Semantically Aware Urban 3D Reconstruction with Plane-Based Regularization - Supplementary Material

Thomas Holzmann, Michael Maurer, Friedrich Fraundorfer, Horst Bischof

Institute of Computer Graphics and Vision, Graz University of Technology
{lastname}@icg.tugraz.at

1 Detailed Evaluation of Proposed Visibility-based Energy

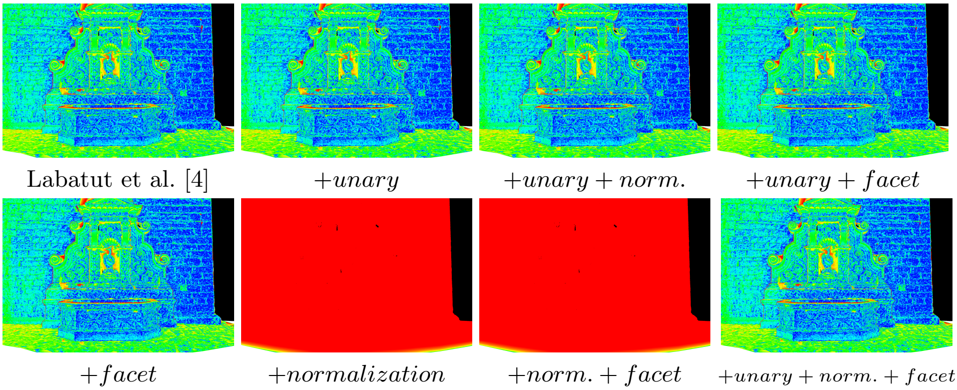
In this section, we evaluate each individual change in the visibility-based energy term compared to Labatut et al. [4] separately and show its effects on the *Fountain-P11* [8] dataset.

In Fig. 1 and Tab. 1 visualized errors and corresponding error and completeness values are depicted. We name the results corresponding to the added energy formulation part w.r.t. Labatut et al. [4] described in Sec. 3.4 in the main paper: *unary* defines the additional unary term in front of a visible vertex, $facet$ means that the pairwise energy terms at facets were assigned in both directions (as opposed to one direction in [4]) and for *normalization* the proposed energy normalization from Eq. 2 and 3 in the main paper was applied. It can be seen visually (Fig. 1 bottom middle) and on the error metrics that the result gets worse when just applying the normalization on the original energy and slightly improves when adding the $facet$ weights in both directions and adding the *unary* term in front of a visible vertex. The combination of all three energy formulation changes, though, delivers the best result in terms of accuracy. Only the completeness is better with normalization only, as then the mesh is just oversmoothed and, hence, the most pixels in the depth map are covered by depth measurements.

To summarize, by adding all three proposed energy formulation changes, we get the best result in terms of accuracy and, simultaneously, deliver a normalized energy which has significant advantages when combining it with additional energy terms.

2 Additional Comparisons for Line-Based Plane Detection

In Fig. 2, the proposed line-based plane detection algorithm is compared against two point-based approaches. In comparison to Schnabel et al. [7], Holzmann et al. [2] does not use point normals and, hence, results in worse detections. However, both point-based methods miss important planes. This is more problematic



Labatut et al. [4] | $+unary$ | $+unary + norm.$ | $+unary + facet$

$+facet$ | $+normalization$ | $+norm. + facet$ | $+unary + norm. + facet$

Fig. 1. *Evaluation of the individual changes compared to [4] in the visibility-based energy on the Fountain-P11 [8] dataset.* As can be seen, when using the normalization without the other two changes, the error is significantly higher due to an oversmoothing of the mesh. All other error images look similar visually.

Table 1. *Error statistics compared to the ground truth.* μ defines the mean and σ the standard deviation of depth values of the result and the ground truth model projected into all cameras. *Completeness* defines the coverage of the projected models in the camera views compared to the ground truth (both defined in [8]). In terms of accuracy, the proposed approach (i.e., adding all three changes in the energy formulation) delivers the best error metrics. Only in terms of completeness other results are slightly better with the tradeoff of lower accuracy and no normalized energy.

	μ (m)	$\sigma(m)$	*completeness*
Labatut et al. [4]	0.0366	0.1458	0.966
+unary	0.0357	0.1434	0.967
+unary+normalization	0.1076	0.2521	0.967
+unary+facet	0.0340	0.1351	0.967
+facet	0.035	0.1376	0.966
+normalization	2.6905	1.2447	**0.977**
+normalization+facet	2.6901	1.2443	**0.977**
+unary+normalization+facet	**0.0338**	**0.1346**	0.966

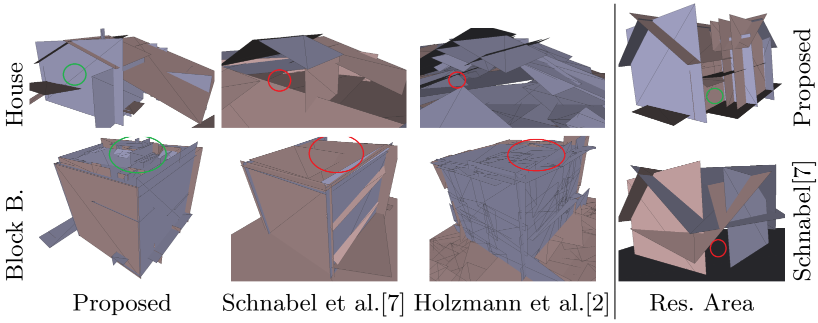


Fig. 2. *Comparison of proposed line-based plane detection with RANSAC-based plane detection of [7] and [2].* Every plane segment is visualized as two triangles. Circles mark planes detected by the proposed approach (green), but not detected by the point-based approaches (red). At the House and the Residential Area dataset several facades were not detected by the point-based approaches which were detected by the proposed approach. At the Block Building, small planes at the constructions on the roof are detected by the proposed approach, where the point-based approaches do not detect any planes. In general, the point-based approaches detect more spurious planes (especially [2]) but also more planes at the ground (less lines reconstructed at the ground).

than detecting too many planes as undetected planar surfaces are not selectable in the result.

3 Comparison to Commercial 3D Reconstruction Software

In Fig. 3 reconstruction results from two commercial reconstruction pipelines are compared with the proposed approach. One can observe that these approaches are not explicitly designed for creating visually appealing urban reconstructions, and up to our knowledge there do not exist commercial products which produce results following our goal definitions.

4 Semantic 3D Data

In this section, we give a more detailed insight into the semantic image labelings which we compute and finally fuse into a semantic 3D point cloud.

In Fig. 4 input images and the corresponding semantically labeled images for all three datasets are illustrated. In Fig. 5 the semantically labeled input point clouds can be seen. One can observe that the point clouds are generally more consistently labeled due to the redundancy from multiple image labels for one 3D point. However, some artifacts at building edges arise, which can be explained by the depth maps used for visibility computation: We are using depth maps in half resolution and with a strong smoothness prior. Additionally, we are using

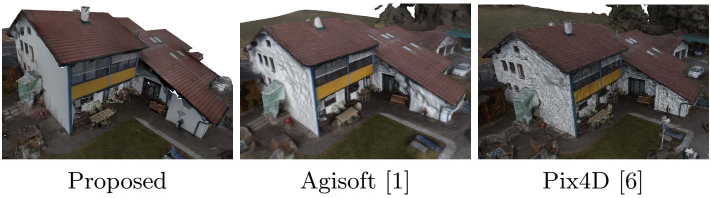

Proposed Agisoft [1] Pix4D [6]

Fig. 3. *Results from commercial reconstruction pipelines of the House dataset compared with the proposed approach.* Compared to the reconstruction of the proposed approach, Agisoft and Pix4D do not generate planar surfaces and straight edges, as they use no shape or semantic priors. The facades and roofs are noisy and scene are smoothed too much and become rounded (e.g., building edges, chimney).

nearest neighbor when back-projecting a 3D point into the depth maps. Hence, when 3D points are slightly noisy but still have visibility in the depth maps, it may happen that they get back-projected to the background and, hence, get assigned the label of the background (which might be, e.g., sky or vegetation).

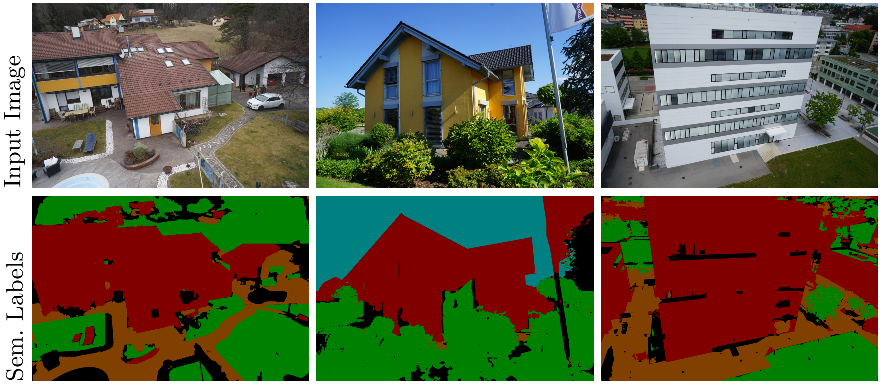


Fig. 4. *Semantically labeled input images* from the *House* (left), *Residential Area* (middle) and *Block Building* (right) datasets. The pixels are labeled as sky (blue), building (red), vegetation (green), road/pavement (brown) and clutter (black). The labeling is not always perfectly correct. However, due to the redundancy, the 3D points have less errors in labeling (see Fig. 5).

5 3D Ground Truth

In this section, we describe the ground truth for the *House* and the *Residential Area* dataset in more detail and discuss the errors of the compared 3D reconstructions.

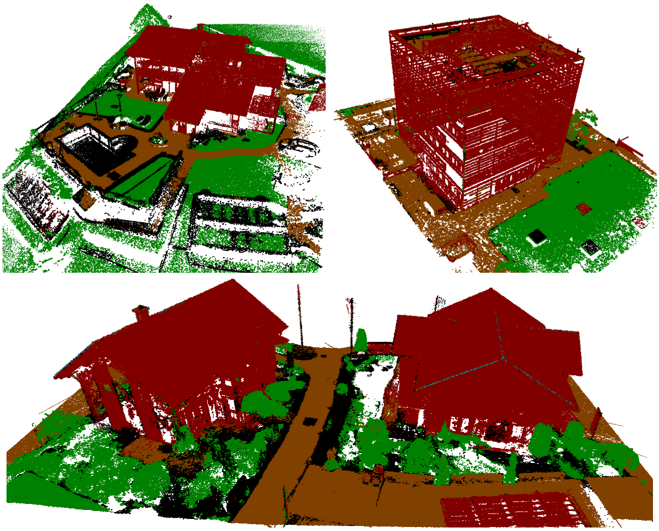


Fig. 5. *Semantically labeled point clouds* from the *House* (top left), *Block Building* (top right) and *Residential Area* (bottom) datasets. The pixels are labeled as sky (blue), building (red), vegetation (green), road/pavement (brown) and clutter (black). Due to the redundancy from multiple image labelings, the point cloud labeling is usually more consistent compared to the single image labelings.

In Fig. 7, the ground truth point clouds are depicted. One can observe, that it is not complete. Especially at parts of the buildings where the proposed approach should outperform others (facades, roofs), there are big holes in the ground truth. In Fig. 6, the error for every ground truth point is illustrated, mean and standard deviation of the errors are depicted in the main paper. It can be seen that the proposed approach has visually low errors on both datasets.

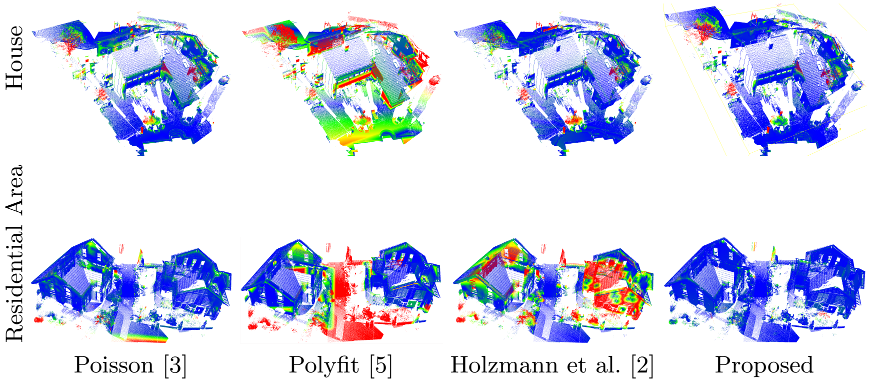
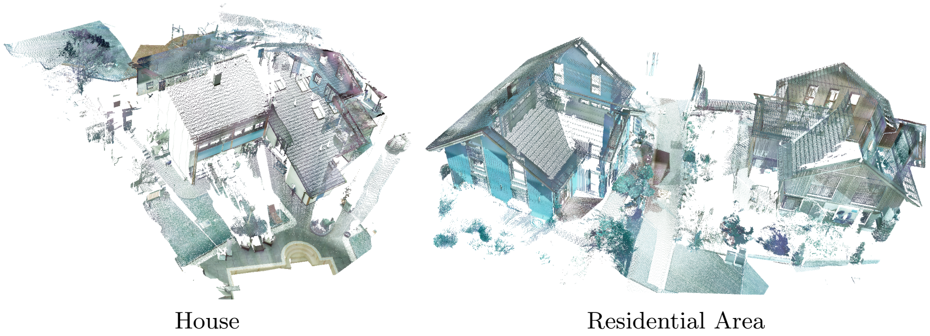


Fig. 6. *Ground truth point errors* measured as the minimum distance from ground truth points to the 3D surface mesh of the results in the main paper. Blue means low error and red means high error, where the maximum error is defined as $1m$. For the *House* dataset (top row), it can be observed that Holzmann et al. has a slightly lower error below the roof and at vegetation compared to Poisson meshing. Polyfit has a much higher error, as it does not succeed in reconstructing the building and the surrounding with an adequate accuracy. For the *Residential Area* dataset (bottom row), Holzmann et al. has the highest error, as it smooths out one building nearly completely and also has big errors at the second building. Also Poisson meshing has significantly higher errors compared to the proposed approach, especially at walls and roofs.



House

Residential Area

Fig. 7. *Ground truth point cloud* captured with a total station. Even though the ground truth was captured from several view points, it is not complete at all parts: E.g., big parts of the roof are missing at both datasets and the facade on the left side of the house is missing in the *House* dataset.

References

1. Agisoft PhotoScan: `http://www.agisoft.com/`
2. Holzmann, T., Oswald, M.R., Pollefeys, M., Fraundorfer, F., Bischof, H.: Plane-based surface regularization for urban 3d reconstruction. In: 28th British Machine Vision Conference. vol. 28 (9 2017)
3. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Eurographics Symposium on Geometry Processing (2006)
4. Labatut, P., Pons, J.P., Keriven, R.: Robust and efficient surface reconstruction from range data. Computer Graphics Forum pp. 2275–2290 (December 2009)
5. Nan, L., Wonka, P.: Polyfit: Polygonal surface reconstruction from point clouds. In: Proceedings International Conference on Computer Vision (2017)
6. Pix4DModel: `https://pix4d.com/`
7. Schnabel, R., Wahl, R., Klein, R.: Efficient ransac for point-cloud shape detection. Computer Graphics Forum **26**(2), 214–226 (Jun 2007)
8. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Proceedings IEEE Conference Computer Vision and Pattern Recognition (2008)