Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to Covid-19 response in Khartoum, Sudan

Dirk Tiede¹ | Gina Schwendemann² | Ahmad Alobaidi² | Lorenz Wendt¹ | Stefan Lang¹

1Christian Doppler Laboratory for Geospatial and EO-Based Humanitarian Technologies, Department of Geoinformatics – Z_GIS, University of Salzburg, Salzburg, Austria
2Spatial Services GmbH, Salzburg, Austria

Correspondence
Dirk Tiede, Christian Doppler Laboratory for Geospatial and EO-Based Humanitarian Technologies, Department of Geoinformatics – Z_GIS, University of Salzburg, Salzburg, Austria.
Email: dirk.tiede@sbg.ac.at

Abstract
Within the constraints of operational work supporting humanitarian organizations in their response to the Covid-19 pandemic, we conducted building extraction for Khartoum, Sudan. We extracted approximately 1.2 million dwellings and buildings, using a Mask R-CNN deep learning approach from a Pléiades very high-resolution satellite image with 0.5 m pixel resolution. Starting from an untrained network, we digitized a few hundred samples and iteratively increased the number of samples by validating initial classification results and adding them to the sample collection. We were able to strike a balance between the need for timely information and the accuracy of the result by combining the output from three different models, each aiming at distinctive types of buildings, in a post-processing workflow. We obtained a recall of 0.78, precision of 0.77 and F₁ score of 0.78, and were able to deliver first results in only 10 days after the initial request. The procedure shows the great potential of convolutional neural network frameworks in combination with GIS routines for dwelling extraction even in an operational setting.
1 | INTRODUCTION

A central piece of information required for almost any humanitarian intervention are accurate numbers about the population in need, and the locations or areas where the affected population resides (Lang et al., 2020). In particular, in protracted crises, traditional census data are missing or outdated. Building footprint polygons can be used to estimate population numbers, either by distributing known population numbers from larger administrative units onto the individual footprints (top-down, dasymetric mapping approach; cf. Eicher & Brewer, 2001) or by conducting microcensuses to calculate average occupancy rates per building over a small subset of buildings, and then extrapolating across the entire city (bottom-up approach; for an overview see Checchi, Stewart, Palmer, and Grundy (2013). When building footprints are not available or updated, for example in OpenStreetMap (OSM), they can be extracted from very-high resolution satellite images.

In Khartoum, Sudan, a city with an estimated population of the order of 5.1 million inhabitants (Zerboni et al., 2020) and a surface area of around 1,000 km², the latest census dates back to 2008 (United Nations Department of Economic and Social Affairs, 2019). OSM contains a fairly complete street network, but building footprints are available only for a few parts of the city. Therefore, in April 2020 Médecins sans Frontières (MSF) requested a recent map of dwelling density distribution across the city, and later a map of individual dwellings/building footprints. MSF supports four hospitals and five clinics in the city, providing care to Covid-19 patients, promoting infection prevention and control measures, and helping to keep essential medical services up and running. MSF used these maps to estimate health-care demand at the health-care posts they support.

In this article we report on the usage of a deep learning approach in such an operational humanitarian response context in a training-sample-scarce situation, in combination with GIS post-processing routines to combine and refine the results of the individual deep learning classifiers. Machine learning approaches, especially deep convolutional neural networks (CNN), have successfully shown their capabilities in unprecedented object-extraction accuracies also from remotely sensed images. Still, ready-to-use networks face challenges with unfamiliar context and less typical situations, where specific sample databases do not (yet) exist. Training or retraining of established networks needs a lot of effort in these situations, since the networks do not generalize very well (Marcus, 2018). We present our applied approach, balancing out the need for timely information and the accuracy of the result in a sample-scarce and quite complex situation of dense buildings in a highly dynamic, growing mega-city (Zerboni et al., 2020).

2 | RELATED WORK

Automated dwelling and building extraction from very high-resolution (VHR) optical satellite imagery for humanitarian purposes, mainly for population estimation in refugee or internally displaced person camps, but increasingly including informal settlements within fast-growing cities, has about two decades of history (see, for example, Bjorgo, 2000; Checchi et al., 2013; Ehrlich et al., 2009; Witmer, 2015). (Semi-)operational solutions were developed to support humanitarian actors with up-to-date information especially in dynamically developing situations (Lang et al., 2020; Tiede, Füreder, Lang, Hölbling, & Zeil, 2013).

Pixel-based approaches were rapidly replaced due to their limitations in spatial context analysis by object-based image analysis (OBIA) or mathematical morphology-based approaches (Giada, De Groeve, Ehrlich, & Soille, 2003; Kemper & Heinzel, 2014; Kemper, Jenerowicz, Pesaresi, & Soille, 2011; Knoth & PEBESMA, 2017; Lang, Tiede, Hölbling, Füreder, & Zeil, 2010; Spröhnle et al., 2014; Stängel, Tiede, Lüthje, Füreder, & Lang, 2014; Tiede, Krafft, Füreder, & Lang, 2017), while the development of semi-automated and fully automated algorithms took quite some time. Challenges in an operational context of humanitarian actions include: quality issues of the images available in crisis situations; the small scale of the objects/dwellings to be extracted; the unplanned dynamic structure of the settlements; and varying environmental conditions. While knowledge-based approaches
have clear advantages (e.g., sample data are not needed), they are time-consuming to set up and limited in their
transferability with ever-changing environmental and atmospheric conditions of the target scenes.

For some years now, CNNs have become prominent in satellite image analysis (Hoeser, Bachofer, &
Kuenzer, 2020; Ma et al., 2019), and various sample databases have been established with the aim of training
transferable and more generically applicable networks. However, current applications still need a lot of ad-
titional training samples to obtain acceptable results, especially for non-standard tasks in information extraction
(Hoeser & Kuenzer, 2020). CNNs are nowadays also used in humanitarian mapping tasks; for example, Quinn et al. (2018)
reviewed recent developments and conducted initial experiments for selected refugee camps where
manually mapped data were available, concluding that full automation is not yet possible. Ghorbanzadeh, Tiede,
Dabiri, Sudmanns, and Lang (2018) and Ghorbanzadeh, Tiede, Wendt, Sudmanns, and Lang (2020) showed for a
single refugee camp, yet with different VHR satellite sensors and different time-steps, how CNNs can be coupled
with knowledge-based OBIA approaches. Lu, Koperski, Kwan, and Li (2020) extracted tents in a Syrian refugee
camp, comparing their proposed fully convolutional network based on an ImageNet pretrained VGG-16 network
with different existing CNN networks and manually labelled data. Most documented approaches have in common
that they are not benchmarked at an operational setting. Also, manually labelled data for the whole area under
investigation were available and timely results are not considered crucial in these experimental works. We con-
clude that they have limited capacity in sample-scarce situations.

Building extraction not specifically related to humanitarian response is as another broad field in deep learning
applications, and in particular the DeepGlobe 2018 Satellite Image Understanding Challenge should be mentioned
here (Demir et al., 2018). In this challenge, data and training samples for Khartoum were also available, but a
cross-check of the data set provided (WorldView-3 image from 2015 in off-nadir view) and the respective training
samples were not readily useful for this task due to their very detailed building footprints, yet shifted slightly and
outdated compared to our satellite scene.

The special situations we face in humanitarian mapping, namely, fast delivery of results for areas not previ-
ously mapped with large numbers of informal dwelling structures, some of them small in size and attached to each
other, are a challenge using training-data-hungry CNNs. Perfectly transferable, already trained networks are not
(yet) available; so retraining on top of an initially trained network is needed. Also, additional information is often
not available for these areas and the most recent images often do not show perfect conditions in respect of view-
ing angle, seasonality, atmospheric disturbances and so on.

The biggest challenge in this operational work remains, namely, to provide enough samples in a short time-
frame to train the CNN so that results would fulfil two main goals: fast delivery of a map of the area indicating
density and sizes of the buildings; and sufficiently high accuracy to support humanitarian operational work. In the
following we report on our workflow, aiming not so much at the ultimate accuracy CNNs can reach in perfect con-
ditions, but rather to foster operational usage in humanitarian operations, optimizing both timeliness and accuracy
in sample-scarce situations.

3 | MATERIAL AND METHODS

The overall workflow is depicted in Figure 1 and explained in this section in detail.

3.1 | Study area

We used a Pléiades 1A image, acquired on 8 November 2019 and pansharpened to 0.5 m ground sampling distance,
which covered approx. 825 km² of the central and eastern part of Khartoum (see Figure 2). Dwelling extraction
and dwelling density mapping were done under the premise of targeting smaller units and their respective density (hence the term “dwelling”), not the precise mapping of building footprints for cadastral purposes and the like.

3.2 | Deep learning

A Mask R-CNN approach (He, Gkioxari, Dollár, & Girshick, 2017) was employed because of its remarkable results in instance segmentation and object-extraction capability. Mask R-CNN is an extension of Faster R-CNN (Ren, He, Girshick, & Sun, 2017), a class of region-based CNN that has been speed-optimized for classification as well as object detection using proposed regions (bounding boxes) for multiple objects present in an image. Mask R-CNN extends this approach using small, fully connected networks applied to each region, predicting an object-specific segmentation mask per pixel, thus extracting target objects without background. Sampling focuses therefore on target objects only; no background or negative samples are needed, which is an advantage over other approaches in a time-critical, operational setting. Mask R-CNN has been successfully applied in similar satellite image analysis tasks, for example for sparse and multi-sized object detection in VHR images (Wang, Tao, Wang, Wang, & Li, 2019), building extraction (Wen et al., 2019) or within the DeepGlobe Building Extraction Challenge (Zhao, Kang, Jung, & Sohn, 2018).

We used Mask R-CNN as implemented in the Python API for Esri’s ArcGIS environment, which allows access to the underlying deep learning algorithms within a GIS environment to perform sample collection, training sample generation, classification, post-processing and map production in a seamless workflow to save time in the production process. Mask R-CNN provides GIS-ready objects, and it outperforms other approaches in this respect.

3.3 | Training sample generation

Due to the huge number of dwellings and the lack of a priori sample data, we followed a parallel approach: two trained image analysts digitized dwellings on image subsets representing specific generic building footprints; in parallel, a Mask R-CNN network was trained using the growing sample database. Initial results of the Mask R-CNN network for different subsets of the image data were checked and validated by the interpreters against visual
results and the positively evaluated ones were added to the growing sample database without additional digitizing work; several iterations helped to rapidly enhance the sample database much more efficiently compared to a manual approach. The goal was to find the trade-off between time-consuming manual sample extraction and a trained network able to deliver a solid overview of building density and structure distribution across the city. Based on the initial results, three different models were trained: one model for small dwellings using training sample augmentation (rotation by increments of $10^\circ$); a second one for larger dwellings without augmentation; and a third model for very dense blocks of buildings. Initial tests with the second model showed that augmentation did not lead to better results, most likely due to the shadows casted by the buildings. The shadows seem to influence the object detection process, and rotation of directed shadows did not improve the results here. The third model focused on building blocks, meaning larger aggregates of concatenated dwellings, in particular in densely built-up areas with large numbers of very small dwellings. The blocks of buildings (sampled for the densest parts of the city) should help to indicate these areas in the final density maps. Also here the sample size was increased through augmentation. For all three models ResNet-50 was selected as the backbone.

The numbers of training samples used for the training of the three final models (mixture of manually digitized samples and validated initial results of the Mask R-CNN models), and their average size, are summarized in Table 1.
3.4 | Post-processing

We deliberately allowed low-probability thresholds in order to create dwelling footprint layers for each size category; afterwards we employed post-processing based on the intersection of footprints and the probability values of each polygon to reduce double-counting and errors. The aim of this strategy was to "relax" the network to some degree, and to recognize non-typical building instances, for which only a few samples had been collected. Post-processing was done with the help of knowledge-based GIS routines (union of feature, identical removal, elimination/dissolve of slivers) and removal of dwellings smaller than 7 m² (defined as the minimum dwelling size based on visual inspection). For the final map production a visual check and manual cleaning of obvious misclassification/errors were also conducted.

3.5 | Accuracy assessment

For the accuracy assessment a 50 m × 50 m grid was created over the area of Khartoum. Then 355 of these cells were randomly selected (see Figure 2) and dwellings were manually delineated by two trained mapping experts not involved in the study itself. The assessment was conducted on the post-processed data for the dwelling classes before the manual refinement took place, in order to evaluate the automated part of the approach only. Precision, recall, and \( F_1 \) were calculated based on true positives (TP), false positives (FP), and false negatives (FN) detected within the validation cells. Precision indicates the proportion of target dwellings correctly identified by the proposed approach:

\[
\text{precision} = \frac{TP}{TP + FP}
\]

while recall indicates the proportion of target dwellings in the validation data that were correctly detected by the approach:

\[
\text{recall} = \frac{TP}{TP + FN}
\]

\( F_1 \) is used to balance precision and recall parameters:

\[
F_1 = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

For an evaluation of our building blocks result, a comparison with a new data set published after our map production in July 2020 was conducted (Dooley, Boo, Leasure, & Tatem, 2020). This data set is based on recent building footprint analyses, which are not yet publically available.
RESULTS AND DISCUSSION

In total, we extracted approximately 1.4 million dwellings automatically (1,099,753 small dwellings from the first model and 297,770 larger dwellings from the second model) and 26,512 dense building blocks. These figures were reduced to 1.17 million dwellings after the automated post-processing steps (890,074 small dwellings and 283,069 larger dwellings) and 20,091 dense building blocks (Table 2). Figure 3 shows a subset of the area analysed and the dwelling footprints of the small and large dwelling category before post-processing took place. This was achieved using only a few hundred completely manually digitized training samples, complemented by a few thousand automatically extracted and manually verified samples (see Table 1 and Figure 1). Despite limited staff time and computer power (three researchers involved part-time in the analysis process, including manual sample generation and automation; training and analysis conducted on a standard workstation with an NVIDIA Quadro P4000 8 GB graphics card), we were able to deliver first results within 10 days after receiving the request, followed by some more detailed analyses and final map production a few days later. Before the final maps were created some manual refinement took place, mainly to remove clutter such as obvious outliers and misclassifications. Two final maps were produced, a dwelling density map where dwelling area was aggregated to a 1 ha sized hexagon-grid for an overview of different dwelling density zones within the city, and a second map showing the extracted dwellings. We distinguish between six dwelling categories by size, ranging from 7 m² to over 900 m² in surface area, and overlaid the results on the outlines of the additionally extracted building blocks to indicate the highest dwelling density areas (Figure 4).

In 172 out of the 355 validation cells no dwellings—either in the validation data or in the automated extraction—were present (see the yellow cells in Figure 2). This shows that false positives and false negatives are in this case not mixed up with non-dwelling classes outside of the urban area; results are a good match to the city structure. Instead, they are more related to problems within the dense building areas. For the resulting 183 cells showing dwellings, the accuracy figures are shown in Table 3.

In light of the limited time available for this information request, the validation shows quite good results for these 183 cells. The number of buildings is very similar, and in particular the extent of the area covered by buildings seem to be a very good match, but the real area overlap reveals some problems with the exact locations. A spatially explicit validation was conducted by a simple spatial match of polygons originating from the manual validation and the Mask R-CNN results, counting true positives, false positives, and false negatives. The precision, recall and $F_1$ scores are very similar (0.78, 0.77 and 0.78), showing a good balance between two various aspects of the error.

Figure 5 provides a closer look at selected validation cells, and reveals a generally less detailed delineation of the dwellings as compared to manual digitization, as well as some problems in very dense areas, where many small dwellings are attached to each other (e.g. Figures 5a,c). Underestimation of dwellings in very densely populated, built-up regions (very densely clustered or attached small dwellings) is a finding noticed already in the first feedback cycles of the Mask R-CNN models based on the first training samples. Instead of sampling a lot of these very small dwellings (due to time constraints and limitations expected for the results) a third building type on a different scale level (dense building blocks) was sampled and analysed in addition (third Mask R-CNN model) and served in

| Class          | Number of dwellings before post-processing | Number of dwellings after post-processing | Mean area per dwelling of the post-processed data (m²) |
|----------------|-------------------------------------------|------------------------------------------|------------------------------------------------------|
| Small dwellings| 1,099,753                                  | 890,074                                  | 31                                                   |
| Larger dwellings| 297,770                                    | 283,069                                  | 231                                                  |
| Building blocks| 26,512                                     | 20,091                                   | 3,717                                                 |
our approach as a final indicator map for locations where very dense structures are occurring. Figure 6 shows a subset of such a densely built-up area and the delineated building blocks.

This result acts as an uncertainty layer providing additional information, and has been validated visually before map production. A comparison with a new data set published after our map production (Dooley et al., 2020) is shown in Figure 7; it exhibits a very similar pattern for the highest building density areas.
FIGURE 4  Example of a final map delivered to MSF showing single dwellings, categorized into six different size classes (blue is used for dwellings below 900 m² in size, pink indicates larger buildings most likely in a non-residential area; see Figure 6 for details). Very dense building blocks are overlaid in yellow (outlines). Insets explain the different dwelling density patterns.
The accuracy of our results needs to be evaluated bearing in mind the short time-frame and limited resources available in such an operational mapping task. Comparable studies, even in more experimental, less operational settings, have achieved similar results. For example, Yuan et al. (2018) extracted building footprints in parts of

**TABLE 3** Results of the accuracy assessment for the automatically extracted dwellings (before manual refinement)

| Accuracy assessment for 355 randomly selected cells (50 m × 50 m) |
|---------------------------------------------------------------|
| Number of dwellings (Mask R-CNN)                             | 1,438 |
| Number of dwellings (validation)                            | 1,627 |
| TP (overlap)                                                | 1,127 |
| FP (no overlap)                                             | 311  |
| FN (reference not overlapping Mask R-CNN results)           | 331  |
| Total area overlap (m²)                                     | 53,633 |
| Total area reference (m²)                                   | 84,265 |
| Total area Mask R-CNN (m²)                                  | 85,217 |
| Precision                                                   | 0.78  |
| Recall                                                      | 0.77  |
| $F_1$                                                       | 0.78  |

**FIGURE 5** Visualization of six validation cells (out of 355), green outlines showing the independently manually digitized dwellings, blue showing the automatically extracted dwellings after the post-processing steps (with different blue tones indicating the size according to the legend shown in Figure 6)
Nigeria based on OSM samples and used a similar manual refinement of sample data based on initial deep learning results as in our study. They achieve a final $F_1$ score of 0.717, while not focusing primarily on single building extractions but on a settlement map. Zhao et al. (2018) used a modified Mask R-CNN within the DeepGlobe Building Extraction Challenge, showing high $F_1$ scores for the cities of Las Vegas (0.881) and Paris (0.760), but also much lower values for Khartoum (0.578). These values are not directly comparable to ours, since intersection over union was used in the evaluation procedure of the DeepGlobe Building Extraction Challenge, while we evaluated the spatial match of buildings without looking into intersecting parts only. The reason is, that we were aiming for a correct number of buildings for population estimation in this humanitarian context and less for correct building delineation. Nevertheless, the comparison of the results underpins the challenges of the complex situation of dense buildings in a highly dynamic mega-city like Khartoum. Major problems in our study occurred therefore mainly due to very dense and attached nature of buildings in some parts of the city. The introduction of an additional dense building block detection layer helped in identifying these dense areas, showing also the areas with higher uncertainty in our approach. Visual inspection prior to map production revealed that larger rectangular dwellings were mixed up with smaller dwellings surrounded by a rectangular wall and a similar cast shadow. These errors were minor, but should be addressed in the sampling process in the future.

The iterative sampling strategy, interacting with initial training results from the Mask R-CNN approach, helped significantly in reducing time and effort required for the training sample production. Accuracy limitations need to be taken into consideration whenever a balance between delivery of fast initial results and reliability is needed to match the user expectations in the humanitarian response of concern. In this case the user, MSF, was satisfied with the accuracy of the map delivered; the information rapidly gained on some 1.2 million dwelling footprints could not have been achieved by other methods in such a short time.

**FIGURE 6** Dense building blocks (red outlines) delineated by the third Mask R-CNN model to compensate for underestimation of small dwellings in very densely built-up areas. Dwelling types are categorized according to their size (see legend).
FIGURE 7  Comparison of the extracted dense building blocks (right) with a building density layer produced by Dooley et al. (2020, left), aggregating building density to approximately 100 m grid cells.
5 | CONCLUSIONS

Our workflow for Mask R-CNN-based dwelling extraction from VHR data shows great potential in operational workflows for humanitarian response. The ultimate goal of a trained network that convincingly performs with less or no additional sampling has not fully been achieved. However, as shown in our study, interactive sampling with feedback loops for refinement improved the speed of the sampling process and helped address limitations identified already in the early stage of the workflow.

Future work is needed to elaborate on the reuse of existing models (e.g. from the DeepGlobe Building Extraction Challenge). Similar exercises can help in improving results and reducing the amount of training data needed for transferability on different images and areas.

ACKNOWLEDGEMENTS

We acknowledge the support of Vitoria Barbosa and Tanya Singh in producing the validation data set. This research was supported by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, the Christian Doppler Research Association, and MSF Austria (Ärzte ohne Grenzen Sektion Österreich).

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Dirk Tiede  https://orcid.org/0000-0002-5473-3344

ENDNOTE

1 See, for example, https://github.com/chrieke/awesome-satellite-imagery-datasets.

REFERENCES

Bjorgo, E. (2000). Using very high spatial resolution multispectral satellite sensor imagery to monitor refugee camps. *International Journal of Remote Sensing, 21*(3), 611–616. https://doi.org/10.1080/014311600210786

Checchi, F., Stewart, B. T., Palmer, J. J., & Grundy, C. (2013). Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations. *International Journal of Health Geographics, 12*, 4. https://doi.org/10.1186/1476-072X-12-4

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., ... Raskar, R. (2018). DeepGlobe 2018: A challenge to parse the Earth through satellite images. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT (pp. 172–181). Piscataway, NJ: IEEE.

Dooley, C. A., Boo, G., Leasure, D. R., & Tatem, A. J. (2020). *Gridded maps of building patterns throughout sub-Saharan Africa (Version 1.1)*. Southampton, UK: University of Southampton.

Ehrlich, D., Lang, S., Laneve, G., Mubareka, S., Schneiderbauer, S., & Tiede, D. (2009). Can Earth observation help to improve information on population? Indirect population estimations from EO derived geo-spatial data: Contribution from GMOSS. In B. Jasani, M. Pesaresi, S. Schneiderbauer, & G. Zeug (Eds.), *Remote sensing from space* (pp. 211–237). Berlin, Germany: Springer.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science, 28*(2), 125-138. https://doi.org/10.1559/152304001782173727

Ghorbanzadeh, O., Tiede, D., Dabiri, Z., Sudmanns, M., & Lang, S. (2018). Dwelling extraction in refugee camps using CNN—First experiences and lessons learnt. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*(1), 161–168. https://doi.org/10.5194/isprs-archives-XLII-1-161-2018

Ghorbanzadeh, O., Tiede, D., Wendt, L., Sudmanns, M., & Lang, S. (2020). Transferable instance segmentation of dwellings in a refugee camp—Integrating CNN and OBIA. *European Journal of Remote Sensing, 54*(Suppl. 1), 127–140. https://doi.org/10.1080/22797254.2020.1759456
Giada, S., De Groeve, T., Ehrlich, D., & Soille, P. (2003). Information extraction from very high resolution satellite imagery over Lukole refugee camp, Tanzania. *International Journal of Remote Sensing*, 24(22), 4251–4266. https://doi.org/10.1080/014311602100035021

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175

Hoeser, T., Bachofer, F., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications. *Remote Sensing*, 12(18), 3053. https://doi.org/10.3390/rs12183053

Hoeser, T., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review—Part I: Evolution and recent trends. *Remote Sensing*, 12(10), 1667. https://doi.org/10.3390/rs12101667

Kemper, T., & Heinzel, J. (2014). Mapping and monitoring of refugees and internally displaced people using EO data. In Q. Weng (Ed.), *Global urban monitoring and assessment through Earth observation* (pp. 195–216). London, UK: Taylor & Francis.

Kemper, T., Jenerowicz, M., Pesaresi, M., & Soille, P. (2011). Enumeration of dwellings in Darfur camps from GeoEye-1 satellite images using mathematical morphology. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(1), 8–15. https://doi.org/10.1109/JSTARS.2010.2053700

Knoth, C., & Pebesma, E. (2017). Detecting dwelling destruction in Darfur through object-based change analysis of very high-resolution imagery. *International Journal of Remote Sensing*, 38(1), 273–295. https://doi.org/10.1080/01431161.2015.1035412

Lang, S., Rüfer, P., Riedler, B., Wendt, L., Braun, A., Tiede, D., ... Hochschild, V. (2020). Earth observation tools and services to increase the effectiveness of humanitarian assistance. *European Journal of Remote Sensing*, 53(Suppl. 2), 67–85. https://doi.org/10.1080/22797254.2019.1684208

Lang, S., Tiede, D., Höbling, D., Füreder, P., & Zeil, P. (2010). Earth observation (EO)-based ex post assessment of internally displaced person (IDP) camp evolution and population dynamics in Zam Zam, Darfur. *International Journal of Remote Sensing*, 31(21), 5709–5731. https://doi.org/10.1080/01431161.2010.496803

Lu, Y., Koperski, K., Kwan, C., & Li, J. (2020). Deep learning for effective refugee tent extraction near Syria-Jordan border. *IEEE Geoscience and Remote Sensing Letters*, 20(6), 1–5. https://doi.org/10.1109/LGRS.2020.2999354

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Marcus, G. (2018). *Deep learning: A critical appraisal*. Preprint, arXiv:1801.00631.

Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L., & Luengo-Oroz, M. (2018). Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society A*, 376, 20170363. https://doi.org/10.1098/rsta.2017.0363

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Spröhnle, K., Tiede, D., Schoepfer, E., Füreder, P., Svanberg, A., & Rost, T. (2014). Earth observation-based dwelling detection approaches in a highly complex refugee camp environment: A comparative study. *Remote Sensing*, 6(10), 9277–9297. https://doi.org/10.3390/rs6109277

Stengel, M., Tiede, D., Lütjhe, F., Füreder, P., & Lang, S. (2014). Object-based image analysis using VHR satellite imagery for monitoring the dismantling of a refugee camp after a crisis: The case of Lukole, Tanzania. In R. Vogler, A. Car, J. Strobl, & G. Griesebner (Eds.), *Gi_Forum 2014—Geospatial innovation for society* (pp. 45–48). Vienna, Austria: Austrian Academy of Sciences Press.

Tiede, D., Füreder, P., Lang, S., Höbling, D., & Zeil, P. (2013). Automated analysis of satellite imagery to provide information products for humanitarian relief operations in refugee camps: From scientific development towards operational services. *FGF Photogrammetrie – Fernerkundung – Geoinformation*, 2013(3), 185–195. https://doi.org/10.1127/1432-8364/2013/0169

Tiede, D., Krafft, P., Füreder, P., & Lang, S. (2017). Stratified template matching to support refugee camp analysis in OBIA workflows. *Remote Sensing*, 9(4), 326. https://doi.org/10.3390/rs9040326

United Nations Department of Economic and Social Affairs. (2019). *World population prospect 2019: Data sources*. Retrieved from https://population.un.org/wpp/Download/Files/4_Metadata/WPP2019_DataSources.pdf

Wang, L., Tao, D., Wang, R., Wang, R., & Li, H. (2019). Big Map R-CNN for object detection in large-scale remote sensing images. *Mathematical Foundations of Computing*, 2(4), 299–314. https://doi.org/10.3934/mfc.2019019

Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., & Wang, P. (2019). Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. *Sensors*, 19(2), 333. https://doi.org/10.3390/s19020333

Witmer, F. D. W. (2015). Remote sensing of violent conflict: Eyes from above. *International Journal of Remote Sensing*, 36(9), 2326–2352. https://doi.org/10.1080/01431161.2015.1035412
Yuan, J., Roy Chowdhury, P. K., McKee, J., Yang, H. L., Weaver, J., & Bhaduri, B. (2018). Exploiting deep learning and volunteered geographic information for mapping buildings in Kano, Nigeria. *Scientific Data, 5*(1), 180217. https://doi.org/10.1038/sdata.2018.217

Zerboni A., Brandolini F., Mariani G. S., Perego A., Salvatori S., Usai D., Pelfini M., & Williams M. A. J. (2020). The Khartoum-Omdurman conurbation: a growing megacity at the confluence of the Blue and White Nile Rivers. *Journal of Maps*, 1–14. https://doi.org/10.1080/17445647.2020.1758810

Zhao, K., Kang, J., Jung, J., & Sohn, G. (2018). Building extraction from satellite images using mask R-CNN with building boundary regularization. In *Proceedings of the 2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT (pp. 242–246). Piscataway, NJ: IEEE.

**How to cite this article:** Tiede D, Schwendemann G, Alobaidi A, Wendt L, Lang S. Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to Covid-19 response in Khartoum, Sudan. *Transactions in GIS*. 2021;25:1213–1227. https://doi.org/10.1111/tgis.12766