Augmenting Crowd-Sourced 3D Reconstructions using Semantic Detections: Supplementary Material

True Price[1] Johannes L. Schönberger[2] Zhen Wei[1] Marc Pollefeys[2,3] Jan-Michael Frahm[1]
[1]Department of Computer Science, UNC Chapel Hill [2]Department of Computer Science, ETH Zürich [3]Microsoft
{jtprice,zhenni,jmf}@cs.unc.edu {jsch,pomarc}@inf.ethz.ch

In this supplementary material, we present additional scale estimation results and overhead visualizations for datasets from Wilson and Snavely [2] and Heinly *et al*. [1]; we also provide ablative analyses of the various parts of our approach. See also our supplementary video, which contains flyovers of our reconstructed crowds and ground surfaces for the four larger datasets analyzed in our paper. This video contains comparative visualizations of 1) the representative subset of individuals selected by our set cover formulation, 2) all people and photographers placed by our method, and 3) the original static reconstructions from multi-view stereo that lack people and ground surfaces. Our visualizations clearly show the benefits of recovering people and ground in these otherwise "lifeless" and incomplete reconstructions.

1. Results on Additional Datasets

Here, we present additional quantitative results of the scene scale estimates produced by our method (Table 1). These results cover 15 scenes in addition to those presented in our paper: three from [2] (Tower of London, Trafalgar Square, and Union Square Park), and 12 from [1]. For each dataset, we also provide a top-down view of our person placements based on the gravity direction estimated by our method, and we show a comparative aerial image from Google Earth. As in the paper, green dots show the placement of detected individuals, red dots show locations for photographers, and black dots show static scene structure.

In general, our placements for detected people into the scene reflect the actual structures where people walk, particularly along sidewalks. Places where people do not walk (*e.g*., the fountains in Trafalgar Square) contain low densities of (likely mis-detected) people. The accurate scale estimates presented in the paper and above provide additional evidence as to the correctness of these placements. We also note that there were failure cases on other scenes, such as the Statue of Liberty (not shown), that were primarily caused by a large number of false person detections on human-like statues. These false detections are also visible in the water of the Trevi Fountain, below; however, the scene conditions in that case did not appear to negatively influence the result. We also empirically find that our method's accuracy is generally higher in scenes having 1) a larger number of person detections and 2) more complete static reconstructions obtained via Structure-from-Motion. The former condition provides greater support for approximate semantic triangulation, while the latter is important for enforcing visibility constraints, which are helpful in avoiding under-estimation of the length of one unit in the reconstruction space.

Scene	**Error**	n_p	n_c
Brandenburg Gate	-7.2%	5115	1131
British Museum	+0.3%	2925	507
Buckingham Palace	-5.9%	4972	1257
Hōzōmon Temple, Tokyo	-1.5%	1768	230
Lincoln Memorial	+6.5%	875	183
Palace of Westminster	-8.8%	331	496
Pike Place Market, Seattle	+8.5%	1081	312
Sacré Cœur, Paris	-0.3%	1705	782
Taj Mahal	-1.1%	395	805
Tōdai-ji Temple, Nara	-2.1%	2419	733
Tower Bridge, London	-2.6%	213	125
Tower of London	-4.7%	551	381
Trafalgar Square	+3.2%	13306	4328
Trevi Fountain	-3.3%	4934	2343
Union Square Park, NYC	-4.5%	2833	1023

Table 1. Quantitative results on our method for scale and placement. "% Error" gives the amount that we over/under-estimated the distance of one unit in the reconstruction. n_p and n_c show the number of placed detected people and photographers, respectively, recovered by our method.

2. Ablative Analysis

We provide additional analysis on the various parts of our reconstruction pipeline. Our algorithm has two general stages: scale voting and scale refinement. The scale voting stage serves to initialize the subsequent refinement. Here, we demonstrate that both stages are necessary to produce a satisfactory result, and we also show how different parame-

ter selections affect the end result in both stages.

2.1. Visibility Constraint During Voting

We first analyze the effect of removing the visibility constraint (Eq. (7)) during our scale voting procedure. The visibility constraint is necessary at this stage, but using the constraint alone is not sufficient to obtain the scene scale. Fig. 1 shows the effect of turning off the constraint for our Campitelli model. Because the (model-space) neighborhood radius in Eq. (5) generally grows faster w.r.t scale than pairwise person distances, using the neighborhood term alone will result in artificially high overlap at larger scales. The visibility constraint is thus important to rule out impossible person placements.

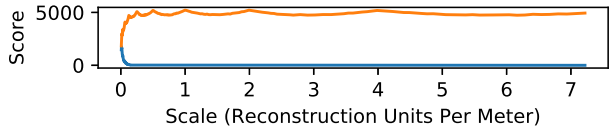


Figure 1. Our scale voting scheme with (blue) and without (orange) the visibility constraint. The ground-truth scale is near 0.01 reconstruction units per meter.

Fig. 2 demonstrates that the visibility constraint alone is not sufficient for determining the scene scale. For each detection in the Campitelli model, we compute the ratio of our estimated neck distance $s||N_i||$ to the visibility threshold $v_i(s)$ (c.f. Eq. (7)) for the ground-truth scene scale, and for $\pm 10\%$ and $\pm 20\%$ of this scale. We sort these ratios across all individuals and plot them. At the correct scale, individuals adjacent to static structures will have a ratio of $\sim$1. We observe that false detections and mis-estimations of the neck distance (having ratios much greater than one) make this condition ambiguous. Our approximate triangulation approach is thus necessary to obtain an initial scale.

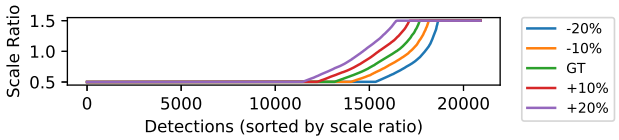


Figure 2. Ratio of our estimated neck distance $s||N_i||$ to the visibility threshold $v_i(s)$ for the ground-truth scale (GT), and for larger/smaller scales. Values are sorted and clipped to $[0.5, 1.5]$.

2.2. Effect of Scale Refinement Terms in Eq. (14)

There are three optimization terms in our scale refinement stage: a height prior, a local planarity penalty, and a visibility constraint. Our algorithm requires the local planarity term – without it, the optimal solution is to set the scale to an infinitesimal positive value (maximizing Eq. (13)) and each h_i to the most probable height. Table 2 shows our estimated scales with the height and visibility terms removed. The effect of the height prior varies between datasets, but we generally find better scale estimates when the constraint is included. The visibility constraint is intended for scenes with fewer individuals, to help prevent scale over-estimation caused by fewer well-supported neighborhoods.

2.3. Effect of Parameters during Refinement

To investigate the sensitivity of our algorithm to parameter changes, Table 2 further shows results after modifying the four major tunable parameters of Section 3.3 (photographer camera height β_c, "overshooting" threshold τ_o, planarity penalty λ, and the xz neighbor threshold) by $\pm 10\%$. The relative scale differences are generally small, and we observe only minor changes in the estimated 3D positions of the detected individuals.

2.4. Comparing Scale Voting and Scale Refinement

Finally, the 3rd and 4th columns of Table 2 show the scale improvement of our refinement stage vs. our initial voting. For many datasets, the refined scale estimate is closer to the ground truth. Since the local planarity term is the driving factor in our refinement step, this result supports the notion that the person placement (including the initial 3D triangulation) is an important component of our approach.

Scene	GT	Initial	Final	No Height	No Vis.	-10%	+10%
Cornell Quad	0.0269	0.0259	0.0280	0.0278	0.0294	0.0282	0.0272
Dubrovnik	0.0200	0.0183	0.0200	0.0199	0.0195	0.0197	0.0198
Pantheon	0.0913	0.0799	0.0873	0.0912	0.0877	0.0877	0.0874
Campitelli	0.0104	0.0097	0.0102	0.0102	0.0103	0.0102	0.0102
San Marco	0.0379	0.0336	0.0380	0.0375	0.0367	0.0383	0.0385
Alamo	0.1350	0.1253	0.1346	0.1323	0.1363	0.1320	0.1351
NYC Library	0.1437	0.1262	0.1418	0.1553	0.1442	0.1429	0.1403
Piccadilly	0.1216	0.1263	0.1290	0.1442	0.1329	0.1289	0.1275
Brandenburg Gate	0.1266	0.1287	0.1365	0.1433	0.1369	0.1356	0.1356
British Museum	0.3913	0.2793	0.3900	0.3434	0.4014	0.3923	0.3877
Buckingham Palace	0.0629	0.0604	0.0668	0.0776	0.0663	0.0662	0.0658
Hōzōmon Temple	0.5651	0.5070	0.5739	0.4642	0.5941	0.5797	0.5689
Lincoln Memorial	0.1161	0.1086	0.1090	0.1217	0.1093	0.1100	0.1080
Palace of Westmin.	0.0259	0.0280	0.0284	0.0298	0.0287	0.0289	0.0296
Pike Place Market	0.1840	0.1314	0.1696	0.1462	0.1754	0.1678	0.1704
Sacré Cœur	0.0507	0.0477	0.0509	0.0512	0.0503	0.0499	0.0502
Taj Mahal	0.0475	0.0420	0.0481	0.0488	0.0497	0.0491	0.0475
Tōdai-ji Temple	0.1340	0.1251	0.1369	0.1563	0.1380	0.1369	0.1354
Tower Bridge	0.2166	0.2391	0.2223	0.2391	0.2238	0.2244	0.2205
Tower of London	0.0484	0.0479	0.0507	0.0497	0.0517	0.0513	0.0498
Trafalgar Square	0.0700	0.0628	0.0678	0.0679	0.0673	0.0700	0.0671
Trevi Fountain	0.3179	0.2538	0.3288	0.3571	0.3278	0.3335	0.3213
Union Square	0.1380	0.1276	0.1430	0.1568	0.1427	0.1422	0.1416

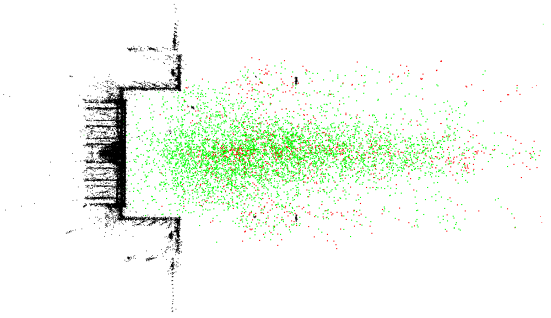
Table 2. GT: Ground-truth scene scales (reconstruction units per meter). Initial/Final: Estimates from our voting and refinement stages. No Height/Vis.: Height/visibility terms removed. ± 10: With modified parameters. Red cells: Results where the estimated length of one unit in the reconstruction was incorrect by $>10\%$.

References

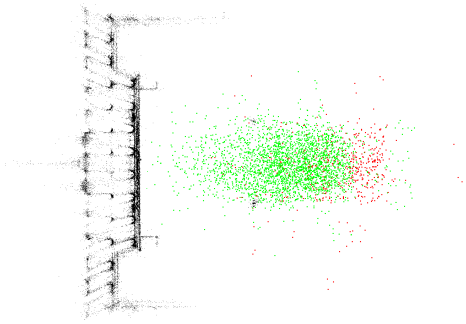
[1] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[2] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *European Conference on Computer Vision (ECCV)*, 2014. 1

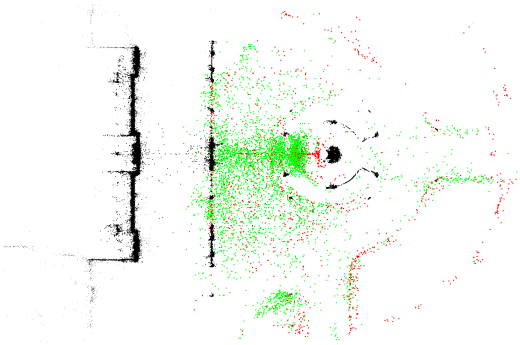
Brandenburg Gate



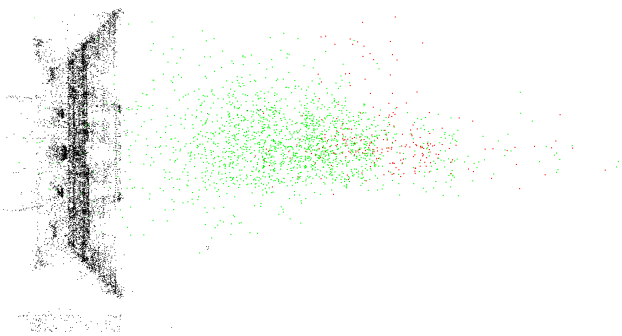
British Museum



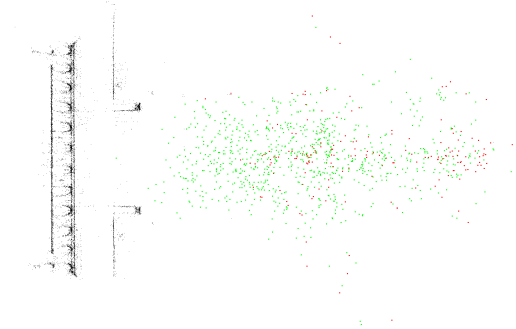
Buckingham Palace



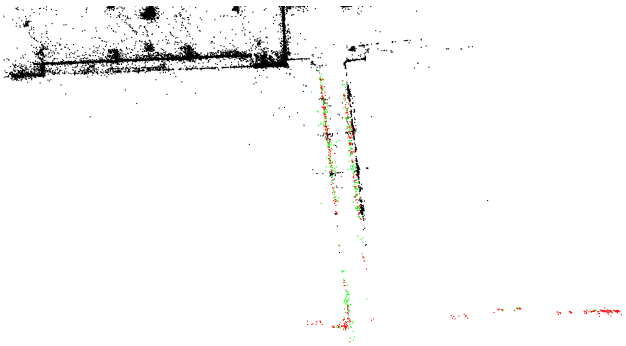
Hōzōmon Temple, Tokyo



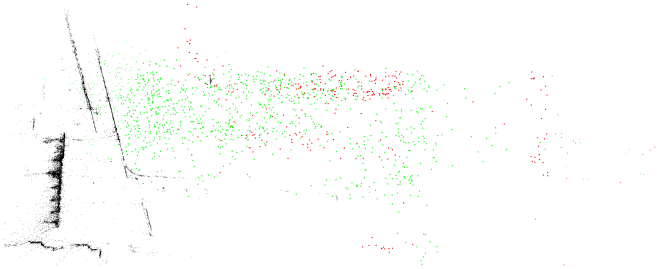
Lincoln Memorial



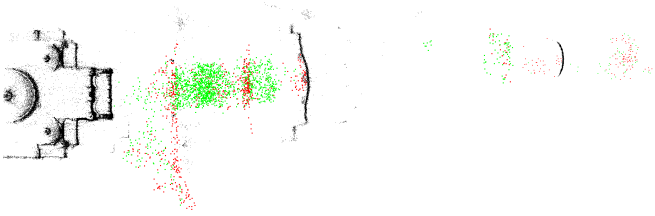
Palace of Westminster

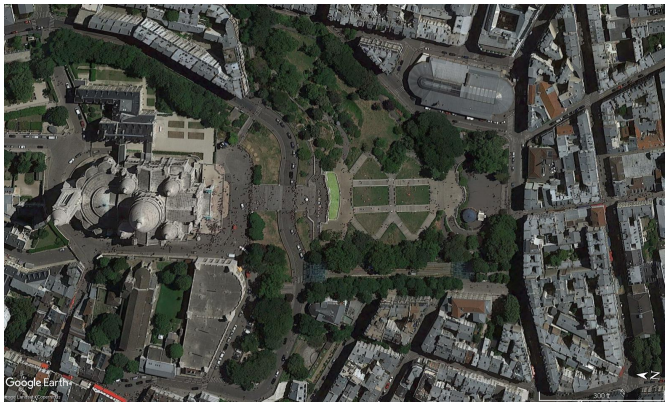


Pike Place Market, Seattle

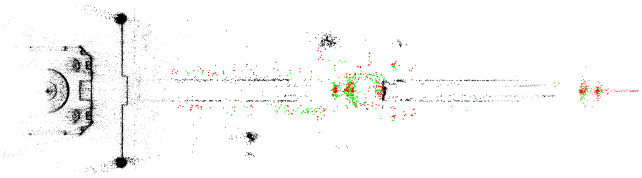


Sacré Cœur, Paris

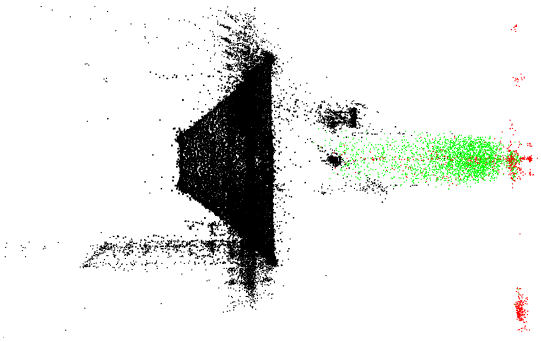




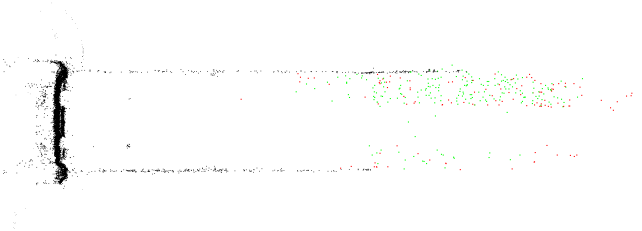
Taj Mahal

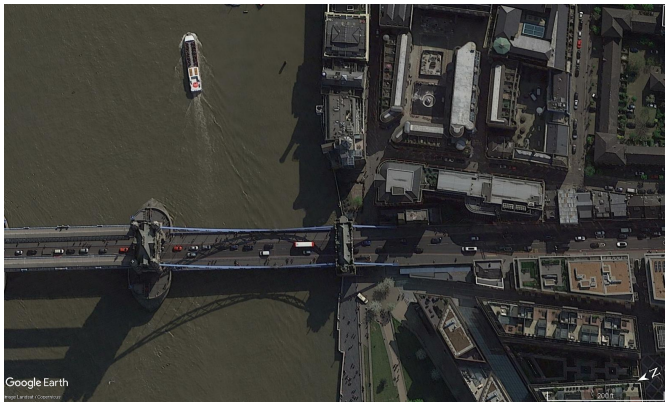


Tōdai-ji Temple, Nara

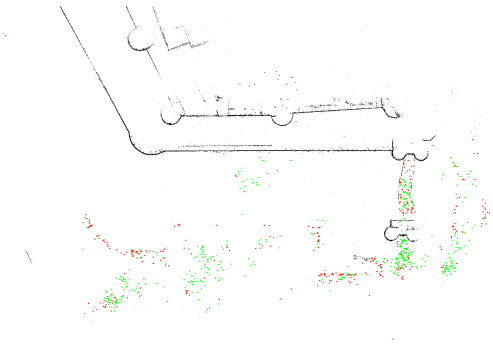


Tower Bridge, London

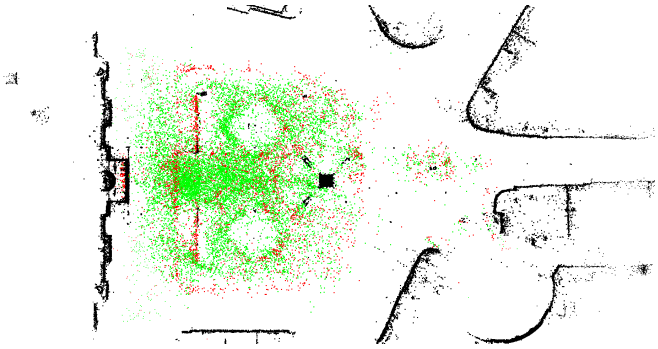




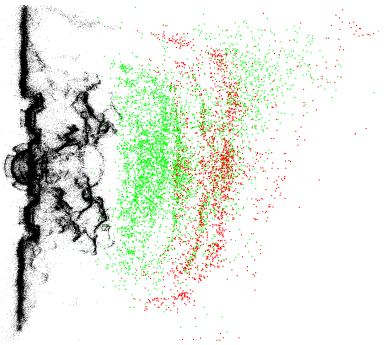
Tower of London



Trafalgar Square



Trevi Fountain



Union Square Park, NYC