Semantic evaluation of 3D city models

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

Semantic evaluation of 3D city models

Oussama Ennafii[1,2], Arnaud Le-Bris[1] Florent Lafarge[2], and Clément Mallet[1]

[1] Univ. Paris-Est, LaSTIG MATIS, IGN, ENSG, 94160 Saint-Mandé, France
[2] Inria, Titane, 06902 Sophia Antipolis, France

Abstract. The automatic generation of 3D urban models from geospatial data is now a standard procedure. However, practitioners still have to visually assess, at city-scale, the correctness of these models and detect inevitable reconstruction errors. Such a process relies on experts, and is highly time-consuming (2 h/km^2/expert). In this work, we propose an approach for automatically evaluating the quality of 3D building models. Potential errors are compiled in a hierarchical, versatile and parameterizable 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, image and depth data. A baseline of handcrafted, yet generic features, is fed to a Random Forest classifier. Both multi-class and multi-label cases are considered. Due to the interdependence between classes of errors, we have the ability to retrieve all errors at the same time while predicting erroneous buildings. We tested our framework on an urban area with more than 1,000 building models. We can satisfactorily detect, on average 96% of the most frequent errors.

Keywords: 3D urban modeling, buildings, quality assessment, taxonomy, classification, error detection, geometry, geospatial imagery, depth.

1 Introduction

3D urban models have a wide range of applications. They can be used for ludic purposes (video games or tourism) as much as they can be vital in more critical domains with significant societal challenges (e.g., run-off water or microclimate simulation, urban planning or security operations preparation) [1], [2]. Therefore, automatic urban reconstruction focuses efforts of both scientific research and industrial activities. However, the problem remains unsolved [2], [3]. In fact, besides the seamless nature of reconstituted models, current algorithms lack of generic capacity. They cannot handle the high heterogeneity of urban scenes. As such, human intervention is needed either in interaction within the reconstruction pipeline or as a post-processing refinement and correction step. The latter is based on a highly tedious task which requires individual visual inspection of buildings [2]. Consequently, for all stakeholders (from researchers up to end-users), the automatic evaluation of 3D building models remains a critical step, especially in a production environment. It has been barely investigated in the literature. This paper addresses this issue.

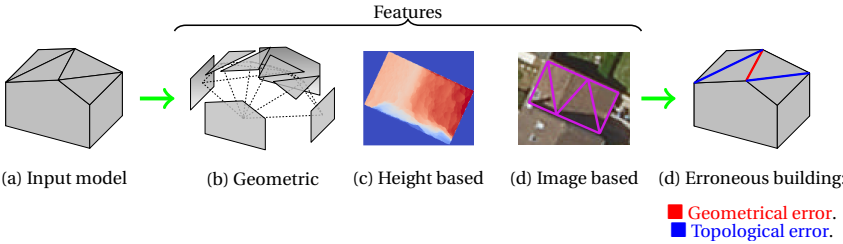


Fig. 1: The semantic evaluation paradigm proposal: in addition to the input model topological structure depicted in (b), features are extracted from comparison to height maps, as represented by the difference computed between the model height and the Digital Surface Model in (c). Images can also be used to characterize models by comparing their projected edges to local gradients (cf. (d)). Based on the computed features, semantic errors affecting the building are predicted using a pretrained classifier.

Our work focuses on assessing polyhedral structured models, representing building architectures. These models result from a given urban reconstruction method, e.g. [4]. Compared to triangle meshes that are extracted from multiview images or point clouds, the studied objects are, by design, more compact but less faithful to input data. In counterpart, they hold more semantic information as each polygonal facet typically corresponds to a façade, a roof or any well defined architectural feature. $3D$ modeling algorithms traditionally build a compromise between compactness of the representation and fidelity to the input data (meshes or 3D points). Depending on its spatial resolution, the urban environment, and the targeted application, the reconstituted result achieves a certain **Level of Detail (LoD)** [5]. A LoD-1 model is a simple building extrusion (flat roof). A LoD-2 model considers geometric simplification of buildings, ignoring superstructures, such as dormer windows and chimneys. These are taken into account in LoD-3. The LoD rational is still open for debate [6]. Nevertheless, in this paper, we will follow the LoD categorization introduced above, which is standard in the computer vision literature [7].

A large body of papers has addressed the 3D building modeling issue and subsequently tried to find the trade-off between fidelity and compactness [4], [8], [9], [7]. Few works investigate the issue of assessing the quality of the derived models, especially out of a given reconstruction pipeline. Usually, quality assessment is based on visual inspection [10], [11] or geometric fidelity metrics [12] without any localized semantic dimension. Only one benchmark dataset [3] exists and is not widely used [13], [14], [15]. This motivates the need for a well suited quality assessment paradigm. Since the models to be diagnosed display strong structural properties, an unconstrained evaluation based on data fidelity metrics, as

in [16], is too general. The evaluation should also ignore format issues or geometric consistencies as proposed in [17], as they must have been ruled out well before this stage. Instead, we target a *semantic* evaluation in which building semantics should be taken into account through the detection and categorization of modeling errors at the facet level for each 3D building. The framework should be independent from the Level of Detail and the modeling method which is regularly evaluated based on the minimized metrics during the reconstruction process. Thus, we define an evaluation framework that can be used for:

- **Building model correction**: for the automatic or interactive [18] refinement of building models using the detected errors.
- **Change detection**: modeling errors can straightforwardly stem from changes, which often occur in urban environments [19]. Conversely, changes can be implicitly detected from other defects.
- **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** [20]: by categorizing user behaviors during crowd-sourced modeling and vandalism detection process [21].

This work proposes an adaptable and flexible framework indifferent to input urban scenes and reconstruction methods (Figure 1). For that purpose, our contributions are three-fold:

- A new **taxonomy of errors**, hierarchical, adapted to all LoD, 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** which are extracted both from the model and 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 results of experiments conducted over an urban area are presented in Section 4. Main conclusions are drawn in Section 5.

2 Related Work

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

Reference data types. Existing methods rely on two types of reference data in order to compare models to. First comes manually plotted ground truth data with very high spatial accuracy. These models can be obtained either from field measurements [4], [12] with the highest possible precision ($\sigma(\text{error}) \approx 0.05\,\text{m}$), or using stereo-plotting [22], [12], [8], [23]. However, such an approach does not scale

well. The other alternative is the direct comparison with raw data. For instance, models can be compared to LiDAR point clouds, height maps [24], [25], [13] or geospatial multi-view images as in [26], [27]. They are, however, not always helpful: these are the input data used by modeling methods and such comparisons are often the basis for their fidelity criterion.

Evaluation outputs. The quality assessment methods can produce two kinds of outputs. **Geometric fidelity metrics** summarize the quality of the whole assessed model. These indices are computed at different levels: specific points of interest (such as corners or edge points) average precision [28], [12], surface dissimilarity [22], [4], [12], [8], [25], [23], [13], [14] or volume discrepancy to reference data [22], [23]. The obtained outputs have the drawback of being too general for the special case of urban structured models. Indeed, their diagnosis, far from surface reconstruction evaluation [16], needs to pinpoint specific types of errors that can be easily corrected once identified [29]. On the other hand, **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 "rejected") [26]. However, these errors depend on a definition of the end-user 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 [27] between footprint errors ("erroneous outline", "inexistent building", "missing inner court" and "imprecise footprint"), intrinsic reconstruction errors ("over segmentation", "under segmentation", "inexact roof" and "Z translation") and "vegetation occlusion" errors. In the latter methods [26], [27], the evaluation is casted 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 (0.2 m to 0.25 m) images and Digital Surface Models (DSMs), like 3D segment or texture correlation score comparisons. In spite of their semantic contribution in quality evaluation, such taxonomies are prone to overfitting to specific urban scenes or modeling algorithms.

Main objective. This work defines a new quality evaluation paradigm when only the most accessible unstructured data could be provided. It should also be capable of detecting semantic localized errors independently from the used reconstruction method(s) and the urban environment.

3 Problem formulation

To evaluate reconstituted $3D$ models, a hierarchical error taxonomy is established. From the latter, we deduce, depending on the evaluation objectives, error labels that can pinpoint defects altering a building model. A set of buildings are thus annotated in order to train a supervised classifier that will be used for prediction on other models.

The error taxonomy is **parameterizable** and **agnostic** towards reconstructed inputs, no matter which modeling method or urban scenes are studied. Furthermore, it does not require onerous reference data aside from the annotated objects on which the classifier is trained.

The quality assessment pipeline is also **modular**. Building models are represented by intrinsic geometric features extracted from the model facets graph. If available, the classifier can also be fed with additional depth related features, based on the comparison of the model altimetry and the DSM, in case of geospatial reconstruction, or, in general, any depth map comparison. Eventually, image information can be incorporated into the pipeline through spectral or textural information available in satellites, aerial or street view 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 to correct a model. *Atomic* errors can thus be heuristically correlated to independent actions that an operator or an algorithm needs to choose to correct building models.

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 the detailed error description.

At a first level, model qualifiability is studied. In fact, aside from formatting issues or geometric inconsistencies [17], other reasons make building models unqualifiable. For instance, buildings can be occluded by vegetation. 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 to be 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 $\cup$ LoD-1. At the next LoD-2, the family "*Facet Errors*" assembles defects that can damage façade or roof fidelity. The last error family, *i.e.*,"*Superstructure Errors*", describes errors

that involve superstructures modeled at LoD-3. Only the first two families are represented in Figure 2. The last one will not be studied in further experiments.

Each family contains *atomic* errors of maximal *finesse* equal to 3. Although they can co-occur in the same building model and across different error 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 predefined 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 **exclusivity**. It is based on family error hierarchization. If errors of a certain LoD family are detected, the ones with higher LoD orders are considered meaningless and thus are not reported.

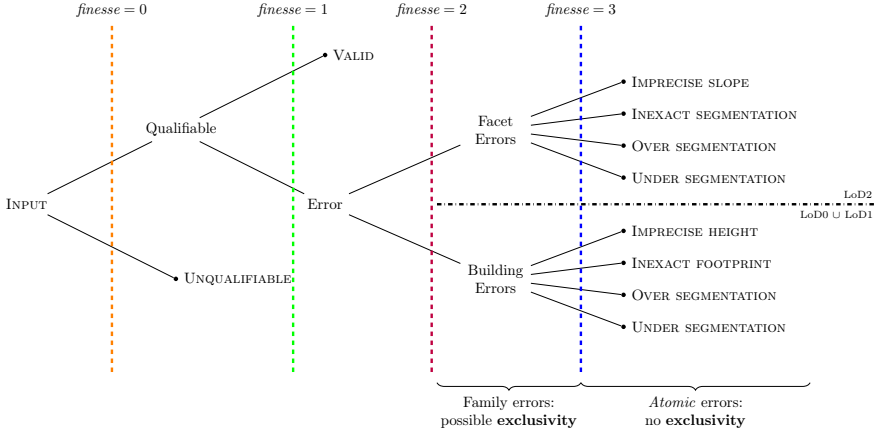


Fig. 2: The proposed taxonomy structure. In our case of very high resolution overhead modeling, only two family errors are depicted. At *finesse* level 2, hierarchization is possible: the **exclusivity** parameter can thus act. However, it is not the case at the *atomic* errors level since they are independent.

The geospatial overhead modeling case. This study is, henceforth, narrowed to the satellite and aerial reconstruction case. The objective is to reconstruct large urban scenes using Very High Resolution geospatial images or, if

available, LiDAR point clouds. These data could also be used later to assess modeled buildings.

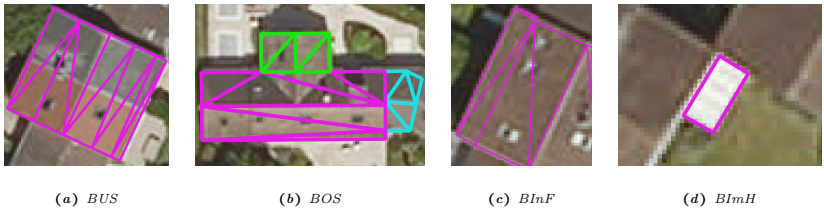
In the present case, $2.5D$ buildings are evaluated. The next *atomic* errors are proposed:

- **Building errors** family (cf. Figure 3.i):
  - *Under segmentation* (*BUS*): two or more buildings are modeled as one;
  - *Over segmentation* (*BOS*): one building is subdivided into two or more buildings;
  - *Inexact footprint* (*BInF*): erroneous building footprint, grouping geometric inaccuracies and topological defects as missing inner courts (≡ not the right number of polygon holes);
  - *Imprecise height* (*BImH*): wrong building height estimation;
- **Facet errors** family (cf. Figure 3.ii):
  - *Under segmentation* (*FUS*): two or more facets are modeled as one;
  - *Over segmentation* (*FOS*): one facet is subdivided into two or more facets;
  - *Inexact segmentation* (*FIS*): facet edges are inaccurate;
  - *Imprecise slope* (*FImS*): wrong facet slope estimation.

These errors are illustrated in Figure 3. In (3.i.(a)), two distinct buildings can be visually identified while they are grouped into one building. The contrary happens when a single building is subdivided in three parts as in (3.i.(b)). *BInF* can be detected easily when it ensues a wrong outline as illustrated in the top right corner of (3.i.(c)). Depth information is hard to convey using only one image, as shown in (3.i.(d)). The balcony, besides being detached from its building, has its height wrongly estimated due to the influence of the main building. It is also impossible to deduce the slope mis-evaluation without a depth map as depicted in (3.ii.(d)). Another fidelity error can be seen in (3.ii.(c)), as the central edge that links the two main roof sides does not correspond to the image position. Facets can also suffer from over segmentation as both roof sides are in (3.ii.(b)). To complete the picture, (3.ii.(a)) illustrates how a model roof facet can be under segmented.

3.2 Feature baseline

In order to predict errors, models need to be described using relevant attributes. Since there is no comparable work that studies the previously defined errors, we propose a new baseline of features. They are kept simple so as to be used in most situations relying on generally available data. Indeed, they are based on depth map comparison, segment and image gradient pairing or the model intrinsic structural characteristics. We avoid computing and comparing $3D$ lines [27], correlation scores [26] or, in general, any Structure-from-Motion (SfM) based metric [18]. In addition of being very costly, these features are methodologically endogamous to the $3D$ modeling techniques used to produce the assessed models. In other words, they are vulnerable to the same defects that may be overlooked during modeling in the first place.



(a) *BUS* (b) *BOS* (c) *BInF* (d) *BImH*

i. Building errors family samples.

(a) *FUS* (b) *FOS* (c) *FIS* (d) *FImS*

ii. Facet errors family samples.

Fig. 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.

The presented approach offers another flexibility lever related to the input data. The model itself can be directly used in order to discover flaws based on its geometrical structure compared to the dataset statistics. Dense depth information can be added, through for instance a DSM, in order to help detecting defects that can be hardly discriminated otherwise. Eventually, optical images can bring additional information critical for semantic heavy segmentation evaluation (high frequencies and texture). Each modality is described herein in detail.

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 [30], the input building model can be seen as a facet (dual) graph:

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

For each facet $f \in \mathsf{F}$, we compute its degree (*i.e.*, number of vertices; $d(f) \triangleq |\{v : v \text{ is a vertex of } f\}|$), area $\mathscr{A}(f)$, circumference $\mathscr{C}(f)$, centroid $\mathscr{G}(f)$ and normal $\boldsymbol{n}(f)$. Statistical characteristics are then computed over building model facets using specific functions S, like a histogram $S^p_{hist} : l \mapsto histogram(l,p)$, with p standing for histogram parameters. Another simple option could be $S_{synth} : l \mapsto \big[\max(l)\ \min(l)\ \bar{l}\ \text{median}(l)\ \sigma(l)\big]$ where $\bar{l}$ (*resp.* $\sigma(l)$) represents the mean (*resp.* the standard deviation) over a tuple l.

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

$$v_{geometric}(\mathsf{M}) = \begin{bmatrix} S\Big(\big(d(f)\big)_{f\in\mathsf{F}}\Big) \\ S\Big(\big(\mathscr{A}(f)\big)_{f\in\mathsf{F}}\Big) \\ S\Big(\big(\mathscr{C}(f)\big)_{f\in\mathsf{F}}\Big) \\ S\Big(\big(||\mathscr{G}(f)-\mathscr{G}(g)||\big)_{(f,g)\in\mathsf{E}}\Big) \\ S\Big(\big(\arccos(\boldsymbol{n}(f),\boldsymbol{n}(g))\big)_{(f,g)\in\mathsf{E}}\Big) \end{bmatrix}. \tag{2}$$

Additionally to individual facet statistics, regularity is taken into account by looking into adjacent graph nodes as in [31]. Such features express only a small 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 been considered for the moment.

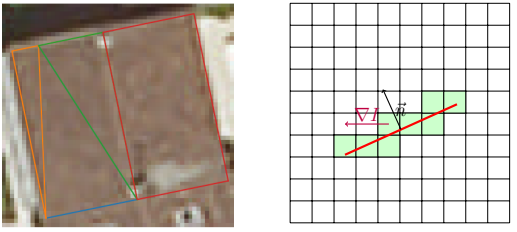
Height based features. For this modality, raw depth information is provided by a Digital Surface Model as a $2D$ height grid dsm. It must have been produced around the same time of the $3D$ reconstruction so as to avoid probable temporal changes. It is compared to the model altimetry like in [32], [8]. The latter is inferred from its facets plane equations. It is then rasterized into the image alt at the spatial resolution of the dsm. Their difference reveals a discrepancy map that can be exploited for the prediction (cf. Figure 1.c). A baseline approach is proposed relying on pixel values statistics computed using previously defined functions S.

$$v_{height}(\mathsf{M}) = S\big(dsm - alt\big) \tag{3}$$

Equation 3 summarizes how building height based features are computed. Different from a root mean square metric [25], [33], the histogram captures the the discrepancy distribution. However, as for the previous geometric attributes, structural information coming from the model is lost.

Image features. We can benefit from high frequencies existing in Very High Spatial Resolution optical images. Building edges correspond to sharp discontinuities in images [34]. The idea is to compare these edges to local gradients in order to look for inconsistencies. In an ideal setting, in an image I, gradients computed at pixels g that intersect any segment s from the building projection $(4.a)$ will almost be collinear with its normal. In consequence, we qualify, applying the same statistical functions S, the distribution of the normalized gradient scalar product with the normal all along a facet edge:

$$\mathsf{D}_S(s,I) \triangleq S\bigg(\Big(\frac{\nabla I(g)\cdot\boldsymbol{n}(s)}{\|\nabla I(g)\|}\Big)_{g\in I \text{ and } g\cap s}\bigg). \tag{4}$$



(*a*) A building model projection superimposed on the aerial image.

(*b*) The green squares represent intersecting pixels with the red segment. The gradient vector in purple is compared to the segment normal in black.

Fig. 4: Illustration of features that can be derived from optical images. For each model facet, the corresponding polygon projection edges are compared to local gradients as in (b).

Once the distribution is computed over a segment, it is compiled 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 segments s forming the polygon projection $q(f)$ of the facet f on as the image I. In order to take into account the variability of segment dimensions, this sum is weighted by segment lengths.

$$D_{S^p_{hist}}(f, I) \triangleq \sum_{s \in q(f)} \|s\| \cdot \mathsf{D}_{S^p_{hist}}(s, I). \tag{5}$$

The same can be done over all facets of a building M, resulting in equation 6. 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 formulation in [13]. Once again, the model structure is partially lost when simply summing histograms over all segments.

$$v_{image}(\mathsf{M}) = D_{S^p_{hist}}(\mathsf{M}, I) \triangleq \sum_{f \in \mathsf{F}} \mathscr{A}(q(f)) \cdot \mathsf{D}_{S^p_{hist}}(f, I). \tag{6}$$

3.3 Classification process

When designing the classification process, two sources of flexibility are to be taken into account: the parametric nature of the taxonomy and the feature vector heterogeneity. The first means that labels to predict are not fixed but depend on the specified parameters. The second means that the classifier must adapt well to different input vectors types and sizes.

Classification problems. Both the classification problem nature and the set of labels to work with are determined by the three previously defined taxonomy

parameters (cf. Table 1). The first **target *finesse*** $= 1$ level corresponds to a binary classification problem: ' 'Valid" or "Erroneous". At the next one *finesse* $=$ 2, the other parameters intervene. The **eLoD** can take then two values in the aerial reconstruction case: LoD-1 or LoD-2. If fixed at LoD-1, it is a binary classification problem: "Valid" or "Building error". For LoD-2, if the **exclusivity** is on, it turns into a multi-class problem: "Valid", "Building error" or "Facet errors", while, if set off, it becomes a multi-label one: "Valid", "Building error" and "Facet errors". At the last *finesse* $= 3$ level, if the **exclusivity** is on, it is a 2-stage classification problem. In the first stage, a multi-class, or simply binary in case $eLoD = LoD1$, problem, like in the previous semantic degree, predicts the error family, after which a second multi-label problem decides between the predicted error family children. If the **exclusivity** is off, it changes into 1-stage multi-label problem that guesses the existence of each atomic error corresponding to the chosen eLoD.

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).

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 binary classification). The multi-label problem decides, for multiple labels, the most probable state: present $(+1)$ or absent (-1). In the multi-stage setting, one decides, at each level, the most probable class or labels which impact the next stages of prediction. Errors are easily propagated in the last case. That is why, the rest of the study is not interested in this case. Since it focuses on semantic evaluation, the first *finesse* degree is not experimented on. Furthermore, it can be inferred from higher levels of *finesse*.

Classifier choice. The highly modular nature of proposed features involving a great number of parameters restricts the choice of classifiers. Random forest classifiers [35], [36] were retained in this setting. In fact, they can manage a great 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 also to any 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 Experiments

4.1 Data

We evaluate our approach using a 3D city model over the city of Elancourt (France). The studied scene spans an area of 15 km^2. It exhibits a high diversity of building types and errors: residential districts with mostly bi-level buildings, industrial areas with flat roof buildings, a stadium, a petrol station and administrative edifices such as schools. These were modeled, using the algorithm described in [10], out of existing building footprints and an aerial multi-view DSM with a 0.06 m spatial resolution. The modeling algorithm simulates possible constrained roof structures. The best one is selected after scoring the extrapolated roofs. Finally, orthogonal building façades connect the best roof to the ground. The produced $2.5D$ models have a LoD-2 level. This method is adapted to roof types of low complexity and favors symmetrical models. Therefore, a high number of errors exist in our test case. 1,501 buildings are considered in these experiments. They were annotated according to the atomic errors list provided by our taxonomy. Table 2 reports statistics over the annotated dataset.

Error Family	**Occurrence ratio**	***Atomic* error**	**Family conditional occurrence ratio**	**Absolute occurrence ratio**
Unqualifiable	0.0180	—	—	—
Building Errors	0.8235	Over segmentation	0.8285	0.6822
		Under segmentation	0.2824	0.2325
		Imprecise footprint	0.1521	0.1252
		Imprecise height	0.0057	0.0047
Facet Errors	0.7249	Over segmentation	0.8971	0.6502
		Under segmentation	0.1314	0.0953
		Imprecise segmentation	0.1351	0.9793
		Imprecise slope	0.0193	0.0140

Table 2: Ground truth statistics over the dataset containing 1501 buildings. *Atomic* errors are miscellaneously represented.

Both error families are highly present in the dataset. Only a small fraction (27 samples) of instances are unqualifiable, being occluded, completely or partially, by vegetation. At the atomic level, apart from building and facet over segmentation cases, most errors are under-represented with the extreme case of height imprecision error (7 samples). The unbalanced nature of our datasets obviously affects our results, as highlighted later in the next Section.

4.2 Results

	Geometry		Geom. ∪ Height		Geom. ∪ Image		All	
	Recall	**Prec.**	**Recall**	**Prec.**	**Recall**	**Prec.**	**Recall**	**Prec.**
Building errors	99.76	**84.16**	99.76	84.11	99.92	83.84	**100**	83.85
Facet errors	90.81	98.02	**91.08**	**98.41**	90.44	97.62	90.53	97.82

Table 3: Test results expressed in percentage for the *finesse* = 2 case. All four configurations are compared across both family errors. **Geom.** stands for geometric

Based on the devised pipeline, four feature configurations were tested: "geometric features" only, "geometric and height features", "geometric and image features" as well as "geometric, height and image features". Each feature modality produces a 20 dimension vector. A 0.06 m spatial resolution DSM and a 0.2 m pixel size orthorectified image are used to derive height and image features. Labels are extracted from a non **exclusive** and **eLoD** = LoD-2 taxonomy. Both *finesse* levels 2 and 3 are tested. We perform a 10-fold cross validation. The overall accuracy is not interesting regarding the highly unbalanced nature of labels.

	Geometry		Geom. ∪ Height		Geom. ∪ Image		All	
	Recall	**Prec.**	**Recall**	**Prec.**	**Recall**	**Prec.**	**Recall**	**Prec.**
BOS	**94.34**	77.09	93.36	**77.66**	92.38	77.04	92.19	76.69
BUS	34.38	76.43	31.52	**78.57**	**42.98**	76.53	41.55	77.96
BInf	**22.34**	**68.85**	22.34	67.74	18.09	64.15	18.62	64.81
BImH	0	—	0	—	0	—	0	0
FOS	98.77	**98.77**	**98.87**	98.67	98.67	98.47	98.67	98.37
FUS	**0.70**	**50.00**	0.70	33.34	**0.70**	**50.00**	**0.70**	**50.00**
FIS	**1.36**	**66.67**	1.36	50.00	1.36	28.57	1.36	40.00
FImS	0	—	0	—	0	—	0	—

Table 4: Test results reported in percentage for the *finesse* level 3. All *atomic* errors are considered over all possible configurations. Ratios in bold represent the higher ones for each error.

4.3 Discussion

Three criteria are considered when investigating qualitative results (Tables 3-4). The first analysis dimension involves *finesse*. Scores are compared, between both *finesse* levels, by averaging across all feature configurations. In terms of

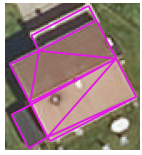
precision, at *finesse* level 3, compared to *finesse* level 2, "Facet errors" family looses 7.2%, while "Building errors" gains only 0.6%. The same *finesse* level comparison reveals that recall is diminished by 2.0% for "Facet errors" and 8.3% for "Building errors". Thus, if limited to determining the error family, training the model at the *finesse* level 2 is the best alternative. Feature configurations are also studied in order to assess image and height attributes contributions. In general, results vary within a 2% margin except in three cases. First, the precision instability for facet inexact and under segmentation can be spotted, as they are under detected: only one or two are rightfully identified as errors, while the number of valid instances, that are predicted as erroneous, vary between 1 and 5. Secondly, inexact footprint recall decreases by around 4% in precision, when image features are added. This could result from the insufficient spatial resolution of optical images. However, in the last case, image features help "Building under segmentation" error gaining around 9% in recall while loosing only 1.5% in precision. This may be explained by the high radiometric heterogeneity of building roofs. Finally, scores are distinguished according to *atomic* errors. As predicted, defects that are well represented in the dataset achieve high recall and precision values. The others, being present at rates lower than 23%, are not well predicted, especially in the case of very rare errors: height imprecision (7) or slope errors (21).

Qualitative assessment is also performed in order to illustrate some failure cases (Figure 5, from left to right). In the first image, the similarity of the building outline to over segmented buildings cases induces an overdetection. In the second example, the building is wrongfully detected as being under segmented due to the presence of a balcony and a smaller annex building. In the third building, while correctly predicting *BOS*, our algorithm fails to detect the under segmented roof. Finally, in the last depiction, except the well caught footprint error, defects are overlooked as there are few comparable samples in the dataset. To alleviate these issues, more robust features could be introduced taking into account higher order information. Dataset enrichment could be another option which provides more instances of underrepresented errors. In the end, we can also add the human in the loop through manual interactive evaluation which can adapt well to user-specific needs.

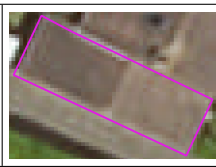
5 Conclusion

We proposed a framework to semantically evaluate automatically modeled buildings. For that purpose, errors are hierarchically organized into a flexible and parametrized taxonomy. It aims to handle the large diversity of urban enviroments 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 classifier. Each model provides intrinsic geometrical characteristics that are compiled in a feature vector. Other modalities can help describing building

Errors	G.T.	Pred.
BOS	✗	✓
Valid	✓	✗



Errors	G.T.	Pred.
BUS	✗	✓
FImS	✓	✗
FOS	✓	✗



Errors	G.T.	Pred.
BOS	✓	✓
FUS	✓	✗

Errors	G.T.	Pred.
BOS	✓	✗
FOS	✓	✗
BUS	✓	✗
BInF	✓	✓

Fig. 5: Predicted (Pred.) errors compared to ground truth (G.T.) labels are illustrated for some pathological cases. Knowing how each error is represented in the dataset helps interpreting mispredictions.

models, as attributes can also be extracted from model comparison to images or depth data. It helps detecting hard cases.

This new framework was applied to the case of aerial urban reconstruction, where features are extracted from geospatial images and a DSM. A dataset containing 1,501 aerial reconstructed building models with high diversity was used to test the devised evaluation method associated to multimodal baseline features. Although being mitigated over under-represented errors, results are satisfactory in the well balanced cases. As a next step, more structurally aware features (based on graph comparison, for instance) could be proposed so as to be applied on a richer and more diverse dataset (potentially involving data augmentation) under a deep-based framework.

References

1. Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Çöltekin, A.: Applications of 3D city models: state of the art review. ISPRS International Journal of Geo-Information **4**(4) (2015) 2842–2889
2. Musialski, P., Wonka, P., Aliaga, D.G., Wimmer, M., van Gool, L., Purgathofer, W.: A survey of urban reconstruction. EUROGRAPHICS 2012 State of the Art Reports **XX** (2012) 1–28
3. Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., Breitkopf, U., Jung, J.: Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing **93** (2014) 256–271
4. Dick, A.R., Torr, P.H., Cipolla, R.: Modelling and interpretation of architecture from several images. International Journal of Computer Vision **60**(2) (2004) 111–134
5. Kolbe, T.H., Gröger, G., Plümer, L.: CityGML: Interoperable access to 3D city models. In: Geo-information for disaster management. Springer (2005) 883–899
6. Biljecki, F., Ledoux, H., Stoter, J.: An improved LOD specification for 3D building models. Computers, Environment and Urban Systems **59** (2016) 25–37
7. Verdie, Y., Lafarge, F., Alliez, P.: LoD generation for urban scenes. ACM Transactions on Graphics **34** (2015) 30

8. Zebedin, L., Bauer, J., Karner, K., Bischof, H.: Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. In: European Conference on Computer Vision. (2008)
9. Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M.: Structural approach for building reconstruction from a single DSM. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(1) (2010) 135–147
10. Durupt, M., Taillandier, F.: 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) (2006) 142–147
11. Macay Moreia, J.M., Nex, F., Agugiaro, G., Remondino, F., Lim, N.J.: From DSM To 3D building models: a quantitative evaluation. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XL-1/W1**(1) (2013) 213–219
12. Kaartinen, H., Hyyppä, J., Gülch, E., Vosselman, G., Hyyppä, H., Matikainen, L., Hofmann, A., Mäder, U., Persson, Å., Söderman, U., et al.: Accuracy of 3d city models: Eurosdr comparison. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **36**(3/W19) (2005) 227–232
13. Li, M., Wonka, P., Nan, L.: Manhattan-world urban reconstruction from point clouds. In: European Conference on Computer Vision. (2016)
14. Nan, L., Wonka, P.: Polyfit: Polygonal surface reconstruction from point clouds. In: International Conference on Computer Vision, Venice, Italy. (2017)
15. Nguatem, W., Mayer, H.: Modeling urban scenes from pointclouds. In: International Conference on Computer Vision. (2017)
16. Berger, M., Levine, J.A., Nonato, L.G., Taubin, G., Silva, C.T.: A benchmark for surface reconstruction. ACM Transactions on Graphics **32**(2) (2013) 20
17. Ledoux, H.: val3dity: validation of 3D GIS primitives according to the international standards. Open Geospatial Data, Software and Standards **3**(1) (2018) 1
18. Kowdle, A., Chang, Y.J., Gallagher, A., Chen, T.: Active learning for piecewise planar 3d reconstruction. In: Conference on Computer Vision and Pattern Recognition. (2011)
19. Taneja, A., Ballan, L., Pollefeys, M.: Geometric change detection in urban environments using images. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(11) (2015) 2193–2206
20. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K., et al.: Crowdsourcing in computer vision. Foundations and Trends in Computer Graphics and Vision **10**(3) (2016) 177–243
21. Neis, P., Goetz, M., Zipf, A.: Towards automatic vandalism detection in openstreetmap. ISPRS International Journal of Geo-Information **1**(3) (2012) 315–332
22. Jaynes, C., Riseman, E., Hanson, A.: Recognition and reconstruction of buildings from multiple aerial images. Computer Vision and Image Understanding **90**(1) (2003) 68–98
23. Zeng, C., Member, S., Zhao, T., Wang, J.: A multicriteria evaluation method for 3D building reconstruction. IEEE Geoscience and Remote Sensing Letters **11**(9) (2014) 1619–1623
24. Akca, D., Freeman, M., Sargent, I., Gruen, A.: Quality assessment of 3D building data. The Photogrammetric Record **25**(132) (2010) 339–355
25. Lafarge, F., Mallet, C.: Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. International Journal of Computer Vision **99**(1) (2012) 69–85

26. Boudet, L., Paparoditis, N., Jung, F., Martinoty, G., Pierrot-Deseilligny, M.: 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) (2006) 136–141
27. Michelin, J.C., Tierny, J., Tupin, F., Mallet, C., Paparoditis, N.: 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** (2013) 161–166
28. Vögtle, T., Steinle, E.: 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) (2003) W13
29. Oude Elberink, S.: Quality measures for building reconstruction from airborne laser scanner data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XXXVIII** (2010) 198–203
30. Verma, V., Kumar, R., Hsu, S.: 3D building detection and modeling from aerial LIDAR data. In: Conference on Computer Vision and Pattern Recognition. (2006)
31. Zhou, Q.Y., Neumann, U.: 2.5D building modeling by discovering global regularities. In: Conference on Computer Vision and Pattern Recognition. (2012)
32. Brédif, M., Boldo, D., Pierrot-Deseilligny, M., Maître, H.: 3D building reconstruction with parametric roof superstructures. In: International Conference on Image Processing. (2007)
33. Poullis, C.: A Framework for Automatic Modeling from Point Cloud Data. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(11) (2013) 2563–2575
34. Ortner, M., Descombes, X., Zerubia, J.: Building outline extraction from digital elevation models using marked point processes. International Journal of Computer Vision **72**(2) (2007) 107–132
35. Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32
36. Criminisi, A., Shotton, J.: Decision forests for computer vision and medical image analysis. Springer Science & Business Media (2013)