Plane-based Surface Regularization for Urban 3D Reconstruction

Thomas Holzmann[1]
holzmann@icg.tugraz.at
Martin R. Oswald[2]
martin.oswald@inf.ethz.ch
Marc Pollefeys[2 3]
marc.pollefeys@inf.ethz.ch
Friedrich Fraundorfer[1]
fraundorfer@icg.tugraz.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute of Computer Graphics and Vision
Graz University of Technology
Graz, Austria

[2] Department of Computer Science
ETH Zurich
Zurich, Switzerland

[3] Microsoft
Redmond, USA

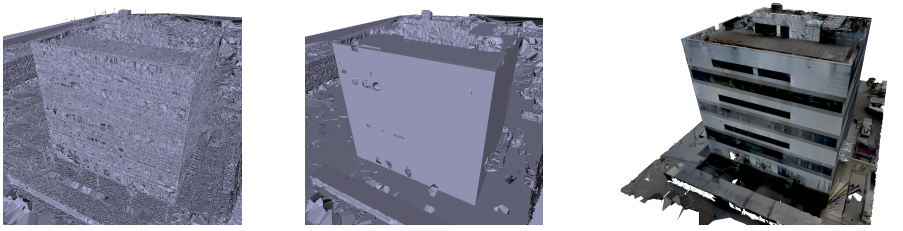
Abstract

We propose a method for urban 3D reconstruction that is a hybrid between a volumetric 3D reconstruction approach and a plane fitting approach in order to obtain a denoised and compact representation of the scene. In our hybrid approach, a single global optimization, using visibility as main information, defines whether the final reconstructed surface should align with a detected plane or rather follow the details of the input data. Our method is based on an established tetrahedral occupancy labeling approach which we taylor for urban reconstruction by adding the possibility to favor an alignment of the surface with detected planes. We further add novel regularization terms that favor Manhattan-like structures and which allow to control the level of detail of the output model. A variety of experiments demonstrate state-of-the-art performance and show that our approach is suitable for both indoor and outdoor environments.

1 Introduction

The 3D reconstruction of urban scenes is an increasingly important topic for the engineering and construction industry for planning, building and verification purposes, because original building plans are often either not available or outdated. Especially with the increasing accuracy and popularity of image-based 3D reconstruction techniques and the availability of low-cost scanning devices, these approaches are increasingly included into project work cycles. Additionally, map services like Google Maps want to offer 3D views for visualization and need a representation suitable for internet transmission. However, the output of typical dense 3D reconstruction approaches is often not directly useful for many applications for which compact and simplified 3D models are often preferred or required. Therefore, scanned 3D models often require large amounts of tedious manual post-processing in order to obtain building information models which are useful for industry applications.

In this work, we aim to minimize manual 3D model post-processing by including typical model simplification steps into the 3D reconstruction process. We present a reconstruction



Labatut et al. [20] proposed, $\alpha_{\text{LoD}} = 375K$ proposed, textured (using [49])

Figure 1: Our approach combines generic 3D reconstruction (like Labatut et al. [20]) with plane fitting and level-of-detail adjustment. Our method produces sharp edges and planar surfaces while simultaneously keeping details like the structures on the roof.

method which favors planar and orthogonal structures like building facades without affecting free-form shapes like vegetation. An example output of our method is depicted in Fig. 1.

Contributions. We present a hybrid approach between generic 3D reconstruction and plane-based urban reconstruction: In contrast to many works that focus only on urban structure reconstruction, our approach is able to deal with a mixture of urban and natural surface structures. Our method robustly deals with noisy, missing and outlier data by computing a consistent watertight surface of arbitrary topology via volumetric occupancy labeling of tetrahedrons via graph-cuts. We present a unified 3D reconstruction framework to jointly favor planar surfaces and orthogonal structures - without explicitly enforcing a Manhattan structure, as well as enforcing user-specified smoothness and level of detail properties. We provide a consistent and minimal approach to account for previously detected planes in the tetrahedral labeling graph by splitting the tetrahedra into smaller ones and adapting the graph and corresponding cost values accordingly. Without any learning, the priors of our method are very generic which makes it well suited for reconstructing any mixture of urban and nature scenes as well as both indoor and outdoor scenarios.

2 Related Work

Urban reconstruction has attracted a large amount of research in the past years, a broad overview is given in the survey by Musialski et al. [30]. The most common property in urban environment is that object surfaces often follow simple geometric shapes such as planes, cuboids, cylinders, spheres and cones. Once detected, they can serve as a strong local shape prior in order to deal with the most common problems in 3D reconstruction, namely: noise, outliers, as well as inconsistent and missing data.

Primitive Fitting. The most common approach is to detect typical primitives in the input point cloud. General methods for fitting arbitrary primitives are presented in [24, 41], but there are also methods for fitting particular primitives like boxes [23], or cylinders, spheres, and cones [19, 22]. The vast majority of works that search for primitives in point cloud data focus on detecting or fitting planes [9, 10, 16, 21, 22, 23, 27, 35, 36, 40, 45]. Some of these approaches are capable of performing real-time plane fitting and geometry simplification [10, 45]. Most approaches that detect planes, simply remove all plane inlier points and replace them with a single polygon. In contrast, our approach combines primitive information with a generic 3D reconstruction approach and lets a global optimization decide whether and where the reconstructed surface should follow a detected plane. To simplify the search and regularization problem a further common assumption for urban scene reconstruction is to only look for axis-aligned piecewise planar structures in an purely orthogonal arrangement,

commonly known has the **Manhattan-world assumption** [13, 23, 40, 50]. In [43], this assumption is slightly relaxed and a mixture of Manhattan frames is computed within a probabilistic approach. Although the above mentioned assumptions are very generic, the majority of works on urban 3D reconstruction focus on either indoor or outdoor scenes.

Indoor Urban Scene Reconstruction. While most primitive fitting approaches use RANSAC [11] for detection, especially methods for indoor reconstruction often rely on a horizontal slicing approach in which all 3D points are vertically projected and collected in a histogram which then shows all walls and major structures as easily detectable maxima [2, 29, 34, 35, 37, 46, 50]. Therefore, these methods additionally require information about the vertical direction. Beyond the reconstruction of walls, several works also segment rooms [6, 16, 29, 32, 46], doors and windows [16, 25, 33], focus on the reconstruction of objects and furniture in indoor environments [17, 26, 42], or on their semantic classification [6, 31].

Outdoor Urban Scene Reconstruction. In [38], a real-time 3D reconstruction system is presented that is mostly generic, but plane sweeping directions for stereo and a depth map hole filling leverage urban structure assumptions. In [36], an iterative approach between plane fitting and plane relationship regularization minimizes the amount of plane directions and has shown to work well on both indoor and outdoor urban environments. The algorithm is specialized to perform fast plane detection and alignment but does not work well on free-form shapes. Similarly, the Manhattan-world reconstruction approach [23] is very fast and works well on outdoor scenes as long as there are no oblique angles or slanted surfaces in the scene. Duan and Lafarge [9] present a work on city reconstruction from satellite images. Starting from a super-pixel segmentation, the algorithms assigns each of the super-pixels a height value, resulting in a very efficient algorithm which allows arbitrary building ground shapes, but similar to other 2.5D approaches [8, 51, 52] they disallow any real 3D structures like overhanging roofs, balconies or bridges.

Approaches like [14, 15] follow the horizontal slicing idea in order to detect walls and cannot model vertically slanted surfaces which leads to stair-casing artifacts for slanted structures like roofs. Moreover, these approaches are specialized to only reconstruct buildings and are not suitable for a hybrid 3D reconstruction of mixed urban and natural scenes. Some building reconstruction approaches explicitly handle level of detail (LoD) control [8, 47, 51].

Generic 3D Reconstruction Approaches. Since we follow a hybrid approach we briefly mention related volumetric 3D reconstruction approaches which typically perform global optimization to compute the scene topology and a surface geometry which best explains the input images for given photometric constraints and pre-defined surface regularity properties. Our work is based on [18, 20, 48] which compute a tetrahedral Delaunay tessellation of the scene from 3D points that have been found as matches in the input images. Subsequently, a graph-cut approach labels each of the tetrahedra as either occupied or empty. Generally, generic reconstruction approaches do not leverage the information that urban scenes are mostly composed of simple geometric shapes.

Hybrid 3D Reconstruction Approaches. In a series of works, Lafarge et al. [19, 21, 22] presented several hybrid primitive fitting and 3D reconstruction approaches. In [19], they first fit primitives and subsequently mesh remaining unfitted scene parts using Delaunay triangulation, but this is not robust to outliers and often fails to reconstruct free-form parts well. In [21], the tetrahedral solution space is augmented with pre-detected planes. We propose an improved plane augmentation which does not require a dense over-sampling of planes and avoids many unnecessary additional objective variables. A substantially different hybrid reconstruction approach was proposed in [22], in which free-form mesh patches and

primitives are jointly optimized within a non-convex setting via a jump-diffusion process.

In sum, detected primitives often only partially explain the data at hand and the majority of urban reconstruction methods rely on simply heuristics to locally decide whether data deviations from a primitive are due to noise or indeed represent important shape details that should be kept in the final reconstruction. We therefore strive for a hybrid reconstruction approach which is suitable for both indoor and outdoor scenes and in which model simplifications are part of a global optimization process which adheres to consistency with the input data as well as to user-defined properties for surface regularity and level of detail.

3 Surface Reconstruction with Plane-based Regularization

In this section, we describe our method for creating regularized 3D reconstructions of urban environments. After detecting planes in the input point cloud, we partition the point cloud (which can be obtained from any stereo reconstruction algorithm) into tetrahedra via Delaunay triangulation. Although our approach can theoretically deal with any kind of shape primitive, we focus only on planar structures since they represent the most common case in urban environments and more complex shapes can often be decomposed or well approximated with piecewise planar structures. Similar to the generic tetrahedra-based 3D reconstruction approach [20] we compute the reconstructed surface as the interface of a volumetric inside/outside labeling. In order to ensure that detected planes can be part of the solution, we augment the solution space by intersecting the tetrahedralization with detected planes in the scene. We add a pairwise smoothness term which favors Manhattan-like structures and a further data term to adjust the level of detail of the reconstruction.

3.1 Plane Detection

As a first step, we detect planes in the input point cloud using a RANSAC [11]-based approach and use this planes later to denoise the point cloud and further subdivide tetrahedra. The inlier points supporting the plane hypotheses are defined as points with a maximum distance of d_{inlier} to the plane, which is computed as the median minimum point-to-point distance of the whole point cloud multiplied by 5. For every plane, point clusters on the computed plane are estimated using Mean Shift [8]. Finally, there are several detected plane segments for every detected plane.

We chose this RANSAC-based plane detection method, because it estimates the geometric structure of urban or indoor environments sufficiently well. However, our approach could use any shape detection method as preprocessing step which can be approximated by a piecewise planar structure (triangles).

3.2 Plane-based Point Cloud Denoising

We consider all inlier points within the distance d_{inlier} around the plane as noisy samples of the plane. To remove this small noise, we project all inlier points onto the plane. This roughly maintains the original point density in the point cloud and is more efficient than the dense over-sampling of the plane in [21] to enforce the plane to be part of the tetrahedralization. In contrast, we subdivide tetrahedra if necessary for this purpose (see Sec. 3.6).

3.3 Point Cloud Tetrahedralization

We compute a tetrahedralization T of the point cloud via Delaunay triangulation which is defined as follows [6]: Given a point set $\mathcal{P} = \{p_1, ..., p_n\}$, the Voronoi cell associated to each point p_i is the region surrounding the point p_i in which every point is closer to p_i than to any other point in $\mathcal{P}$. The Delaunay triangulation $Del(\mathcal{P})$ of $\mathcal{P}$ is defined as the geometric dual of the Voronoi diagram. Thus, there is an edge between two points if and only if their corresponding Voronoi cells have a non-empty intersection. Such a Delaunay triangulation leads to a partition of the convex hull of $\mathcal{P}$ into d-dimensional simplices, corresponding to triangles in 2D and to tetrahedra in 3D space.

3.4 Tetrahedra Occupancy Labeling

We aim to compute a dense watertight surface as the interface between two disjoint sets labeling every tetrahedron in the scene as either inside or outside. Hence, the surface is fully described by a binary labeling $\ell : T \to \{0,1\}$. To this end, we formulate the following energy minimization problem which expresses each of our goals with a particular energy term:

$$\underset{\ell}{\text{minimize}} \quad E_{\text{Vis}}(\ell) + \alpha_{\text{Man}} E_{\text{Man}}(\ell) + \alpha_{\text{LoD}} E_{\text{LoD}}(\ell) \ . \tag{1}$$

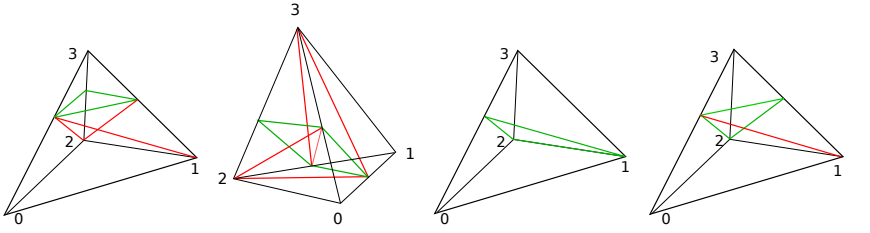
Each of these terms enforces or favors a different property. In particular, E_{Vis} scores the face visibility of tetrahedra, E_{Man} favors Manhattan-like solutions and E_{LoD} allows for level of detail adjustment. The corresponding weights $\alpha_{\text{Man}}, \alpha_{\text{LoD}} \in \mathbb{R}_{\geq 0}$ balance the impact of each term. We compute the globally optimal solution of this energy minimization problem by using Graph Cuts [7]. The visibility-based energy corresponds to the energy minimized in [18], but we slightly modify the overall energy to suit our needs. The following subsections will detail these modifications and will explain all of the terms in Eq. (1).

3.5 Visibility-based Unary and Pairwise Costs

For each cell and for each face, costs are computed using the visibility information that can be derived from given camera-point correspondences. These visibility-based costs $E_{\text{Vis}}(\ell)$ are defined as described in [18]. We define unary costs providing a point-wise prior on the cell occupancy, as well as pairwise costs which locally favor or penalize labeling transitions.

Unary Costs. The unary costs derived from the visibility information are defined as follows: Every cell containing a camera and every infinite cell is labeled as outside by adding infinite weights. Contrarily, every point which is directly behind a vertex (seen from the camera) is labeled as inside. For this, the cell behind the point gets a finite weight for each camera where the point is visible in. Here, we explicitly avoid using infinite weights, since point measurements are prone to noise and may contain outliers.

Pairwise Costs. The pairwise costs are set to penalize ray conflicts for every camera to point correspondence. The pairwise costs are only added in one direction, since faces cannot exist in front of a measured point. Hence, every intersection of a camera-to-point ray with a face gets a constant penalty. In addition to the visibility-based costs, we add constant pairwise costs as a simple regularization term. Although more complicated regularization terms exist in literature (*e.g.*, the beta skeleton term [20]), adding constant costs has shown to be the simplest and most effective regularization term [28].



(a) 3-point intersection (b) 4-point intersection (c) Edge intersection (d) Vertex intersection

Figure 2: Subdivision schemes of tetrahedra intersected by a plane. Depending on the amount of edge intersection points and their locations, the cell needs to be divided differently. The cut by the plane is illustrated with *green* edges, additional edges which need to be inserted are illustrated in *red*. In (a) and (b) there is an edge-plane intersection which results in a new vertex for each intersection. Four new cells are created in (a) and six new cells are created in (b). In (c), the cell-plane intersection follows exactly an edge. Therefore, just one new vertex and two new cells are created. In (d), the cell-plane intersection comprises one cell vertex. The cell gets divided into three new cells by adding two new vertices.

3.6 Tetrahedra Subdivision

After removing the noise of plane inlier points by moving them onto the plane (Sec. 3.2), the plane faces are not necessarily part of the tetrahedralization. We therefore compute intersections of planes and tetrahedra to ensure that all planes are represented as faces in the tetrahedralization. Note that in contrast to [21] that augment the tetrahedralization with a densely sampled representation of every plane, our approach minimizes the amount of added points and tetrahedra. We now briefly describe how we consistently subdivide tetrahedra, for more details we refer to the supplementary material.

For every edge which intersects a plane segment, we subdivide all incident cells of this edge by dividing the cell along the plane. However, a cell division into two parts is not sufficient as cells need to be further subdivided into multiple tetrahedra and kept consistent with their neighbor cells. Therefore, we define subdivision schemes for all possible intersection cases of a cell by a plane (see Fig. 2). Depending on its neighbors, the correct subdivision orientation is selected and, if necessary, the subdivision is adopted to be consistent with all neighbors. Note that the triangulation might not fulfill the Delaunay property after the tetrahedra subdivision, but we do not require this property in the forthcoming processing steps.

Cost Re-Computation for Divided Tetrahedra. Using the visibility-based unary and pairwise costs described in Sec. 3.5, it is necessary to have visibility information available for each point (i.e. camera-point correspondences). As there is no known visibility information for the points created by the tetrahedra subdivision, we compute the visibility-based cost using the original tetrahedralization and propagate them to the subdivided cells. For the unary costs we assign the original cost scaled by the volume of the new tetrahedron and for the pairwise costs the weight is scaled according to the area of the face, if the face of a new cell is part of a face of the old cell. For all other faces (i.e.,faces inside of the old cell), the biggest face from the original cell with a sufficiently small enclosing angle with the new face is selected and scaled according to the area:

$$E_{\text{unary}}(t) = E_{\text{unary}}(t_{\text{orig}})\frac{v_t}{v_{t_{\text{orig}}}} \qquad E_{\text{pairwise}}(f) = E_{\text{pairwise}}(f_{\text{orig}})\frac{a_f}{a_{f_{\text{orig}}}} \quad , \tag{2}$$

where t is the new tetrahedron and v_t its corresponding volume, t_{orig} and $v_{t_{\text{orig}}}$ are the original

tetrahedron and its corresponding volume, respectively.

Further, f, f_{orig} and $a_f, a_{f_{\text{orig}}}$ are the new and original faces with their corresponding areas. In order to select a corresponding original face for new faces not lying on an original one, we retrieve all faces with an enclosing angle smaller than 0.3 *rad* (approx. 17.2 *deg*) and take the biggest one of the retrieved faces. If no face fulfilling this property exists, we increase the maximum enclosing angle by 0.1 *rad* until an appropriate face is found.

With this propagation scheme, the unary costs overall stay the same and the pairwise costs are propagated from faces which are as similar as possible to the new faces and, by taking the biggest one, carry as much as possible information.

3.7 Plane-Aware Regularization

We introduce new regularization terms which allow a continuous choice between generic 3D reconstructions and reconstructions in which the pre-detected planes and Manhattan-like structures are increasingly replacing surface noise and details of the surface structure.

Manhattan Regularity Term. Similar to [22] we introduce a regularity term which favors orthogonal and parallel scene structures. The following term favors label transitions with Manhattan-like surface structures, that is, neighboring faces with enclosing angles similar to 0 or multiples of 90 degrees. More exactly, this term penalizes faces which cannot be part of a Manhattan-like surface structure:

$$E_{\text{Man}}(\ell) = \sum_{f \in T} \mathbf{1}_{\{\ell_{t_1} \neq \ell_{t_2}\}} \frac{a_f}{3} \sum_{e \in f} \min_{g \in \mathcal{N}_e} \left\{ \left| \sin\left(2 \angle(f,g)\right) \right| \right\} \quad , \tag{3}$$

where f denotes a face in the tetrahedralization T, $\mathbf{1}_{\{\cdot\}}$ is the indicator function, ℓ_{t_1} and ℓ_{t_2} are the labels of the adjacent tetrahedra of face f, a_f the area of f, e are the edges of f and $\mathcal{N}_e$ are all incident faces of edge e. As a major advantage, the term favors Manhattan-like structures, but does not strictly enforce them and is therefore applicable to any kind of surface type. Moreover, the term acts completely locally and the surface does not need to be aligned with any world coordinate axis.

Level of Detail Term. In order to control the amount of detail which is removed by the Manhattan term and due to the favoring of pre-detected planes, we introduce another term controlling the amount of removed structure according to its size. For instance, we want to remove the noise on a building roof but keep the chimney. To achieve this, we introduce a new term penalizing the volume deviation of the plane-based reconstruction with respect to the original non-regularized reconstruction. Using the visibility-based energy defined in [18], the original generic 3D reconstruction without favoring planes is defined as $\ell^{\text{Labatut}} = \arg\min_\ell \left[E_{\text{Vis}}(\ell)\right]$. A natural error measure to control structure removal upon plane replacement is the volume difference between the two models. We hence define the level of detail term as follows:

$$E_{\text{LoD}}(\ell) = \sum_{t \in T} v_t \mathbf{1}_{\{\ell_t \neq \ell_t^{\text{Labatut}}\}} \quad , \tag{4}$$

where t defines a tetrahedron in the tetrahedralization T, $\mathbf{1}_{\{\cdot\}}$ is the indicator function, ℓ_t is the labeling of t, and v_t denotes the volume of tretrahdron t. This term acts as the counterpart of the Manhattan regularity term: While the Manhattan regularity term removes details not supported by any plane, this term allows to control the amount of details to be removed.

Plane Intersection Artifacts Removal. The terms work well in most cases, but artifacts may arise near plane intersections, typically along sharp edges in the scene like building outlines. Cells which contain a unary term voting for being inside while actually being outside of the

object may exist enclosed by two planes due to noise within the 3D reconstruction. However, as some of the cells' faces are lying on the planes and, hence, are not penalized by the Manhattan term, the smoothness term is not strong enough to enforce sharp edges.

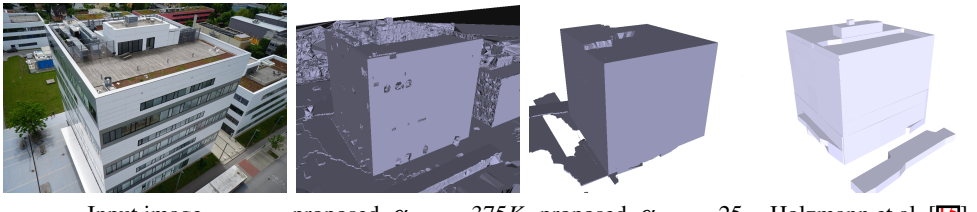
To avoid such artifacts, we reduce the influence of the unary and pairwise terms in scene parts around plane intersections. For every tetrahedron for which its center point has a normal distance d to the plane intersection smaller than $3d_{\text{inlier}}$ (with d_{inlier} being the plane inlier distance defined in Sec. 3.1), we update the unary costs by: $E(t) = E_{\text{orig}}(t)\left(1 - \exp\left(\frac{-d^2}{3d_{\text{inlier}}^2}\right)\right)$, with t being the tetrahedron to update and $E_{\text{orig}}(t)$ being the initial unary cost before the update. In the same way, the pairwise costs at faces are updated.

4 Experiments

In our experiments, we evaluated our method on multiple urban outdoor and indoor datasets and compared it to other state-of-the-art methods. We show that our method yields comparable results while being more flexible than others and that the level of detail of the reconstruction can be adjusted easily with the parameter α_{LoD}. For all the datasets except the *Entry-P10* dataset, we computed the camera poses using the input images within our own Structure-from-Motion pipeline. To compute a (semi-)dense 3D reconstruction, we used PMVS2 [12] and Sure [39]. Further, we used the CGAL library [1] for computing the 3D Delaunay triangulation. For all the experiments, we selected $\alpha_{\text{Man}} = 250K$, as with this setting the whole reconstruction just consists of planar surfaces in combination with a low value for α_{LoD} (see Fig. 3).

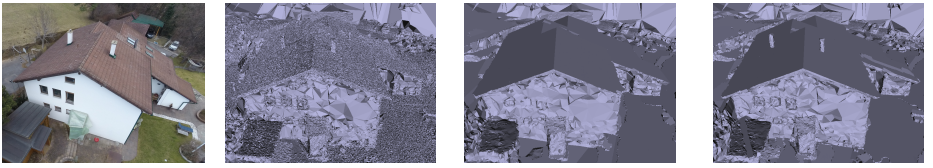
The outdoor dataset *Block Building*, consists of a block shaped building with some additional details on the roof. Therefore, it is well suited to show different reconstruction results when adjusting α_{LoD}. As input, we use a semi-dense reconstruction created by PMVS2 [12] with approx. 5.2M points. Fig. 1 shows a textured result and a result of Labatut et al. [20]. Our method produces well regularized results and generates models with sharp edges and planar surfaces while still containing details which were not supported by any plane.

In Fig. 3, we compare results with varying α_{LoD} to the result of [15]. For high α_{LoD} (middle left) more details are kept in the reconstruction, while for low α_{LoD} (middle right) mostly only plane supported faces are kept. Compared to Holzmann et al. [15], the simplification of planar surfaces is similar, but structures on the roof are represented with more details by our approach. It is worth mentioning that [15] cannot deal with slanted roof sections and will also simplify non-building geometry like vegetation or irregular ground structure.



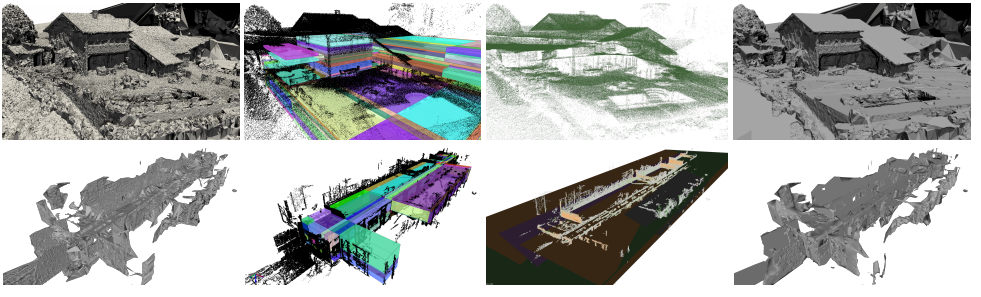
Input image | proposed, $\alpha_{\text{LoD}} = 375K$ | proposed, $\alpha_{\text{LoD}} = 25$ | Holzmann et al. [15]

Figure 3: Results with varying level of detail on the *Block Building* dataset. For high values α_{LoD}, many details of the reconstruction from Labatut et al. [20] which are not supported by a plane are still included in the reconstruction. Contrarily, when setting α_{LoD} low, mostly only plane-supported surfaces are kept. Compared to the proposed method, Holzmann et al. [15] (on the same input) approximates planar surfaces well but misses details on the roof.



Input image | Labatut et al. [20] | proposed, $\alpha_{\text{LoD}} = 250K$ | proposed, $\alpha_{\text{LoD}} = 500K$

Figure 4: Our method with different level of detail settings in comparison to Labatut et al. [20] on the *House* dataset. For a low level of detail (*middle right*), the chimneys are not reconstructed while for high level of detail (*right*) they are included in the reconstruction.



Labatut et al. [20] | Li et al. [23] | Monszpart et al. [27] | proposed

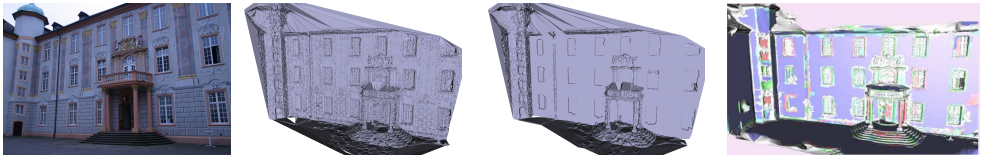
Figure 5: Qualitative comparison to other methods on the *House* dataset (*Top row*), and an *Indoor* dataset (*Bottom row*). While Labatut et al. [20] does not simplify any geometry, Li et al. [23] over-simplifies the scene with only very few boxes. The results of Monszpart et al. [27] were not useful on our datasets as detected planes (only shown in bottom row) did not well align with the geometry. Our approach provides a hybrid reconstruction between primitive-fitted planar parts and generic reconstruction for the free-form parts.

The next outdoor dataset, in the following referred as *House*, is a family house containing a sloped roof with chimneys. This scene is a dense reconstruction created with Sure [39] and down-sampled to 1M points. In Fig. 4, one can observe the varying level of detail, which is especially well observable at the chimneys. Additionally, an example input image and a result from Labatut et al. [20] can be seen in Fig. 4. In Fig. 5, further views of this dataset are depicted and shown in comparison to state-of-the-art mesh simplification methods [23, 27].

We also evaluated on an *Indoor* dataset, which is a reconstruction of a hall with planar walls, some tables and chairs. Again, this is a semi-dense reconstruction created with PMVS2 [12] and consists of approx. 4.6M points. Results and a comparison with other methods can be found in Fig. 5. While Li et al. [23] over-simplifies the geometry by only fitting boxes, Monszpart et al. [27] did not produce any meaningful results on our datasets.

Finally, we compared our method with the method proposed in [22] using the *Entry-P10* dataset [44]. This dataset consists of 10 images captured at ground level. We used the provided camera poses and computed a semi-dense point cloud using PMVS2 [12] consisting of approx. 0.4M points. As can be seen in the results in Fig. 6, our proposed method smoothes the planar surfaces very well while still keeping most of the important details of Labatut et al. [20]. Compared to Lafarge et al. [22], our method delivers a comparable regularization of the planar surfaces.

Runtimes. We executed our algorithm on a computer with an Intel Xeon E5-2680 running at 2.8GHz with 40 cores and 264 GB of RAM. Not all computing steps were parallelized, only the visibility casting used to compute the visibility costs made use of multiple CPUs and



Input image | Labatut et al. [20] | proposed, $\alpha_{\text{LoD}} = 1500K$ | Lafarge et al. [22]

Figure 6: Results of the *Entry-P10* dataset [44]. While the result without shape priors (Labatut et al. [20]) contains a very noisy facade, the proposed approach reconstructs a completely planar surface and simultaneously keeps most of the significant details. The proposed approach also produces a comparable planar regularization of the facade to [22]. The image for Lafarge et al. is taken from [22].

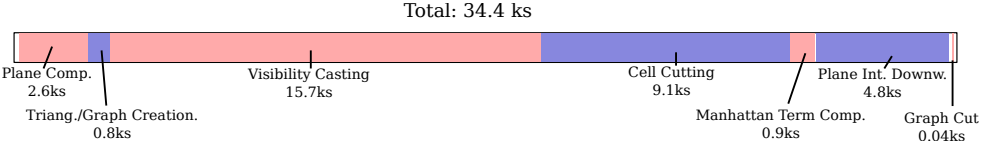


Figure 7: Runtime breakdown for an execution on the *Block Building* dataset, which has a total runtime of approx. 9.5 h (34.4 ks). Most of the processing time is needed for visibility casting and cell cutting. Important processing parts are depicted in red and blue, remaining small processing parts (e.g., data loading, LoD weight computation) are depicted in white.

multiple instances were executed at the same time, which also diminished the parallelization effect. The runtime heavily depends on the amount of input points and cameras and varies from approx. 5 min for the *Entry-P10* dataset (0.4M points, 10 cameras) to approx. 9.5 h for the *Block Building* dataset (5.2M points, 232 cameras). Though, as most of the processing time is needed for preprocessing steps, the final optimization can be rerun with different parameters within less than one minute. A breakdown of the runtime can be found in Fig. 7. The runtimes for Li et al. [23] were in the range of 10 sec, for Holzmann et al. [15] around 50 min, and for Monszpart et al. [27] more than 16 h.

5 Conclusion

We presented a hybrid method for 3D reconstruction of natural scenes containing arbitrary surface structure as well as man-made structures which often exhibit planar shapes in orthogonal and planar alignment. Given the output of an image-based reconstruction approach, we compute a tetrahedral tessellation of the scene and build a corresponding graph to reason about free and occupied space within a graph-cut framework. We provided a consistent and minimal approach to include previously detected planes into the graph structure by splitting intersected tetrahedra into smaller ones and by updating the graph structure and costs accordingly. We introduced a novel combination of Manhattan-like regularization as well as level of detail adjustment to define the level of surface simplification. Our method efficiently computes compact as well as detailed models with state-of-the-art reconstruction quality.

Acknowledgements. This research was funded by the Austrian Science Fund (FWF) in the project V-MAV (I-1537) and by the Austrian Research Promotion Agency (FFG) in the project FreeLine (Bridge1/843450). Martin R. Oswald received funding from the European Union Horizon 2020 research and innovation programme (grant No. 637221).

References

[1] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[2] Antonio Adan and Daniel Huber. 3d reconstruction of interior wall surfaces under occlusion and clutter. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pages 275–281. IEEE, 2011.

[3] H. Arefi, J. Engels, M. Hahn, and H. Mayer. Levels of detail in 3d building reconstruction from lidar data. In *ISPRS*, 2008.

[4] Murat Arikan, Michael Schwärzler, Simon Flöry, Michael Wimmer, and Stefan Maierhofer. O-snap: Optimization-based snapping for modeling architecture. *ACM Trans. Graph.*, 32(1):6:1–6:15, 2013.

[5] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, 1998.

[7] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (12):1222–1239, 2001.

[8] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[9] Liuyun Duan and Florent Lafarge. Towards large-scale city reconstruction from satellites. In *Proceedings European Conference on Computer Vision*, pages 89–104, 2016.

[10] M. Dzitsiuk, J. Sturm, R. Maier, L. Ma, and D. Cremers. De-noising, stabilizing and completing 3D reconstructions on-the-go using plane priors. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017.

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL http://doi.acm.org/10.1145/358669.358692.

[12] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[13] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Manhattan-world stereo. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE Computer Society, 2009. ISBN 978-1-4244-3992-8.

[14] Thomas Holzmann, Christof Hoppe, Stefan Kluckner, and Horst Bischof. Geometric abstraction from noisy image-based 3d reconstructions. In *Proceedings of The 38th Annual Workshop of the Austrian Association for Pattern Recognition (ÖAGM)*, 2014.

[15] Thomas Holzmann, Friedrich Fraundorfer, and Horst Bischof. Regularized 3d modeling from noisy building reconstructions. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 528–536, 2016.

[16] Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. Structured indoor modeling. In *Proceedings International Conference on Computer Vision*, pages 1323–1331, 2015.

[17] Young Min Kim, Niloy J. Mitra, Dong-Ming Yan, and Leonidas J. Guibas. Acquiring 3d indoor environments with variability and repetition. *ACM Trans. Graph.*, 31(6): 138:1–138:11, 2012.

[18] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *Proceedings International Conference on Computer Vision*, 2007.

[19] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Hierarchical shape-based surface reconstruction for dense multi-view stereo. In *International Workshop on 3-D Digital Imaging and Modeling (3DIM), ICCV Workshops*, pages 1598–1605, Kyoto, Japan, October 2009.

[20] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Robust and efficient surface reconstruction from range data. *Computer Graphics Forum*, pages 2275–2290, December 2009.

[21] Florent Lafarge and Pierre Alliez. Surface reconstruction through point set structuring. *Comput. Graph. Forum*, 32(2):225–234, 2013.

[22] Florent Lafarge, Renaud Keriven, Mathieu Brédif, and Hoang-Hiep Vu. A hybrid multiview stereo algorithm for modeling urban scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):5–17, 2013.

[23] Minglei Li, Peter Wonka, and Liangliang Nan. Manhattan-world urban reconstruction from point clouds. In *Proceedings European Conference on Computer Vision*, 2016.

[24] Yangyan Li, Xiaokun Wu, Yiorgos Chrysanthou, Andrei Sharf, Daniel Cohen-Or, and Niloy J. Mitra. Globfit: consistently fitting primitives by discovering global relations. *ACM Trans. Graph.*, 30(4):52:1–52:12, 2011.

[25] H Macher, T Landes, and P Grussenmeyer. Point clouds segmentation as base for as-built bim creation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5/W3(5):191–197, 2015.

[26] Oliver Mattausch, Daniele Panozzo, Claudio Mura, Olga Sorkine-Hornung, and Renato Pajarola. Object detection and classification from large-scale cluttered indoor scans. *Comput. Graph. Forum*, 33(2):11–21, 2014.

[27] Aron Monszpart, Nicolas Mellado, Gabriel Brostow, and Niloy Mitra. RAPter: Rebuilding man-made scenes with regular arrangements of planes. *ACM SIGGRAPH 2015*, 2015.

[28] Christian Mostegel and Markus Rumpler. Robust Surface Reconstruction from Noisy Point Clouds using Graph Cuts. Technical report, Graz University of Technology, Institute of Computer Graphics and Vision, June 2012. https://www.tugraz.at/institute/icg/Media/mostegel_2012_techreport.

[29] Claudio Mura, Oliver Mattausch, Alberto Jaspe Villanueva, Enrico Gobbetti, and Renato Pajarola. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers & Graphics*, 44:20–32, 2014.

[30] Przemyslaw Musialski, Peter Wonka, Daniel G. Aliaga, Michael Wimmer, Luc J. Van Gool, and Werner Purgathofer. A survey of urban reconstruction. *Comput. Graph. Forum*, 32(6):146–177, 2013.

[31] Liangliang Nan, Ke Xie, and Andrei Sharf. A *search-classify* approach for cluttered indoor scene understanding. *ACM Trans. Graph.*, 31(6):137:1–137:10, 2012.

[32] Sebastian Ochmann, Richard Vock, Raoul Wessel, Martin Tamke, and Reinhard Klein. Automatic generation of structural building descriptions from 3d point cloud scans. In *Computer Graphics Theory and Applications (GRAPP), 2014 International Conference on*, pages 1–8. IEEE, 2014.

[33] Sebastian Ochmann, Richard Vock, Raoul Wessel, and Reinhard Klein. Automatic reconstruction of parametric building models from indoor point clouds. *Computers & Graphics*, 54:94–103, 2016.

[34] Sven Oesau, Florent Lafarge, and Pierre Alliez. Indoor scene reconstruction using primitive-driven space partitioning and graph-cut. In *Proceedings of the Eurographics Workshop on Urban Data Modelling and Visualisation*, UDMV, pages 9–12. Eurographics Association, 2013. ISBN 978-3-905674-46-0.

[35] Sven Oesau, Florent Lafarge, and Pierre Alliez. Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut. *ISPRS Journal of Photogrammetry and Remote Sensing*, 90:68–82, 2014.

[36] Sven Oesau, Florent Lafarge, and Pierre Alliez. Planar shape detection and regularization in tandem. *Comput. Graph. Forum*, 35(1):203–215, 2016.

[37] Brian Okorn, Xuehan Xiong, Burcu Akinci, and Daniel Huber. Toward automated modeling of floor plans. In *Proceedings of the Symposium on 3D Data Processing, Visualization and Transmission*, volume 2, 2010.

[38] Marc Pollefeys, David Nistér, Jan-Michael Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, Seon Joo Kim, Paul Merrell, C. Salmi, Sudipta N. Sinha, B. Talton, Liang Wang, Qingxiong Yang, Henrik Stewénius, Ruigang Yang, Greg Welch, and Herman Towles. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2-3):143–167, 2008.

[39] Mathias Rothermel, Konrad Wenzel, Dieter Fritsch, and Norbert Haala. Sure: Photogrammetric surface reconstruction from imagery. In *Proceedings LC3D Workshop, Berlin*, 2012.

[40] Victor Sanchez and Avideh Zakhor. Planar 3d modeling of building interiors from point cloud data. In *ICIP*, pages 1777–1780. IEEE, 2012.

[41] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. Efficient ransac for point-cloud shape detection. *Computer Graphics Forum*, 26(2):214–226, June 2007.

[42] Tianjia Shao, Weiwei Xu, Kun Zhou, Jingdong Wang, Dongping Li, and Baining Guo. An interactive approach to semantic modeling of indoor scenes with an RGBD camera. *ACM Trans. Graph.*, 31(6):136:1–136:11, 2012.

[43] Julian Straub, Guy Rosman, Oren Freifeld, John J. Leonard, and John W. Fisher III. A mixture of manhattan frames: Beyond the manhattan world. In *CVPR*, 2014.

[44] Christoph Strecha, Wolfgang von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2008.

[45] Sebastian Thrun, Christian Martin, Yufeng Liu, Dirk Hahnel, Rosemary Emery-Montemerlo, Deepayan Chakrabarti, and Wolfram Burgard. A real-time expectation-maximization algorithm for acquiring multiplanar maps of indoor environments with mobile robots. *IEEE Transactions on Robotics and Automation*, 20(3):433–443, 2004.

[46] Eric Turner and Avideh Zakhor. Multistory floor plan generation and room labeling of building interiors from laser range data. In *Computer Vision, Imaging and Computer Graphics-Theory and Applications*, pages 29–44. Springer, 2014.

[47] Yannick Verdie, Florent Lafarge, and Pierre Alliez. LOD generation for urban scenes. In *ACM Transactions on Graphics*, 2015.

[48] Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(5):889–901, 2012.

[49] Michael Waechter, Nils Moehrle, and Michael Goessele. Let there be color! - large-scale texturing of 3d reconstructions. In *Proceedings European Conference on Computer Vision*, 2014.

[50] Jianxiong Xiao and Yasutaka Furukawa. Reconstructing the world's museums. *International Journal of Computer Vision*, 110(3):243–258, 2014.

[51] Lukas Zebedin, Joachim Bauer, Konrad F. Karner, and Horst Bischof. Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In *Proceedings European Conference on Computer Vision*, 2008.

[52] Qian-Yi Zhou and Ulrich Neumann. 2.5d building modeling by discovering global regularities. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, pages 326–333, 2012.