Supplementary Material for "Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data"

Michael Danielczuk[1], Matthew Matl[1], Saurabh Gupta[1],
Andrew Li[1], Andrew Lee[1], Jeffrey Mahler[1], Ken Goldberg[1,2]

This document describes supplementary experiments for the ICRA 2019 submission "Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data."

I. WISDOM Dataset Statistics

The real dataset has 3849 total instances with an average of 4.8 object instances per image, fewer than the 7.7 instances per image in the Common Objects in Context (COCO) dataset and the 6.5 instances per image in WISDOM-Sim, but more instances per image than both ImageNet and PASCAL VOC (3.0 and 2.3, respectively). Additionally, it has many more instances that are close to, overlapping with, or occluding other instances, thus making it a more representative dataset for tasks such as bin picking in cluttered environments. Since it is designed for manipulation tasks, most of the objects are much smaller in area than in the COCO dataset, which aims to more evenly distribute instance areas. In the WISDOM dataset, instances take up only 2.28% of total image area in the simulated images, 1.60% of the total image area on average for the high-res images, and 0.94% of the total image area for the low-res images. Figure 1 compares the distributions of these metrics to the COCO dataset.

II. Precision-Recall Evaluation

We performed additional experiments to analyze the precision-recall performance of SD Mask R-CNN along with the baseline methods for category-agnostic instance segmentation on RGB-D images: RGB object proposals [1, 2], cluster-based geometric segmentation methods from PCL [3], and Mask R-CNN fine-tuned for instance segmentation from the WISDOM-Real dataset. We also include a variant of SD Mask R-CNN fine-tuned on real depth images from WISDOM-Real. We evaluate performance using the widely-used COCO instance segmentation benchmarks [4].

The RGB object proposal baselines were based on two algorithms: Geodesic Object Proposals (GOP) [2] and Multi-scale Combinatorial Grouping (MCG) [1]. GOP identifies critical level sets in signed geodesic distance transforms of the original color images and generates object proposal masks based on these [2]. MCG employs combinatorial grouping of multi-scale segmentation masks and ranks object proposals in the image. For each of these methods, we take the top 100 detections. We then remove masks where less than half of the area of the mask overlaps with the foreground segmask of the image and apply non-max suppression with a threshold of 0.5 Intersection-over-Union (IoU).

Method	AP	AR
Euclidean Clustering	0.161	0.252
Region Growing	0.172	0.274
SD Mask R-CNN	***0.664***	***0.748***

TABLE I: Average precision and average recall (as defined by COCO benchmarks) on the WISDOM-Sim dataset for the PCL baselines SD Mask R-CNN.

Figure 2 shows precision-recall curves on three test datasets: 2000 images from the WISDOM-Sim validation set and 300 test images each from the Primesense and Phoxi cameras. Our learning-based method produces a ranked list of regions, that can be operated at different operating points. Not only does SD Mask-RCNN achieve higher precision at the same operating point than PCL, it is able to achieve a much higher overall recall without any compromise in precision at the cost of only a handful of extra regions.

We also evaluate the performance of SD Mask R-CNN and several baseline on the WISDOM-Sim test set.

III. Details of Instance-Specific Grasping Experiment

To evaluate the effectiveness of SD Mask R-CNN in a robotics task, we performed experiments in which we used segmentation as the first phase of an instance-specific grasping pipeline. In the experiment, an ABB YuMi robot was presented a pile of ten known objects in a bin and instructed to grasp one specific target object from the bin using a suction gripper. An attempt was considered successful if the robot lifted the target object out of the bin and successfully transported the object to a receptacle.

One approach to this problem is to collect real images of the items piled in the bin, labeling object masks in each image, and using that data to train or fine-tune a deep neural network for object classification and segmentation [5, 6]. However, that data collection process is time consuming and must be re-performed for new object sets, and training and fine-tuning a Mask R-CNN can take some time. Instead, our experimental pipeline uses a class-agnostic instance

[1]Department of Electrical Engineering and Computer Science
[2]Department of Industrial Engineering and Operations Research
[1−2]The AUTOLAB at UC Berkeley; Berkeley, CA 94720, USA
{mdanielczuk, mmatl, sgupta, andrewyli, andrew_lee, jmahler, goldberg}@berkeley.edu

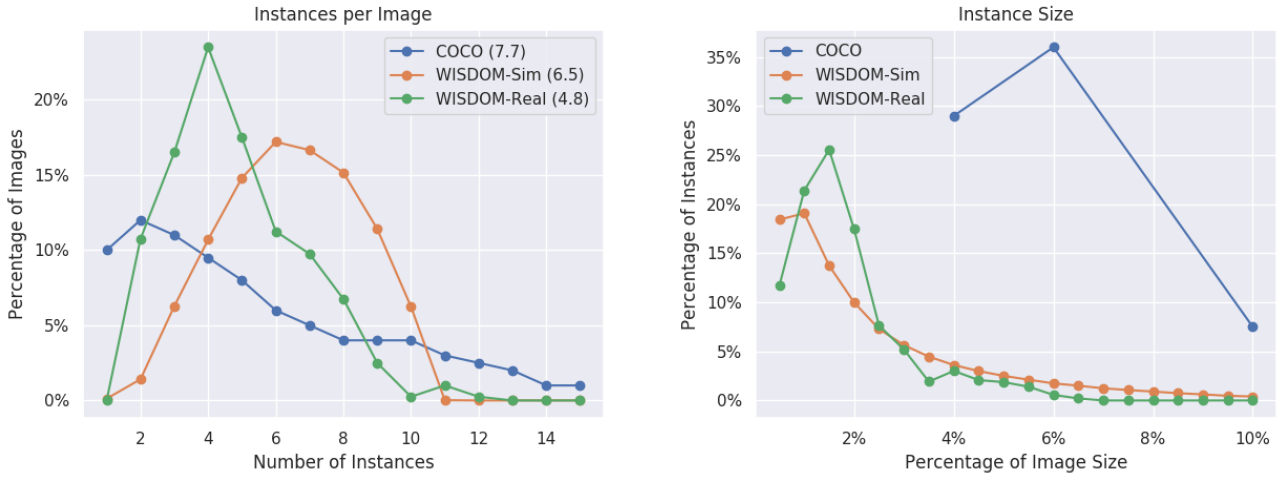

Fig. 1: Distributions of the instances per image and instance size for the WISDOM dataset, with comparisons to the COCO dataset. Average number of instances are listed in parentheses next to each dataset. The number of instances and relative object size make this dataset more applicable to manipulation tasks.

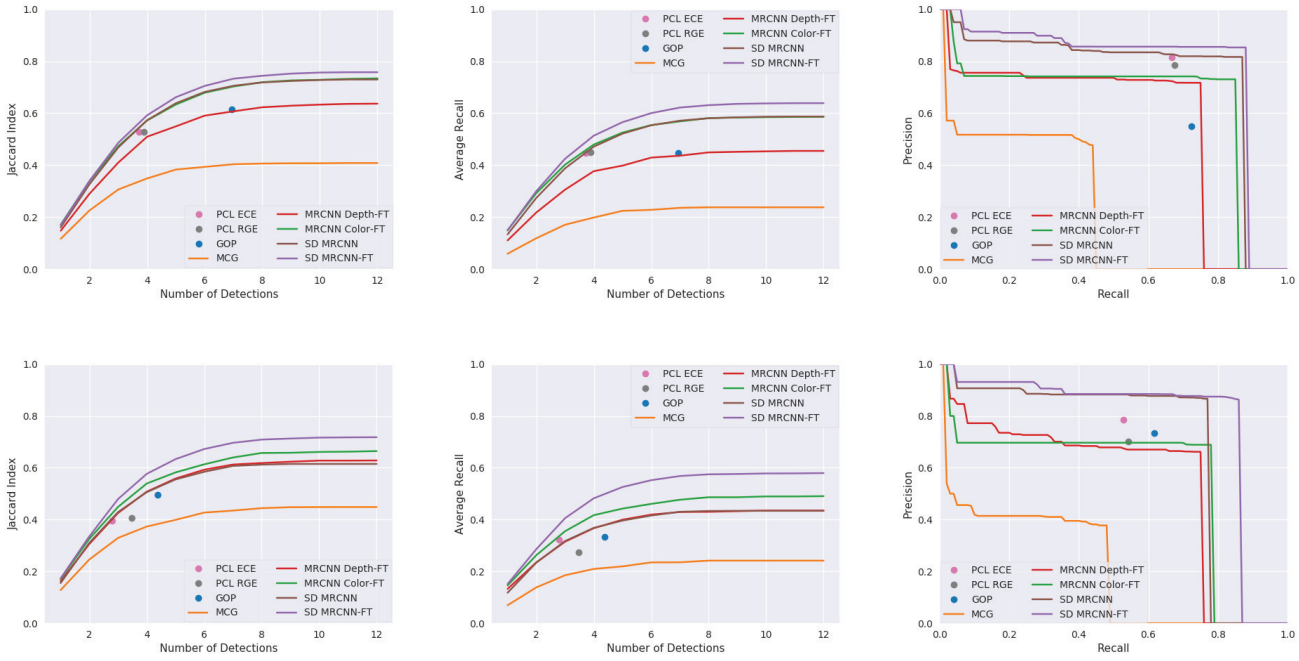

Fig. 2: Average Jaccard Index, Average Recall, and Precision-Recall (at IoU = 0.5) curves for each method and the high-res (top row) and low-res (bottom row) dataset, using segmentation metrics. The fine-tuned SD Mask R-CNN implementation outperforms all baselines on both sensors in the WISDOM-real dataset. The precision-recall curves suggest that the dataset contains some hard instances that are unable to be recalled by any method. These instances are likely heavily occluded objects whose masks get merged with the adjacent mask or flat objects that cannot be distinguished from the bottom of the bin.

segmentation method followed by a standard CNN classifier, which is easier to generate training data for and faster to train.

To train the classifier, we collected ten RGB images of each target object in isolation. Each image was masked and cropped automatically using depth data, and then each crop was augmented by randomly masking the crop with overlaid planes to simulate occlusions. From the initial set of 100 images, we produced a dataset of 1,000 images with an 80-20 train-validation split. We then used this dataset to fine-tune the last four layers of a VGG-16 network [7] pre-trained on Imagenet. Fine-tuning the network for 20 epochs took less than two minutes on a Titan X GPU, and the only human intervention required was capturing the initial object images.

Given a pre-trained classifier, the procedure used to execute instance-specific suction grasps was composed of three phases. First, an RGB-D image was taken of the bin with the PhoXi and a class-agnostic instance segmentation method is

Method	Success Rate (%)	Prec. @ 0.5 (%)	# Corr. Targets
Euclidean Clustering	56 ± 14	63 ± 19	35
Fine-Tuned Mask R-CNN (Color)	78 ± 11	85 ± 12	44
SD Mask R-CNN	74 ± 12	87 ± 11	39

TABLE II: Results of semantic segmentation experiments, where success is defined as grasping and lifting the correct object. **(Success Rate)** Number of successful grasps of the correct object over 50 trials. **(Prec. @ 0.5)** Success rate when the classifier was $> 50\%$ certain that the selected segment was the target object. **(# Corr. Targets)** Number of times the robot targeted the correct object out of 50 trials.

used to detect object masks. Then, the classifier was used to choose the mask that is most likely to belong to the target object. Finally, a suction grasp was planned and executed by constraining grasps planned by Dex-Net 3.0 to the target object mask [8].

We benchmarked SD Mask R-CNN on this pipeline against two segmentation methods. First, we compared against the PCL Euclidean Clustering method to evaluate baseline performance. Second, we compared with Mask R-CNN fine-tuned on the WISDOM-Real training dataset to evaluate whether SD Mask R-CNN is competitive with methods trained on real data.

Each segmentation method was tested in 50 independent trials. Each trial involved shaking the ten objects in a box, pouring them into the bin, and allowing the system to select a target object uniformly at random to attempt to locate and grasp. The results of these instance-specific grasping experiments are shown in Table II.

SD Mask R-CNN outperforms the PCL baseline and achieves performance on par with Mask R-CNN fine-tuned on real data, despite the fact that SD Mask R-CNN was training on only synthetic data. This suggests that high-quality instance segmentation can be achieved without expensive data collection from humans or self-supervision. This could significantly reduce the effort needed to take advantage of object segmentation for new robotic tasks.

References

[1] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 328–335.

[2] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *European Conference on Computer Vision*. Springer, 2014, pp. 725–739.

[3] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*. IEEE, 2011, pp. 1–4.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[5] D. Morrison, A. Tow, M. McTaggart, R. Smith, N. Kelly-Boxall, S. Wade-McCue, J. Erskine, R. Grinover, A. Gurman, T. Hunn *et al.*, "Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge," *arXiv preprint arXiv:1709.06283*, 2017.

[6] M. Schwarz, A. Milan, C. Lenz, A. Munoz, A. S. Periyasamy, M. Schreiber, S. Schüller, and S. Behnke, "Nimbro picking: Versatile part handling for warehouse automation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3032–3039.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust robot suction grasp targets in point clouds using a new analytic model and deep learning," *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2017.