ISPRS Journal of Photogrammetry and Remote Sensing 150 (2019) 59–69

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks

Michael Wurm[a,*], Thomas Stark[b], Xiao Xiang Zhu[b,c], Matthias Weigand[a,d], Hannes Taubenböck[a]

[a] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen 82234, Germany
[b] Technical University of Munich (TUM), Signal Processing in Earth Observation (SiPEO), 80333 Munich, Germany
[c] German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), Oberpfaffenhofen 82234, Germany
[d] University of Würzburg, Department for Remote Sensing, 97074 Würzburg, Germany

ARTICLE INFO

Keywords:
Slums
FCN
Convolutional neural networks
Deep learning
Transfer learning

ABSTRACT

Unprecedented urbanization in particular in countries of the global south result in informal urban development processes, especially in mega cities. With an estimated 1 billion slum dwellers globally, the United Nations have made the fight against poverty the number one sustainable development goal. To provide better infrastructure and thus a better life to slum dwellers, detailed information on the spatial location and size of slums is of crucial importance. In the past, remote sensing has proven to be an extremely valuable and effective tool for mapping slums. The nature of used mapping approaches by machine learning, however, made it necessary to invest a lot of effort in training the models. Recent advances in deep learning allow for transferring trained fully convolutional networks (FCN) from one data set to another. Thus, in our study we aim at analyzing transfer learning capabilities of FCNs to slum mapping in various satellite images. A model trained on very high resolution optical satellite imagery from QuickBird is transferred to Sentinel-2 and TerraSAR-X data. While free-of-charge Sentinel-2 data is widely available, its comparably lower resolution makes slum mapping a challenging task. TerraSAR-X data on the other hand, has a higher resolution and is considered a powerful data source for intra-urban structure analysis. Due to the different image characteristics of SAR compared to optical data, however, transferring the model could not improve the performance of semantic segmentation but we observe very high accuracies for mapped slums in the optical data: QuickBird image obtains 86–88% (positive prediction value and sensitivity) and a significant increase for Sentinel-2 applying transfer learning can be observed (from 38 to 55% and from 79 to 85% for PPV and sensitivity, respectively). Using transfer learning proofs extremely valuable in retrieving information on small-scaled urban structures such as slum patches even in satellite images of decametric resolution.

1. Introduction

Poverty is considered one of the major challenges for our society in the upcoming decades, making it the number one issue of the Sustainable Development Goals as defined by the United Nations (UN, 2017). In urban areas, slums are the most visible, distinct manifestation of poverty (Amnesty International, 2016). Unprecedented processes of urbanization over the past decades have transformed mankind into an urban species with two thirds of the global population being expected to live in urban areas by the year 2050 (UN, 2015). This rural-urban migration is especially intense in mega cities of the global south, such as Mumbai in India which grew at a pace of up to 300,000 inhabitants per year (Burdett and Rhode, 2010). Since formal urban development cannot keep up with this pace of rural-urban migrants, many new urban dwellers are forced to find their new homes in settlements of informal nature with poor living conditions, lack of basic services such as access to safe water and sanitation facilities. Today, these *slums* are home to almost an estimated billion dwellers on a global scale (UN Habitat, 2015). In some cities, the share of slum dwellers accounts for up to 42% of the city's total population in official numbers (and a significantly higher number in estimations) such as it is the case for Mumbai (Taubenböck and Wurm, 2015). Various strategies for dealing with slums have been developed by local authorities, however a recent change can be observed towards a strategy of integrating the 'invisible city' into governing structures is today for many cities the accepted way to deal with those informal areas since the presence of slums cannot be neglected anymore (Wurm and Taubenböck, 2018). Thus, the derivation of reliable, spatial information on the size and location of slum

* Corresponding author.
E-mail address: michael.wurm@dlr.de (M. Wurm).

https://doi.org/10.1016/j.isprsjprs.2019.02.006
Received 25 May 2018; Received in revised form 10 December 2018; Accepted 11 February 2019
Available online 21 February 2019

areas by mapping approaches has gained much of interest over the past.

1.1. Morphological characteristics of slums from a remote sensing perspective

As it can be observed for many applications in the context of urban remote sensing, the turn of the millennium marks an important date with the advent of very high resolution satellites providing images at resolutions of 1 m or better. Especially for the discrimination of very small, heterogeneous objects such as buildings within the urban environment, high image resolutions are of crucial importance. Thus, also in the context of slum mapping, an increased interest in the utilization of VHR satellite images can be observed since then. This goes in parallel with the advent of more sophisticated image analysis techniques such as object-based image analysis or, recently, deep learning methods. Thus, in the following we review previous works on remote sensing-based slum mapping based on different methods and image features in the light of the complex nature of slum morphology.

From a synoptic perspective, urban poverty finds its physical expression in many different ways which usually do not follow a strict and universal concept (Taubenböck et al., 2018; Kuffer et al., 2017). However, some forms of urban poverty in particular can be directly associated with the morphology of the built environment, though (Sandborn and Engstrom, 2016; Jean et al., 2016; Wurm and Taubenböck, 2018). Most commonly, organic, irregular arrangements of buildings are associated with slum areas, as well as low building heights, poor construction materials and a generally high building density in often hazardously exposed areas (Baud et al., 2010; Kuffer et al., 2016a; Graesser et al., 2012; Jain, 2007). These characteristic morphologic features are extensively exploited in remote sensing-based image analysis for slum mapping. Since recently thorough studies on the state of slum mapping have been released (Kuffer et al., 2016a; Mahabir et al., 2018), we only summarize below past research efforts based on significant cornerstones in methods or data. While generally, very high mapping accuracies are achieved by visual image interpretation (Wurm and Taubenböck, 2018; Taubenböck et al., 2018) or knowledge-based methods using object-based image analysis (OBIA) relying on tuned parameters (Kuffer et al., 2014; Baud et al., 2010), large-area mapping of slums is usually based on machine learning algorithms which aim at generalizing specific semantic knowledge in the images based on labeled elements and image descriptors to provide transferability of the learned knowledge into unknown areas. One key feature in the identification of slums is their sharp contrast in their physical appearing compared to formal developed urban areas. Therefore, contextual image features such as the grey-level-co-occurrence-matrix (GLCM) was used extensively in slum mapping in combination with machine learning techniques such as random forests (e.g. Kuffer et al., 2016b; Graesser et al., 2012; Wurm et al., 2017; Owen and Wong, 2013) or support vector machines (Huang et al., 2015). Besides the extensive use of VHR optical data, only few studies were dedicated to the exploitation of actively acquired data, e.g. such as dual-polarized X-band SAR data from TerraSAR-X (Wurm et al., 2017; Schmitt et al., 2018). Only recently, the current trend in machine learning for semantic segmentation of images has been taken up by the application of deep learning for the detection of slums in VHR images confirming current trends in deep learning methods to outperform state-of-the-art machine learning techniques (Persello and Stein, 2017). The next subsequent step to learning and applying a network on the same data set is to transfer a pretrained network to sensors of different resolutions. Thus, deeper networks consisting of more hidden layers need to be considered (Oquab et al., 2014).

1.2. Transfer learning for semantic segmentation using convolutional neural networks

Generally, most machine learning methods work well because human-designed representations and features are used to optimize weights for an accurate prediction. Representation learning attempts to automatically learn good features or representations, which works well for small problems. In contrast, manually designed features are often over-specified, incomplete, and are very time-consuming for design and validation. Deep learning algorithms attempt to automatically learn multiple levels of representations exclusively from its input data, without the need of additional user input (Zhu et al., 2017). Besides its effectiveness, this can be regarded as one of the reasons for the big success of deep learning in machine learning since the task of training and prediction is facilitated. Recent advances in the field have proven deep learning a very successful set of tools, sometimes even able to surpass human ability to solve highly computational tasks (Zhu et al., 2017). Especially for image representations, convolutional neural networks have proven to excel at extracting mid- and high level abstract features from raw images. Recent studies indicate that the feature representations learned by CNNs are greatly effective in large scale image recognition (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), object detection (Girshick et al., 2016) and semantic segmentation (Long et al., 2015).

Image segmentation aims at understanding an image at pixel level, i.e. each pixel of an image is assigned a semantic class. Initially, images of a fixed size were required for classification, but soon fully convolutional networks (FCNs) without fully connected layers popularized CNN architectures for dense predictions of images of any size and significantly increased speed (Long et al., 2015). Apart from fully connected layers, one of the main challenges using CNNs for semantic segmentation are the 'pooling layers'. They increase the field of view and are able to aggregate the context while discarding the location information. However, semantic segmentation requires the exact alignment of class maps and thus, needs the spatial information to be preserved. This issue can be tackled by encoder-decoder architectures where an encoder gradually reduces the spatial dimension with pooling layers and a decoder which gradually recovers the object details and spatial dimension using transposed/fractionally strided convolutions. While FCNs can learn the interpolation during the decoding process, upsampling produces coarse segmentation maps because of loss of information during pooling. Therefore, skip connections are introduced from higher resolution feature maps.

In Long et al. (2015), the authors describe the key observation that fully connected layers in classification networks can be viewed as convolutions with kernels that cover their entire input regions. This is equivalent to evaluating the original classification network on overlapping input patches but is much more efficient because computation is shared over the overlapping regions of patches. In remote sensing, the use of deep learning brings up new challenges, since satellite image analysis raises some unique issues that need to be considered, e.g. geolocation of satellite images, sensor specifics (resolution, incidence angles, data quality etc.) or the big data challenge (Zhu et al., 2017).

In the context of remote sensing, scene classification of satellite images, which aims to automatically assign a semantic label to each pixel in an image, has recently been an active research topic in the field of VHR satellite images. Generally, scene classification can be divided into two steps: *feature extraction* and *classification*. With growing numbers of images, training a complicated non-linear classifier is very time consuming. Hence, to extract a holistic and discriminative feature representation is the most significant part for scene classification. Traditional approaches are mostly based on the Bag-of-Visual-Words model (Sivic and Zisserman, 2003; Zhu et al., 2016), but their potential for improvement was limited by the ability of experts to design the feature extractor and the expressive power encoded. In contrast, deep learning architectures have been successfully applied to the problem of scene classification of high-resolution satellite images outperforming state-of-the-art image classifiers (Zou et al., 2015; Penatti et al., 2015; Castelluccio et al., 2015; Mou et al., 2017).

As deep learning is a multi-layer feature learning architecture, it can

learn more abstract and discriminative semantic features with growing depth. Thus, it has been shown that it can achieve far better classification performances compared to mid-level approaches (Zhu et al., 2017). Training of neural networks, is usually performed using pretrained networks on large image datasets, e.g., COCO (Lin et al., 2014), Pascal VOC (Everingham et al., 2010) or ImageNet (Deng et al., 2009) which in general reach impressive accuracies (Hu et al., 2015; Zou et al., 2015). Expanding the three channel input limitations of traditional deep learning algorithms (Kemker et al., 2018; Marmanis et al., 2018) use specific architectures to use elevation information and multispectral imagery to boost performance in semantic segmentation frameworks. Training networks from scratch is an extremely elaborate and time consuming method which is usually employed only if the data has completely different characteristics compared to internet images, for example hyperspectral images (Mou et al., 2018; Pan et al., 2018) or SAR (Gong et al., 2017; Hughes et al., 2018).

In general, transfer learning builds upon learned knowledge from one dataset to improve learning in another dataset. More specifically, it can be described as a method which aims to improve learning the target predictive function $f_T(\cdot)$ in the new target dataset D_T using the knowledge learned in the source dataset D_S. As described by Pan and Yang (2010), transfer learning can be divided into three categories: (a) In inductive transfer learning the target task is different from the source task, no matter if the source or target datasets are the same or not. In this case labeled data is required to induce the target learning task. (b) For transductive transfer learning both target and source learning tasks are the same while their datasets are different. In this situation no labeled data in the target dataset are required. Lastly, (c) unsupervised transfer learning is used when the target and source tasks are different, and no labeled data is available in both source and target datasets.

Selection of transfer learning strategies not only depends on the availability of existing labels in both source and target datasets and the similarity of the source and target dataset but also if weights learned in the source task can be adjusted or shared in the target task. Transfer learning can be achieved using multiple strategies. Multi-task learning has been used to improve object detection accuracy by transferring knowledge from one object class to another using a support vector machine's (SVM) discriminative training framework for HOG template models (Aytar and Zisserman, 2011) or using a hierarchical classification model that allows rare objects to borrow statistical strength from related objects (Salakhutdinov et al., 2011). Two multi-task classifiers are used to obtain a more robust classifier for object detection in videos (Ma et al., 2014). In hyperspectral remote sensing domain adaption technology can be applied to share knowledge between different geographical domains when using support vector machines (Sun et al., 2012) or random forest classifiers with transfer component analysis (Xia et al., 2017). Impressive results could be observed using unsupervised feature representation using pretrained CNNs for scene classification in very high resolution remote sensing imagery (e.g. Castelluccio et al., 2015; Hu et al., 2015). Inductive transfer learning enables to further improve the learning task where backpropagation successfully re-weights labeled data from natural image datasets, e.g. ImageNet to solve new problems in remote sensing datasets (e.g. Maggiori et al., 2017; Marmanis et al., 2016; Nogueira et al., 2017; Kang et al., 2018). Therefore, in this study inductive transfer learning of a FCN is used due to relative large labeled datasets where the fine tuning of weights during backpropagation aims to achieve best possible results.

1.3. Transferring deep features between various remote sensing data sets

In slum mapping, in particular approaches using remotely sensed data from satellite images with varying characteristics were used extensively for assessing image processing and analysis techniques (Kuffer et al., 2016a; Mahabir et al., 2018). Both scientific meta-studies state that while previous work on remote sensing-based slum mapping has acknowledged the advances of recent machine learning techniques for locating slums in satellite images, they lack transferability between various data sets. Costs for the large-area availability of very high resolution (VHR) optical satellite imagery at a geometric resolution of 1 m and below are a limiting factor and thus, multi-sensor approaches with data sets of varying origins are proposed.

In this study we want to address these identified issues by using state-of-the-art machine learning techniques from the family of convolutional neural networks (CNN) which need no tuning of parameters and have therefore better capabilities for transferring a trained network to another data set, as long as the training data set is sufficiently large and representative. Specifically, we want to explore the capabilities of this process of '*transfer learning*' to adopt a pretrained CNN from VHR optical Quickbird imagery to be applied to satellites with larger mapping areas but lower geometric resolution such as Sentinel-2. Further, in a second experiment we want to assess the capabilities of transfer learning from optical imagery to active SAR imagery from TerraSAR-X.

The remainder of this article is structured as follows: in the following Section 2 we present the methodological framework of fully convolutional networks (FCN), transfer learning for slum mapping and used data sets among the experimental set-up. In Section 3 we present the results and discussion of the performed experiments, while Section 4 concludes the paper.

2. Methods and experimental set-up

2.1. Method: The fully convolutional network FCN-VGG19

FCNs, first introduced by Long et al. (2015) allow for semantic segmentation to train end-to-end and pixel-to-pixel for the prediction of dense outputs from arbitrary sized input images. Learning and inference are performed on the entire image by dense feedforward computation and backpropagation. Within the network upsampling layers enable a pixelwise prediction and learning with subsampled pooling. For our experiments, we use the CNN based on the classification architecture VGG19 by the Visual Geometry Group of Oxford University (Simonyan and Zisserman, 2014). The CNN relies on rather small receptive fields of 3×3 pixels which are convolved with the input at every pixel. In this way a stack of two 3×3 convolutional layers has an effective receptive field of 5×5. Consequently, four layers have a 9×9 effective receptive field. This strategy has the advantage of incorporating four nonlinear rectification layers instead of a single one, making the decision function more discriminative. Furthermore, it decreases the number of parameters: $4(3^2C^2) = 36C^2$ produces less trainable weights than a single 9×9 convolutional layer: $9^2C^2 = 81C^2$.

To adapt the CNN-VGG19 architecture to an FCN some modifications are required: The final classification layer is discarded and replaced with a 1×1 convolution and with the channel dimension of the number of used classes. Further, deconvolutional layers are introduced for bilinear upsampling of the coarse outputs to pixel-dense outputs. In this case, upsampling through deconvolutional layers means using transpose convolutions. This operation simply reverses the forward and backward passes of the convolution. Upsampling is performed for end-to-end learning by backpropagation from a pixelwise loss (Long et al., 2015).

A graphical representation of the used FCN-VGG19 architecture is depicted in Fig. 1. It shows that the FCN uses skips, which combines the final prediction layer with lower level layers with finer strides. Fusing fine layers and coarse layers lets the model make local predictions that respect a global structure. The FCN fuses the upsampled output of the VGG19 network architecture with predictions computed on top of the third and fourth pooling layer.

2.2. Method: Transfer learning approach

Training the FCN was performed using an inductive transfer

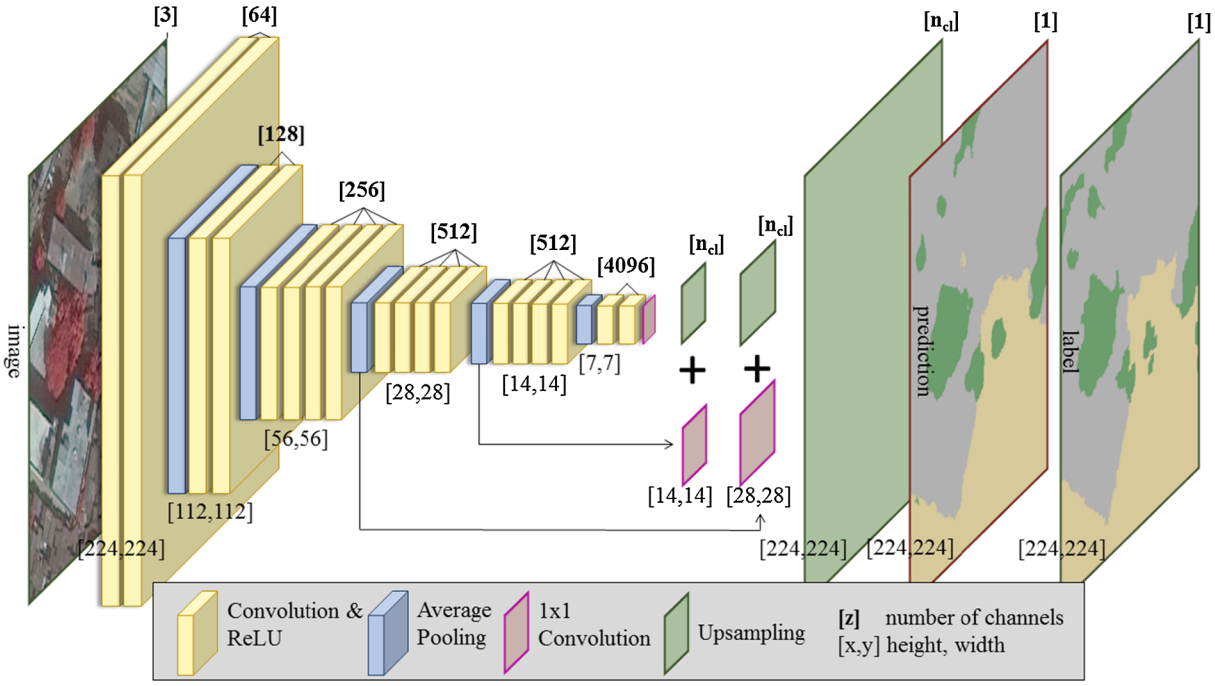

Fig. 1. Architecture of the FCN-VGG19 adapted from Long et al. (2015) which learns to combine high level information with fine, low level information using skips from the third and fourth pooling layer. Hidden layers are equipped with rectified linear units (ReLUs) and the number of channels for the convolutional layers increases with the depth of the network. During training the input image is a fixed size of 224 × 224 pixels, while receptive fields for all filters are 3 × 3 pixels throughout the whole network. This configuration allows the FCN to learn approximately 140 million parameters. Prediction is performed using upsampling layers with four channels for the all classes [n_{cl}] in the reference data. Upsampling layers are fused with 1 × 1 convolutions of the third and fourth pooling layers with the same channel dimension [x,y,n_{cl}]. The final upsampling layer predicts fine details using fused information from the last convolutional layer, third and fourth pooling layer upsampled at stride 8.

Table 1
Characteristics of satellite images for testing transfer learning techniques for the FCN-VGG19.

	GSD	Scene size	Bands/Polarization	Date	Incidence Angle	Image tiles
QuickBird	0.5 m	103 km^2	blue, green, red, nir	Nov 17, 2008	16.6°	7487
Sentinel-2	10 m	781 km^2	blue, green, red, nir	Nov 19, 2017	4.8°	219
TerraSAR-X	6 m	242 km^2	HH/VV	Sep 29, 2013	33.7°	2113
	6 m	242 km^2	VV/VH	Dec 11, 2013	33.7°	
	6 m	308 km^2	HH/VV	Oct 10, 2013	34.7°	
	6 m	308 km^2	VV/VH	Dec 04, 2013	34.7°	

learning approach (cf. Section 1.3). When given a source domain dataset D_S and a learning task T_S, a target domain dataset D_T and learning task T_T aims to improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S where $T_S \neq T_T$ (Pan and Yang, 2010). In this case the target domain dataset D_T and the learning task $T_T$benefit from using the knowledge learned in the source domain dataset D_S. We present two groups of experiments. In our first approach weights from a vgg19 CNN which was pretrained on the ImageNet dataset are transfer learned for 100 epochs with all weights available for tuning during the backpropagation algorithm on all three remote sensing datasets where the source domain is the ImageNet dataset $D_S^{ImageNet}$ and the target domain is QuickBird, Sentinel-2 and TerraSAR-X imagery (FCN QB, FCN S2, FCN TX). Instance transfer allows to re-label weights from the source domain to the target domain and ensures adapting the backpropagation algorithm to improve the target learning task. Table 1 indicates a small dataset in the target domain for Sentinel-2 D_T^{S2} with only 219 image tiles and also in the TerraSAR-X target domain D_T^{TX} with only 2113 image tiles. A small target domain in $D_T^{S2,TX}$is usually insufficient for finding good feature representations between the source learning task $T_S^{ImageNet}$ and the target learning task $T_T^{S2,TX}$ for which reason a second group of transfer learning experiments was performed. It aims to reduce differences between the source and target domain where both domains are based on satellite images. Thus, the FCN trained on the QuickBird dataset (FCN QB) from the first group of experiments acts as a new source domain D_S^{QB} for the second group. The target learning task for Sentinel D_T^{S2} benefits from a better feature representation since both datasets D_S^{QB} and D_T^{S2} are optical remote sensing images. In the same way, the experiment is performed for the TerraSAR-X target domain D_T^{TX}. For both transfer learning experiments all trainable variables of the FCN are available during backpropagation to ensure adapting all parameters for the different resolutions and image sensing methods of the remote sensing data.

2.3. Material: Satellite images for slum mapping

For our experiments, space-borne satellite images of three different sensors (QuickBird, Sentinel-2, TerraSAR-X) with entirely different specifications are investigated. Since we aim at testing the capabilities of transfer learning of pretrained models between different images, we briefly introduce the used satellite images for our experiments below (Table 1). In general, our main image data set is from QuickBird. For transfer learning we use Sentinel-2 and TerraSAR-X.

QuickBird: was the first VHR commercial space-borne sensor with a sub-meter resolution of 0.5 m in the panchromatic band. The four

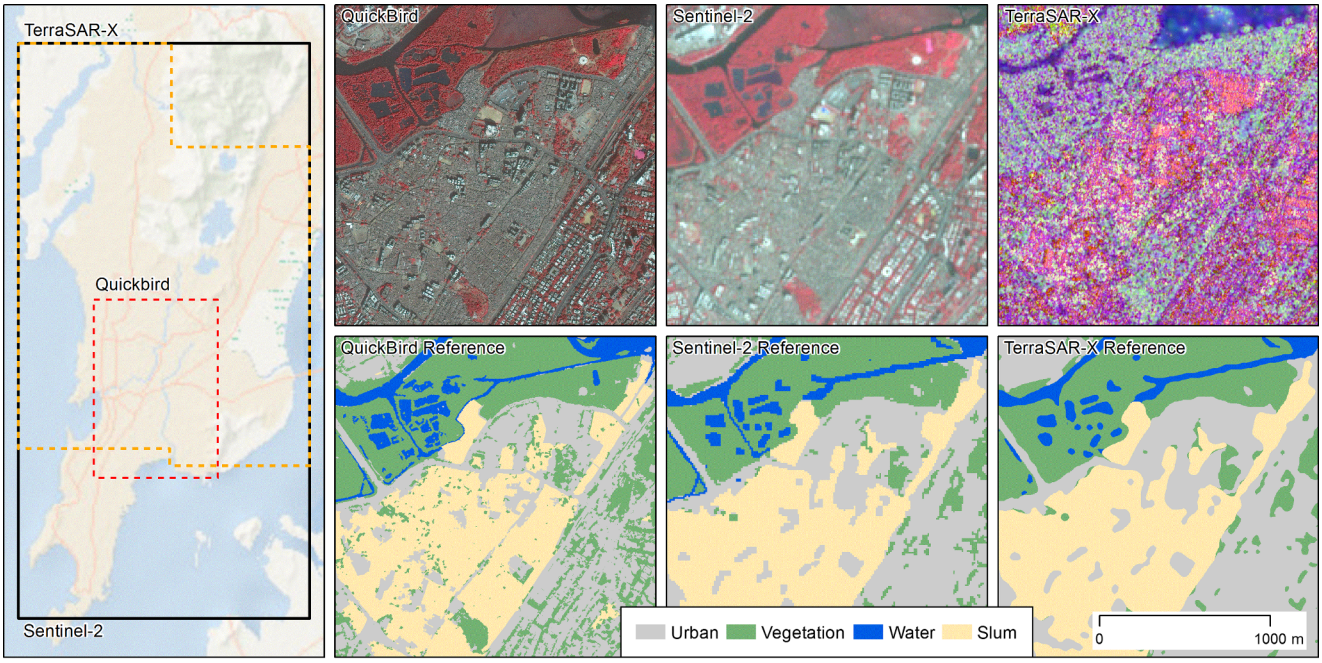

Fig. 2. Composites and reference labels for all datasets: QuickBird and Sentinel-2 in false color and TerraSAR-X as PCA composite for a subset of central Mumbai.

multispectral bands *blue*, *green*, *red* and *nir* are acquired at 2 m resolution. Scenes usually have a swath width of ~17 km.

Sentinel-2: is the high resolution optical sensor of the European Copernicus Programme with 12 spectral and thermal bands at varying resolutions. The *blue*, *green*, *red* and *nir* bands are acquired at 10 m resolution. The swath width is 290 km.

TerraSAR-X: is an active SAR sensing system with various imaging modes of polarizations and resolution. For the commonly used stripmap mode, dual and cross polarized images are acquired at a ground sampling distance (GSD) of 6 m. The swath width is 11 km.

Satellite images are split into image tiles of 224 × 224 pixels with an overlap of 28 pixels to increase the amount of input data and to counter classification problems near edges. Since semantic segmentation performs classification of the entire images, four semantic classes are defined which cover the entire scenes: 'urban', 'vegetation', water' and 'slums'. For training and evaluation, fully labeled images are created for each data set (Fig. 2). Labeling of reference data is based on a multi-step image analysis procedure through a combination of hierarchical, knowledge-based and object-based classification, machine learning and visual image interpretation: in a first step, image objects are generated through a combined workflow of quad-tree and multi-resolution image segmentation methods. Further, spectral and spatial image features are calculated for each image object and basic landcover classes such as water and vegetation are classified using a random forest classifier based on visually derived training objects. In a subsequent step, slum patches are derived by visual image interpretation from image analysts and cross-validated. The reference map is controlled by a stratified spatial random sample of 800 test points over the image with a resulting overall accuracy of 93% and a kappa value of 0.91. Accuracy for the slum class is reported with sensitivity of 92% and a positive prediction value of 95%. For the transfer learning experiments, the reference map was adapted to the geometric resolution of each target image data set.

2.4. Experimental set-up

The FCNs are trained on an Nvidia Titan X GPU using the 'adam optimizer' (Kingma and Ba, 2014) and a batch size of two image tiles. All FCNs use fixed learning rates of 10^{-5} and a dropout value of 15%. The training methodology for the FCNs was as following: *first*, a pretrained model is initially trained for 100 epochs on all three datasets (QuickBird, Sentinel-2 and TerraSAR-X) to set-up the FCN. *Second*, two transfer learning experiments are conducted: the pretrained QuickBird-FCN is transferred on Sentinel-2 and TerraSAR. The implementation of the FCN is based on the TensorFlow™ framework of Shekkizhar (2017).

Performance of the FCN is evaluated within a 4-fold cross validation procedure where each scene is split into four equal data strips. Out of the four data strips, three strips are used as training samples which are randomly shuffled after each epoch and the remaining strip is used for validation. The cross-validation process is repeated four times, with each of the four strips used exactly once for validation. Finally, the four results of the folds are mosaicked to produce a single output covering the entire scene with each strip being the result of one of the four classification experiments and thus allowing for assessment of independent results.

For quantitative assessment of the accuracy of the outputs of semantic segmentation, some commonly accepted performance measures are used: *First*, overall measures assess the general performance and *second*, class-specific measures reveal specific insights. The kappa index is applied as a measure to define to what extent the classification outcome differs from a random result with ranges between 0 and 1; where 0 corresponds to a completely random result and 1 corresponds to a completely nonrandom result. The overall accuracy (OA) and intersection over union (IoU; also known as Jaccard Index) are calculated in addition. OA is generated from an error matrix between the classification map and the reference map and allows for a general assessment of the agreement between the two maps; however, OA can be subject to a strong bias for very imbalanced semantic class distributions.

Class-specific accuracy measures are calculated to assess the proportion of correctly classified pixels from the reference (sensitivity) and the fraction of correctly classified pixels from the output (positive prediction value; PPV). These multiple standard measures are used for comparison with other classification experiments and are, much like OA, subject to well-known biases due to class-imbalance. Therefore, IoU is used in deep learning such as PASCAL VOC and CITYSCAPES challenge (Long et al., 2015). This accuracy measure compares the similarity between two maps and is calculated by the sum of true positives divided by the sum of true positives, false positives and false negatives

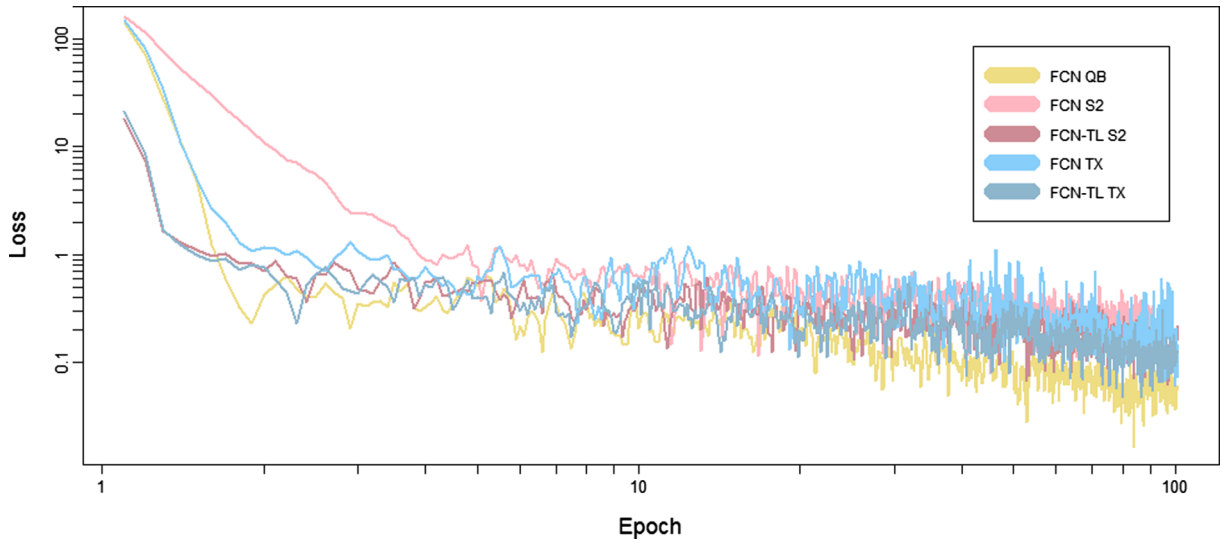

Fig. 3. Logarithmic learning curves for training five FCNs. The x-axis shows all FCNs trained for an equal duration of 100 epochs. The y-axis shows the cross entropy loss computed during training.

over the whole data set. It can be viewed as a precise indicator to the success of a classifier.

Besides the above introduced pixel-based performance evaluation strategies, a patch-based accuracy assessment is applied to account for a dependency of the slum patch area and the accuracy of the FCN. In this way, slum patch sizes are grouped into three size-based classes: smaller than 5 ha, 5–25 ha and larger than 25 ha. Accuracy assessment is performed for each slum patch size and analyzed (see Section 3.3).

3. Results and discussion

In this section, the capabilities of deep learning for slum mapping in different remotely sensed data sets with varying characteristics are analyzed subject to the quantitative results of the performed semantic segmentation experiments. Performance of the FCN is first evaluated for all four semantic classes in general and second for the slum class in particular. In total, five experiments were performed in two groups:

(1) training a pretrained model on the high resolution QuickBird image (*FCN QB*), on Sentinel-2 (*FCN S2*) and on TerraSAR-X (*FCN TX*).
(2) transfer learning of the pretrained FCN on Quickbird to Sentinel-2 (*FCN-TL S2*) and TerraSAR-X (*FCN-TL TX*).

Training the FCN is performed using a sparse softmax cross entropy loss function within TensorFlow™ to measure the performance of the model. The loss is a summation of the errors made for each example during the training stage, which implies how well or poorly a certain model behaves after each iteration of optimization. The respective loss curves are presented in Fig. 3 where all five FCNs indicate an interpretation on how well the model performs for the training datasets. All networks show convergence towards zero with some minimal jitter between 0.01 and 0.5. Both transfer learned FCNs (*FCN-TL S2 and FCN-TL TX*) reach a low loss value much faster than the pretrained FCNs, while the FCN trained on Sentinel-2 data takes considerably longer to converge against zero.

Semantic segmentation based on the FCN is performed on all total scenes (cf. extents in Fig. 2) according to the above described experimental set-up (cf. Section 2.4). A graphical depiction of the results for the same subset of a central area in Mumbai is depicted in Fig. 4. Visual interpretation of the results indicates very fine-structured patches for *FCN QB* as it is also the case in the reference data set. For that reason high accuracies are to be expected for the QB data set. As regards with the Sentinel-2 data, the effects of transfer learning become clearly visible: from large-structured patches of the results for *FCN S2*, a major increase in granularity using the transfer learning approach *FCN-TL S2* is observed: even at a geometric resolution of 10 m, small fractions of vegetation and slum patches are successfully detected. For TerraSAR-X (*FCN TX*), no significant alteration of the classification result is observed through transfer learning.

3.1. Overall accuracies

Quantitative results in terms of overall performance for the semantic segmentation are presented in Table 2 for all five experiments. With regards to overall measures, all five experiments obtained considerable accuracies with Kappa values between 0.72 and 0.85. The best performing set-up is reported, as expected, for QuickBird (*FCN-QB*). The Kappa value (0.85) and the Overall Accuracy (90.62%) show a very high agreement. This is followed by the Sentinel-2 experiment (*FCN-TL S2*) with the same Kappa value (0.85) and marginally lower OA (89.64%). Interestingly, highest IoU (87.43%) is reported for Sentinel-2 (*FCN–TL S2*) which can be considered as being mostly related to the substantially larger area of interest for Sentinel-2 (cf. Fig. 2) and the respectively larger shares of water bodies (cf. Table 3) which impact significantly the overall measures in general and the IoU in particular.

Transfer learning from the ImageNet domain $D_S^{ImageNet}$ to the remote sensing domains $D_T^{QB,S2,TX}$ performs well for the QuickBird learning task. This can be accounted for by a sufficient quantity of training data in D_T^{QB} (cf. Table 1). The second transfer task $D_T^{S2,TX}$ with less training data performs significantly poorer. Two possible reasons can explain this aspect: for the Sentinel-2 target learning task there is just not enough data available for a good knowledge transfer from $D_S^{ImageNet}$ to D_T^{S2}. The same accounts the for transfer learning task to TerraSAR-X data including another difficulty of a stark difference in feature representation of optical image data in $D_S^{ImageNet}$ and radar data in D_T^{TX}.

As regards with the performance of transfer learning against the performance of pre-trained networks, we observe remarkable differences among the transfer between QB/S2 and QB/TX: the transfer learning approach could significantly increase all overall performance measures for S2; however, no relevant change in accuracy is observed for the transfer between QuickBird and TerraSAR-X data. In fact, accuracy is even marginally lower for the transfer learning approach in this particular setting. We interpret this effect by difficulties of the network in transferring the learned model from optical features to SAR image features (cf. Hughes et al., 2018). Thus, no additional improvement of the model can be achieved.

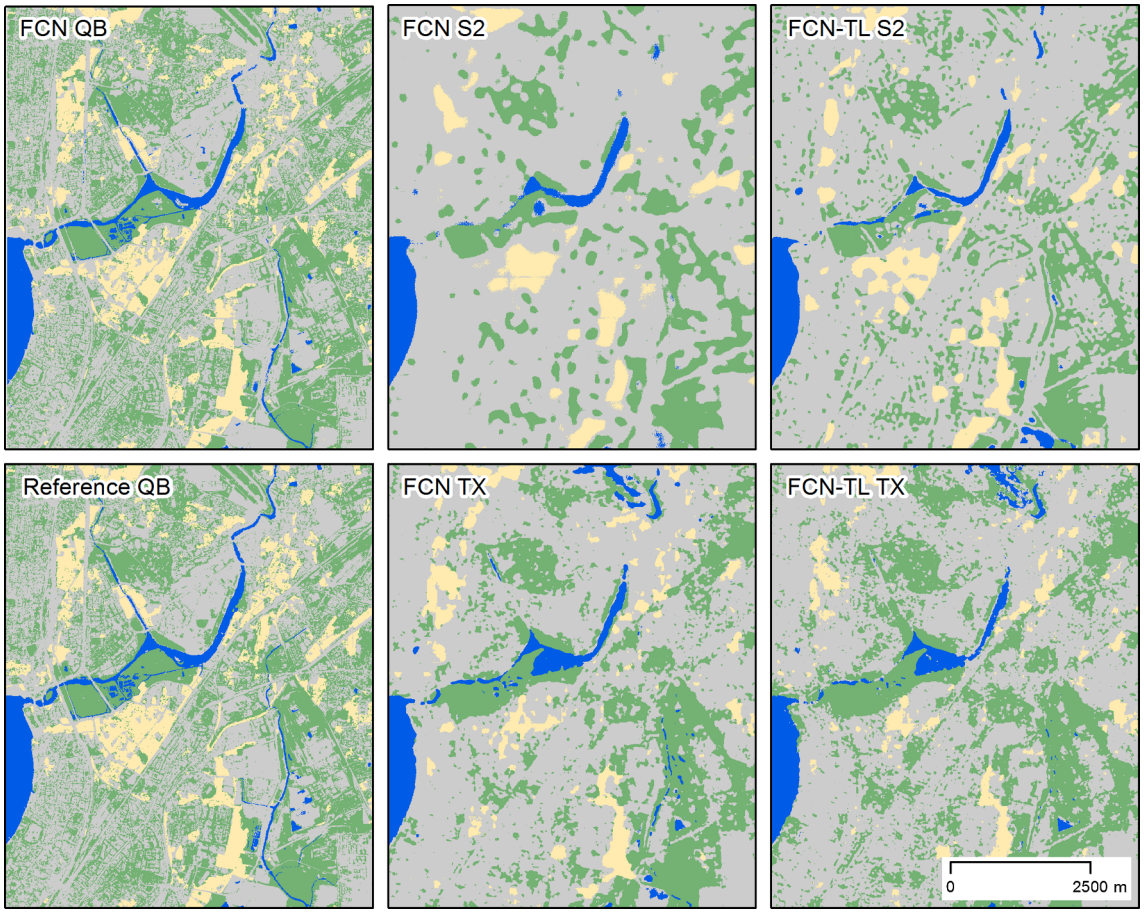

Fig. 4. Results of the semantic segmentation for the five experiments on the three data sets: QuickBird [QB], Sentinel-2 [S2] and TerraSAR-X [TX] on pre-trained FCNs and transfer learned FCNs [FCN-TL].

Table 2
Performance Evaluation of the FCN For all Classes. OA: Overall Accuracy; IoU: Intersection over Union; TL: Transfer Learned.

Approach	Kappa	OA (%)	IoU (%)
FCN-QB	**0.85**	**90.62**	84.12
FCN-S2	0.81	86.71	83.94
FCN-TL S2	**0.85**	89.64	**87.43**
FCN-TX	0.73	80.68	73.96
FCN-TL TX	0.72	80.03	73.02

Transfer learning from the QuickBird domain D_S^{QB} to the Sentinel-2 domain D_T^{S2} improves performance for all accuracy measurements significantly due to the similar feature representation in both the source and the target domain. Performance when using transfer learning techniques from the QuickBird domain D_S^{QB} to TerraSAR-X 2 domain D_T^{TX} stagnates or decreases to about 1–2% in the accuracy measurements. Prior studies have already pointed out this observation when dealing with SAR data (Zhu et al., 2017). We can confirm these issues where the upper limit of SAR classification accuracy is reached when only 2113 image tiles are available. The knowledge transfer is too difficult when transfer learning from either ImageNet or QuickBird to SAR data due to the significantly different image information representation

3.2. Class-based accuracies

While overall performance measures allow for a general assessment of the conducted experiments, detailed interpretation of class-based performance evaluation shed more light on the segmentation results. Thus, class-specific performance measures are presented in Table 3. With respect to the individual semantic classes, we observe the following: by far the highest accuracies in all performance measures for the classes 'urban' and 'slum' are obtained by QuickBird (*FCN-QB*). For

Table 3
Performance Evaluation of the FCN for the Individual Semantic Classes for the total scenes. IoU: Intersection over Union; TL: Transfer Learned; PPV: Positive Prediction Value; Sens: Sensitivity; A: area (percentage of scene coverage). Best results are marked in bold.

	Urban			Vegetation			Water			Slum		
Approach	Sens (%)	PPV (%)	IoU (%)	Sens (%)	PPV (%)	IoU (%)	Sens (%)	PPV (%)	IoU (%)	Sens (%)	PPV (%)	IoU (%)
FCN-QB	**91.37**	**90.34**	**83.24**	92.90	95.35	88.88	**90.78**	**90.97**	**83.28**	**85.70**	**88.39**	**77.02**
FCN-S2	87.47	75.87	68.43	96.42	98.44	94.97	85.35	89.72	77.75	38.21	78.82	35.51
FCN-TL S2	87.62	82.00	73.49	**97.47**	**98.57**	**96.12**	90.14	90.61	82.44	55.47	85.25	51.23
FCN-TX	84.29	83.13	71.99	93.86	94.03	88.59	78.46	75.65	62.63	51.64	72.50	46.27
FCN-TL TX	85.78	80.21	70.80	93.49	93.58	87.85	75.82	75.64	60.94	43.64	78.43	38.42

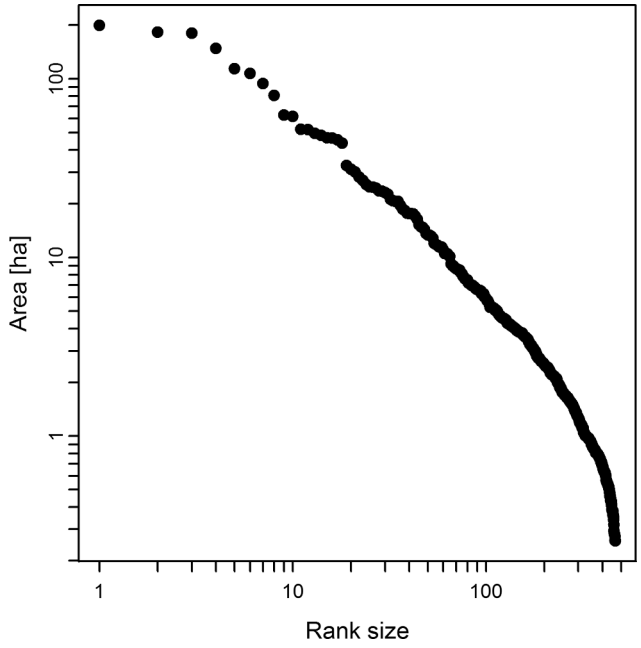

Fig. 5. Rank size distribution of slum patch sizes in Mumbai in a loglog plot.

Table 4
Proportions of number of slum patches and area for three size-based classes.

	Small slums [< 5 ha]	Medium slums [5–25 ha]	Large slums [> 25 ha]
Patches	84.63%	13.62%	1.75%
Area	26.10%	36.40%	37.50%

Table 5
Sensitivity measurement as a function of varying slum patch size.

Approach	Small slums [< 5 ha]	Medium slums [5–25 ha]	Large slums [> 25 ha]
FCN-QB	78.57	83.63	88.39
FCN-S2	09.32	28.19	47.18
FCN-TL S2	24.67	50.64	62.46
FCN-TX	31.26	47.36	55.34
FCN-TL TX	20.78	37.98	48.36

the 'vegetation' class, Sentinel-2 (*FCN-TL S2*) obtained best results, most likely due to the aggregation of information in Sentinel-2 and the consequential less small-structured vegetation fraction. Accuracies for the water class are quite similar between QuickBird (*FCN-QB*) and Sentinel-2 (*FCN-TL S2*) with only marginal differences. While for the urban class, QuickBird (*FCN-QB*) performs considerably better than Sentinel-2 (*FCN-TL S2*). The effect for the 'slum' class is most striking: the small-scaled buildings and their very organic arrangements are best segmented by the sensor with the highest geometric resolution being also capable of identifying individual buildings or shacks. Both, positive prediction value (88.4%) as well as sensitivity (85.7%) reach very high accuracies, i.e. the majority of slum areas as classified in the reference data set could be detected and only very few false positives occur. These effects are underpinned by high very IoU values (77%) which can be seen a very conservative measure of accuracy.

Comparing the results for pretraining and transfer learning, we observe a significant gain in accuracy in all semantic classes for Sentinel-2. Especially the performance of slums is increased remarkably making the effect of transfer learning in this case extremely valuable. As already reported in literature (Hughes et al., 2018), no positive effect is observed for TerraSAR-X data. Here, almost all classes are better represented by the pretraining approach (*FCN-TX*) than the transfer learning approach (*FCN-TL TX);* however, with one exception: PPV of slums is increased. If considering only the slum class, however, very competitive results in comparison to Sentinel-2 are obtained (55.47% vs. 51.64%).

All in all, we can state the following:

(1) the pretrained network on QuickBird performs very well in classifying heterogeneous urban environments.
(2) transfer learning for Sentinel-2 can significantly improve the results.
(3) for TerraSAR-X performance is reported lower than for the optical data.
(4) Transfer learning for TerraSAR-X could not improve the performance.

3.3. The impact of slum patch size

As stressed already in prior studies slum patch sizes vary significantly within cities (e.g. Wurm et al., 2017). Friesen et al. (2018) found that slum patch size distribution in several mega cities in the world follow very closely Zipf's law and can be analyzed via rank size distribution (Zipf, 1941). The case for Mumbai is presented in Fig. 5. We observe a majority of small slums with areas below 5 ha and only a handful of large slums above 25 ha. Their respective contribution to the total slum area is, however, inverse, as presented in Table 4.

Based on these observations, we additionally perform a patch size-based accuracy assessment for the specific class of 'slums' to analyze the impact of slum patch size on the resulting classification performance. Both, a visual comparison for all approaches, and a quantitative assessment of sensitivity are conducted (Table 5). Small slum patches (< 5 ha) are presented in Fig. 6 with very good slum mapping capabilities for QuickBird (*FCN-QB:* 78.57%). Further, a significant increase of sensitivity for Sentinel-2 between pretrained and transfer learned is observed (9.32 vs. 24.67%). Prior discussed effects for TerraSAR-X images are also observed for the smallest group of patches: decreasing sensitivity between pretrained and transfer learned (31.26 vs. 20.78%). Both, Sentinel-2 and TerraSAR-X, however, perform very poor for this smallest group of patch sizes which is to be expected at image resolutions of 10 m and 6 m, respectively.

Medium-sized slum patches are presented in Fig. 7. Here the same trend is identified as for small patches: highest sensitivity is obtained by QuickBird (*FCN-QB*: 83.63%) and transfer learning significantly improves slum patch detection for Sentinel-2 against pre-training (28.19 vs. 50.64%). Again, a decrease is measured for the approach using TerraSAR-X (47.36 vs. 37.98%).

Finally, results for large slum patches (Fig. 8) are reported highest for all performed experiments. In QuickBird 88.39% of the reference slum pixels are detected (*FCN-QB*). For Sentinel-2, again, transfer learning significantly enhances mapping capabilities (47.18 vs. 62.46%) and a decrease in a performance is observed for TerraSAR-X (48.36 vs. 55.34%). Summarizing these observations, a strong effect of slum patch size on the detection rate is reported for all experiments (cf. Wurm et al., 2017).

4. Conclusion

In this paper, we perform a series of experiments to analyze the capabilities of fully convolutional neural networks for semantic segmentation of slums for the example of Megacity Mumbai using satellite images with different characteristics. As a result, we observe the following effects:

(1) very high geometric resolution of 0.5 m in QuickBird imagery allows for the best results of all experiments.
(2) transfer learning of a pre-trained network from QuickBird to

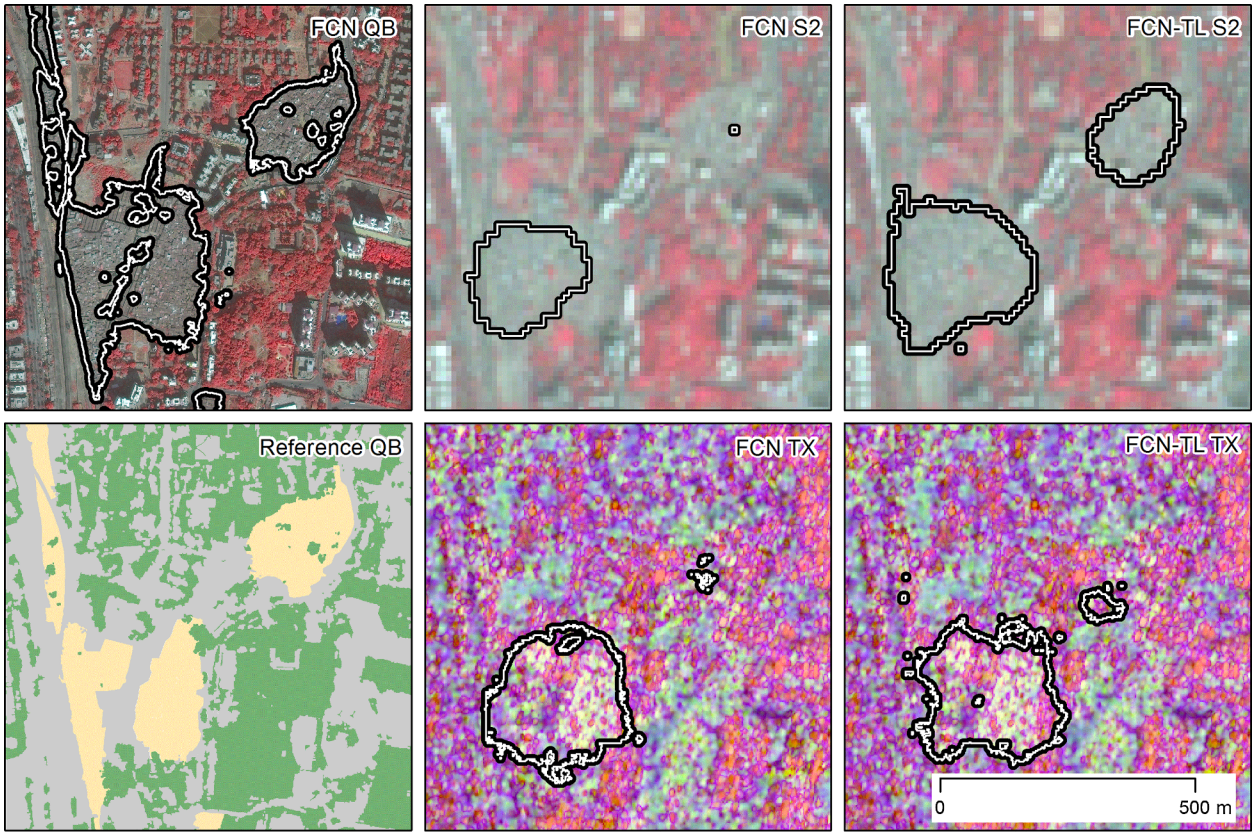

Fig. 6. Comparative alignment of small slum patches [< 5 ha] showing differences in segmentation results obtained by pre-trained FCNs and transfer learned FCNs (FCN-TL) on QuickBird, Sentinel-2 and TerraSAR-X images. Slum patches in the reference map are depicted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

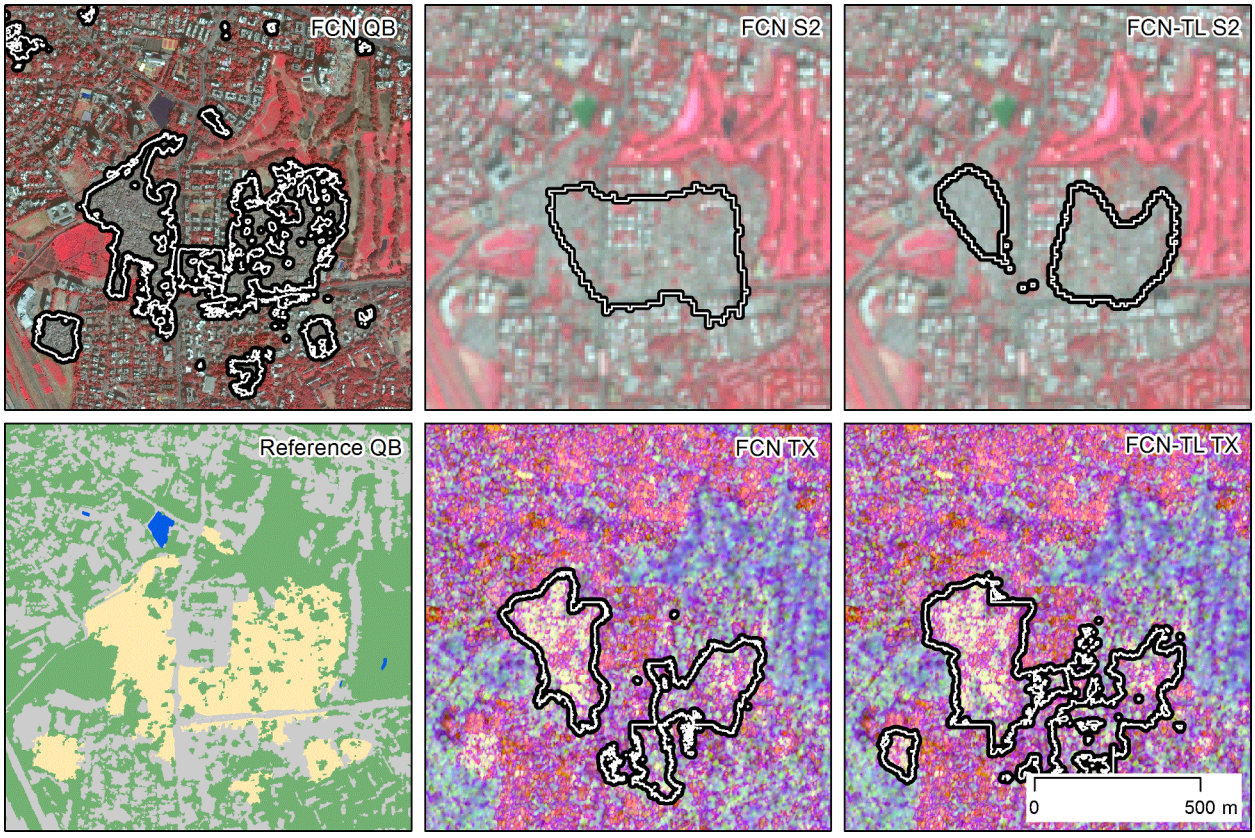

Fig. 7. Comparative alignment of medium sized slums [5 ha–25 ha] showing differences in segmentation results obtained by pre-trained FCNs and transfer learned FCNs (FCN-TL) on QuickBird, Sentinel-2 and TerraSAR-X images. Slum patches in the reference map are depicted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

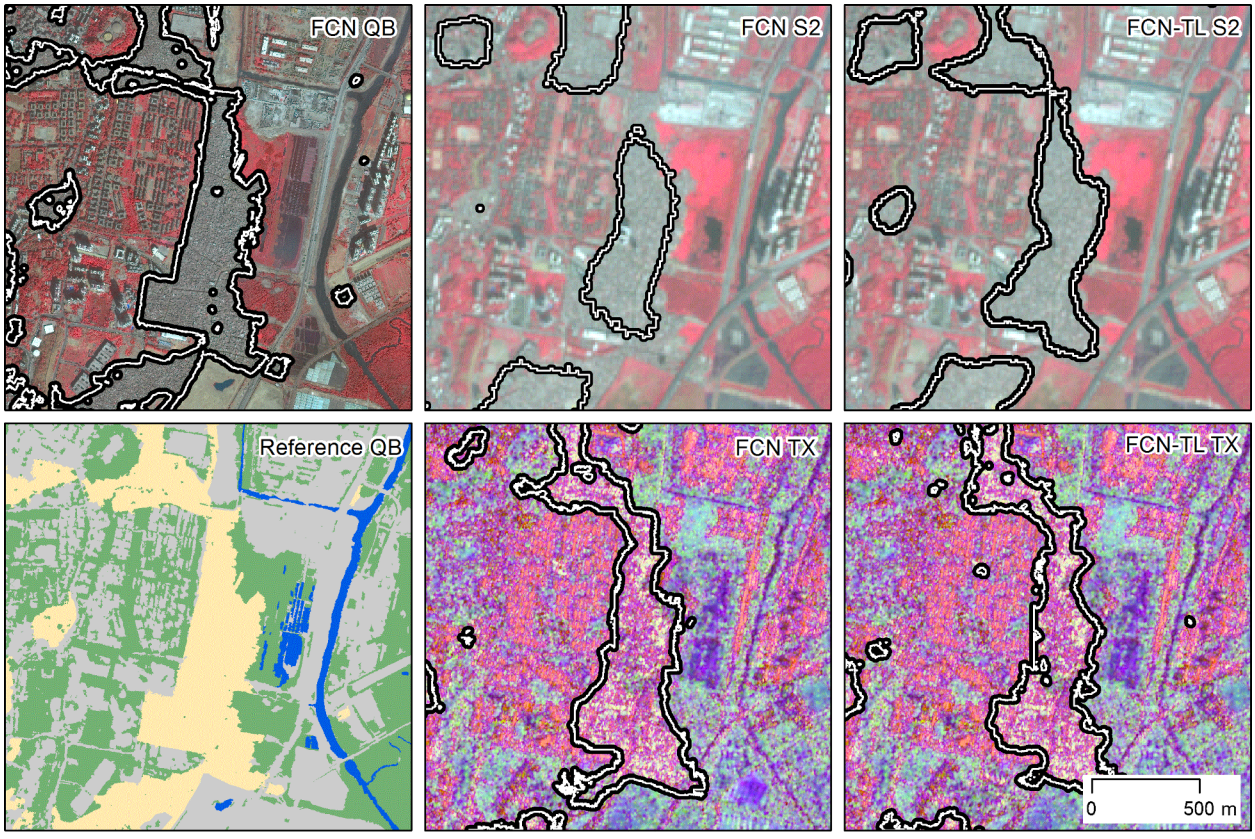

Fig. 8. Comparative alignment of a large slum patch [> 25 ha] showing differences in segmentation results obtained by pretrained FCNs and transfer learned FCNs (FCN-TL) on QuickBird, Sentinel-2 and TerraSAR-X images. Slum patches in the reference map are depicted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Sentinel-2 images significantly improves the segmentation results. This makes medium resolution sensors at 10 m GSD an opportunity for very large-area mapping of slums for entire countries or sub-continents.

(3) for active satellite imagery such as TerraSAR-X, the transfer learning approach does not improve the results, but even decrease the performance. We relate this observation to the fact that the network is not able to transfer the learned image features from optical imagery to the SAR representation of urban structures.

(4) Further, we observe a strong effect of slum patch size for being detected by the segmentation approaches. While this effect is smallest for high resolution QuickBird imagery which already performs at a very high level: from 79.57% for < 5 ha to 88.39% for > 25 ha, an increase from 9.32 to 47.18% in sensitivity is obtained for Sentinel-2 pretrained (*FCN-S2*) and from 24.67 to 62.46% for Sentinel-2 transfer learned (*FCN-TL S2*). The same effect is also observed for TerraSAR-X: from 31.26 to 55.34% for pre-trained (*FCN-TX*) and 20.78 to 48.36% for transfer learned, respectively (*FCN-TL TX*).

Finally, segmentation outcomes are extremely promising and encouraging for further experiments using transfer learning and fully convolutional networks for slum mapping in satellite imagery. Further experiments need to focus on large-area approaches and the transfer between different geographical regions. This challenging task needs to address the morphological representations of slums in different cultural areas as shown by Taubenböck et al. (2018), since the physical nature of slums is represented by a large variety of morphological structures.

Funding

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No [ERC-2016-StG-714087], Acronym: So2Sat).

References

Amnesty International, 2016. Eine Milliarde Menschen Leben in Slums. https://www.amnesty.de/mit-menschenrechten- gegen-armut/wohnen-wuerde/eine-milliarde-menschen-weltweit-leben-slums.

Aytar, Y., Zisserman, A., 2011. Tabula rasa: Model transfer for object category detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 2252–2259.

Baud, I., Kuffer, M., Pfeffer, K., Sliuzas, R.V., 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.

Burdett, R., Rhode, P., 2010. Living in the urban age. In: Living in the Endless City. Phaidon, pp. 8–43.

Castelluccio, M., Poggi, G., Sansone, C., Verdoliva., L., 2015. Land use classification in remote sensing images by convolutional neural networks. arXiv preprint arXiv:1508.00092.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. IEEE Computer Vision and Pattern Recognition (CVPR).

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

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

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 38 (1), 142–158.

Gong, M., Yang, H., Zhang, P., 2017. Feature learning and change feature classification

based on deep learning for ternary change detection in SAR images. ISPRS J. Photogramm. Remote Sens. 129, 212–225. https://doi.org/10.1016/j.isprsjprs.2017.05.001.
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 (4), 1164–1176. https://doi.org/10.1109/JSTARS.2012.2190383.
Hu, F., Xia, G.S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7 (11), 14680–14707.
Huang, X., Liu, H., Zhang, L., 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 53, 3639–3657.
Hughes, L., Schmitt, M., Mou, L., Wang, Y., Zhu, X., 2018. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. IEEE Geosci. Remote Sens. Lett. 15 (5), 784–788.
Jain, S., 2007. Use of IKONOS satellite data to identify informal settlements in Dehradun, India. Int. J. Remote Sens. 28, 3227–3233.
Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), 790–794. https://doi.org/10.1126/science.aaf7894.
Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145 (Part A), 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.
Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens. 145, 60–77. https://doi.org/10.1016/j.isprsjprs.2018.04.014.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kuffer, M., Barros, J., Sliuzas, R., 2014. The development of amorphological unplanned settlement index using very-high-resolution (VHR) imagery. Comput. Environ. Urban. Syst. 48, 138–152. https://doi.org/10.1016/j.compenvurbsys.2014.07.012.
Kuffer, M., Pfeffer, K., Sliuzas, R., 2016a. Slums from space – 15 years of slum mapping using remote sensing. Remote Sens. 8 (6), 455. https://doi.org/10.3390/rs8060455.
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 (5), 1830–1840. https://doi.org/10.1109/JSTARS.2016.2538563.
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 (4), 384. https://doi.org/10.3390/rs9040384.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, Cham, pp. 740–755.
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.
Ma, Z., Yang, Y., Nie, F., Sebe, N., Yan, S., Hauptmann, A.G., 2014. Harnessing lab knowledge for real-world action recognition. Int. J. Comput. Vision 109 (1–2), 60–73.
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 (2), 645–657.
Mahabir, R., Croitoru, A., Crooks, A.T., Agouris, P., Stefanidis, A., 2018. A Critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: trends, challenges and emerging opportunities. Urban Sci. 2 (1), 8. https://doi.org/10.3390/urbansci2010008.
Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geosci. Remote Sens. Lett. 13 (1), 105–109.
Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009.
Mou, L., Ghamisi, P., Zhu, X.X., 2018. Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 56 (1), 391–406.
Mou, L., Zhu, X., Vakalopoulou, M., Karantzalos, K., Paragios, N., Le Saux, B., Moser, G., Tuia, D., 2017. Multi-temporal very high resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (8), 3435–3447.
Nogueira, K., Penatti, O.A., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn. 61, 539–556.
Owen, K.K., Wong, D.W., 2013. An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics. Appl. Geogr. 38, 107–118. https://doi.org/10.1016/j.apgeog.2012.11.016.
Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Pan, B., Shi, Z., Xu, X., 2018. MugNet: deep learning for hyperspectral image classification using limited samples. ISPRS J. Photogramm. Remote Sens. 145, 108–119. https://doi.org/10.1016/j.isprsjprs.2017.11.003.
Penatti, O.A., Nogueira, K., dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 44–51.
Persello, C., Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in VHR images. IEEE Geosci. Remote Sens. Lett. 14 (12), 2325–2329. https://doi.org/10.1109/LGRS.2017.2763738.
Salakhutdinov, R., Torralba, A., Tenenbaum, J., 2011. Learning to share visual appearance for multiclass object detection. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR, pp. 1481–1488.
Sandborn, A., Engstrom, R.N., 2016. Determining the relationship between census data and spatial features derived from high-resolution imagery in Accra, Ghana. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (5), 1970–1977. https://doi.org/10.1109/JSTARS.2016.2519843.
Schmitt, A., Sieg, T., Wurm, M., Taubenböck, H., 2018. Investigation on the separability of slums by multi-aspect TerraSAR-X dual-co-polarized high resolution spotlight images based on the multi-scale evaluation of local distributions. Int. J. Appl. Earth Obs. Geoinform. 64, 181–198. https://doi.org/10.1016/j.jag.2017.09.006.
Shekkizhar, S., 2017. FCN.tensorflow. GitHub https://github.com/shekkizh/FCN.tensorflow.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: Proc. International Conference on Learning Representations.
Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos. Computer Vision. Proceedings. Ninth IEEE International Conference on.
Sun, Z., Wang, C., Li, P., Wang, H., Li, J., 2012. Hyperspectral image classification with SVM-based domain adaption classifiers. In: IEEE International Conference on Computer Vision in Remote Sensing (CVRS), pp. 268–272.
Taubenböck, H., Wurm, M., 2015. Ich weiß, dass ich nichts weiß – Bevölkerungsschätzung in der Megacity Mumbai. In: Taubenböck, Wurm, Esch, Dech (Eds.), Globale Urbanisierung. Perspektive aus dem All: 171–178. Springer Spektrum.
Taubenböck, H., Kraff, N.J., Wurm, M., 2018. The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. Appl. Geogr. 92, 150–167. https://doi.org/10.1016/j.apgeog.2018.02.002.
United Nations, 2017. The sustainable Development Goals Report. https://unstats.un.org/sdgs/files/report/2017/TheSustainableDevelopmentGoalsReport2017.pdf.
UN, 2015. The world urbanization prospects. The 2014 revision. http://esa.un.org/unpd/wup/FinalReport/WUP2014-Report.pdf.
UN Habitat 2015: Slum Almanac 2015-2016. https://unhabitat.org/slum-almanac-2015-2016/#.
Wurm, M., Taubenböck, H., Weigand, M., Schmitt, A., 2017. Slum mapping in polarimetric SAR data using spatial features. Remote Sens. Environ. 194, 190–204. https://doi.org/10.1016/j.rse.2017.03.030.
Wurm, M., Taubenböck, H., 2018. Detecting social groups from space –Assessment of remote sensing-based mapped morphological slums using income data. Remote Sens. Lett. 9 (1), 41–50. https://doi.org/10.1080/2150704X.2017.1384586.
Xia, J., Yokoya, N., Iwasaki, A., 2017. Ensemble of transfer component analysis for domain adaptation in hyperspectral remote sensing image classification. In: In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4762–4765.
Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5 (4), 8–36.
Zhu, Q., Zhong, Y., Zhao, B., Xia, G.S., Zhang, L., 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci. Remote Sens. Lett. 13 (6), 747–751.
Zipf, G.K., 1941. National unity and disunity - the nation as a bio-social organism. Bloomington, Indiana. https://babel.hathitrust.org/cgi/pt?id=mdp.39015057175484;view=1up;seq=5.
Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 12 (11), 2321–2325.