The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-6-2025
ISPRS, EARSeL & DGPF Joint Istanbul Workshop “Topographic Mapping from Space” dedicated to Dr. Karsten Jacobsen’s 80th Birthday
29–31 January 2025, Istanbul, Türkiye

Building Segmentation and Modelling from Space-Borne and Aerial Imagery

Thomas Krauß, Ksenia Bittner, Pablo d’Angelo, Philipp Schuegraf, Peter Reinartz, Rupert Müller

DLR, German Aerospace Center, 82234 Oberpfaffenhofen, Germany – thomas.krauss@dlr.de

Keywords: DSM generation, True-Ortho-Mosaic, Building segmentation, Building modeling

Abstract

Accurate 3D building reconstruction is essential for urban planning, disaster management, and environmental applications. However, current methods often struggle to achieve geometric precision and topological consistency, particularly when when processing satellite or aerial imagery. This paper presents a comprehensive workflow that addresses these challenges, enabling the generation of multiple outputs—including digital surface models (DSMs), digital terrain models (DTMs), true-orthophotos, 2D building segments, and vectorized 3D LoD-2 building models. Our approach leverages very high-resolution (VHR) imagery to derive precise DSM and DTM data, which are used in conjunction with orthorectified imagery to accurately segment buildings and delineate roof planes. By focusing on planar building components and employing robust vectorization techniques, our workflow ensures consistent 3D model construction while avoiding the challenges of fine-detail extraction.Validated on diverse urban datasets, our method demonstrates high accuracy, scalability, and potential to advance building reconstruction workflows in remote sensing, contributing significantly to geospatial and environmental research.

1. Introduction

Building segmentation and modeling are critical for urban planning, real estate management, population estimation, disaster response, and environmental monitoring. These tasks rely on the accurate extraction of building information, which acts as the foundation for critical applications such as city modeling, infrastructure development, and emergency response planning. The advent of high-resolution remote sensing data from space-borne platforms like Pléiades, GeoEye-1, and WorldView, as well as aircrafts and helicopters, has significantly enhanced the precision and scalability of these processes.

Despite these advancements, building information extraction from remote sensing imagery presents numerous challenges. Variations in building size, shape, height, and function, coupled with occlusions and shadow effects in complex urban environments, complicate accurate information extraction. Traditional methods for building information extraction have relied on manual interpretations or semi-automated low-level image processing techniques, providing initial solutions. These include threshold-based approaches (Chen and Chen, 2009), region-based (Karthick et al., 2014), edge-based (Canny, 1986; Chen et al., 1987; Kanopoulos et al., 1988), classification-based methods using feature extraction algorithms such as SIFT (Lowe, 1999), SURF (Bay, 2006), and HOG (Dalal and Triggs, 2005). While effective in specific contexts, these methods are time-consuming, difficult to scale for large datasets and limited in generalizability.

In recent years, the emergence of deep learning has revolutionized the field of building information extraction. Convolutional neural networks (CNNs) have set new standards by enabling end-to-end semantic segmentation, which bypasses the need for manual feature extraction. Multiple different neural networks based on the fully convolutional network (FCN) (Long et al., 2015), U-Net (Ronneberger et al., 2015), ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) architectures, as examples, demonstrated state-of-the-art performance in building segmentation (Bittner et al., 2018; Schuegraf and Bittner, 2019; Khan et al., 2020), leveraging high-resolution spatial features to achieve pixel-wise classification with unprecedented accuracy.

Figure 1. 3D building model of the first (aerial imagery) test area in Braunschweig, Tostmannplatz (Germany).

While semantic segmentation assigns class labels to each pixel, instance segmentation goes further by distinguishing individual building structures, even in densely built areas. Mask R-CNN (He et al., 2017) has set the standard for instance segmentation, enabling precise delineation of individual buildings in remote sensing imagery. Building on this, TernausNetV2 (Iglovikov et al., 2018) proposed an encoder-decoder architecture with skip connections, incorporating key modifications to support both semantic and instance segmentation tasks. Building on the foundational work of (Iglovikov et al., 2018), Schuegraf et al. (2022) extended the concept of building instance segmentation to focus on building section instance segmentation, particularly addressing the challenges of identifying individual roof tiles with varying heights or forms in complex structures. This approach demonstrated that FCNs, such as SkipFuse-U-DenseNet12, could effectively integrate RGB and digital surface model (DSM) images to predict a three-class map comprising background, building regions, and rooftop touching borders. To refine these predictions, post-processing techniques like the watershed algorithm and morphological operations were employed, bridging gaps between instances generated by neural networks and improving segmentation accuracy. Expanding this line of inquiry, Schuegraf et al. (2024b) explored the application of building section instance segmentation for identifying individual

buildings in both formal and informal settlements. Utilizing a SkipFuse-UResNet34 model, their approach produced more comprehensive building masks for Medellín, Colombia, outperforming conventional official data sources. Girard et al. (2021) innovatively added a frame field learning (FFL) to the deep-segmentation model to generate a vector field that encodes useful boundary information alongside the corresponding segmentation mask. Zorzi and Fraundorfer (2023) propose Re:PolyWorld that leverages both vertex features and the visual appearance of edges. The edge-aware graph neural network (GNN) efficiently predicts connections between vertex pairs forming rectangular shapes for every building section instance.

In this paper, we present a workflow for extracting vectorized models of building segments from aerial or space-borne imagery. Our approach leverages stereo or multi-stereo imagery acquired from very high-resolution (VHR) sensors with a ground sampling distance (GSD) of 20 to 30 centimeters. From this data, a dense DSM is derived, along with a digital terrain model (DTM) that represents the ground surface exclusively. Using the DSM, we generate a true-ortho-mosaic, which serves as a critical input for further processing.

Unlike methods that rely on precise delineation of minute structural lines—which can be particularly challenging in satellite imagery—our workflow focuses on spatial embeddings to detect building segments and roof planes as primary components for level of detail (LoD)-2 building models (Schuegraf et al., 2024a). These planar elements are robustly extracted and serve as the foundation for generating 3D building models. Furthermore, our workflow produces multiple valuable outputs, including DSMs, DTMs, and true-ortho-mosaics, which have wide-ranging applications in remote sensing, urban planning, and environmental monitoring.

2. Methodology

In the first step of our workflow, a dense DSM is generated using the semi-global matching (SGM) method developed at the German Aerospace Center (DLR) by Krauß et al. (2013). The process begins with multi- or stereo satellite or aerial images, where bundle block adjustment is applied to refine the rational polynomial coefficients (RPC), enhancing the sensor model and enabling accurate stereo matching. Tie points across all input images are detected using the scale-invariant feature transform (SIFT) algorithm and refined to sub-pixel precision with local least squares matching. Stereo pairs are processed with SGM, and the resulting disparity maps are merged through a robust integration method that incorporates area-based outlier filtering to reject mismatches in occluded or unmatchable regions, such as water bodies or cloud cover (d’Angelo and Reinartz, 2011). Residual occluded areas are interpolated using nearby ground height values, while larger voids caused by extensive cloud or water coverage are filled with the Copernicus digital elevation model (DEM) to ensure spatial completeness in the final DSM. Fig. 2 shows the DSM derived from aerial imagery for the test-area Braunschweig Tostmannplatz.

From the filled DSM, the DTM representing the bare ground surface is extracted using a modified morphological filtering approach, as outlined in Krauß et al. (2011). The normalized digital elevation model (nDEM), which captures the height of above-ground objects, is then obtained by subtracting the derived DTM from the DSM. This step provides a detailed representation

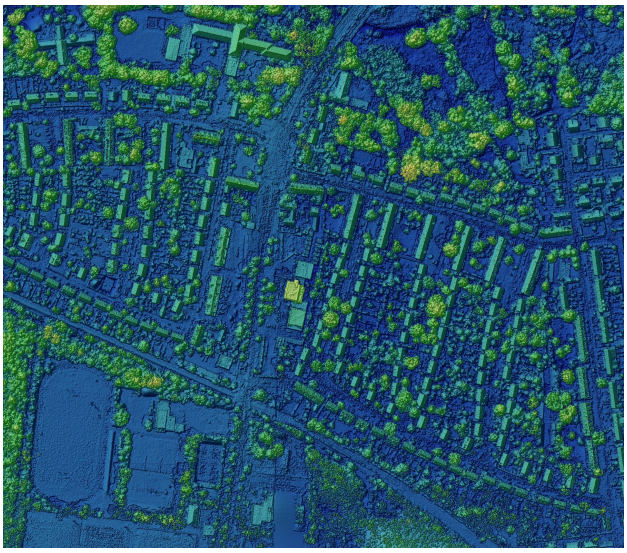

Figure 2. Derived DSM for test-area Braunschweig Tostmannplatz, 720×630 m^2.

of surface features such as vegetation canopy heights or building structures, enabling further analyses of above-ground elements.

A true ortho mosaic is generated from the DSM using all available images of the area. Each image is orthorectified by applying the sensor model (RPCs) to project every pixel onto the DSM. For elevated objects, such as buildings, multiple intersections typically occur—one on the object’s surface (e.g., the roof) and another on the ground behind it. To address this, a visibility map is computed to identify and exclude occluded regions in the ortho image, such as projections onto hidden areas like roads obscured by buildings. When multiple images are available, a median merge is applied to combine them, leveraging the overlapping data to fill occluded areas with content from other ortho images. This approach also removes moving objects, provided there are sufficient overlapping images—commonly the case with aerial datasets—ensuring a seamless and artifact-free mosaic. As an example the true-ortho-mosaic for the test area Braunschweig is shown in fig. 3.

To segment buildings and identify their corresponding roof planes, we employ advanced deep learning techniques that have demonstrated state-of-the-art performance across a variety of tasks. Specifically, we adopt an approach capable of simultaneously predicting multiple adjacent instances. One such method is described by Neven et al. (2019), which predicts instance-specific vectors pointing to the centers of objects, along with seed points indicating instance centers and additional shape parameters. These outputs are processed in a shape-conditioned clustering step to generate distinct instances of building sections and roofs.

Fig. 4 shows the workflow of the implented method following Schuegraf et al. (2024a) for deriving building segments from a provided DSM and a fitting panchromatic image.

The detailled workflow of the preparation of building footprints and seeds for the segmentation from the results of the FCN is shown in Fig. 5.

Using these footprints, the segmentation map and the seed as inputs for a watershed transformation gives the final building segmentation as shown in fig. 6.

Figure 3. True-ortho-mosaic for test-area Braunschweig Tostmannplatz, 720 × 630 m^2.

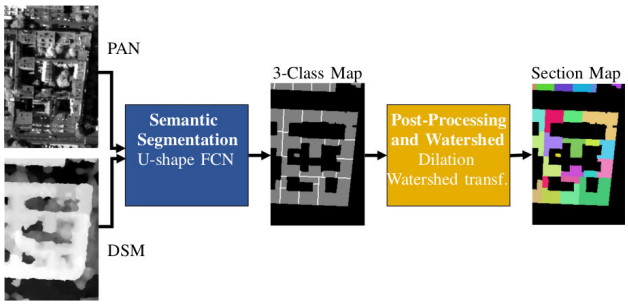

Figure 4. Overview of the implemented method deriving building segments from the DSM and the pan-image.

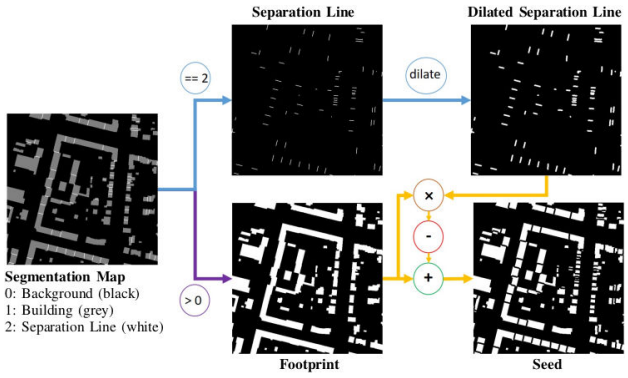

Figure 5. Detail of the workflow used for splitting building segments.

Building on this, we follow the method described by Schuegraf et al. (2023) to vectorize the obtained instances. Here, the instance masks are transformed into polygons by tracing the edges of each instance at the pixel level. These polygons serve as the foundation for further analysis and processing of the building roof planes.

In the final step, the instance polygons are simplified and rectified using information from the true ortho-mosaic and the nDEM. This refinement ensures geometric accuracy and alignment with the underlying data. Using the finalized simplified polygons and height values extracted from the DSM, a detailed three-dimensional vector model is generated for each building segment. This process produces accurate and structured 3D representations suitable for further analysis or integration into geospatial applications.

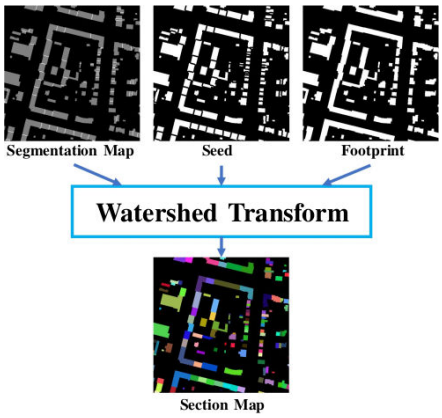

Figure 6. Final splitting of the footprints and the segmentation map to building segments using the watershed transformation.

3. Experiments

3.1 Data

During the training and validation phases, we utilize a WorldView 1 panchromatic image and a photogrammetric DSM of Berlin, Germany, with dimensions of 30,733 × 45,999 pixels. Publicly available data from the Berlin Senate[1] serves as the ground truth for building sections, roof planes, and building heights.

Figure 7. DSM of a 700 m × 500 m section from the second test area (from WorldView-3 satellite) in Lyon (France).

We use two separate datasets for evaluation, from World View 3 depicting Lyon city, France with a GSD of 0.3 m and 3K aerial data of Braunschweig city, Germany of a resolution of 0.2 m. For metric computation, we use public ground truth of both Lyon[2] and Braunschweig[3] in vector format.

During training, we divide the data into non-overlapping patches, each measuring 256 × 256 pixels. To enhance data diversity,

[1] https://daten.berlin.de/tags/geodaten
[2] https://data.grandlyon.com/
[3] https://www.lgln.niedersachsen.de/

Figure 8. True-Ortho-Image of the test area in Lyon (France).

random shifts of up to 256 pixels are applied horizontally and vertically. For validation, patches of the same size are extracted without overlap. During testing, 256 × 256 pixel patches are generated with an overlap of 128 pixels in both horizontal and vertical directions. Predictions are made per patch, and a complete map is constructed by averaging overlapping areas to ensure smooth transitions between patches.

3.2 Implementation Details

The network parameters are initialized randomly, and the model is trained using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.0002 and momentum terms of 0.5 and 0.999. Training is performed over 300 epochs, with the learning rate reduced by a factor of 0.1 after the 100th and 200th epochs. A batch size is set to 8.

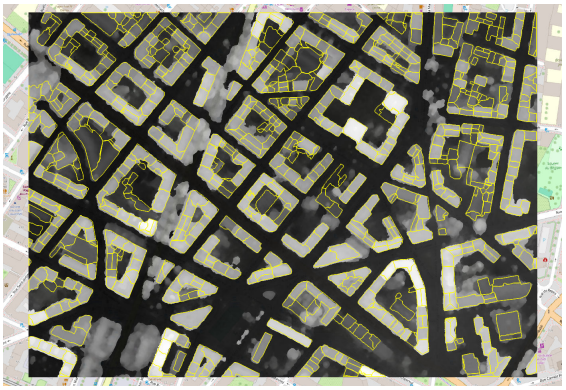

Figure 9. Extracted building segments of the test area in Lyon (France).

4. Results and Discussion

We evaluate the model performance by assessing the vectorized roof planes in 2D using the intersection over union (IoU) metric as described by Schuegraf et al. (2024a). For each ground truth polygon, the corresponding predicted polygon with the highest IoU is selected, and the average IoU across all polygons is calculated. Additionally, the accuracy of the rasterized predicted LoD-2 model is assessed in 3D using root mean square error (RMSE) and the median absolute deviation (MAD), which are derived from per-pixel differences between the predicted model and the ground truth. The results are summarized in table 1.

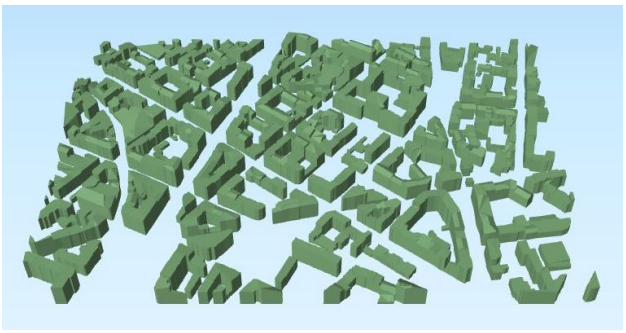

Figure 10. 3D building model of the test area in Lyon (France).

Table 1. Quantitative results for two test areas. The values for Lyon city are taken from Schuegraf et al. (2024a). ↑ indicates that higher values are superior, ↓ indicates that lower values correspond to higher accuracy.

Test Area	IoU^{gt}_{inst} ↑	MAD [m] ↓	RMSE [m] ↓
Lyon	0.769	1.74 m	4.98 m
Braunschweig	0.762	0.31 m	1.64 m

Figs. 11 and 12 shows visually the results of the correctness described by the IoU where white areas correspond to correctly detected areas whereas green areas are false positives, i.e. detected areas which are not existing in reference and red areas as false negative, i.e. existing in referece but missing in results.

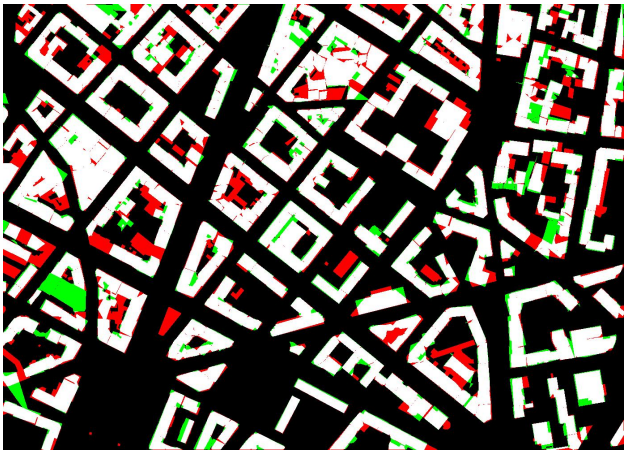

Figure 11. Results of test-area Lyon compared to OSM building mask, 700 × 500 m^2 section of Lyon (see fig. 10).

5. Conclusions and Outlook

In the presented work we describe a novel method to derive three-dimensional models of building segments directly from a few airborne or space-borne very high resolution (multi-)stereo images. After pre-processing and generation of a DSM, a DTM and the true-ortho-mosaic the extraction of building segment instances is performed using a deep-learning approach. Finally, the results are simplified and a 3D model of these segments is derived. Since the method is based only on at least two stereo images it allows the modelling of urban areas from any place on earth.

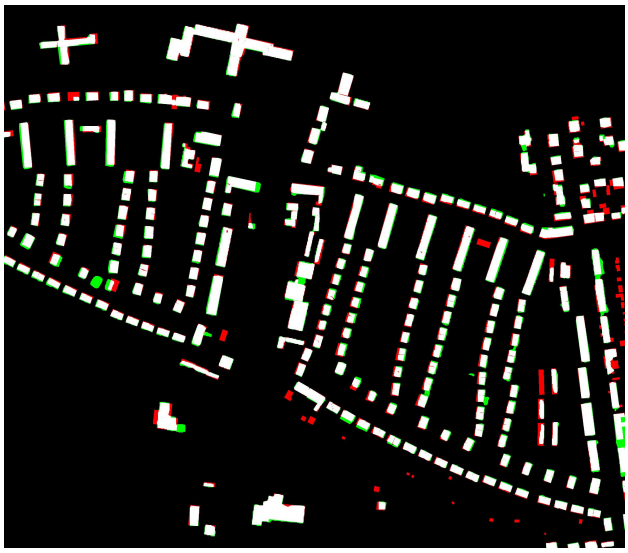

Figure 12. Results of test-area Braunschweig compared to OSM building mask, 720 × 630 m^2 section of Braunschweig (see fig. 1).

References

Bay, H., 2006. Surf: Speeded up robust features. *Computer Vision—ECCV*.

Bittner, K., Adam, F., Cui, S., Körner, M., Reinartz, P., 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. *IEEE J. of Select. Topics in Appl. Earth Observ.s and Remote Sens.*, 11(8), 2615–2629.

Canny, J., 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 679–698.

Chen, J.-S., Huertas, A., Medioni, G., 1987. Fast convolution with Laplacian-of-Gaussian masks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 584–590.

Chen, Y. B., Chen, O. T., 2009. Image segmentation method using thresholds automatically determined from picture contents. *Eurasip journal on image and video processing*, 2009, 1–15.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, 1, Ieee, 886–893.

d’Angelo, P., Reinartz, P., 2011. Semiglobal Matching Results on the ISPRS Stereo Matching Benchmark. *International Archives of Photogrammetry and Remote Sensing*, XXXVIII-4/W19, 79–84.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5891–5900.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. *CVPR*, 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

Iglovikov, V., Seferbekov, S., Buslaev, A., Shvets, A., 2018. Ternausnetv2: Fully convolutional network for instance segmentation. *CVPRW*, 233–237.

Kanopoulos, N., Vasanthavada, N., Baker, R. L., 1988. Design of an image edge detection filter using the Sobel operator. *IEEE Journal of solid-state circuits*, 23(2), 358–367.

Karthick, S., Sathiyasekar, K., Puraneeswari, A., 2014. A survey based on region based segmentation. *International Journal of Engineering Trends and Technology*, 7(3), 143–147.

Khan, A., Sohail, A., Zahoora, U., Qureshi, A., 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intell. Rev.*, 1 - 62.

Kingma, D., Ba, J., 2017. Adam: A Method for Stochastic Optimization.

Krauß, T., Arefi, H., Reinartz, P., 2011. Evaluation of selected methods for extracting digital terrain models from satellite born digital surface models in urban areas.

Krauß, T., d’Angelo, P., Schneider, M., Gstaiger, V., 2013. The fully automatic optical Processing System CATENA at DLR. *ISPRS Journal of Photogrammetry and Remote Sensing*, 40-W1, 177–181.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *CVPR*, 3431–3440.

Lowe, D. G., 1999. Object recognition from local scale-invariant features. *Proceedings of the seventh IEEE international conference on computer vision*, 2, Ieee, 1150–1157.

Neven, D., Brabandere, B. D., Proesmans, M., Van Gool, L., 2019. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8829–8837.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Schuegraf, P., Bittner, K., 2019. Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN. *ISPRS Int. J. of Geo-Inform.*, 8(4).

Schuegraf, P., Gui, S., Qin, R., Fraundorfer, F., Bittner, K., 2024a. Sat2building: Lod-2 Building Reconstruction from Satellite Imagery using Spatial Embeddings. *Submitted to ISPRS Journal of Photogrammetry and Remote Sensing*. to be appeared.

Schuegraf, P., Schnell, J., Henry, C., Bittner, K., 2022. Building section instance segmentation with combined classical and deep learning methods. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 407–414.

Schuegraf, P., Stiller, D., Tian, J., Stark, T., Wurm, M., Taubenböck, H., Bittner, K., 2024b. Ai-based building instance segmentation in formal and informal settlements. *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 1558–1561.

Schuegraf, P., Zorzi, S., Fraundorfer, F., Bittner, K., 2023. Deep Learning for the Automatic Division of Building Constructions Into Sections on Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 7186–7200.

Zorzi, S., Fraundorfer, F., 2023. Re:polyworld - a graph neural network for polygonal scene parsing. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16762–16771.