Roof Segmentation Towards Digital Twin Generation in LoD2+Using Deep Learning

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-11 (2022) 173–178

Roof Segmentation Towards Digital Twin Generation in LoD2+ Using Deep Learning

N. Kolibarov * **D. Wästberg** **,‡ **V. Naserentin** ***,**** **D. Petrova-Antonova** * **S. Ilieva** * **A. Logg** ***

* *GATE Institute, Sofia University "St. Kliment Ohridski", Bulgaria*
** *Chalmers Industriteknik, Sweden*
*** *Chalmers University of Technology, Sweden*
**** *Aristotle University of Thessaloniki, Greece*
† *Fraunhoffer Chalmers Research Institute for Industrial Mathematics, Sweden*
‡ *Corresponding author: dag.wastberg@chalmersindustriteknik.se*

Abstract: There is an increasing need for digital twins of cities and their base maps, 3D city models. Creating and updating these twins is not an easy task, so automating and streamlining the process is a field of active research. A significant part of the urban geometry is residential buildings and their roofs. Modeling of roofs for urban buildings can be divided into three main areas – building detection, roof recognition and building reconstruction. The building and roofs are segmented with the help of machine learning and image processing. Afterwards the extracted information is used to generate parametric models for the roofs using methods from computational geometry. The goal is to create correct virtual models of roofs belonging to many different types of buildings. In this study, a supervised deep learning approach is proposed for the segmentation of roof edges from a single orthophoto. The predicted features include the linear elements of roofs. The experiments show that, despite the small amount of training data, even in the presence of noise, the proposed method performs well on semantic segmentation of roofs with different shapes and complexities. The quality of the extracted roof elements for the test area is about 56% and 71% for mean intersection over union (IOU) and Dice metric scores, respectively.

Keywords: Digital Twin Cities, LoD2+, Deep Learning, Convolutional Neural Networks, Roof Segmentation

1. INTRODUCTION

The need for digital models of existing physical entities has been around for decades (Liu et al., 2021) and especially in the urban context, the term *digital twin* has been getting increased research attention (Ketzler et al., 2020). The basis for such a twin of cities is a *3D city model* (Biljecki et al., 2014; Ledoux et al., 2019; Stoter et al., 2019). Creating and maintaining such a model is a tedious and time-consuming task (Stoter et al., 2020) especially in higher *Level of Detail* (LOD) (Biljecki et al., 2014), so automating the process in a general and robust way is a topic of research. Since the most represented assets of the city scale are buildings in various forms and shapes, there is a lot of focus of reconstructing the geometries from existing raw data, i.e., photogrammetry, cadastral footprints, aerial point clouds (LiDAR) (Yao et al., 2018; Logg and Naserentin, 2021, 2022). Existing building reconstruction approaches are based on the data sources they use (single, multi-sensor) or the amount of user interaction (manual, semi-automatic, automatic).

Automatic detection and reconstruction of buildings have become essential in many remote sensing and computer vision applications. The desired outcome is an automatically generated detailed 3D model of a building from aerial imagery, footprints, LiDAR, or a fusion of them. It can be applied in fields like architecture, civil engineering, urban planning, construction, real estate, GIS, and lots of others. The process is complex and has multiple steps for achieving a complete solution. The following literature review is intended to provide an overview of the needed background, the problem, the data, the techniques, and further research directions for automatic 3D building model generation. Depending on the data sources and methods used, the steps may vary. For a complete, end-to-end solution, the steps involve developing robust systems for building detection, rooftop recognition, and geometry generation.

⋆ This work is part of the Digital Twin Cities Centre (DTCC, 2022) supported by Sweden's Innovation Agency Vinnova under Grant No. 2019-421 00041, and GATE Project supported by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under Grant Agreement No. 857155, the Swedish Research Council for Sustainable Development Formas (grants 2019-01169 and 2019-01885) and by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01. P.O. Hristov was funded by the National Scientific Program "Petar Beron i NIE" under the AUDiT project, no. KP-06-DB/3.

10.1016/j.ifacol.2022.08.068

1.1 Building Detection

Most approaches for building detection generally are segmentation-based, classification-based, or hybrid methods. The system must efficiently separate buildings from non-ground objects even in the presence of noise (clouds, rain, snow) and low-resolution data. Pre-trained CNN-based on the VGG-F architecture neural network (Simonyan and Zisserman, 2014) is combined with transfer learning using aerial images of buildings, roads, and trees results in the location of buildings surrounded by bounding boxes with a quality rate of approximately 97% (Alidoost and Arefi, 2018). Before training the network, resizing, normalization, mean subtraction, and augmentation must be applied to the image tiles.

1.2 Roof Recognition

One possible desired outcome is to classify the roof into a set of pre-defined roof types. Another is to split it into planar segments belonging to a set of pre-defined roof segments. This can be achieved using various approaches.

This line-and region-based watershed segmentation method is based on a combination of edge- and region-based segmentation techniques and consists of three steps (El Merabet et al., 2015). First is the pre-processing step where the image is simplified in order to limit illumination changes. The next step has two parallel operations: the simplified image is segmented both by watershed regions and watershed lines. The last step integrates both watershed segmentation strategies into a single cooperative segmentation leading to accurate results with a quality rate of approximately 96%.

The building detection approach using CNNs (Alidoost and Arefi, 2018) is further continued where two networks are trained. The first one is used to classify buildings in the building detection step and the second network is used to classify roofs in the recognition step. The second network is fine-tuned using aerial images of flat, gable, and hip roof types. The building regions are the input of the second network which finally defines the roof shape of each building as flat, gable, or hip with a quality rate of approximately 92%.

Linear elements of individual roofs are derived from the RGB image using an optimized multi-scale convolutional–deconvolutional network (Alidoost et al., 2019). The predicted features include the normalized digital surface models (nDSMs) and linear elements of roofs in three classes of eave, ridge, and hip lines with a quality rate of approximately about 91.31% and 83.69%.

1.3 Building Reconstruction

Most approaches for 3D building reconstruction generally are data-driven or model-driven. In data-driven methods, point or image based segmentation techniques are used for extracting corresponding points of roof planes and the 3D shapes of roofs are generated by merging different planes. In the model-driven approaches, the primitives of buildings are extracted,and the most appropriate models are fitted to the buildings points. Building footprint or location, roof shape, and the height of buildings are important parameters for 3D building model reconstruction which reduces the complexity of the reconstruction procedure.

Orthogonal point cloud projections. The building orientation is derived from analysing height histogram bins. Orthogonal 2D projections of point clouds are generated using the orientation, where roof segments occur as lines of points which are extracted using a line tracking algorithm. Then the lines are extended to planes, and they are analysed for deviations from rectangular shape. To generate 3D building models, two or more neighbouring planes are grouped together (Partovi et al., 2019).

Straight skeleton computation. Straight skeleton is defined as the union of the pieces of angular bisectors traced out by polygon vertices during a continuous shrinking process in which edges of the polygon move inward, parallel to themselves at a constant speed (Aichholzer et al., 1995). Using straight skeleton computation (Sugihara, 2013), 3D building models with general shaped roofs are automatically generated. It can be applied to constructing general shaped roofs based on any simple building polygon.

Binary polygon enhancement. The roof outlines extraction approach using CNNs (Alidoost et al., 2019) is further continued where for each building, the prismatic and parametric models are reconstructed based on the estimated nDSM. The prismatic models of buildings are generated by analyzing the eave lines. The parametric models of individual roofs are reconstructed using the predicted ridge and hip lines. Next, the 3D lines of roofs are integrated into the prismatic models to generate the 3D parametric models. The experiments show that, even in the presence of noises in height values, the proposed method performs well on 3D reconstruction of buildings with different shapes and complexities. A similar approach using a Y-net is also proposed by (Alidoost et al., 2020) where the input of the network is a single RGB image, while the outputs are predicted height information of buildings as well as the rooflines in three classes of eave, ridge, and hip lines. The extracted knowledge is utilized for 3D reconstruction of buildings in LoD2 (Fig. 1).

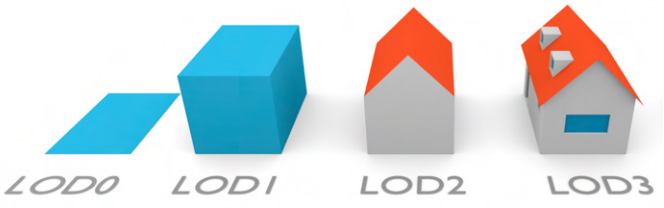

Fig. 1. Building detail at different LoD levels (Biljecki et al., 2016)

Automated detection, recognition and reconstruction of buildings from remotely sensed data are some of the most important tasks in 3D city model generation. The desired result is automatic generation of parametric LoD2+ models of buildings. To meet high application requirements like full automation, high accuracy, less computational power, and less operational time, more research in building detection, rooftop recognition and geometry generation needs to be conducted.

This study focuses on the roof recognition step using a deep learning approach with convolutional neural networks

and aims to segment roof edges from any other objects in the image. The rest of the paper is organized as follows. Section 2 describes the research method applied. Section 3 presents the obtained results. Finally, Section 4 concludes the paper and gives directions for future work.

2. RESEARCH METHOD

2.1 Training Data

The data set for generating the training data needed to contained both high resolution orthophotos and 3D roof geometry. This limits our choices, since most open data sets available are missing one or both of these. The City of Helsingborg in Sweden has an open data web portal (Helsinborg, 2022) with data that fulfilled both of the requirements, as they had both high resolution orthophotos (8cm pr pixel) and LoD2 3D building models in DWG format (see Fig. 2). The training data used was generated from open data over the areas of Eneborg, Fältbacken and Husensjö.

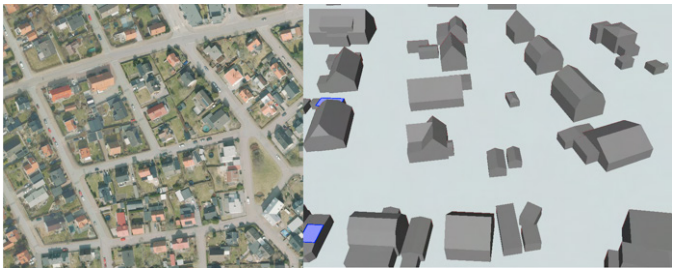

Fig. 2. Orthophoto and 3D model from Helsingborg

Upon the examining the data, it became clear that the outlines of the 3D roofs didn't always line up with the image of the corresponding roof in the orthophotos. To correct for this, 2D footprints of the 3D buildings were created and, together with the orthophotos, loaded into QGIS. Here each polygon was manually offset to more accurately line up with the roof images in the orthophotos. In addition roof polygons which didn't have a corresponding building in the orthophoto or where the building in the images clearly didn't match, were removed. This happened in situations where the building had been either demolished or extended in the time between the 3D model being made and the orthophoto being taken.

Once this process had been done, the individual offset applied to each footprint in QGIS was calculated and the same XY offset was applied to the 3D model of the corresponding building. Using FME the 3D roof surfaces where converted to vector lines that where projected to 2D and rasterized (background was black and the lines white) to a georeferenced raster with the same pixel size and coordinate system as the orthophotos, creating a segmentation mask for the roof.

The orthophotos and segmentation masks where then tiled into 512x512 pixel images and only image tiles that showed part of a building were included. The final dataset was made up of 600 3x512x512 RGB images with their corresponding 512x512 pixel gray scale segmentation masks; see Fig. 3.

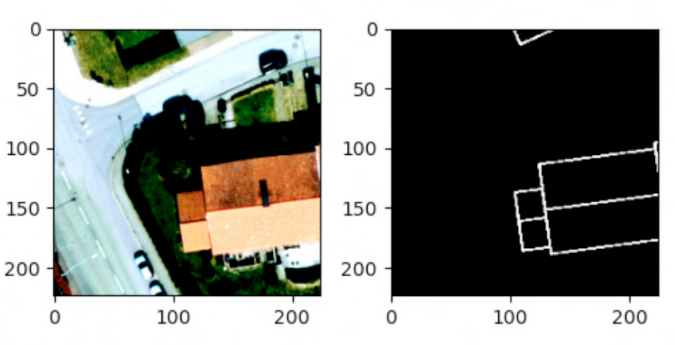

Fig. 3. RGB image with corresponding segmentation mask (Resized to 224 x 224).

Later in the project we also tested using different thicknesses of the rasterized lines in segmentation map (Fig. 4), and found that buffering the lines to a thickness of 20 cm prior to rasterizing showed some promising improvements in the final model. Going forwards this approach will be explored more.

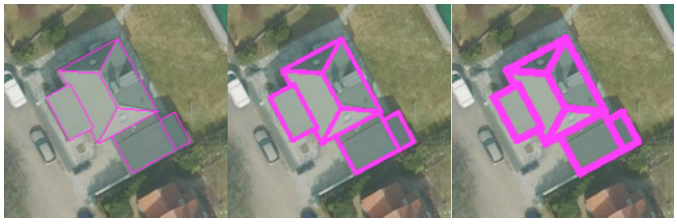

Fig. 4. Different rasterized line thickness

2.2 Dataset Problems

The dataset included images with buildings without any annotations for them. Furthermore, it contained images that were only partially correctly annotated. This happened in cases where we didn't have the 3D model for the corresponding building in the image. There were also some labeling differences - some of the samples include roof window annotations, while others don't. Since the training data didn't include sufficient data without any buildings and only buildings where annotated the classifier would often misclassify non-building structures as roofs. Highways for example can be inaccurately segmented as buildings with flat roofs. The observed issues result in less accurate predictions and need to be addressed in the future to improve model performance.

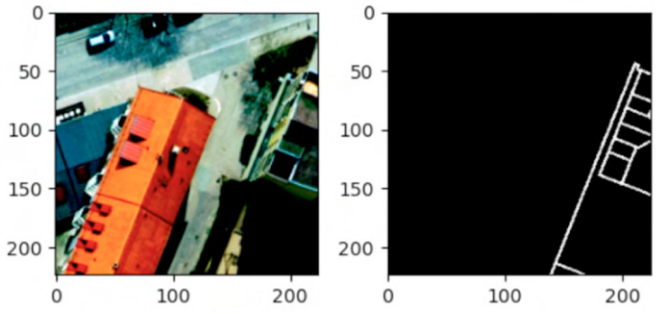

Fig. 5. Partially annotated training image.

2.3 Data Preprocessing

Data augmentation is a technique that can be used to artificially expand the size of a training dataset by creating modified versions of samples in the dataset. More data can result in better inference accuracy and the augmented variations of the samples can improve the generalization of the model.

The following data augmentation techniques are used to artificially enlarge the dataset:

- All samples are horizontally flipped.
- All samples are vertically flipped.
- All samples are cropped in the center.

2.4 Model Development

To separate buildings from other objects and extract roof edges information, a convolutional neural network for semantic segmentation is trained. U-Net was originally invented and first used for biomedical image segmentation (Ronneberger et al., 2015). The approach follows a popular encoder-decoder structure (Fig. 6). The encoder downsamples the spatial resolution of the input, developing lower-resolution feature mappings and then the encoder upsamples the feature representations into a full-resolution segmentation map. For semantic segmentation of roof edges, the used model architecture was a ResNet18 encoder (He et al., 2015) and a U-Net decoder.

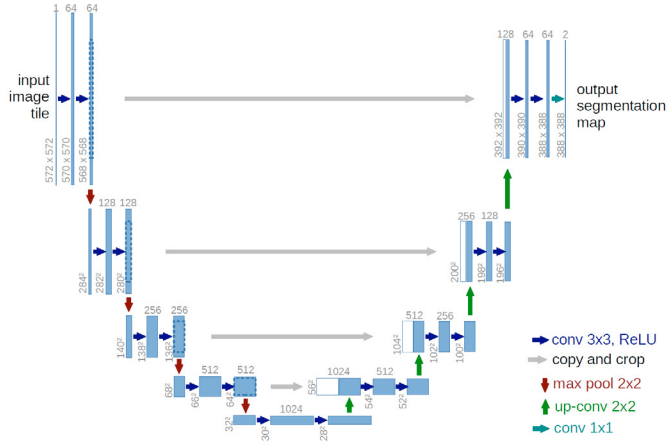

Fig. 6. U-Net architecture.

Transfer learning is used to learn features from a separate problem and then adapt the model to the roof segmentation problem. It has the benefit of decreasing the training time and can result in lower generalization error. The encoder is pre-trained on a large-scale hierarchical image database - ImageNet. The model weights are publicly available and easily accessible on PyTorch Hub. It was trained on more than 1,000,000 images for 1000 categories and it has obtained knowledge on how to detect generic features. Such pre-trained models offer state of the art performance and remain effective on the initial image recognition task.

To learn parameters for roof segmentation, the proposed network weights are initialized uniformly and trained on the dataset mentioned in the previous sub-section. For the loss function, a weighted combination of Binary Cross-Entropy (BCE) with Dice coefficient is applied. The Binary Cross-Entropy loss, shown in Eq. (1) examines each pixel individually and then averages over all pixels, asserting equal learning to each pixel, where $\hat{y}$ is the predicted class probability and y is the reference class probability.

$$L_{bce}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (1)$$

The Sørensen–Dice coefficient, shown in Eq. (2) is a statistical tool which measures the similarity between two sets of data. It measures the overlap between the predicted mask and ground truth mask. This measure ranges from 0 to 1 where a coefficient of 1 denotes complete overlap. To avoid division by zero and reduce overfitting, a smoothing technique is used by adding a variable close or equals to 1 denoted as ϵ to the numerator and denominator.

$$L_{dice}(y, \hat{y}) = \frac{2 \sum y\hat{y} + \epsilon}{\sum y^2 + \sum \hat{y}^2 + \epsilon} \quad (2)$$

We have an unbalanced class representation with the prevalent class being non-building and using BCE loss on its own is not suitable. Combining the two loss functions and adding a weight β as shown in Eq. (3) allows for some diversity in the loss, while benefiting from the stability of binary cross-entropy.

$$L = \beta L_{bce} + L_{dice}(1 - \beta) \quad (3)$$

Mini-batch stochastic gradient descent with momentum and Adam optimizer are employed as training optimizers for roof edge segmentation.

After stitching the predicted tiles together, the image undergoes several image processing techniques which help to increase the possibility of obtaining the correct corner points and geometry that allows building polygon formation. Small pixel islands that are redundant are removed by filtering using contour area then applying morpholgical closing to fill the small holes in the image.

3. RESULTS

3.1 Roof edge segmentation network

The roof edge segmentation network was trained using Google Colab on a single NVIDIA T4 with a batch size of 16 for 100 epochs. Mini-batch gradient descent was used with a combination of binary cross-entropy and Dice coefficient for loss function and Adam algorithm with learning rate, beta 1, beta 2, and epsilon parameters selected as 0.001, 0.9, 0.999 and 1e-8.

3.2 Performance evaluation

To evaluate the performance of the trained network, a test area was selected outside the training area composed of different shapes and types of buildings. Fältbacken and Husensjö were used for training and Eneborg was used for testing. The number of training samples was increased to about 2400 image tiles after data augmentation. The input size of the proposed network was 224 x 224, while

the size of the test area was about 7835 x 7680. If the entire test area was resized to the needed input size of the network, then the accuracy would be significantly degraded. Therefore, the test area was divided into smaller tiles (Fig. 7) and the predicted tiles were stitched together afterwards (Fig. 8). The predicted roof elements include valuable knowledge of building and roof boundaries.

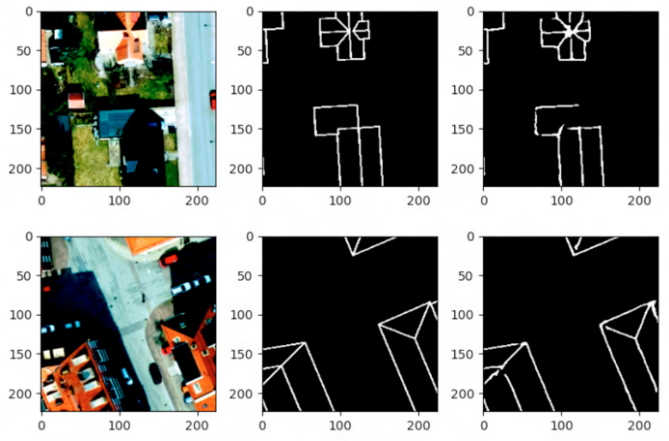

Fig. 7. Tiled image, ground truth mask, predicted mask.

Fig. 8. Stitched predicted mask.

The accuracy of the estimated roof edges is evaluated based on standard metrics such as pixel accuracy, Dice and IOU scores. The pixel accuracy for the test area is about 97%. However, this metric provides misleading results since the roof edge class representation is small within the image. Thus, the measure will be biased towards the negative class. In such cases it is more suitable to use other metrics for model evaluation. The IOU metric measures the number of pixels common between the target and prediction masks divided by the total number of pixels present across both masks. The quality of the extracted roof elements for the test area is about 56% and 71% for mean IOU and Dice metric scores, respectively.

Another experiment was to access the performance of the model on unseen data from another source. The data used was acquired from Lantmäteriet (Swedish Mapping and Land Registration Authority) and consisted of orthophotos over Hammarkullen in Gothenburg, Sweden. These images were at a lower resolution (25 cm per pixel) than our training data. To compensate this, the images were upsampled to match the resolution of the training data prior to running them through the model. Since there is no labeled data available it is not possible to accurately measure the overall performance of the model. Insights on the model performance are gained through visual inspection (Fig 9). Given the small amount of training samples and the different data distribution, the model seems to be able to recognize roofs from non-roof objects in the image.

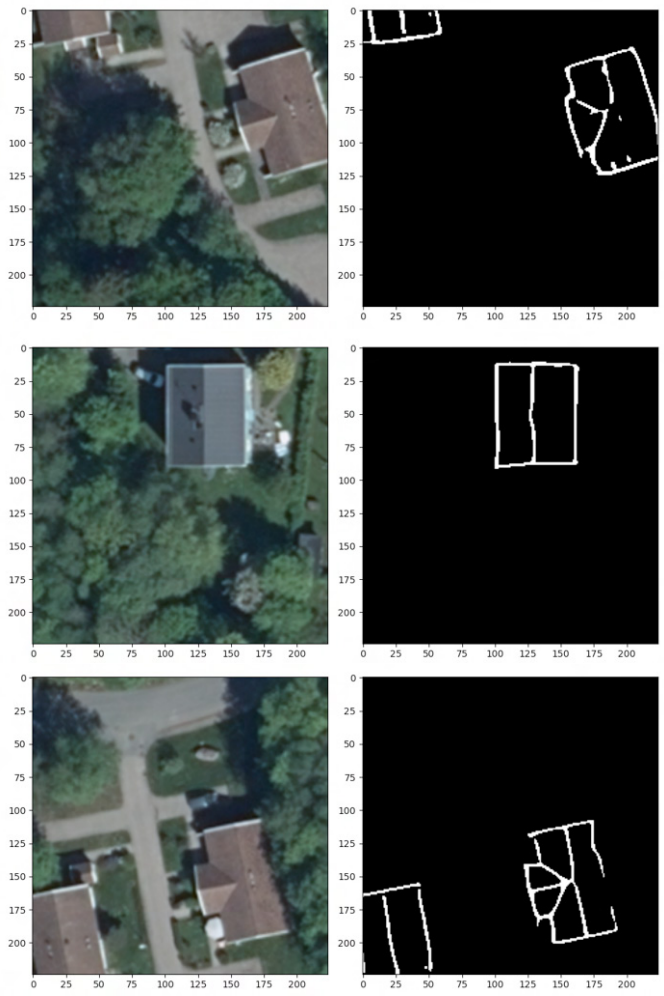

Fig. 9. Tiled image and predicted mask.

4. CONCLUSIONS

In this study, we presented an approach based on supervised deep learning techniques to extract roof edges of buildings from a single orthophoto towards generating parametric LoD2+ models. Although there were limitations for the dataset in terms of size and noise, a neural network was trained for roof edge binary semantic segmentation. The results over the test area showed the reasonable

performance and promising possibilities of the proposed method, despite the issues of the provided dataset. Applying transfer learning using an encoder pre-trained on orthophotos of buildings instead of photos of common objects, should also increase the accuracy and decrease training time significantly. Expanding the dataset by acquiring samples from different sources including variation in cities, illumination, weather conditions, as well as cleaning it from inaccurate labels, needs to be done in future studies to improve the generalization and transferability of the trained network.

This project is part of the larger Digital Twin Cities Centre (https://dtcc.chalmers.se) and all code will be made available at https://gitlab.com/dtcc-platform

5. ACKNOWLEDGEMENT

This research was funded by the H2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155 and by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01.

REFERENCES

Aichholzer, O., Aurenhammer, F., Alberts, D., and Gärtner, B. (1995). A novel type of skeleton for polygons. *J. Univers. Comput. Sci.*, 1, 752–761.

Alidoost, F. and Arefi, H. (2018). A cnn-based approach for automatic building detection and recognition of roof types using a single aerial image. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 86, 235–248.

Alidoost, F., Arefi, H., and Hahn, M. (2020). Y-shaped convolutional neural network for 3d roof elements extraction to reconstruct building models from a single aerial image. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 321–328.

Alidoost, F., Arefi, H., and Tombari, F. (2019). 2d image-to-3d model: Knowledge-based 3d building reconstruction (3dbr) using single aerial images and convolutional neural networks (cnns). *Remote Sensing*, 11(19). doi:10.3390/rs11192219. URL `https://www.mdpi.com/2072-4292/11/19/2219`.

Biljecki, F., Ledoux, H., and Stoter, J. (2016). An improved lod specification for 3d building models. *Computers, Environment and Urban Systems*, 59, 25–37. doi: https://doi.org/10.1016/j.compenvurbsys.2016.04.005.

Biljecki, F., Ledoux, H., Stoter, J., and Zhao, J. (2014). Formalisation of the level of detail in 3D city modelling. *Computers, Environment and Urban Systems*, 48, 1–15.

DTCC (2022). Digital Twin Cities Centre — Homepage. URL `https://dtcc.chalmers.se`.

El Merabet, Y., Meurie, C., Ruichek, Y., Sbihi, A., and Touahni, R. (2015). Building roof segmentation from aerial images using a lineand region-based watershed segmentation technique. *Sensors*, 15(2), 3172–3203. doi:10.3390/s150203172. URL `https://www.mdpi.com/1424-8220/15/2/3172`.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385. URL `http://arxiv.org/abs/1512.03385`.

Helsinborg (2022). Helsingborg.io – Öppen data portal. URL `https://helsingborg.io`.

Ketzler, B., Naserentin, V., Latino, F., Zangelidis, C., Thuvander, L., and Logg, A. (2020). Digital Twins for Cities: A State of the Art Review. *Built Environment*, 46(4), 547–573. doi:10.2148/benv.46.4.547.

Ledoux, H., Arroyo Ohori, K., Kumar, K., Dukai, B., Labetski, A., and Vitalis, S. (2019). CityJSON: a compact and easy-to-use encoding of the CityGML data model. *Open Geospatial Data, Software and Standards*. doi:10.1186/s40965-019-0064-0.

Liu, M., Fang, S., Dong, H., and Xu, C. (2021). Review of digital twin about concepts, technologies, and industrial applications. *Journal of Manufacturing Systems*, 58, 346–361. doi:10.1016/J.JMSY.2020.06.017.

Logg, A. and Naserentin, V. (2021). Digital Twin Cities Platform — Builder. URL `https://gitlab.com/dtcc-platform/dtcc-builder`.

Logg, A. and Naserentin, V. (2022). Digital twins for cities: Automatic, efficient, and robust mesh generation for large-scale city modeling and simulation. Submitted in *Nature Computational Science*.

Partovi, T., Fraundorfer, F., Bahmanyar, R., Huang, H., and Reinartz, P. (2019). Automatic 3-d building model reconstruction from very high resolution stereo satellite imagery. *Remote Sensing*, 11(14). doi:10.3390/rs11141660. URL `https://www.mdpi.com/2072-4292/11/14/1660`.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556. URL `http://arxiv.org/abs/1409.1556`.

Stoter, J., Ho, S., and Biljecki, F. (2019). CONSIDERATIONS FOR A CONTEMPORARY 3D CADASTRE FOR OUR TIMES. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W15, 81–88. doi: 10.5194/isprs-archives-XLII-4-W15-81-2019.

Stoter, J., Ohori, A., Dukai, B., Labetski, A., , Kumar, K., Vitalis, S., and Ledoux, H. (2020). State of the art in 3d city modelling. URL `https://www.gim-international.com/content/article/state-of-the-art-in-3d-city-modelling-2`.

Sugihara, K. (2013). Straight skeleton for automatic generation of 3-d building models with general shaped roofs.

Yao, Z., Nagel, C., Kunde, F., Hudra, G., Willkomm, P., Donaubauer, A., Adolphi, T., and Kolbe, T.H. (2018). 3DCityDB - a 3D geodatabase solution for the management, analysis, and visualization of semantic 3D city models based on CityGML. *Open Geospatial Data, Software and Standards*. doi:10.1186/s40965-018-0046-7.