A learning approach to evaluate the quality of 3D city models

Oussama Ennafii, Arnaud Le Bris, Florent Lafarge, Clément Mallet

A learning approach to evaluate the quality of 3D city models

Oussama Ennafii[a,b,*], Arnaud Le Bris[a], Florent Lafarge[b], Clément Mallet[a]

[a] *Univ. Paris-Est, LaSTIG STRUDEL, IGN, ENSG, F-94160 Saint-Mandé, France*
[b] *INRIA, Titane, 06902 Sophia Antipolis, France*

Abstract

The automatic generation of 3D building models from geospatial data is now a standard procedure. An abundant literature covers the last two decades and several softwares are now available. However, urban areas are very complex environments. Inevitably, practitioners still have to visually assess, at city-scale, the correctness of these models and detect frequent reconstruction errors. Such a process relies on experts, and is highly time-consuming with approximately two hours/km^2 per expert. This work proposes an approach for automatically evaluating the quality of 3D building models. Potential errors are compiled in a novel hierarchical and versatile taxonomy. This allows, for the first time, to disentangle fidelity and modeling errors, whatever the level of details of the modeled buildings. The quality of models is predicted using the geometric properties of buildings and, when available, Very High Resolution images and Digital Surface Models. A baseline of handcrafted, yet generic, features is fed into a Random Forest classifier. Both multi-class and multi-label cases are considered: due to the interdependence between classes of errors, it is possible to retrieve all errors at the same time while simply predicting correct and erroneous buildings. The proposed framework was tested on three distinct urban areas in France with more than 3,000 buildings. 80%-99% F-score values are attained for the most frequent errors. For scalability purposes, the impact of the urban area composition on the error prediction was also studied, in terms of transferability, generalization, and representativeness of the classifiers. It shown the necessity of multimodal remote sensing data and mixing training samples from various cities to ensure a stability of the detection ratios, even with very limited training set sizes.

Keywords: 3D urban modeling, buildings, quality assessment, taxonomy, classification, error detection, geometry, aerial imagery, Very High Spatial Resolution, Digital Surface Model.

[*] Corresponding author
Email address: oussama.ennnafii@ign.fr (Oussama Ennafii)
URL: http://recherche.ign.fr/labos/matis/~ennafii (Oussama Ennafii)

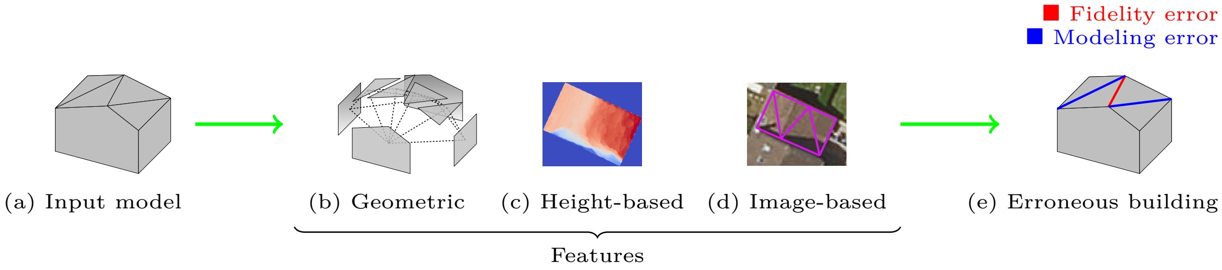
1. Introduction

1.1. Context and objectives

3D urban models have a wide range of applications. They can be used for consumer purposes (video games or tourism) as much as they can be vital in more critical domains with significant societal challenges (*e.g.*, disaster control, run-off water or microclimate simulation, urban planning or security operations preparation) (Musialski et al., 2012; Biljecki et al., 2015). Therefore, automatic urban reconstruction from geospatial imagery (spatial/airborne sensors) focuses efforts on both scientific research and industrial activities. 3D city modeling has therefore been deeply explored in the photogrammetric, GIS, computer vision, and computer graphics literature with an emphasis on compactness, full automation, robustness to acquisition constraints, scalability, and inevitably, quality (Müller et al., 2006; Over et al., 2010; Vanegas et al., 2010; Lafarge and Mallet, 2012; Poli and Caravaggi, 2013; Stoter et al., 2013; Zhou and Neumann, 2013; Cabezas et al., 2015; Monszpart et al., 2015; Kelly et al., 2017; Nguatem and Mayer, 2017). However, the problem remains partly unsolved (Sester et al., 2011; Musialski et al., 2012; Rottensteiner et al., 2014). In fact, besides the seamless nature of reconstituted models, current algorithms lack of generalization capacity. They cannot handle the high heterogeneity of urban landscapes. As such, for operational purposes, human intervention is needed either in interaction within the reconstruction pipeline or as a post-processing refinement and correction step. The latter is highly tedious: it consists in an individual visual inspection of buildings (Musialski et al., 2012). Consequently, automatizing the last step remains, for all stakeholders (from researchers up to end-users), a critical step, especially in a production environment. It has been barely investigated in the literature. This paper addresses this issue by expanding earlier work (Ennafii et al., 2019).

1.2. Qualifying 3D building models

Our work focuses on assessing polyhedral structured 3D models, representing building architectures (Haala and Kada, 2010). These models result from a given urban reconstruction method, that is unknown from our evaluation pipeline. We discard triangle meshes that are standardly generated from multiview images or point clouds with state-of-the-art surface reconstruction methods. Here, the studied objects are, by design, more compact but less faithful to input data. In counterpart, they hold more semantic information: each polygonal facet typically corresponds to a façade, a roof, or any other architecturally atomic feature of a building. 3D modeling algorithms traditionally build a trade-off between representation compactness and fidelity to the input data (meshes or 3D points).



(a) Input model (b) Geometric (c) Height-based (d) Image-based (e) Erroneous building

Figure 1: Our semantic evaluation framework for 3D building models (a). Semantic errors affecting the building are predicted using a supervised classifier and handcrafted features. In addition to the input model topological structure (b), features are extracted from Very High resolution overhead data. It can be based on a comparison with the Digital Surface Model (DSM) (c). Optical images can also be used through, for instance, local gradient extraction (d). Several errors can be detected at the same time, in a hierarchical manner (e). Fidelity errors correspond to geometrical imprecision as shown in red. On the other hand, modeling errors denote morphological inconsistancies with the real object.

Depending on its spatial accuracy, the urban setting, and the targeted application, the reconstituted result achieves a certain **Level of Detail (LoD)** (Kolbe et al., 2005). A LoD-1 model is a simple building extrusion (flat roof) (Ledoux and Meijers, 2011; Biljecki et al., 2017). A LoD-2 model considers geometric simplification of buildings, ignoring superstructures, such as dormer windows and chimneys (Taillandier and Deriche, 2004). These are taken into account in LoD-3 (Brédif et al., 2007). The LoD rational is still open for debate (Biljecki et al., 2016b). Nevertheless, in this paper, we will follow the LoD categorization introduced above, which is also standard in the computer vision and graphics literature.

A large body of papers has addressed the 3D building modeling issue and subsequently tried to find the trade-off between fidelity and compactness (Dick et al., 2004; Zebedin et al., 2008; Lafarge et al., 2010; Verdie et al., 2015). Conversely, few works investigate the issue of assessing the quality of the derived models, especially out of a given reconstruction pipeline (Schuster and Weidner, 2003). Usually, quality assessment is based on visual inspection (Durupt and Taillandier, 2006; Macay Moreia et al., 2013), geometric fidelity metrics (Kaartinen et al., 2005), or by extending standard 2D object detection criteria (Karantzalos and Paragios, 2010), without any semantic dimension. Only one benchmark dataset has addressed the issue (Rottensteiner et al., 2014). It remains focused on very few areas and a geometric comparison with manually extracted roof structures (Li et al., 2016; Nan and Wonka, 2017; Nguatem and Mayer, 2017). Consequently, it cannot be easily extended. Similar conclusions can be drawn for indoor reconstruction (Tran et al., 2019).

1.3. Positioning and contributions

The current situation motivates the need for a well suited quality assessment paradigm. Since the building models display strong structural properties, an unconstrained evaluation based on data fidelity metrics, as in (Berger et al., 2013), is too general. The evaluation should also ignore format issues or geometric consistencies as proposed in (Ledoux, 2018). Although being a serious issue and clean 3D models are usually not the norm (Biljecki et al., 2016a; Hu et al., 2018), we rule out, at this stage, these cases for simplicity. Instead, we target a *semantic* evaluation in which building semantics is taken into account through the detection and categorization of modeling errors at the facet level for each 3D building. The framework is independent from the LoD and the modeling method. The standard criteria used in the reconstruction process (*e.g.*, L_1 norm between the model and a DSM) will not be taken into account, as they are usually chosen as minimization targets in the modeling procedure. Thus, we define an evaluation framework that can be used for:

- **Building model correction**: for the automatic or interactive (Kowdle et al., 2011) refinement of building models using the detected errors.

- **Change detection**: modeling errors can straightforwardly stem from changes, which frequently occur in urban environments (Taneja et al., 2015). Conversely, changes can be implicitly detected from other defects (Tran et al., 2019).

- **Reconstruction method selection**: evaluating models from various reconstruction algorithms can allow assessing which method(s) is(are) the most adapted for a specific LoD and building type.

- **Crowd-sourcing evaluation** (Kovashka et al., 2016): categorizing user behaviors during crowd-sourced modeling and vandalism detection process (Neis et al., 2012).

This work proposes an adaptable and flexible framework indifferent to input urban scenes and reconstruction methods. For that purpose, our contributions are three-fold:

- A new **taxonomy of errors**, hierarchical, adapted to all LoDs, and independent from input models;

- A **supervised classification** formulation of the evaluation problem which predicts all errors affecting the building model;

- A multimodal **baseline of features** that are extracted from the model itself as well as from Very High Resolution external data (optical images and height data).

Section 2 introduces the problem of the evaluation of 3D building models and discusses existing methods. Section 3 details the proposed approach, while data and experiments conducted over three urban areas are presented in Section 4. A more comprehensive set of experiments studying the scalability of the proposed method is reported in Section 5. The same experiments are conducted at other semantic levels and recorded in Section 6. Main conclusions are drawn in Section 7.

2. Related Work

Quality assessment methods can be classified according to two main criteria: reference data and output type.

2.1. Reference data types

Existing methods rely on two types of reference data.

Manually plotted ground truth data with very high spatial accuracy: these models can be obtained either from field measurements (Dick et al., 2004; Kaartinen et al., 2005) with the highest possible precision ($\sigma(\text{error}) \approx 0.05\,\text{m}$), or using stereo-plotting techniques (Jaynes et al., 2003; Kaartinen et al., 2005; Zebedin et al., 2008; Zeng et al., 2014). Generally, the criterion is the root mean squared error on the height values. Such a strategy does not scale well, does not straightforwardly bring semantics, and requires a 3D matching procedure (overlapping ratio between surfaces, minimal roof areas, integration of superstructures) that can be complex in dense urban environments.

Raw remote sensing data: models can either be compared to the source that allowed the generation of the models or remote sensing data of superior geometric accuracy: LiDAR point clouds, height maps (*i.e.*, DSMs) (Akca et al., 2010; Lafarge and Mallet, 2012; Li et al., 2016; Zhu et al., 2018) or multi-view Very High Resolution images as in (Boudet et al., 2006; Michelin et al., 2013). Despite the fact such strategy better samples the area of interest, it may not always be helpful. On one hand, they have been exploited by the modeling methods and such comparisons are often the basis for their fidelity criterion. On the other hand, additional remote sensing data is not easy to obtain, especially at large scales under operational constraints.

2.2. Evaluation outputs

The quality assessment methods can produce two kinds of outputs: geometric fidelity metrics and labels of errors.

Geometric fidelity metrics summarize the quality at the building level. These criteria are computed at different levels: average precision of specific points of interest (corners or edge points, (Vögtle and Steinle, 2003; Kaartinen et al., 2005), surface dissimilarity (Jaynes et al., 2003; Dick et al., 2004; Kaartinen et al., 2005; Zebedin et al., 2008; Lafarge and Mallet, 2012; Zeng et al., 2014; Li et al., 2016; Nan and Wonka, 2017), average mean absolute distance (Duan and Lafarge, 2016; Zeng et al., 2018), tensor analysis of residuals (You and Lin, 2011) or volume discrepancy to reference data (Jaynes et al., 2003; Zeng et al., 2014; Nguatem and Mayer, 2017). Evaluation can also be performed according to compactness, which is complementary to fidelity metrics: number of faces/vertices in the model (Lafarge and Mallet, 2012; Zhang and Zhang, 2018). For both cases, the obtained outputs have the drawback of being too general for the special case of urban structured models. Far from surface reconstruction evaluation (Berger et al., 2013), it is preferred that a diagnosis pinpoints specific types of errors that can be easily corrected with specific procedures (Elberink and Vosselman, 2011).

Semantic errors identify topological and geometric errors that affect building models. One example of such defects is the traffic light paradigm ("correct", "acceptable/generalized", and "incorrect") (Boudet et al., 2006). However, these errors depend on the definition of the end-user oriented nomenclature and a specific "generalization" level at which models are rejected. In addition, this taxonomy does not help in localizing the model shortcomings. Another solution is to look at the issue at hand through the used reconstruction algorithm perspective. For instance, defects are discriminated in Michelin et al. (2013), between footprint errors ("erroneous outline", "inexisting building", "missing inner court" and "imprecise footprint"), intrinsic reconstruction errors ("over-segmentation", "under segmentation", "inexact roof", and "Z translation"), and "vegetation occlusion" errors or are considered only for roof topology as in (Xiong et al., 2014) ("Missing Node", "False Node", "Missing Edge", and "False Edge"). In most of these methods, the evaluation is cast as a supervised classification process: the predicted classes are defects listed in an established taxonomy. Features used for this classification are extracted from very high spatial resolution (VHR, 0.1 m to 0.5 m) images and DSMs, like 3D segments or texture correlation score comparisons. In spite of their semantic contribution in quality evaluation, such taxonomies are prone to overfitting to specific urban scenes/ modeling algorithms or require the computation of complex features on

VHR data that do not scale well.

2.3. Problem statement

This work aims to propose a new quality evaluation paradigm that detects and describes semantic errors that affect 3D building models. Two important characteristics must be taken into account. First, the definition of the semantic errors should not vary from one urban scene to another and guaranty independence to the underlying 3D reconstruction method. Secondly, the scalability of the method should be addressed in order to ensure the ability to correctly classify unseen areas and to define the minimal amount of data required. This is all the more necessary in case of limited training sets in order to avoid overfitting to a specific problem and environment.

3. Problem formulation

We start by establishing a novel hierarchical error taxonomy. It is **parameterizable**[1] and **agnostic** towards reconstructed models. Independence from the modeling method and the urban scenes is mandatory for generalization and transferability capacities. Depending on the evaluation objectives, we deduce error labels that can pinpoint defects altering the models. Their presence is predicted using a **supervised classifier**, trained with manually annotated data.

The quality assessment pipeline is constructed in order to be **modular**. Building models are represented by intrinsic geometric features extracted from the model facet graph. If available, the classifier can also be fed with remote sensing data: depth features based on the comparison of the model altimetry and a DSM or image information with spectral or textural features available in satellite, aerial or street view optical images.

3.1. Error taxonomy

In order to build a generic and flexible taxonomy, we rely on two criteria for error compilation: the building model LoD and the error semantic level, named henceforth *finesse* (cf. Figure 2). Different degrees of *finesse* describe, from coarse to fine, the specificity of defects. Errors with maximal *finesse* are called *atomic* errors. Multiple *atomic* errors can affect the same building. For instance, topological defects induce, almost always, geometrical ones. In practice, only independently coexisting *atomic* defects are reported. The idea is to provide the most relevant information to be able

[1] Parameters determine the considered errors

to correct a model. *Atomic* errors can thus be intuitively correlated to independent actions to be chosen by an operator or an algorithm so as to correct the model.

3.1.1. The general framework

The main idea of error hierarchization is to enable modularity in the taxonomy, and thus achieve a strong flexibility towards input urban scenes and desired error precision. A general layout is first drawn, followed by a more detailed error description.

At a first level, model qualifiability is studied. In fact, aside from formatting issues or geometric inconsistencies (Ledoux, 2018), other reasons make building models unqualifiable. For instance, buildings can be occluded by vegetation and thus cannot be assessed with most of the remote sensing data sources. Generally speaking, input models can be impaired by some pathological cases that are outside our evaluation framework. In consequence, *qualifiable* models are distinguished here from *unqualifiable* buildings. This first level corresponds to a *finesse* equal to 0. At the *finesse* level 1, we predict the correctness of all qualifiable buildings. It is the lowest semantization level at which the evaluation of a model is expressed. Then, a model is either *valid* or *erroneous*. Most state-of-the-art evaluation methods address this level.

Model errors are grouped into three families depending on the underlying LoD. The first family of errors "*Building Errors*" affects the building in its entirety. It corresponds to an accuracy evaluation at LoD-0 (footprint errors) ∪ LoD-1 (height/geometric error). At the next LoD-2, the family "*Facet Errors*" gathers defects that can alter façade or roof fidelity. The last error family, "*Superstructure Errors*", describes errors that involve superstructures modeled at LoD-3. Only the first two families are studied in this paper.

Each family contains *atomic* errors of maximal *finesse* equal to 3. Although they can co-occur in the same building model and across different families, these errors are semantically independent. They represent specific topological or geometric defects. Topological errors translate inaccurate structural modeling, while geometric defects raise positioning infidelity.

At evaluation time, three parameters play a role in determining which error labels to consider. The first is the **evaluation Level of Detail (eLoD)**. Every reconstruction method targets a certain set of LoDs. In consequence, when assessing a reconstruction, a LoD must be specified. At a given eLoD, all error families involving higher orders will be ignored. Depending on the target of the qualification process, a **finesse** level might be preferred. This second evaluation parameter specifies the appropriate semantic level at which errors will be reported. The last one is error

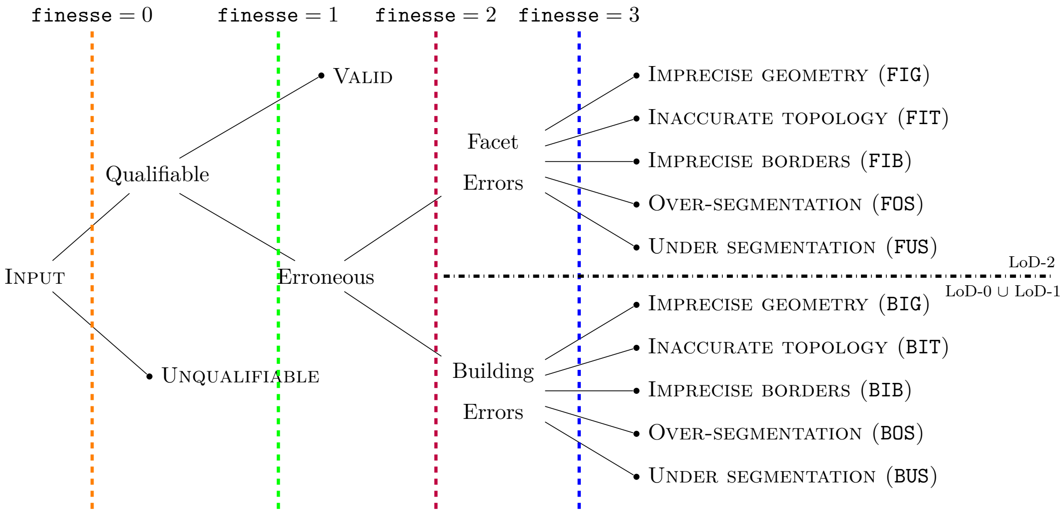


Figure 2: The proposed taxonomy structure. In our case of very high resolution overhead image modeling, only two family errors are depicted. At *finesse* level 2, hierarchization is possible: the **exclusivity** parameter can thus act. However, it is not the case at the *atomic* errors level since they are independent.

exclusivity. It conveys family error hierarchy. Errors of a given LoD$= l$ family are prioritized over ones with higher LoD$> l$.

3.1.2. Application to the geospatial overhead case

This paper tackles the aerial reconstruction case where the objective is to reconstruct large urban scenes using Very High Resolution geospatial images or, if available, lidar point clouds. The framework is general enough to encompass both orthorectified images and oblique ones. In this paper, we only used orthorectified images. In an ideal scenario, using oriented images is better for edge verification (as already shown in (Michelin et al., 2013)) as orthoimages are a byproduct of earlier ones. However, in practice, oblique imagery would give rise to other issues, especially, registration problems. Hereaafter, 3D buildings are evaluated. The *atomic* errors are (Figures 2 and 3):

Building errors family:

- *Under segmentation* (BUS): two or more buildings are modeled as one. In Figure 3.i.a, two distinct buildings were identified as one building, eventhough they can be visually distinguished.
- *Over-segmentation* (BOS): one building is subdivided into two or more buildings. This is the opposite of the previous situation. Figure 3.i.b shows a single building that, when modelled, was subdivided into three parts.

- *Imprecise borders* (BIB): at least one building footprint border is incorrectly located. A sample is shown in Figure 3.i.c.

- *Inaccurate topology* (BIT): the building footprint suffers from topological defects as missing inner courts or wrong primitive fitting (for instance, a circular footprint approximated by a polygon). In Figure 3.i.d, we illustrate how the footprint morphology can be erroneous. This error, as the earlier ones, result either from defective building identification process, or from an outdated cadastral map.

- *Imprecise geometry* (BIG): inaccurate building geometric estimation. In case **eLoD** $>$ LoD-0 $\cup$ LoD-1, this error is not reported as it becomes redundant with below delineated errors.

Facet errors family:

- *Under segmentation* (FUS): two or more facets are modeled as one, as illustrated in Figure 3.ii.a.

- *Over-segmentation* (FOS): one facet is subdivided into two or more facets. Refer to Figure 3.ii.b for an example.

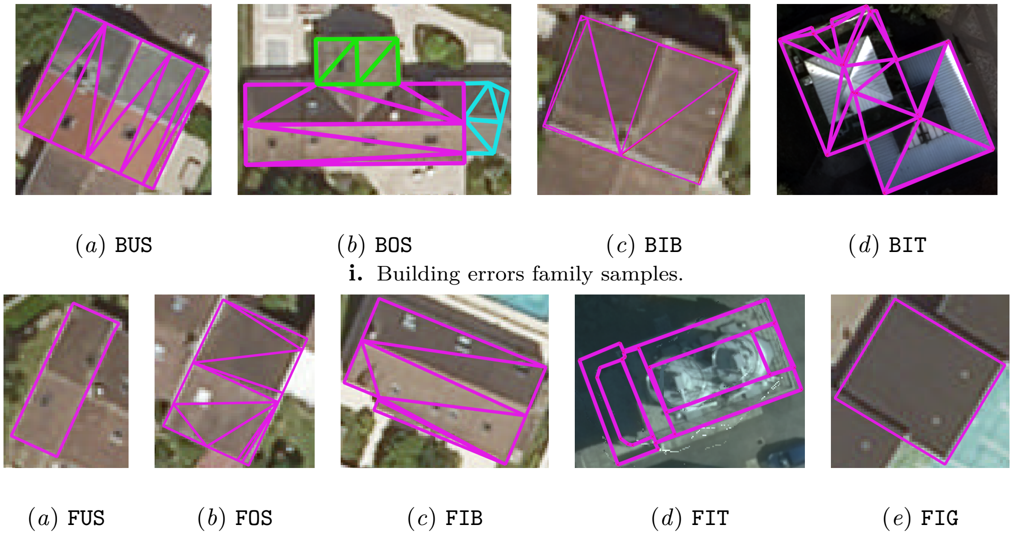
- *Imprecise borders* (FIB): at least one facet border is incorrectly located. As en example, Figure 3.ii.c shows that the central edge that links the two main roof sides does not correspond to the one on the image position.

- *Inaccurate topology* (FIT): the facet suffers from topological defects such as wrong primitive fitting (for example, a dome approximated by planar polygons). In Figure 3.ii.d), we can observe how two cylindrical towers were reconstructed as a rectangular parallelepiped.

- *Imprecise geometry* (FIG): inaccurate facet geometric estimation :*e.g.*, wrong height or inaccurate slope. The latter is depicted in Figure 3.ii.e[2]. All these errors stem either from the modeling approach, or from the poor spatial resolution of the input data (DSM or point cloud).

These errors can be related to state-of-the-art labels. For instance, “Missing Node” (*resp.* “False Node”, “Missing Edge” and “False Edge”) in Xiong et al. (2014) correspond to, or are

[2] There is a problem of slope. The model corresponds to a flat roof whereas in reality the slope is *ca.* 25°. The error could only be shown if we provided the height residual. However, for the sake of homogeneity, we only provided orthoimages as background. It motivates also the need for a height-based modality.



(*a*) BUS (*b*) BOS (*c*) BIB (*d*) BIT

i. Building errors family samples.

(*a*) FUS (*b*) FOS (*c*) FIB (*d*) FIT (*e*) FIG

ii. Facet errors family samples.

Figure 3: Illustration of various errors of our taxonomy. One can see that geometric, spectral and height information are required for an accurate detection of all kinds of errors.

included in, the topological *atomic* errors from the *Facet Errors* family: FUS (*resp.* FOS, FIT, and FIT). The difference is that we distinguish flaws that can affect superstructure facets (LoD-3) from the ones that impair building facets (LoD-2). The taxonomy developed by Michelin et al. (2013), on the other hand, is closer to ours. In fact, while footprint errors is reshuffled into *Building Errors* as BIB ("erroneous outline" and "imprecise footprint") and BIT ("missing inner court") , intrinsic reconstruction errors ("over-segmentation", "under segmentation", "inexact roof" and "Z translation") can be re-adapted into both family errors. Finally, "vegetation occlusion" and " non existing" are gathered into the *unqualifiable* label at *finesse* level 0. Boudet et al. (2006), however, studied the acceptability of a model in a whole. Their taxonomy cannot directly fit with our labels. The acceptability dimension can be incorporated into our framework by attributing an confidence score to each error: for example, a prediction probability.

3.2. Feature baseline

In order to detect such specific labels while guarantying a certain flexibility towards reference data, multiple modalities are necessary. The structure of the 3D model can be directly used to extract geometrical features. Dense depth information can be added, through for instance a DSM, so as to help detecting geometric defects that can be hardly discriminated otherwise (as in Figure 3.ii.(e)), in particular for the outer part of buildings. VHR optical images bring additional information (high frequencies and texture) that is particularly suited for inner defect detection.

Since there is no comparable work that studies the previously defined errors, we propose a baseline for each modality. Attributes are kept simple so as to be used in most situations relying on generally available data. We avoid computing and comparing 3D lines (Michelin et al., 2013), correlation scores (Boudet et al., 2006) or any Structure-from-Motion (SfM) based metric (Kowdle et al., 2011). In addition of being very costly, these features are methodologically redundant with the 3D modeling techniques: they are vulnerable to the same defects. Conversely, evaluation metrics used in the 3D building reconstruction literature (*e.g.*, Root Mean Square Error (RMSE)) are too weak for such a complex task.

3.2.1. Geometric features

The model facet set is denoted by F. $\forall (f,g) \in \mathsf{F} \times \mathsf{F} \quad f \sim g$ correspond to facets f and g being adjacent: *i.e.*, they share a common edge. As the roof topology graph in (Verma et al., 2006), the input building model M can be seen as a facet (dual) graph:

$$\mathsf{M} \triangleq \Big(\mathsf{F}, \mathsf{E} \triangleq \{\{f,g\} \subset \mathsf{F} : f \sim g\}\Big). \tag{1}$$

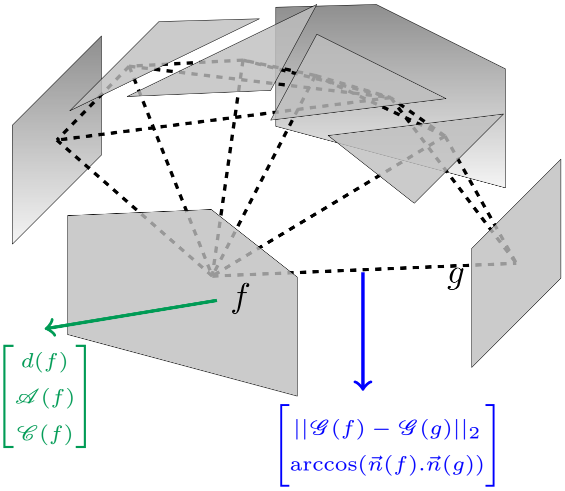

f
g
$d(f)$
$\mathscr{A}(f)$
$\mathscr{C}(f)$
$||\mathscr{G}(f) - \mathscr{G}(g)||_2$
$\arccos(\vec{n}(f).\vec{n}(g))$

Figure 4: Computed geometric attributes represented on the dual graph, for facets f and g. The green vector groups the node (facet) attributes while the blue one shows the edge features.

The dual graph is illustrated in Figure 4. For each facet $f \in \mathsf{F}$, we compute its degree (*i.e.*, number of vertices; $d(f) \triangleq |\{v : v \text{ is a vertex of } f\}|$), its area $\mathscr{A}(f)$ and its circumference $\mathscr{C}(f)$. For each graph edge $e = \{f,g\} \in \mathsf{E}$, we look for the distance between facet centroids $||\mathscr{G}(f) - \mathscr{G}(g)||$ and the angle formed by their normals $\arccos(\vec{n}(f), \vec{n}(g))$. Statistical characteristics are then computed over building model facets using specific functions S, like a histogram:

$$S = S^p_{hist} : l \mapsto histogram(l, p), \tag{2}$$

with p standing for histogram parameters. A simpler option could be:

$$S = S_{synth} : l \mapsto \Big[\max(l) \quad \min(l) \quad \bar{l} \quad \text{median}(l) \quad \sigma(l) \Big] , \tag{3}$$

where $\bar{l}$ (*resp.* $\sigma(l)$) represents the mean (*resp.* the standard deviation) over a tuple l.

These features are designed for general topological errors. For instance, over-segmentation may result in small facet areas and small angles between their normals. Conversely, an undersegmented facet would have a large area. Later on, the importance of these features will be discussed in details based on experimental results.

Each building M can consequently be characterized by a geometric feature vector that accounts for its geometric characteristics:

$$v_{geometric}(\mathsf{M}) = \begin{bmatrix} S\Big(\big(d(f)\big)_{f \in \mathsf{F}} \Big) \\ S\Big(\big(\mathscr{A}(f)\big)_{f \in \mathsf{F}} \Big) \\ S\Big(\big(\mathscr{C}(f)\big)_{f \in \mathsf{F}} \Big) \\ S\Big(\big(||\mathscr{G}(f) - \mathscr{G}(g)||\big)_{(f,g) \in \mathsf{E}} \Big) \\ S\Big(\big(\arccos(\vec{n}(f), \vec{n}(g))\big)_{(f,g) \in \mathsf{E}} \Big) \end{bmatrix} . \tag{4}$$

Additionally to individual facet statistics, regularity is taken into account by looking into adjacent graph nodes as in (Zhou and Neumann, 2012). Such features express a limited part of structural information. Taking this type of information into account would implicate graph comparisons which are not genuinely simple tasks to achieve. Since our objective is to build a baseline, this approach has not yet been considered.

3.2.2. Height-based features

For this modality, raw depth information is provided by a DSM as a 2D height grid: $dsm \in \mathbb{R}^{w \times h 3}$. It must have been produced around the same time of the 3D reconstruction so as to avoid temporal discrepancies. It is compared to the model height (Brédif et al., 2007; Zebedin et al., 2008). The latter is inferred from its facets plane equations. It is then rasterized into the image: $alt \in \mathbb{R}^{w \times h}$ at the same spatial resolution as dsm. Their difference generates a discrepancy map (Figure 1.c). A baseline approach is proposed relying on the statistics of pixel values computed using the S functions (Figure 5).

$$v_{height}(\mathsf{M}) = S(dsm - alt). \tag{5}$$

[3] w (*resp.* h) is the grid width (*resp.* height)

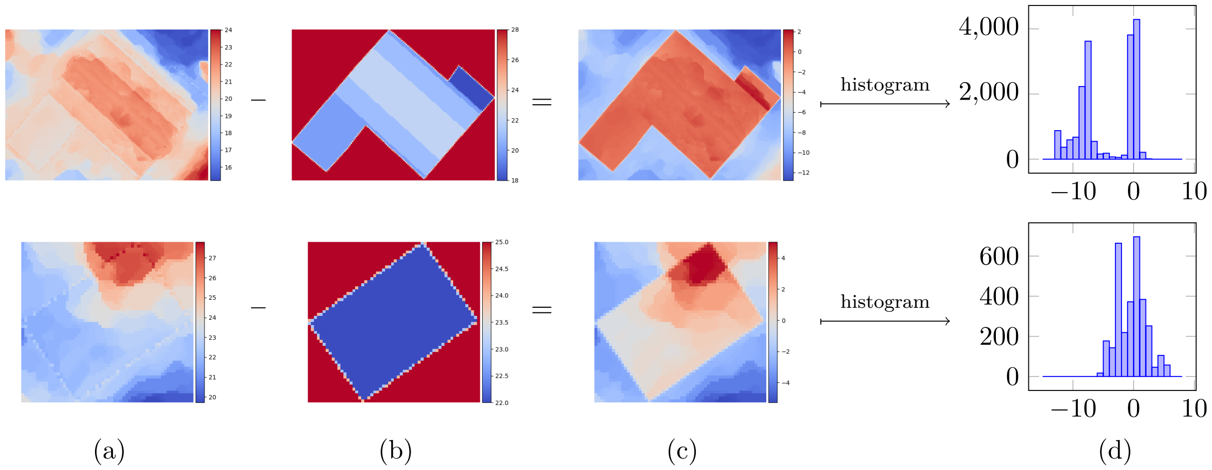


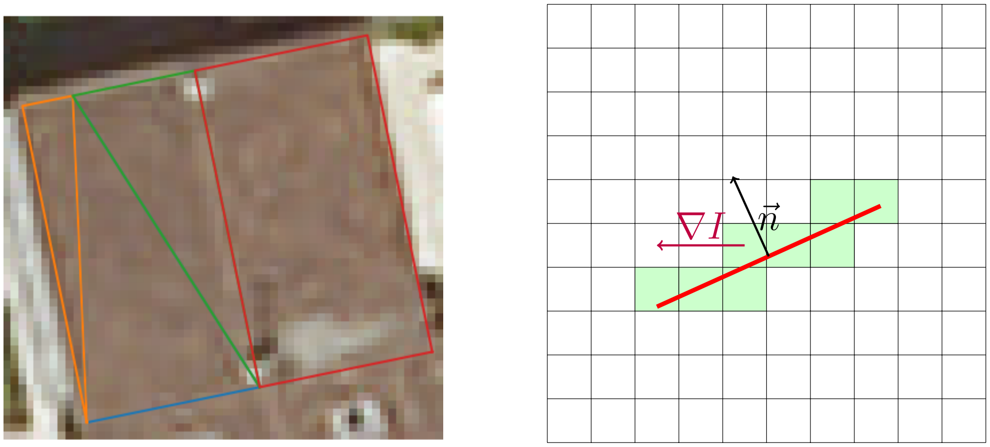
Figure 5: Histogram height-based features computed from the DSM residuals. (a) DSMs. (b) Height maps extracted from the 3D model. (c) Difference between (a) and (b). The difference is transformed into a vector using a histogram (d).

Equation 5 summarizes how building height-based features are computed. Different from a root mean square metric (Lafarge and Mallet, 2012; Poullis, 2013), the histogram captures the discrepancy distribution, which is particularly helpful in detecting undersegmentation defects or geometric imprecision. However, as for the previous geometric attributes, the grid structure of information coming from the model is lost. Errors cannot be spatialized and linked to a specific facet.

3.2.3. Image-based features

We aim to benefit from the high frequencies in Very High Spatial Resolution optical images. Building edges correspond to sharp discontinuities in images (Ortner et al., 2007). We verify this by comparing these edges to local gradients. We start by projecting building models on the orthorectified image I (Figure 6.a). For each facet, we isolate an edge s (Figure 6.b). In an ideal setting, gradients computed at pixels g that intersect s need to be almost be collinear with its normal $\vec{n}(s)$. In consequence, applying the same statistics functions S, we compute the distribution of the cosine similarity between the local gradient and the normal all along that s:

$$\mathsf{D}_S(s, I) \triangleq S\left(\left(\frac{\nabla I(g) \cdot \vec{n}(s)}{\|\nabla I(g)\|}\right)_{g \in I \text{ and } g \cap s \neq \emptyset}\right). \tag{6}$$



(a) (b)

Figure 6: Illustration of how features are derived from optical images. Model facets (each represented by a specific color) are projected onto the aerial image (a). (b) local gradients (in purple), on intersecting pixels (in green), are compared to the edge (in red) normal (in black).

Once the distribution is computed over a segment, it is stacked over all facet edges to define the distribution over projected facets. In the case of histograms S^p_{hist} with the same parameters (and thus the same bins), it is equivalent to summing out the previous vectors $\mathsf{D}_{S^p_{hist}}(s, I)$ over edges s from the projection $q(f)$ of the facet f. In order to take into account the variability of segment dimensions, this sum is normalized by segment lengths.

$$D_{S^p_{hist}}(f, I) \triangleq \sum_{s \in q(f)} \|s\| \cdot \mathsf{D}_{S^p_{hist}}(s, I). \tag{7}$$

The same can be done over all facets of a building M (Equation 8). The weights are added in order to take into account the geometry heterogeneity. The gradient to normal comparison is similar to the 3D data fitting term formulated in (Li et al., 2016). Once again, the model structure is partially lost when simply summing histograms over all segments.

$$v_{image}(\mathsf{M}) = D_{S^p_{hist}}(\mathsf{M}, I) \triangleq \sum_{f \in \mathsf{F}} \mathscr{A}(q(f)) \cdot \mathsf{D}_{S^p_{hist}}(f, I). \tag{8}$$

These image-based attributes are helpful for precision error detection. As example, facet imprecise borders can be detected as local gradients direction will be expected to differ greatly from the inaccurate edge. It can also be detrimental in under-segmentation detection as colors can change considerably from one facet or one building to another inducing an gradient orthogonal to edge normals.

3.3. Classification process

Two sources of flexibility are taken into account. First, the parametric nature of the taxonomy leads to a varying set of label. Secondly, the classifier should be able to handle the heterogeneity of the feature vector and must adapt to different input vectors types and sizes.

3.3.1. Classification problems.

We first define two terms used afterwards. In a *multi-class* classification problem, each instance has only one label that takes only one value amongst multiple ones (two in the case of a *binary* problem). The *multi-label* problem decides, for multiple labels, the most probable state: present (+1) or absent (−1). Both the classification problem nature and the label set are determined by the three previously defined taxonomy parameters (Table 1).

finesse	**eLoD**	**exclusivity**	**Classification output**
1	—	—	Binary(Valid, Erroneous)
2	LoD-1	—	Binary(Valid, Building error)
2	LoD-2	on	MultiClass(Valid, Building error, Facet error)
2	LoD-2	off	MultiLabel(Valid, Building error, Facet error)
3	LoD-1	on	MultiLabel(children(Binary(Valid, Building error)))
3	LoD-2	on	MultiLabel(children(MultiClass(Valid, Building error, Facet error)))
3	LoD-1	off	MultiLabel(children(Building error))
3	LoD-2	off	MultiLabel(children(Building error)∪ children(Facet error))

Table 1: The summary of all possible classification problem types. children(*error*) lists the children of *error* from the taxonomy tree (Figure 2).

Finesse = 1 level corresponds to the standard binary classification problem: ‘ ‘Valid” or “Erroneous”. At *finesse* = 2, the **eLoD** can then take two values in the aerial reconstruction case: LoD-1 or LoD-2. If set at LoD-1, it is a binary classification problem: “Valid” or “Building error”. For LoD-2, if the **exclusivity** is on, it will be a multi-class problem: “Valid”, “Building error” or “Facet errors”. If set off, it becomes a multi-label one with the same labels. At *finesse* = 3 level, if the **exclusivity** is on, it is a 2-stage classification problem. In the first stage, a multi-class task predicts the error family, after which a second multi-label problem decides between the predicted error family children. If the **exclusivity** is off, it turns into 1-stage multi-label problem that predicts the existence of each atomic error corresponding to the chosen eLoD.

3.3.2. Classifier choice

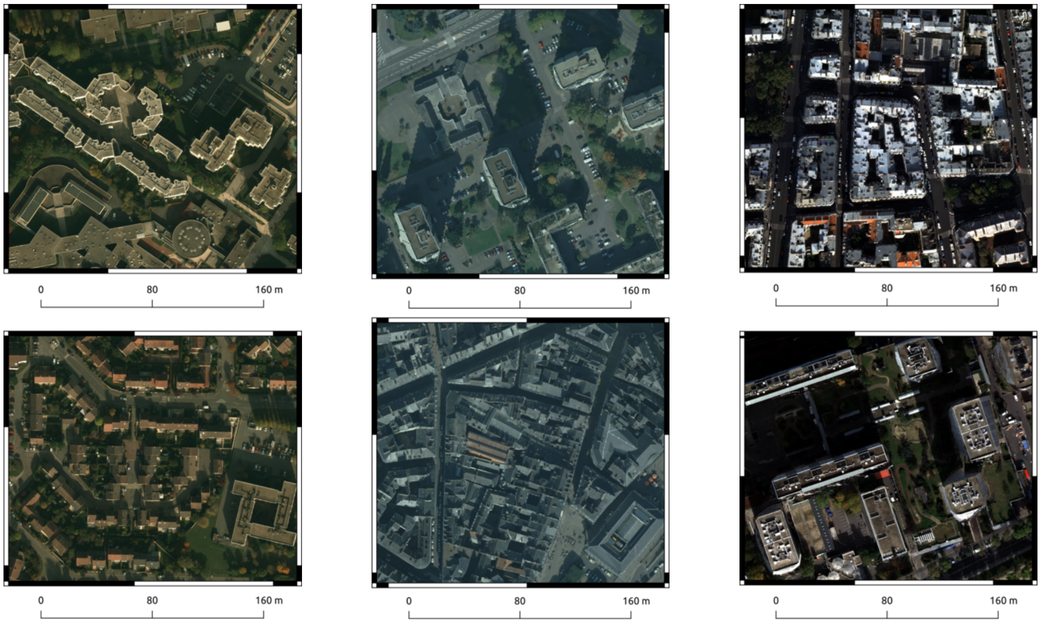
The highly modular nature of the framework with multimodal features involving many parameters restricts the choice of classifiers. Random Forest classifiers (Breiman, 2001; Criminisi and Shotton, 2013) are selected. They can manage a large number of features with different dynamics and coming from multiple modalities. Relying on their bagging property, a high number of trees ($1,000$ elements) is necessary to cover most of the feature space, while a limited tree depth (4) helps avoiding overfitting during training. It adapts to any of our classification paradigm: multi-class or multi-label. In the latter case, a one-vs-all approach is adopted in addition so as to address each label separately.

4. Results

4.1. Dataset

3D models from three different cities of France are selected in order to assess the performance of our framework: **Elancourt**, **Nantes**, and the XIII$^{\text{th}}$ district of Paris (**Paris-13**) (Figure 6.i). **Elancourt** is a small city exhibiting a high diversity of building types: residential areas (hipped roof buildings), and industrial districts (larger buildings with flat roofs). **Nantes** represents a denser urban setting with lower building diversity. In Paris-13, high towers, with flat roof, coexist with Haussmann style buildings that typically exhibit highly fragmented roofs. The Elancourt (*resp.* Nantes and Paris-13) scene contains 2,009 (*resp.* 748 and 478) annotated building models. The DSM and orthorectified image resolution is 6 cm while it is 10 cm for the two other areas.

3D models were generated using the algorithm described in (Durupt and Taillandier, 2006), out of existing building footprints and aerial VHR multi-view DSMs. The modeling algorithm simulates possible roof structures with facets satisfying some geometric constraints. The best configuration is selected using a scoring system on the extrapolated roofs. Finally, vertical building façades connect the optimal roof to the building footprint. These models have a LoD-2 level. This method is adapted to roof types of low complexity and favors symmetrical models (residential areas). It has been selected to ensure a varying error rate for the three areas of interest, especially since models were generated with partly erroneous cadastral maps. 3,235 buildings in total are considered. They were annotated according to the atomic errors list provided by our taxonomy. Figure 6.ii reports modeling errors statistics over the annotated buildings.

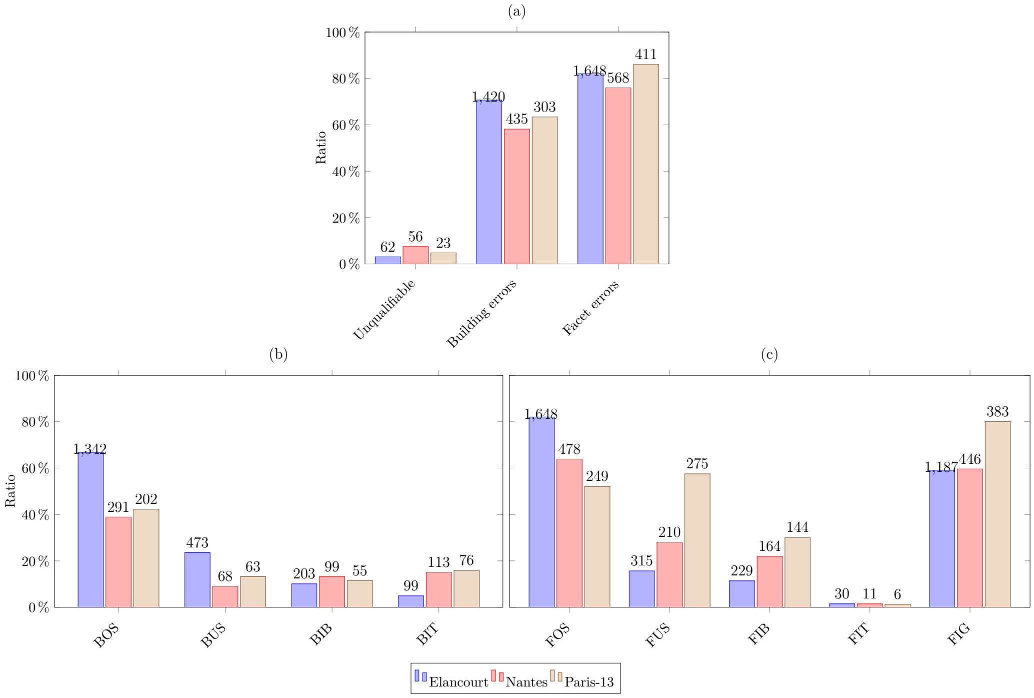


(a) Elancourt.

(b) Nantes.

(c) Paris-13.

i. Selection of two areas of interest in the three datasets.



ii. Statistics: in (a) Unqualifiable ∪ *finesse* = 2 statistics are represented, while (b) illustrates “Building errors” and (c) “Facet errors”.

Figure 7. Statistics per urban scene and error type. Almost similar situations can be noticed.

Unqualifiable buildings represent a small[4] fraction of the dataset ($< 7.5\%$). Only a small fraction of buildings are valid[5]: 57 (2.84%) in Elancourt, 55 (7.35%) for Nantes and 21 (4.39%) in Paris-13. Most buildings are affected by the *Building Errors* family ($> 58.16\%$) and the *Facet Errors* one ($> 75.94\%$). At the *finesse* level 3, more differences are noticed across. Over-segmentation errors are generally well represented, for all LoDs, with at least 38.9% and at most 66.8%. The same is true for `FIG` (59.8-80%). Otherwise, the presence ratio is within the percentage interval of [10, 30], except for topological defects. This negatively impacts the detection of such rare labels. In general, all errors have the same frequency across datasets, apart from `FUS`, `BUS`, and `BIT`. They greatly change from Elancourt (less dense and more heterogeneous) to Paris and Nantes (compact and uniform patterns).

4.2. Experimental set-up

Four feature configurations were tested: “geometric features” (**Geom.**) only, “geometric and height features”(**Geom.** $\cup$ **Hei.**), “geometric and image features”(**Geom.** $\cup$ **Im.**) as well as “geometric, height and image features”(**All.**). Each feature modality generates a 20 dimension vector. The DSMs and orthorectified images used to derive height and image features have the same spatial resolution as the reconstruction input data. Labels are extracted from a non **exclusive** and **eLoD** = LoD-2 taxonomy. All *finesse* levels were tested. The overall accuracy is not interesting due to the highly unbalanced label distribution. We prefer reporting recall (*Rec*) and precision (*Prec*) ratios. Recall expresses, from a number of samples of a given class, the proportion that was rightfully detected as such. Precision indicates how much samples, amongst the detected ones, were, in truth, part of the studied class (Powers, 2011). We also summarize these two ratios with their harmonic mean, the F-score.

4.3. Feature analysis

We assess the added value of each modality. Various feature configurations are studied. They are compared with a baseline consisting in predicting the errors using only the RMSE, which is the standard measure in most of 3D reconstruction methods. We conclude the analysis by studying the feature importance per training zone. All experiments are conducted performing a 10-fold cross validation to avoid overfitting/underfitting issues.

[4] Geometrically inconsistent 3D models were filtered out in a preprocessing (nadir projection) step. This fraction corresponds only to the occluded (partially or completly) buildings that could not be qualified.

[5] Valid means the absence of errors for a specified building

4.3.1. RMSE predictive capacity

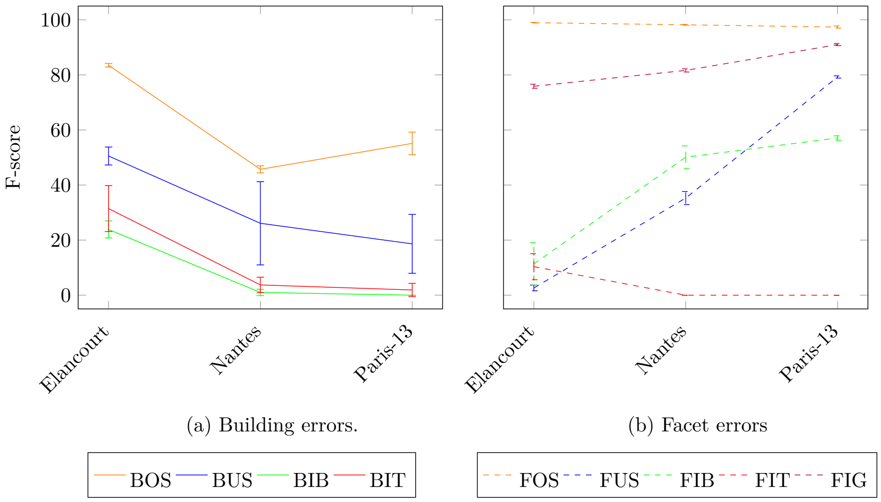
We train the classifier on Elancourt with a one dimensional feature vector $RMSE$. Mean test results are shown in Table 2 We can conclude that the RMSE is not able to detect our errors. We can distinguish two clusters: high recall and low precision and overall accuracy (BOS, FOS and FIG) and low recall and precision (BUS, BIB, BIT, FUS, FIB and FIT). The first group consists of the most numerous errors (Figure 6.ii). This explains how the classifier assigns to almost all samples the positive class: we end up with a high ratio of false positives (false alarms) and hence a high recall ratio but coupled with a weak precision and overall accuracy. The inverse happens with the rest of the errors as we obtain a high percentage of false negative.

	BOS	BUS	BIB	BIT	FOS	FUS	FIB	FIT	FIG
Rec	99.55	0.21	0	0	98.68	0.63	0	0	98.15
Prec	68.78	33.33	—	0	66.60	0.25	—	0	61.15
$\boldsymbol{F_{score}}$	81.35	0.42	0	0	79.52	1.24	0	0	75.36
Acc	68.46	75.65	89.57	94.66	66.36	83.62	88.24	98.36	60.86

Table 2. Finesse 3 experiment results using RMSE on Elancourt. ***Acc*** expresses the overall accuracy ratio.

4.3.2. Feature ablation study

We tested the different feature configurations, at finesse level 3 and in all urban zones. Mean precision and recall test results are reported in Table 3. F-scores are averaged across all feature configurations and represented in Figure 8.



(a) Building errors.

(b) Facet errors

Figure 8. Mean F-score and standard deviation for the feature ablation study.

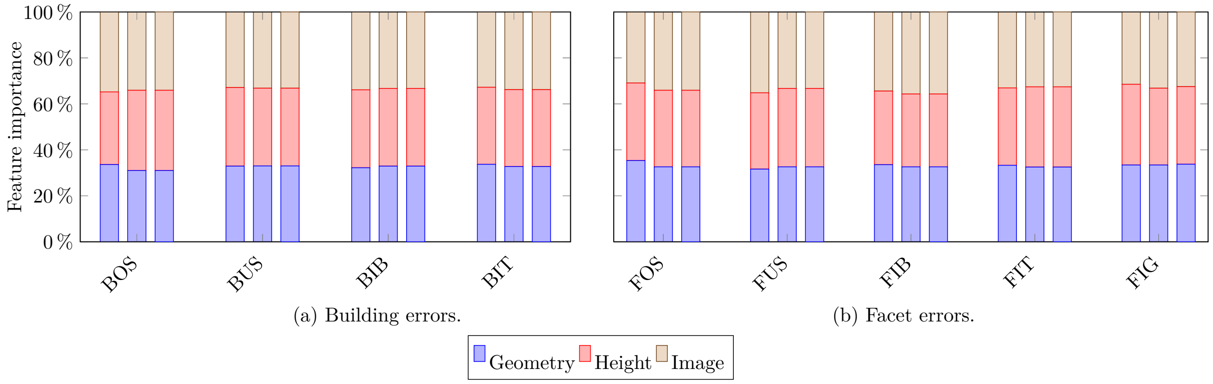
We can first conclude that geometric features alone are generally sufficient. It is the best alternative for topological error detection as shown for `BOS`, `FOS`, `FUS`, `FIT` and `BIT` in Table 3. This is confirmed also by the low variance observed in Figure 8.(a). An exception is noticed with `BUS` in

Elancourt, where height-based features allow an increase of around 9% in recall without, practically, any loss in precision. Similar behaviour is noticed for Nantes and Paris-13 with image-based features (+20% in recall). The first case can be explained by the discrepancy in height that can be observed between under-segmented buildings. The second is made clear by the difference in roof colors, in dense uniform settings (Figure 3.i.a). This helps identifying different instances of buildings.

Figure 8 shows all “Building errors” family labels are better detected for Elancourt. It is also the case of `FOS` and `FIT`. A certain monotony can be noticed, at the exception of `BOS`. Better resultas are obtained for Paris-13 than for Nantes, while having around half the number of models to train on. This means that `BOS` cannot be easily learnt in Nantes. It is coherent with the fact that the dataset represents a part of the dense downtown of the city. The same monotony is observed, this time in reverse, with the rest of “Facet errors” defects. Paris-13 is much better with less training samples. For geometric defects (`FIG` and `FIB`), Nantes is comparable to Paris-13, but, with `FUS`, it is way much worse. This may result from the highly heterogeneous aspect of this dataset that encompasses high tower buildings with a densely populated city district. FInally, well represented errors are more easily detected than the less frequent ones, especially the rare ones like `FIT` in Nantes and Paris-13.

4.3.3. Feature importance

Random forest classifiers can easily infer feature importance at training time. These were here computed and aggregated by modality in all urban scenes (Figure 9).



(a) Building errors.

(b) Facet errors.

Figure 9. Modality importance computed by stacking single feature importances retrieved from the Random Forest classifier. The first (*resp.* second and third) column represents Elancourt (*resp.* Nantes and Paris-13).

At first, we observe how much individual attributes are important before being gathered. For geometric features, all attributes are equally important. However, concerning image- and height-

Elancourt

	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***
BOS	**93.96**	76.15	91.43	**77.76**	91.51	76.08	90.83	76.14
BUS	32.98	**76.47**	**41.86**	75.57	40.38	71.00	39.32	71.81
BIB	12.32	67.57	12.81	**68.42**	16.26	67.35	**16.75**	68.0
BIT	**25.25**	92.59	20.20	90.91	20.20	**95.24**	11.11	91.67
FOS	98.91	99.07	98.91	**99.30**	**98.99**	98.84	98.91	98.84
FUS	**1.90**	54.55	0.63	**66.67**	1.61	50	1.27	**66.67**
FIB	**9.17**	87.5	0	—	8.30	82.61	7.42	**100**
FIT	6.67	**100**	**8.73**	95.24	3.33	**100**	3.33	**100**
FIG	**80.54**	73.14	80.45	**72.62**	78.69	72.12	79.02	71.82

Nantes

	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***
BOS	**38.14**	61.67	36.43	60.23	36.77	**62.21**	34.71	60.48
BUS	7.35	62.5	7.35	55.56	**29.41**	**66.67**	26.47	64.29
BIB	0	—	0	—	**1.01**	**50.0**	**1.01**	**50.0**
BIT	1.77	22.22	**3.54**	44.44	0	0	2.65	**50.0**
FOS	**98.54**	**98.13**	**98.54**	**98.13**	98.33	97.92	98.12	97.91
FUS	27.62	55.24	**27.62**	**59.18**	24.76	54.74	23.33	53.85
FIB	37.80	62.0	36.59	**63.16**	**49.39**	60.90	46.39	60.90
FIT	0	—	0	—	0	—	0	—
FIG	86.32	78.09	**86.77**	78.02	84.53	**78.71**	83.86	78.08

Paris-13

	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***
BOS	45.54	65.25	46.53	68.61	**50.0**	68.24	46.53	**70.15**
BUS	6.35	66.67	7.94	71.43	**22.22**	**77.78**	7.94	62.5
BIB	0	—	0	—	0	0	0	—
BIT	**2.63**	**50.0**	0	—	1.32	50.0	0	0
FOS	97.19	97.19	97.19	97.19	**97.59**	**98.38**	97.19	97.19
FUS	**85.09**	**75.0**	84.36	74.12	85.09	74.52	84.36	74.12
FIB	53.47	62.10	51.39	61.67	**53.47**	**63.11**	52.78	61.79
FIT	0	—	0	—	0	—	0	—
FIG	97.65	84.62	**98.96**	**84.79**	97.65	84.62	**98.96**	**84.79**

Table 3. Feature ablation study preformed on the three areas at *finesse* level 3. Test results are expressed in percentage. All *atomic* errors are considered over all possible configurations.

based features, only a few are relevant (higher feature importance ratio). Indeed, these few attributes correspond to the highest and lowest values of the histograms. As described earlier, image and height features consist of a histogram of distances between the model and the real measured signals: vector cosine similarity, for the first, and the L_2 norm for the last. It is clear that the presence of errors would result in saturating the high values in the histogram, while an absence of defects would imply a big number of low values. This intuitively explains the observed phenomenon.

In a second time, we notice that no modality is more important than the others, contrarily to what was observed in Table 3. In fact, for most atomic errors, test results using geometric features are comparable to those obtained with more modalities. However, during training, all modalities are relevant ($\sim 1/3$ in Figure 9). This explains why all configurations are kept for subsequent analysis.

5. Scalability analysis

It is established that the scene composition can affect greatly model defect detection. This fact motivates studying training the classifier and testing prediction on different scenes. The goal is to prove the resilience of the prediction to unseen urban scenes. As the annotation process require a lot of effort, this trait is crucial to guarantee the scalability of this method. Different configurations are possible (Figure 10). In a first experiment, we train on one urban scene and test on another one (*transferability* of the classifier model). In a second configuration, the classifier is trained on two scenes and tested on the last one: the goal is to investigate *generalization*. The last experiment targets the *representativity* of a single 3-area dataset by trying multiple train-test splits.

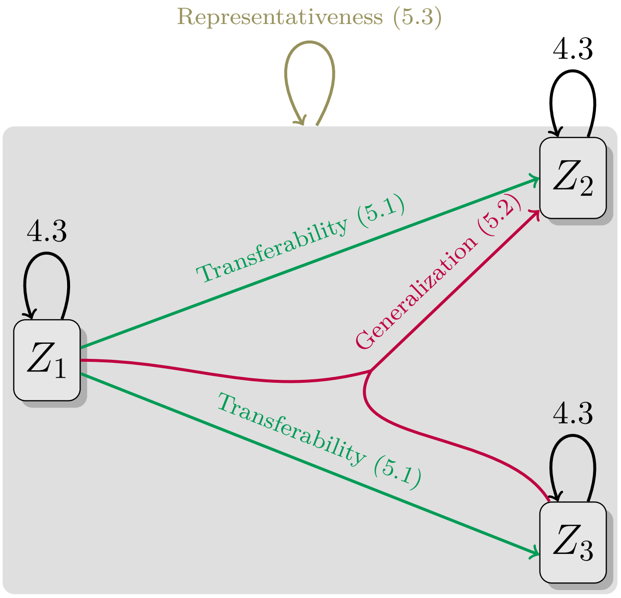
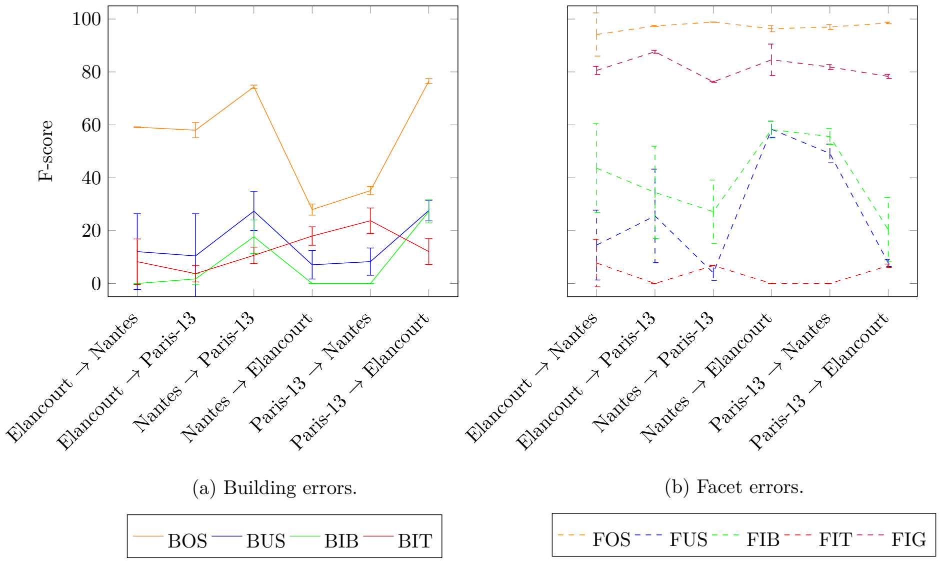


Figure 10. A graph representing possible experiments: arrow origins represent training scenes while test ones are depicted as targets. $Z_i, i = 1, 2, 3$ represent the urban zones. All these nodes are assembled in one, meaning that all urban scenes were aggregated in on train/test node. The numbers indicate in which section each experiment is analyzed.



(a) Building errors.

(b) Facet errors.

Figure 11. Mean F-score and standard deviation for the transferability study.

We will see how “Building errors” depend on the training scene, in contrast to “Facet errors”. The latter will prove to be more transferable and generalizable than the first one. We will also discuss how every modalities play a role in error prediction. Image-based features will demonstrate to be the most valuable compared to height-based ones. Eventually, we will review each *atomic* error prediction sensitivity provided the training set.

5.1. Transferability study

In this configuration, we test how transferable are the learned classifiers from one urban scene to another. We train on a zone Z_i and test on another one Z_j. We will denote each transferability experiment by the couple (Z_i, Z_j) or by $Z_i \to Z_j$. Six transferability couples are possible. F-scores are shown, per label, and per experiment, in Figure 11.

First, a *coherence* analysis is performed. We compare the results of the transferability experiments to the ablation results with the same training scene (for a given area Z_i in all couples $(Z_i, Z_j)_{\forall j \neq i}$, differences between Figure 11 and Table 3/Figure 8). Secondly, we investigate how an urban scene composition helps predicting defects in an unseen one. This is called the *projectivity* comparison. For a given test scene Z_j in couples $(Z_i, Z_j)_{\forall i \neq j}$, we compare results from Figure 11 with Table 3/Figure 8. Analysis is provided in Table 4. In both settings, if a feature type appears, it means it is, by a large margin, the most decisive one. A color scheme was devised to encode the amplitude of change. All various feature configurations are tested these experiments. If a modality stands out, in terms of the F-score, it is mentioned in the corresponding cell in Table 4.

To summarize the comparisons, error family wise, out of 22 “Building errors” possible projectivity comparisons, 14 yield worse results. This proves how hard it is, for this error family, to transfer learned classifiers. It is, however, the contrary for the “Facet errors”. Only 8 out of 27 projectivity errors are worse than training on the same test area.

			BOS	BUS	BIB	BIT	FOS	FUS	FIB	FIT	FIG
Transferability	Coherence	Elancourt → Nantes	🟥	🟧	🟧	🟧	▒	🟩 Im.	▩ Im.	🟨 Im.	🟩
		Elancourt → Paris-13	🟥	🟧	🟧	🟧	▒	🟩 Im.	▩ Im.	🟨	▩
		Nantes → Paris-13	🟨	🟧	□	🟩 Geom.	▒	▩	🟩	□	▒ Hei.
		Nantes → Elancourt	▩	▒	▩	🟩 Geom.	▒	🟧	🟧	🟩	🟨
		Paris-13 → Nantes	🟨	🟨	□	▩ Geom.	▒	🟥	🟧	□	🟧
		Paris-13 → Elancourt	▩	🟩	▩	🟩 Geom.	▒	🟫	▒	🟩	🟨
	Projectivity	Elancourt → Nantes	🟨	🟧	🟨	🟧	▒	▒	▩ Im.	🟨	▒
		Elancourt → Paris-13	🟨	🟧	▒	🟧	▒	🟩 Im.	🟩 Im.	🟨	▒
		Nantes → Paris-13	🟨	🟧	□	▩	▒	🟩	🟩	□	▒
		Nantes → Elancourt	🟩	🟨	🟨	🟩 All	▒	🟧	🟨 Im.	🟩 Im.	▒
		Paris-13 → Nantes	🟧	🟨	□	🟩	▒	🟥	▒	□	🟨 Hei.
		Paris-13 → Elancourt	▒	🟨	🟩 Im.	▒	▒	🟫	🟧 Im.	□	▒
General.		Elancourt	🟧	🟧 Im.	🟧	🟧	▒	🟩 Im.	▩ Im.	🟨 Geom.	🟨 Hei
		Nantes	▒ All	🟨 Im.	🟨 Im.	▩	▒	▒	🟧 Im.	□	▒
		Paris-13	🟧 All	🟧	□	🟩 Hei.	▒	🟫	🟧 Im.	□	🟨

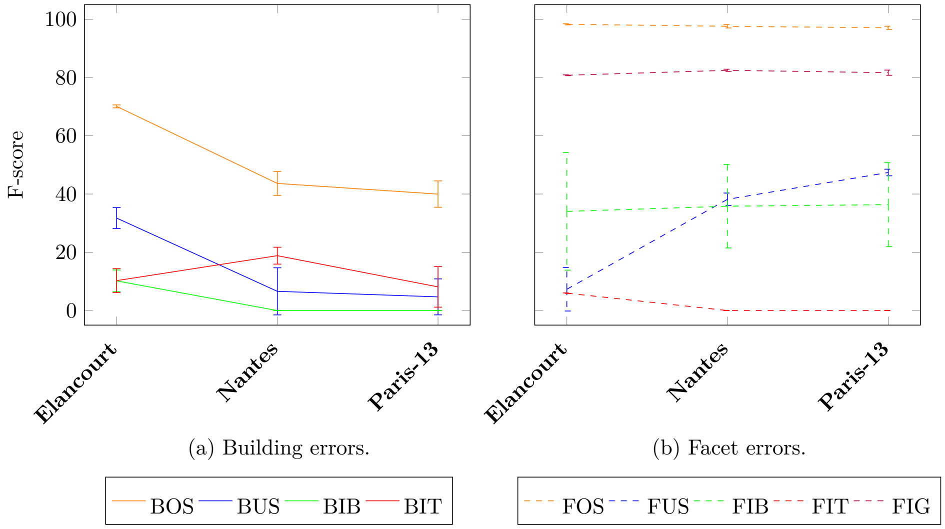
Table 4. Evolution of the F-score value, for each error, between each tested configuration and the best result per area (Section 4.3.2). Feature sets having a significant impact on the classification results are mentioned. Otherwise, Geom. Im., and Hei. contribute equally. The color indicates the magnitude: 🟫: [−45, −35%[– 🟥: [−35, −25%[– 🟧: [−25, 15%[– 🟨: [−15, 5%[– ▒: [−5, 5%[– 🟩: [5, 15%[– ▩: [15, 25%] – □: statistics cannot be computed.

As mentioned earlier, additional modalities play a important role in prediction accuracy. We start with image-based attributes. In some cases, they were pivotal in obtaining better results for geometric errors (FIB, BIB), as well as for topological ones (FUS, FIT). These features have a significant coherence power when trained over Elancourt (FIB and FUS), and projects very well to other scenes (FIB, FUS, BIB and FIT, Table 4). On the other hand, as expected, geometric features alone are best for topological errors, when trained on dense areas, especially BIT (Table 4). Finally, although stiking out for FIG in a minor capacity (*cf.* Table 4), height-based features proved to be less transferable. In fact, adding height-based features leads, in most cases, to a small decrease in accuracy (≈ 2%) for *atomic* errors. All these previous findings further justify why we did not leave out any modality, as they are more frequently critical for transferability than in the ablation study (Table 3).

An analysis can also be drawn for *atomic* errors with respect to the best training scene. We can see that for BOS, training on a dense urban scene like Nantes, is the best solution, as for topology errors (FIT and BIT). Paris-13 represents also a dense downtown scene but with even more diverse building types. This is instrumental to achieve transferability for BUS and BIB. Conversely, Elancourt offers more heterogeneity on the LoD-2 level. As a consequence, it is the best training zone for FUS, FIB and FIG. Finally, as one can obviously suspect, FOS learning is evenly transferable, as it is well detected when training on any scene.

5.2. Generalization study

We try to find out how omitting one urban zone from the training dataset affects the test results on that same area. An other way to look at it is, from an operational point of view, to find out how much learning on a union of many urban scenes is helpful when testing on an unseen one. We also seek to confirm the outcome of the transferability experiments. Experiments that merge all zones except Z_i ($\bigcup_{\forall j\neq i} Z_j$) for training and test on Z_i are noted by the couple ($\bigcup_{\forall j\neq i} Z_j, Z_i$) or by $\bigcup_{\forall j\neq i} Z_j \rightarrow Z_i$. There are three possibilities: Elancourt ∪ Nantes → Paris-13, Paris-13 ∪ Nantes → Elancourt and Paris-13 ∪ Elancourt → Nantes. The F-score evolution per experiment and error is depicted in Figure 12.



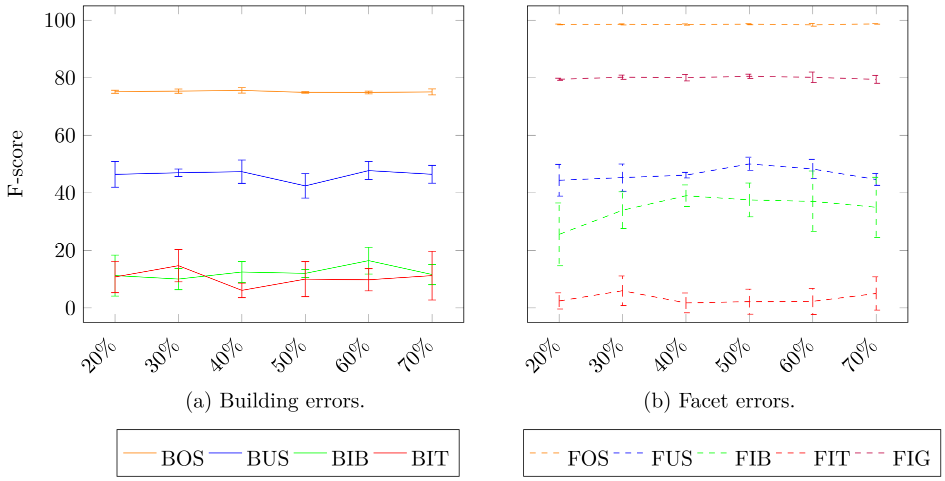
(a) Building errors.

(b) Facet errors.

Figure 12. Mean F-score and standard deviation for the generalization study per test zone.

We compare these experiments with the ablation study on the same area (*cf.* Table 4). We analyse result along the same criterea as the transferability study.

We start again with a comparison depending on error families. Out of the 11 possibilities for the “Building errors” family, 8 yield worse results. For the ”Facet” family, 6 out of 13 comparisons exhibit the same trend. This is worst than the transferability comparisons in ratio. This results from the fact that fusing more datasets, that are not tailored for a specific error detection, does not



(a) Building errors.

(b) Facet errors.

Figure 13. Mean F-score and standard deviation for the representativeness experiments depending on the training set size.

help alleviating the problem. It only evens out the best(*resp.* worst) performances by including the best (*resp.* worst) urban scene in the training set.

Similarly to the previous study, image and height modalities play a major role in error detection. Image-based features are crucial for `FIB`, `BIB`, `FUS` and `BUS` detection (Table 4). Height-based attributes, however, induce a larger impovement in predicting `FIG` and `BIT`, while geometric ones are relegated to playing a minor role. Otherwise, a curiosity can be noticed: only when fusing all modalities together, in Paris-13 and Nantes, does predictions improve for `BOS`.

We also confirm the observations about the best urban scene for error prediction training. In this case, the best zone should always give the worst scores. It is mostly the case with all atomic errors, with the exception of `BIT`. This outlier can be explained by the resemblance of the Paris-13 3D model to Nantes (which was established to be the best) samples. Indeed, for most labels, Nantes and Paris-13 reach the same scores. However, the discrepancy in F-scores proves the added value for each dataset.

5.3. Representativeness study

The objective is to find out, after merging all training samples from all datasets, what is the minimal amount of data that can guaranty stable predictions. We can, thereafter, understand how much learning on one scene typology can affect results compared to a mixed training set. Figure 13 depicts F-score as a function of training ratios (between 20-70%) and *atomic* errors.

We note the high stability of the F-score. This indicates that having a small heterogeneous dataset is not detrimental to the learning capacity and can be even the most suitable solution. `BOS`, `FOS`, and `FIG` have a standard deviation under 2%, as opposaed to `FIB`, `BIT` and `FIT`. Indeed, they have large variance, and even a larger standard deviation than mean value. Scalability is, hence,

Elancourt								
	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***
Building errors	99.76	**85.96**	99.82	85.88	99.88	85.57	**100**	85.55
Facet errors	91.79	**89.79**	92.65	89.40	93.21	89.45	**93.46**	89.16

Nantes								
	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***
Building errors	85.98	67.27	87.59	67.79	85.75	68.32	**86.90**	**69.23**
Facet errors	91.20	94.01	91.37	**94.36**	91.20	94.35	**91.73**	94.21

Paris-13								
	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***	***Rec***	***Prec***
Building errors	97.36	68.76	97.36	68.76	97.36	68.76	97.36	68.76
Facet errors	99.03	91.26	99.03	91.26	99.03	91.26	99.03	91.26

Table 5. Feature ablation study on the three datasets for the *finesse* = 2 case. Results are expressed in percentage.

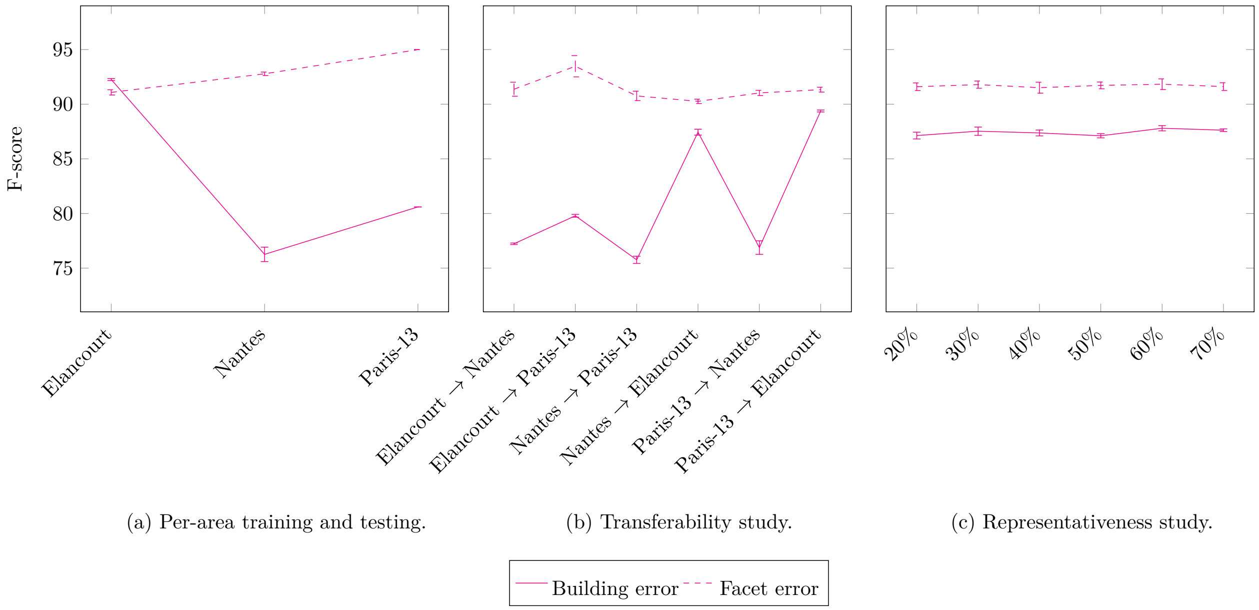
ensured with a limited training set. No standard logarithmic behaviour can be found at the studied scales. 20% of the full label set is sufficient so as to retrieve results with a performance similar to the initial ablation study. The best results are observed for BOS, BUS, and FUS. These errors are topological defects of building roof facets which require a high diversity of training samples for their detection. More sophisticated features are however still required to help predicting less frequent and more semantic labels.

6. *Finesse* study

In this section, we reproduce the experimental settings described in Sections 4 and 5. This time, the *finesse* level is fixed at 2. The goal is to find out how good, transferable and stable are the model quality predictions at the semantic level of error families (*i.e.*, “Building errors” vs. “Facet errors”).

6.1. Error family detection

We start by the ablation study. Table 5 reveals that inserting more remote sensing modalities do not change the prediction results dramatically. This is perfectly illustrated by the low variance in Figure 14 a for the three areas of interest. These results are in line with the conclusions at the *finesse* level 3. In the higher *finesse* level, only BUS, from all *atomic* errors, was highly impacted by a change in the feature configuration. This may explain the observed low variability. We also note the prevalence of FOS errors (between 60% and 85%) and FIG (between 70% and 95%) in the “Facet errors” family. This added to the fact that they are, in a large capacity, easily detected individually (>90%, in F-score, for the first and 80% for the second, see Figure 8b) helps understanding why the F-score reaches at least 90% for this family (Figure 14a). As with *finesse* level 3 experiments,



(a) Per-area training and testing.

(b) Transferability study.

(c) Representativeness study.

Figure 14. F-score mean and standard deviation for the feature ablation study outcomes per zone for *finesse* level 2. (a) corresponds to the ablation study, (b) to the transferability experiments and (c) to the representativity setting.

"Facet errors" yields higher prediction scores than on "Building errors". Indeed, we can see a smaller discrepancy between F-scores on different scenes for "Facet errors"' (below 5%) than for "Building errors" (15%).

The transferability study (Figure 14b) compares the F-scores with the ablation study provided in Figure 14a. Out of all 12 possible comparisons, only 2 exhibit a decrease in error discrimination. Both affect the "Building errors" family when trained on Nantes. "Facet errors", on the other hand, confirms, its transferability and stability (less than 5% of discrepancy between the two extremal values). For this reason, we skip the generalization study, all together, at this section.

The representativeness study conducted for the *finesse* level 2 results in the F-scores that are illustrated in Figure 14c. Family detection scores are very stable across all different tested split ratios. Moreover, in contrast to *atomic* errors results (cf. Figure 13), F-scores do not vary by more than 1% in mean and standard deviation. This proves that at *finesse* level 2, error family prediction is evened out independent of different split ratios, as opposed to higher order errors. Again, it benefits from the higher heterogeneity of the training set with multiple areas.

6.2. Detection of erroneous models

Now, we work at *finesse* level 1, first on feature ablation. Since valid samples are very rare in our case, it is expected that it will be very difficult to detect these instances. In consequence, in

Elancourt								
	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	**Valid**	***Rec***	**Valid**	***Rec***	**Valid**	***Rec***	**Valid**
Erroneous	99.95	$\frac{1}{57}$	99.95	$\frac{1}{57}$	99.95	$\frac{0}{57}$	99.95	$\frac{1}{57}$

Nantes								
	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	**Valid**	***Rec***	**Valid**	***Rec***	**Valid**	***Rec***	**Valid**
Erroneous	99.84	$\frac{0}{55}$	99.84	$\frac{0}{55}$	100	$\frac{0}{55}$	**100**	$\frac{0}{55}$

Paris-13								
	Geom.		**Geom. ∪ Hei.**		**Geom. ∪ Im.**		**All**	
	Rec	**Valid**	***Rec***	**Valid**	***Rec***	**Valid**	***Rec***	**Valid**
Erroneous	99.77	$\frac{3}{21}$	99.77	$\frac{3}{21}$	99.77	$\frac{3}{21}$	99.77	$\frac{3}{21}$

Table 6. Test results expressed in percentage for the *finesse*= 1 case. All four configurations are compared across both family errors.

Table 6, we choose to report correctly "Valid" buildings instead of computing the precision score in percentage.

At this level, even more that the error family semantic degree, feature configurations have virtually no impact on test results: Elancourt was the only exception when image features are added to geometric ones. Furthermore, we confirm expectations as, at most, only 1 out of 57 (*resp.* 0 out of 55 and 3 out of 21) valid instances are detected for Elancourt (*resp.* Nantes and Paris-13). As a consequence, we do not report the rest of previously conducted experiments for this *finesse* level. Indeed, it is senseless to compare detection transferability, generalization or representativeness if we hardly detect them at all on the same training scene.

7. Conclusion

A learning framework was proposed to semantically evaluate the quality of 3D models of buildings. For that purpose, errors were hierarchically organized into a novel flexible taxonomy. It aims to handle the large diversity of urban environments and varying requirements stemming from end-users (geometric accuracy and level of details). Based on the desired LoD, exclusivity and semantic level, an error collection is considered. Model quality is then predicted using a supervised Random Forest classifier. Each model provides intrinsic geometrical characteristics that are compiled in a handcrafted feature vector. Remote sensing modalities can be introduced. This helps better describing building models and detecting errors. Attributes can indeed be extracted by comparing the 3D model with optical images or depth data at the spatial resolution at least similar to the input 3D model. Experiments shows it helps detecting hard cases both for geometrical and topological errors.

This new framework was applied to the case of aerial urban reconstruction, where features are extracted from Very High Resolution airborne images and a DSM. A fully annotated dataset containing 3,235 aerial reconstructed building models with high diversity and from three distinct areas was used to test our method. It was associated with multimodal RGB optical and Digital Surface Model features. Although being mitigated over under-represented errors, results are satisfactory in the well balanced cases. More importantly, we proved that the urban scene composition affects greatly error detection. In fact, some predictions scores are not only stable, when training on a different urban scene, they even outperform when learning on the same scene. Additionally, we reported how, for a heterogeneous training dataset, the size of the training set have, practically no effect, as test score stay stable for all errors. This demonstrates that the proposed framework can be easily scaled with the right choice of training samples with little manually generated data. This exactly answers to the raised problematic, contrarily to the present state-of-the-art literature. As a next step, more structure-aware features (based on graph comparison, for instance) could be proposed (Boguslawski et al., 2011) so as to be applied on a richer and more diverse dataset (potentially involving data augmentation) under a deep-based framework.

References

Akca, D., Freeman, M., Sargent, I., Gruen, A., 2010. Quality assessment of 3D building data. The Photogrammetric Record 25 (132), 339–355.

Berger, M., Levine, J. A., Nonato, L. G., Taubin, G., Silva, C. T., 2013. A benchmark for surface reconstruction. ACM Transactions on Graphics 32 (2), 20.

Biljecki, F., Ledoux, H., Du, X., Stoter, J., Soon, K. H., Khoo, V. H. S., 2016a. The most common geometric and semantic errors in CityGML datasets. Vol. IV-2/W1 of ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Athens, Greece, pp. 13–22.

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

Biljecki, F., Ledoux, H., Stoter, J., 2017. Generating 3D city models without elevation data. Computers, Environment and Urban Systems 64, 1–18.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Çöltekin, A., 2015. Applications of 3D city models: state of the art review. ISPRS International Journal of Geo-Information 4 (4), 2842–2889.

Boguslawski, P., Gold, C. M., Ledoux, H., 2011. Modelling and analysing 3d buildings with a primal/dual data structure. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2), 188–197.

Boudet, L., Paparoditis, N., Jung, F., Martinoty, G., Pierrot-Deseilligny, M., 2006. A supervised classification approach towards quality self-diagnosis of 3D building models using digital aerial imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (3), 136–141.

Brédif, M., Boldo, D., Pierrot-Deseilligny, M., Maître, H., 2007. 3D building reconstruction with parametric roof superstructures. In: International Conference on Image Processing, IEEE, 16-19 September 2007, San Antonio, USA. pp. 537–540.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Cabezas, R., Straub, J., Fisher, J. W., 2015. Semantically-aware aerial reconstruction from multi-modal data. In: IEEE International Conference on Computer Vision, IEEE, 11-18 December 2015, Santiago, Chile. pp. 2156–2164.

Criminisi, A., Shotton, J., 2013. Decision forests for computer vision and medical image analysis. Springer Science & Business Media.

Dick, A. R., Torr, P. H., Cipolla, R., 2004. Modelling and interpretation of architecture from several images. International Journal of Computer Vision 60 (2), 111–134.

Duan, L., Lafarge, F., 2016. Towards large-scale city reconstruction from satellites. In: European Conference on Computer Vision, 8-16 October 2016, Amsterdam, The Nederlands. Springer, pp. 89–104.

Durupt, M., Taillandier, F., 2006. Automatic building reconstruction from a Digital Elevation Model and cadastral data: an operational approach. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (3), 142–147.

Elberink, S. O., Vosselman, G., 2011. Quality analysis on 3d building models reconstructed from airborne laser scanning data. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2), 157 – 165.

Ennafii, O., Le Bris, A., Lafarge, F., Mallet, C., 2019. The necessary yet complex evaluation of 3D city models : a semantic approach. In: Joint Urban Remote Sensing Event (JURSE), IEEE/ISPRS, 22-24 May 2019, Vannes, France.

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

Hu, Y., Zhou, Q., Gao, X., Jacobson, A., Zorin, D., Panozzo, D., jul 2018. Tetrahedral meshing in the wild. ACM Transactions on Graphics 37 (4), 1–14.

Jaynes, C., Riseman, E., Hanson, A., 2003. Recognition and reconstruction of buildings from multiple aerial images. Computer Vision and Image Understanding 90 (1), 68–98.

Kaartinen, H., Hyyppä, J., Gülch, E., Vosselman, G., Hyyppä, H., Matikainen, L., Hofmann, A., Mäder, U., Persson, Å., Söderman, U., et al., 2005. Accuracy of 3d city models: EuroSDR comparison. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (3/W19), 227–232.

Karantzalos, K., Paragios, N., 2010. Large-scale building reconstruction through information fusion and 3-d priors. IEEE Transactions on Geoscience and Remote Sensing 48 (5), 2283–2296.

Kelly, T., Femiani, J., Wonka, P., Mitra, N. J., 2017. Bigsur: Large-scale structured urban reconstruction. ACM Transactions on Graphics 36 (6), 204:1–204:16.

Kolbe, T. H., Gröger, G., Plümer, L., 2005. CityGML: Interoperable access to 3D city models. In: Geo-information for disaster management. Springer, pp. 883–899.

Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K., et al., 2016. Crowdsourcing in computer vision. Foundations and Trends in Computer Graphics and Vision 10 (3), 177–243.

Kowdle, A., Chang, Y.-J., Gallagher, A., Chen, T., 2011. Active learning for piecewise planar 3d reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 20-25 June 2011, Colorado Springs, USA. pp. 929–936.

Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M., 2010. Structural approach for building reconstruction from a single DSM. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (1), 135–147.

Lafarge, F., Mallet, C., 2012. Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. International Journal of Computer Vision 99 (1), 69–85.

Ledoux, H., 2018. val3dity: validation of 3D GIS primitives according to the international standards. Open Geospatial Data, Software and Standards 3 (1), 1.

Ledoux, H., Meijers, M., 2011. Topologically consistent 3d city models obtained by extrusion. International Journal of Geographical Information Science 25 (4), 557–574.

Li, M., Wonka, P., Nan, L., 2016. Manhattan-world urban reconstruction from point clouds. In: European Conference on Computer Vision, 8-16 October 2016, Amsterdam, The Nederlands. pp. 54–69.

Macay Moreia, J. M., Nex, F., Agugiaro, G., Remondino, F., Lim, N. J., 2013. From DSM To 3D building models: a quantitative evaluation. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1/W1 (1), 213–219.

Michelin, J.-C., Tierny, J., Tupin, F., Mallet, C., Paparoditis, N., 2013. Quality evaluation of 3D city building models with automatic error diagnosis. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-7/W2, 161–166.

Monszpart, A., Mellado, N., Brostow, G. J., Mitra, N. J., 2015. Rapter: Rebuilding man-made scenes with regular arrangements of planes. ACM Transactions on Graphics 34 (4), 103:1–103:12.

Müller, P., Wonka, P., Haegler, S., Ulmer, A., Van Gool, L., 2006. Procedural modeling of buildings. ACM Transactions on Graphics 25 (3), 614–623.

Musialski, P., Wonka, P., Aliaga, D. G., Wimmer, M., van Gool, L., Purgathofer, W., 2012. A survey of urban reconstruction. EUROGRAPHICS 2012 State of the Art Reports XX, 1–28.

Nan, L., Wonka, P., 2017. Polyfit: Polygonal surface reconstruction from point clouds. In: International Conference on Computer Vision and Pattern Recognition (ICCV), IEEE/CVF, 22-29 October 2017, Venice, Italy. pp. 2353–2361.

Neis, P., Goetz, M., Zipf, A., 2012. Towards automatic vandalism detection in OpenStreetMap. ISPRS International Journal of Geo-Information 1 (3), 315–332.

Nguatem, W., Mayer, H., 2017. Modeling urban scenes from pointclouds. In: Internatinal Conference on Computer Vision (ICCV), IEEE/CVF, 22-29 October 2017, Venice, Italy. pp. 3837–3846.

Ortner, M., Descombes, X., Zerubia, J., 2007. Building outline extraction from Digital Elevation Models using marked point processes. International Journal of Computer Vision 72 (2), 107–132.

Over, M., Schilling, A., Neubauer, S., Zipf, A., 2010. Generating web-based 3D city models from OpenStreetMap: The current situation in Germany. Computers, Environment and Urban Systems 34 (6), 496 – 507.

Poli, D., Caravaggi, I., 2013. 3d modeling of large urban areas with stereo VHR satellite imagery: lessons learned. Natural Hazards 68 (1), 53–78.

Poullis, C., 2013. A Framework for Automatic Modeling from Point Cloud Data. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11), 2563–2575.

Powers, D., 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. International Journal of Machine Learning Technology 2 (1), 37–63.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing 93, 256–271.

Schuster, H.-F., Weidner, U., 2003. A new approach towards quantitative quality evaluation of 3d building models. In: ISPRS Commission IV Joint Workshop on Challenges in Geospatial Analysis, Stuttgart, Germany. pp. 614–629.

Sester, M., Harrie, L., Stein, A., 2011. Theme issue ”Quality, scale and analysis aspects of urban city models. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2), 155–156.

Stoter, J., van den Brink, L., Beetz, J., Ledoux, H., Reuvers, M., Janssen, P., Penninga, F., Vosselman, G., Oude Elberink, S., 2013. Three-dimensional modeling with national coverage: case of The Netherlands. Geo-spatial Information Science 16 (4), 267–276.

Taillandier, F., Deriche, R., 2004. Automatic buildings reconstruction from aerial images: a generic Bayesian framework. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 35 (3A).

Taneja, A., Ballan, L., Pollefeys, M., 2015. Geometric change detection in urban environments using images. IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (11), 2193–2206.

Tran, H., Khoshelham, K., Kealy, A., 2019. Geometric comparison and quality evaluation of 3d models of indoor environments. ISPRS Journal of Photogrammetry and Remote Sensing 149, 29–39.

Vanegas, C. A., Aliaga, D. G., Beneš, B., 2010. Building reconstruction using Manhattan-world grammars. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 13-18 June 2010, San Francisco, USA. pp. 358–365.

Verdie, Y., Lafarge, F., Alliez, P., 2015. LoD generation for urban scenes. ACM Transactions on Graphics 34, 30.

Verma, V., Kumar, R., Hsu, S., 2006. 3d building detection and modeling from aerial lidar data. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 17-22 June 2006, New York, USA. pp. 2213–2220.

Vögtle, T., Steinle, E., 2003. On the quality of object classification and automated building modeling based on laser scanning data. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 34 (Part 3), W13.

Xiong, B., Elberink, S. O., Vosselman, G., 2014. A graph edit dictionary for correcting errors in roof topology graphs reconstructed from point clouds. ISPRS Journal of Photogrammetry and Remote sensing 93, 227–242.

You, R., Lin, B., 2011. A Quality Prediction Method for Building Model Reconstruction Using LiDAR Data and Topographic Maps. IEEE Transactions on Geoscience and Remote Sensing 49 (9), 3471–3480.

Zebedin, L., Bauer, J., Karner, K., Bischof, H., 2008. Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. In: European conference on computer vision. Springer, pp. 873–886.

Zeng, C., Member, S., Zhao, T., Wang, J., 2014. A multicriteria evaluation method for 3D building reconstruction. IEEE Geoscience and Remote Sensing Letters 11 (9), 1619–1623.

Zeng, H., Wu, J., Furukawa, Y., 2018. Neural procedural reconstruction for residential buildings. In: European Conference on Computer Vision (ECCV), 8-14 September 2018, Munich, Germany. pp. 737–753.

Zhang, L., Zhang, L., 2018. Deep learning-based classification and reconstruction of residential scenes from large-scale point clouds. IEEE Transactions on Geoscience and Remote Sensing 56 (4), 1887–1897.

Zhou, Q.-Y., Neumann, U., 2012. 2.5 D building modeling by discovering global regularities. In: IEEE Conference on Computer Vision and Pattern Recognition, 16-21 June 2012, Providence, USA. pp. 326–333.

Zhou, Q.-Y., Neumann, U., 2013. Complete residential urban area reconstruction from dense aerial lidar point clouds. Graphical Models 75 (3), 118 – 125.

Zhu, L., Shen, S., Gao, X., Hu, Z., 2018. Large scale urban scene modeling from mvs meshes. In: European Conference on Computer Vision (ECCV), 8-14 September 2018, Munich, Germany. pp. 614–629.