. With such ideal setups, the potential of DCNNs along with many other machine learning-based techniques are insufficiently displayed. For example, in rapidly growing cities, the locations, and in particular the boundaries of deprivation, are not available or very outdated in municipal maps. In urban planning practice, large patches of deprivation are not as common as we assume. A recent study shows that the typical size of slums can be as small as 0.016 km^2 with many concentrated towards the small end of the size distribution (Friesen et al., 2018). Given the fact that many small deprivation areas are not well captured in previous studies (Kit and Lüdeke, 2013; Wurm et al., 2017), neglecting the very small ones across the entire city, will leave deprivation dynamics largely unknown and exclude such areas from improvement programs.

This study uses Bangalore, India, as an empirical case, to explore the potential of DCNN in mapping very small deprivation areas, also referred to as deprivation pockets. The design of the study considers the characteristics of a typical city in the global south experiencing rapid urban transformation and growth, where such areas are highly dynamic, and the reference data is outdated. These deprivation pockets are commonly packed with very dense and small slum shacks with heterogeneous morphology. Many pockets are too small to meet the official minimum size criteria to be recognized by the slum map produced by the city government (India, 2015; T. Saharan, 2018). "A compact area of at least 300 populations or about 60–70 households of poorly built congested tenements, in unhygienic environment usually with inadequate infrastructure and lacking in proper sanitary and drinking water facilities" in the State/UT are categorized as identified Slums (India, 2011). This study departs from the situation of limited data accessibility, commonly found in global south cities, where only RGB images equivalent to Google Earth images are publicly and freely available for training the model. Such data are more commonly found in local planning offices compared to expensive pansharpened multi-spectral VHR images (Duque et al., 2017; Guo et al., 2016; Klaufus, 2010; Kohli et al., 2016a). Several studies addressed data accessibility issues and used the freely accessible Google Earth images (Jenerette et al., 2016; Kalma et al., 2008; Li et al., 2017), however, these relied on ideal situations where either large proportions, normally over 60%, of the deprivation pockets are known and available for model training, or where spectral bands other than RGB are used. None of the above studies addressed data accessibility restrictions, amount of known deprivation pockets, the small size of such areas, and unknown features of deprivation morphology jointly. The presented research differs from the existing studies by assuming multiple practical limitations found in a typical global south city and using DCNN as a representation learning model to resolve these limitations in support of city level small size deprivation monitoring. Collectively, the study aims to answer two questions related to data and model architecture: (1) How can limited training data incrementally bring the information of deprivation morphology to a DCNN model? (2) How can the model architecture be optimized to utilize the information contained in limited data?

2. Methodology

This study is set in the context of EO-based deprivation monitoring, where the morphology of urban deprivation is only fuzzily defined (Taubenböck and Kraff, 2014). Spatial indicators such as building size (object), density and settlement shape (settlement) or geographic location (environ) as defined by the generic slum ontology constitute a conceptual schema, which, however, needs to be adapted to local slum characteristics (Kohli et al., 2012). A recent study found large variations in urban deprivation morphologies across the globe in terms of building or shack density, height, size, orientation and settlement heterogeneity (Taubenböck et al., 2018a). Given the absence of a consistent morphological quantification for urban deprivation, this study explores the potential of DCNN in detecting very small deprivation areas through representative learning without predefined morphological indicators in a typical city in the global south. The methodological design of the study recognizes multiple practical limitations, commonly missing in existing studies, by satisfying the following real-world boundary conditions: (1) with only one deprivation pocket above the typical size of 0.016 km^2 (Friesen et al., 2018), all deprivation pockets in the study area are very small (under the typical size), (2) only very few large ones are properly labeled on an outdated reference map and can be used for training the model, (3) only regular RGB images without pansharpening are accessible, and (4) the computational cost of model training should be handled by consumer laptops/desktops.

2.1. Study area and data

The case study is set in the city of Bangalore, the administrative capital of the state of Karnataka, India (Fig. 1(a)). The most recent census report shows that the population has already reached over 8.5 million in 2011 (Chandramouli and General, 2011). In the past three decades, the officially reported population in deprivation pockets doubled comprising 8.39% of the total city population (Chandramouli and General, 2011; Roy et al., 2018), whereas potentially a large number remains unidentified (Roy et al., 2018). The large population in deprivation is scattered around the entire city, often in very small deprivation pockets with blue tent roofs (Fig. 1(b)) (Krishna et al., 2014). The average size of deprivation pockets in Bangalore, marked in 2017 by local experts, is only 1,483 m^2, being less than one-tenth of the typical size of slums (Friesen et al., 2018). The study area is covered by a tile of a WorldView-2 scene acquired by DigitalGlobe (one of the Google Earth image providers) (Fig. 1(c)). The tile covers the Bangalore East Core zone and its eastern suburbs of Mahadevapura zone (Fig. 1(c)), with a horizontal extent at the scale of 10^4 m. Such a scale approximates the definition of meso- or city-scale in many urban environments

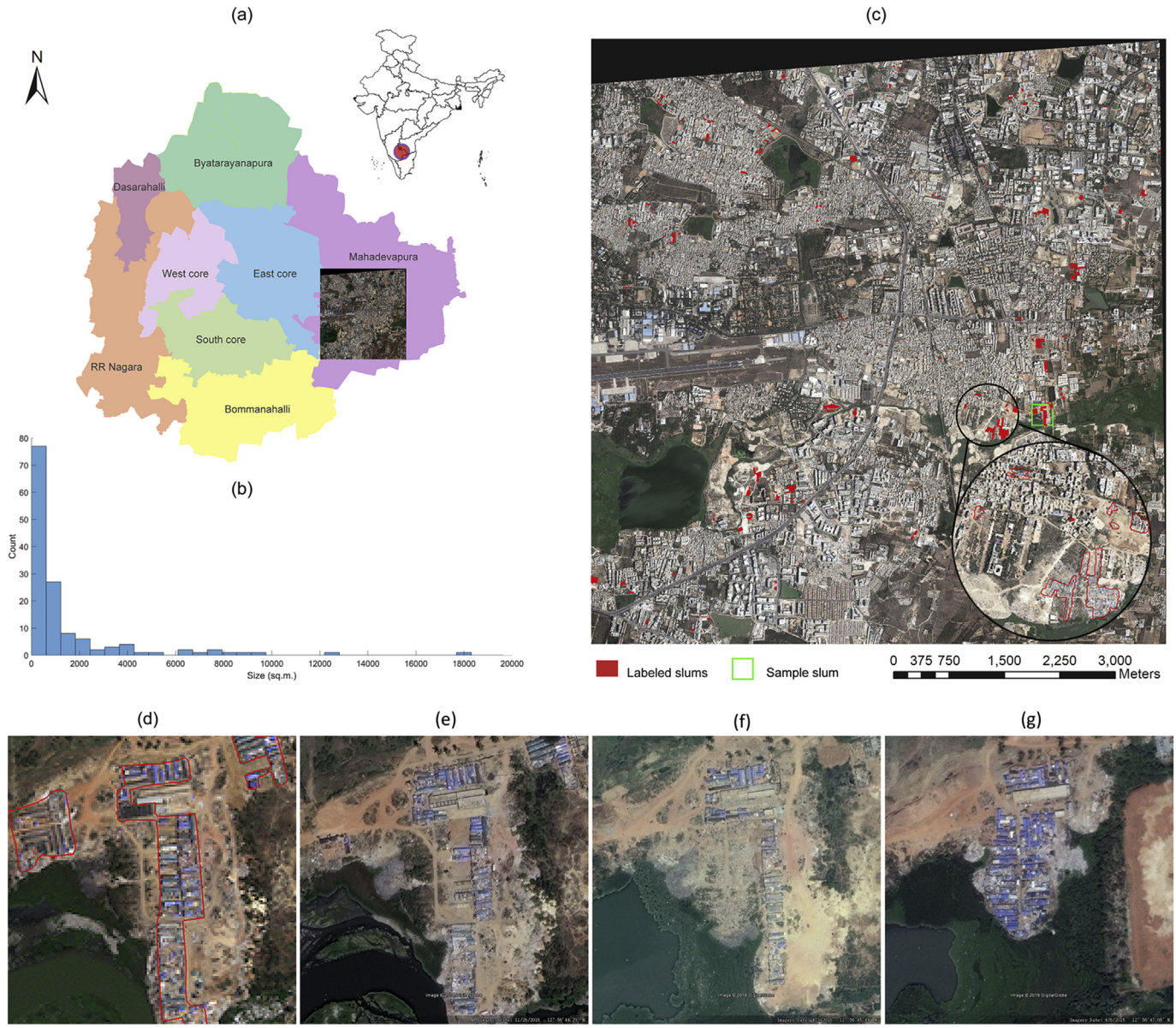

Fig. 1. Study area in Bangalore, India shown in the WorldView-02 multispectral image in February, 2017. (a) Location of Bangalore in India and the spatial extent of the study area relative to the Bangalore metropolitan area. (b) Histogram of deprivation pocket size in the study area. (c) Surveyed pockets in the study area with zoom-in snippet of contrast between deprivation and their surroundings. (d)–(g) Sample pocket morphology at the same location (from left to right) at the time of image acquisition in February 2017, June 2016, April 2016, and November 2015.

around the globe entitling this study to be city-level deprivation detection (Muller et al., 2013; Oke, 2002).

The WorldView-02 multispectral RGB provided by DigitalGlobe acquired in February 2017 for the entire study area is used as the base image for model training and testing. As the WordView-2 imagery is one of the source images of Google Earth images, we used the RGB bands of the Worldview-2 images to "simulate" Google Earth images. Within our research project, we had access to Worldview-2 images. However, most researchers in the Global South will not have easy access to such commercial images. Therefore, we restricted our methodology to work with RGB images. The spatial extent of this image is 3888 × 4096 pixels (approximately 8 × 8 km) (Fig. 1(c)) with a resolution of 2 m. The ground truth data (used for training, validation and testing) is comprised of the Google Earth image mosaic (year 2017) as the base image and associated vector reference labels, available as a base map. To obtain the most up-to-date information of existing pockets of deprivation, the DynaSlum project recruited local experts to map all deprivation areas within the city (see description at: https://www.esciencecenter.nl/project/dynaslum). The project focuses on modeling city and slum dynamics, for which a base map of deprivation areas (including pockets) is generated by a local survey using Google Earth imges combined with on-site inspection in May 2017 and used as our input data (Roy et al., 2017). Yet the base map of labeled pockets is subject to several uncertainties. The expert knowledge varies among experts in defining the boundaries of pockets even same set of visual elements such as tone, shape, size and texture on either the image or the ground are adopted (Kohli et al., 2012). In addition, the labels based upon a mosaic of multiple source images acquired at different times in a year may present inconsistent deprivation information. And a gap of few months in a highly dynamic city can cause many differences, which leads to a potential risk of feeding the DCNN with poorly labeled information and misleading the model in learning the morphology of deprivation. Among all of the surveyed 141 pockets in the study area, the average size in this area is 1,472 m^2, while the minimum and

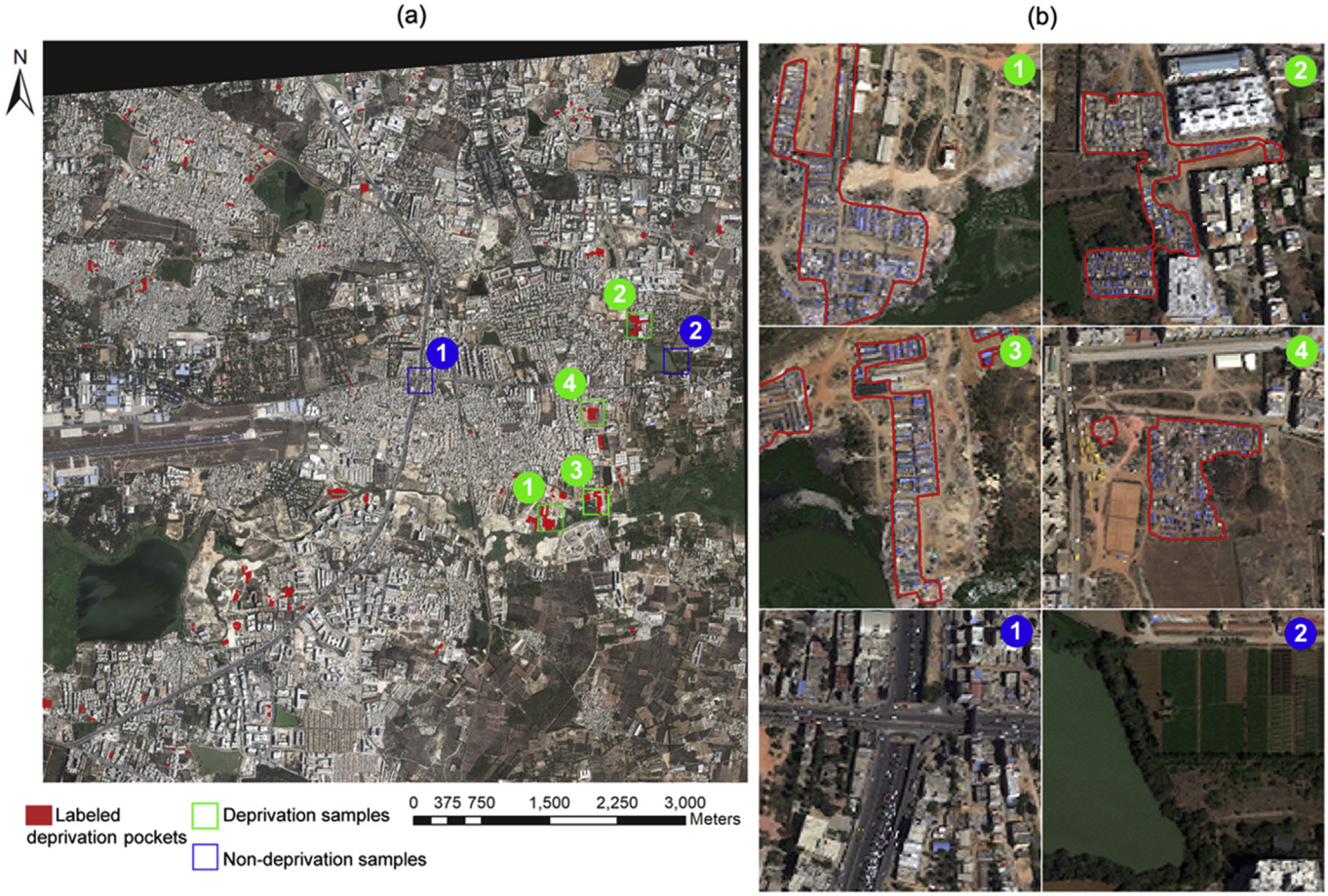

Fig. 2. Input data for model training. (a) The location of training samples numbered and highlighted by green and blue boxes in the study area, (b) zoomed-in illustration of the samples. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

maximum values are 31 m^2 and 18,052 m^2, respectively. The largest pocket is thus only slightly larger than the typical size (0.016 km^2) as found by Friesen et al. (2018). Due to the very small size, the physical appearance of pockets can be sensitive to the change of even a few numbers of shacks. Even the third largest pocket in the study area highlighted by a green box (Fig. 1(c)) displays a significant difference due to the change of few shelters. While this pocket is dominated by elongated shelters in February 2017 (Fig. 1(d)), its morphology is significantly different in mid-2016 with few incomplete shelters (Fig. 1(e)). And it, in fact, evolved into an entirely different morphology within only half a year between the end of 2015 (Fig. 1(f)) and mid-2016 (Fig. 1(g)).

To explore the potential of DCNN in learning the deprivation morphology and mapping, this study adopts the 'typical size of slums' (Friesen et al., 2018) and considers it as a rough threshold to choose training samples from the study area. Thus only the four largest pockets with the size at the level of $S \sim 10^{-2} km^2$ are selected (labeled in green in Fig. 2(a)) for the following rationale: (1) the selections are significantly larger and more likely to be identified (also known in official data) than other pockets in the study area, and (2) reference data in cities like Bangalore with many small deprivation pockets will be more reliable for larger pockets than for the smaller ones as a change of few shacks in a pocket can significantly modify its physical appearance. Another two patches of non-deprivation sample areas are also selected as training data to inform the model about non-deprivation morphology (Fig. 2(a)). An zoomed-in visualization of the samples are shown in Fig. 2(b). The selected deprived pockets comprise 3% of the total number of such areas in the study area (Fig. 2(a)). These areas are 18,052 m^2, 12,749 m^2, 9,691 m^2, and 9,008 m^2, respectively and marked sequentially from 1 to 4 (Fig. 2(b)). The size of the fifth largest pockets drops to 8,051 m^2. Overall, 115 out of the total of 141 pockets in the selected study area are well below 2,000 m^2 and 94 are below 700 m^2 (Fig. 1(b)).

The experiment starts with a test of the model performance at the local level, where prediction in a small area is needed with large fraction of deprivation areas is known. Then the challenge of detecting small deprivation pockets is rendered by involving the entire study area at city level. The major steps of the experimental workflow are shown in Fig. 3.

2.2. Deprivation pockets mapping through the U-Net-CPD

To learn the information from limited training samples, the U-Net DCNN is chosen as the starting point as its architecture has been proved to be efficient in dealing with limited training samples of either medical or satellite image data (Iglovikov et al., 2017; Iglovikov and Shvets, 2018; Ronneberger et al., 2015). Here, the original U-Net is modified by adding a series of dilated convolutional operations right at the beginning of the network to produce multi-scale low-level feature maps before information loss through the convolutional and max pooling operations (Fig. 4). The U-Net in its compound form (U-Net-CPD) is a fully convolutional network (FCN) that takes input image patches of arbitrary size and generates an output of dense pixel level segmentation maps of equal size (Long et al., 2015). Since the original FCN upsamples predictions directly back to the size of the input image patch, the final segmentation may suffer from coarse prediction boundaries. Instead, the U-Net-CPD, as compared to the original U-Net, takes advantage of the encoder-decoder architecture, which has been applied in other

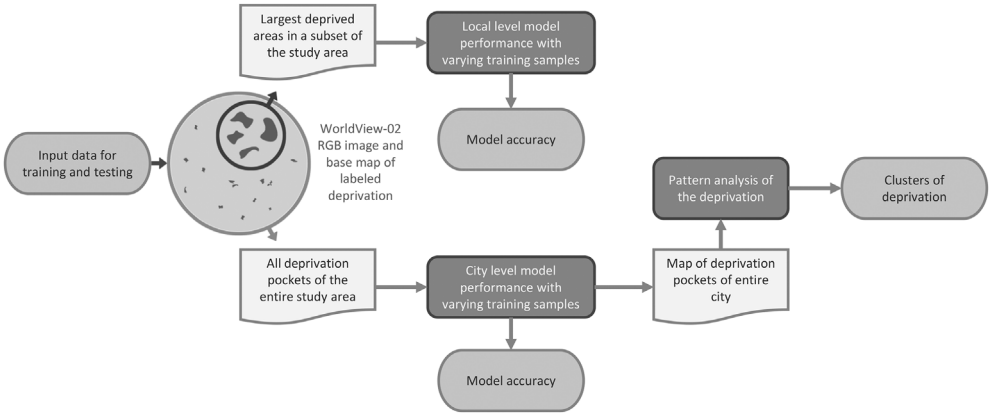

Fig. 3. Workflow with major steps involved in the experiment.

networks such as the SegNet for semantic segmentation (Badrinarayanan et al., 2015). The encoder component continuously applies 3 × 3 convolutional kernels and 2 × 2 max pooling operations to extract maps with hierarchical features such as edges, shapes, and objects, while the decoder incrementally upsamples the feature map by using the extracted feature maps as guidelines to resolve the segmentation boundaries. By copying and concatenating the hierarchical feature maps to each of the upsampling steps, the U-Net-CPD recovers the predictions to the size of the input image with a dense pixel level segmentation and clear boundaries. The encoding comprises of 3 × 3 convolutional kernels, which may be insufficient to capture the edge information of objects with different sizes. For instance, the kernel may capture the edges of dwellings yet fail in delineating the boundaries of pockets. Thus dilated kernels are employed to capture the low-level features such as edges at the input block of the model. These dilated kernels maintain the number of weights in the kernel while expanding the field-of-view of the kernel by inserting zeros in the kernel. In this way, the dilated kernel with an expanded field-of-view is capable of

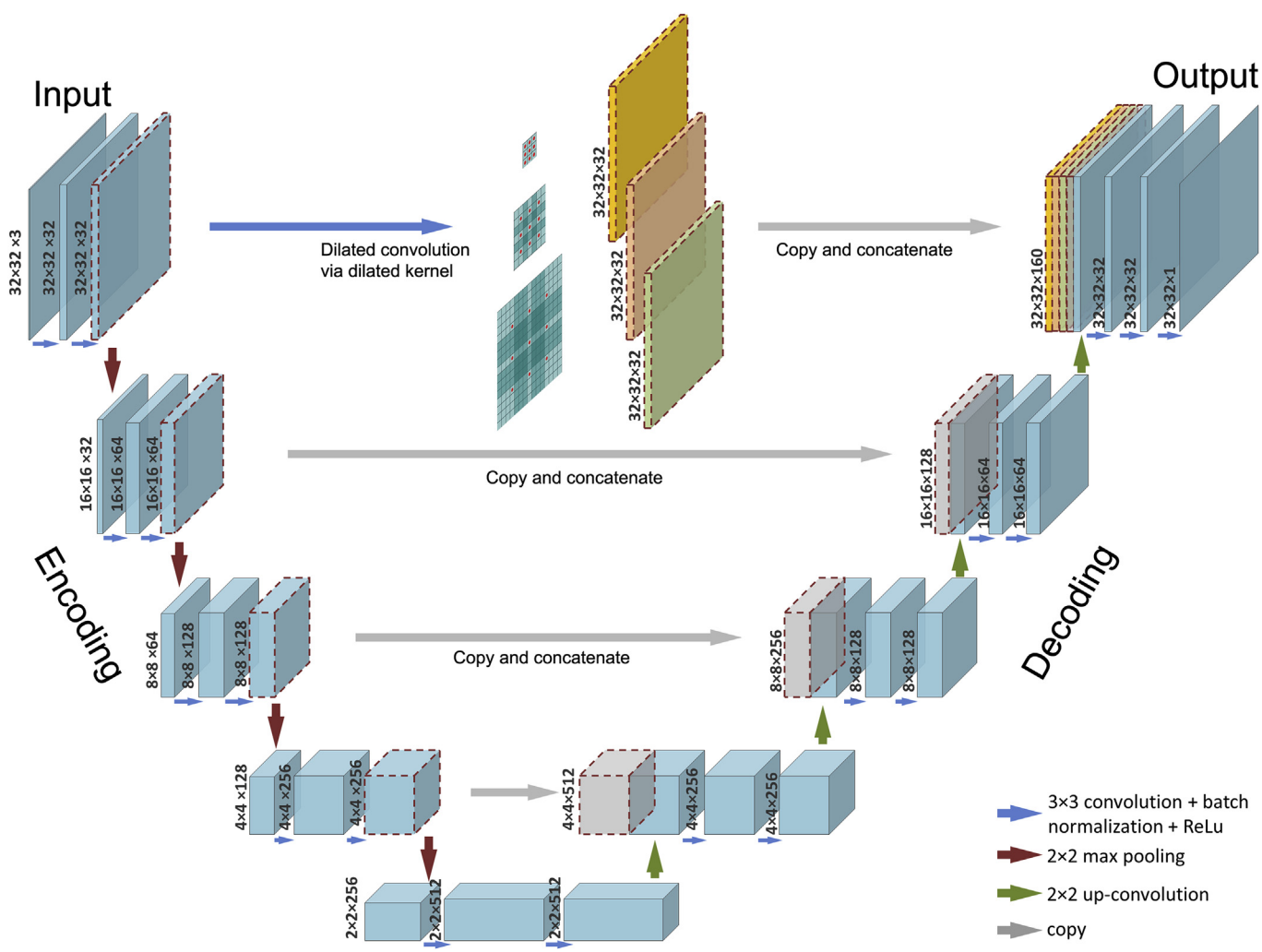

Fig. 4. The U-Net-CPD DCNN with encoder-decoder architecture combined with dilated kernels for multi-scale low-level feature extraction.

capturing low-level edge features at different sizes or scales (Yu and Koltun, 2015). Imposing such low-level edge information is expected to improve the mapping of boundaries in the output as prediction accuracy can be very sensitive to small pockets of different sizes. As the input image patch is of 32 × 32, only three dilated kernels with dilation rates of 1, 2 and 4 are used to produce field-of-views of 3 × 3, 7 × 7 and 15 × 15, respectively (Yu and Koltun, 2015). The dilation rates ensure the field-of-views are restricted by the size of the input image patch.

With a resolution of 2 m, even the largest deprivation pocket is comprised of only a few thousands of pixels, and limited numbers of discriminative image patches with the size of 32 × 32 can be drawn. Therefore, intense data augmentation is applied to generate discriminative samples (Iglovikov et al., 2017; Ronneberger et al., 2015). The samples are produced by first drawing a large amount of 32 × 32 samples without considering whether the samples overlap or not and then by applying augmentation to increase the variation of the samples. Augmentation includes random rotation, shifting, flipping, minimal shearing and stretching, and is restricted affine transformations. It simulates the variations of deprivation morphologies and a small amount of sensor distortion. During training, the input data is split into 70% and 30% for training and validation, respectively.

2.3. The strengths and weaknesses of the U-Net-CPD

The performance of the proposed U-Net-CPD is first evaluated for small fractions of the study area containing the larger deprivation pockets. This local-scale analysis focuses only on the four largest samples and evaluates how the proposed model responds to incremental information contained in the training data. It is similar to a few previous studies where only small and homogenous areas of deprivation were used to evaluate the model performance (Mboga et al., 2017; Persello and Stein, 2017). These studies assumed that most areas of deprivation in a small urban area are known, and only a small part had to be predicted. These assumptions are ideal to reach high prediction accuracy but are not very realistic for providing information to urban planning and decision support. Yet they can set the starting point to understand the learning and predicting power of the U-Net-CPD.

Next, we investigate the prediction power of the U-Net-CPD at the city scale by adding incremental information of deprivation. Fully convolutional neural networks (FCN) with dilated kernels (DK) as well as the original U-Net used in a previous study (Demir et al., 2018; Iglovikov et al., 2017; Iglovikov and Shvets, 2018; Li et al., 2018; Seferbekov et al., 2018) for local level slum prediction and land use classification are employed for benchmarking. These models are FCN with 4 and 6 layers of dilated kernels (FCN-DK4 and FCN-DK6) and U-Net. The performance of the U-Net-CPD will be visualized to examine the morphology of correctly and falsely predicted deprivation pockets.

2.4. Accuracy assessment

Assuming limited quality in the reference data caused by temporal changes and manual delineation, two scenarios are formulated to capture the deprivation on the ground:

(1) The prediction shows agreement with the reference data in capturing deprived areas on the ground (Fig. 5(a)), and (2) both the prediction and reference data partially capture parts of the deprived areas without full agreement (Fig. 5(b)).

Therefore, accuracy assessment metrics regarding how the prediction resembles the location and extent of areas delineated by the reference data are used. The primary accuracy metrics is the Jaccard Index (Jaccard, 1912), also known as intersection over union. It is a very restrictive metric evaluating the similarity between two datasets and has been applied as an area-based accuracy assessment in image analysis (Hernandez-Stefanoni and Ponce-Hernandez, 2004; Singh and Garg, 2013). Here, the accuracy of prediction regarding the extent of deprived areas as denoted by the reference data is measured through:

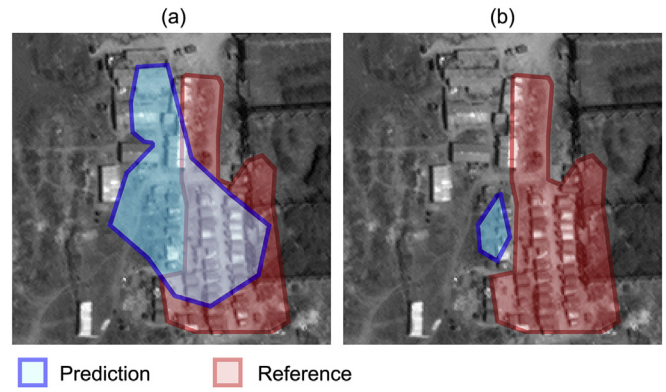

Fig. 5. Two scenarios of prediction: (a) prediction partially resembles the extension of a deprivation pocket denoted in the reference data, and (b) prediction fails to capture the extension of the deprivation pocket denoted by the reference data yet correctly locates the existence of the pocket which had not been included in the reference data.

$$J(\text{Prediction, Reference}) = \frac{|\text{Prediction} \cap \text{Reference}|}{|\text{Prediction} \cup \text{Reference}|}. \quad (1)$$

The second metric is the existence accuracy of prediction, assessed by searching within a buffer zone at the location of the reference label. The search area is the circumcircle of the smallest bounding box over the reference label. Once the prediction is found in that search area, it is considered as a correct existence prediction. To compare achieved accuracies with those of previous studies, a third metric, the more conventional producer accuracy (PA) is employed for comparison.

2.5. Pattern analysis

Besides mapping individual areas of deprivation in terms of extent and location (Kuffer et al., 2018), investigating the model performance from a geographic perspective helps to understand the collective patterns of deprivation process. Since deprivation information should possibly not be made publicly available at resolutions that could harm individual and group privacy, spatial clustering is deployed at different scales to study the deprivation distribution captured by the model from local to city scale. The multi-scale distribution of predicted deprivation is compared to the one of the reference data by (1) using the Ripley's K-function to investigate the level of concentration of deprivation compared to a random distribution, and (2) visualizing kernel density of clusters at different scales in the study area.

3. Results

The results show the model performance at both local and city scale, the impacts of incremental training samples, model performance comparison with a fixed amount of training samples, model performance from a geographic perspective, model operation through the lens of convolutional kernels, and the weakness and strength of the model performance.

3.1. Model performance at the local level

For local scale analysis, the largest deprivation pocket (out of the four largest ones) is used for training, and the remaining three are used for testing. It means that around 37% of the information about the deprivation pockets is available for training the U-Net-CPD to predict the remaining 63%. The training follows the 70/30 rule to further partition the known 37% deprived areas into 70% and 30% for training and validation. Two samples without deprivation, i.e. negative samples,

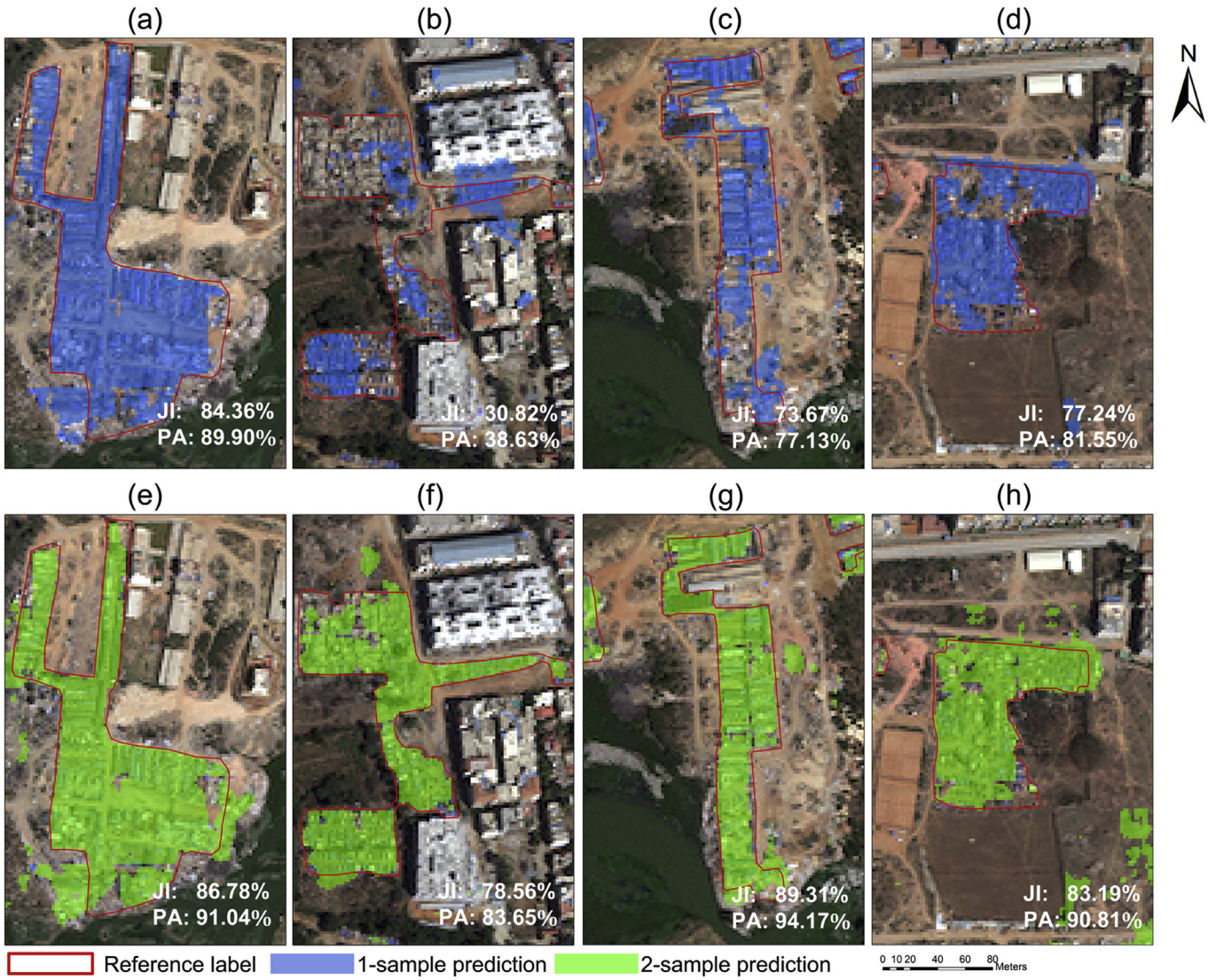

Fig. 6. The local level analysis of model performance on large deprivation pockets. Training with the largest pocket shown in (a) with augmentation and prediction of all the top four largest areas shown in (a)–(d). Training with the top two largest pockets shown in (e) and (f) with augmentation and prediction of all the four largest pockets shown in (e)–(h).

are also selected as shown in Fig. 2. Using the Jaccard Index as accuracy metrics, the training and validation converge at around 98% after 100 epochs of training. Since the training data has been augmented and can be slightly different from the original inputs, the study also examines how the trained model performs on the original data. Then the largest pocket used for training is also fed into the trained model along with the test data of the other three largest deprivation pockets. Similarly, the largest two areas equivalent to 62% information of all the four largest areas are then used to train the model, achieving a training accuracy of 98%.

Although only the largest pocket is used for training, the predictions for the third and fourth largest pockets accuracy are above 70% (Fig. 6(a)–(d)). As the model achieves training accuracy of 98% on augmented data, the prediction accuracy on the actual largest pocket is 89.9%. The poor prediction in Fig. 6(b) (Table 1) is caused by the model's failure in learning relevant deprivation morphology from the training data. However, the data augmentation helps to generalize the spatial morphology (Fig. 6(a)) so that the model can still partially capture varying deprivation morphologies (Fig. 6(b) and (d)). As the morphology in Fig. 6(c) is similar to the training data (Fig. 6(a)), most of the deprivation pockets are successfully predicted and labeled. The relatively low prediction accuracy of 73.67% in Fig. 6(c) compared to

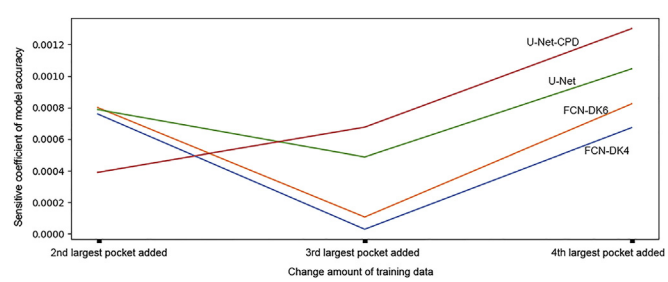

Fig. 7. Local sensitivity coefficient of model prediction over a changing amount of training data.

the accuracy obtained in Fig. 6(d) can be attributed to the non-deprivation area falsely included in the reference data. This highlights the influence of the uncertainties in the reference data (see Fig. 7).

When the largest two pockets (Fig. 6(e) and (f)) are used for training, the spatial information is better captured in Fig. 6(h) than in Fig. 6(d) with an accuracy of 83.19% as the model is able to learn a similar morphology shown in Fig. 6(f). This highlights the DCNN's sensitivity to the spatial morphology in the image. The accuracies

Table 1
The performance of the U-Net-CPD at the local level shown as Jaccard Index (JI) and producer accuracy (PA).

	Training with 1 sample		Training with 2 samples	
	PA	JI	PA	JI
Largest pocket	89.90%	84.36%	91.04%	86.78%
2nd largest pocket	38.63%	30.82%	83.65%	78.56%
3rd largest pocket	77.13%	73.67%	94.17%	89.31%
4th largest pocket	81.55%	77.24%	90.81%	83.19%

achieved in this local level experiment are summarized in Table 1. In addition to the Jaccard Index (JI), the producer accuracy (PA) is provided.

3.2. Model performance at the city level

This study found that the prediction accuracy can be quite low at the city level especially when training data include only a small fraction of the large number of deprivation pockets, commonly of much smaller size. When the training sample is increased, namely from the largest pocket to using the largest two, the prediction accuracies of all the models improve significantly. For instance, the Jaccard index (JI) of U-Net and FCN-DK6 increased by 10% from 22.64% and 11.19% to 32.68% and 21.39%, respectively (Table 2). This indicates that the second largest pocket as shown in Fig. 6(b) and (c) adds abundant information for the model to learn 10% more about all the pockets in the study area. In contrast, the third largest pocket contributes less information to the model to learn the spatial morphology of the areas. This is particularly prominent for the FCN-DK models as the JI only improves from 21.05% and 21.39% to 21.33% and 22.42%, respectively. Since the amount of training data is the only parameter guiding the training of the model, it is common to conduct a local sensitivity analysis of the effect of changing training data as opposed to a global sensitivity analysis to reduce computational expense (UN, 2018). The local sensitivity is measured by the local sensitivity coefficient approximated by the first-order coefficient in the Taylor series expansion of the changing accuracy against the changing amount of training data. Often, the coefficient is denoted as $\frac{\partial Y}{\partial p}$, where Y is the accuracy output measured by the JI, and p is the model parameter measuring the amount of training data in this case. In Fig. 7, the U-Net and U-Net-CPD seem to be more sensitive to the added information of the third largest pocket. Then the fourth largest pocket introduces additional morphological information to capture deprivation in the study area, which again can be observed through the improved performance of FCN-DK models. Although the U-Net-CPD is the least sensitive to the extra information brought by the second largest pocket, the model already obtains significantly higher accuracy by using only the largest pocket (Table 2). Another potential reason of limited sensitivity to the added information of the second largest pocket is that the U-Net-CPD is a more complicated architecture, which demands more training data for improving the prediction accuracy. Through the process of increasing training samples, the performance of U-Net and U-Net-CPD improves steadier than the FCN-DK models implying the U-Net models learn and generalize added and augmented information more efficiently.

Analyzing the prediction accuracy of all the deprivation pockets individually in the study area brings insights into the learning and mapping mechanisms of the models. In the scenario of using all the four largest pockets as training data, the U-Net-CPD performs quite uniformly on all the 141 deprivation pockets (Fig. 8(d)), which leads to an average accuracy of 53.99% over the entire study area (Table 2). In comparison, the other models display a major weakness in predicting small pockets (Fig. 8(a)–(c)). In particular, the FCN-DK models perform similarly with slight improvements in predicting larger pockets by increasing dilated convolutional layers from 4 to 6. However, these models still miss most of the small pockets with zero JI accuracy (Fig. 8(a) and (b)). At this point, it can be confirmed that the U-Net-CPD outperforms the other models mainly on predicting small pockets, which is attributed to the multi-scale low-level feature extractor. The extracted low-level features such as edges help to resolve the prediction boundaries, which significantly impact the accuracy in predicting very small pockets.

3.3. Insights through the U-Net-CPD

Apart from comparing the models by both varying and fixed number of training samples, a visual interpretation is provided to investigate how the model sees and learns from the data through the lens of the convolutional kernel (Fig. 9). An input patch with deprivation pockets is used for illustration (Fig. 9(a)). The patch is located in the south part of the second largest pocket and can be identified in Figs. 1(b), Fig. 6(b) and (f). The pocket is highlighted by a red line. A panchromatic image of the same patch is also provided for visualizing the details (Fig. 9(b)). When the patch is fed into the U-Net-CPD trained on the largest pocket, 32 feature patches are produced (Fig. 9(c)) by the first convolutional block with a size of 32 × 32 (Fig. 4). These low-level features extracted or "seen" by the model are expected to be edges, shapes or brightness. However, since the model is only trained on one area with images of 2-m resolution, the features seem to be blurred and the boundaries between deprivation and non-deprivation are also unclear. The model trained by using only the largest pocket insufficiently maps the second largest one in Fig. 6(b). Once the model is trained by all four largest pockets, the feature patches produced by the same convolutional block of the model are less blurred and more meaningful for interpretation. For instance, the number 0 feature patch in Fig. 9(d) highlights the lighter roofs while number 1 and 9 highlight most of the vertical edges. Some kernels may have learnt to be sensitive to colors thus producing feature patches as number 20 and 24 in Fig. 9(d), where blue roofs likely activate brighter pixels in the feature patches. These low-level features could be further weighted and combined to produce high-level features such as shack clusters and neighborhoods, where clusters of shelters are recognized as deprivation pockets.

3.4. Distribution patterns of slums

Due to extensional uncertainties in predicting the boundaries of deprivation pockets, this study further analyzes the possibility to

Table 2
The U-Net-CPD performance on predicting deprivation pockets at city level benchmarked by FCN-DK models and the original U-Net. Metrics are producer accuracy (PA), Jaccard index (JI) and existence accuracy (EA).

	Training with 1 sample			Training with 2 samples			Training with 3 samples			Training with 4 samples		
	PA	JI	EA	PA	JI	EA	PA	JI	EA	PA	JI	EA
FCN-DK4	13.22%	11.37%	63/141	24.82%	21.05%	73/141	28.74%	21.33%	84/141	30.61%	27.39%	92/141
FCN-DK6	16.21%	11.19%	57/141	25.15%	21.39%	79/141	29.37%	22.42%	82/141	34.13%	29.84%	88/141
U-Net	31.09%	22.64%	68/141	42.41%	32.68%	89/141	43.76%	37.40%	88/141	61.55%	46.82%	102/141
U-Net-CPD	36.69%	30.75%	74/141	44.36%	35.72%	94/141	52.95%	42.27%	106/141	70.41%	53.99%	131/141

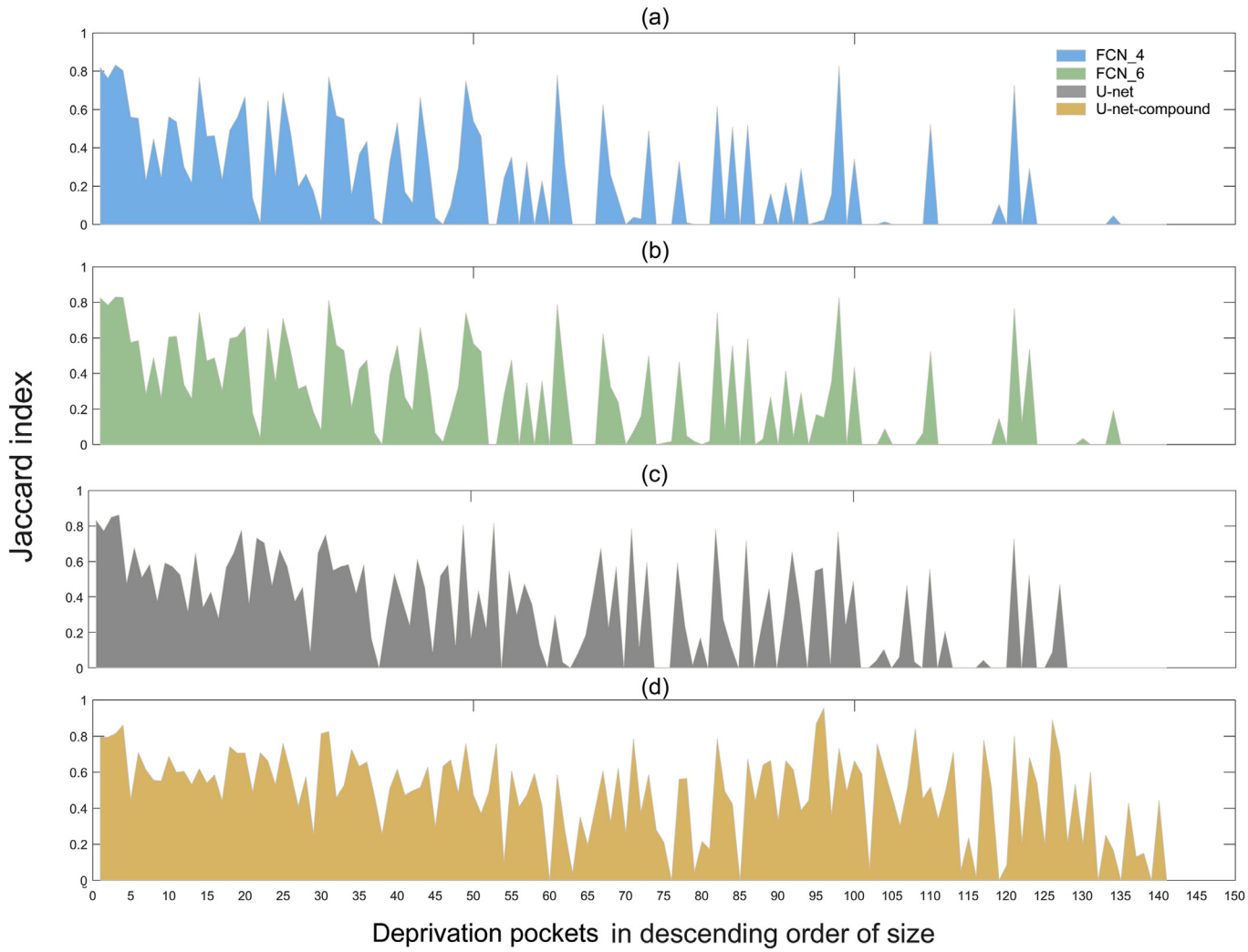

Fig. 8. Model performance on individual pockets in the study area by using the four largest pockets as training data. The FCN-DK4 and FCN-DK6 prediction accuracies (JI) are shown in (a) and (b). The U-Net and U-Net-CPD performances are shown in (c) and (d), respectively.

capture the existence of deprivation in the form of the spatial distribution density. In the evaluation of the model performance of capturing the distribution patterns of deprivation pockets, the Ripley's K-function shows that areas labeled in the reference data display a strong concentrated pattern compared to a random pattern below a scale of 1800 m, where the measured K value falls under the expected K value (Fig. 10(a)). At scales larger than 1800 m, the red curve is below the expected random distribution denoted by the blue line, indicating a more sparse distribution than random patterns. The predicted pockets in the study area show a similar distribution compared to the reference data. However, the concentration is only valid at scales below 1500 m (Fig. 10(b)), and the level of concentration is slightly lower than the one in the reference data and exhibits patterns close to a random distribution. Thus, inference with regards to the clustering patterns of deprivation pockets is only valid within 1500–1800 m, where distribution is not sparse and random. At the scale of the study area, the pocket distribution can be considered as sparse and random implying deprivation as a pervasive phenomenon around the entire city.

Clusters of deprivation pockets can be visually explored through a kernel density analysis given a properly selected kernel size. According to the results from the Ripley's K-function, three kernel sizes are used within 1500 m with increments of 500 m. The cluster density is also weighted by the size of deprivation pockets so that a high probability value indicates the concentration of deprivation with large size. For each of the kernel sizes, the patterns in reference data (Fig. 11(a)–(c)) and prediction (Fig. 11(d)–(f)) are visually similar, meaning geographic patterns observed in reference data matches the prediction.

Choosing 500 m as the kernel size means that all deprivation pockets within a distance of 500 m are considered as one cluster. While few clusters can be found to the south of the study area in both the reference data (Fig. 11(a)) and prediction (Fig. 11(d)), many high density spots only highlight individual pockets as “self-evident” patterns. If 1500 m is selected as the kernel size, clusters can still be identified as the kernel size is within the threshold found by the Ripley's K-function. However, in both the reference data (Fig. 11(c)) and prediction (Fig. 11(f)), the density is rather flat. The spread of the contour lines indicates that only a weak concentration is found with the large kernel size. The kernel size of 1000 m appears to be neutral compared to the larger and smaller kernel sizes. The clusters are quite preeminent in both the reference data (Fig. 11(b)) and prediction (Fig. 11(e)) indicating that clustering of deprivation pockets can be found at the scale of 1000 m. Thus, it is more likely to observe pockets within than beyond 1000 m from any existing pocket in the study area.

3.5. Weakness and strength of the U-Net-CPD

Missed pockets in the prediction highlight the weakness of the U-Net-CPD. Typical samples are displayed where the model failed to capture deprivation pockets (Fig. 12). These missed areas are evaluated with JI accuracy of 0, which can be observed in Fig. 8(d). The largest

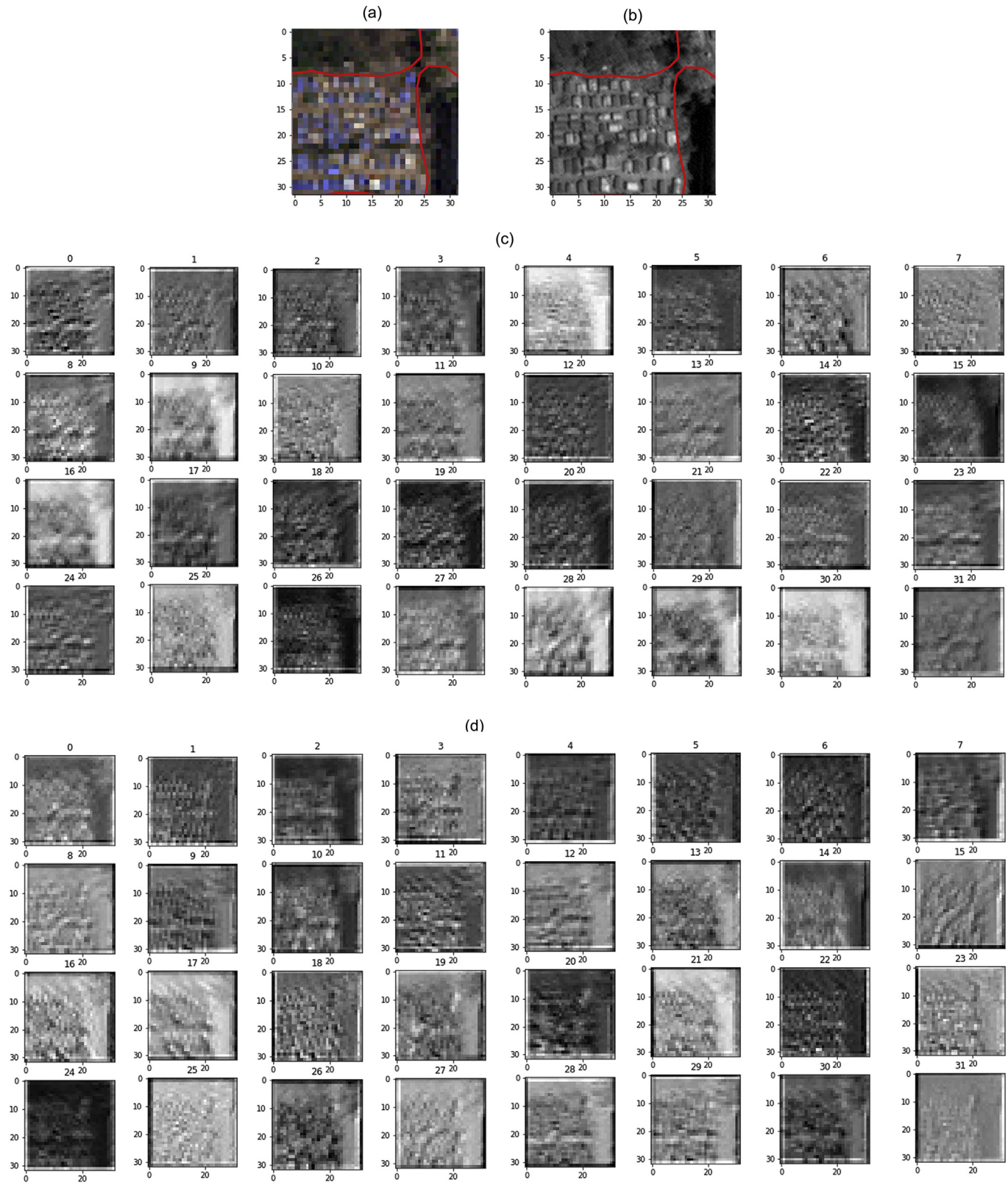

Fig. 9. Sample 32 × 32 patch seen by trained model. (a) Sample patch fed into the model with reference label. (b) Same sample patch shown by panchromatic image with a resolution of 0.5 m for visualizing the details of the morphology of deprivation. Low level features seen through the 32 kernels at the first convolutional block of the U-Net-CPD trained by (c) the largest pocket and (d) both the largest and second largest pockets.

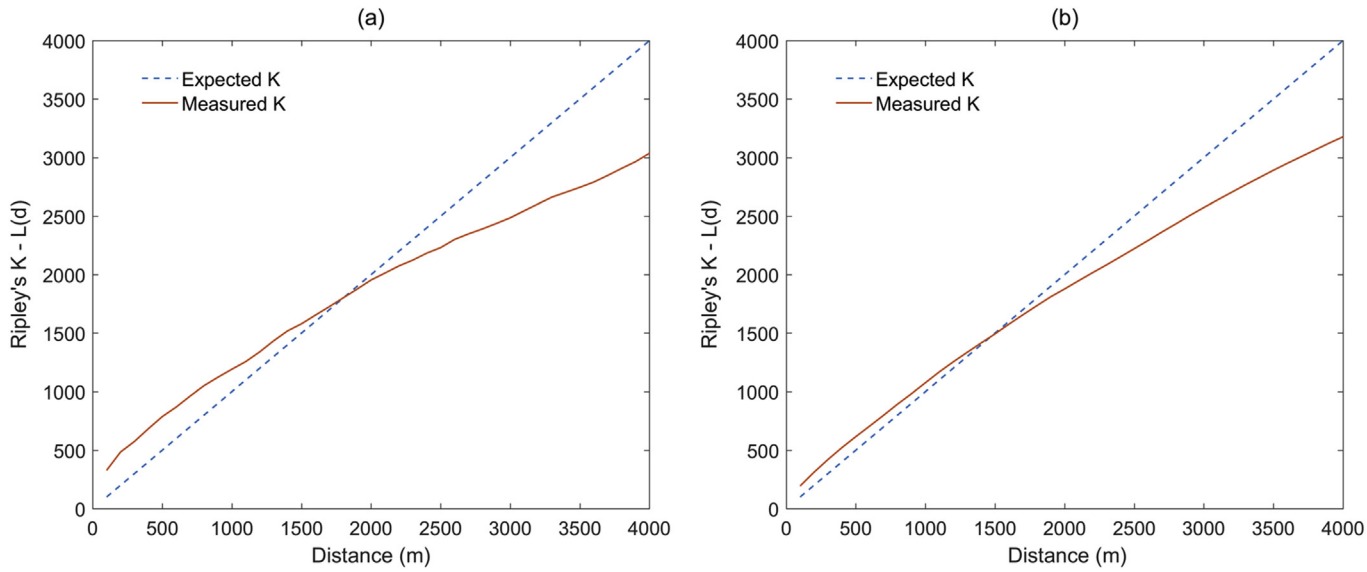

Fig. 10. Multi-scale cluster pattern analysis through the Ripley's K-function. (a) Level of clustering of deprivation pockets at different scales in reference data; (b) Level of clustering of predicted pockets at different scales.

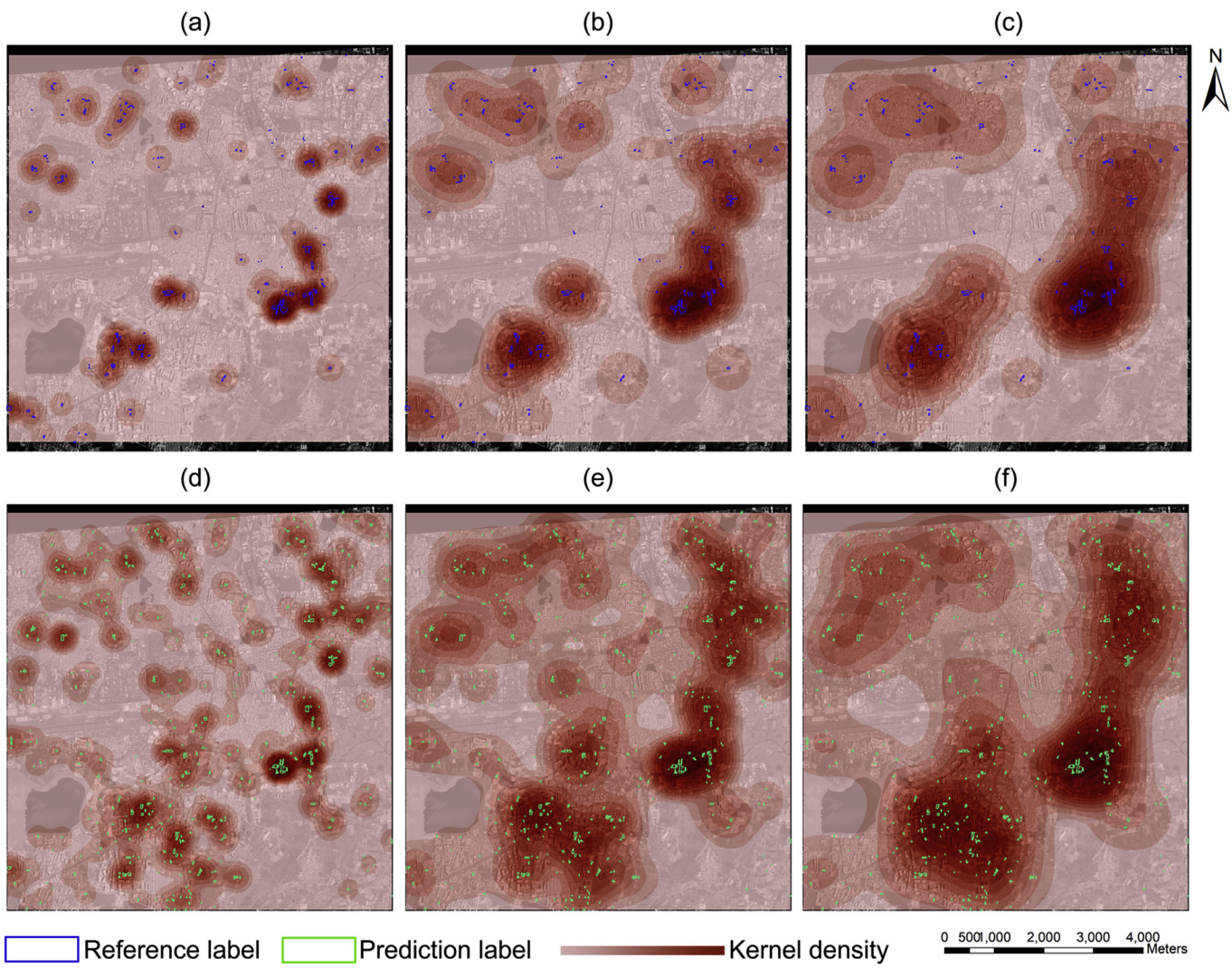

Fig. 11. Kernel density analysis of deprivation pocket clustering overlaid on the satellite image with a kernel size of 500 m, 1000 m, and 1500 m for reference data in (a)–(c) and prediction in (d)–(f).

Fig. 12. Four examples of omissions in predicting deprivation pockets labeled by the reference data: (a) Few scattered dwellings labeled roughly by reference data, (b) extremely small dwellings interleaved with trees, (c) narrow stripes of deprived areas comprising of extremely small dwellings, and (d) dwellings within formal built-ups.

missed pocket is ranked at 60th in size among all 141 areas (Fig. 12(a)). Other missed ones shown in Fig. 12(b)–(c) are ranked at 76th, 85th, and 119th. A reason for this omission might be that the model has been trained on typical data with low intra-class variance. The first row in Fig. 12 are original images with a resolution of 2 m fed into the model while the second row shows corresponding areas in panchromatic images with a resolution of 0.5 m. The 60th largest pocket is in fact roughly labeled by the reference data with only three small dwellings in the labeled extent (Fig. 12(a)). The pocket in Fig. 12(b) is interleaved with trees, and there is no similar morphology in the training samples for the model to learn. While the model fails to capture the narrow pocket stripe mixed with trees, the model cannot sufficiently capture or overestimates small pockets to the north part of the image (Fig. 12(c)). When very few small deprived dwellings are surrounded by non-deprivation built-up areas with a different morphology, the model may still fail to distinguish the deprivation pockets (Fig. 12(d)).

The strength of the U-Net-CPD is highlighted by samples that are insufficiently labeled or omitted in the reference data but detected by the model (Fig. 13(a) and (b)). Similarly, original images and corresponding panchromatic images are provided for visualization. The model performance is difficult to assess when predictions are morphologically similar to deprivation but omitted in reference data. These poorly built shacks may be located at either construction sites (Fig. 13(c)) or surrounded by bare land (Fig. 13(d)) at the periphery. It is thus difficult to assess if these model outputs are false predictions.

4. Discussions

4.1. Data and model performance

The model performance largely depends on the amount of training data relative to the size of the study area. Local level prediction of large deprivation pockets is much more promising than predictions of morphologically diverse small pockets scattered across the city. At local level, it is very likely to achieve a high producer accuracy once one third or even more than half of all the deprivation pockets are known. On the one hand, this confirms the strength of DCNN models. On the other hand, sufficient high-quality training data limited the difference among the performances of models as the dataset is always likely to bring acceptable results through several models.

In practice, training data for urban deprivation detection at a large scale is commonly very limited. The conclusion about model performance at a local scale can hardly be generalized to a city level. Compared to other natural image segmentation (Chen et al., 2018; He et al., 2016; Hoo-Chang et al., 2016; Martin et al., 2001; Pal and Pal, 1993), satellite image segmentation is restricted by data availability mainly due to limited access to VHR image data in economically resource-constrained areas. Segmentation with regard to deprivation detection is further restricted by a lack of reliable ground truth data for model training. However, DCNN is very sensitive to the spatial information contained in data. As shown in section 3.1, a slight change in the diversity in even a very small amount of training data can impact the way the model sees and predicts samples. Information increments in

Fig. 13. Four examples of capturing deprivation pockets omitted in reference data: (a) partially labeled areas complemented by model prediction, (b) completely omitted area detected by the model, (c) construction site with morphological similarity detected by the model, and (d) group of morphological deprivation pockets with elongated shapes are captured.

training data would help the model to learn more generalized information about the target deprivation morphology, while misleading information would send the model into the trap of "garbage in, garbage out" (GIGO). So dealing with limited training data is an ongoing discussion in computer science and the machine learning community (Cui et al., 2015; Dundar et al., 2015; Wang et al., 2015). Many of the most recent application of the U-Net explicitly addressed intense augmentation of limited training data to generate more diverse and generalized training samples (Çiçek et al., 2016; Dong et al., 2017; Iglovikov and Shvets, 2018; Ronneberger et al., 2015). Unfortunately, in the application of DCNN to deprivation detection, such discussion is rarely found. Studies exclusively display the strength of proposed models and tend to show only part of the story and miss to discuss application relevant limitations for deprivation mapping (Ibrahim et al. 2018a, 2018b; Li et al., 2017; Mboga et al., 2017).

4.2. Learnt features and model performance

Resolving the prediction boundaries produced by DCNN models is a major theme in improving the semantic segmentation results. Utilizing low level features learnt at the first few blocks of DCNN to reconstruct the details of inputs has proven to be beneficial in not only natural image segmentation (Kavukcuoglu et al., 2010; Lee et al., 2016) but also deprivation mapping in this study. The benefit is more prominent in mapping small deprivation pockets than large ones as few pixels of the boundary shift may significantly impact the extent and existence of very small pockets. Thus concatenating multi-scale low-level features in the U-Net-CPD largely improves the JI accuracy of small deprivation pocket prediction. So far there is hardly any evidence that bias exists in the model in detecting the morphology of small pockets as the model can either underestimate or overestimate the extension of small pockets (Fig. 10(c)). Yet it is quite convincing that the DCNN models inaccurately predict many small pockets because small pockets are sensitive to falsely predicted boundaries.

Using low-level features is non-trivial since they may include many specifications other than edges such as colors, contrasts and brightness. Thus, it is expected that low-level features can be further explored, better understood and used more efficiently. The further application of learnt features for other classification tasks directly reduces to sufficient understanding and interpretation of the learnt features. Fortunately, the information seen by the kernel can also be seen by humans for visual exploration. In this study, only features produced at low level are visualized. A more systematic investigation is recommended to understand how low-level features are weighted and combined to activate a segmentation of objects with clear boundaries. Understanding the features can be also useful to produce a rich and discriminative feature space because features are automatically learnt as opposed to artificially designed with potentially insufficient prior knowledge. These features can be used for feature-based classification and tasks.

4.3. The uncertainties in deprivation mapping

Uncertainties arise in input data, model training, and prediction in terms of extensional and existential uncertainties. The uncertainty

produced at one step also propagates and is mixed with successive uncertainties. For instance, the uncertainty in reference data brings further uncertainties in learning deprived area morphology during model training. Data augmentation also produces uncertainties. These uncertainties ultimately accumulated in the final prediction map. Among all types of uncertainties, the boundary uncertainty of reference data is most critical as it flows in as manually digitized boundaries and independent of the scope of model design, which renders the uncertainty intrinsic in the data during model training and testing. As discussed in section 4.1, poor training data leads to poor results, and therefore also confuse machine learning techniques. Although one may expect an improvement of model performance by either working on the input data such as augmentation or better designing the model to utilize low-level features, the improvement is expected to hit its limit due to the quality of input information. There are also different views on the improvement of boundary uncertainty. While some study suggests that uncertainties in reference data can be reduced by including additional local knowledge (Kohli et al., 2016b), another treated uncertainties in boundary definition as a reflection of multi-dimensionally deprived areas apart from the morphological definition (Pratomo et al., 2017). If the boundary uncertainty is a manifestation of multi-dimensionality in defining deprivation, then one needs to recognize the trade-off between intrinsic uncertainties in the input data and machine learning-based model performance. In this sense, the output of a DCNN can be weighted by its significance in capturing existential and extensional information of deprived areas.

In dealing with the model output, the uncertainties can, if intrinsic and not controllable, be encoded so that potential end-users of the output can be informed about the reliability of the outputs. One potential option is to encode uncertainties as probabilities by providing the prediction probability map as a heat map instead of rigid "deprivation and non-deprivation" binary map. Correspondingly, accuracy assessment of the output should be adapted, which relates to the purpose of showing either the location or the extension of deprivation. Although the JI accuracy is a rigid metric that evaluates exact deprived area boundaries in the output map, the feasibility of proposing less rigid metrics is worth further discussion.

Besides mapping urban deprivation, using the mapped information triggers further uncertainties. The level of aggregation needs to be defined to provide collective distribution patterns of slums at neighborhood, city or regional level.

4.4. Scaling and transferring deep learning based deprivation mapping

All the issues of input data, learnt features and uncertainties discussed above can magnify when the deep learning based mapping of deprivation scales from local or city level up to regional or continental level, as well as been transferred to other geographic regions. At the local level within a same city, data availability as well as relatively similar deprivation morphology render the efficiency of deep learning based deprivation mapping. However, the varying and complex deprivation morphology across cities, countries and continents (Kuffer et al., 2017; Taubenböck et al., 2018a) is largely ignored at local level analysis as machine learning techniques have only been applied to very homogeneous small areas, leaving the performance assessment of deep learning technique biased. When mapping at larger scales, data availability becomes the primary concern as the requirement of input data specifications varies across cities and regions. For instance, many local governments have limited access to very high spatial resolution imagery, nevertheless, high spatial resolution does not necessarily guarantee optimal mapping results in all situations (Wang et al., 2019). Furthermore, the transferability of deep learning based deprivation mapping needs to be considered (Duque et al., 2017). Given the fact that mapping results are sensitive to input data and model architecture as shown in this study, at least two levels of transferability need to be addressed in large scale deprivation mapping: (1) the transferability of features learnt from one city or region to others, and (2) the transferability, if not found in the features, but in the model architecture applied in one city or region to others. The transferability of either the learnt features or the model architecture determines the computation resources as whether to use pre-trained model, or train a existing model, or fine-tune the existing model architecture before training, or even design a model from scratch (Wurm et al., 2019a). Apart from computation, less transferability also means extra workload to treat each city or region as a special case, and more labeling activities and uncertainties are introduced.

5. Conclusions

This study is a first attempt to use DCNN to map very small deprivation pockets in a larger area while considering several practical issues such as limited data accessibility, unreliable/generalized ground truth data and insufficient computational resources. Although DCNN is capable of capturing deprivation morphology by using limited training samples, the city level mapping result is significantly worse than the one obtained at the local level. The DCNN is sensitive to not only the amount of training samples provided but also the morphological information contained in the training sample. Thus, providing training data with rich and well-generalized information is important for DCNN to learn and capture precise spatial specifications of deprivation. The situation of limited data can be complemented by data augmentation to provide more variations in the training data. However, the improvements brought by the augmentation may not be significant.

Apart from the limitation of training data, the prediction accuracy largely depends on the boundary prediction of deprivation pockets. Inaccurate segmentation is manifested in poor boundary prediction and can especially impact the accuracy of very small deprivation pocket prediction. The boundary issue can be effectively resolved by optimizing the model architecture to utilize low-level features in recovering object boundaries. The proposed U-Net-CPD explicitly concatenates low-level features to the last block of the model leading to improved boundary predictions. The improvement is preeminent in the prediction of small pockets. A slight boundary shift may significantly impact the predicted extension of small pockets. Leveraging the power of learnt features in other feature-based classifications worth further research as understanding and interpreting the learnt features are non-trivial tasks. Evaluating the accuracy of prediction is difficult when only unreliable reference data is available. For instance, predicted areas that are morphologically similar to deprivation, but omitted in the reference data, require further effort in accuracy assessment.

From a pragmatic point of view, deprivation mapping highlights the potential of efficiently monitoring the multi-dimensionality of an urban phenomenon by using ground validated remote-sensed based information. The mapping allows supporting planning and policy development as well as monitoring the implementation of policies in large and fast-growing cities in the global south, where information on deprivation locations and dynamics is often scarce. Furthermore, such information could help in the calibration and validation of micro-simulation computer models.

Acknowledgements

The authors would like to acknowledge the support of the SimCity project (contract number: C.2324.0293) and Dynaslum (Data Driven Modelling and Decision Support for Slums) project (contract number: 27015G05), which are managed by the Dutch national research council (NWO) and the Dutch organization for ICT in education and research (SURF) to provide resources for this research. We acknowledge the European Space Agency (ESA) and DigitalGlobe Foundation for providing the image data through its Third Party Missions.

References

Albert, A., Kaur, J., Gonzalez, M.C., 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1357–1366.

Arribas-Bel, D., Patino, J.E., Duque, J.C., 2017. Remote sensing-based measurement of Living Environment Deprivation: improving classical approaches with machine learning. PLoS One 12, e0176684.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. 1511.00561.

Baud, I., Kuffer, M., Pfeffer, K., Sliuzas, R., Karuppannan, S., 2010. Understanding heterogeneity in metropolitan India: the added value of remote sensing data for analyzing sub-standard residential areas. Int. J. Appl. Earth Obs. Geoinf. 12, 359–374.

Benediktsson, J.A., Pesaresi, M., Amason, K., 2003. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Trans. Geosci. Remote Sens. 41, 1940–1949.

Chandramouli, C., General, R., 2011. Census of India 2011. Provisional Population Totals. *Government of India*, New Delhi.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Cui, X., Goel, V., Kingsbury, B., 2015. Data augmentation for deep neural network acoustic modeling. IEEE Trans. Audio Speech Lang. Process. 23, 1469–1477.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. Deepglobe 2018: A Challenge to Parse the Earth through Satellite Images. *(ArXiv e-prints)*.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Annual Conference on Medical Image Understanding and Analysis. Springer, pp. 506–517.

Dundar, M., Kou, Q., Zhang, B., He, Y., Rajwa, B., 2015. Simplicity of kmeans versus deepness of deep learning: a case of unsupervised feature learning with limited data. In: Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, pp. 883–888.

Duque, J.C., Patino, J.E., Betancourt, A., 2017. Exploring the potential of machine learning for automatic slum identification from VHR imagery. Remote Sens. *9*.

Ezeh, A., Oyebode, O., Satterthwaite, D., Chen, Y.-F., Ndugwa, R., Sartori, J., Mberu, B., Melendez-Torres, G., Haregu, T., Watson, S.I., 2017. The history, geography, and sociology of slums and the health problems of people who live in slums. Lancet 389, 547–558.

Friesen, J., Taubenböck, H., Wurm, M., Pelz, P.F., 2018. The similar size of slums. Habitat Int. 73, 79–88.

Graesser, J., Cheriyadat, A., Vatsavai, R.R., Chandola, V., Long, J., Bright, E., 2012. Image based characterization of formal and informal neighborhoods in an urban landscape. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5, 1164–1176.

Guo, Z., Shao, X., Xu, Y., Miyazaki, H., Ohira, W., Shibasaki, R.J.R.S., 2016. Identification of Village Building via Google Earth Images and Supervised Machine Learning Methods 8. pp. 271.

Habitat, U., 2003. Slums of the World: the Face of Urban Poverty in the New Millennium. *United Nations Human Settlements Programme*, Nairobi.

Hachmann, S., Arsanjani, J.J., Vaz, E., 2018. Spatial data for slum upgrading: volunteered Geographic Information and the role of citizen science. Habitat Int. 72, 18–26.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hernandez-Stefanoni, J.L., Ponce-Hernandez, R., 2004. Mapping the spatial distribution of plant diversity indices in a tropical forest using multi-spectral satellite image classification and field measurements. Biodivers. Conserv. 13, 2599–2621.

Hofmann, P., Taubenböck, H., Werthmann, C., 2015. Monitoring and modelling of informal settlements-A review on recent developments and challenges. In: 2015 Joint Urban Remote Sensing Event (JURSE). IEEE, pp. 1–4.

Hoo-Chang, S., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35, 1285.

Huang, X., Zhang, L., 2013. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 51, 257–272.

Ibrahim, M.R., Haworth, J., Cheng, T., 2018a. URBAN-i: from Urban Scenes to Mapping Slums, Transport Modes, and Pedestrians in Cities Using Deep Learning and Computer Vision. 1809.03609.

Ibrahim, M.R., Titheridge, H., Cheng, T., Haworth, J., 2018b. predictSLUMS: A New Model for Identifying and Predicting Informal Settlements and Slums in Cities from Street Intersections Using Machine Learning. 1808.06470.

Iglovikov, V., Mushinskiy, S., Osin, V., 2017. Satellite Imagery Feature Detection Using Deep Convolutional Neural Network: A Kaggle Competition. 1706.06169.

Iglovikov, V., Shvets, A., 2018. TernausNet: U-Net with VGG11 Encoder Pre-trained on ImageNet for Image Segmentation. 1801.05746.

India, G.o, 2011. Slums in India: a statistical compendium. In: Government of India, Ministry of Housing and Poverty Alleviation New Delhi.

India, G.o, 2015. Slums in India: a statistical compendium. In: Government of India, Ministry of Housing and Poverty Alleviation New Delhi.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. New Phytol. 11, 37–50.

Jain, S., 2008. Remote sensing application for property tax evaluation. Int. J. Appl. Earth Obs. Geoinf. 10, 109–121.

Jenerette, G.D., Harlan, S.L., Buyantuev, A., Stefanov, W.L., Declet-Barreto, J., Ruddell, B.L., Myint, S.W., Kaplan, S., Li, X., 2016. Micro-scale urban surface temperatures are related to land-cover features and residential heat related health impacts in Phoenix, AZ USA. Landsc. Ecol. 31, 745–760.

Kalma, J.D., McVicar, T.R., McCabe, M.F., 2008. Estimating land surface evaporation: a review of methods using remotely sensed surface temperature data. Surv. Geophys. 29, 421–469.

Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., Cun, Y.L., 2010. Learning convolutional feature hierarchies for visual recognition. In: Advances in Neural Information Processing Systems, pp. 1090–1098.

Kit, O., Lüdeke, M., 2013. Automated detection of slum area change in Hyderabad, India using multitemporal satellite imagery. ISPRS J. Photogrammetry Remote Sens. 83, 130–137.

Klaufus, C., 2010. Watching the city grow: remittances and sprawl in intermediate Central American cities. Environ. Urbanization 22, 125–137.

Kohli, D., Sliuzas, R., Kerle, N., Stein, A., 2012. An ontology of slums for image-based classification. Comput. Environ. Urban Syst. 36, 154–163.

Kohli, D., Sliuzas, R., Stein, A.J.J.S., 2016a. Urban Slum Detection Using Texture and Spatial Metrics Derived from Satellite Imagery. J. Spat. Sci. 61, 405–426.

Kohli, D., Stein, A., Sliuzas, R., 2016b. Uncertainty analysis for image interpretations of urban slums. Comput. Environ. Urban Syst. 60, 37–49.

Krishna, A., Sriram, M., Prakash, P., 2014. Slum types and adaptation strategies: identifying policy-relevant differences in Bangalore. Environ. Urbanization 26, 568–585.

Kuffer, M., Pfeffer, K., Sliuzas, R., 2016a. Slums from space—15 years of slum mapping using remote sensing. Remote Sens. 8, 455.

Kuffer, M., Pfeffer, K., Sliuzas, R., Baud, I., 2016b. Extraction of slum areas from VHR imagery using GLCM variance. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 1830–1840.

Kuffer, M., Pfeffer, K., Sliuzas, R., Baud, I., van Maarseveen, M., 2017. Capturing the diversity of deprived areas with image-based features: the case of Mumbai. Remote Sens. *9*.

Kuffer, M., Wang, J., Nagenborg, M., Pfeffer, K., Kohli, D., Sliuzas, R., Persello, C., 2018. The scope of earth-observation to improve the consistency of the SDG slum indicator. ISPRS Int. J. Geo-Inf. 7, 428.

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

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436.

LeCun, Y., Kavukcuoglu, K., Farabet, C., 2010. Convolutional networks and applications in vision. In: ISCAS, pp. 253–256.

Lee, G., Tai, Y.-W., Kim, J., 2016. Deep saliency with encoded low level distance map and high level features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 660–668.

Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., Li, W., 2018. Deepunet: a deep fully convolutional network for pixel-level sea-land segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.

Li, Y., Huang, X., Liu, H., 2017. Unsupervised deep feature learning for urban village detection from high-resolution remote sensing images. Photogramm. Eng. Remote Sens. 83, 567–579.

Liu, H., Huang, X., Wen, D., Li, J., 2017. The use of landscape metrics and transfer learning to explore urban villages in China. Remote Sens. 9, 365.

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

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657.

Mahabir, R., Croitoru, A., Crooks, A., Agouris, P., Stefanidis, A., 2018. A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums. Urban Science 2, 8.

Mahabir, R., Crooks, A., Croitoru, A., Agouris, P.J.R.S., 2016. Regional Science. In: The Study of Slums as Social and Physical Constructs: Challenges and Emerging Research Opportunities 3. pp. 399–419.

Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. IEEE, pp. 416–423.

Martinez, J., Mboup, G., Sliuzas, R., Stein, A., 2008. Trends in urban and slum indicators across developing world cities, 1990–2003. Habitat Int. 32, 86–108.

Mboga, N., Persello, C., Bergado, J., Stein, A., 2017. Detection of informal settlements from VHR images using convolutional neural networks. Remote Sens. 9, 1106.

Muller, C.L., Chapman, L., Grimmond, C., Young, D.T., Cai, X., 2013. Sensors and the city: a review of urban meteorological networks. Int. J. Climatol. 33, 1585–1600.

Oke, T.R., 2002. Boundary Layer Climates. Routledge.

Pal, N.R., Pal, S.K., 1993. A review on image segmentation techniques. Pattern Recognit. 26, 1277–1294.

Persello, C., Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in VHR images. IEEE Geosci. Remote Sens. Lett. 14, 2325–2329.

Pesaresi, M., 2000. Texture analysis for urban pattern recognition using fine-resolution panchromatic satellite imagery. Geogr. Environ. Model. 4, 43–63.

Pesaresi, M., Gerhardinger, A., Kayitakire, F., 2008. A robust built-up area presence index by anisotropic rotation-invariant textural measure. IEEE J. Sel. Top. Appl. Earth Obs.

Remote Sens. 1, 180–192.
Pratomo, J., Kuffer, M., Martinez, J., Kohli, D., 2017. Coupling uncertainties with accuracy assessment in object-based slum detections, case study: jakarta, Indonesia. Remote Sens. 9, 1164.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Roy, D., Lees, M.H., Palavalli, B., Pfeffer, K., Sloot, M.A., 2014. The emergence of slums: a contemporary view on simulation models. Environ. Model. Softw 59, 76–90.
Roy, D., Lees, M.H., Pfeffer, K., Sloot, M.A., 2017. Modelling the impact of household life cycle on slums in Bangalore. Comput. Environ. Urban Syst. 64, 275–287.
Roy, D., Lees, M.H., Pfeffer, K., Sloot, M.A., 2018. Spatial segregation, inequality, and opportunity bias in the slums of Bengaluru. Cities 74, 269–276.
Seferbekov, S., Iglovikov, V., Buslaev, A., Shvets, A., 2018. Feature pyramid network for multi-class land segmentation. In: The IEEE Conference On Computer Vision And Pattern Recognition (CVPR) Workshops.
Singh, P.P., Garg, R., 2013. Automatic road extraction from high resolution satellite image using adaptive global thresholding and morphological operations. J. Indian Soc. Remote Sens. 41, 631–640.
Roy, D., Palavalli, B., Menon, N., King, R., Pfeffer, K., Lees, M., Sloot, M.A., 2018. Survey-based socio-economic data from slums in Bangalore, India. Sci. Data 5, 170200.
Saharan, T., B, I., Pfeffer, Karin, 2018. 'Slum' and the City: exploring relations of informal settlements comparatively in Chennai, India and Durban, South Africa. In: Faculty of Social and Behavioural Sciences (FMG). University of Amsterdam, pp. 182.
Taubenböck, H., Kraff, N., 2014. The physical face of slums: a structural comparison of slums in Mumbai, India, based on remotely sensed data. J. Hous. Built Environ. 29, 15–38.
Taubenböck, H., Kraff, N., Wurm, M., 2018a. The morphology of the Arrival City-A global categorization based on literature surveys and remotely sensed data. Appl. Geogr. 92, 150–167.
Taubenböck, H., Kraff, N.J., Wurm, M., 2018b. The Blind Spot–Reducing Knowledge Gaps in Urban Poverty with Earth Observation Interacting with Structured and Unstructured Geodata.
Taubenböck, H., Staab, J., Zhu, X., Geiß, C., Dech, S., Wurm, M., 2018c. Are the poor digitally left behind? ndications Urban Divides Based Remote Sensing Twitter Data. ISPRS International Journal of Geo-Information 7, 304.
Taubenböck, H., Wegmann, M., Roth, A., Mehl, H., Dech, S., 2009. Urbanization in India–Spatiotemporal analysis using remote sensing data. Comput. Environ. Urban Syst. 33, 179–188.
Taubenböck, H., Wurm, M., 2015. Ich weiß, dass ich nichts weiß–Bevölkerungsschätzung in der Megacity Mumbai. Globale Urbanisierung. Springer, pp. 171–178.
Turok, I., Borel-Saladin, J., 2018. The theory and reality of urban slums: pathways-out-of-poverty or cul-de-sacs? Urban Stud. 55, 767–789.
UN, 2018. 2018 revision of world urbanization prospects. In: United Nations Department of Economic and Social Affairs.
Veljanovski, T., Kanjir, U., Pehani, P., Oštir, K., Kovačič, P., 2012. Object-based Image Analysis of VHR Satellite Imagery for Population Estimation in Informal Settlement Kibera-Nairobi, Kenya. Remote Sensing-Applications. InTech.
Wang, H., Chen, S., Xu, F., Jin, Y.-Q., 2015. Application of deep-learning algorithms to MSTAR data. In: Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International. IEEE, pp. 3743–3745.
Wang, J., Kuffer, M., Pfeffer, K., 2019. The role of spatial heterogeneity in detecting urban slums. Comput. Environ. Urban Syst. 73, 95–107.
Wurm, M., Stark, T., Zhu, X.X., Weigand, M., Taubenböck, H.J.I.J.H., Sensing, R., 2019. Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks. ISPRS Journal of Photogrammetry and Remote Sensing 150, 59–69.
Wurm, M., Taubenböck, H., Weigand, M., Schmitt, A., 2017. Slum mapping in polarimetric SAR data using spatial features. Remote Sens. Environ. 194, 190–204.
Yu, F., Koltun, V., 2015. Multi-scale Context Aggregation by Dilated Convolutions. 1511.07122.