SAT2BUILDING: LoD-2 Building Reconstruction from Satellite Imagery using Spatial Embeddings

Philipp Schuegraf[a], Shengxi Gui[b], Rongjun Qin[b], Friedrich Fraundorfer[c], Ksenia Bittner[a]

[a] *German Aerospace Center, Münchener Straße 20 Weßling, 82234, Bavaria, Germany*
[b] *The Ohio State University, W Lane Ave 281 Columbus, 43210, Ohio, USA*
[c] *Graz University of Technology, Innfeldgasse 16 Graz, 8010, Styria, Austria*

Abstract

The reconstruction of buildings in level of detail (LoD)-2, according to the CityGML standard (Kolbe et al., 2005), is an important task in applications like urban planning, environmental simulations and virtual reality. Yet, existing methods either do not accurately separate individual buildings, work primarily on aerial data, or depend on external digital terrain model (DTM). In this work, we present SAT2BUILDING, a method that predicts roof planes, building sections, and building heights in a single fully convolutional neural network (FCN).The network relies on only orthorectified panchromatic imagery and photogrammetric digital surface model (DSM). The three outputs are jointly processed in an LoD-2 reconstruction pipeline that generates seamlessly connected, geometrically accurate and complete, and topologically correct building models. We use spatial embeddings, which enable accurate segmentation of building sections and roof planes from satellite imagery. The model generalizes to data from Bonn, Germany and Lyon, France after being trained on data from Berlin, Germany. The training and test data differ in lighting conditions, architectural styles and ground sampling distances (GSDs). Thorough comparative evaluation shows the superiority of SAT2BUILDING over three baseline methods.

Keywords: Building Reconstruction, Images, Digital Surface Model, Instance

1. Introduction

The modelling of level of detail (LoD)-2 buildings constitutes an important task in the field of geospatial and architectural representation. LoD-2 specifically enhances the geometric complexity of building models beyond basic shapes, incorporating roof structure, which is essential for applications ranging from urban planning and environmental simulations to virtual reality. We consider LoD-2 as it is defined by the CityGML standard (Kolbe et al., 2005). LoD-2 data is often unavailable or of poor quality and hence profits from methods that can generate high quality building geometry.

Optical data includes texture information, which makes it easy to distinguish between various objects above terrain height, such as vegetation and building. Furthermore, it can be used to extract photogrammetric digital surface models (DSMs). It has been leveraged for LoD-2 reconstruction by Nex and Remondino (2012). Their method uses hand-crafted features and is suitable to reconstruct low-complexity buildings. Arefi and Reinartz (2013) also use hand-crafted features from orthorectified image and DSM data. Even though it generates more regular buildings, it lacks robustness to large variations in the input data. Peters et al. (2022) presented an approach that reconstructs buildings in LoD-2 using building sections and lidar point clouds. We define a building section as a component of a building with homogeneous roof type that is visually divisible from the rest of the building. In many cases, this definition coincides with building adresses. Yet, to reconstruct buildings in LoD-2, sub-parts of building sections have to be identified. Hence, region growing is used to create a roof plane (i.e. main planar components) partition of each building section and delineate the lines of intersection between the roof planes. In Li and Shan

(2022), LoD-2 buildings are obtained from normalized point clouds by extracting building primitives and reconstructing them based on a roof type that is determined using the roof point cloud. This approach leads to symmetric and regular LoD-2 models, but is limited by the library of pre-defined roof types. Furthermore, building model generation based on point clouds was improved by several works (Gruen and Wang, 1998; Henricsson, 1998; Gruen, 1998; Sinning-Meister et al., 1996).

Recently, deep learning was introduced to the field of LoD-2 reconstruction. Recently, Alidoost et al. (2019) propose a methodology with two separate neural networks for LoD-2 reconstruction from a single aerial image. One of the networks predict building heights above ground and the other extracts eave-, ridge-, and hip-lines. On top of that, a model-based approach reconstructs LoD-2 buildings. Another work that uses deep learning is keypoint inference by segmentation (KIBS) (Lussange et al., 2023). There, two consecutive Mask-RCNNs (He et al., 2017) are leveraged to first segment roof planes and subsequently predict roof plane corners accompanied by their discrete-valued elevation above ground. This approach shows good accuracy, but it lacks the context of building sections. Both deep learning-based methods above predict building heights from only image data, which is an ill-posed task.

On the contrary, SAT2LOD2 (Gui and Qin, 2021; Gui et al., 2022) uses photogrammetric DSMs together with satellite imagery to reconstruct buildings in LoD-2. They use the images to segment building footprints, which are fed to a geometrical reconstruction pipeline together with the DSM. The reconstruction pipeline uses optimization to obtain regularized LoD-2 models, but does not produce accurate roof geometries. Another limitation is the usage of footprints, which do not allow to reconstruct each building section individually. Gui et al. (2024) overcome this issue by using building section segmentation and the reconstruction pipeline of SAT2LOD2 to get a fine-grained LoD-2 building model.

In our previous work (Schuegraf et al., 2024), we presented PLANES4LOD2, a method which first predicts roof planes and building sections using a single U-shape neural network and afterwards uses conventional vectorization and LoD-2 reconstruction to obtain geometrically and semantically accurate LoD-2 models from aerial imagery and photogrammetric DSM. Even though the obtained results are appealing in the aerial domain, the satellite domain was not explored. Furthermore, PLANES4LOD2 uses an external digital terrain model (DTM) to normalize building heights.

Our first main contribution of this work aims on improved instance segmentation. Since PLANES4LOD2 segments building sections and roof planes based on line features, it is prone to incomplete delineations in the case of occluded or low-contrast object boundaries. On the contrary, Neven et al. (2019) present a method for instance segmentation that is based on spatial embeddings. Spatial embeddings use 2D direction vectors, which point to the center of an instance. We use this idea to segment building sections and roof planes, but add several skip-connections to the respective decoders and add hierarchical skip-connections (Roggiolani et al., 2023) to allow flow of information from the building section to the roof plane task. Overall, we leverage spatial-embedding based instance segmentation in the much more challenging realm of remote sensing, where objects are tiny and often occluded.

Our second main contribution of this work is to predict building heights, sharing the encoder with the instance segmentation network. Building heights are a normalized digital surface model (nDSM), but without vegetation. We leverage a regularization loss, which enforces surface normal consistency of the predicted depth with the ground truth.

Overall, the contributions of this paper are the following:

- Building section and roof plane segmentation based on spatial embeddings.
- Building height estimation with an encoder that is shared with the instance segmentation network.
- Experimental evaluation on test areas in Bonn, Germany and Lyon, France with varying lighting conditions, architectural styles and ground sampling distances (GSDs).
- Comparative evaluation with three baseline methods for LoD-2 reconstruction.

The remainder of this paper is organized as follows: In Section 2, we describe a new method for LoD-2 reconstruction in detail and derive the corresponding equations. In Section 3, we give details about the used datasets, experimental setup and evaluation procedure. In Section 4, we present results to proof the superiority of our method in comparison to baseline methods, which we call SAT2BUILDING. In Section 5, we discuss limitations and possible improvements of the novel SAT2BUILDING method. Finally, Section 6 concludes this paper.

2. Methods

SAT2BUILDING consists of three stages. First, it segments building sections and roof planes by a single U-shape fully convolutional neural network (FCN) alongside LoD-2 building heights from the concatenation of an orthorectified panchromatic image (PAN) and a patch of a photogrammetric DSM. Second, it vectorizes the segments. Third, it generates an LoD-2 model based on the vectorized segments and the building heights.

2.1. Architecture

We utilize ResNet50 (He et al., 2016) as the backbone network (yellow layers in Figure 1) together with two decoders for the building section task, two decoders for the roof plane task and one decoder for the building height task (blue layers in Figure 1). Each of the decoders gradually up-samples the feature map from the last encoder layer, using higher-resolution features maps from the encoder as guidance (indicated by green arrows in Figure 1). This helps in combining fine geometrical details with deep semantic features. Further information is passed from the building section decoders to the roof plane decoders (red arrows in Figure 1). This helps in introducing knowledge from the coarse building section segmentation to the finer roof plane segmentation, which is similarly done for hierarchical plant segmentation in Roggiolani et al. (2023). In all places where multiple feature maps flow to the same skip-connection, we use summation to aggregate them, which is more memory efficient then concatenation.

2.2. Instance Segmentation

We have two instance segmentation sub-tasks, which are building section and roof plane segmentation. For each of the tasks, our network produces three outputs, similar to Neven et al. (2019). We formulate the losses only for a single instance segmentation task, but all losses are computed once for each building section and roof plane.

The first output is the 2D direction vector $o_i \in \mathbb{R}^2$, pointing to the center of an instance $S_k \in \{S_1, S_2, ..., S_K\}$, where k is the cluster index and K is the number of clusters. A vanilla loss function to guide the training of o_i would be

$$\mathcal{L}_{mse} = \sum_{i=1}^{n} \|o_i - \hat{o}_i\|^2, \tag{1}$$

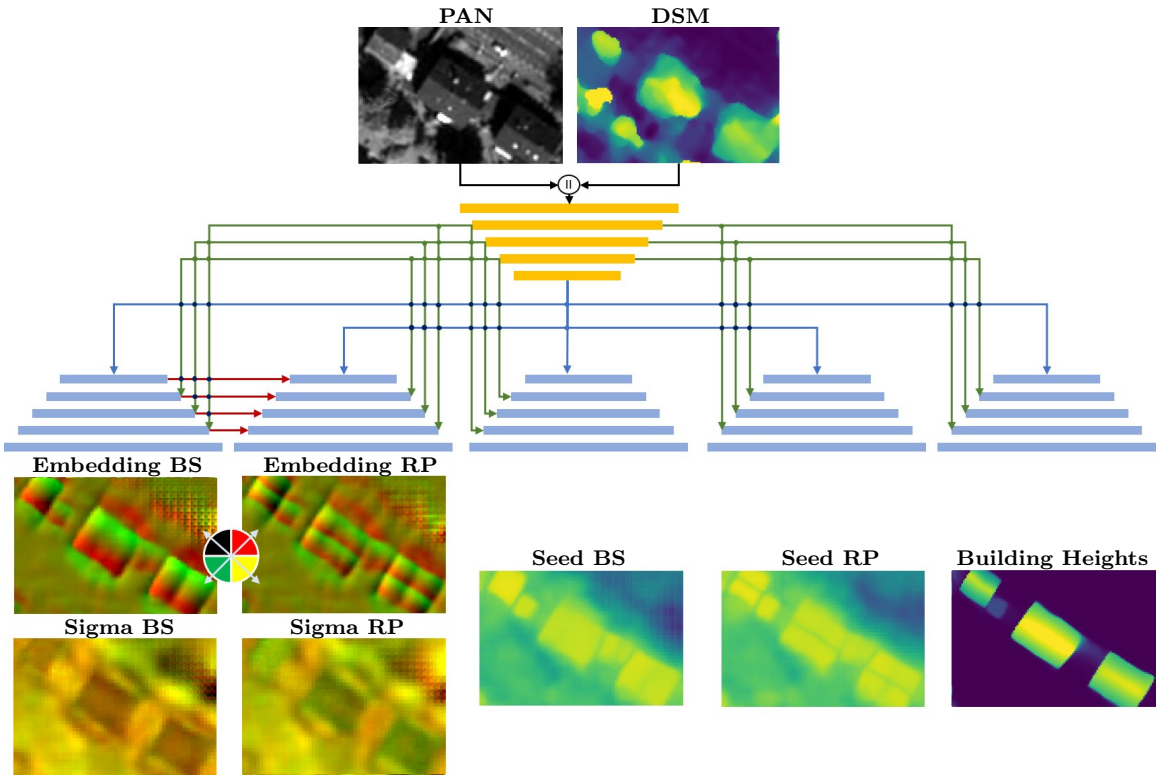


Figure 1: The structure of our proposed network architecture. **BS** and **RP** are abbreviations for "building section" and "roof plane". The yellow and blue rectangles indicate encoder and decoder layers. "‖" is the concatenation operation, the black arrows show the flow of the input data, blue arrows show the flow of the final encoder feature map to the decoder, green arrows are skip connections from the encoder to the decoders, and red arrows are hierarchical skip connections. The circle diagram is the legend of offset directions in the embedding map. The roof planes in Sigma RP are longer than the building sections in Sigma BS. Hence, the sections have a reddish color, whereas the planes have a green tint.

where $\hat{o}_i = C_k - x_i$ for $x_i \in S_k$. $C_k = \frac{1}{n}\sum_{x \in S_k} x$ is the centroid of all pixels x belonging to instance S_k.

However, during inference, we need to determine the centers of the clusters $C = \{C_1, C_2, ..., C_K\}$ and assign all pixels x to one of the cluster centers. A common method to do that is density-based clustering, where the density of the embeddings $e_i = x_i + o_i$ is computed and local maxima are selected as cluster centers. Then, pixels are assigned to the instance with the shortest distance from their corresponding embedding to the center of the instance. But this includes post-processing, which we

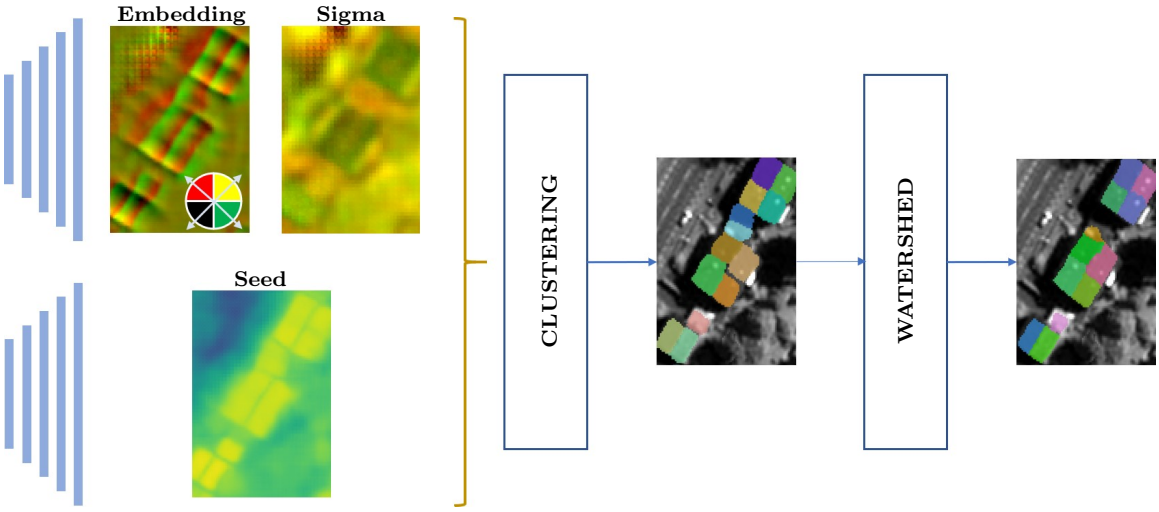


Figure 2: The process of spatial embedding-based roof plane segmentation. Two decoders output spatial embeddings, cluster shape parameters sigma and a seed map. These elements are passed to a simple clustering algorithm. Gaps between instances are closed using the watershed transformation. The definition of a cluster is closely related to the ground truth (see Figure 5).

want to avoid.

Hence, we use the Lovasz-Hinge loss (Yu and Blaschko, 2015) $\mathcal{L}_{lh}$, which is a continuous and differentiable extension of the hinge loss

$$\mathcal{L}_{hinge} = \sum_{k=1}^{K} \sum_{e_i \in S_k} max(\|e_i - C_k\| - \delta, 0), \tag{2}$$

where δ controls the cluster size. Instead of using a fixed cluster size, we replace $\|e_i - C_k\| - \delta$ by $2 \times \phi_k(e_i) - 1$, where

$$\phi_k(e_i) = exp(-t_{kx} \times (e_{ix} - C_{kx})^2 - t_{ky} \times (e_{iy} - C_{ky})^2), \tag{3}$$

and $t_k = exp(10 \times \sigma_k)$. Now, the cluster size is parameterized by $\sigma_k = \frac{1}{|S_k|} \sum_{\sigma_i \in S_k} \sigma_i$ and σ_i is the second output of our network for instance segmentation. Each σ_i has two components, σ_{ix} for the horizontal and σ_{iy} for the vertical direction. This makes

it easier to learn non-square instances. To enforce smoothness of the σ_i, we use

$$\mathcal{L}_{smooth} = \frac{1}{|S_k|} \sum_{\sigma_i \in S_k} \|\sigma_i - \sigma_k\|^2. \tag{4}$$

During inference, we can get a hint on where pixels belonging to instances are located by using the seed s_i of pixel i, which is the third output of our network. The seed score indicates how close a pixel is to the center of any instance. This distance is equivalent to ϕ_k and should be zero in the background bg. Hence, we use the loss function

$$\mathcal{L}_{seed} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{x_i \in S_k\}} \|s_i - \phi_k(e_i)\|^2 + \mathbb{1}_{\{x_i \in bg\}} \|s_i - 0\|^2, \tag{5}$$

where the gradient of $\mathcal{L}_{seed}$ is only computed with respect to s_i. The indicator functions $\mathbb{1}_{\{x_i \in S_k\}}$ and $\mathbb{1}_{\{x_i \in bg\}}$ constitute a mask of pixels belonging to any instance S_k or to the background bg. In the term $\|s_i - \phi_k(e_i)\|^2$, k denotes the number of the cluster S_k that pixel x_i belongs to according to $\mathbb{1}_{\{i \in S_k\}}$. The final loss function for both instance segmentation tasks is

$$\mathcal{L}_{inst} = \mathcal{L}_{lh} + \mathcal{L}_{smooth} + \mathcal{L}_{seed}. \tag{6}$$

To obtain instances during inference time (see Figure 2), we sequentially select the pixels with the highest seed values s_k as the cluster centers C_k. We furthermore select σ_k at those pixels as the sigma value. Then, we assign all pixels i to cluster S_k if

$$e_i \in S_k \iff \phi_k(e_i) < 0.35. \tag{7}$$

Note that the value of $\phi_k(e_i)$ is specific to the center C_k of cluster S_k. The threshold in the implementation of Neven et al. (2019) was 0.5, but we find that the smaller

threshold 0.35 leads to more complete instances in our task. After the assignment of each cluster, we mask out the pixels assigned to cluster S_k and proceed with the new highest seed score. If a cluster contains less then 12 pixels, we discard it as noise. This limit lays far below the 128 pixels used in the original implementation, which is too high for small instances like roof planes in satellite imagery. The clustering process proceeds until there are less than 128 pixels which are not yet clustered. This clustering procedure leaves gaps between adjoining instances open. To solve this issue, we apply a variant of the watershed transformation (Beucher and Meyer, 2018), that makes each instance grow inside the limits of the binary building mask, obtained by thresholding the predicted building heights at 2 m, until it meets another instance.

2.3. Building Height Estimation

To obtain an LoD-2 model, the building height is required. We define the predicted building height as function $f(p) \in \mathbb{R}$ at some pixel $p \in P$, where P is the set of all pixels in an image. This definition implies, that the building height is defined as a pixel-wise, scalar field, where non-building pixels are set to 0. The function f is implemented by the shared encoder (yellow, top in Figure 1) and a decoder (right side in Figure 1). Since the encoder receives not only a DSM, but also a panchromatic image, the sharpness of the building heights can profit from high quality building edge information. The decoder consists of stacked feature fusion modules (FFMs) (Patil et al., 2022), which has a high-resolution feature map from the encoder and a low-resolution feature map from the previous decoder layer or the bottleneck as inputs. Bilinear up-sampling brings the low-resolution feature map to the same resolution as the high-resolution feature map. In parallel, the high-resolution feature map is passed to a residual block. Consecutively, the sum of the output of the residual block and the up-sampled feature map is passed to another residual block. The ground truth

height $\hat{f}(p)$ serves as the learning target in the mean squared error loss

$$\mathcal{L}_{mse} = \frac{1}{|P|} \sum_{p \in P} \|f(p) - \hat{f}(p)\|^2. \tag{8}$$

To improve regularity of the predicted height, we enforce it to have similar normals like the ground truth by utilizing the loss

$$\mathcal{L}_{normal} = \frac{1}{|P|} \sum_{p \in P} \|\nabla_{x,y} f(p) - \nabla_{x,y} \hat{f}(p)\|^2. \tag{9}$$

The final loss for depth estimation is

$$\mathcal{L}_{depth} = \mathcal{L}_{mse} + \mathcal{L}_{normal}. \tag{10}$$

2.4. Vectorization

To reconstruct an LoD-2 model, it is necessary to obtain vectorized roof structure information. We accomplish this by extracting all border pixels of building sections and roof planes. The border pixels are considered vertices and are connected by starting at an initial pixel. From there on, a search finds the closest neighbors among the vertices iteratively along both paths until a cycle exists. We then refine by utilizing the douglas peucker polygon simplification algorithm (Douglas and Peucker, 1973). Hence, we remove pixels that, if excluded, lead to an error of less than 1.0 m with respect to the initial polygon. In CityGML, the representation of an LoD-2 building is a collection of 3D roof planes with a single building ID. We assign building ids to roof planes by selecting the ID of the building section with the highest intersection over union (IoU) with that roof plane. The approach for vectorization and index assignment is outlined in Figure 3.

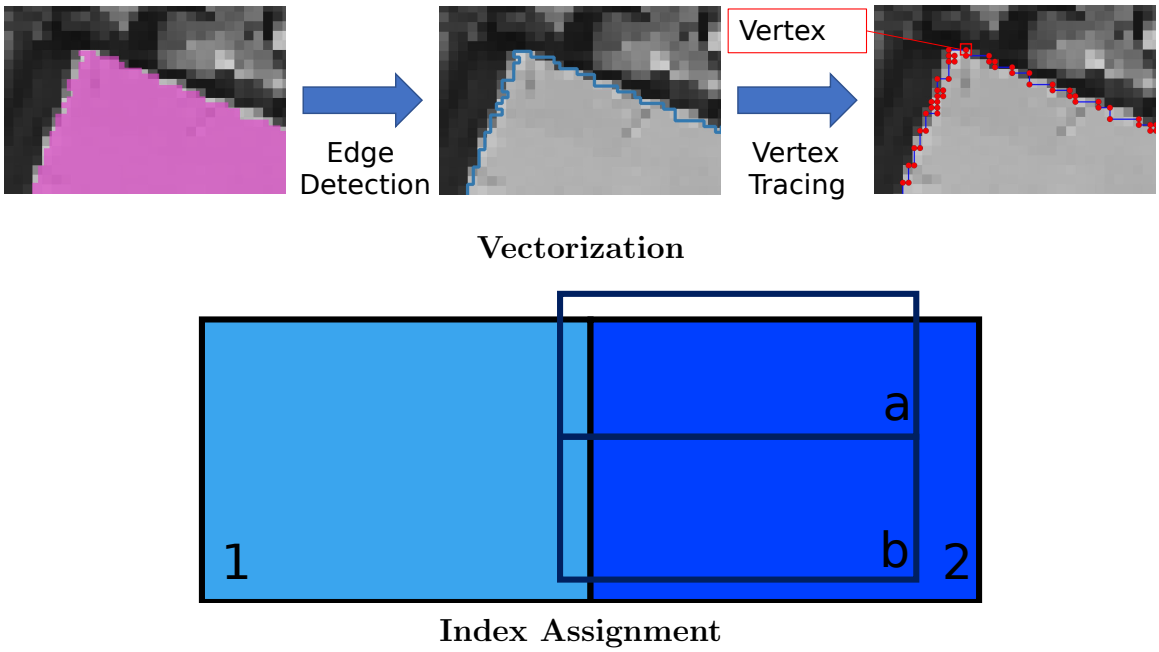


Figure 3: A visualization of the vectorization and index assignment. Building section ids (numbers) are assigned to individual roof planes (letters). Since building section "2" has the higher IoU with both roof planes "a" and "b", the ID "2" will be assigned to them.

2.5. LoD-2 Reconstruction

As the final step, we use random sample consensus (RANSAC) (Fischler and Bolles, 1981) for robustly projecting the 2D roofplanes to 3D. To achieve that, we first consider the scalar building height field. At each pixel that falls inside the area surrounded by the 2D roofplane, we sample the corresponding height value from the building heights. In this way, we obtain a set of 3D points. The 3D points are passed to RANSAC, which estimates plane parameters in a way that is robust to outliers. We project each 2D vertex of the roof plane polygons to 3D by sampling the plane at the vertex location. The above vectorization and LoD-2 reconstruction procedure was originally presented in our previous work (Schuegraf et al., 2024).

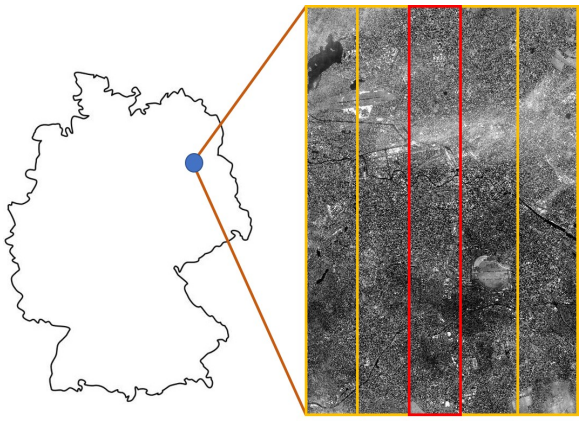


Figure 4: Visualization of the whole Berlin dataset and how it is split into training and validation areas. The red vertical box contains the validation data, the orange boxes contain the training data.

3. Experiments

3.1. Data

For training and validation, we use a World View-1 panchromatic image and photogrammetric DSM of Berlin, Germany of size 30733 × 45999 pixels (see Figure 4). We split the image into five vertical stripes of equal size and use the middle one for validation and the remaining four for training. As the ground truth for building sections, roof planes and building height we use public data provided by the senate of Berlin [1]. In the study area, it contains overall 479,626 building sections with 729,524 roof planes. Some samples of the test area are visualized in Figure 5.

We use two separate datasets for evaluation, one from Bonn, Germany of size 1023 × 896 pixels from Pleíades and the other from Lyon, France of size 1387 × 994 pixels from World View-1. For metric computation, we use public ground truth of both Bonn [2] and Lyon [3] in vector format. The ground truth of Bonn contains 508 building sections with 1,141 roof planes, and that of Lyon 778 building sections with

[1] `https://daten.berlin.de/tags/geodaten`
[2] `https://www.opengeodata.nrw.de/produkte/geobasis`
[3] `https://data.grandlyon.com/`

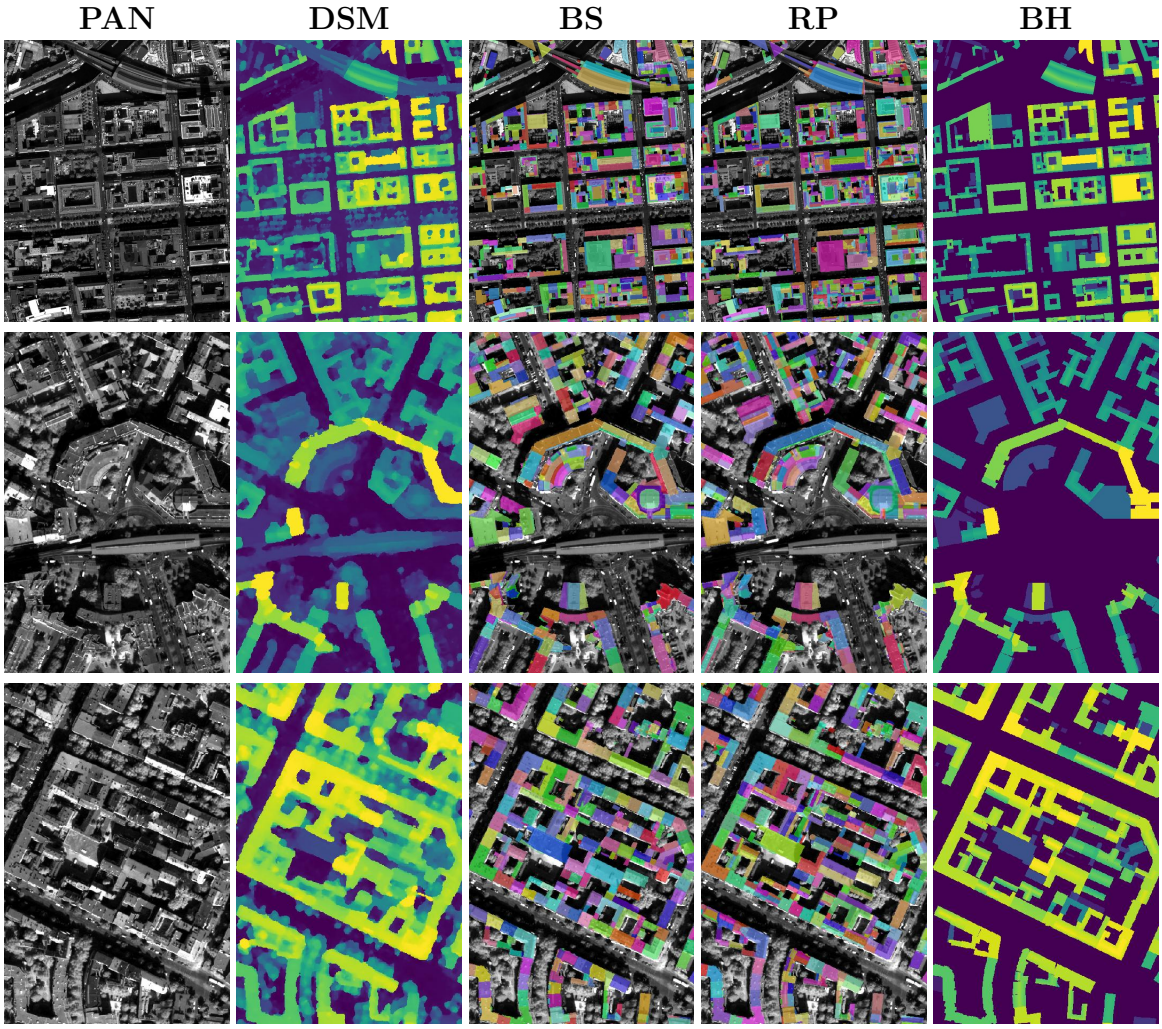


Figure 5: Parts from our training data in Berlin. **PAN** abbreviates panchromatic image, **BS** building sections, **RP** roof planes and **BH** building heights.

2,575 roof planes. In Figure 6, samples from Bonn and Lyon are visualized. In Lyon, a building section contains on average more than 3 roof planes, which is much higher than the ratio of ~ 2 in Bonn and ~ 1.5 in Berlin. Looking at the bottom row of Figure 6, it becomes clear that the ground truth in Lyon contains more details than the other areas.

During training, we crop patches of size 256×256 pixels without overlap. We do random window shifting of up to 256 pixels in horizontal and vertical direction

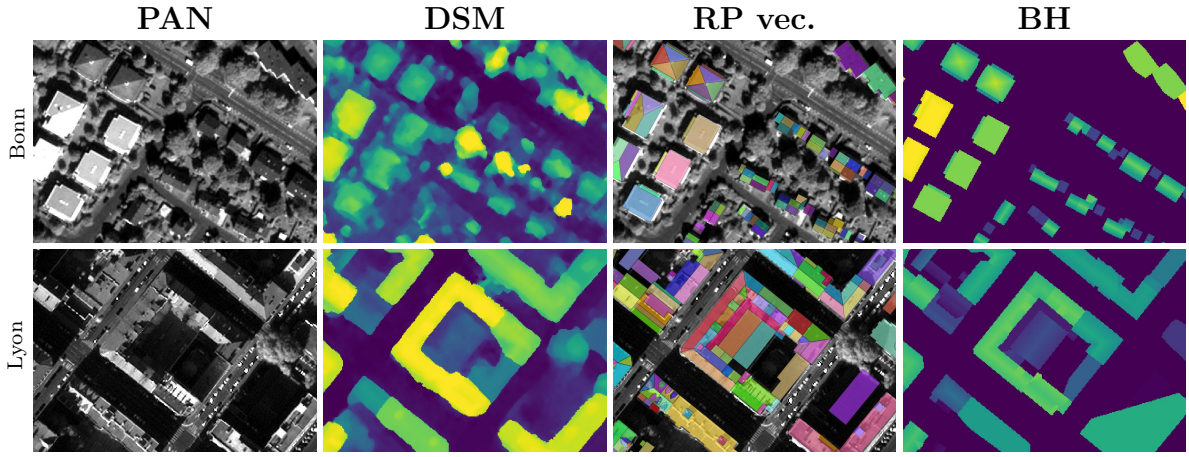


Figure 6: Parts from our test data in Bonn (top row) and Lyon (bottom row). **RP vec.** abbreviates vectorized roof planes.

to increase the data diversity during training. In the validation phase, patches of size 256 × 256 pixels are cropped without overlap. While testing, the crop size is also 256 × 256 pixels and the overlap is 128 pixels horizontally and vertically. The network predicts per-patch and a large map is created by averaging the patches at the overlapping areas.

The data from Berlin and Lyon has GSD 0.5 m, whereas that of Bonn has GSD 0.7 m. Hence, we up-sample the data from Bonn to GSD 0.5 m. We generate all ground truth in GSD 0.5 m. During evaluation, we use two kinds of ground truth, which is raster ground truth of building heights and vector ground truth of roof planes. The building heights are rasterized LoD-2 models. They contain height values for each pixel in meters above ground. As opposed to photogrammetric DSM, building heights do neither contain trees nor terrain information.

3.2. Training Details

We use random initialization of the network parameters and train them using Adam optimizer (Kingma and Ba, 2017) with learning rate 0.0002, momenta 0.5 and 0.999. The model is trained for 300 epochs and the learning rate is multiplied by 0.1

after the 100th and 200th epoch. We use batch size 8 and combine the loss functions from Equations (6) and (10) to the final multi-task loss

$$\mathcal{L}_{total} = \mathcal{L}_{inst,bs} + \mathcal{L}_{inst,rp} + \mathcal{L}_{depth}. \tag{11}$$

3.3. Evaluation Metrics

To quantify the performance of the trained models, we evaluate both the vectorized roof planes in 2D and the rasterized predicted LoD-2 model in 3D. One commonly used metric for segmentation is the IoU. Since it doesn't take into account individual instances, it is not suitable for roof plane segmentation. On the other hand, the common objects in context (COCO) metrics are too challenging for tiny objects like roof planes in satellite imagery. Hence, we provide the new metric

$$IoU^{gt}_{inst} = \frac{1}{|\hat{P}|} \sum_{\hat{p} \in \hat{P}} \max_{p \in P} IoU(p, \hat{p}), \tag{12}$$

where $p \in P$ is a predicted polygon and $\hat{p} \in \hat{P}$ is a ground truth polygon. In IoU^{gt}_{inst}, we iterate over the ground truth polygons and select the respective predicted polygon with the highest IoU. Then, these IoUs are averaged.

For the evaluation of the 3D models, we rasterize their height values and use the root-mean-squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_i |\hat{h}_i - h_i|^2}{N}}, \tag{13}$$

where i is a specific pixel, N is the number of pixels, $\hat{h}_i$ is the ground truth height at pixel i and h_i is the predicted height at pixel i. Furthermore, we use mean absolute

error (MAE)

$$\mathrm{MAE} = \frac{\sum_i |\hat{h}_i - h_i|}{N}. \tag{14}$$

for evaluation. Note that the MAE is less sensitive to outliers than the RMSE. To gain more insight into the obtained results, we also carry out qualitative inspection on both the roof plane polygons and the rasterized LoD-2 model.

3.4. Experiments

Several experiments are done to show the superiority of SAT2BUILDING over competing methods. As the reference method, we use PLANES4LOD2 (Schuegraf et al., 2024), which was originally trained for LoD-2 reconstruction of aerial imagery. To improve comparability, we re-train PLANES4LOD2 on the same satellite data of Berlin that we use for SAT2BUILDING. For PLANES4LOD2, we use an external DTM together with the DSM to derive building heights instead of predicting them. The next experiment is the method of Gui et al. (2024) (SAT2LOD2-LineSep). SAT2LOD2-LineSep requires normalized building height information and building sections. Since SAT2LOD2 does not have a specification on the source of the building heights, we use the building height from our proposed method (SAT2BUILDING) and the building sections from PLANES4LOD2, which are obtained based on the prediction of separation lines between sections. Moreover, we feed the building height and sections from SAT2BUILDING to the method of Gui et al. (2024) and call that experiment SAT2LOD2-Embed, because the input building sections are obtained using spatial embedding-based instance segmentation. Using the building height from SAT2BUILDING lays the focus on the comparison of using roof planes for LoD-2 reconstruction, since both SAT2LOD2 and SAT2BUILDING use the same building heights. We compare the above approaches to SAT2BUILDING.

To showcase the effectiveness of using a shared encoder for both instance seg-

Table 1: Quantitative results of comparison between a single shared, and two separate encoders for two test areas. ↑ indicates that higher values are superior, ↓ indicates that lower values correspond to higher accuracy.

Test Area	Shared Encoder	IoU^{gt}_{inst} ↑	MAE ↓	RMSE ↓
Bonn		0.300	0.56 m	1.92 m
Bonn	X	**0.323**	**0.53** m	**1.83** m
Lyon		0.199	1.81 m	5.22 m
Lyon	X	**0.205**	**1.74** m	**4.98** m

mentation and height estimation, we compare that to a setting with two separate networks without inter-connection.

4. Results

4.1. Quantitative Results

In this subsection, we analyse the quantitative results of the experimental study. In Table 1 we can see that it is advantageous to use a shared encoder instead of two separate networks. The unified network performs better in both instance segmentation and LoD-2 model geometrical accuracy on both test areas, showing that one encoder can effectively learn features that are more useful for both tasks as compared to two separate encoders. We attribute this advantage to the regularizing effect of the multi-task setting. Since both models are trained on data from Berlin that has different lighting conditions and architectural styles, the results highlight their capability to generalize to unseen data.

In Table 2, we observe the quantitative results of SAT2BUILDING with a shared encoder in comparison to three baseline methods. The comparison between the related methods SAT2LOD2-SepLine and SAT2LOD2-Embed shows that instance segmentation based on spatial embeddings is either an equally as good (Bonn) or a better

Table 2: Comparative results on Bonn and Lyon. ↑ indicates that higher values are superior, ↓ indicates that lower values correspond to higher accuracy.

Test Area	NAME	IoU^{gt}_{inst} ↑	MAE ↓	RMSE ↓
Bonn	SAT2LOD2-SepLine	-	0.89 m	2.66 m
Bonn	SAT2LOD2-Embed	-	0.88 m	2.67 m
Bonn	PLANES4LOD2	0.1826	0.64 m	2.14 m
Bonn	SAT2BUILDING	**0.323**	**0.53** m	**1.83** m
Lyon	SAT2LOD2-SepLine	-	4.00 m	8.60 m
Lyon	SAT2LOD2-Embed	-	3.16 m	7.50 m
Lyon	PLANES4LOD2	0.162	2.66 m	6.35 m
Lyon	SAT2BUILDING	**0.205**	**1.74** m	**4.98** m

(Lyon) basis for the LoD-2 reconstruction using the SAT2LOD2 method. The advantage of spatial embeddings over separation lines becomes particularly clear in Lyon, where the buildings are densely built. A higher density of buildings and a larger quantity of buildings with joint borders makes it more critical to discern building sections. Furthermore, separating buildings based on a thin line is very challenging in satellite imagery, as compared to aerial imagery with smaller GSDs below 0.3 m.

The comparison between SAT2LOD2-SepLine and PLANES4LOD2 shows that the LoD-2 reconstruction becomes geometrically more accurate if it focuses on reconstructing roofs based on individual roof planes instead of primitives. PLANES4LOD2 outperforms SAT2LOD2-SepLine with a large margin, remarkably in Lyon. The complex building roofs in Lyon make it even more important to accurately reconstruct each single roof plane.

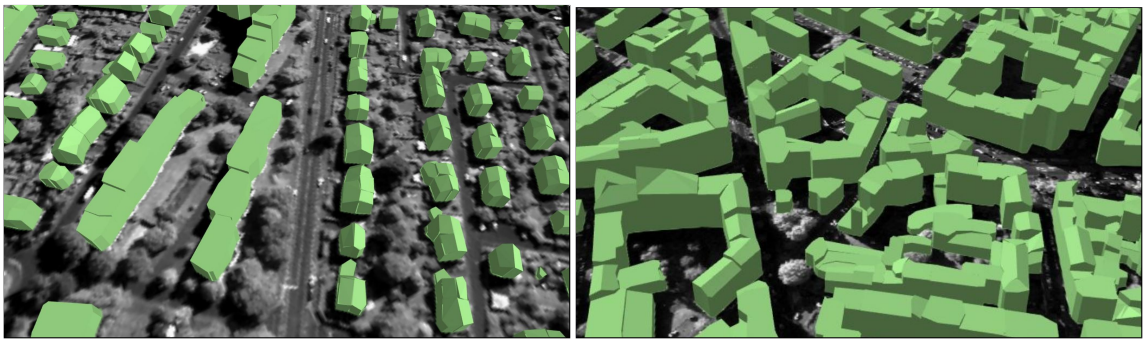
Comparing PLANES4LOD2 and SAT2BUILDING highlights the advantages of using spatial embeddings over separation lines for instance segmentation of roof planes. SAT2BUILDING consistently surpasses PLANES4LOD2 across all metrics

and in both test areas. This difference in performance, particularly the larger gap in IoU^{gt}_{inst} observed in Lyon compared to Bonn, can be attributed to the GSD of each area. The GSD in Lyon is smaller than that in Bonn, which has a GSD of 0.7 m. A larger GSD makes it challenging to segment the thin lines that separate adjoining roof planes, thus reducing the performance of PLANES4LOD2. This happens, because PLANES4LOD2 uses the separation line to segment roof planes and this separation line becomes more unclear as the GSD increases. In contrast, SAT2BUILDING, which segments instances based on spatial embeddings, groups pixels by the center of each instance. This approach is more robust on larger GSD images, as it does not rely on clearly visible separation lines, giving SAT2BUILDING an advantage over PLANES4LOD2 on lower-resolution satellite imagery. Since the IoU^{gt}_{inst} metric is influenced by the accuracy of roof plane separation, SAT2BUILDING demonstrates superior performance compared to PLANES4LOD2 on this measure.

Another advantage of SAT2BUILDING is its independence from external DTM information. The strong performance of SAT2BUILDING is achieved using its own predicted building height map, whereas PLANES4LOD2 includes external terrain information to obtain heights above ground. The improved values of MAE and RMSE indicate that using the predicted building heights has no negative effect on the performance as compared to using an external DTM.

4.2. Qualitative Results

In Figure 7, it can be seen that SAT2BUILDING generates geometrically accurate building models at LoD-2. Even under highly challenging conditions like large shadows, high GSDs and complex building structures, SAT2BUILDING correctly identifies individual roof planes. Nevertheless, in some places, roof planes are misaligned. The LoD-2 reconstruction pipeline can sometimes estimate incorrect plane parameters, if



Bonn **Lyon**

Figure 7: A 3D visualization of the results on our two test sites in Bonn and Lyon.

RANSAC randomly selects an inadequate subset of points on the plane.

As we inspect Figure 8, we see that SAT2BUILDING reconstructs roof more similar to the ground truth than all other tested methods in Bonn. In the middle of the upper sample, the positive effect of the improved instance segmentation leads to correctly generated gable roofs. PLANES4LOD2 incorrectly generates flat roofs, which is caused by missing separation lines. SAT2LOD2-Embed and SAT2LOD2-SepLine produce regularized buildings, but they often do not accurately reflect the structure of the roof. SAT2LOD2 reconstructs building roofs based on fixed roof templates, which are often incorrectly inferred in case of complex buildings. Furthermore, SAT2LOD2 does not keep seamless neighboring relations while refining the boundaries of building sections, which leads to incorrect gaps between them. On the other hand, SAT2BUILDING and PLANES4LOD2 do not have such a gap because they refine outlines of adjoining building sections and roof planes jointly. In the lower example of Figure 8, SAT2BUILDING is the only method that gets the pyramid shape of the four squared building in the middle correct. Since we use training data from a public source, which contains many inconsistencies, this can sometimes cause incorrect predictions.

In Figure 10 we can see that SAT2BUILDING is more accurate than the other

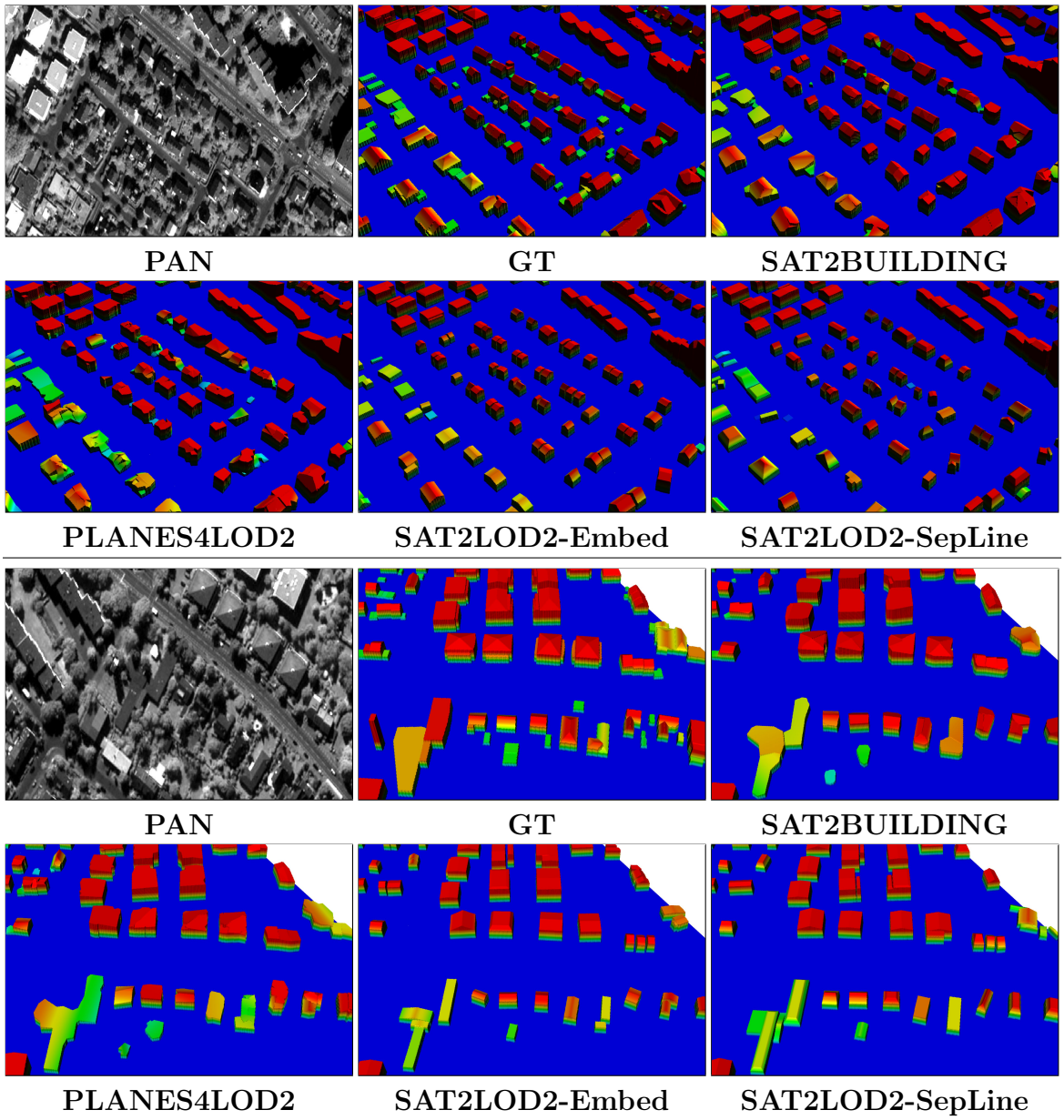


Figure 8: Visual results of the comparative study from Bonn.

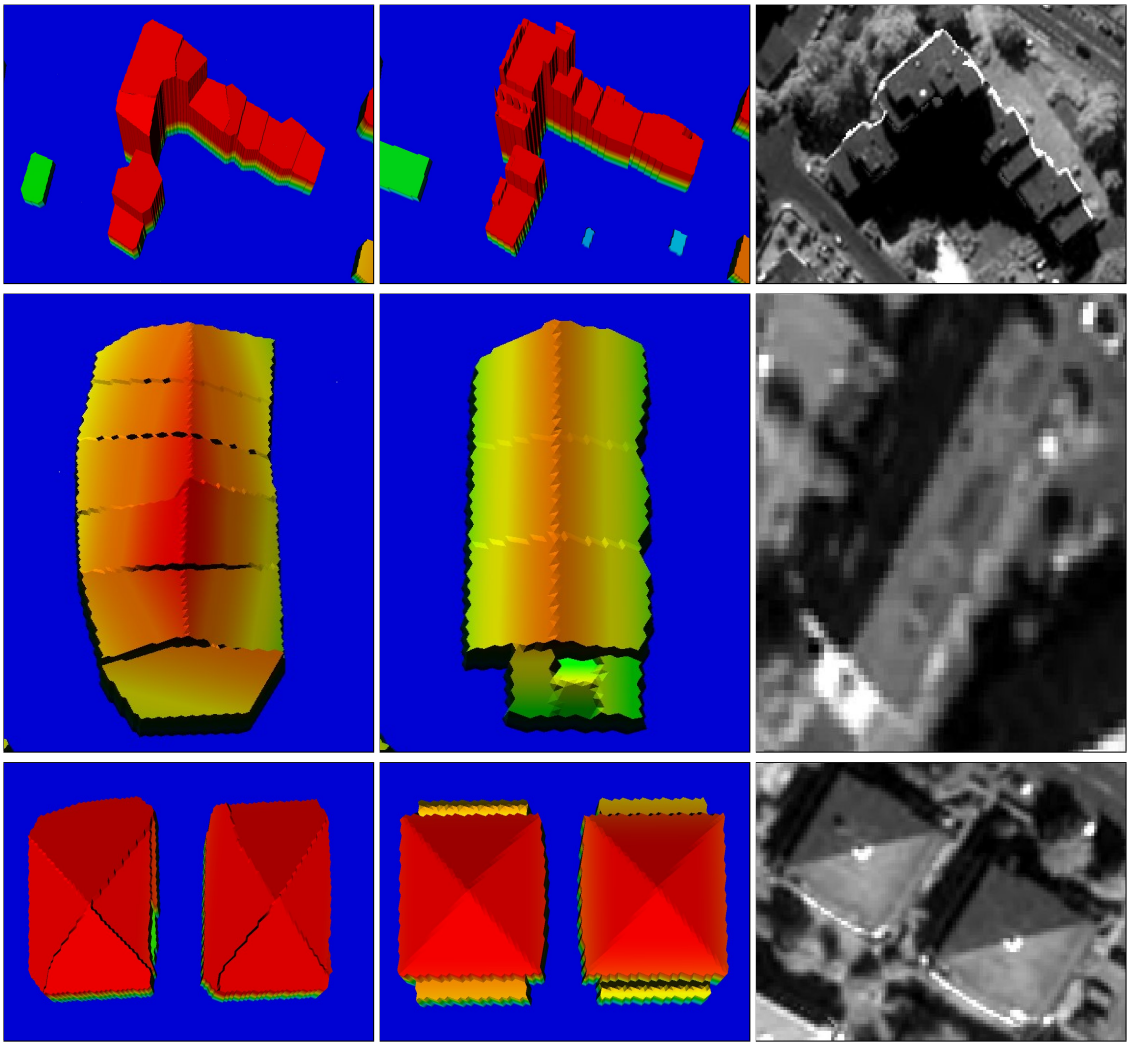


Figure 9: Detailed visual results of SAT2BUILDING.

three methods even in highly complex scenarios like the city of Lyon. For example, the hipped roof in the middle of the scene is only accurately reconstructed by SAT2BUILDING. Furthermore, SAT2LOD2 generates even more incorrect gaps than in Figure 8.

We take a closer look on several buildings in Figure 9. In the first row, a complex building structure is visualized as reconstructed by our SAT2BUILDING approach and the corresponding ground truth. While SAT2BUILDING is capable of reconstructing the overall structure of this building accurately, roof details are not reconstructed according to the ground truth. Those details are small and are hardly recognizable in the panchromatic image. Even though in all three detailed examples the overall structure of the roof is accurately reconstructed, several details are incorrect or geometrically distorted. Those errors can be mostly explained by the missing information (low GSD, shadows) in the panchromatic image. In the third row in Figure 9, the position where the four roof planes predicted by SAT2BUILDING intersect is not perfectly central. This position is largely influenced by the estimated plane parameters, which are based on the predicted building heights. Since the elevation profile in the building heights depend strongly on the DSM, inaccuracies in the DSM lead to topological errors in the prediction of SAT2BUILDING.

5. Discussion

Although SAT2BUILDING outperforms the other methods in quantitative evaluation, the metric IoU^{gt}_{inst} does not exceed 0.323 in Bonn and 0.205 in Lyon. Particularly the value for Lyon is very low. We can observe in the third column bottom row of Figure 6, that the roof plane ground truth in Lyon contains many details, that are very hard to recognize in the panchromatic image (first column) and impossible to detect in the DSM (second column). On the other hand, SAT2BUILDING extracts

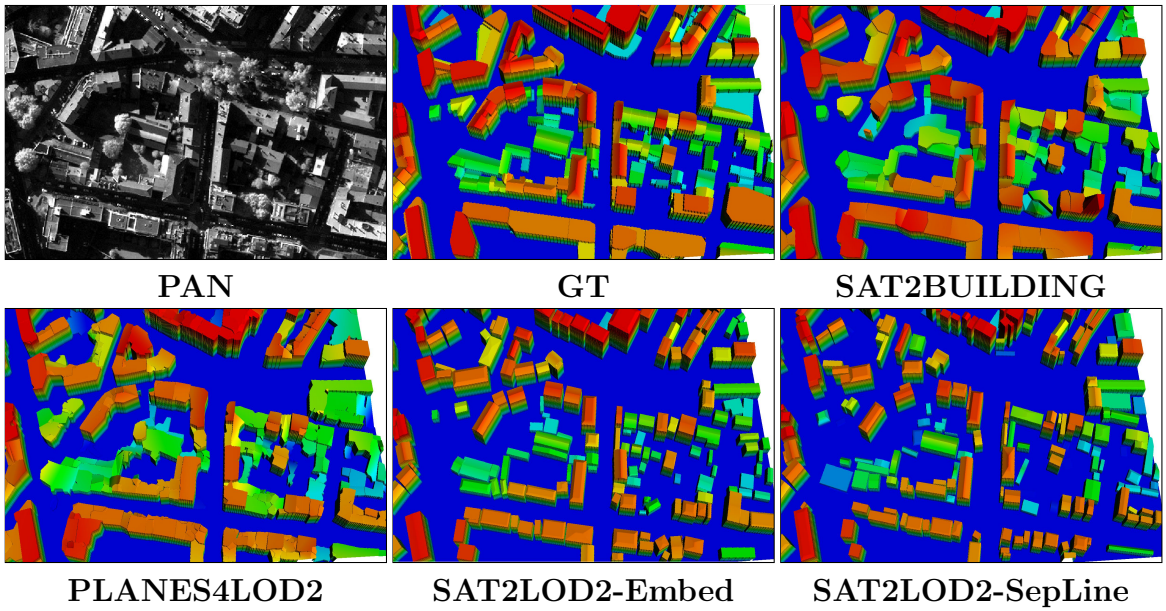


Figure 10: Visual results of the comparative study from Lyon.

roof plane polygons, representing larger planar structures of building sections. But the IoU^{gt}_{inst} metric averages IoU scores across ground truth instances, leading to a low score.

Moreover, Table 2 shows much higher MAE and RMSE in Lyon than in Bonn. The evaluation of the generated 3D models was done based on rasterized height maps. Since our test area in Bonn is more sparse, it leads to a lot of background pixels, which causes better metrics. In Lyon on the other hand, a high building density, with buildings at various heights, and complex building structures have more room for error and cause worse 3D metrics.

Furthermore, we want to highlight the importancy of including photogrammetric DSM data as an input to our proposed model. DSM is a substantial hint on the desired 3D building model. But without the height context of the DSM, our model would not be able to extract a meaningful building height field. Other than the DSM, we only present orthorectified panchromatic imagery to the model, which includes no information about the number of floors of the buildings. One could claim

that shadow information is present in orthorectified imagery and that shadows imply height information. But shadow-behavior is dependent on the season and geographic location.

It is also obvious, that neither SAT2BUILDING nor the comparing methods can achieve very good accuracy, as some applications might require. We attribute this to the quality of the satellite imagery in this study, since it was shown in Schuegraf et al. (2024) that, given higher-resolution aerial imagery with GSD of 0.3, the accuracy of the resulting LoD-2 models are vastly better. Therefore, LoD-2 reconstruction requires improved input data, if very high accuracy is demanded.

6. Conclusion

We presented SAT2BUILDING, a novel method for level of detail (LoD)-2 reconstruction based on the segmentation of main planar roof components. Our method utilizes deep learning and conventional methods to build a complete 3D reconstruction workflow, only based on panchromatic satellite imagery and photogrammetric digital surface model (DSM). Our method predicts normalized building heights, which makes it independent from external terrain information. SAT2BUILDING leverages spatial embeddings for robust roof plane segmentation. The resulting LoD-2 model is geometrically accurate, even when facing difficulties such as high ground sampling distances (GSDs) of 0.5 m to 0.7 m, large shadows, densely built areas, and complex roof structures. We showed how SAT2BUILDING improves in comparison to existing methods, obtaining considerable performance gains. Furthermore, SAT2BUILDING generalizes well when evaluated on cities with different lighting conditions, architectural styles, and GSDs then are contained in the training area.

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 Sensing 11, 2219 ff.

Arefi, H., Reinartz, P., 2013. Building reconstruction using dsm and orthorectified images. Remote Sensing 5, 1681 ff.

Beucher, S., Meyer, F., 2018. The morphological approach to segmentation: The watershed transformation. Mathematical Morphology in Image Processing .

Douglas, D., Peucker, T., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica: The International Journal for Geographic Information and Geovisualization 10, 112 ff.

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

Gruen, A., 1998. Tobago - a semi-automated approach for the generation of 3-d building models. ISPRS Journal of Photogrammetry and Remote Sensing 53, 108–118.

Gruen, A., Wang, X., 1998. Cc-modeler: a topology generator for 3-d city models. ISPRS Journal of Photogrammetry and Remote Sensing 53, 286–295.

Gui, S., Qin, R., 2021. Automated lod-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto. ISPRS Journal of Photogrammetry and Remote Sensing 181, 1 ff.

Gui, S., Qin, R., Tang, Y., 2022. Sat2lod2: A software for automated lod-2 building reconstruction from satellite-derived orthophoto and digital surface model. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B2-2022, 379 ff.

Gui, S., Schuegraf, P., Bittner, K., Qin, R., 2024. Unit-level lod2 building reconstruction from satellite-derived digital surface model and orthophoto. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences .

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. IEEE International Conference on Computer Vision , 2980 ff.doi:10.1109/ICCV.2017.322.

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

Henricsson, O., 1998. The role of color attributes and similarity grouping in 3-d building reconstruction. Computer Vision and Image Understanding 72, 163–184.

Kingma, D., Ba, J., 2017. Adam: A method for stochastic optimization arXiv:1412.6980.

Kolbe, T.H., Gröger, G., Plümer, L., 2005. CityGML: Interoperable Access to 3D City Models. Springer Berlin Heidelberg, Berlin, Heidelberg. p. 883 ff.

Li, Z., Shan, J., 2022. Ransac-based multi primitive building reconstruction from 3d point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 185, 247 ff.

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.

Neven, D., Brabandere, B., Proesmans, M., Van Gool, L., 2019. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth , 8829 ff.

Nex, F., Remondino, F., 2012. Automatic roof outlines reconstruction from photogrammetric dsm. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences , 257 ff.

Patil, V., Sakaridis, C., Liniger, A., Van Gool, L., 2022. P3depth: Monocular depth estimation with a piecewise planarity prior , 1610 ff.

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. Photogrammetric Engineering and Remote Sensing 88, 165 ff.

Roggiolani, G., Sodano, M., Guadagnino, T., Magistri, F., Behley, J., Stachniss, C., 2023. Hierarchical approach for joint semantic, plant instance, and leaf instance segmentation in the agricultural domain , 9601 ff.

Schuegraf, P., Shan, J., Bittner, K., 2024. Planes4lod2: Reconstruction of lod-2 building models using a depth attention-based fully convolutional neural network. ISPRS Journal of Photogrammetry and Remote Sensing 211, 425 ff.

Sinning-Meister, M., Gruen, A., Dan, H., 1996. 3d city models for caad-supported analysis and design of urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 51, 196–208.

Yu, J., Blaschko, M., 2015. Learning submodular losses with the lovasz hinge , 1623 ff.