Scan2LoD3: Reconstructing semantic 3D building models at LoD3 using ray casting and Bayesian networks

Olaf Wysocki[1] , Yan Xia [1,2] , Magdalena Wysocki [1] , Eleonora Grilli [3] ,
Ludwig Hoegner [1,4] , Daniel Cremers [1,2] , Uwe Stilla [1]

[1] Technical University of Munich, [2] University of Oxford, [3] Bruno Kessler Foundation,
[4] Munich University of Applied Sciences

(olaf.wysocki, yan.xia, magdalena.wysocki, ludwig.hoegner, cremers, stilla)@tum.de
elegrilli5@gmail.com

Abstract

Reconstructing semantic 3D building models at the level of detail (LoD) 3 is a long-standing challenge. Unlike mesh-based models, they require watertight geometry and object-wise semantics at the façade level. The principal challenge of such demanding semantic 3D reconstruction is reliable façade-level semantic segmentation of 3D input data. We present a novel method, called Scan2LoD3, that accurately reconstructs semantic LoD3 building models by improving façade-level semantic 3D segmentation. To this end, we leverage laser physics and 3D building model priors to probabilistically identify model conflicts. These probabilistic physical conflicts propose locations of model openings: Their final semantics and shapes are inferred in a Bayesian network fusing multimodal probabilistic maps of conflicts, 3D point clouds, and 2D images. To fulfill demanding LoD3 requirements, we use the estimated shapes to cut openings in 3D building priors and fit semantic 3D objects from a library of façade objects. Extensive experiments on the TUM city campus datasets demonstrate the superior performance of the proposed Scan2LoD3 over the state-of-the-art methods in façade-level detection, semantic segmentation, and LoD3 building model reconstruction. We believe our method can foster the development of probability-driven semantic 3D reconstruction at LoD3 since not only the high-definition reconstruction but also reconstruction confidence becomes pivotal for various applications such as autonomous driving and urban simulations.

1. Introduction

Reconstructing detailed semantic 3D building models is a fundamental challenge in both photogrammetry [10] and computer vision [39]. Recent developments have shown that reconstruction using 2D building footprints and aerial observations provides building models up to level of detail (LoD) 2 [10, 20, 34], which are characterized by complex roof shapes but display planar façades. Owing to their watertightness and object-oriented modeling, such models have found many applications [4] and are now ubiquitous, as exemplified by around 140 million open access building models in the United States, Switzerland, and Poland [1].

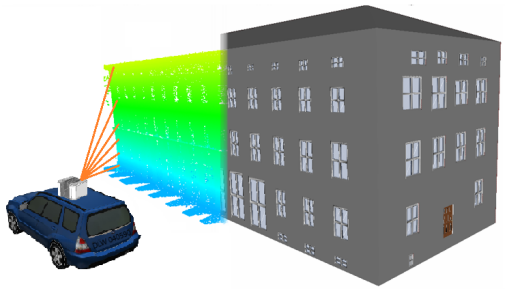


Figure 1. Scan2LoD3: Our method reconstructs detailed semantic 3D building models; Its backbone is laser rays' physics providing geometrical cues enhancing semantic segmentation accuracy.

However, reconstructing façade-detailed semantic LoD3 building models remains an open challenge. Currently, LoD3-specific façade elements, such as windows and doors, are frequently manually modeled [5, 43]; yet at-scale, automatic LoD3 reconstruction is required by numerous applications ranging from simulating flood damage [2], estimating heating demand [27], calculating façade solar potential [47] to testing automated driving functions [36].

The best data source for semantic LoD3 façade modelling [54] appears to be mobile mapping data, as the last years have witnessed a growth in mobile mapping units

[1] https://github.com/OloOcki/awesome-citygml

yielding accurate, dense, street-level image and point cloud measurements. Yet, typically such data necessities robust, accurate, and complete semantic segmentation before it can be applied to semantic reconstruction. In the past decade, various learning-based façade-level 3D point cloud segmentation solutions have achieved promising performance [8, 23, 49]. However, they have limited accuracy of up to 40% [23] when working on translucent (e.g., windows) and label-sparse (e.g., door) objects. Methods based on intersections of laser rays with 3D models are used to improve the accuracy [40, 49]. However, such methods are prone to errors due to the limited semantic information [40] and field-of-view obstacles, such as window blinds [49]. Another approaches employ images for façade segmentation and achieve high performance [22, 33]; yet, their direct application for 3D façade segmentation is limited chiefly owing to the 2D representation [16, 30].

In this paper, we present a novel ray-casting-based multimodal framework for semantic LoD3 building model reconstruction named Scan2LoD3. In contrast to previous methods, we combine multimodalties instead of relying on single modality [40]; and we fuse modalities using their state probabilities, as opposed to mere binary fusion [49]. The key to maintaining geometric detail is to utilize laser ray physical intersections with vector priors to find probability-quantified model conflicts in a Bayesian network, as highlighted in Figure 1; we list our contributions as follows:

- A probabilistic visibility analysis using mobile laser scanning (MLS) point clouds and semantic 3D building models, enabling detection of detailed conflicts by non-binary probability masks and L2 norm;
- A Bayesian network approach for the late fusion of multimodal probability maps enhancing 3D semantic segmentation at the façade-level;
- An automatic, watertight reconstruction of LoD3 models with façade elements of windows and doors compliant with the CityGML standard [9];
- An open LoD3 reconstruction benchmark comprising LoD3 and façade-textured LoD2 building models, and façade-level semantic 3D MLS point clouds [2].

2. Related work

The key to reconstructing the LoD3 building model is to achieve an accurate 3D façade segmentation. Here, we provide insights into visibility- and learning-based methods.

Visibility analysis using ray casting and 3D models. In the context of 3D building models, ray casting from the sensor's origin yields deterministic information about measured, unmeasured, and unknown model parts [24, 41], but also provides geometric cues, so-called conflicts, for the façade elements reconstruction [13, 40, 49]. For example, Tuttas et al. [40] exploit the fact that laser scanning rays traverse glass objects to identify building openings: They assume that the intersection points of rays and found building planes indicate the position of windows, which are then reconstructed by minimum bounding boxes. Hoegner & Gleixner [13] pursue this idea using mobile laser scanning and, besides rays intersections, they analyze empty regions in point clouds. Due to the methods' assumption that each visible opening is a window, they do not distinguish between other openings, such as doors or underpasses. To overcome this issue, Wysocki et al. [50] propose the conflict classification method, which infers the semantics of ray intersections with 3D models using 2D vector maps to detect and reconstruct building underpasses. However, conflict-based methods are prone to occlusions and are limited in identifying openings that are concealed by non-translucent objects, such as blinds.

Machine learning in 3D façade reconstruction. Early learning-based façade segmentation methods [6, 19, 33, 39, 42] typically rely on ubiquity of 2D image façade segmentation datasets and represent façade elements as 2D objects (discussed in detail in [25]). Recent works utilize well-established 2D image-based neural networks to identify façade elements in images and then project them onto 3D point clouds or their derivatives, such as 3D models [12, 16, 29, 30]. However, these methods frequently assume full point cloud coverage of buildings and correctly co-referenced multiple image observations from various angles. For example, Huang et al. [16] propose a method employing FC-DenseNet56 [17], trained with ortho-rectified façade images, to recognize façade openings. The labels are projected onto LoD2 building model, which is reconstructed from a drone-based photogrammetric point cloud. The projected window and door labels are approximated to bounding boxes, which cut openings in LoD2 solids, thereby upgrading 3D models to LoD3.

An alternative strategy concentrates on direct 3D façade modeling from laser scanning point clouds since MLS point clouds provide detailed and accurate depth information [53]. Recently, it has been demonstrated that great advances of point-wise, learning-based methods [31, 55] are applicable in the context of 3D façade segmentation [8, 23], where an early fusion of geometric features into DGCNN [45] enhances façade segmentation accuracy. Nevertheless, sparsely represented classes, such as windows and doors, remain challenging [23]. This issue is further exacerbated by the lack of comprehensive 3D façade-level training and validation data: to the best of our knowledge, no 3D façade-level reconstruction benchmark includes textures, point clouds, and ground-truth LoD3 models [51].

One recent work [49] pursues the idea of combining ge-

[2] https://sites.google.com/view/olafwysocki/papers/scan2lod3

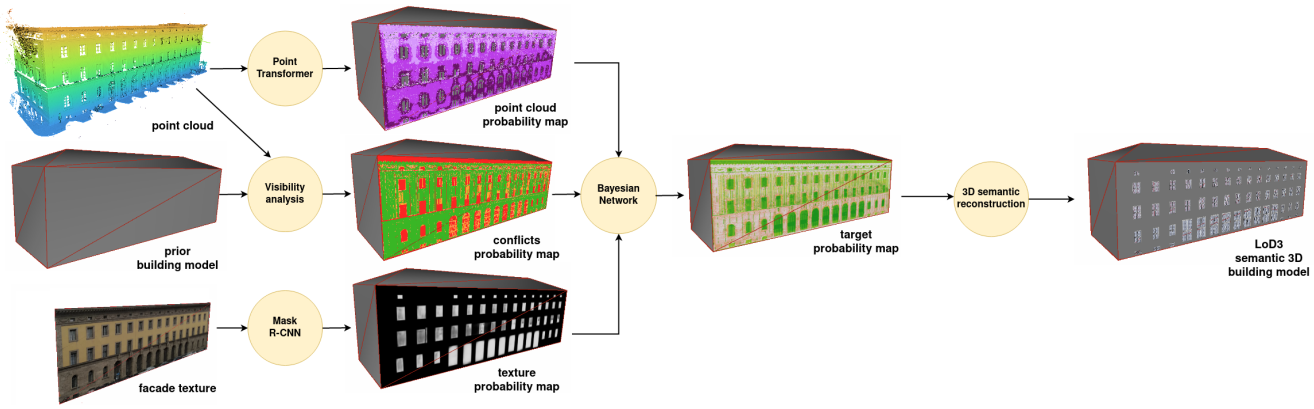


Figure 2. The workflow of the proposed Scan2LoD3 consists of three parallel branches: The first is generating the point cloud probability map based on a modified Point Transformer network (top); the second is producing a conflicts probability map from the visibility of the laser scanner in conjunction with a 3D building model (middle); and the third is using Mask-RCNN to obtain a texture probability map from 2D images. We then fuse three probability maps with a Bayesian network to obtain final facade-level segmentation, enabling a CityGML-compliant LoD3 building model reconstruction.

ometric features and visibility analysis. The authors merge model conflicts and inferred semantics from a modified Point Transformer architecture [55]. The output is added to a 3D building model face using a projection, and respective window and door openings are 3D-modeled by 3D bounding box fitting of pre-defined models. The method, however, is limited in reconstructing windows with partially closed blinds owing to simplified probabilities to binary masks comprising only high-probability conflicts and semantics. Additionally, the visibility analysis concerns uncertainties using L1 distance, which generalizes L2 distance measurements, rendering it less sensitive for detailed conflicts.

3. Methodology

Our Scan2LoD3 method comprises two interconnected steps: semantic 3D segmentation that yields input for semantic 3D reconstruction. As shown in Figure 2, we first generate a ray-based conflicts probability map consisting of three states (*conflicted*, *confirmed*, and *unknown*), analyzing the visibility of the laser scanner in conjunction with 3D building models (Sec. 3.1). However, this map is limited to the laser field-of-view and does not provide façade-specific semantics. To address this limitation, we additionally introduce two probability maps derived from point clouds and images: The former is generated by a modified Point Transformer network [49, 55] (top branch), while the latter is produced using Mask-RCNN [11] (bottom branch), as described in Sections 3.2 and 3.3, respectively. We then fuse these three probability maps via a Bayesian network, resulting in a target probability map that represents the occurrence of openings and their associated probability score (Sec. 3.4). The opening labels yield detailed 3D opening geometries for reconstruction, which is conducted with the input 3D building model and a pre-defined 3D library of openings (Sec. 3.5). Finally, we assign the respective semantics to the reconstructed parts along with the final probability score, resulting in the CityGML-compliant LoD3 building model [9].

3.1. Visibility analysis concerning uncertainties

We perform ray tracing on a 3D voxel grid to determine areas that are measured by a laser scanner and analyze them with a 3D building model (Fig. 3). The total grid size adapts to the input data owing to the utilized octree structure with leaves represented by 3D voxels of size v_s dependent on the relative accuracy of the scanner.

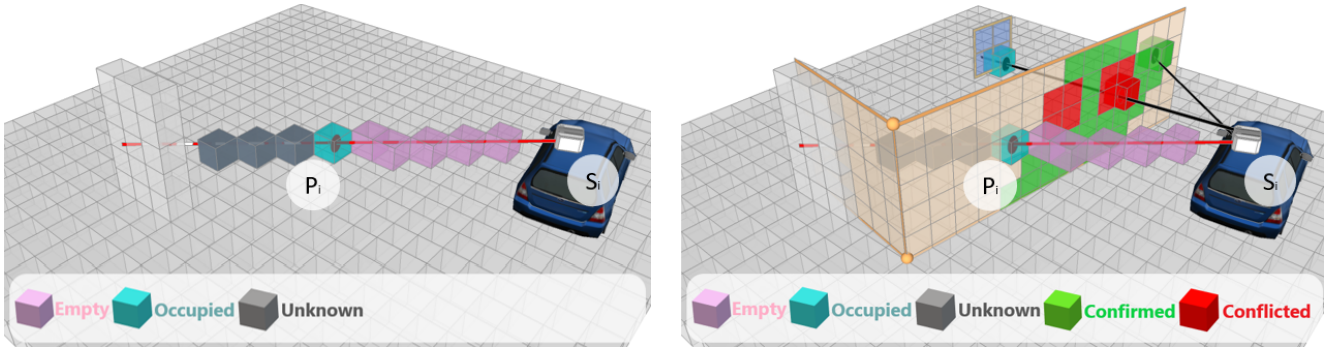
As shown in Figure 3a, the laser rays are traced from sensor position s_i, using orientation vector r_i, to hit point $p_i = s_i + r_i$. Our approach leverages MLS trait of multiple laser observations z_i to decide upon the laser occupancy states (i.e., *empty*, *occupied*, and *unknown*) and includes the respective occupancy probability score. The states' update mechanism uses prior probability $P(n)$, current estimate $L(n|z_i)$, and preceding estimate $L(n|z_{1:i-1})$ to calculate and assign the final state. The mechanism is controlled by log-odd values $L(n)$ along with clamping thresholds l_{min} and l_{max} [14, 49, 50]:

$$L(n|z_{1:i}) = max(min(L(n|z_{1:i-1}) + L(n|z_i), l_{max}), l_{min}) \quad (1)$$

where

$$L(n) = log[\frac{P(n)}{1 - P(n)}] \quad (2)$$

As illustrated in Figure 3a, in the visibility analysis process of laser observations, voxels encompassing p_i are



(a) Ray casting of laser observations

(b) Rays analyzed with 3D model

Figure 3. Visibility analysis using laser scanning observations and 3D models on a voxel grid. The ray is traced from the sensor position s_i to the hit point p_i. The voxel is: *empty* if the ray traverses it; *occupied* when it contains p_i; *unknown* if unmeasured; *confirmed* when *occupied* voxel intersects with vector plane; and *conflicted* when the plane intersects with an *empty* voxel [49].

deemed as *occupied* (light-blue), those traversed by a ray as *empty* (pink), and unmeasured as *unknown* (gray). Then, as shown in Figure 3b, we assign further voxel states by analyzing occupancy voxels and building model: Voxels are *confirmed* (green) when *occupied* voxels intersect with the building surface and are *conflicted* (red) when a ray traverses a building surface and reflects inside a building. The final probability estimate, however, also concerns 3D model uncertainties.

Specifically, we address the uncertainties of global positioning accuracy of building model surfaces and of point clouds along the ray. Let us assume that the probability distribution of global positioning accuracy of a building surface $P(A)$ is described by the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1)$, where μ_1 and σ_1 are the mean and standard deviation of the Gaussian distribution. Analogically, let us assume that the probability distribution of global positioning accuracy of a point in point cloud $P(B)$ is described by the Gaussian distribution $\mathcal{N}(\mu_2, \sigma_2)$. To estimate the probability of the confirmed $P_{confirmed}$ and conflicted $P_{conflicted}$ states of the voxel V_n, we use the joint probability distribution of two independent events $P(A)$ and $P(B)$:

$$V_n = \begin{Bmatrix} P_{confirmed}(A,B) = P(A) * P(B) \\ P_{conflicted}(A,B) = 1 - P_{confirmed}(A,B) \end{Bmatrix} \quad (3)$$

We obtain a *conflicts probability map* (Fig. 4) by projecting the vector-intersecting voxels to the vector plane, where the cell spacing is consistent with the voxel grid; each pixel receives probability values of the states *conflicted*, *confirmed*, and *unknown*, accordingly.

3.2. 3D semantic segmentation on point clouds

We semantically segment 3D point clouds using the enhanced Point Transformer (PT) network [49, 55]. The enhancement involves fusing geometric features at the early

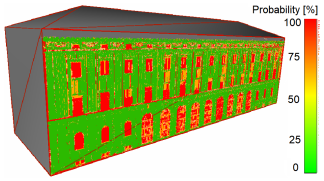


Figure 4. Exemplary *conflict probability map*: high probability pixels present high conflict probability, whereas low probability pixels show high confirmation probability.

training stage to increase 3D façade segmentation performance [23, 49]. In this work, we consider seven geometric features: *height of the points, roughness, volume density, verticality, omnivariance, planarity*, and *surface variation* [8, 46, 49], which are calculated within an Euclidean neighborhood search radius d_i. We define eight pertinent classes for the façade segmentation task: *arch*, *column*, *molding*, *floor*, *door*, *window*, *wall*, and *other* [49].

The final softmax layer of the modified PT network provides a per-point vector of probabilities of each class as an output (Fig. 5). Notably, in contrast to [49], we do not discard points based on a probability threshold but consider each point and its class probability score for further processing. Finally, we create the *point cloud probability map* (Fig. 7) by projecting the points onto the face of a building while preserving the probabilities and following the cell

Figure 5. Exemplary results of the modified network: point cloud colors according to the probability vector of the class *window*.

spacing of the *conflict probability map* (Sec. 3.1).

3.3. 2D semantic segmentation on images

As demonstrated by Hensel et al., 2019, [12], Faster R-CNN [32] effectively identifies approximate façade openings positions. In our approach, we utilize Mask-RCNN [11], which builds upon the concept of Faster R-CNN and identifies probability masks within proposed bounding boxes. This trait allows us to obtain later a more accurate instances that are not necessarily restricted to a rectangular shape. For the proposed façade opening detection, we focus on two classes: windows and doors. Analogically to the 3D semantic segmentation stage (Sec. 3.2), we preserve the pixel-predicted probabilities. To generate the texture probability map (Fig. 6), we project the pixels and their probabilities onto the building face, aligning with the cell spacing of the other probability maps (Secs. 3.1 and 3.2).

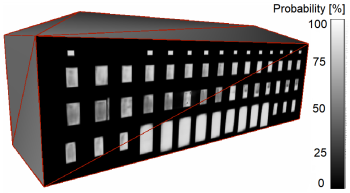


Figure 6. Exemplary *texture probability map*: high probability pixels stand for a high probability of opening.

3.4. Final segmentation with Bayesian network

To calculate the final shape, semantics, and probability score of opening instances, the multimodal probability maps are fused using a Bayesian network. The network quantifies uncertainties and assigns weights based on evidence when calculating the target probability map. Figure 7 shows the network architecture, including three input nodes for each probability map, to infer the probability of opening occurrence. The X and Y nodes exhibit a causal relationship, forming directed acyclic links. We utilize a conditional probability table (CPT) to assign weights to combinations of each node and state. The target node estimates two mutually exclusive states: opening and non-opening. The probability of node Y (opening space) being in the state y (opening) is calculated using the marginalization process, which combines the conditional probabilities of the parent nodes' X states x (i.e., of point cloud probability, conflicts probability, texture probability maps) [38, 50].

The probability maps serve as pieces of evidence updating the joint probability distribution $P(X, Y)$ of the compiled network. The inference mechanism performs the update and estimates the posterior probability distribution (PPD), which provides the states' probability [38, 50]. In

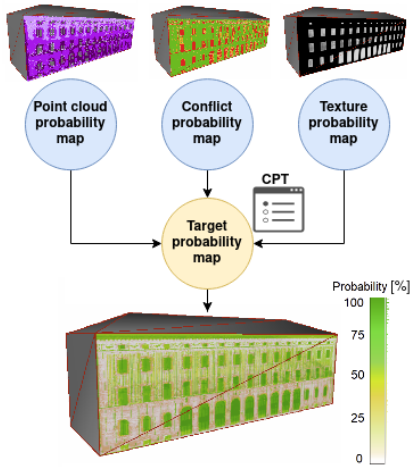


Figure 7. The Bayesian network architecture comprising three input nodes (blue), one target node (yellow), and a conditional probability table (CPT) with the assigned combinations' weights.

general, the network favors situations where there is a high probability of an opening occurring if at least two pieces of high-probability evidence co-occur; otherwise, it yields a low opening probability. For example, a very high *conflict* probability overlying high texture *opening* probability and medium point cloud *opening* probability should yield a high *opening* probability.

As an output from the Bayesian network, we extract the high probability clusters P_{high}, which have a neighbor in any of the eight directions of the pixel. To distinguish between doors and windows, we compare overlying per-pixel class probabilities and select the more probable pixel class. The pixel-wise probability scores are then averaged per instance and kept for the final 3D model. Since the extraction can include noisy clusters, we employ their post-processing to obtain final, noise-free opening shapes. To this end, we apply morphological opening to reduce the effect of small distortions and weak-connected shapes. We also calculate a modified rectangularity index [3, 49], on which basis we reject erroneously elongated shapes using upper PE_{up} and lower PE_{lo} percentiles of the index score.

3.5. Semantic 3D reconstruction

Since it is crucial to preserve the 3D model's watertightness and its given semantics, we use the prior building solid as the basis for the modeling. Specifically, the openings are cut automatically in the prior model using the constructive solid geometry (CSG) difference operation: the bounding boxes of found windows and doors cut the openings in the outer boundaries of the given solid. Then, we use these 3D cuts as matching and fitting geometries for automatically queried 3D models from a pre-defined library of LoD3 façade objects. To ensure the watertightness and prevent

self-intersections, each object is aligned with the respective face and scaled to the 3D cut shape.

We leverage the CityGML's traits to create a hierarchical semantic model structure [9]. Specifically, the prior solid and its constituting faces preserve their unique identifiers and associated semantic classes. The new unique identifiers are assigned to openings, which point to the respective solid's faces; each window and door obtain the standard *Window* and *Door* class, respectively. As it is pivotal to preserve the final detection confidence, we also add an attribute named *confidence*, keeping the final detection confidence of the shape opening. Ultimately, the model's LoD attribute value is upgraded to LoD3.

4. Experiments

In this section, we describe experiments concerning the proposed Scan2LoD3 method, which necessitated acquiring existing and creating new datasets. Within the scope of this work, we publish in the repository [2]: textured LoD2 and modeled LoD3 building models, enriched TUM-FAÇADE point clouds, implementation, and settings.

4.1. Datasets

To showcase the performance of Scan2LoD3, we evaluated the method on the public datasets: TUM-MLS-2016 [56], TUM-FAÇADE [51,52] and textured CityGML building models at LoD2 [44] representing the Technical University of Munich main campus in Munich, Germany. Additionally, we used a proprietary MLS point cloud of the TUM area called MF. To validate the reconstruction, segmentation, and detection performance, we manually modeled a CityGML-compliant LoD3 building model [9] based on the combination of point clouds and LoD2 building model, serving as ground-truth; the LoD2 building models were additionally textured.

The TUM-MLS-2016 dataset. The point clouds in TUM-MLS-2016 were collected via obliquely mounted two Velodyne HDL-64E LiDAR sensors mounted on the Mobile Distributed Situation Awareness (MODISSA) platform. The entire point cloud covered an urban area with an inner and outer yard of the campus. The inertial navigation system was supported by the real-time kinematic (RTK) correction data of the the German satellite positioning service (SAPOS), which ensured geo-referencing.

The TUM-FAÇADE dataset. The TUM-FAÇADE dataset is derived from the TUM-MLS-2016 point clouds, where the former enriches the latter in 17 façade-level semantic classes. The dataset comprises 17 annotated and 12 non-annotated façades totalling 256 million façade-level labeled and geo-referenced points. Within the scope of this work, we additionally annotated four of the open-access non-annotated façades. As discussed in Section 3.2, we define seven façade classes as pertinent for the reconstruction. Therefore, we combined 17 TUM-FAÇADE's classes into seven by merging: *molding* with *decoration*; *drainpipe* with *wall*, *outer ceiling surface* and *stairs*; *floor* with *terrain* and *ground surface*; *other* with *interior* and *roof*; *blinds* with *window*; whereas *door* remained intact.

The MF dataset. The MF point clouds were acquired at the TUM campus and covered an approximately the same area as the TUM-MLS-2016 dataset. The point cloud was geo-referenced by proprietary mobile mapping platform, supported by the German SAPOS RTK system [37].

Textured LoD2 and LoD3 semantic building models. We acquired open data CityGML-compliant building priors at LoD2 from the state open access portal of Bavaria, Germany [44], which were created using 2D cadastre footprints in combination with aerial observations [34]; comparable results can be achieved with methods such as PolyFit [26]. The textures were acquired manually at an approximately 45° horizontal angle using a 13MP rear camera of a Xiaomi Redmi Note 5A smartphone and projected to the respective faces: this approach simulated terrestrial acquisition of a mobile mapping unit or street view imagery where no ortho-rectifications were applied [15]. The LoD3 building model was created manually based on a combination of TUM-FAÇADE and textured LoD2 models. We modeled the so-called *building 23* as it has been commonly used as a validation object for various methods [13, 40, 49, 51, 52]. The pre-defined library of openings was downloaded from the open dataset of LoD3 building models of Ingolstadt, Germany [35].

4.2. Implementation details

Visibility analysis. We set the size of voxels to $v_s = 0.1$ m and initialized them with a uniform prior probability of $P = 0.5$ to perform the ray casting on an efficient octree structure [14]; we used the standard [14, 41, 49] clamping and log-odd values. The uncertainty of building models and point clouds was assigned considering their reported global positioning accuracy. As such, the parameters of building models were set to $\mu_1 = 0$ and $\sigma_1 = 3$, while for the TUM-MLS-2016 and MF point clouds were set to $\mu_2 = 0$, $\sigma_2 = 2.85$ and to $\mu_2 = 0$, $\sigma_2 = 1.4$, respectively.

Semantic segmentation. For the modified Point Transformer data pre-processing, we followed [49] and removed redundant points within a 5 cm radius, which resulted in 10 million points; the point cloud was split into 70% training and 30% validation subsets. We chose the optimal geometric features search radius d_i following [7, 49]: As for the features *roughness, volume density, omnivariance, planarity*, and *surface variation* the radius was set to $d_i = 0.8$ m; whereas for *verticality* to $d_i = 0.4$ m. For the image segmentation, we deployed a pre-trained Mask-RCNN on the COCO dataset [21]. The inference was fine-tuned with 378 base images of the CMP façade database [42], where

we selected two classes for training: *door* and *window* including *blinds*. As P_{high} pixels in the Bayesian network, we deemed values higher than $P_{high} = 0.7$. To reject outliers, we fixed the modified rectangularity percentiles to $PE_{up} = 95$ and $PE_{low} = 5$.

4.3. Results and Discussion

Detection rate. The methods of Hoegner & Gleixner, 2022, [13] and Wysocki et al., 2022 [49] were both tested on the three façades of the *building 23* at the TUM campus using the TUM-MLS-2016 data; thus we validated the detection accuracy using the same setup and our manually modeled LoD3 building (Tab. 1). To show the ratio of the detection rate to the laser-covered rate, we introduced metrics for all existing façade openings (AO) and only laser-measured façade openings (MO).

Our multimodal fusion enabled a higher detection rate and still maintained a low false alarm rate. If compared to the Hoegner & Gleixner (H&G) [13] and CC [49] methods, Scan2LoD3 achieved higher detection rate on the TUM dataset by 10% and 6%, respectively (Tab. 1 and Fig. 8). The MF map provided more accurate results (i.e., 91% of

	H&G [40]				CC [49]				Scan2LoD3 (TUM)				Scan2LoD3 (MF)			
	A	B	C	Tot	A	B	C	Tot	A	B	C	Tot	A	B	C	Tot
AO	66	17	20	**103**	66	17	20	**103**	66	17	20	**103**	66	17	20	**103**
MO	60	17	10	**87**	60	17	12	**87**	60	17	12	**89**	66	12	18	**96**
D	60	15	4	**75**	60	15	6	**81**	60	16	11	**87**	65	16	16	**97**
TP	60	12	4	**76**	60	15	5	**80**	60	16	11	**87**	65	14	15	**94**
FP	0	3	0	**3**	0	0	1	**1**	0	0	0	**0**	0	0	1	**3**
FN	6	5	16	**27**	6	2	15	**23**	6	1	9	**16**	1	3	5	**9**
DA	91	71	20	**74**	91	88	25	**78**	91	94	55	**84**	98	82	75	**91**
FA	0	0	0	**4**	0	0	17	**1**	0	0	0	**0**	0	12	6	**3**
DM	100	71	40	**87**	100	88	42	**90**	100	94	92	**98**	98	117	83	**98**

Table 1. Detection rate for all openings (DA) and laser-measured openings (DM) and the respective false alarm rate (FA) for façades A, B, and C (AO = all openings, MO = laser-measured openings, D = detections, TP = true positives, FP = false positives, FN = false negatives).

all openings correctly detected) owing to higher point cloud global accuracy and complete façade A coverage; also other maps complemented the MF's laser-observed openings, as exemplified by façade B (Tab. 1).

Semantic segmentation. To measure the accuracy of the segmentation, we selected the median per-instance intersection over union (IoU) metric for all openings of *building 23* (Tab. 2 and Fig. 8). This setup enabled us the comparison to the introduced modified Point Transformer network (Pt+Ft.) working only on point clouds; Mask-RCNN (M-RCNN) using only images [11]; method using ray-casting and binary point cloud masks (CC) [49]; our method fusing three maps (i.e., conflicts, point clouds, images), once with the TUM-MLS-2016 conflict map (TUM) and on the higher accuracy conflict map of MF (MF).

Our experiments corroborate that, in contrast to the tested methods, our proposed solution identifies even closed openings, their full shapes, and reaches higher accuracy (Tab. 2 and Fig. 8). This fact enabled the whole-shape reconstruction of, for example, covered by blinds windows, which resulted in up to 20% higher IoU on the TUM-MLS-2016 dataset (red boxes, Fig. 8). Similarly to the detection

	median IoU ↑			
Façade	A	B	C	Total
Openings	66	17	20	103
PT+Ft. [55]	7.3	4.6	3.7	7.3
M-RCNN [11]	63.7	47.4	38.6	58.4
CC [49]	66.5	56.4	**53.2**	60.6
Scan2LoD3 (TUM)	63.9	52.9	38	62.1
Scan2LoD3 (MF)	**78.4**	**62.3**	40.6	**76.2**

Table 2. Comparison of opening segmentation using only: 3D point clouds (Pt+Ft.), images (M-RCNN), binary masks (CC), and our method with TUM and MF conflict maps.

results, the accuracy of laser measurements significantly influenced the IoU results: Our method tended to overestimate opening shapes on the TUM point cloud, whereas on MF the shapes were approximately 14% more accurate. On the other hand, Scan2LoD3 was sensitive to poor segmentation results (façade C, Tab. 2).

3D reconstruction. We measured the accuracy of reconstruction by comparing our method using the TUM-FAÇADE data to the well-established and mesh-oriented Poisson reconstruction [18] and to the second-best-IoU performing CC method (Figs. 8 and 9 and Tab. 3). To highlight the influence of point cloud accuracy, we also added the results for MF point clouds. As shown in Table 3 and

Method	vs. GT LoD3 ↓		
	μ	RMS	WT
Poisson (TUM) [18]	0.35	0.54	✗
CC [49]	0.31	0.34	✓
Scan2LoD3 (TUM)	0.23	0.26	✓
Scan2LoD3 (MF)	**0.13**	**0.25**	✓

Table 3. Comparison of mesh-based Poisson, building-prior-driven CC, and our proposed method using the ground-truth LoD3 model and measuring watertightness (WT).

in Figure 9, the 3D building priors provided more accurate reconstruction results than the standard Poisson reconstruction (i.e., RMS lower by 52%); the former also achieved the watertightness. Among the prior-driven methods, the improvement related to higher detection rate and IoU was noticeable: Scan2LoD3 had lower mean and RMS scores by up to 26% and 24%, respectively, compared to CC (Tab. 3).

Figure 8. Comparison of different reconstruction results for the façade A: Our method reconstructs complete window shapes despite the presence of window blinds (red boxes).

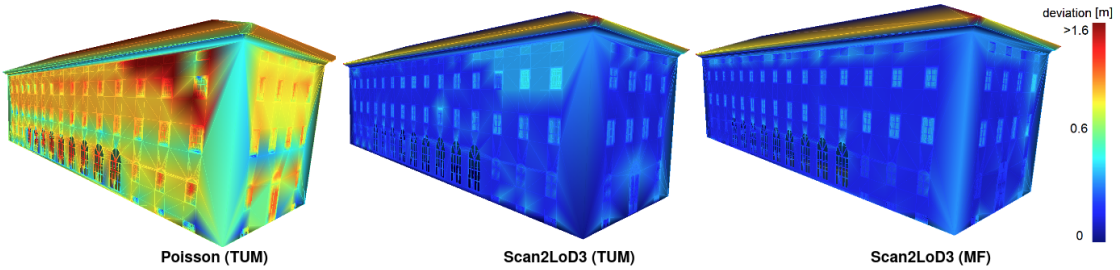


Figure 9. Comparison of the Poisson to our reconstruction approach: Deviations are projected onto the ground-truth LoD3 model.

It is worth noting that the eaves were incorrectly reconstructed in any of the presented methods.

5. Conclusions

In this paper, we introduce Scan2LoD3, a multimodal probabilistic fusion method for the high-detail semantic 3D building reconstruction. Our work has led us to the conclusion that the multimodal probabilistic fusion can maximize the advantages of ray-casting- and learning-based methods for the LoD3 reconstruction. The findings of this study indicate that while joining images, point clouds, and model conflicts, a Bayesian network reveals a very high-level detection rate (i.e., 91%); and robustness as the false alarm rate is negligible (i.e., 3%). Crucially, our method segments and reconstructs complete opening shapes, even when closed by blinds, which can provide up to around 76% shape accuracy. By such detection and segmentation, we minimize the final reconstruction deviations by 54% and 24% when compared to mesh-based and other prior-driven methods, respectively. Such method's characteristics are of great importance for applications necessitating object-oriented semantics, high robustness, and completeness, such as automated driving testing [36] or façade solar potential analysis [47], among others [28, 48]. Furthermore, an upshot of keeping reconstruction confidence score can be pivotal for confidence-based navigation algorithms, such as in autonomous cars [1,48,57]. It is worth noting that our method focuses on upgrading facades to LoD3; refining roofs to LoD3 would require additional, airborne data.

As the late fusion results so far have been very encouraging and do not require any training data, we deem Bayesian networks suitable for the task. Future work will concentrate on comparing the Bayesian network's generalization capabilities to deep neural networks, which, however, require extensive training data. Moreover, we expect the method's performance to be comparable on similar architecture styles; considering selected classes and small sample size. To tackle these issues, we plan to extend our open library of textured LoD2 and LoD3 models to foster the methods' development.

Acknowledgments This work was supported by the Bavarian State Ministry for Economic Affairs, Regional Development and Energy within the framework of the IuK Bayern project *MoFa3D - Mobile Erfassung von Fassaden mittels 3D Punktwolken*, Grant No. IUK643/001. Moreover, the work was conducted within the framework of the Leonhard Obermeyer Center at the Technical University of Munich (TUM).

References

[1] Christian R Albrecht, Jenny Behre, Eva Herrmann, Stefan Jürgens, and Uwe Stilla. Investigation on robustness of vehicle localization using cameras and LiDAR. *Vehicles*, 4(2):445–463, 2022. 8

[2] Sam Amirebrahimi, Abbas Rajabifard, Priyan Mendis, and Tuan Ngo. A bim-gis integration method in support of the assessment and 3d visualisation of flood damage to a building. *Journal of spatial science*, 61(2):317–350, 2016. 1

[3] Melih Basaraner and Sinan Cetinkaya. Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in GIS. *International Journal of Geographical Information Science*, 31(10):1952–1977, 2017. 5

[4] Filip Biljecki, Jantien Stoter, Hugo Ledoux, Sisi Zlatanova, and Arzu Çöltekin. Applications of 3D city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4):2842–2889, 2015. 1

[5] Konstantinos Chaidas, George Tataris, and Nikolaos Soulakellis. Seismic damage semantics on post-earthquake lod3 building models generated by uas. *ISPRS International Journal of Geo-Information*, 10(5), 2021. 1

[6] Raghudeep Gadde, Renaud Marlet, and Nikos Paragios. Learning grammars for architecture-specific facade parsing. *International Journal of Computer Vision*, 117(3):290–316, 2016. 2

[7] Eleonora Grilli, Elisa Mariarosaria Farella, Alessandro Torresani, and Fabio Remondino. Geometric features analysis for the classification of cultural heritage point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W15:541–548, 2019. 6

[8] Eleonora Grilli and Fabio Remondino. Machine learning generalisation across different 3D architectural heritage. *ISPRS International Journal of Geo-Information*, 9(6):379, 2020. 2, 4

[9] Gerhard Gröger, Thomas H Kolbe, Claus Nagel, and Karl-Heinz Häfele. OGC City Geography Markup Language CityGML Encoding Standard, 2012. Open Geospatial Consortium: Wayland, MA, USA, 2012. 2, 3, 6

[10] Norbert Haala and Martin Kada. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):570 – 580, 2010. 1

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 5, 7

[12] Simon Hensel, Steffen Goebbels, and Martin Kada. Facade reconstruction for textured LoD2 CityGML models based on deep learning and mixed integer linear programming. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5:37–44, 2019. 2, 5

[13] Ludwig Hoegner and Georg Gleixner. Automatic extraction of facades and windows from MLS point clouds using voxelspace and visibility analysis. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022:387–394, 2022. 2, 6, 7

[14] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Auton. Robots*, 34(3):189–206, 2013. 3, 6

[15] Yujun Hou and Filip Biljecki. A comprehensive framework for evaluating the quality of street view imagery. *International Journal of Applied Earth Observation and Geoinformation*, 115:103094, 2022. 6

[16] Hai Huang, Mario Michelini, Matthias Schmitz, Lukas Roth, and Helmut Mayer. LoD3 building reconstruction from multi-source images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020:427–434, 2020. 2

[17] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017. 2

[18] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, page 0, 2006. 7

[19] Filip Korč and Wolfgang Förstner. eTRIMS Image Database for interpreting images of man-made scenes. Technical Report TR-IGG-P-2009-01, Department of Photogrammetry, University of Bonn, April 2009. 2

[20] Binyu Lei, Rudi Stouffs, and Filip Biljecki. Assessing and benchmarking 3d city models. *International Journal of Geographical Information Science*, 0(0):1–22, 2022. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence C Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[22] Hantang Liu, Yinghao Xu, Jialiang Zhang, Jianke Zhu, Yang Li, and Steven CH Hoi. DeepFacade: A deep learning approach to facade parsing with symmetric loss. *IEEE Transactions on Multimedia*, 22(12):3153–3165, 2020. 2

[23] Francesca Matrone, Eleonora Grilli, Massimo Martini, Marina Paolanti, Roberto Pierdicca, and Fabio Remondino. Comparing machine and deep learning methods for large 3d heritage semantic segmentation. *ISPRS International Journal of Geo-Information*, 9(9):535, 2020. 2, 4

[24] Theresa Meyer, Ansgar Brunn, and Uwe Stilla. Change detection for indoor construction progress monitoring based on bim, point clouds and uncertainties. *Automation in Construction*, 141:104442, 2022. 2

[25] Przemyslaw Musialski, Peter Wonka, Daniel G Aliaga, Michael Wimmer, Luc Van Gool, and Werner Purgathofer. A survey of urban reconstruction. *Computer graphics forum*, 32(6):146–177, 2013. 2

[26] Liangliang Nan and Peter Wonka. Polyfit: Polygonal surface reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2353–2361, 2017. 6

[27] Romain Nouvel, Claudia Schulte, Ursula Eicker, Dirk Pietruschka, and Volker Coors. CityGML-based 3D city model for energy diagnostics and urban energy policy support. *IBPSA World*, 2013:1–7, 2013. 1

[28] Ankit Palliwal, Shuang Song, Hugh Tiang Wah Tan, and Filip Biljecki. 3D city models for urban farming site identification in buildings. *Computers, Environment and Urban Systems*, 86:101584, 2021. 8

[29] Hui E Pang and Filip Biljecki. 3D building reconstruction from single street view images using deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 112:102859, 2022. 2

[30] Bryan G Pantoja-Rosero, Radhakrishna Achanta, Mateusz Kozinski, Pascal Fua, Fernando Perez-Cruz, and Katrin Beyer. Generating LoD3 building models from structure-from-motion and semantic segmentation. *Automation in Construction*, 141:104430, 2022. 2

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5

[33] Hayko Riemenschneider, Ulrich Krispel, Wolfgang Thaller, Michael Donoser, Sven Havemann, Dieter Fellner, and Horst Bischof. Irregular lattices for complex shape grammar facade parsing. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA, June 16-21 2012, pp. 1640–1647, 2012. 2

[34] Robert Roschlaub and Joachim Batscheider. An INSPIRE-conform 3D building model of Bavaria using cadastre information, LiDAR and image matching. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B4:747–754, 2016. 1, 6

[35] Benedikt Schwab, Sophie Haas Goschenhofer, and Olaf Wysocki. LoD3 road space models, release v0.8.1. https://github.com/savenow/lod3-road-space-models, 2021. Accessed: 2023-01-30. 6

[36] Benedikt Schwab and Thomas H Kolbe. Requirement analysis of 3D road space models for automated driving. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-4/W8:99–106, 2019. 1, 8

[37] 3D Mapping Solutions. MoSES mobile mapping platform - technical details. https://www.3d-mapping.de/ueber-uns/unternehmensbereiche/data-acquisition/unser-vermessungssystem/, 2023. Accessed: 2023-01-30. 6

[38] Ana Stritih, Sven-Erik Rabe, Orencio Robaina, Adrienne Grêt-Regamey, and Enrico Celio. An online platform for spatial and iterative modelling with bayesian networks. *Environmental Modelling & Software*, 127:104658, 2020. 5

[39] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 1, 2

[40] Sebastian Tuttas and Uwe Stilla. Reconstruction of façades in point clouds from multi aspect oblique ALS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W3:91–96, 2013. 2, 6, 7

[41] Sebastian Tuttas, Uwe Stilla, Alexander Braun, and André Borrmann. Validation of BIM components by photogrammetric point clouds for construction site monitoring. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4:231–237, 2015. 2, 6

[42] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition, Saarbrücken, Germany, September 3-6, 2013*, pages 364–374. Springer, 2013. 2, 6

[43] Maria Uggla, Perola Olsson, Barzan Abdi, Björn Axelsson, Matthew Calvert, Ulrika Christensen, Daniel Gardevärn, Gabriel Hirsch, Eric Jeansson, Zuhret Kadric, Jonas Lord, Axel Loreman, Andreas Persson, Ola Setterby, Maria Sjöberger, Paul Stewart, Andreas Rudenå, Andreas Ahlström, Mikael Bauner, Kendall Hartman, Karolina Pantazatou, Wenjing Liu, Hongchao Fan, Gefei Kong, Hang Li, and Lars Harrie. Future Swedish 3D city models - specifications, test data, and evaluation. *ISPRS International Journal of Geo-Information*, 12(2), 2023. 1

[44] Bayerische Vermessungsverwaltung. 3D-Gebäudemodelle (LoD2). https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=lod2, 2023. Accessed: 2023-01-30. 6

[45] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph CNN for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 2

[46] Martin Weinmann, Boris Jutzi, and Clément Mallet. Feature relevance assessment for the semantic interpretation of 3D point cloud data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5/W2:313–318, 2013. 4

[47] Bruno Willenborg, Martin Pültz, and Thomas H Kolbe. Integration of semantic 3D city models and 3D mesh models for accuracy improvements of solar potential analyses. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W10:223–230, 2018. 1, 8

[48] Kelvin Wong, Yanlei Gu, and Shunsuke Kamijo. Mapping for autonomous driving: Opportunities and challenges. *IEEE Intelligent Transportation Systems Magazine*, 13(1):91–106, 2020. 8

[49] Olaf Wysocki, Eleonora Grilli, Ludwig Hoegner, and Uwe Stilla. Combining visibility analysis and deep learning for refinement of semantic 3D building models by conflict classification. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W2-2022:289–296, 2022. 2, 3, 4, 5, 6, 7

[50] Olaf Wysocki, Ludwig Hoegner, and Uwe Stilla. Refinement of semantic 3D building models by reconstructing underpasses from MLS point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 111:102841, 2022. 2, 3, 5

[51] Olaf Wysocki, Ludwig Hoegner, and Uwe Stilla. TUM-FAÇADE: Reviewing and enriching point cloud benchmarks

for façade segmentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVI-2/W1-2022:529–536, 2022. 2, 6

[52] Olaf Wysocki, Jiarui Zhang, and Uwe Stilla. TUM-FAÇADE: A database of annotated façade point clouds. https://mediatum.ub.tum.de/1636761?v=2, 2021. Accessed: 2023-01-08. 6

[53] Yan Xia, Yusheng Xu, Cheng Wang, and Uwe Stilla. VPC-Net: Completion of 3D vehicles from MLS point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:166–181, 2021. 2

[54] Yusheng Xu and Uwe Stilla. Towards building and civil infrastructure reconstruction from point clouds: A review on data and key techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2857–2885, 2021. 1

[55] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2, 3, 4, 7

[56] Jingwei Zhu, Joachim Gehrung, Rong Huang, Björn Borgmann, Zhenghao Sun, Ludwig Hoegner, Marcus Hebel, Yusheng Xu, and Uwe Stilla. TUM-MLS-2016: An annotated mobile LiDAR dataset of the TUM City Campus for semantic point cloud interpretation in urban areas. *Remote Sensing*, 12(11):1875, 2020. 6

[57] Qianqian Zou and Monika Sester. Uncertainty representation and quantification of 3d building models. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022:335–341, 2022. 8