ISPRS Journal of Photogrammetry and Remote Sensing 211 (2024) 425–437

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

PLANES4LOD2: Reconstruction of LoD-2 building models using a depth attention-based fully convolutional neural network

Philipp Schuegraf [a,*], Jie Shan [b], Ksenia Bittner [a]

[a] German Aerospace Center, Münchener Straße 20, Weßling, 82234, Bavaria, Germany
[b] School of Civil Engineering, Purdue University, 550 Stadium Mall, West Lafayette, IN, 47907, USA

ARTICLE INFO

MSC:
0000
1111

Keywords:
Building reconstruction
Images
Digital surface model
Instance segmentation
Depth attention module

ABSTRACT

Level of detail (LoD)-2 reconstruction is an inevitable task in digital twin-related applications such as disaster management, flood simulation, landslide simulation and solar panel recommendation. However, there is a lack of capable methods that can exploit fine details in RGB imagery and mitigate noise in photogrammetric digital surface models (DSMs). Our investigation is focused on the use of roof planes to achieve a geometrically complete and correct, and topologically consistent LoD-2 building reconstruction. Using UNet with the EfficientNet-B3 backbone, the developed approach starts with jointly predicting building sections and roof planes from the orthorectified RGB imagery and a photogrammetric DSM. The detected sections and planes are then vectorized by employing tree search and simplified with the Douglas Peucker algorithm. Subsequently, height values from the noisy input DSM and the vectorized image-based (and simplified) roof planes are used to derive 3D-planes. Finally, the building model is formed by computing plane intersections as the ridge lines. This study demonstrates that a well-designed depth attention module (DAM), which is the bottleneck of the UNet, can achieve a very good use of both spectral and depth features. The resultant 1-to-n correspondence between building section and roof plane benefits accurate and consistent building model reconstruction. Furthermore, it leads to a superior generalization capability of the proposed method. Experiments with 1437 buildings from the cities Cologne and Braunschweig, Germany, demonstrate the success of the proposed workflow in reconstructing compound buildings with complex roof structures. The achieved geometric mean absolute error (MAE) is 1.06 m and 0.24 m respectively. Comprehensive comparative evaluations showcase the superiority of the approach in terms of geometric completeness and accuracy, and topological consistence with. The improvement over SAT2LOD2 (Gui and Qin, 2021) is 1.12 m in Cologne (data accessible at https://github.com/dlrPHS/GPUB) and 0.47 m in Braunschweig in geometrical MAE.

1. Introduction

Urbanization is one of the mega-trends that pose massive challenges for humanity. Many of these challenges are linked to buildings, the main structural elements of cities. From disaster management, flood simulation, landslide simulation to solar panel recommendation, all need precise knowledge of building locations, dimensions and appearances. A 3D building model at level of detail (LoD)-2, according to the CityGML standard (Kolbe et al., 2005), is required in those applications. One possible way to obtain LoD-2 city models is to scan these structures with terrestrial laser scanning. Yet, this is a very time and energy consuming approach, and cannot quickly take into account changes in the housing stock of a city or historic buildings in large quantities. Laser-scanning from the air involves a lidar sensor, which provides robust geometrical information but lacks spectral information and is much more expensive and less efficient than an optical camera. Photos from multiple angles of a scene allow the derivation of a photogrammetric digital surface model (DSM). Although it is more noisy than a Lidar DSM, it is less cost-intensive and accompanied by spectral information. To make use of these data, a key step is to extract features from them. Conventional methods rely on hand-crafting such features to detect buildings and their components (Nex and Remondino, 2012; Arefi and Reinartz, 2013; Peters et al., 2022), but these features are often not robust to strong variations in the data. On the other hand, deep learning allows to automatically learn features from high-dimensional data, making it ideal for image recognition in remote sensing.

Although several studies have carried out LoD-2 reconstruction from airborne sensor data (Nex and Remondino, 2012; Arefi and Reinartz, 2013; Alidoost et al., 2019; Gui and Qin, 2021; Peters et al., 2022;

* Corresponding author.
E-mail address: philipp.schuegraf@dlr.de (P. Schuegraf).

https://doi.org/10.1016/j.isprsjprs.2024.04.015
Received 18 December 2023; Received in revised form 3 April 2024; Accepted 4 April 2024
Available online 24 April 2024

Lussange et al., 2023) only few of them use deep learning (Alidoost et al., 2019; Gui and Qin, 2021; Lussange et al., 2023) and none of them predicts the main planar components (i.e. roof planes) of each roof, directly based on an image and photogrammetric DSM. As such, there is a need to uniquely identify building sections even if they have common borders. Note that we regard building sections as parts of a building with distinguishable roof-structure according to a single roof type.

The work in this paper extends our previous work reported in Schuegraf et al. (2023a). In that paper, we introduced a dataset for instance segmentation of buildings and their respective roof planes, named Roof3D. Along with the data, we presented a method that jointly segments building sections and roof planes using a Unet with a ResNet-34 backbone. Although this method showed promising results when operating solely on the image data, integrating the DSM led to a drop in performance. In the current paper, we will use a depth attention module (DAM) to improve the prediction performance for both building sections and roof planes. Here, attention refers to a mechanism that models the interactions in a feature map by learning weights for computing a weighted sum of the input. In our case, the weights are calculated from DSM features. We show that the Efficient-NetB3 is a more suitable backbone for the task at hand. In the meantime, we are able to reduce the number of primitive classes in the preliminary semantic segmentation task from 5 to 4 by removing the outer boundary of building. Additionally, we make use of building sections and planes to derive an LoD-2 reconstruction of our test region in Cologne, Germany. To demonstrate the generalization capability of our method to dissimilar architectural styles and geographical locations, we perform an inference on a separate test region in Braunschweig, Germany.

The rest of this paper is organized as follows. In Section 2, we provide an overview of existing works, draw a boundary to previous research, and highlight our contributions. Section 3 describes our method, noted as PLANES4LOD2, in detail. In Section 4, we explain the data we used and the experimental scheme we employed to evaluate our method. Section 5 includes the qualitative and quantitative results of the experiments. Section 6 discusses limitations of our method and recommends future improvements. Finally, Section 7 concludes this paper and presents an outlook.

2. Related works

2.1. Building instance segmentation

Numerous studies address the challenge of extracting rooftop structures. To illustrate, the PolyMapper approach (Li et al., 2019) directly forecasts buildings and road networks in vector form, but its efficacy on the CrowdAI dataset (Mohanty et al., 2020) is not satisfactory. Conversely, approximating shapes in images with polygons (ASIP) (Li et al., 2020) surpasses the performance of PolyMapper. ASIP initiates polygons by segmenting the image into convex cells, followed by polygon refinement through an energy function. This function minimizes disparities between each polygon's fidelity to the input image and its complexity.

Another method based on frame field learning for enhancing rooftop polygonization is outlined in Girard et al. (2021). The authors train the network to predict building and building border classifications, alongside the frame field representing possible tangent directions at each border pixel. The frame field is employed for regularization during training and can facilitate the polygonization process. Notably, this frame field learning approach disregards the use of DSM.

Furthermore, various endeavors address the challenge of segmenting building sections. PolyWorld (Zorzi et al., 2022) even surpasses the achievements of Girard et al. (2021) and PolyMapper on the CrowdAI dataset (Mohanty et al., 2020) by training layered models for multiple building polygonization sub-tasks. However, PolyWorld falls short in predicting distinct roof sections. This specific issue is delineated in Schuegraf et al. (2023b), which first segments satellite images into background, buildings, and touching borders. Mathematical morphology is then utilized to refine and transform these results into instance segments. In comparison with the Mask-R-CNN (He et al., 2017), the approach of Schuegraf et al. (2023b) is focused on touching borders and produces seamlessly interconnected neighboring building sections. Consequently, this method proves suitable for forecasting individual building rooftops, yet it does not deduce roof planes.

2.2. LoD-2 reconstruction

The derivation of roof planes from imagery and/or height information is often tightly connected to the reconstruction of buildings in LoD-2. Hence, roof plane segmentation is studied in terms of a secondary task for LoD-2 reconstruction. Therefore, we do not distinguish the works dedicated to LoD-2 reconstruction from those dedicated to roof plane segmentation.

Reconstructing buildings in LoD-1 is a well-studied field (Schuegraf et al., 2023b; Yu et al., 2021; Dukai et al., 2019; Peters et al., 2022; Bagheri et al., 2019), whereas the LoD-2 reconstruction has received relatively limited attention in remote sensing research. Nex and Remondino (2012) presented a study that does not involve machine learning but depends on manually designed features to recreate 3D building rooftops. This approach relies on utilizing the near-infrared channel, which may not be universally available. Additionally, the method struggles to accurately handle highly complex building structures. Arefi and Reinartz (2013) also employ a learning-free technique to create LoD-2 building reconstructions by utilizing both the DSM and the orthorectified image. Despite generating improved regular reconstructed buildings, this learning-free approach depends on manually designed features and consequently lacks robustness when encountering significant variations in the input data.

Peters et al. (2022) proposed a method for the reconstruction of buildings in LoD-2 with building sections and lidar point clouds as input. They use a region growing algorithm to partition the footprints into roof planes and detect their intersection lines.

Another work that relies on normalized point clouds for LoD-2 reconstruction is that of Li and Shan (2022), where building primitives from a list of rooftypes are optimized given the point cloud at hand. In this paper, we use a fully convolutional neural network (FCN) to extract building sections and roof planes first and then pair them with a normalized photogrammetric DSM for LoD-2 reconstruction. Furthermore, we rely on RGB image and raster DSM information.

In the study conducted by Alidoost et al. (2019), a solitary aerial image is employed to create LoD-2 building models. Their methodology involves initially training two distinct neural networks: one for estimating building heights and the other for extracting roof features such as eaves, ridges, and hips. Subsequently, a model-based technique is utilized to generate 3D building models. Recently, LoD-2 reconstruction was performed using deep learning methods by Lussange et al. (2023), that use two consecutive Mask-RCNNs, called keypoint inference by segmentation (KIBS). The first Mask-RCNN performs roof plane segmentation, while the second detects roof plane corners and their respective heights in a categorical manner. Although the LoD-2 reconstruction results look promising, the resulting 3D geometries are not necessarily connected to individual buildings. Furthermore, KIBS is dependent on the oblique view image and hence does not generalize to dissimilar viewing angles than those in the training set. Even though learning-based approaches can achieve consistent city models, it is worth noting that the accuracy of the predicted heights solely from an image remains to be a potential limitation. Therefore, there is a need to use heights estimated by robust stereo matching for 3D reconstruction.

In Gui and Qin (2021), buildings are segmented by a semantic segmentation neural network. LoD-2 models are derived using learning-free methods and allowing the integration of open street map (OSM)

Table 1
Memory requirements of a forward and backward pass using the cross entropy loss on an NVIDIA Titan RTX GPU with 24 190 MB memory and the PyTorch deep learning library in python. We provide the memory requirement for batch size 4. The numbers are obtained by performing the computations for batch size 2 and multiplying them by 2. The size of each patch is 512 pixels in width and height.

Model	Modality	Memory requirements
SkipFuse-Unet-3+	RGB+DSM	44.0 GB
UResNet34	RGB	4.0 GB
EfficientUnetB3	RGB	5.7 GB
DepthAtt-EfficientUnetB3 channel & spatial	RGB+DSM	5.9 GB

data. Their method is also based on an ortho image and a DSM as input. Different from Gui and Qin (2021) that uses a rule-based approach, we train the network to predict the separation lines between building sections for improved generalization and robustness. In addition, we also learn the prediction of roof planes, whereas roof-type based models are fitted in Gui and Qin (2021). Gui et al. (2022) provide software as open access, which they describe in detail. This allows comparison with our method.

2.3. FCN architecture

FCNs are convolutional neural networks (CNNs) without fully connected layers. They usually are comprised of two parts, an encoder and a decoder. The encoder, also known as the backbone, is often chosen to be a ResNet (He et al., 2016). It has been successfully utilized for many image recognition tasks, including building segmentation (Liu et al., 2020). Yet, the EfficientNet backbone could improve the performance of ResNet, even requiring lower computational resources (Tan and Le, 2019). Furthermore, EfficientNet exists in different sizes from B0, the smallest, to B7, the largest. The decoder of FCNs often contains as many upsampling layers as downsampling operations in the encoder, connecting levels of identical spatial resolution in the encoder and decoder by skip-connections. This scheme was first introduced in the Unet architecture (Ronneberger et al., 2015). In contrast to the Unet, Unet-3+ (Huang et al., 2020) makes use of the feature maps at different levels of resolution in multiple different skip-connections at the same time. This way, the information flow between the encoder and decoder is even larger than in the Unet. In Schuegraf et al. (2023b), the authors have leveraged the SkipFuse-Unet-3+ architecture for the segmentation of building sections. This modified design demonstrated superiority over other architectures compared in that study. From Table 1 we can observe, that for the forward and backward pass, using only the simple cross entropy loss, our NVIDIA Titan RTX GPU with 24 190 MB memory could not meet the computation requirement. Hence, we could not further investigate the SkipFuse-Unet-3+ architecture for the task of roof plane extraction, because we could not train it properly. Instead, we used the UResNet34 as a baseline, since it had been successfully employed for building section segmentation and roof plane segmentation in Schuegraf et al. (2023a) and compared to the promising EfficientUnet at different scales (Baheti et al., 2020). The SkipFuse scheme to fuse two different data sources, using two separate encoders and one joint decoder, was applied to remote sensing by Henry et al. (2021). Even though it has shown promising results for building section segmentation on some datasets (Schuegraf et al., 2023b), it did not lead to an improvement in the overall performance of joint building section segmentation and roof plane extraction on the Roof3D dataset (Schuegraf et al., 2023a).

2.4. Attention in building segmentation

In recent years, different flavors of attention have been implemented for building segmentation. One such work is Chen et al. (2021), where the authors use self-attention for the semantic segmentation of buildings in optical remote sensing imagery. In Dai et al. (2023), the authors use a location channel attention module to improve the segmentation of building edges in building and water segmentation. Another work that uses a combination of spatial and channel attention is Pan et al. (2019). Besides these cases of using attention in CNNs modules, Sun et al. (2022) use a multi-resolution transformer that heavily depends on the attention mechanism for building and road segmentation. In Wang et al. (2022), a hybrid model combines hierarchical feature extraction of CNNs with global context modeling of transformers for urban scene semantic segmentation. Yet, all works introduced here use only spectral features for attention computations. In contrast, we will show that the introduction of DAM can utilize the height information to extract salient regions in the scene.

2.5. Contributions

Based on the above literature review, we introduce a new approach, PLANES4LOD2, which has the following contributions:

- It predicts building sections and roof planes jointly, such that each roof plane is uniquely connected to a building section.
- It utilizes the predicted building sections and roof planes to achieve a complete LoD-2 reconstruction, which is represented both as a 3D shapefile and an LoD-2 DSM.
- A special attention module, DAM, is able to effectively and efficiently utilize the geometric features of a photogrammetric DSM in a Unet architecture with an EfficientNetB3 backbone.
- By using two independent datasets, we show the superiority of the combination of spatial and spectral attention. Furthermore, we demonstrate the generalization capability of our approach to a test region that is dissimilar in architectural style and geographical location from the primary test region.

3. Methods

We will first give an overview of our workflow, the PLANES4LOD2 method, and then describe its three major steps, including instance segmentation, polygonization and LoD-2 reconstruction.

3.1. Overview

The LoD-2 reconstruction of buildings can be achieved using three main inputs: **(1)** building sections, **(2)** building planes and **(3)** a normalized digital surface model (nDSM).

The definition of building section is often ambiguous. It often refers to a building that has a primitive roof structure, but it can also be interpreted as the building belonging to a building address. In the end, the definition is tightly connected to the ground truth. The data from a public source that we use for training is based on the address definition. On the other hand, addresses are not always visibly discernible. Hence, for the hand-labeled data in the inference, we use the roof primitive definition. In the rest of the paper, we also refer to building section as a roof.

The nDSM is obtained by subtracting a DTM, acquired from a public source, from the photogrammetric DSM. We derive the polygons of building sections and building planes by a two-step procedure. The first step consists of passing an RGB image together with a photogrammetric DSM to FCN, which then produces a 4-class segmentation map. The four classes are background, separation lines between building sections, separation lines between roof planes that do not lie at the junctions between sections, and building segments. In the second step, holes in the line classes are filled using morphological dilation. Then, raster instances are obtained using the watershed transform. Afterwards, the resulting raster instances are polygonized and simplified. As the last step, the polygons and the nDSM are used to generate the LoD-2 model. In that step, random sample consensus (RANSAC) is used to fit 3D roof planes, while ridge lines are generated by intersecting roof planes. Fig. 1 shows the overall workflow of PLANES4LOD2 as described above.

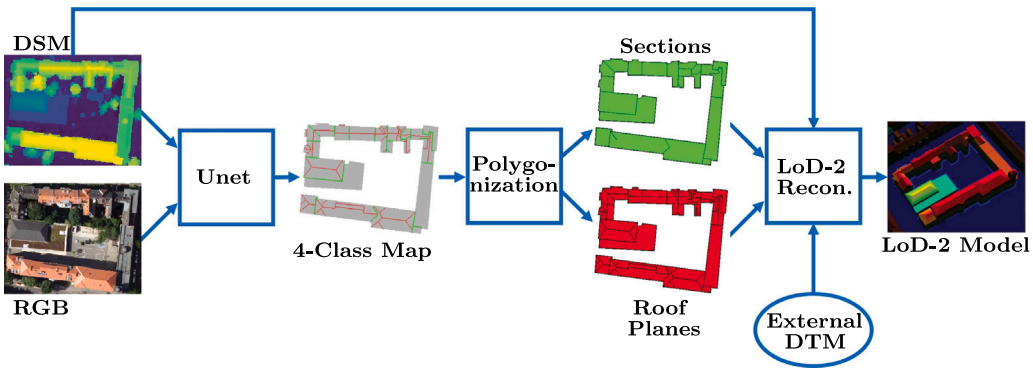

Fig. 1. The overall workflow of PLANES4LOD2. The RGB imagery and DSM patches are passed to Unet to produce a 4-class map. Polygonization yields building sections and roof planes. Using an external DTM, LoD-2 reconstruction generates a vectorized 3D building model.

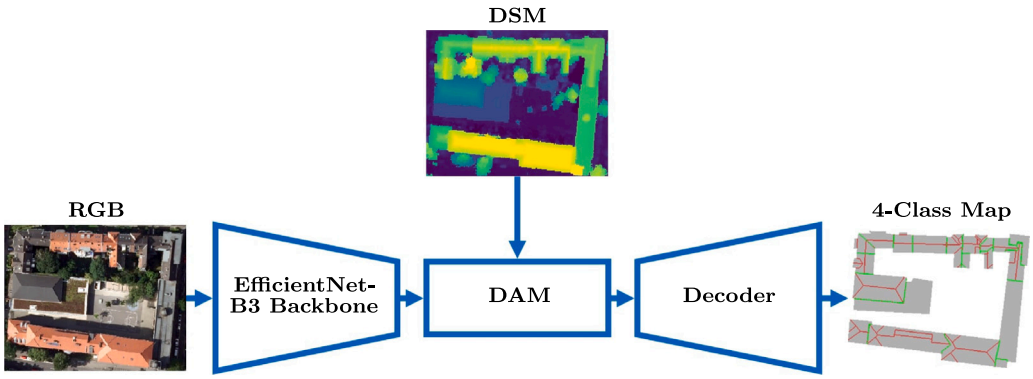

Fig. 2. Our proposed DepthAtt-EfficientUnetB3 architecture. The EfficientNet-B3 backbone extracts features from the RGB data, which are then enriched in the DAM module by DSM information. The decoder reconstructs geometrical details to produce a 4-class map.

3.2. Network architecture

For the task of building section segmentation and roof plane segmentation, UResNet34 has been leveraged in Schuegraf et al. (2023a). Yet, this architecture has multiple drawbacks. First of all, the ResNet architecture has been outdated by the success of the EfficientNet architecture. Second, UResNet34 does not gain from the inclusion of height information, since neighboring buildings may not vary in height, but only in spectral appearance. Thereby, the network is confronted with confusing information. This observation also holds when including the SkipFuse-scheme to the UResNet34 (Schuegraf et al., 2023a). Consequently, we propose the DepthAtt-EfficientUnetB3 architecture. In Fig. 2, the individual parts of our architecture are outlined. The first part of the name DepthAtt refers to an attention mechanism that we call DAM, which is in the center of Fig. 2. DAM is applied at the last layer of the encoder of a Unet architecture. It receives a photogrammetric DSM patch as the input. DAM uses two different attention mechanisms based only on the DSM, leveraging height features at the deep part of the network. We apply a sequence of strided convolutional layers, ReLU activations and maximum pooling layers to the DSM, as is visualized in the upper part of Fig. 3. The convolution operations allow automatic learning on features from the raw height information. ReLU introduces non-linearity to the network. The stride in the convolutions and the maximum pooling layers bring the height features to the same resolution as the feature maps of the bottleneck of the image network. These height features are then used in two types of attention layers. Both attention layers consist of a convolutional layer followed by a sigmoid activation. Yet, one of the attention layers uses a channel-wise convolution to place attention on features other than regions, whereas the other attention layer places the attention on pixels, to enhance features in certain spatial locations, which are derived from the DSM. DAM enables fusion of the RGB data and DSM at a coarse spatial resolution (see lower part of Fig. 3), namely at the bottleneck. Hence, small spatial shifts between the two inputs affect little to the extracted feature maps. Moreover, since the DSM is only used in the attention mechanism, the network focuses on the features from the RGB image, but can use height features to suppress noise and guide the training process. The EfficientUnetB3 receives only an RGB image patch and extracts features sequentially in the encoder, leveraging the EfficientNetB3 architecture (Tan and Le, 2019), which is shown on the left side of Fig. 2. As for the decoder (right part of Fig. 2), what we used is similar to Unet in Baheti et al. (2020). This includes skip-connections to allow for better information flow from the encoder to the decoder. Our implementation of the EfficientUnetB3 is mostly based on an implementation that is publicly available on github.[1] During the training, we use a *softmax* activation, since it is required by our loss functions. When doing inference, we use the argmax of the network outputs to produce class predictions for each pixel.

3.3. Polygonization

Although the raster results are valuable for some applications, most further applications, e.g. LoD-2 reconstruction, require vector data as the input. Hence, we convert our 4-class maps to two different vector layers. Note that when we refer to simplification algorithms in the following paragraphs, we always simplify common borders of polygons and the rest of the polygons separately to avoid irregular gaps between

[1] https://github.com/zhoudaxia233/EfficientUnet-PyTorch.

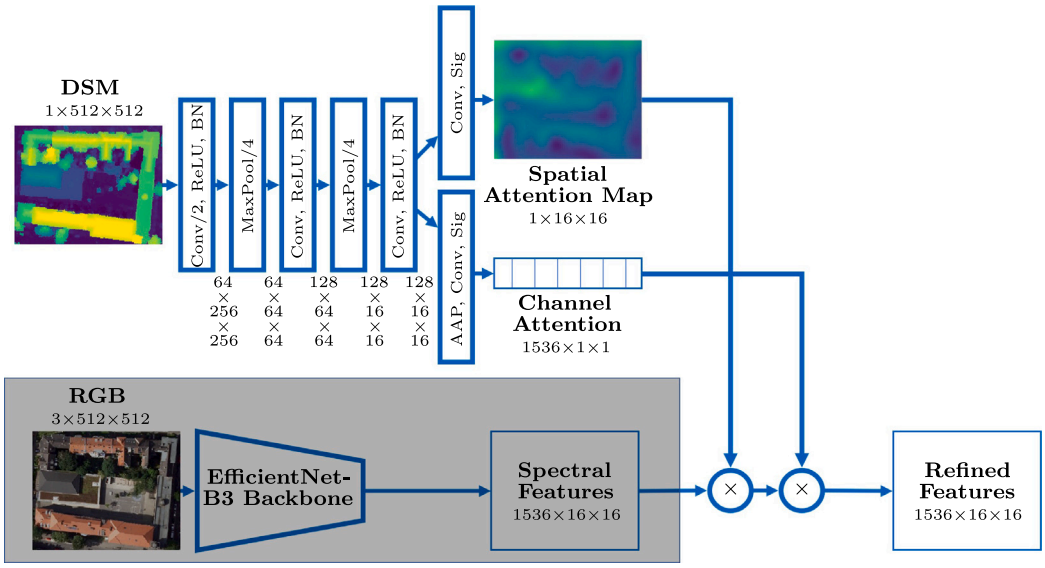

Fig. 3. The structure of DAM. The shadowed area is not part of DAM but is visualized to show the origin of the spectral features. Conv and Conv/2 refer to convolutional layers with stride 1 and 2. ReLU means rectified linear unit, BN represents batch normalization, and MaxPool/4 refers to a maximum pooling layer with stride 4. AAP refers to adaptive average pooling and Sig stands for the sigmoid function, which maps its inputs to the range $[0, 1]$. The spatial attention map is visualized with 32-times the original resolution using bi-cubic interpolation.

neighboring instances. Section polygons with an area smaller than $4\,\mathrm{m}^2$ are dropped, since they most likely correspond to false positive noise.

Building section layer. We achieve the separation of buildings into sections by using the same learning-free post-processing scheme as in Schuegraf et al. (2023b). We treat the plane separation line as part of a section. Hence, this leaves us with three classes: background, building segment and section separation line. Then we use the watershed transform to infer instances. As the seed for the watershed transform, we dilate the section separation line, using a disk with radius $R_{\mathrm{sec}} = 6$ as the structuring element, and remove it from the building segment. The mask element for the watershed transform is the inverse of the background class raster. The surface map will be the segmentation raster with value 0 for background, value 1 for building segment, and value 2 for building section separation. To obtain boundary pixels, we use tree search and simplify the resulting polygon by utilizing the Douglas Peucker algorithm (Douglas and Peucker, 1973) with tolerance $\epsilon_{\mathrm{sec}} = 0.5\,\mathrm{m}$.

Roof plane layer. For the generation of a roof plane vector layer, we follow the same procedure as for the building section layer. The only difference is that we reconstruct the plane separation line by using both the building section separation and plane separation as the separation line. We again apply dilation to improve separation between sections, but with a disk of radius $R_{\mathrm{plane}} = 6$. To simplify the roof plane polygons with the Douglas Peucker algorithm, we use the tolerance $\epsilon_{\mathrm{plane}} = 0.5\,\mathrm{m}$.

3.4. LoD-2 model generation

The next task is to generate the LoD-2 model based on our predicted roof plane geometries. As the first step, we count the number of predicted planes of each building section. For a single plane, we estimate the roof plane parameters (a_i, b_i, c_i, d_i) using RANSAC (Fischler and Bolles, 1981). The parameters define a plane with the equation $a_i x + b_i y + c_i z + d = 0$ for roof plane i. We then check whether the plane is nearly horizontal or parallel to the xy-plane ($a_i \sim 0$, $b_i \sim 0$, $z_i \sim 1$). In that case, we improve regularization by assuming complete flatness of the roof plane and average the height value of all vertices inside the roof polygon to obtain a single height value at all vertices.

If there are two planes for a roof, we assume that it is a gable roof. Even though not all roofs with two planes are of roof type gable, this assumption holds for most buildings in our datasets. We estimate a plane for each of the roof planes using RANSAC. For plane estimation, we sample all height values from the nDSM that lie in the area surrounded by the roof plane polygon. Next, we use the two sets of plane parameters (a_1, b_1, c_1, d_1) and (a_2, b_2, c_2, d_2) to compute their intersection line in the point-slope expression $\vec{l}(t) = \overrightarrow{p_0} + t \times \vec{s}$, where $p_0 = [x_0, y_0, z_0]^\intercal$, $\vec{s} = [\delta_x, \delta_y, \delta_z]^\intercal$ with

$$\delta_x = b_1 c_2 - b_2 c_1, \tag{1}$$

$$\delta_y = a_2 c_1 - a_1 c_2, \tag{2}$$

$$\delta_z = a_1 b_1 - a_2 c_1, \tag{3}$$

$$x_0 = \begin{cases} 0, & \text{if } \delta_x \neq 0 \\ (d_1 c_2 - d_2 c_1) \div \delta_y, & \text{if } \delta_x = 0 \wedge \delta_y \neq 0 \\ (d_2 b_1 - d_1 b_2) \div \delta_z, & \text{if } \delta_x = 0 \wedge \delta_y = 0 \wedge \delta_z \neq 0, \end{cases} \tag{4}$$

$$y_0 = \begin{cases} (c_1 d_2 - c_2 d_1) \div \delta_x, & \text{if } \delta_x \neq 0 \\ 0, & \text{if } \delta_x = 0 \wedge \delta_y \neq 0 \\ (d_1 a_2 - d_2 a_1) \div \delta_z, & \text{if } \delta_x = 0 \wedge \delta_y = 0 \wedge \delta_z \neq 0, \end{cases} \tag{5}$$

$$z_0 = \begin{cases} (b_2 d_1 - b_1 d_2) \div \delta_x, & \text{if } \delta_x \neq 0 \\ (d_2 a_1 - d_1 a_2) \div \delta_y, & \text{if } \delta_x = 0 \wedge \delta_y \neq 0 \\ 0, & \text{if } \delta_x = 0 \wedge \delta_y = 0 \wedge \delta_z \neq 0 \end{cases} \tag{6}$$

and $t \in \mathbb{R}$. We intersect this line with the union polygon of the two roof planes using a line search. There, we iteratively evaluate the point-slope expression for different pairs (t_0, t_1), check whether the line that passes through the two resulting points $\vec{l}(t_0)$ and $\vec{l}(t_1)$ intersects the union polygon, until we find a pair (t_0, t_1). At the two intersection points, we use their average height according to the point-slope expression. The heights of the remaining vertices of the union polygon are complemented using the initial planes parameters. Next, we split the union polygon through the intersection line defined by the two intersection points. As a result, we yield two roof plane 3D-polygons with consistent height at the ridge line, i.e. avoiding vertical jumps of elevation.

For buildings with more than two planes, we use RANSAC to determine the plane parameters similar as for two planes. If the normal of a roof plane indicates a non-inclined plane, we model it as a flat roof

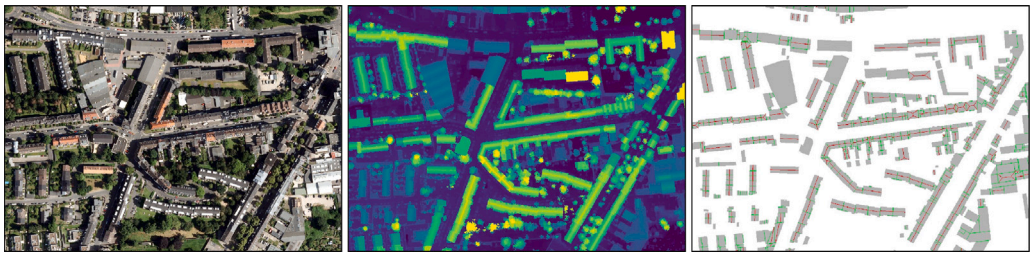

Fig. 4. Excerpt from the training data of Roof3D. The RGB imagery was captured with a GSD of 0.1 m, whereas the DSM was computed with 0.5 m GSD. Before being passed to the network, both of them are resampled at 0.3 m GSD using bicubic interpolation, since the ground truth is generated at 0.3 m GSD.

with the average elevation at all vertices. We model the non-flat roof planes by using their estimated plane parameters to complement the height values at the vertices.

To complete the building models, we further include ground and wall polygons.

4. Experiments

4.1. Data

We use two different datasets for the experiments in this paper. The first dataset is Roof3D (Schuegraf et al., 2023a), with data from the cities Cologne and Berlin for training and Cologne for evaluation. The RGB imagery and photogrammetric DSMs in Roof3D are comprised of real and synthetic pairs. The addition of the synthetic data increases the size of the training dataset and comes along with perfectly matching ground truth. Next to the perfect annotations of the synthetic data, Roof3D includes two more sources of ground truth. One source includes building outlines from the German building cadastre and coarse roof plane annotation from a semi-automatic method based on laser-scanning. The other source is manual annotation of real image and DSM pairs. The first testing region is that of Roof3D in Cologne, Germany, which we use for ablation, is annotated manually and has exclusively non-synthetic inputs. See Fig. 4 for a visualization of an area in the training data. Furthermore, the testing region does not geographically overlap with the training data. Refer to Schuegraf et al. (2023a) for further details about Roof3D. For the construction of a reference LoD-2 DSM, we use public data.[2] This reference data stems from a semi-automatic method that uses cadastre data and laser scanning, which often leads to erroneous annotations. The testing set of Roof3D originates from the same flight campaign as some of the images used for training.

As for our second dataset, showing a part of the city Braunschweig, Germany, is solely used for testing and stems from a different flight campaign, with different lighting conditions, architectural styles and viewing angles, leading to dissimilar artefacts in the orthorectified imagery. We use the same tiling scheme for all our tests as in Roof3D. Both the RGB data and DSM in the two datasets have 0.3 m GSD after resampling.

4.2. Training details

It is important to train an FCN according to the requirements of the task at hand. One important aspect is the choice of the loss function, which defines the learning objective together with the ground truth. We use the weighted cross-entropy loss, which is a standard choice for semantic segmentation tasks, with weight 1 for the background class, 6 for the roof plane separation, 6.2 for the building section separation and 1.5 for the building segment class. We obtained these values by using the median frequency weighting heuristic

$$w_{cl} = \frac{\text{freq}_m}{\text{freq}_{cl}}, \tag{7}$$

where freq_m is the median of the frequencies of pixels of each class and freq_{cl} is the pixel frequency of class cl. However, the cross-entropy loss is known to generate models producing blurry objects. To obtain sharper object boundaries, we combine the cross-entropy loss with the generalized dice loss (Sudre et al., 2017), which has inverse frequency class weights. Where noted ("Topo") we also use the topological loss (Mosinska et al., 2018) to regularize the semantic raster output of the respective network. Topological loss was previously applied to regularize building footprints (Zhang et al., 2020; Schuegraf et al., 2023b). We apply it to the building segment class (weighted with 0.05 in the loss function) and the union of building section separation and roof plane separation lines (0.1), as two separate terms in the loss function. As the optimization algorithm, we leverage AdamW (Loshchilov and Hutter, 2019) with weight decay of 0.0001, as it is a common choice for training FCNs.

4.3. Evaluation metrics

For evaluation, we use two kinds of metrics. The first ones are for the evaluation in 2D, and the second in 3D.

Metrics 2D. To quantitatively evaluate the two instance segmentation tasks, building section and roof plane segmentation, we use average precision (AP) and average recall (AR). The harmonic mean of these two is

$$F1_{INST} = 2 \times \frac{\text{AP} \times \text{AR}}{\text{AP} + \text{AR}}. \tag{8}$$

AP and AR are two commonly used metrics for instance segmentation. The two metrics highly depend on the overlap between the predicted instances and ground truth instances and are thus highly discriminative. Furthermore, ambiguous ground truth can lead to low values of these metrics. AP focuses on the quality of the predicted results by considering both precision and recall, while AR focuses solely on the proportion of relevant items that are successfully retrieved. Hence, AR responds better to over-segmenting methods, whereas AP has a higher score on under-segmenting methods. Both metrics are based on the polygonized results and polygonized ground truth. Since these metrics only give insight to quantitative aspects of the results, we also carry out a visual inspection for qualitative evaluation in some of the experiments.

Metrics 3D. For the quantitative evaluation of the reconstructed LoD-2 DSM with our and reference methods, we use the root-mean-squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_p |\hat{h}_p - h_p|^2}{N}}, \tag{9}$$

where p is the respective pixel, N is the total number of pixels, $\hat{h}_p$ is the predicted elevation at pixel p and h_p is the reference elevation at pixel

[2] https://www.opengeodata.nrw.de/produkte/geobasis.

Table 2
Results of various models for the building section segmentation task on the Roof3D dataset. ↑ indicates that the higher values of the metrics correspond to better quality.

Architecture	Modality	AP ↑	AR ↑	$F1_{INST}$ ↑
UResNet34	RGB	0.183	0.371	0.245
Fuse-UResNet34	RGB+DSM	0.176	0.365	0.237
DepthAtt-UResNet34 channel & spatial	RGB+DSM	0.201	0.390	0.265
DepthAtt-UResNet34 spatial	RGB+DSM	0.179	0.365	0.240
DepthAtt-UResNet34 channel	RGB+DSM	0.170	0.359	0.231
SpecAtt-UResNet34 channel & spatial	RGB	0.194	0.379	0.257
SpecDepthAtt-UResNet34 channel & spatial	RGB+DSM	0.183	0.359	0.242
DepthAtt-EfficientUnetB3 channel & spatial	RGB+DSM	**0.207**	**0.398**	**0.272**
DepthAtt-EfficientUnet-B3-Topo channel & spatial	RGB+DSM	0.197	0.361	0.255

p. Yet, the RMSE is sensitive to the scale of the values and to outliers. Hence, we use a more robust regression metric, mean absolute error (MAE)

$$\mathrm{MAE} = \frac{\sum_p |\hat{h}_p - h_p|}{N}. \tag{10}$$

Another metric, which originates from stereo matching and optical flow, is the T_t-error

$$T_t = \frac{1}{N} \sum_p \begin{cases} 1 & \text{if } |\hat{h}_p - h_p| \geq t \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

which gives the percentage of pixels, where the predicted height has an absolute deviation of more than t from the ground truth, where t is expressed in meters. We use the strict T_1-error and the T_3-error to gain a better overall understanding of the quality of the predictions of our method.

4.4. Experiment descriptions

For the analysis of our method, we perform multiple sets of experiments.

4.4.1. Roof3D

The public Roof3D dataset is suitable for the evaluation of algorithms on the tasks of segmenting building sections and roof plane extraction. Hence, we use it to carry out an ablation study to find the best setting of our architecture.

Polygonization. As a baseline model, we train UResNet34 for the 4-class semantic segmentation task using only RGB imagery. Post-processing techniques, as outlined in Section 3, are applied to obtain building sections and roof planes. In Schuegraf et al. (2023a), it was shown that Fuse-UResNet34 does not improve UResNet34, even though it has auxiliary height information as input. To address such drawbacks, we experiment with DepthAtt-UResNet34 with channel and spatial attention, leveraging DAM. To discern the impact of attention mechanisms, we evaluate DepthAtt-UResNet34 with only spatial attention, only channel attention, and both channel and spatial attention. Spectral attention uses only the features from the RGB image to derive attention maps. In an effort to determine the efficacy of depth and spectral attention, we introduce SpecAtt-UResNet34 with both channel and spatial attention. Additionally, we test the combination of spectral and depth attention in SpecDepthAtt-UResNet34. Given the success of EfficientNet in various image recognition tasks, we explore the performance of the DepthAtt-EfficientUnetB3 with channel and spatial attention architecture. We evaluate the EfficientNetB3 backbone and compare them to DepthAtt-UResNet34. To enhance regularization in the segmentation outputs, we introduce the topology loss to DepthAtt-EfficientUnetB3-Topo channel & spatial architecture.

LoD-2 reconstruction. We leverage DepthAtt-EfficientUnetB3-Topo channel & spatial to derive building sections and roof planes from pairs of RGB imagery and photogrammetric DSMs. The input photogrammetric DSM is normalized using a DTM from a public source to extract heights above ground. Then, we apply our LoD-2 reconstruction method from Section 3.4. We use SAT2LOD2 (Gui and Qin, 2021) for comparison to our method with the software described in Gui et al. (2022). We feed only the ortho image and photogrammetric DSM to SAT2LOD2, omitting the OSM data.

4.4.2. Generalization

One of the great promising properties of deep learning-based algorithms is their generalization capability. To test this, we apply our DepthAtt-EfficientUnetB3-Topo channel & spatial to a dataset that does not geographically overlap with the Roof3D dataset. Since this dataset stems from an entirely different campaign, this implies not only different architectural styles, but also different viewing angles and lighting conditions leading to a different appearance of buildings in the ortho image than those in the Roof3D dataset, as well as different architectural styles.

5. Results

5.1. Roof3D

In this subsection, we compare the quantitative results as in Tables 2 and 3 and the qualitative results from a visual inspection of the models trained and evaluated on the Roof3D dataset. The regression metrics for the 3D reconstruction task are provided in Table 4.

5.1.1. Quantitative results

RGB-based building and roof plane segmentation. UResNet-34 successfully segments building sections and roof planes.

Polygonization. Comparing Fuse-UResNet34 and DepthAtt-UResNet34 channel & spatial, we observe that the latter is scoring higher metric values. Hence, the noise suppression and feature refinement of DAM lead to improved metrics. On the contrary, directly incorporating DSM makes it harder for the network to focus on RGB data, which contains the most important spectral information. Inspecting metric scores of the three models with different depth attention settings, the combination of channel and spatial attention outperforms spatial attention and channel attention. Using only one of the two attention mechanisms is not sufficient (Woo et al., 2018), but the combination of both leads to a more effective use of the features provided by the RGB image encoder.

We observe in Table 2 that the depth attention improves the performance of spectral attention in SpecAtt-UResNet34 channel & spatial. The combination of spectral and depth attention in SpecDepthAtt-UResNet34 is ranking even behind SpecAtt-UResNet34 channel & spatial. The attention provided by the input RGB data is helpful, but does not provide as much additional information as the depth attention does.

Table 3
Results of various models in the roof plane segmentation task on the Roof3D dataset. The second-last and third-last row correspond to identical metric values. ↑ indicates that the higher values of the metrics correspond to better quality.

Architecture	Modality	AP ↑	AR ↑	$F1_{INST}$ ↑
UResNet34	RGB	0.115	0.279	0.163
Fuse-UResNet34	RGB+DSM	0.119	0.289	0.169
DepthAtt-UResNet34 channel & spatial	RGB+DSM	0.127	0.295	0.178
DepthAtt-UResNet34 spatial	RGB+DSM	0.100	0.267	0.146
DepthAtt-UResNet34 channel	RGB+DSM	0.117	0.283	0.166
SpecAtt-UResNet34 channel & spatial	RGB	0.123	0.282	0.171
SpecDepthAtt-UResNet34 channel & spatial	RGB+DSM	0.109	0.265	0.154
DepthAtt-EfficientUnetB3 channel & spatial	RGB+DSM	0.138	0.303	0.190
DepthAtt-EfficientUnetB3-Topo channel & spatial	RGB+DSM	**0.149**	**0.312**	**0.202**

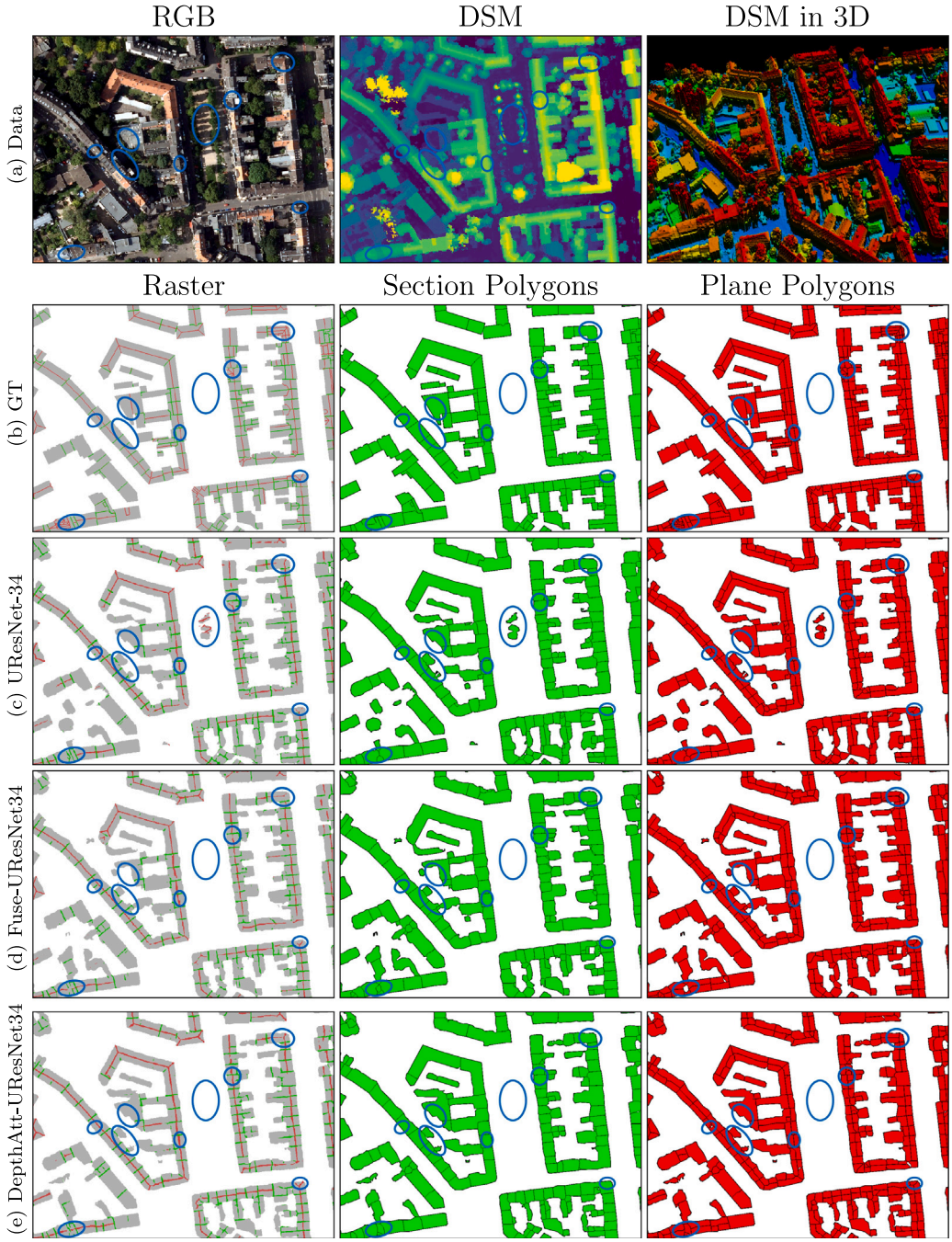

Fig. 5. Visualization of the 2D results on a crop of the Roof3D test region. Row (a) shows the input data. Row (b) shows the reference ground truth and (c) the prediction of the UResNet-34. Row (d) presents the results derived from the Fuse-UResNet-34 and (e) those of the DepthAtt-UResNet34 channel & spatial. Blue ovals highlight the differences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

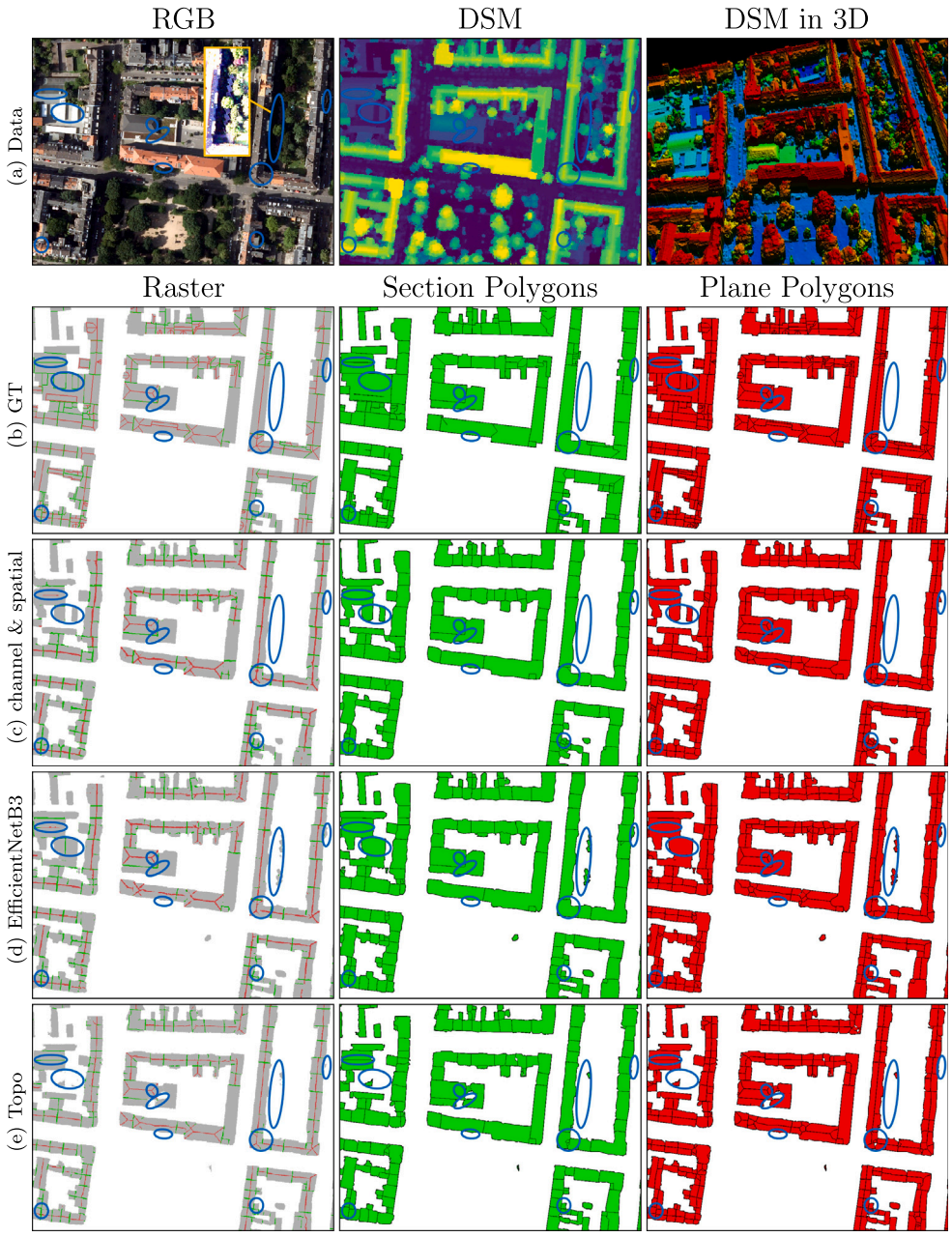

Fig. 6. Visualization of the 2D results on another crop of the Roof3D test region. Row (a) shows the input data. Row (b) presents the reference ground truth and (c) the prediction of the DepthAtt-UResNet34 channel & spatial. Row (d) shows the results of the DepthAtt-EfficientUnetB3 channel & spatial and (e) those of the DepthAtt-EfficientUnetB3-Topo channel & spatial. Blue ovals highlight the differences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of the LoD-2 reconstruction results on the two test regions. ↓ indicates that the lower values of the metrics correspond to better quality.

Method	Dataset	RMSE ↓	MAE ↓	T_1 ↓	T_3 ↓
SAT2LOD2	Roof3d	5.28 m	2.18 m	0.26	0.15
PLANES4LOD2	Roof3d	**3.34 m**	**1.06 m**	**0.18**	**0.07**
SAT2LOD2	Braunschweig	2.52 m	0.71 m	0.10	0.08
PLANES4LOD2	Braunschweig	**1.39 m**	**0.24 m**	**0.04**	**0.02**

Averaging the attention maps from the spectral and depth information does not seem to be the best way to make use of both mechanisms. We compared eight different settings of EfficientNet as the backbone and EfficientNetB3 outperforms all other versions on both tasks. The replacement of the ResNet34 backbone by the EfficientNetB3 backbone in DepthAtt-EfficientUnetB3 channel & spatial consistently outperforms all other backbones on all metrics. The most likely reason for the superiority of EfficientUnet over ResNet34 is its fine-grained scalability as compared to UResNet34. This allows us to choose a properly dimensioned feature extractor.

Training DepthAtt-EfficientUnetB3 spatial & channel with the topology loss leads to a drop in performance on building segmentation, but to a rise in performance on roof plane segmentation. The strength of introducing the topology loss is that it makes the predictions visually more similar to the ground truth. Because of the complex junctions of separation lines between roof planes, the model profits strongly if it is pushed to segment thin and topologically correct lines and junctions. On the other hand, building segmentation profits more from thicker lines, which avoids gaps better and hence leads to less missed separations between resulting building section polygons.

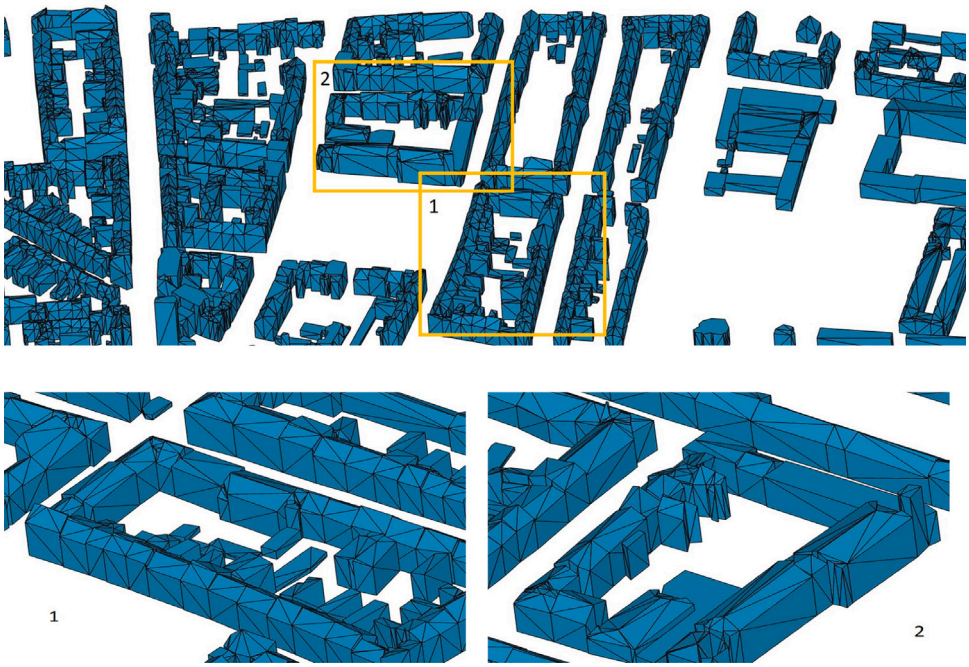

Fig. 7. The resulting 3D LoD-2 model in vector format of a scene in the Roof3D test region. The image in the top row shows an overview, whereas the bottom row gives two detailed views.

LoD-2 reconstruction. Evaluating the LoD-2 DSM with reference to the ground truth raster, our method achieves better values than SAT2LOD2 on all metrics, indicating more accurate geometrical results. In addition, our PLANES4LOD2 has accurate presentation about roof planes. Furthermore, PLANES4LOD2 recognizes inner yards and can properly handle such topological structure of buildings, whereas SAT2LOD2 regards them as parts of the buildings. All these factors demonstrate that PLANES4LOD2 performs superior.

5.1.2. Qualitative results

Polygons. Fig. 5 visualizes the predictions of models with different modalities and fusion strategies. We highlight multiple places where we noted significant visual deviations. In row (e), DepthAtt-UResNet34 channel & spatial produces separation lines in the raster segmentation. The results are more complete than in the other rows, which leads to more accurate and regular building sections and roof plane polygons than UResNet34 in row (c) and Fuse-UResNet34 in row (d). Furthermore, Fuse-UResNet34 sometimes produces false positives. Fig. 6 presents the comparisons of the results obtained from models with the backbone architectures ResNet-34 (row (c)) and EfficientNetB3 (rows (d) and (e)) and under the addition of the topology loss (row (e)). In the RGB image, we highlight a rectangle by rescaling it to the lowest 30% of pixel values, which correspond to shadows in the original RGB image. In the highlighted box, regarding the low corresponding elevation in the DSM, the visible building structure in the middle most likely corresponds to garages. This structure is detected as buildings by DepthAtt-EfficientUnetB3 channel & spatial, whereas DepthAtt-UResNet34 channel & spatial segments it as background. In most parts of the visualization, DepthAtt-EfficientUnetB3-Topo channel & spatial produces thinner and more complete lines than the other two models, though it sometimes fails to detect building segments. Overall, the two models with EfficientNetB3 as backbones produce slightly more complete separation lines.

LoD-2 model. In Fig. 7, a resulting LoD-2 model is visualized in vector format. From Fig. 8 it becomes clear that the 3D building model of our method looks more similar to the ground truth than the one from SAT2LOD2. In the first row of Fig. 9, we provide a more detailed visualization of the reconstruction performed by SAT2LOD2, our method and the reference ground truth. The rooftops generated by SAT2LOD2 look very regular because they are based on roof type reconstruction. This induces symmetry into the resulting roof of the building model. Our method generates building models that are visually much closer to the ground truth, but does not enforce symmetric properties similar to SAT2LOD2. Furthermore, SAT2LOD2 cannot reconstruct buildings with inner yards correctly, because it is based on binary building segmentation. In contrast, we reconstruct buildings based on individual sections and directly segment their roof planes. Hence, our method can capture inner yards well, which is an advantage in scenarios with complex building structure, as it is typical in European cities.

5.2. Generalization

To test the capability of our LoD-2 reconstruction method to adapt to an entirely new scene with different lighting conditions and different architectural styles, we evaluated it on a test region in Braunschweig, Germany. We also evaluated the SAT2LOD2 method on the same data for comparison. Quantitatively, Table 4 shows that our method scores RMSE 1.39 m, MAE 0.24 m, T_1 0.04 and T_3 0.02, whereas SAT2LOD2 achieves RMSE 2.52 m, MAE 0.71 m, T_1 0.10 and T_3 0.08. Hence, our method quantitatively outperforms the reference method compared by a factor of ~2 to 3. Since SAT2LOD2 fits roof tops based on roof type primitives, it does not produce rooftops that are structurally accurate. On the other hand, our method fits a plane to each segmented roof plane polygon, which leads to more accurate, but less mathematically symmetric roof tops. Visually, in Fig. 10, both our method and SAT2LOD2 show a reconstruction that looks quite similar to the ground truth. Taking a closer look in the second row of Fig. 9, the impression remains that both results are similar to the ground truth. Even though our method also outscores SAT2LOD2 on the simple Braunschweig test area, the advantages of PLANES4LOD2 are most significant when studying more complex scenes like the test region of Roof3D.

6. Discussion

The quality of the LoD-2 resulting from PLANES4LOD2 is affected by multiple factors. If there is high vegetation covering the roof plane, the

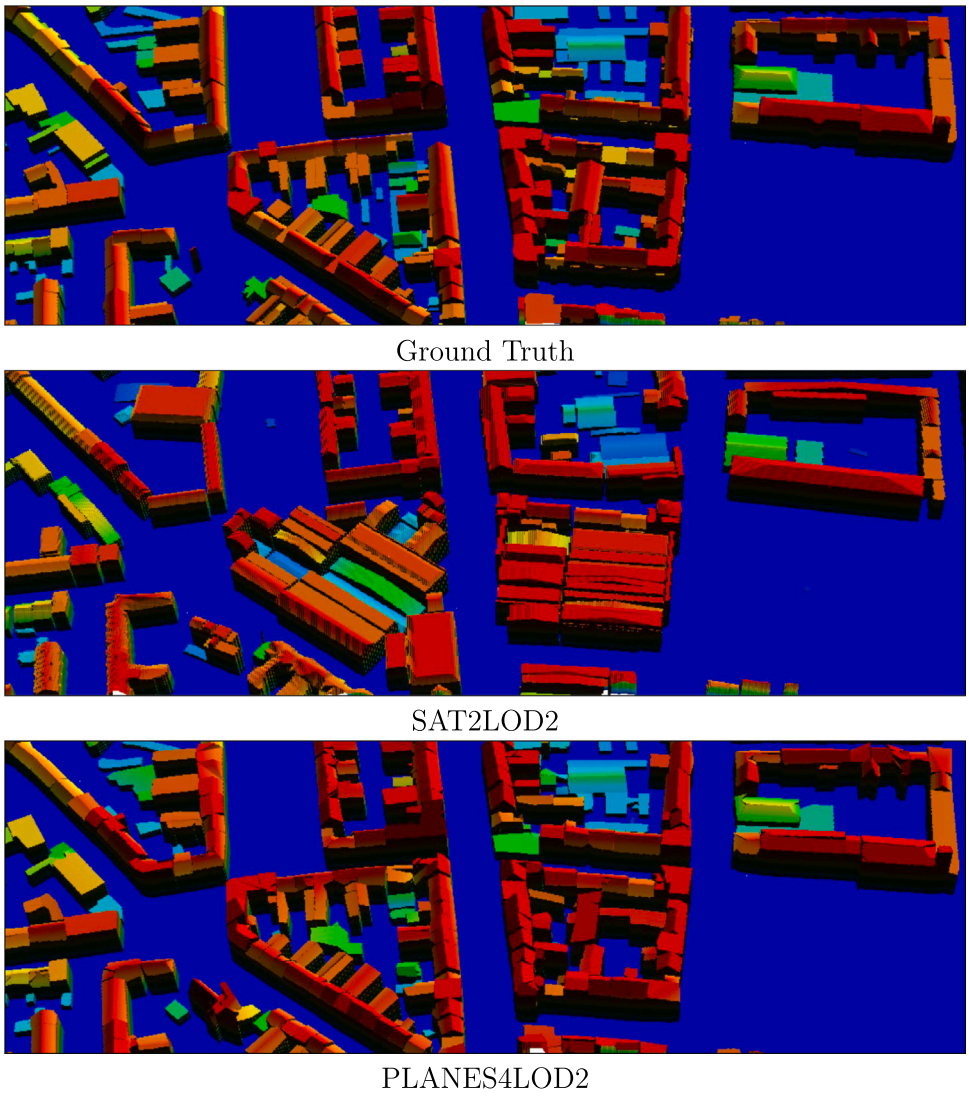

Fig. 8. Visualization of the results of our and a reference method for LoD-2 reconstructions of the test region of Roof3D. For the visualization of height features, we use a color mapping from blue (low) to red (high). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

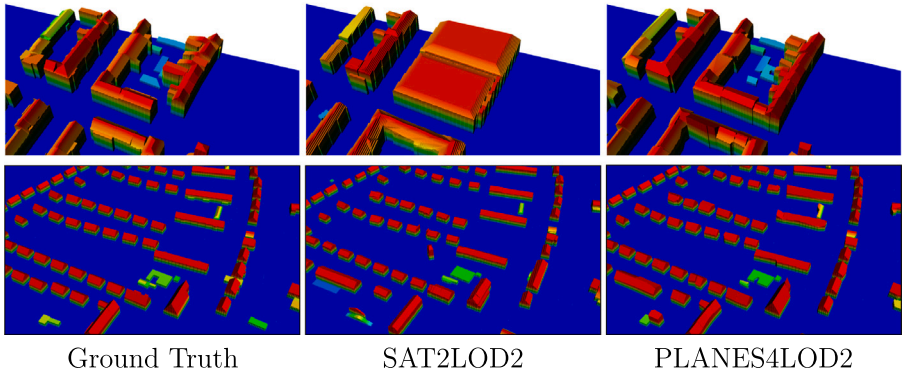

Fig. 9. An example building in the test region of Roof3D (first row) that is reconstructed as a large block by a reference method and reconstructed in detail in our reconstruction and some example results in the test area in Braunschweig, Germany (second row). Both methods achieve practically identical results for simple roof shapes, as can be seen in the second row. However, PLANES4LOD2 can handle more complex buildings as visualized in the first row.

accuracy of the associated plane parameters might be decreased. One possible way to address this issue would be to use a separate network to remove trees from the DSM (Bittner et al., 2019, 2020; Stucker and Schindler, 2022). Furthermore, we do not enforce symmetry between roof planes for any roof type other than for a single plane. Since we assume the roof type to be gable for buildings with two predicted roof planes, roof tops that have vertical gaps between roof planes will be modeled as if they intersect at the ridge line. Buildings with roof

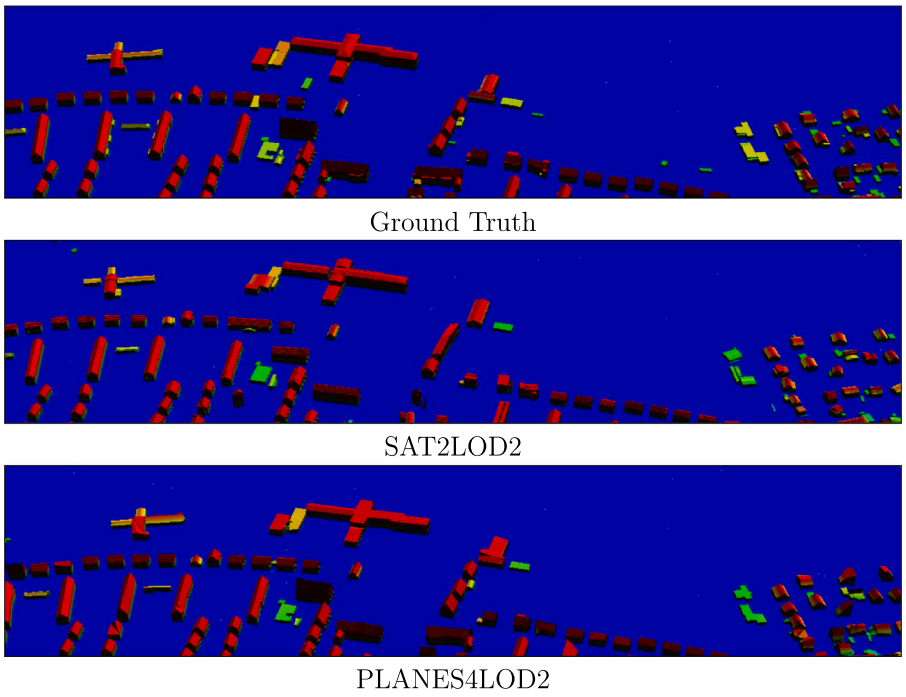

Fig. 10. Visualization of the results of our and a reference method for LoD-2 reconstructions of buildings in Braunschweig. For the visualization of height features, we use a color mapping from blue (low) to red (high). The model is trained with data from Cologne and Berlin. PLANES4LOD2 profits most from the high resolution RGB image, allowing it to separate connected or close building sections. Furthermore, it is capable to filter noise from the DSM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

types like hip and half-hip will not be reconstructed in a regularized style, since we do not assure either symmetry or intersection of the roof planes at the identical height at junctions between them. On the other hand, primitive-based approaches like that of Li and Shan (2022) fit roof models that are a-priori symmetrical, but are less flexible than PLANES4LOD2. In practice, one could combine a primitive-based approach for simple roof types with PLANES4LOD2 for the remaining roof structures.

Further restrictions are induced by the GSD. We decided to use 0.3 m. A smaller GSD leads to better visibility of the roof lines and it would be easier to distinguish roof planes. On the other hand, it would cause more noise, since more details are visible, which the network would have to learn. Regarding a larger GSD, it would cause blurrier lines and PLANES4LOD2 is sensitive to the visibility of separation lines.

We also observed that PLANES4LOD2 predicts the instances of roof planes and building sections more accurate than what the metrics suggest. The reason for this is that the common objects in context (COCO)-metrics, including AP, AR and $F1_{INST}$ we are using, are very sensitive. For one ground truth polygon, if the highest overlap with a predicted polygon is 0.4999, it will not be recognized as a true positive, but a false negative. Another effect, that makes metrics underestimate, is possible ambiguous ground truth. Many small roof planes that exist in the ground truth can hardly be seen by bare eyes, or are so small that even a fine-grained neural network cannot detect them as a separate object. Moreover, the COCO-metrics compute the average precision AP and AR not only for the threshold 0.5, but also for much higher thresholds up to 0.95. While this is a reasonable threshold for large buildings or large objects on multi-media imagery, it is hard to achieve a good score in building section or roof plane segmentation. On the other hand, those metrics are commonly used in instance segmentation task and we argue that they are sufficient and realistic to compare different experimental setups.

7. Conclusion

We presented PLANES4LOD2, a method that uses planar roof components to reconstruct buildings as level of detail (LoD)-2 models. The PLANES4LOD2 pipeline relies on deep learning as well as conventional approaches to implement a full 3D reconstruction pipeline from an RGB image, a photogrammetric digital surface model (DSM) and a digital terrain model (DTM). The method makes use of the DSM in the novel depth attention module (DAM) to enhance building plane prediction and in the roof surface reconstruction. PLANES4LOD2 robustly interpolates roof surfaces from sampled height values and initial roof planes. The resulting LoD-2 building model appears visually similar, or close to be identical to the ground truth, even when the test region contains very complex building structures and is densely built. Furthermore, we demonstrated the advantages of our method for LoD-2 reconstruction compared to other software. We also evaluated PLANES4LOD2 on a test region in a different city. The results reveal superior generalization capability of our method being adaptive to lighting conditions and architectural styles different from the ones the model is trained.

CRediT authorship contribution statement

Philipp Schuegraf: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jie Shan:** Investigation, Supervision, Writing – review & editing. **Ksenia Bittner:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Prof. Dr. Rongjun Qin and Mr. Shengxi Gui for their support in processing our testregions with the SAT2LOD2 software to provide comparing results. Dr. Shan got involved in this work when he was a visiting scientist at DLR.

References

Alidoost, F., Arefi, H., Tombari, F., 2019. 2D image-to-3D model: Knowledge-based 3D building reconstruction (3DBR) using single aerial images and convolutional neural networks (CNNs). Remote Sens. 11, 2219.

Arefi, H., Reinartz, P., 2013. Building reconstruction using DSM and orthorectified images. Remote Sens. 5 (4), 1681.

Bagheri, H., Schmitt, M., Zhu, X., 2019. Fusion of multi-sensor-derived heights and OSM-derived building footprints for urban 3D reconstruction. ISPRS Int. J. Geo-Inf. 8 (4).

Baheti, B., Innani, S., Gajre, S., Talbar, S., 2020. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. p. 1473.

Bittner, K., Liebel, L., Körner, M., Reinartz, P., 2020. Long-short skip connections in deep neural networks for dsm refinement. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLIII-B2-2020, 383.

Bittner, K., Reinartz, P., Korner, M., 2019. Late or earlier information fusion from depth and spectral data? large-scale digital surface model refinement by hybrid-CGAN.

Chen, Z., Li, D., Fan, W., Guan, H., Wang, C., Li, J., 2021. Self-attention in reconstruction bias U-net for semantic segmentation of building rooftops in optical remote sensing images. Remote Sens. 13 (13).

Dai, X., Xia, M., Weng, L., Hu, K., Lin, H., Qian, M., 2023. Multiscale location attention network for building and water segmentation of remote sensing image. IEEE Trans. Geosci. Remote Sens. 61, 1.

Douglas, D., Peucker, T., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartogr.: Int. J. Geogr. Inf. Geovis. 10 (2), 112.

Dukai, B., Ledoux, H., Stoter, J., 2019. A multi-height LoD1 model of all buildings in the netherlands. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-4/W8, 51.

Fischler, M., Bolles, R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 5891.

Gui, S., Qin, R., 2021. Automated LoD-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto. ISPRS J. Photogramm. Remote Sens. 181, 1.

Gui, S., Qin, R., Tang, Y., 2022. Sat2lod2: a software for automated lod-2 building reconstruction from satellite-derived orthophoto and digital surface model. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLIII-B2-2022, 379.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: IEEE International Conference on Computer Vision. p. 2980. http://dx.doi.org/10.1109/ICCV.2017.322.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 770.

Henry, C., Hellekes, J., Merkle, N., Azimi, S.M., Kurz, F., 2021. Citywide estimation of parking space using aerial imagery and OSM data fusion with deep learning and fine-grained annotation. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 479.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J., 2020. Unet 3+: A full-scale connected unet for medical image segmentation.

Kolbe, T.H., Gröger, G., Plümer, L., 2005. Citygml: Interoperable access to 3D city models. In: Geo-Information for Disaster Management. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 883.

Li, M., Lafarge, F., Marlet, R., 2020. Approximating shapes in images with low-complexity polygons. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 8630.

Li, Z., Shan, J., 2022. RANSAC-based multi primitive building reconstruction from 3D point clouds. ISPRS J. Photogramm. Remote Sens. 185, 247.

Li, Z., Wegner, J.D., Lucchi, A., 2019. Topological map extraction from overhead images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 1715.

Liu, Z., Chen, B., Zhang, A., 2020. Building segmentation from satellite imagery using U-net with ResNet encoder. p. 1967.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: International Conference on Learning Representations.

Lussange, J., Yu, M., Tarabalka, Y., Lafarge, F., 2023. 3D detection of roof sections from a single satellite image and application to LOD2-building reconstruction. arXiv:2307.05409.

Mohanty, S.P., Czakon, J., Kaczmarek, K.A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S., et al., 2020. Deep learning for understanding satellite imagery: An experimental survey. Front. Artif. Intell. 3.

Mosinska, A., Márquez-Neila, P., Koziński, M., Fua, P., 2018. Beyond the pixel-wise loss for topology-aware delineation. p. 3136.

Nex, F., Remondino, F., 2012. Automatic roof outlines reconstruction from photogrammetric DSM. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 257.

Pan, X., Yang, F., Gao, L., Chen, Z., Zhang, B., Fan, H., Ren, J., 2019. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. Remote Sens. 11 (8).

Peters, R., Dukai, B., Vitalis, S., van Liempt, J., Stoter, J., 2022. Automated 3D reconstruction of LoD2 and LoD1 models for all 10 million buildings of the netherlands. Photogramm. Eng. Remote Sens. 88 (3), 165.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Schuegraf, P., Fuentes Reyes, M., Xu, Y., Bittner, K., 2023a. Roof3D: A real and synthetic data collection for individual building roof plane and building sections detection. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 971.

Schuegraf, P., Zorzi, S., Fraundorfer, F., Bittner, K., 2023b. Deep learning for the automatic division of building constructions into sections on remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1–16.

Stucker, C., Schindler, K., 2022. ResDepth: A deep residual prior for 3D reconstruction from high-resolution satellite images. ISPRS J. Photogramm. Remote Sens. 183, 560.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. p. 240.

Sun, Z., Zhou, W., Ding, C., Xia, M., 2022. Multi-resolution transformer network for building and road segmentation of remote sensing image. ISPRS Int. J. Geo-Inf. 11 (3).

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. p. 6105.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. UNetFormer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196.

Woo, S., Park, J., Lee, J.-Y., Kweon, I., 2018. CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision. ECCV, p. 3.

Yu, D., Ji, S., Liu, J., Wei, S., 2021. Automatic 3D building reconstruction from multi-view aerial images with deep learning. ISPRS J. Photogramm. Remote Sens. 171, 155.

Zhang, Y., Li, W., Gong, W., Wang, Z., Sun, J., 2020. An improved boundary-aware perceptual loss for building extraction from VHR images. Remote Sens. 12 (7).

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 1848.