CUMULATIVE ASSESSMENT FOR URBAN 3D MODELING

Shea Hagstrom[1], Hee Won Pak[1], Stephanie Ku[1], Sean Wang[1], Gregory Hager[2], and Myron Brown[1]

[1]JHU Applied Physics Laboratory   [2]The Johns Hopkins University

ABSTRACT

Urban 3D modeling from satellite images requires accurate semantic segmentation to delineate urban features, multiple view stereo for 3D reconstruction of surface heights, and 3D model fitting to produce compact models with accurate surface slopes. In this work, we present a cumulative assessment metric that succinctly captures error contributions from each of these components. We demonstrate our approach by providing challenging public datasets and extending two open source projects to provide an end-to-end 3D modeling baseline solution to stimulate further research and evaluation with a public leaderboard.

Index Terms— Semantic segmentation, multiple view stereo, urban 3D modeling, metrics, benchmarks

1. INTRODUCTION

Urban 3D models play an important role in a variety of applications such as city planning, tourism, navigation, and emergency management. Automated methods for producing these models from multiple satellite images are especially important for use cases involving coverage of large geographic areas or remote geographic regions inaccessible by airborne cameras or lidar.

The component challenges for urban 3D geometric modeling include semantic segmentation to delineate urban features, multiple view stereo (MVS) for 3D reconstruction of surface heights, and 3D model fitting to produce compact mesh models with accurate roof slopes. Comprehensive evaluation must account for error contributions from each to enforce model quality. For example, Rottensteiner et al. [1] assessed semantic, geometric, and topological error for 3D building reconstruction from airborne images. Similarly, Bosch et al. [2] assessed semantic, volumetric, and curvature error for building modeling from satellite images. Both derived reference values for evaluation from airborne lidar and assessed each metric separately. To the best of our knowledge, there has been no published work to define a holistic metric for end-to-end evaluation accounting for all error contributions with a single value.

Li et al. [3] and Leotta et al. [4] have recently proposed end-to-end pipelines for 3D building reconstruction from satellite images. The Danesfield pipeline [4] is open source
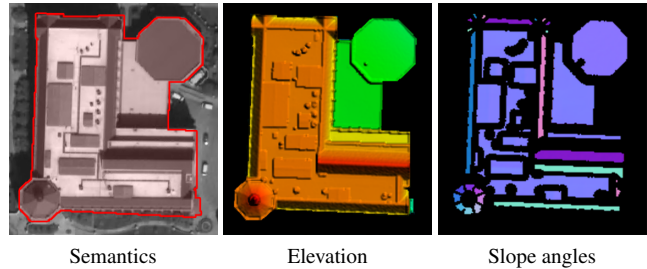

Semantics   Elevation   Slope angles

Fig. 1: Our metrics for cumulative assessment capture errors in semantic segmentation, multi-view stereo, and 3D model fitting. Image is courtesy DigitalGlobe.

except for a commercial MVS algorithm, while Li et al.'s pipeline [3] employs open source S2P [5] for MVS but is otherwise not public. We are not aware of any end-to-end pipeline that is entirely open source.

The ISPRS benchmark [1] provides public data for assessing urban 3D building reconstruction from airborne images. While public data is widely available individually for semantic segmentation [6, 7], MVS [8, 9], and semantic stereo [10, 11], we are not aware of any public datasets for evaluation of end-to-end building modeling pipelines with satellite images. Notably, reference data used for evaluations reported in [3] and [4] were not previously publicly available.

We begin to address these observed limits of prior work with the following contributions:

- We propose a cumulative assessment metric for urban 3D modeling from overhead images that explicitly captures error contributions from semantic segmentation, multiple view stereo, and 3D model fitting based on elevation and surface orientations (Fig. 1).
- To provide an end-to-end open source baseline solution for public evaluation, we extend the Danesfield pipeline [4] with open source VisSat MVS [12] to replace the commercial MVS component which is not publicly available.
- We provide challenging test sets and evaluate both Danesfield [4] and our extended baseline solution. To stimulate further research, we make our code and test sets publicly available [13] with an online leaderboard for evaluation.

2. METHODS

2.1. Cumulative Assessment

Standard 2D IOU evaluation using semantic labels is our starting point for building a single assessment metric incorporating both labeling and modeling. 3D mesh models are converted to rasterized DSMs for the purpose of evaluating geometry. Rasterization limits our evaluation to 2.5D simplifications, though in practice we prefer this for the simplicity in large-scale evaluation. Prior to evaluation, the test model is registered to our ground truth to remove any translation offset. IOU for 2D semantic labels is

$$TP_c = \sum_i^N TP_c^i \text{ where } TP_c^i = 1 \text{ iff } \hat{C}_i = 1 \text{ and } C_i = 1$$

$$FP_c = \sum_i^N FP_c^i \text{ where } FP_c^i = 1 \text{ iff } \hat{C}_i = 1 \text{ and } C_i = 0$$

$$FN_c = \sum_i^N FN_c^i \text{ where } FN_c^i = 1 \text{ iff } \hat{C}_i = 0 \text{ and } C_i = 1$$

$$IOU_c = \frac{TP_c}{TP_c + FN_c + FN_c} \tag{1}$$

where each pixel has a known binary label for category C in addition to predicted category $\hat{C}$, and we define true positive (TP), false positive (FP), and false negative (FN) binary counts from these. An additional binary label can be derived for error in elevation Z by accepting only values where the Z error is less than a defined threshold, as in Equation (2). The Z constraint helps ensure building model height and volume are correct, but it does not account for the correctness of localized slopes. For this purpose we introduce the slope error, defined as the angle between the reference and test model surface normals as in Equation (3). Local normals using Z values within a few meters of each pixel are computed from both for comparison using eigendecomposition. While this is well-defined for planar regions, at sharp edges or complex geometry the reference normals become more ambiguous. We use the ratios of the three eigenvalues to exclude these unstable reference normals and use only values passing the constraints

$$\lambda_3 / (\lambda_1 + \lambda_2 + \lambda_3) < 0.005$$
$$(\lambda_2 - \lambda_3) / \lambda_1 > 0.2$$

In our tests we found reference pointing normal errors to be under 3 degrees in 90% of samples, and Fig. 2 shows that errors are often even lower. The slope plot in Fig. 1 shows the regions with sufficient planarity to use in evaluation and excludes edges and small features.

The TP binary semantic values are pixel-wise multiplied by the corresponding binarized Z or angle errors, each successively reducing the TP counts. The full equation becomes

$$TP_z^i = 1 \text{ iff } |\hat{Z}_i - Z_i| < 1.0 \text{ meter} \tag{2}$$

$$TP_\theta^i = 1 \text{ iff } \cos^{-1}(\hat{\vec{u}_i} \cdot \vec{u}_i) < 5.0 \text{ degrees} \tag{3}$$

$$TP_m = \sum_i^N TP_c^i TP_z^i TP_\theta^i \tag{4}$$

$$IOU_m = \frac{TP_m}{TP_c + FN_c + FN_c} \tag{5}$$

We also define an additional IOU_z as an intermediate product not incorporating slope information. This gives a cumulative progression from IOU_c to IOU_z to IOU_m where each incorporates an additional constraint on the model accuracy.

In practice, both Z error and orientation error are not evaluable at every sample due to lack of valid reference values. In these cases we have chosen to assume that the respective measurements are correct in the IOU totals. Another approach would be to only evaluate models where all reference values are valid, however this would reduce the total area evaluated, in some cases significantly.

2.2. Public Baseline

To provide an open source baseline solution, we extended the Danesfield pipeline [4]. Danesfield performs semantic segmentation to delineate buildings, with inputs from multi-spectral image channels, normalized difference vegetation index (NDVI) values computed from those channels, and a normalized digital surface model (nDSM). The nDSM is derived by first segmenting ground and producing a digital terrain model (DTM) and then subtracting the DTM from the MVS digital surface model (DSM) to get height above ground. The MVS solution in Danesfield is commercial and not available publicly. MVS point clouds for segmented roofs are further segmented to produce roof shapes that are merged to produce building models.

We replaced the commercial MVS solution with open source VisSat [12], an adaptation of the state-of-the-art COLMAP MVS pipeline [14] for use with satellite images. Point clouds were initially noisier than those produced by the commercial MVS solution, so we added a point cloud bilateral filter [15] in post-processing. Details for reproducing our results are available in [13].

3. EXPERIMENTS

3.1. Dataset

Our test set consists of multiple DigitalGlobe WorldView-3 satellite images over areas of interest (AOIs) from Jacksonville (2 sq. km), San Diego (1 sq. km), and Omaha (1 sq. km) [13] (Fig. 3). Jacksonville and San Diego AOIs were also evaluated in [3, 4]. These AOIs are well-suited for semantic segmentation using available public training datasets, and each has dozens of images suitable for accurate MVS. AOIs also vary in modeling difficulty, with Jacksonville roofs being mostly large and flat and Omaha and San Diego having much more variety in roof complexity and pitch (Fig. 2). Omaha in particular has a variety of standard roof pitches in this
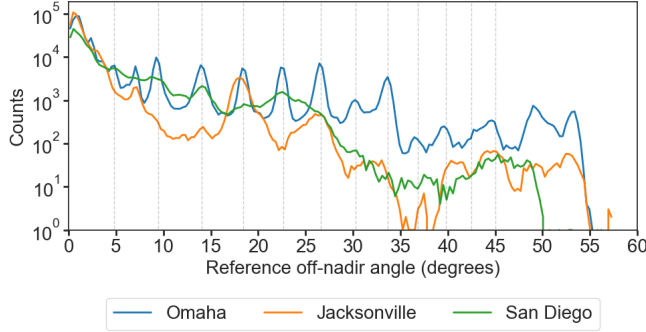
Fig. 2: Distribution histograms of reference roof angles from flat (0 degrees) are shown for our three test sites. Vertical lines represent roof pitches from 1/12 to 12/12 in 1/12 increments. Counts are plotted with a log scale.

region, made apparent by the many histogram peaks. Evaluation is conducted with manually edited building semantic labels with height and slope derived from airborne lidar.

3.2. Results

In Tab. 1, we compare metrics for Danesfield 3D reconstruction accuracy using open source VisSat MVS [12] and using commercial MVS from [4]. Cumulative metrics from section 2.1 are reported along with root mean square (RMS) error statistics for roof height and slope. For both pipelines, we assessed building models produced directly from MVS 3D points and also models produced after fitting those points to roof primitives. Improvement in IOU_m and RMS_θ values clearly indicate the efficacy of model fitting for achieving more accurate roof slopes. Upon inspection, we see that flat roofs are modeled very well and sloped roofs much less so, accounting for increased RMS_z values. To achieve a high IOU_m score, semantic segmentation, MVS, and 3D model fitting algorithms must all perform well. We hope that our work will encourage more attention to each, but particularly model fitting which has received less attention.

Our filtered VisSat MVS point clouds achieve consistently lower RMS_z and RMS_θ values than the commercial MVS solution from [4]; however, relative horizontal accuracy is inferior for the San Diego test site, leading to less accurate ground segmentation which contributes to error in building segmentation and thus lower IOU_c values. Building segmentation methods relying only on RGB images and trained using public data have been shown to score higher for IOU-2D on our test sites. For example, [3] reports 0.86 for Jacksonville and 0.82 for San Diego. We expect our baseline can be improved by replacing the semantic segmentation module with a high-performing solution from the SpaceNet competition [6].

Textured building models for the three AOIs are shown in Fig. 3 along with photos of the full modeled areas. An example of cumulative error statistics is shown in Fig. 4 for a building in Jacksonville. Semantic segmentation errors are
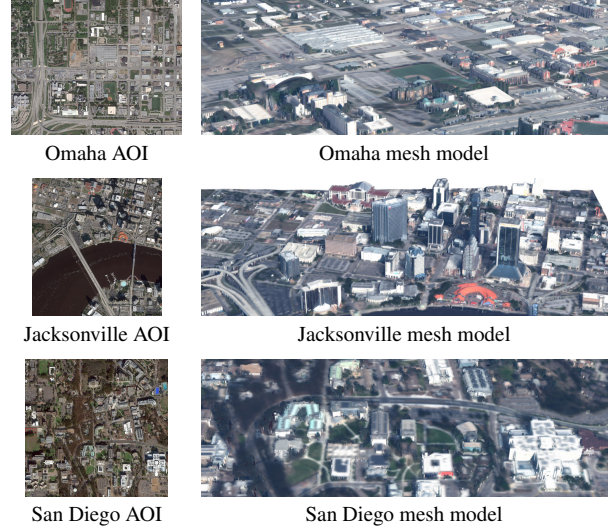

Fig. 3: AOI images and renders of textured building models from Danesfield / VisSat. Images are courtesy DigitalGlobe.

apparent, such as trees mislabeled as building shown in red. The MVS DSM has accurate height but incorrect slopes on the flat roof. This is much improved after successful model fitting; however, modeling with the baseline is less successful for complex and pitched roofs, as shown in Tab. 1. We hope that the cumulative metrics we have presented will inspire research to more robustly address these challenges.

	IOU_c	IOU_z	IOU_m	RMS_z	RMS_θ
JACKSONVILLE					
Commercial MVS	0.63	0.52	0.33	0.72	13.39
Commercial MVS + Modeling	0.63	0.51	0.44	1.39	6.99
VisSat MVS	0.67	**0.56**	0.41	**0.50**	8.80
VisSat MVS + Modeling	**0.68**	**0.56**	**0.50**	0.71	**3.53**
SAN DIEGO					
Commercial MVS	**0.67**	**0.53**	0.32	0.86	17.77
Commercial MVS + Modeling	**0.67**	0.52	**0.40**	1.88	15.14
VisSat MVS	0.58	0.47	0.32	**0.70**	15.89
VisSat MVS + Modeling	0.59	0.47	0.38	0.99	**10.32**
OMAHA					
VisSat MVS	**0.58**	0.52	0.34	**0.37**	9.32
VisSat MVS + Modeling	**0.58**	0.50	**0.40**	0.57	**8.35**

Table 1: IOU_m and RMS_θ metrics for Danesfield 3D reconstruction accuracy using open source VisSat MVS [12] and using commercial MVS from [4] show the value of roof modeling for improving 3D reconstruction.

4. CONCLUSION

We have presented cumulative assessment metrics for urban 3D modeling leading to a single score capturing error contributions from semantic segmentation, multiple view stereo, and 3D model fitting. To encourage 3D model fitting research, we produced challenging test data and combined two open
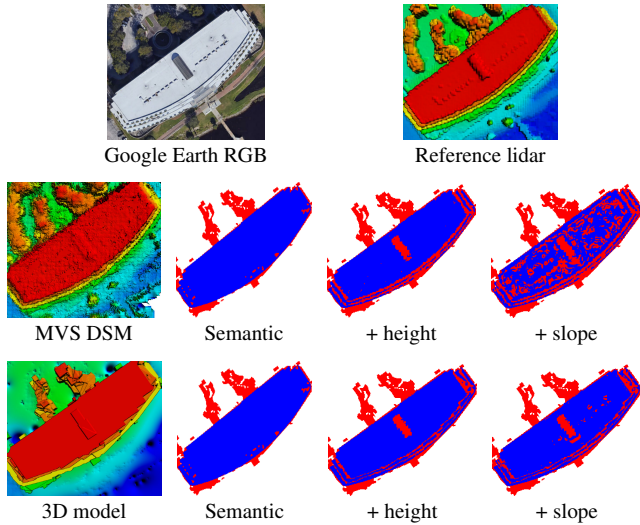
Fig. 4: Model heights from MVS and after 3D model fitting are compared. Cumulative evaluation constraints show the increasing failed areas (red) and passed areas (blue). Lidar and Google Earth RGB images are shown for reference.

source projects to provide an end-to-end 3D modeling baseline solution. We will make our code and data publicly available with an online leaderboard for evaluation.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Franz Rottensteiner, Gunho Sohn, Markus Gerke, Jan Dirk Wegner, Uwe Breitkopf, and Jaewook Jung, "Results of the isprs benchmark on urban object detection and 3d building reconstruction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 93, pp. 256 – 271, 2014.

[2] Marc Bosch, A. Leichtman, D. Chilcott, H. Goldberg, and Myron Z. Brown, "Metric evaluation pipeline for 3d modeling of urban scenes," *ISPRS Archives*, pp. 239– 246, 2017.

[3] X. Li, Y. Lin, J. Miller, A. Cheon, and W. Dixon, "Primitive-based 3d building modeling, sensor simulation, and estimation," in *IGARSS 2019*, 2019, pp. 5148– 5151.

[4] Matthew J. Leotta, Chengjiang Long, Bastien Jacquet, Matthieu Zins, Dan Lipsa, Jie Shan, Bo Xu, Zhixin Li, Xu Zhang, Shih-Fu Chang, Matthew Purri, Jia Xue, and Kristin Dana, "Urban semantic 3d reconstruction from multiview satellite imagery," in *CVPRW*, June 2019.

[5] G. Facciolo, C. De Franchis, and E. Meinhardt-Llopis, "Automatic 3d reconstruction from multi-date satellite images," in *CVPRW*, 2017, pp. 1542–1551.

[6] A. V. Etten, Dave Lindenbaum, and Todd M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *ArXiv*, vol. abs/1807.01232, 2018.

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska, "DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images," in *CVPRW*, 2018.

[8] M. Bosch, Z. Kurtz, S. Hagstrom, and M. Brown, "A multiple view stereo benchmark for satellite imagery," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2016.

[9] Sonali Patil, Bharath Comandur, Tanmay Prakash, and Avinash C. Kak, "A new stereo benchmarking dataset for satellite images," *CoRR*, vol. abs/1907.04404, 2019.

[10] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown, "Semantic Stereo for Incidental Satellite Images," in *WACV*, 2019.

[11] S. Kunwar, H. Chen, M. Lin, H. Zhang, P. Dangelo, D. Cerra, S. M. Azimi, M. Brown, G. Hager, N. Yokoya, R. Hansch, and B. Le Saux, "Large-scale semantic 3d reconstruction: Outcome of the 2019 ieee grss data fusion contest - part a," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020.

[12] Kai Zhang, Jin Sun, and Noah Snavely, "Leveraging Vision Reconstruction Pipelines for Satellite Imagery," in *ICCV Workshop on 3D Reconstruction in the Wild (3DRW)*, 2019.

[13] *CORE3D Public Baseline and Metrics, GitHub repository*, 2021, https://github.com/pubgeo/core3d-open.

[14] Johannes Lutz Schönberger and Jan-Michael Frahm, "Structure-from-Motion Revisited," in *CVPR*, 2016.

[15] Julie Digne and Carlo de Franchis, "The Bilateral Filter for Point Clouds," *Image Processing On Line*, vol. 7, pp. 278–287, 2017.