Semantically Informed Multiview Surface Refinement

Supplemental Material

Maroš Bláha[1] Mathias Rothermel[1] Martin R. Oswald[2] Torsten Sattler[2]
Audrey Richard[1] Jan D. Wegner[1] Marc Pollefeys[2,3] Konrad Schindler[1]
[1] Institute of Geodesy and Photogrammetry, ETH Zurich
[2] Department of Computer Science, ETH Zurich [3] Microsoft

1. Overview

The supplemental material is structured as follows: we first demonstrate the impact of the introduced energy terms (Section 2, 3). Therefore, the single energies are minimized individually for a test area and resulting surface models are analyzed qualitatively. In Section 4, we give details on the method used to predict the class-conditional probabilities and describe the utilized feature set. Section 5 presents additional qualitative evaluations of the generated semantic surface models to complement the evaluation section of the paper. In particular, we explain the key improvements achieved by our method in comparison to the input and baseline models. Finally, in order to clarify the precision gain of our method, we scale the results of the quantitative evaluation presented in the paper (Table 3) to pixel units.

2. Geometric Refinement - Impact of the Individual Energy Terms

To begin with, we demonstrate the impact of the individual geometric energy terms used in our approach. Thereby the single data, intra- and inter-class energies are minimized independently and the generated surface models are evaluated qualitatively. This experiment was performed on the Enschede A dataset comprising 15 nadir and oblique viewing angles and images, respectively. For each energy term we apply five iterations of gradient decent in a standalone fashion and analyze the resulting 3D models. As shown in Fig. 1(a,b), both data terms E_{photo} and E_{sem} contribute to recovering fine details such as windows on facades or thin rails on roofs. The intra-class regularization enables semantically adaptive surface smoothness, *e.g.*, stronger regularization for roofs compared to vegetation as highlighted in Fig. 1(c). Fig. 1(d) shows that the presented inter-class energy enforces smooth, anti-aliased transitions between two classes.

3. Label Refinement - Impact of the Individual Energy Terms

In this section, we illustrate the effects of the single energies used within the presented relabeling method. Column (a) of Fig. 2 shows an example label map (top), the input model (middle), and the semantically labeled surface (bottom) derived by re-projecting pixel-wise label likelihoods onto the surface, face-wise integration of the labels, and labeling the faces according to the class with the highest score. As expected, solely relying on the data term results in scattered labels. However, the data term also allows to recover fine details in some parts that were lost by the input method. Fig. 2(b) displays the surface labeling obtained by additionally applying the pairwise term E_{smooth}, leading to more a homogeneous labeling and better performance at class transitions. Additionally taking into account the surface normal dependent energy E_{geo} improves the correctness of the labeling, in particular for surface areas where normals are reliable (Fig. 2(c)). For the experiments we applied 40 iterations of loopy belief propagation [5].

4. Features for the Computation of Semantic Likelihood Images

The utilized input likelihoods are computed using a multiclass boosting classifier trained on a few manually labeled images [1]. In total, 94 features are extracted per pixel from the intensity images and the depthmaps. The depthmaps were generated from the multiview RGB images using semi-global matching [3, 6]. Fig. 3 shows a tuple of an intensity image, the corresponding depthmap and ground truth labels (*ground, roof, vegetation, facade*) at a glance.

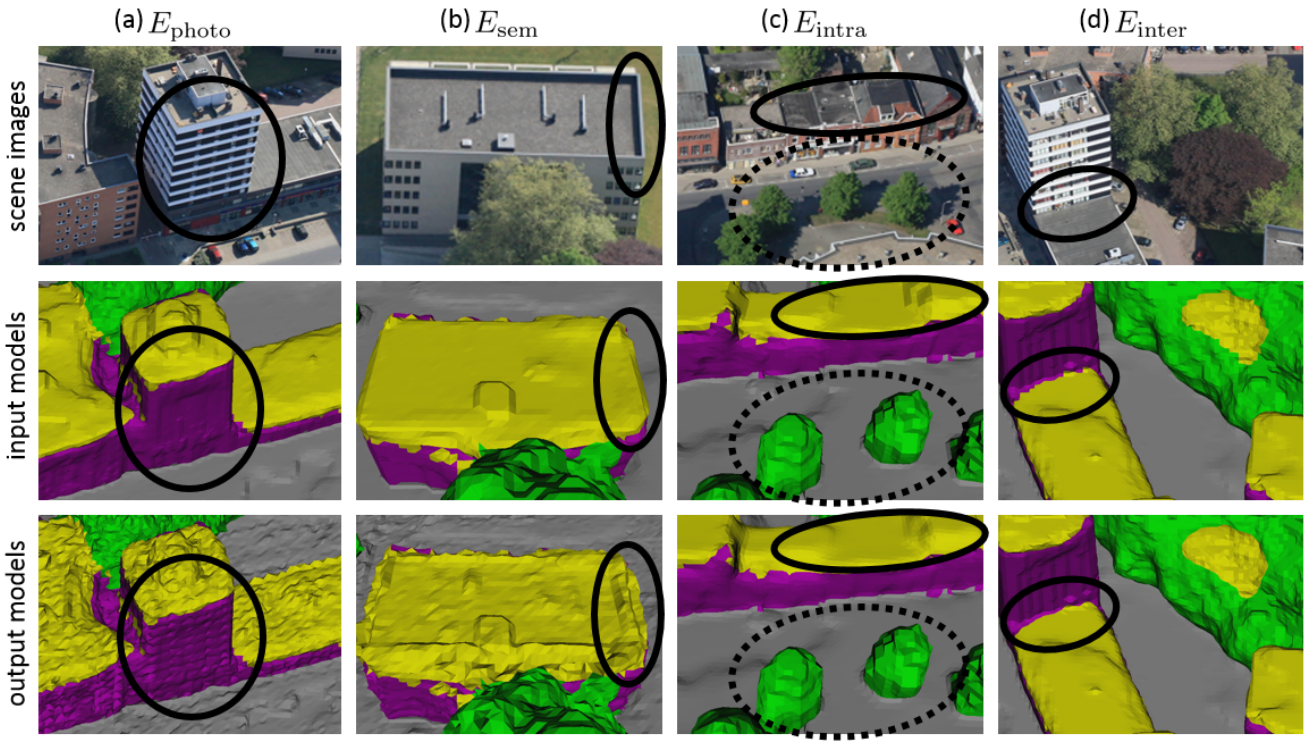


Figure 1: Illustrative support of our geometric update terms. *Left-to-right*: photometric and semantic data term, intra-class and inter-class smoothness term. *Top-down*: scene images, input and output models. Exemplary improvements are highlighted with solid black circles. The dashed circle for the intra-class smoothing highlights the vegetation as an unchanged scene part because the smoothing has mainly been increased for the *roof* class.

More precisely, the feature vector comprises the appearance of a 5 $\times$ 5 pixel neighborhood, from which 75 features are extracted. Additionally, 19 local geometry features are derived from the 3D point cloud based on the depthmaps [4]. These features take into account the 3D structure tensor (*i.e*. eigenvalues), the height, the local tangent plane, and the point distribution in a vertical column - see Tab. 1 for a detailed list of all geometric features. Fig. 4 (b)-(e) illustrates the per-pixel likelihoods of the respective semantic classes. These likelihoods serve as semantic input for our method.

5. Qualitative Evaluation of All Processed Data Sets

In this section, we provide additional qualitative evaluations highlighting the key improvements of our method in comparison to results obtained by the input [2] and baseline method (*i.e*. our own reimplementation of [7]). Fig. 5 and Fig. 6 illustrate the results of each method for six test scenes and underline the advantages of exploiting both modalities, geometry and semantics, during the surface refinement. Using semantics during geometric refinement enables us to treat different classes individually as assumptions about shape are object specific. Specifically, this enables us to recover fine structures, and, simultaneously apply adequate smoothing to roofs and facades (Fig. 5(1)); preserve the heavily regularized input for streets (Fig. 5(2)); and apply limited smoothing for high fidelity surface structures such as vegetation (Fig. 5(3)). As expected, a more homogeneous labeling can be achieved using MRF inference (Fig. 6(6)). Moreover, geometry is a powerful cue within the labeling process. Surfaces with vertical normals are more likely to be ground or roof than facades and vice versa, horizontal normals indicate building walls. This prior knowledge supports correctness of labeling as depicted by Fig. 6(4) and enables the reconstruction of fine semantic details Fig. 6(5).

6. Complementary Quantitative Evaluation

In order to clarify the improvement regarding accuracy of the presented method, we scale results presented in Table 3 in the paper to pixel units according to $\delta_{pix} = \frac{f}{d_{av}} * \delta_{obj}$. Thereby d_{av} is the average distance from the cameras to the scene, f is the focal length and δ_{obj} is the error in the object space. The corresponding values are given in Tab. 2. As discussed in the

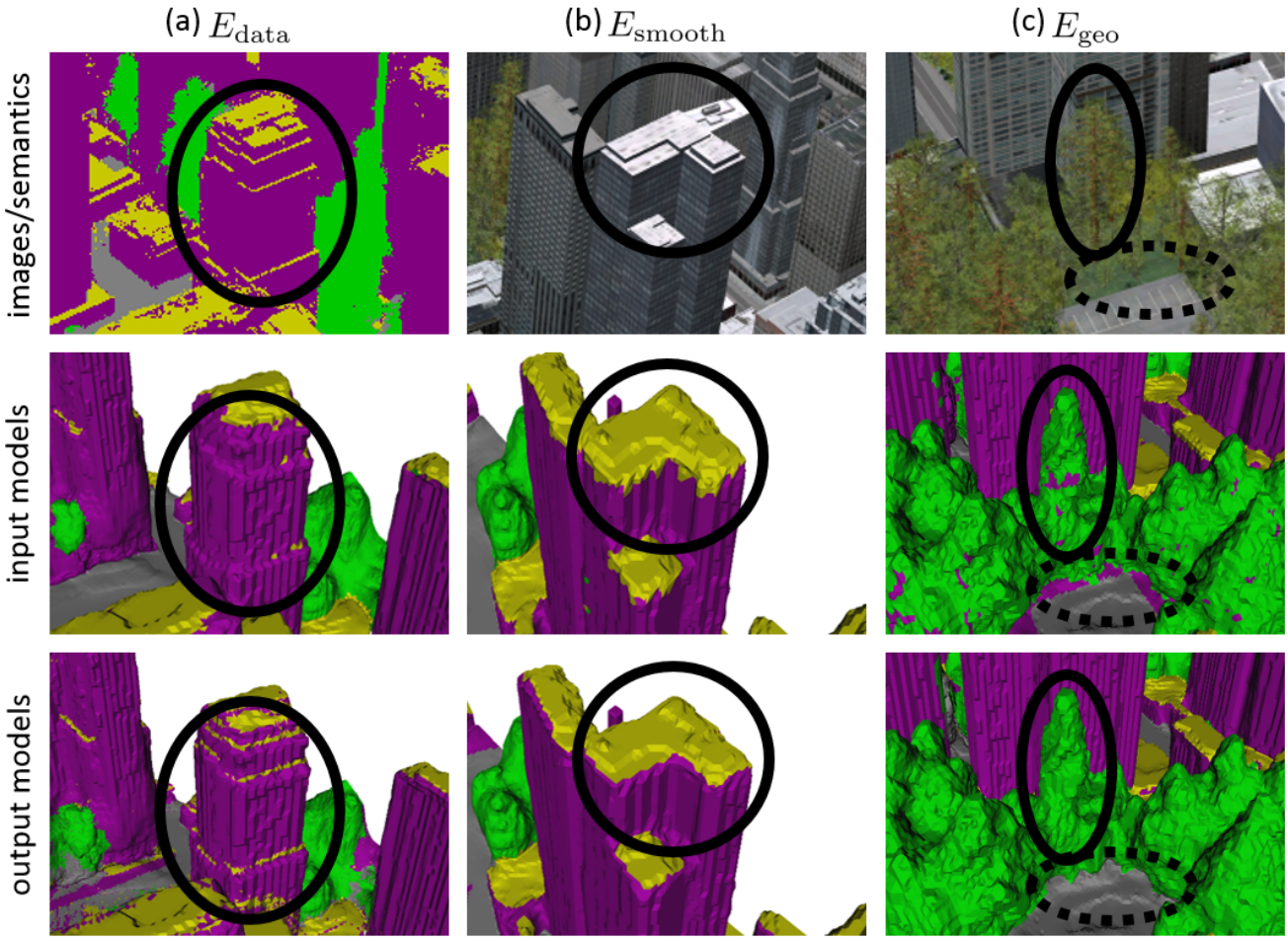


Figure 2: Illustrative support of our semantic update terms. *Left-to-right*: semantic likelihood data term, smoothing term, and geometric prior. *Top-down*: scene images/semantics, input and output models. Exemplary improvements are highlighted with black circles. The two circles (solid and dashed) in column (c) highlight two different areas where the labeling was improved by taking into account the surface normal dependent energy.

main paper, the use of additional semantic information leads to superior refinement performance on synthetic, as well as on real world data experiments.

References

[1] D. Benbouzid, R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. Kgl. Multiboost: a multi-purpose boosting package. *JMLR*, 2012.

[2] M. Bláha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. *CVPR 2016*.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] N. Chehata, L. Guo, and C. Mallet. Airborne lidar feature selection for urban classification using random forests. *IntArchPhRS*, 2009.

[5] B. J. Frey and D. J. MacKay. A Revolution: Belief Trees: Belief Propagation in Graphs With Cycles. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, 1998.

[6] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30(2), 2008.

[7] H. H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *PAMI*, 34(5), 2012.

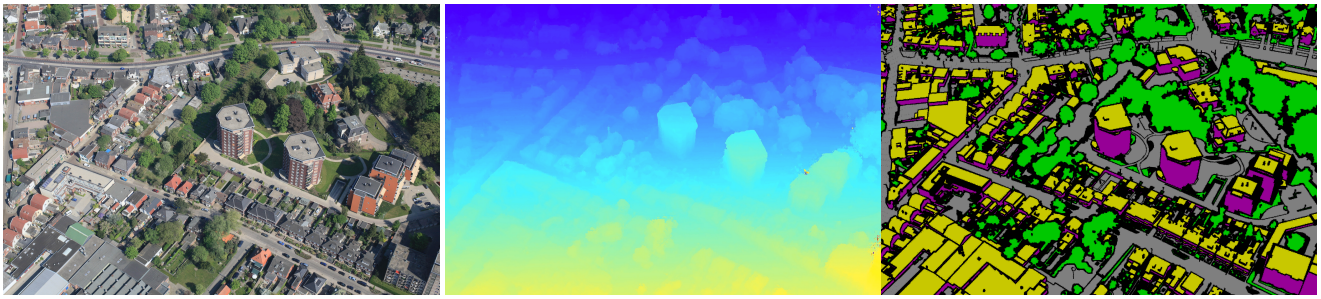
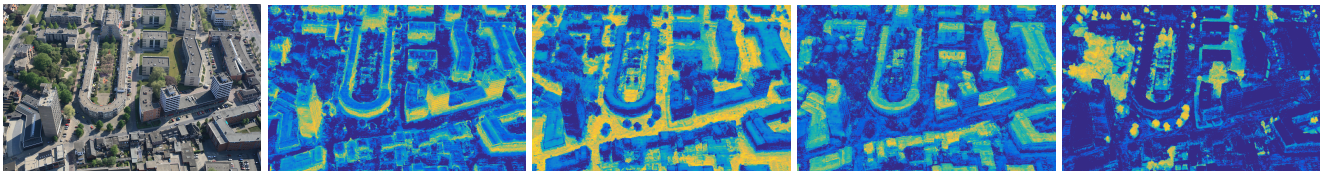


Figure 3: Input data used to compute the per-pixel semantic likelihoods. *Left-to-right*: intensity image, depthmap [3, 6], hand-labeled ground truth. The colors indicate *ground* (gray), *roof* (yellow), *vegetation* (green) and *facades* (purple).



(a) (b) (c) (d) (e)

Figure 4: (a) Intensity image and the resulting per-pixel likelihoods from [1] for the classes *facade* (b), *ground* (c), *roof* (d) and *vegetation* (e). A yellow color indicates high values, while blue color indicates low values.

Type	**Feature definition**
Height features	Height (z-component) Height variance
Eigenvalue features	Anisotropy $(\lambda_1 - \lambda_3)/\lambda_1$ Planarity $(\lambda_2 - \lambda_3)/\lambda_1$ Sphericity λ_3/λ_1 Linearity $(\lambda_1 - \lambda_2)/\lambda_1$
Local tangent plane features	Vertical component of plane normal Deviation angle of plane normal from vertical Variance of deviation angles Distance from point to local plane Variance of point-to-plane distances
Features based on histogram of signed z-differences to other points in a vertical column	# of bins above mean frequency # of bins below mean frequency Difference: # above bins - # below bins # of local frequency maxima Average distance between local maxima Sum of positive values Sum of negative values # of elements in zero-bin

Table 1: Geometric features [4] used for the generation of the per-pixel semantic likelihoods (in combination with the RGB values of the particular images).

Data set	**Modality**	**Performance Measure**	**Input [2]**	**Baseline [7]**	**Ours**
SynthCity3 A	Geometry	δ_{pix} [px]	1.25	1.06	**0.91**
		δ_{pix} [px]	1.45	1.22	**1.05**
SynthCity3 B	Geometry	δ_{pix} [px]	2.00	1.76	**1.48**
		δ_{pix} [px]	2.31	2.05	**1.72**

Table 2: Complementary quantitative evaluation of our method. Best performance is shown in bold.

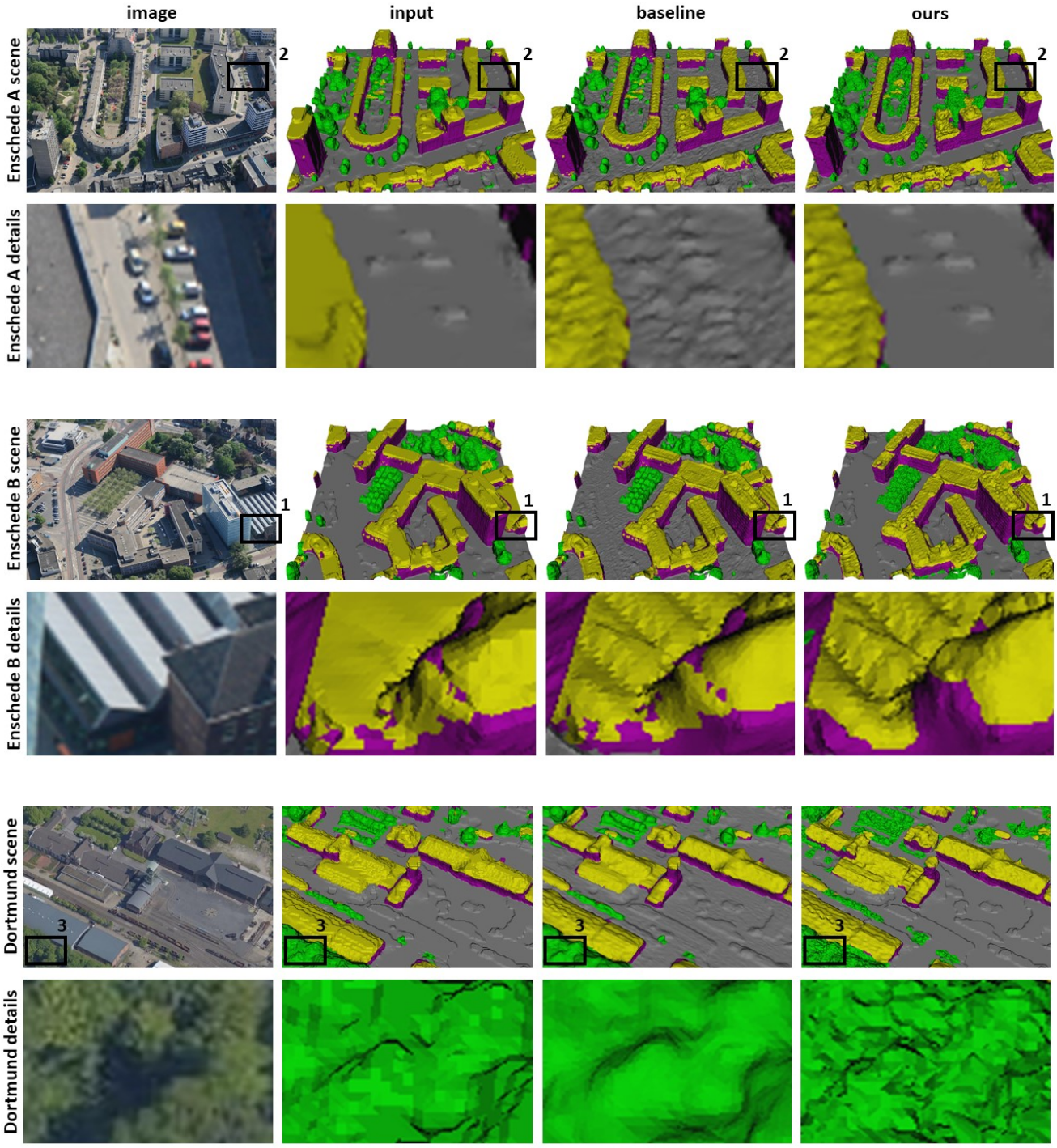


Figure 5: Qualitative evaluation of the proposed method for three data sets. *Left-to-right*: scene image, input model [2], baseline [7], proposed method. Notice the high scene fidelity, and, at the same time an adaptive, class-specific surface regularization, clean class transitions and less noisy semantics in our model. Exemplary improvements are highlighted with black rectangles and corresponding close-ups are visualized. The detailed descriptions are in section Section 5. Colors indicate *ground* (gray), *facade* (purple), *roof* (yellow), and *vegetation* (green).

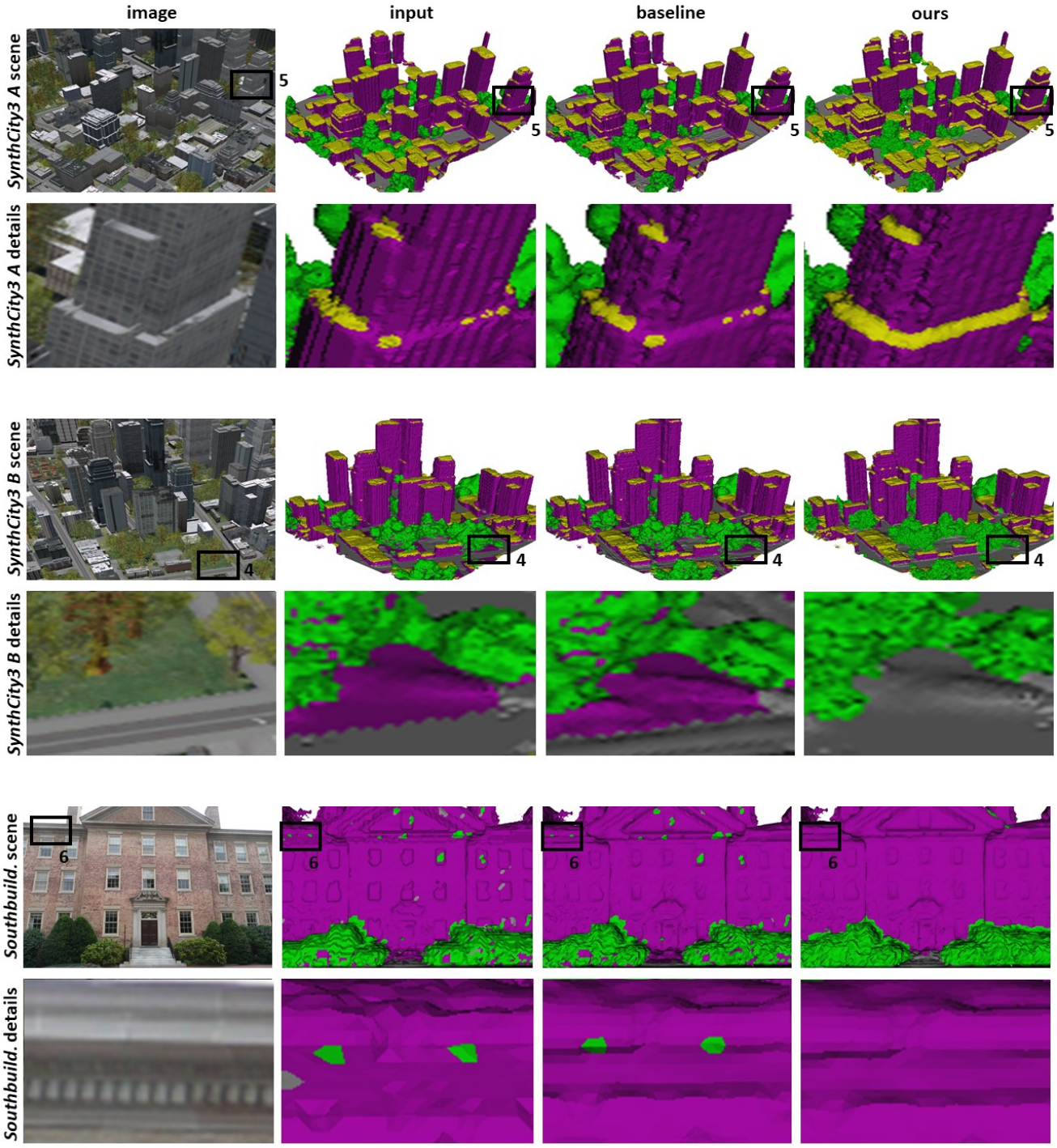


Figure 6: Qualitative evaluation of the proposed method for three data sets. *Left-to-right*: scene image, input model [2], baseline [7], proposed method. Notice the high scene fidelity, and, at the same time an adaptive, class-specific surface regularization, clean class transitions and less noisy semantics in our model. Exemplary improvements are highlighted with black rectangles and corresponding close-ups are visualized. The detailed descriptions are in section Section 5. Colors indicate *ground* (gray), *facade* (purple), *roof* (yellow), and *vegetation* (green).