Supplementary Material

Xiaohan Fei and Stefano Soatto

UCLA Vision Lab
University of California Los Angeles

1 Statistics of VISMA dataset

Upon submission, the VISMA dataset we have constructed contains 8 short sequences of several office scenes to allow cross-modality pseudo-groundtruth validation with an RGB-D sensor. The statistics of each sequence are as follows:

Table 1: *VISMA statistics.*

Sequence	clutter1	clutter2	occlusion1	occlusion2	meeting	swivel	lateral	double
#frames	951	701	936	827	1264	1290	1092	2017
traj. length(m)	12	16	8.9	10.5	54	10	13	18

We plan to augment the VISMA dataset with more sequences including both indoor and outdoor scenarios with deformable and dynamic objects in the future.

2 More qualitative results

Fig. 1 shows more qualitative results on outdoor sequences as well as results on the 4 VISMA sequences which have only coarse annotations.

3 Computational cost

In this section, we give a breakdown of the computational cost of each module of our system. Visual-inertial SLAM runs at $\sim$ 300Hz. Edge extraction runs at $\sim$ 300Hz. Faster R-CNN runs at $\sim$ 10Hz in both proposal generation and hypothesis scoring mode. The bottleneck is the naive implementation of our rendering pipeline in the prediction step: Rendering contour maps of 1K particles takes $\sim$ 300ms. Typically a budget of 500 particles is allocated to each object in the scene to achieve reliable estimation. Once the likelihood terms are gathered, overhead to update the posterior is negligible. All the timings are done on 640×480 imagery and a laptop with a GTX1080 GPU, an i7 CPU @ 2.7GHz and 32GB RAM. We expect a reduction in computational time through more advanced rendering techniques and parallel processing of particles.

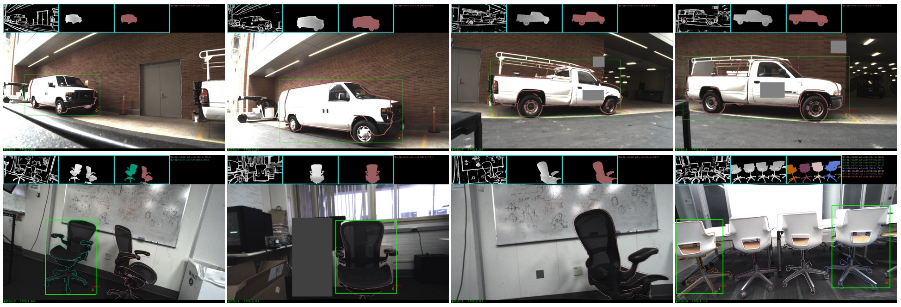

Fig. 1: *Live view* (best in color at 6×) of outdoor sequences (top) and 4 VISMA seq. not used for evaluation (bottom). Top inset shows (L to R): edge map, Z-buffer, projection masks, debug info; bottom shows input RGB with predicted mean object boundary and CNN detection. Though we use generic truck models from ShapeNet, pose estimates are robust to shape variations.

4 Details on one-dimensional search

Edge similarity h takes the hypothesized edge map $\widehat{E}$ and the one extracted by the network E, search along the normal of each edge pixel of $\widehat{E}$ within 80 pixels; if a corresponding edge pixel is found in E, the matched pair is put in set M and distance traveled d is recorded; then $h(\widehat{E}, E) = \frac{|M|/|\widehat{E}|}{\bar{d}+\epsilon}$, where the numerator is matching ratio, $\bar{d}$ is the average matching distance and $\epsilon = 10^{-4}$. For the coefficients of the two likelihood terms, we found that $\alpha = 50$, $\beta = 200$ work well in most cases via cross validation.

5 Video

See the attached MP4 file `video_full_533.mp4` for demos and qualitative results on both VISMA (first part) and SceneNN dataset (second part). `H.254` codec is used to generate the video.

For VISMA dataset, we show the point cloud of the environment, trajectory of the sensor platform and reconstructed objects in the scene with precise shape and pose – all inferred by our visual-inertial object detection and mapping system, causally and over time

For SceneNN dataset, due to the lack of inertial signals, we rely on the ground truth trajectory from RGB-D SLAM provided by SceneNN and only demonstrate the object detection and mapping part of our system. Thus no point cloud is shown. Also the incomplete models and jittery trajectories are due to the imperfect ground truth provided by SceneNN.