The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVI-2/W1-2022
9th Intl. Workshop 3D-ARCH “3D Virtual Reconstruction and Visualization of Complex Architectures”, 2–4 March 2022, Mantua, Italy

TUM-FAÇADE: REVIEWING AND ENRICHING POINT CLOUD BENCHMARKS FOR FAÇADE SEGMENTATION

O. Wysocki, L. Hoegner, U. Stilla

Photogrammetry and Remote Sensing, TUM School of Engineering and Design, Technical University of Munich (TUM), Munich, Germany
(olaf.wysocki, ludwig.hoegner, stilla)@tum.de

KEY WORDS: Point cloud benchmark, Façade segmentation, Semantic segmentation, Review, TUM-FAÇADE, 3D reconstruction

ABSTRACT:

Point clouds are widely regarded as one of the best dataset types for urban mapping purposes. Hence, point cloud datasets are commonly investigated as benchmark types for various urban interpretation methods. Yet, few researchers have addressed the use of point cloud benchmarks for façade segmentation. Robust façade segmentation is becoming a key factor in various applications ranging from simulating autonomous driving functions to preserving cultural heritage. In this work, we present a method of enriching existing point cloud datasets with façade-related classes that have been designed to facilitate façade segmentation testing. We propose how to efficiently extend existing datasets and comprehensively assess their potential for façade segmentation. We use the method to create the TUM-FAÇADE dataset, which extends the capabilities of TUM-MLS-2016. Not only can TUM-FAÇADE facilitate the development of point-cloud-based façade segmentation tasks, but our procedure can also be applied to enrich further datasets.

1. INTRODUCTION

Buildings are one of the most fundamental elements of a city, which is why digital building reconstruction has become such a pivotal issue for the majority of urban studies. Every building possesses a number of façades, so digital building reconstruction inevitably involves façade reconstruction, too. This issue has long been regarded as a challenge within the photogrammetry and computer vision communities (Musialski et al., 2013).

Although state-of-the-art, semantic 3D building models are widely available, they have generalized extruded façades, chiefly owing to their top-view source datasets (Haala and Kada, 2010). This generalized level has been largely regarded as plausible for various applications (Biljecki et al., 2015).

However, recent developments have led to growing demand for detailed façade reconstruction in a wide variety of applications, including calculating heating demand (Nouvel et al., 2013), preserving cultural heritage (Grilli and Remondino, 2019), assessing flood damage (Apel et al., 2009), simulating wind flow (Montazeri and Blocken, 2013), analysing solar potential (Willenborg et al., 2018), and testing automated driving functions (Wysocki et al., 2021a, Schwab and Kolbe, 2019).

Central factors hampering the development of façade reconstruction methods are a lack of generic, façade-grade datasets and shortage of methods that can accommodate a range of architectural façade styles. While the former can be aided by mobile laser scanning (MLS) vehicles, which have recently begun delivering dense, street-level point clouds on an unprecedented scale, the latter requires various benchmark datasets for testing façade segmentation and reconstruction methods. However, this process is cumbersome and involves costly measurement campaigns as well as laborious, manual work to provide reference objects.

Despite this, recent years have witnessed a significant growth in urban point cloud benchmark dataset (Griffiths and Boehm, 2019a), but few of them have addressed the issue of façades segmentation, albeit they frequently include buildings.

In this paper, we present a method that reduces the need for creating new point cloud benchmark datasets by enriching existing benchmarks with façade-related semantics. To this end, our contributions are as follows:

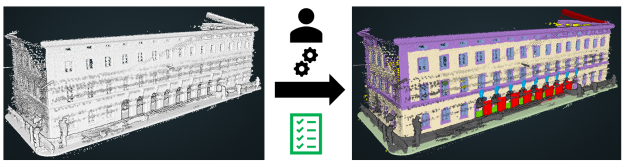

Figure 1. TUM-FAÇADE as a blueprint for enriching existing point cloud benchmarks: a) raw dataset, b) potential assessment, c) extended benchmark by façade classes.

- We review the terrestrial, outdoor point cloud benchmark datasets, with a focus on façade segmentation.
- We identify terrestrial, outdoor point cloud benchmark datasets that can potentially be used as testing datasets for façade segmentation.
- We present a method and classes that can enrich existing point cloud benchmark datasets for façade segmentation methods testing.
- We introduce TUM-FAÇADE[1] (Wysocki et al., 2021c), which enriches the TUM-MLS-2016 (Zhu et al., 2020) point cloud benchmark dataset with façade-related classes.

2. RELATED WORK

As they have a rich history in the domains of computer vision, photogrammetry, and remote sensing communities, there is a considerable amount of literature on benchmark datasets. Despite this level of interest, to the best of our knowledge, nobody has ever published a comprehensive review of point cloud benchmark datasets suitable for façade segmentation.

[1] https://github.com/OloOcki/tum-facade

Table 1. Our proposed classes for point cloud benchmark datasets to facilitate testing of façade segmentation methods.

Index	Class	CityGML building-related class	Description
1	wall	WallSurface	Walls excluding any decorative elements
2	window	Window	Windows excluding any decorative elements
3	door	Door	Including garage doors
4	balcony	BuildingInstallation	Excluding pillars and other supportive structures
5	molding	BuildingInstallation	Decorative static elements adhering to a building (e.g., cornices)
6	deco	BuildingInstallation	Decorative elements mounted to a building (e.g., flags, gargoyles, lights)
7	column	BuildingInstallation	Excluding cornices (cornice → molding class)
8	arch	BuildingInstallation	Only surfaces oriented downwards
9	drainpipe	BuildingInstallation	Pipes and rain gutters of a building
10	stairs	BuildingInstallation	Stairs excluding support structures (e.g., poles)
11	ground surface	GroundSurface	Any other ground surfaces inside a building envelope
12	terrain	-	Any other ground surfaces outside a building envelope (e.g., sidewalks)
13	roof	RoofSurface	Any surfaces relating to a roof structure (incl. dormers)
14	blinds	BuildingInstallation	Window closures open or closed
15	outer ceiling surface	OuterCeilingSurface	Ceilings within a building
16	interior	-	Measurements that reflect in a building
17	other	-	Any other elements

Research has tended to focus on overviews of available point cloud datasets rather than comprehensively reviewing them and focusing on the task of façade segmentation. For instance, Griffiths and Boehm provide a detailed review of deep learning techniques for 3D datasets, including a chapter concerning benchmark datasets (Griffiths and Boehm, 2019a). Not only do they present benchmark datasets for RGB-D, indoor, and outdoor scenes, but they also provide an overview of selected benchmarks. Zhu et al. provide an extensive list of outdoor MLS benchmark datasets, as well as presenting the TUM-MLS-2016 benchmark dataset (Zhu et al., 2020). Li et al. identify such characteristics as the format of the datasets or the number of available classes, albeit only with reference to a few selected benchmark datasets (Li et al., 2020). The work of Matrone et al. elaborates on the lack of 3D heritage datasets and bridges this gap by introducing the ArCH dataset (Matrone et al., 2020).

Façade segmentation methods have been widely studied (Musialski et al., 2013). Much is known about methods using images (Teboul et al., 2012, Mathias et al., 2016, Müller et al., 2007), largely facilitated by rich façade image datasets benchmarks, such as those by (Riemenschneider et al., 2012) or (Tyleček and Šára, 2013). However, as the images are 2D, they have to be processed to facilitate subsequent semantic 3D reconstruction. On the other hand, 3D point clouds are deemed among the best data sources for urban mapping purposes, as they yield an immediate 3D representation (Xu and Stilla, 2021). Of the particular interest are point clouds acquired by MLS vehicles thanks to their, high temporal resolution, and the density of the street-level point clouds (Wysocki et al., 2021a). This has led to a recent growth in interest in developing methods of parsing façades using point clouds (Martinovic et al., 2015, Fan et al., 2021, Zolanvari and Laefer, 2016), especially using machine learning methods (Matrone et al., 2020, Liu et al., 2020).

However, only a few studies have focused on releasing point cloud façade segmentation benchmark datasets (Matrone et al., 2020). In the literature, there are a few examples of methods that enrich existing datasets by adding new semantic information. One of these is the SemanticKITTI benchmark (Behley et al., 2021), which builds upon the KITTI Vision Benchmark (Geiger et al., 2013). Alternatively, the dataset can be enriched by conducting a repeated measurement campaign to provide another epoch, which is done chiefly for change detection purposes, as in the work by (Zhu et al., 2020).

To sum up, most of the existing benchmarks were not created for the purpose of façade segmentation. Moreover, publications either overlook some benchmarks or provide only sparse statistics, which hampers any detailed comparison of their potential for façade segmentation using point clouds. Although some methods of enriching existing benchmark datasets have been implemented, they are scarce, especially in the filed of façade segmentation.

3. DEVELOPED METHODOLOGY

3.1 Assessing existing benchmark datasets

As the majority of the point cloud benchmarks were not created for the purpose of façade segmentation, the benchmark datasets we consider comply with several requirements: They must represent an open-dataset, outdoor scene, depict buildings or at least façades, and consist of point clouds. We therefore exclude 2D image benchmark datasets such as (Tyleček and Šára, 2013, Gadde et al., 2016, Riemenschneider et al., 2012). Furthermore, as a façade represents a front of a building, it implies that indoor-oriented benchmarks, such as (Armeni et al., 2017), are out of scope, too. Due to the limited coverage of façade details, we disregard aerial benchmark datasets, such as (Varney et al., 2020) as well as automotive datasets that primarily focus on road objects, such as (Geiger et al., 2013).

We establish that features crucial to the comparison of datasets for façade segmentation tasks should include the following data: (1) year, (2) sensor type, (3) scalar fields relevant to segmentation (i.e., point position (XYZ), color (RGB), intensity (I), or normals (N)), (4) world (i.e., real or synthetic), (5) total number of points, (6) whether a dataset is georeferenced, (7) in what region it was acquired, (8) number of available classes, (9) whether a building class is available, (10) whether classes relating to façade details are available, and (11) whether the scene is urban or rural.

3.2 Creating an extended benchmark dataset

We propose 17 classes for façade segmentation, following the approach of Matrone et al., which is based on CityGML, Industry Foundation Classes (IFC), and Art and Architecture Thesaurus (AAT) (Matrone et al., 2020). We increase the number of classes introduced by Matrone et al., while maintaining consistency and backwards compatibility (i.e., it is possible to merge classes for testing on the same datasets). To facilitate both segmentation and

reconstruction tasks, our classes are also consistent with the modeling guidelines for CityGML level of detail (LoD)3 building models (Gröger et al., 2012, Special Interest Group 3D, 2020). We present the classes in Table 1, with their names and indices, a respective building-related CityGML class, and a brief description.

Extending point cloud benchmark datasets by adding new ground-truth classes inevitably necessitates manual work to be performed by trained annotators. To minimize the effort involved, supporting algorithms can be used to pre-cluster point clouds, as in (Zhu et al., 2020). In our case, the central aspect is to first cluster objects that belong to the façade and its immediate vicinity and neglect all other objects. Hence, we propose using point clouds that are georeferenced to clip-out buildings. Using the position obtained from the global coordinate reference system (CRS), we obtain point clouds superimposed on geographic information systems (GIS) datasets. This, in turn, allows us to create buffers around building footprints extracted from vector GIS datasets (e.g., CityGML building models or OpenStreetMap (OSM) buildings). This ensures to reject a significant proportion of the point clouds and cluster the building-related points per building object, while addressing global point positioning inaccuracies (Wysocki et al., 2021b). Alternatively, when point clouds are not georeferenced or GIS datasets are unavailable, existing benchmark points, annotated as buildings, can be used as a pre-cluster for façade-related points. If the aforementioned cases are not satisfied, the façades must be extracted manually, or else clustering algorithms must be used, similar to (Zhu et al., 2020).

4. RESULTS

4.1 Potential of existing point cloud benchmarks for façade segmentation

We analyzed 18 point cloud benchmark datasets (i.e., 17 existing ones plus our dataset), the results of which are presented in Table 2. As expected, most of the datasets were not created for façade segmentation tasks, with only TUM-FAÇADE (Wysocki et al., 2021c) and ArCH (Matrone et al., 2020) being designed specifically for this task. However, the rest of the set revealed significant extension potential for façade segmentation testing.

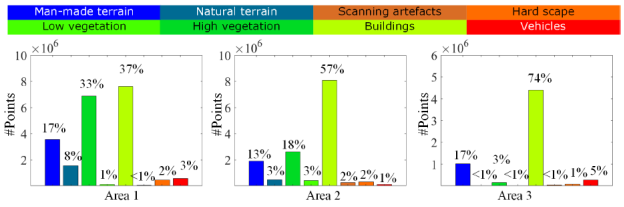

Figure 2. Ratio of annotated semantic classes in the TUM-MLS-2016 benchmark dataset.

Remarkably, we noticed an increase in the number of datasets released in recent years. The earliest benchmark in our set is from 2009 (Munoz et al., 2009), and was the only one to be published that year. In contrast, in the last three years (i.e., 2020-2022) as many as eight were published, which was almost 50% of our set. It is worth noting that no acquisition date is given for the Robotic 3D Scan Repository, either because it is absent in some of the repository’s point clouds or varies between them (Nüchter and Lingemann, 2016).

The analyzed list was dominated by MLS platforms, ranging from an early-type scanner mounted on a car to map a campus in Oakland, Canada (Munoz et al., 2009), through a backpack mapping unit in the city center of Basel, Switzerland (Blaser et al., 2021), to a mixture of dense mapping and simulated point clouds in Paris, France (Deschaud et al., 2021). Despite this, we included four terrestrial laser scanning (TLS) point cloud datasets, as well as the ArCH dataset that combines measurements from TLS, MLS, an unmanned aerial vehicle (UAV), and terrestrial photogrammetry (TP) measurements (Matrone et al., 2020).

As expected, each point cloud dataset consisted of points with their respective XYZ positions. However, additional scalar fields varied across the benchmarks. The intensity values were dominant, being present in 12 datasets of the set. On the other hand, RGB values occurred in eight datasets, while normals in only two datasets.

Interestingly, our set consisted not only of real-world point clouds but also included synthetic-world point clouds. This was the case with both SynthCity (Griffiths and Boehm, 2019b) and Paris-CARLA-3D (Deschaud et al., 2021). The former presented a completely simulated MLS point cloud based on a vector model and covering a combination of European mainland cities and New York, USA (Griffiths and Boehm, 2019b). The latter combined acquired point clouds with simulated ones using the CARLA environment (Deschaud et al., 2021).

One distinct advantage of synthetic point clouds was that they can easily outnumber the real ones: 700 M to 60 M in the case of Paris-CARLA-3D (Deschaud et al., 2021). Still, even simulated total points numbers were lower than the ones of TLS datasets: the semantic3D.net TLS dataset consisted of 4 BN points (Hackel et al., 2017). On the other hand, the KITTI-360 MLS dataset featured 1 BN points, with 73.7 km of roads being measured in Karlsruhe, Germany (Liao et al., 2021). It was thus clear that quantity of points was directly linked to the pace of acquisition (e.g., MLS is intuitively faster than TLS) and the total covered area. The latter is particularly difficult to acquire and compare, since the various datasets have different ways of quantifing this measure, namely: as a number of scenes, the length of road driven, approximate area extent, or else it is unpublished. Moreover, three obtained datasets (Nüchter and Lingemann, 2016, Lande, 2012, De Deuge et al., 2013) did not reveal their total number of points, thereby this statistic is omitted in these cases.

Regarding point clouds georeferencing, ou analysis showed that the set was equally divided between those published in a local CRS and in those in a global CRS, both had a score of eight datasets. It should be noted that semantic3D.net (Hackel et al., 2017), and ArCH (Matrone et al., 2020), provided a description of an acquisition place albeit they were in a local CRS. Thus, it should be possible to obtain a rough georeference of the point clouds.

Curiously, the majority of the datasets were located in Europe, while two were from North America (Munoz et al., 2009, Tan et al., 2020), and there was only one representative both from Asia (Dong et al., 2020) and Australia (De Deuge et al., 2013). It should be noted that the SynthCity dataset represented a mixture of virtual models from New York, USA, and mainland Europe and thus represented the simulated environment of North America and Europe (Griffiths and Boehm, 2019b).

The most remarkable result to emerge from the analysis is that buildings represent the majority of points in the datasets. For example, as we show in Figure 2, ratio of points per building class in the three TUM-MLS-2016 dataset areas outnumbered other classes with scores of 37%, 57%, and 74% (Zhu et al., 2020). On the other hand, along with the rising classes number, the ratio

of points per class vanished. On average, the number of classes equaled 16.5, with a maximum of 50 and a minimum of 0. For instance, the Paris-Lille-3D dataset distinguished between 50 different classes, which resulted in classes such as *table* or *mobile scooter*, represented by only 576 and 131 points, respectively (Roynard et al., 2018). As anticipated, the datasets that focused on the registration of point clouds (Dong et al., 2020, Nüchter and Lingemann, 2016, Blaser et al., 2021, Lande, 2012) excluded semantic classes.

However, even though the buildings were often annotated, most of the façade-level classes were still absent. Apart from our TUM-FAÇADE, only the Oakland 3D, Paris-rue-Madame, and ArCH datasets had incorporated façade-level classes in their repositories (Munoz et al., 2009, Serna et al., 2014, Matrone et al., 2020). Yet, although Oakland 3D had a list of several façade-level classes, they were underrepresented; for example, there were 500 and 100 points per stairs and gate classes, respectively (Munoz et al., 2009). The Paris-rue-Madame dataset had a few classes with façades details limited to wall lights, wall signs, and balcony plants (Serna et al., 2014). On the other hand, the ArCH dataset, which was designed for façade segmentation purposes, had a rich set of façade-related classes (Matrone et al., 2020).

As we had stipulated that our set had to include point clouds encompassing buildings, most of the datasets' scenes are urban. Some, such as A2D2 (Geyer et al., 2020), included rural areas, too. It is worth mentioning that in our set we rejected point clouds from Whu-TLS (Dong et al., 2020), Robotic 3D Scan Repository (Nüchter and Lingemann, 2016), and ETH PRS (Lande, 2012), that were indoor or building-unrelated.

Moreover, we identified several drawbacks in the currently available point cloud benchmark datasets that hinder effective testing of façade segmentation methods, namely:

- *Lack of façade-level classes*: As we present in the *façade-level classes?* column in Table 2, most of the benchmarks do not have any façade-grade classes, which hampers any comparison of methods conducted at such a fine granularity.
- *Lack of standardization in façade-level classes*: The classes are inconsistently named and annotated between the datasets and so the meanings of the objects can be confusing, which hinders methods comparison. This is exacerbated by the significant variation in the numbers of classes, too, as can be seen in the *# Classes?* column in Table 2.
- *Low variability of façades*: The datasets are limited in a number of façades, with often similar architectural styles. This phenomenon can bias algorithms towards overfitting to a particular architectural style and thus limit their generalization. It can also limit distinction capabilities between important classes such as doors and windows (Matrone et al., 2020). For instance, although the TerraMobilita/iQmulus dataset yields high density point clouds, it is limited to a 200 m long survey covering merely a few façades (Vallet et al., 2015).
- *Low ratio of façade-level points per class*: Even when façade-grade classes are available, the number of points per class is low. This means that the algorithms can be biased towards highly represented classes (e.g., walls) and neglect the underrepresented ones (e.g., doors). We illustrate this drawback in Figure 3, by analyzing the Oakland 3D point cloud dataset (Munoz et al., 2009), which is a perfect example of the identified trend.
- *Lack of georeferencing*: Many benchmarks do not contain information about the position with reference to the global CRS; They are often provided in a local CRS, as shown in Table 2. This excludes or at best hinders a comparison of methods using multimodal sources, such as point clouds in conjunction with 2D or 3D GIS datasets, such as in (Murtiyoso and Grussenmeyer, 2019) or (Wysocki et al., 2021a).
- *Lack of 3D reference building models at LoD3*: Point cloud semantic segmentation algorithms can only be validated against ground-truth labels in point clouds. This means that it is impossible to perform a second-tier validation (e.g., for methods addressing occlusions in point clouds). The application of open semantic volumetric- or surface-based models compliant with at LoD3, should enable this process, however. With such models, the benchmarks could be used for both segmentation and 3D reconstruction purposes.

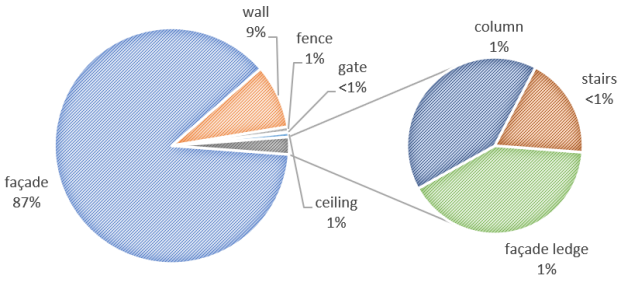

Figure 3. Ratio of annotated façade points per class and low façade classes variability, based on the example of the Oakland 3D point cloud dataset.

4.2 The TUM-FAÇADE benchmark

In this paper, we present TUM-FAÇADE: a point cloud benchmark dataset that aims to facilitate the development of façade segmentation methods (Wysocki et al., 2021c). Not only does it consists of 17 detailed ground-truth classes but it is based on the challenging MLS point cloud dataset, too. We created TUM-FAÇADE on the basis of the TUM-MLS-2016 benchmark dataset (Zhu et al., 2020), as it featured a challenging, urban environment, with realistic, dense, and georeferenced MLS point clouds. TUM-FAÇADE consists of five annotated and five non-annotated buildings replicating 14 and 15 façades, respectively. There are 17 annotated classes that range from features such as windows to drainpipes, as we show in Figure 4 and Table 3. We incorporated local and georeferenced XYZ positions in the scalar fields, together with the respective labels. Optionally, the dataset can be enhanced by adding intensity values from TUM-MLS-2016, too.

To create this dataset, we transformed raw TUM-MLS-2016 point clouds (i.e., 1.7 BN points) to global CRS using the transformation matrix included in the TUM-MLS-2016 benchmark repository (Zhu et al., 2020). Having models aligned in global CRS (EPSG: 25832) allowed us to encircle selected building entities with a 3 m buffer, using accurate, cm-grade footprints of governmental CityGML LoD2 models[2]. To facilitate the annotation process, each of the selected building point cloud clusters was then shifted to a local CRS with an origin in the building's center.

To manually annotate five of these point cloud entities, we used the Semantic Segmentation Editor[3] software by the Hitachi Auto-

[2] `https://www.ldbv.bayern.de/produkte/3dprodukte/3d.html`

[3] `https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor`

Table 2. Analysis of potential point cloud benchmark datasets for façade segmentation methods testing.

Name	Year	Sensor	Scalar fields	World	# points	Georeferenced?	Region	# Classes	Building class?	Façade-level classes?	Scene
Oakland 3D	2009	MLS	X,Y,Z	real	1.6 M	✗	North America	44	✓	~	urban
ETH PRS	2012	TLS	X,Y,Z,I	real	✗	✗	Europe	0	✗	✗	urban
Sydney Urban Objects Dataset	2013	MLS	X,Y,Z,I	real	✗	✗	Australia	26	✓	✗	urban
Paris-rue-Madame database	2014	MLS	X,Y,Z,I	real	20 M	✓	Europe	27	✓	~	urban
iQumulus	2015	MLS	X,Y,Z, I	real	12 M	✓	Europe	8	✓	✗	urban
TUM-MLS-2016	2016	MLS	X,Y,Z,I	real	1.7 BN	✓	Europe	9	✓	✗	urban
semantic3D.net	2017	TLS	X,Y,Z, I, RGB	real	4 BN	✗	Europe	9	✓	✗	rural+urban
Paris-Lille-3D	2018	MLS	X,Y,Z,I	real	143 M	✓	Europe	50	✓	✗	urban
SynthCity	2019	MLS	X,Y,Z, RGB,N	synthetic	368 M	✗	North America/Europe	9	✓	✗	urban
A2D2	2020	MLS	X,Y,Z,I	real	387 M	✗	Europe	38	✓	✗	rural+urban
ArCH	2020	TLS/MLS/UAV/TP	X,Y,Z, RGB, N	real	136 M	✗	Europe	10	✓	✓	urban
Toronto-3D	2020	MLS	X,Y,Z,I, RGB	real	78 M	✓	North America	8	✓	✗	urban
Whu-TLS	2020	TLS	X,Y,Z,I, RGB	real	551 M	✗	Asia	0	✗	✗	rural+urban
BIMAGE Datasets	2021	MLS	X,Y,Z, RGB	real	840 M	✓	Europe	0	✗	✗	urban
KITTI-360	2021	MLS	X,Y,Z, RGB	real	1 BN	✓	Europe	19	✓	✗	urban
Paris-CARLA-3D	2021	MLS	X,Y,Z,I,RGB	real+synthetic	60 + 700 M	✓	Europe	23	✓	✗	urban
Robotic 3D Scan Repository	✗	TLS	X,Y,Z,I	real	✗	✗	Europe	0	✗	✗	rural+urban
TUM-FAÇADE	2021	MLS	X,Y,Z	real	118 M	✓	Europe	17	✓	✓	urban

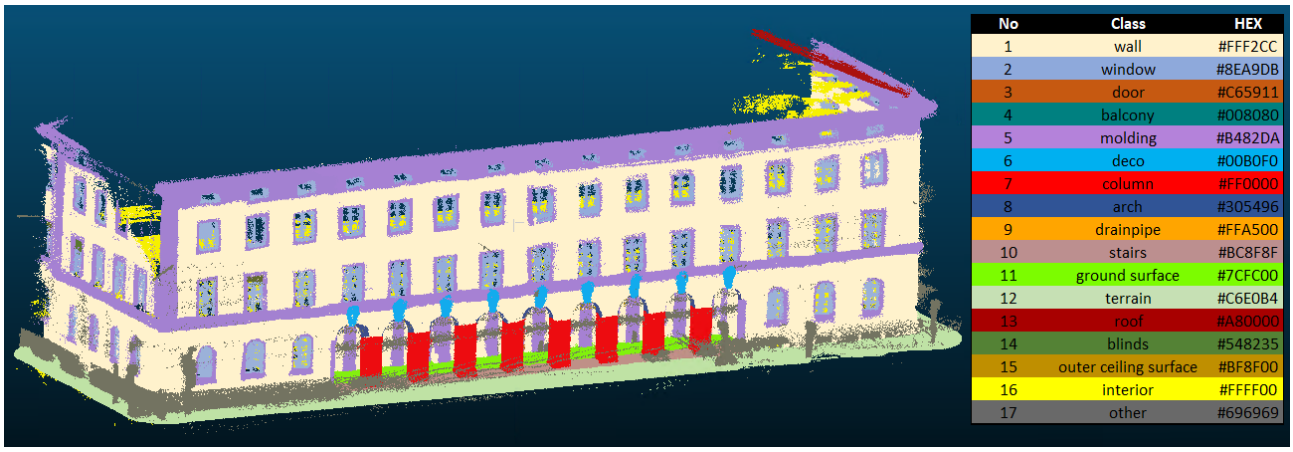

Figure 4. One of the TUM-FAÇADE buildings showing the color-coded points classes.

motive And Industry Laboratory. We extended its capabilities to enable it to accommodate our classes, presented in Table 1. The respective instructions and a configuration file are available under our repository (Wysocki et al., 2021c). We divided the building point clouds into smaller groups of 4 M points to address software, hardware capabilities, and the operator's ability to distinguish between different façade's features. Depending on the complexity of an object, labeling took between seven to 23 hours per building, with an estimated total of 83 hours for five buildings or approximately six hours per façade. It should be mentioned that the operator had little prior experience in working with the software. The second-tier, semi-automatic check was then performed to identify and correct any missing or false annotations. Once the validity check was completed, the previously used center-shift re-aligned the building to the global CRS, and HEX colors were added to the classes as appropiate, as shown in Figure 4. Another set of five non-annotated buildings for testing, as well as shift and HEX values, are published in our repository (Wysocki et al., 2021c).

Table 3. Annotated classes and points distribution in the TUM-FAÇADE dataset

#	Class	# points
1	wall	55,554,783
2	window	9,799,964
3	door	979,958
4	balcony	0
5	molding	13,497,145
6	deco	1,104,554
7	column	1,393,392
8	arch	220,774
9	drainpipe	29,398
10	stairs	419,409
11	ground surface	7,534,665
12	terrain	7,918,790
13	roof	74,035
14	blinds	547,288
15	outer ceiling surface	3,797,046
16	interior	9,477,868
17	other	5,347,086
Total		117,696,155

5. CONCLUSIONS

In this work, we present a comprehensive review of currently available point cloud benchmark datasets with the potential to be used for testing façade segmentation methods. We also name potential areas to be addressed in current and future benchmarks. To encourage further research and to maximize datasets' potential, we present TUM-FAÇADE, our façade-grade benchmark dataset (Wysocki et al., 2021c). It enriches the TUM-MLS-2016 benchmark dataset (Zhu et al., 2020), thereby we show that existing point cloud benchmark datasets can be seamlessly extended by adding façade-grade labels to widen the spectrum of benchmark dataset applications.

We anticipate that the segmentation façade classes we propose, will also facilitate the semantic 3D façade reconstruction process; the classes are derived from the established CityGML modeling standard (Gröger et al., 2012). As such, the classes can be used to identify semantic point cloud clusters and for semantic 3D façade reconstruction. This facilitates assigning CityGML city model functions, too. For example, the class *stairs* corresponds to the CityGML class *BuildingInstallation* and function *stairs 1013* (Special Interest Group 3D, 2020). This enables modeling of CityGML models at LoD3 (Gröger et al., 2012), or at so-called hybrid LoD with a façade at LoD3 and a roof structure at LoD2 (Biljecki et al., 2016). Moreover, this feature can also be used for linking the segmented point clouds to existing building models without explicit reconstruction, as demonstrated by (Beil et al., 2021).

Remarkably, our studies revealed that most of the available point cloud benchmarks not only include a building class but that this class also represents a majority of annotated ground-truth points in the datasets, as we show in Figure 2 and Table 2. Hence, we conclude that most of the existing point cloud benchmarks, although not specifically intended for façade segmentation testing, can be seamlessly extended to serve that purpose.

Moreover, our semantic statistics corroborate that typical MLS point clouds can capture fine façade details. However, they significantly omit roof structures (e.g., only 6% of points cover roofs in the TUM-FAÇADE benchmark), as we show in Table 3. Thus, as anticipated, MLS point clouds are inappropriate for roof segmentation testing purposes.

Nevertheless, we observe that some classes among the various benchmark datasets are inconsistent. This hampers the development of generic methods that can be tested on various datasets. Therefore, to facilitate such developments, we present 17 classes for façade-related annotations. We believe that they can be used as a set of blueprint classes for further research.

It remains the case that the outstanding challenge of having

overlying ground-truth information of surface- or volumetric-based 3D models with terrestrial point clouds, has not yet been solved (Xu and Stilla, 2021). It is worth noting that for several cities and regions city models have been released as open data[4]. Yet, they barely overlap with terrestrial point cloud benchmarks, as in the Ingolstadt's LoD3 building models[5] case. One of the exceptions is the BIMAGE dataset (Blaser et al., 2021) acquired in Basel, Switzerland, which can be superimposed on open, country-wide, semantic building models. However, these models are limited in their façades representation, as they consist of LoD2 and not LoD3 building models. We believe that this challenge will be the subject of future research.

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. The work was also carried out within the framework of the Leonhard Obermeyer Center at the Technical University of Munich (TUM). We gratefully acknowledge the team from the Chair of Geoinformatics TUM for their valuable insights and for providing the CityGML datasets. We are indebted to Jiarui Zhang for his diligent work in the data annotation process. The authors would like to thank (Zhu et al., 2020) for providing us with Figure 2.

REFERENCES

Apel, H., Aronica, G., Kreibich, H. and Thieken, A., 2009. Flood risk analyses—how detailed do we need to be? *Natural hazards* 49(1), pp. 79–98.

Armeni, I., Sax, S., Zamir, A. and Savarese, S., 2017. Joint 2D-3D-semantic data for indoor scene understanding. *ArXiv preprint:1702.01105*.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Gall, J. and Stachniss, C., 2021. Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset. *The International Journal on Robotics Research* 40(8-9), pp. 959–967.

Beil, C., Kutzner, T., Schwab, B., Willenborg, B., Gawronski, A. and Kolbe, T. H., 2021. Integration of 3D point clouds with semantic 3D city models – providing semantic information beyond classification. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* VIII-4/W2-2021, pp. 105–112.

Biljecki, F., Ledoux, H. and Stoter, J., 2016. An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems* 59, pp. 25–37.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S. and Çöltekin, A., 2015. Applications of 3D City Models: State of the Art Review. *ISPRS International Journal of Geo-Information* 4(4), pp. 2842–2889.

Blaser, S., Meyer, J. and Nebiker, S., 2021. Open urban and forest datasets from a high-performance mobile mapping backpack – a contribution for advancing the creation of digital city twins. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B1-2021, pp. 125–131.

De Deuge, M., Quadros, A., Hung, C. and Douillard, B., 2013. Unsupervised feature learning for classification of outdoor 3D scans. In: *Australasian Conference on Robitics and Automation, University of New South Wales: Kensington, Australia, 2013*, Vol. 2, p. 1.

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

[5] https://github.com/savenow/lod3-road-space-models

Deschaud, J.-E., Duque, D., Richa, J. P., Velasco-Forero, S., Marcotegui, B. and Goulette, F., 2021. Paris-CARLA-3D: A real and synthetic outdoor point cloud dataset for challenging tasks in 3D mapping. *Remote Sensing* 13(22), pp. 4713.

Dong, Z., Liang, F., Yang, B., Xu, Y., Zang, Y., Li, J., Wang, Y., Dai, W., Fan, H., Hyyppä, J. et al., 2020. Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* 163, pp. 327–342.

Fan, H., Wang, Y. and Gong, J., 2021. Layout graph model for semantic façade reconstruction using laser point clouds. *Geospatial Information Science* 24(3), pp. 403–421.

Gadde, R., Marlet, R. and Paragios, N., 2016. Learning grammars for architecture-specific facade parsing. *International Journal of Computer Vision* 117(3), pp. 290–316.

Geiger, A., Lenz, P., Stiller, C. and Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research* 32(11), pp. 1231–1237.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A. S., Hauswald, L., Pham, V. H., Mühlegg, M., Dorn, S. et al., 2020. A2D2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*.

Griffiths, D. and Boehm, J., 2019a. A review on deep learning techniques for 3D sensed data classification. *Remote Sensing* 11, pp. 1499–1528.

Griffiths, D. and Boehm, J., 2019b. SynthCity: A large scale synthetic point cloud. *arXiv preprint arXiv:1907.04758*.

Grilli, E. and Remondino, F., 2019. Classification of 3D digital heritage. *Remote Sensing* 11(7), pp. 847.

Gröger, G., Kolbe, T. H., Nagel, C. and Häfele, K.-H., 2012. OGC City Geography Markup Language CityGML Encoding Standard. Open Geospatial Consortium: Wayland, MA, USA, 2012.

Haala, N. and Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing* 65(6), pp. 570 – 580.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K. and Pollefeys, M., 2017. Semantic3d.net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847*.

Lande, M. B., 2012. Automatic registration of partially overlapping terrestrial laser scanner point clouds. `https://prs.igp.ethz.ch/research/completed_projects/automatic_registration_of_point_clouds.html`. Accessed: 2020-10-30.

Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M. A., Cao, D. and Li, J., 2020. Deep learning for LiDAR point clouds in autonomous driving: a review. *IEEE Transactions on Neural Networks and Learning Systems*.

Liao, Y., Xie, J. and Geiger, A., 2021. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *arXiv preprint arXiv:2109.13410*.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y. and Hoi, S. C., 2020. DeepFacade: A deep learning approach to facade parsing with symmetric loss. *IEEE Transactions on Multimedia* 22(12), pp. 3153–3165.

Martinovic, A., Knopp, J., Riemenschneider, H. and Van Gool, L., 2015. 3D all the way: Semantic segmentation of urban scenes from start to end in 3D. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*, pp. 4456–4465.

Mathias, M., Martinović, A. and Van Gool, L., 2016. ATLAS: A three-layered approach to facade parsing. *International Journal of Computer Vision* 118(1), pp. 22–48.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R. and Remondino, F., 2020. Comparing machine and deep learning methods for large 3D heritage semantic segmentation. *ISPRS International Journal of Geo-Information* 9(9), pp. 535.

Montazeri, H. and Blocken, B., 2013. CFD simulation of wind-induced pressure coefficients on buildings with and without balconies: Validation and sensitivity analysis. *Building and Environment* 60, pp. 137–149.

Müller, P., Zeng, G., Wonka, P. and Van Gool, L., 2007. Image-based procedural modeling of facades. *ACM Transactions on Graphics* 26(3), pp. 85.

Munoz, D., Bagnell, J. A. D., Vandapel, N. and Hebert, M., 2009. Contextual classification with functional Max-Margin Markov networks. In: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA , 20–25 June 2009*, pp. 975 – 982.

Murtiyoso, A. and Grussenmeyer, P., 2019. Point cloud segmentation and semantic annotation aided by GIS data for heritage complexes. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W9, pp. 523–528.

Musialski, P., Wonka, P., Aliaga, D. G., Wimmer, M., Van Gool, L. and Purgathofer, W., 2013. A survey of urban reconstruction. *Computer graphics forum* 32(6), pp. 146–177.

Nouvel, R., Schulte, C., Eicker, U., Pietruschka, D. and Coors, V., 2013. CityGML-based 3D city model for energy diagnostics and urban energy policy support. In: *Proceedings of the 5th German-Austrian IBPSA Conference (BauSIM 2014), Aachen, Germany, 22–24 September 2014*, pp. 83–90.

Nüchter, A. and Lingemann, K., 2016. Robotic 3D Scan Repository. `http://kos.informatik.uni-osnabrueck.de/3Dscans/`. Accessed: 2020-10-30.

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D. and Bischof, H., 2012. 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.

Roynard, X., Deschaud, J.-E. and Goulette, F., 2018. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research* 37(6), pp. 545–557.

Schwab, B. and Kolbe, T. H., 2019. Requirement analysis of 3D road space models for automated driving. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-4/W8, pp. 99–106.

Serna, A., Marcotegui, B., Goulette, F. and Deschaud, J.-E., 2014. Paris-rue-Madame database: As 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods. In: *Proceedings of the International Conference on Pattern Recognition Applications and Methods. ACM, Angers, France, 6–8 March*, pp. 819–824.

Special Interest Group 3D, 2020. Modeling guide for 3D objects - Part 2: Modeling of buildings (LoD1, LoD2, LoD3) - SIG3D quality wiki EN. `https://en.wiki.quality.sig3d.org/index.php?title=Modeling_Guide_for_3D_Objects_-_Part_2:_Modeling_of_Buildings_(LoD1,_LoD2,_LoD3)`. Accessed: 2020-10-30.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K. and Li, J., 2020. Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 202–203.

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P. and Paragios, N., 2012. Parsing facades with shape grammars and reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(7), pp. 1744–1756.

Tyleček, R. and Šára, R., 2013. Spatial pattern templates for recognition of objects with regular structure. In: *German Conference on Pattern Recognition, Saarbrücken, Germany, September 3-6, 2013*, Springer, pp. 364–374.

Vallet, B., Brédif, M., Serna, A., Marcotegui, B. and Paparoditis, N., 2015. TerraMobilita/iQmulus urban point cloud analysis benchmark. *Computers & Graphics* 49, pp. 126–133.

Varney, N., Asari, V. K. and Graehling, Q., 2020. DALES: A large-scale aerial LiDAR data set for semantic segmentation. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14-19 June 2020*, pp. 717–726.

Willenborg, B., Pültz, M. and Kolbe, T. H., 2018. 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, pp. 223–230.

Wysocki, O., Schwab, B., Hoegner, L., Kolbe, T. H. and Stilla, U., 2021a. Plastic surgery for 3D city models: A pipeline for automatic geometry refinement and semantic enrichment. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-4-2021, pp. 17–24.

Wysocki, O., Xu, Y. and Stilla, U., 2021b. Unlocking point cloud potential: Fusing MLS point clouds with semantic 3D building models while considering uncertainty. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* VIII-4/W2-2021, pp. 45–52.

Wysocki, O., Zhang, J. and Stilla, U., 2021c. TUM-FAÇADE, Technical University of Munich. `https://mediatum.ub.tum.de/1636761`. Accessed: 2020-12-01.

Xu, Y. and Stilla, U., 2021. 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, pp. 2857–2885.

Zhu, J., Gehrung, J., Huang, R., Borgmann, B., Sun, Z., Hoegner, L., Hebel, M., Xu, Y. and Stilla, U., 2020. 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), pp. 1875.

Zolanvari, S. I. and Laefer, D. F., 2016. Slicing method for curved façade and window extraction from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* 119, pp. 334–346.