Learning Rich Features from RGB-D Images for Object Detection and Segmentation : Supplementary Material

Saurabh Gupta[1], Ross Girshick[1], Pablo Arbeláez [1,2], and Jitendra Malik[1]
{sgupta, rbg, arbelaez, malik}@eecs.berkeley.edu

[1]University of California, Berkeley, [2]Universidad de los Andes, Colombia

1 Detection

1.1 PR Curves - Our 19 class task

In this subsection, we present the Precision Recall curves on the NYUD2 *test* set, comparing the output from our object detectors with that from RGB DPMs [1], and RGB-D DPMs as described in Section 3.3 in the main paper. The Precision Recall curves are plotted in Figure 1.

1.2 Object Detection Visualizations - Our 19 class task

We provide more visualization of the detections from our detectors: bed (Figure 3), chair (Figure 4), sofa (Figure 5), table (Figure 6), toilet (Figure 7), lamp (Figure 8), pillow (Figure 9), counter (Figure 10), night-stand (Figure 11), television (Figure 12), monitor (Figure 13), garbage-bin (Figure 14), door (Figure 15), desk (Figure 16), bookshelf (Figure 17), dresser (Figure 18), box (Figure 19), bathtub (Figure 20) and sink (Figure 21). We uniformly sample 30 detections among the top K detections (K = number of instances for that class). We color code the detection box as true positive (green), duplicate detection (blue), mislocalization (yellow) and confusion with other category (red). We also mark the inferred instance mask in magenta.

1.3 PR Curves - Lin *et al.*'s [5] 21 class task

In this subsection, we compare our object detectors with that of Lin *et al.* [5]. Lin *et al.* produce only a few operation points on the Precision Recall curve (corresponding to 8, 15 or 30 region proposals per image). They also use a different set of categories, and combine 'semantically similar' categories into a single category (we list the category groups in Table 2). They produce 3D detections, but also have a region associated with each 3D detection, which we use to obtain a 2D detection (by putting a tight bounding box around the region in the image space). We then benchmark the output by assigning the detections as being true positives and false positives, and compute precision and recall for each of their setting (of 8, 15, and 30 region proposals per image), and report

the F_1 measure in Table 1. We also report the $\mathrm{F}_{\max}$ measure from our approach, and plot our Precision Recall curves along with their three operating points in Figure 2. We retrained our detector linear SVMs for the task being considered here (but still using features from the network that was finetuned for our 19 class task).

1.4 RGB-D DPMs Baseline

In this subsection, we present empirical evidence for our RGB-D DPMs baseline as described in Section 3.3 in the main paper. Here, we work with the B3DO dataset [3], which has objects in more uncontrolled settings and includes furniture objects like chairs which is our interest in this work. We compare against previously published methods on this dataset [3], and [4], and provide our RGB-D DPM results in Table 3.

2 Instance Segmentation

2.1 List of Feature Channels

We use the following feature channels: x, y coordinate of the pixel in the 50×50 mask, depth, depth corrected by the scale of the detection (we do this by subtracting the median depth in the detection box from the raw depth), height above ground, angle with gravity, azimuth, the N_x, N_y, N_z components for the normal, Luv color channels, and if the pixel was missing in the original input depth map. Also, in addition to question involving difference in values, we also ask questions about angle between normals at a pair of points.

	F_1@8	F_1@15	F_1@30	F_{max}	AP
	Lin *et al.* [5]	Lin *et al.* [5]	Lin *et al.* [5]	Our	Our
mean	**16.6**	**17.9**	**18.1**	**43.7**	**35.8**
bathtub	0.0	6.2	20.5	**50.0**	44.4
bed	30.3	31.2	28.8	**66.5**	67.5
blinds	23.6	19.2	20.5	**42.9**	32.7
board	16.7	15.6	15.6	**35.0**	25.9
cabinet	20.6	18.8	18.4	**36.2**	27.8
chair	23.0	22.8	22.8	**48.6**	43.5
chest	19.7	22.4	21.6	**33.2**	26.1
counter	19.7	22.2	20.1	**51.9**	43.8
curtain	18.2	16.5	15.1	**38.2**	25.0
headboard	14.0	16.7	8.5	**36.4**	21.0
mantel	0.0	13.3	0.0	**40.0**	29.5
microwave	10.0	19.7	23.7	**36.9**	28.5
monitor	24.6	22.8	25.1	**62.1**	59.1
oven	5.4	15.8	11.1	**50.0**	41.0
printer	4.9	0.0	0.0	**30.6**	23.8
refrigerator	6.9	6.1	16.4	**30.2**	21.8
shelf	16.3	17.4	15.3	**29.6**	19.2
sink	17.8	21.0	19.0	**47.0**	36.1
sofa	22.3	23.9	25.4	**58.2**	55.4
table	16.7	17.7	17.3	**40.1**	33.3
toilet	37.9	26.3	35.6	**54.2**	46.8

Table 1. F1 scores for object detection on the 21 class task as introduced by Lin *et al.* [5]: We report the F_1 measure at the three operating points using the precomputed results from Lin *et al.* [5] (see Section 1.3), and the best operating point from our approach F_{max}.

Class Name	Constituent Categories
bathtub	bathtub
bed	bed, mattress, bunk bed
blinds	blinds, reflection of window shutters
board	cork board, whiteboard, blackboard, classroom board, poster board, board, display board
cabinet	cabinet, storage space
chair	chair, stacked chairs, plastic chair
chest	stand, night stand, dresser, drawer, tv stand, chest, desk drawer, storage chest, desser
counter	counter, kitchen island
curtain	curtain, shower curtain, door curtain
headboard	headboard
mantel	mantel, fireplace, mantle
microwave	microwave, toaster oven
monitor	monitor, television
oven	oven
printer	printer, fax machine
refrigerator	refridgerator
shelf	shelves, mail shelf, bookshelf, spice rack, toy shelf, toys shelf, storage shelvesbooks, toys rack, storage rack, shelf frame
sink	sink
sofa	sofa, furniture
table	table, desk, coffee table, table runner, game table, foosball table, pool table, ping pong table
toilet	toilet

Table 2. Category groups as used by Lin *et al.* [5]

	DPM [1]	Janoch *et al.* [3]-Prn	Janoch *et al.* [3]-Rscr	Kim *et al.* [4]	RGB-D DPMs
mean	28.2	29.7	30.7	31.2	**39.4**
bottle	10.1	10.4	10.1	10.1	**21.9**
bowl	37.8	38.8	38.0	45.4	**47.8**
chair	16.8	21.8	23.0	17.1	**39.9**
cup	30.9	33.6	35.6	38.3	**47.0**
keyboard	22.3	24.2	25.0	25.6	**25.7**
monitor	66.8	64.8	66.7	**68.2**	64.9
mouse	22.8	25.2	27.6	25.4	**48.8**
phone	18.0	19.2	19.7	**19.8**	19.4

Table 3. Performance on B3DO: Comparison of our RGB-D DPM baseline with [3], [4] on B3DO dataset [3].

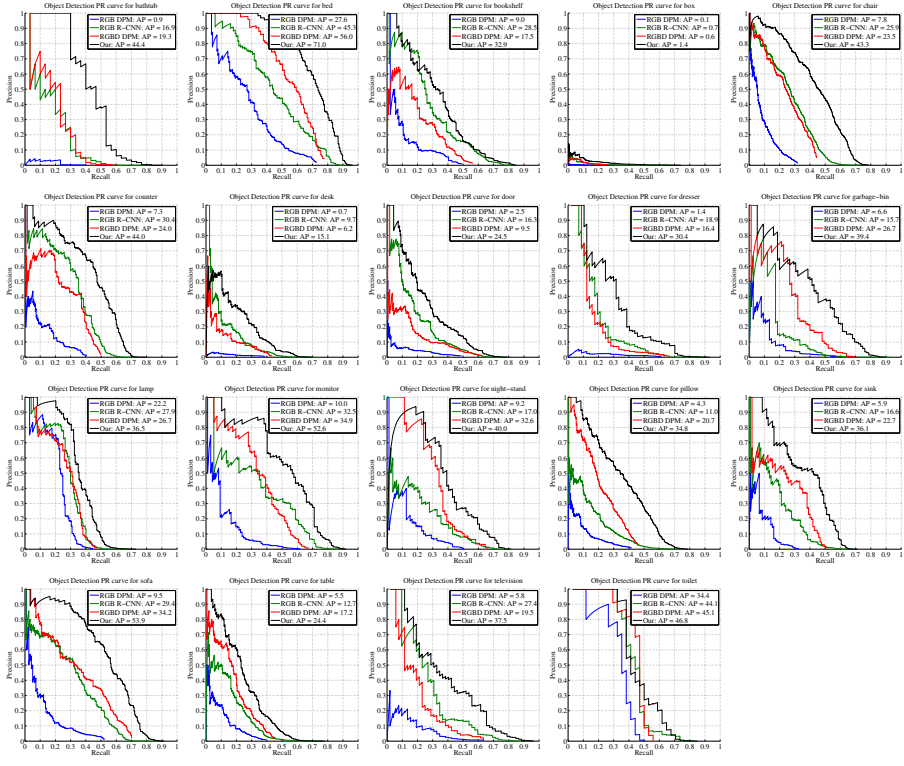


Fig. 1. Precision Recall curves for Object Detection task on NYUD2: We plot the Precision and Recall curves for two baselines, RGB DPMs (blue), RGB R-CNN from [2] (green), RGB-D DPMs (red) and our (black) approach, for the object categories we study in this paper: bathtub, bed, bookshelf, box, chair, counter, desk, door, dresser, garbage-bin, lamp, monitor, night-stand, pillow, sink, sofa, table, television and toilet.

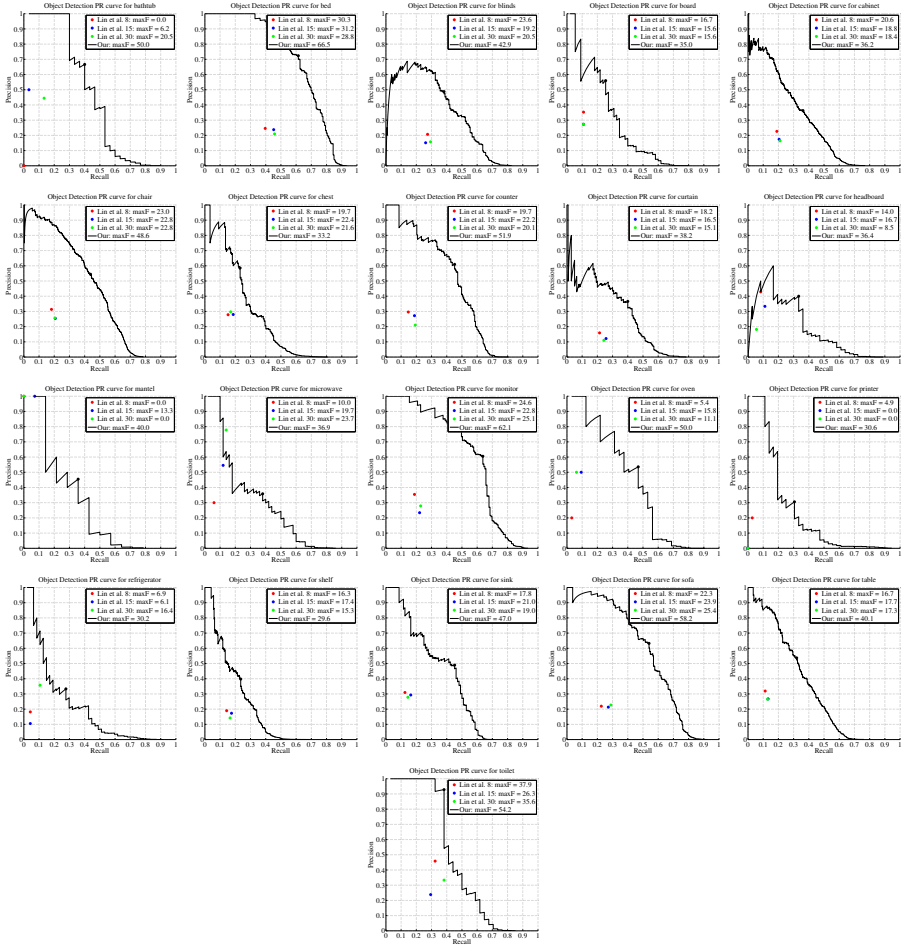


Fig. 2. Precision Recall curves comparing with Lin *et al.* [5]: We plot the Precision and Recall curves for our approach (black) and show the various points of operation as picked by Lin *et al.* in [5] (red, blue, green for 8, 15 and 30 region proposals for each image). For this Figure, we use the categories proposed by Lin *et al.* in [5] (Lin *et al.* group certain categories together, see Table 2), and benchmark for the task of 2D detection that we study in this paper. We use their publicly available results, and use a tight bounding box around their proposed region as the detection box.

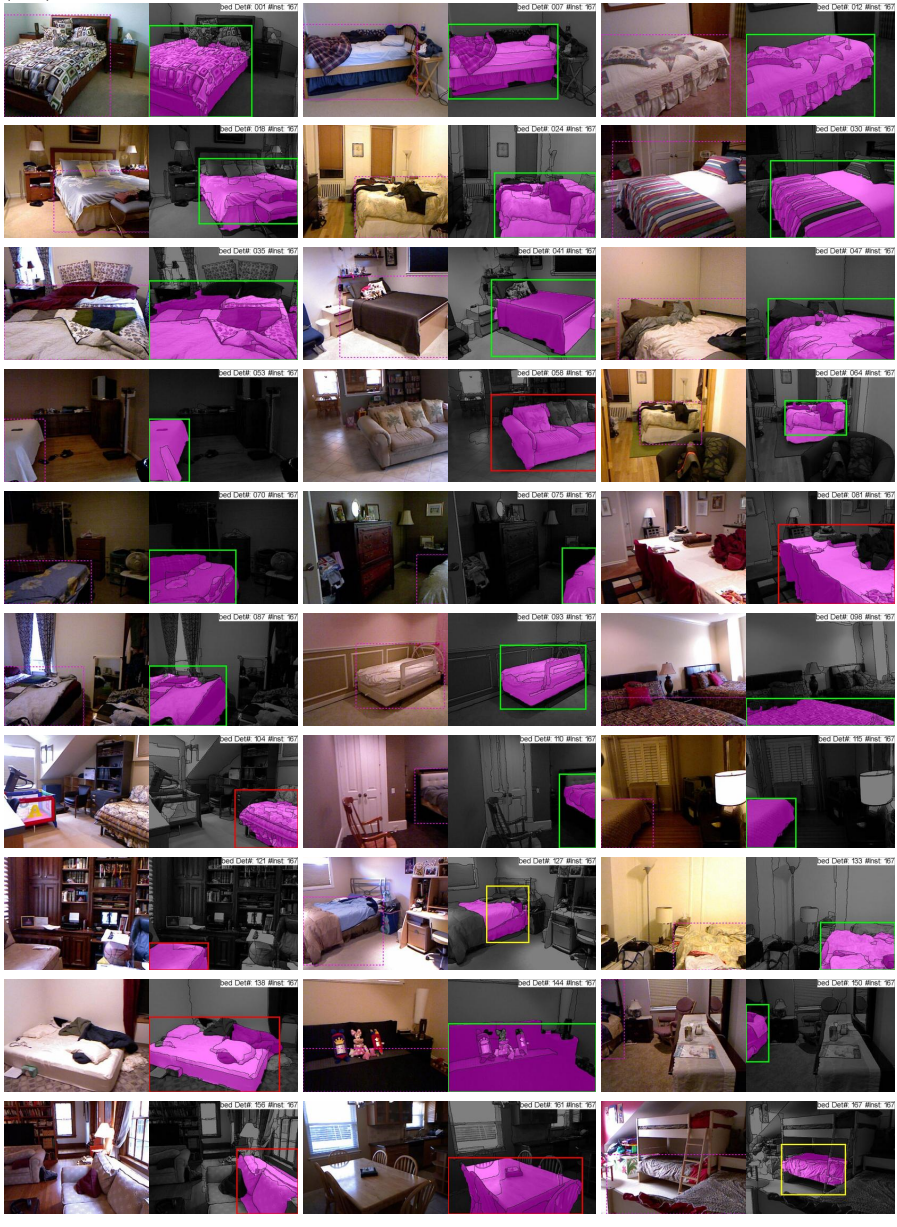


Fig. 3. Output of our bed detector: We visualize 30 detections uniformly sampled from the first $numInsts$ detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

Fig. 4. Output of our chair detector: We visualize 30 detections uniformly sampled from the first $numInsts$ detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

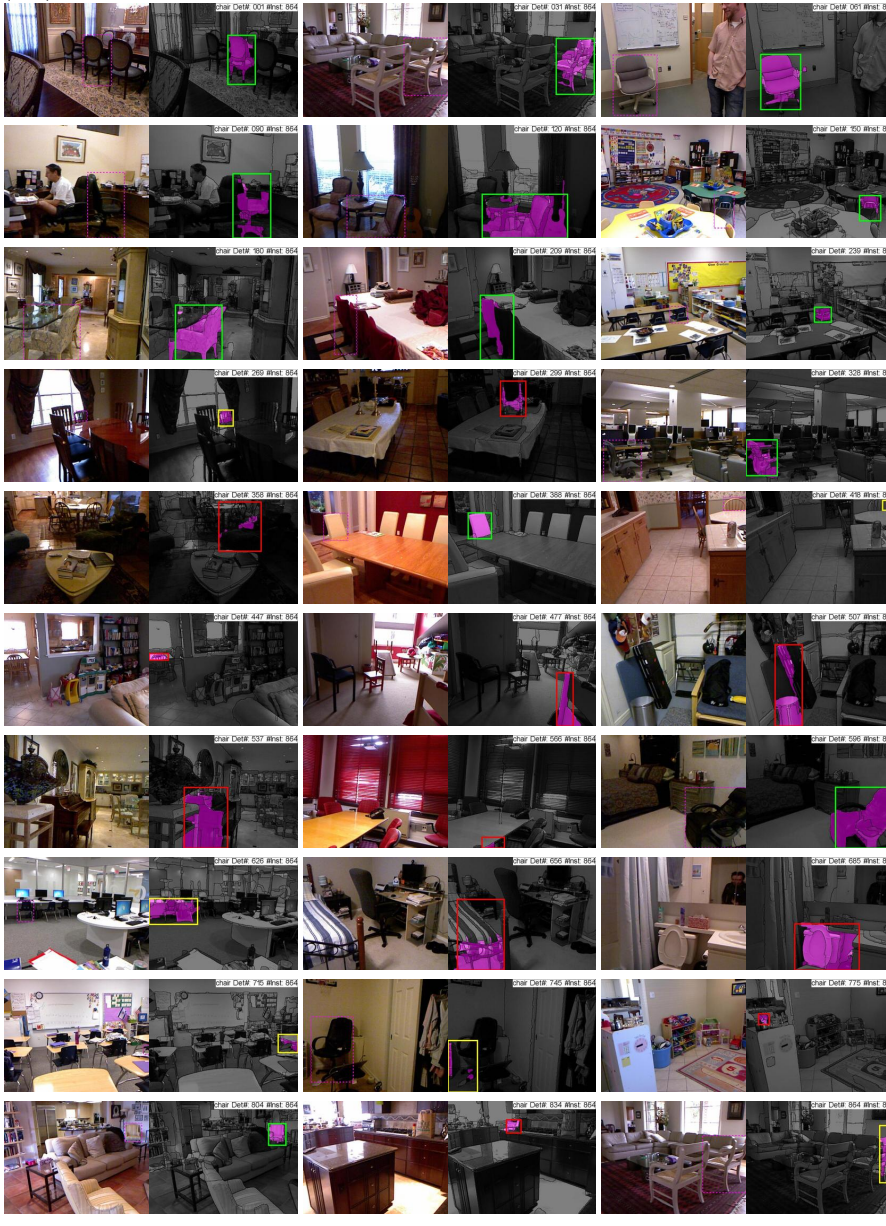


Fig. 5. Output of our sofa detector: We visualize 30 detections uniformly sampled from the first $numInsts$ detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

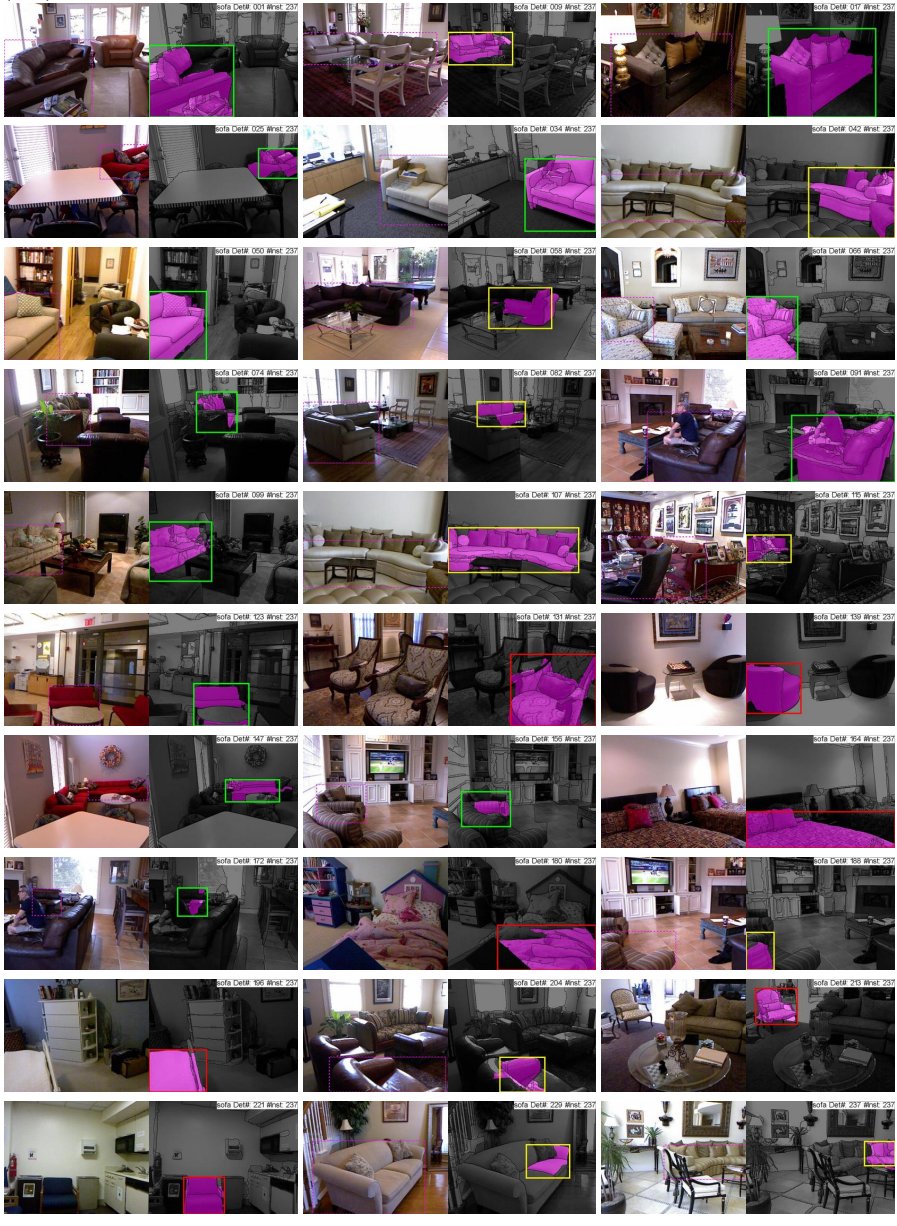


Fig. 6. Output of our table detector: We visualize 30 detections uniformly sampled from the first $numInsts$ detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

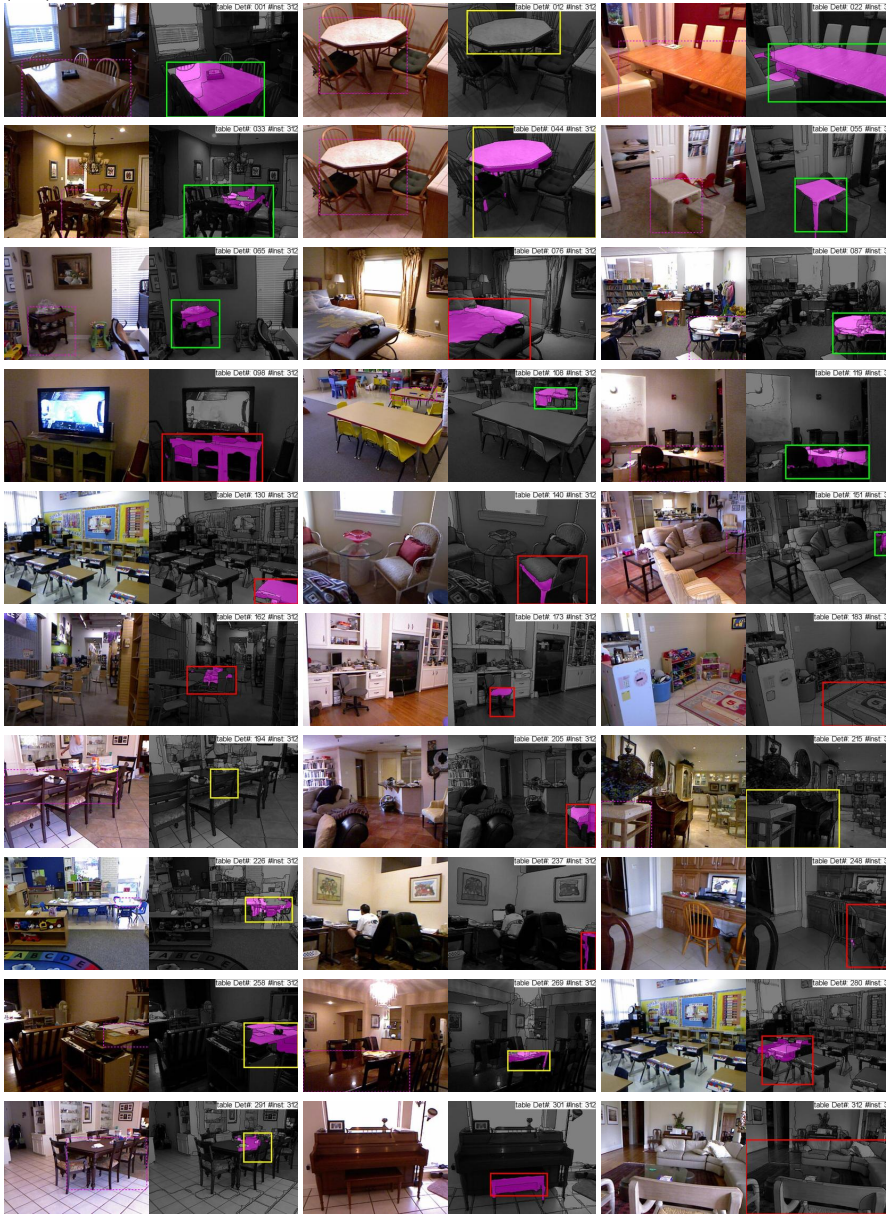


Fig. 7. Output of our toilet detector: We visualize 30 detections uniformly sampled from the first $numInsts$ detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

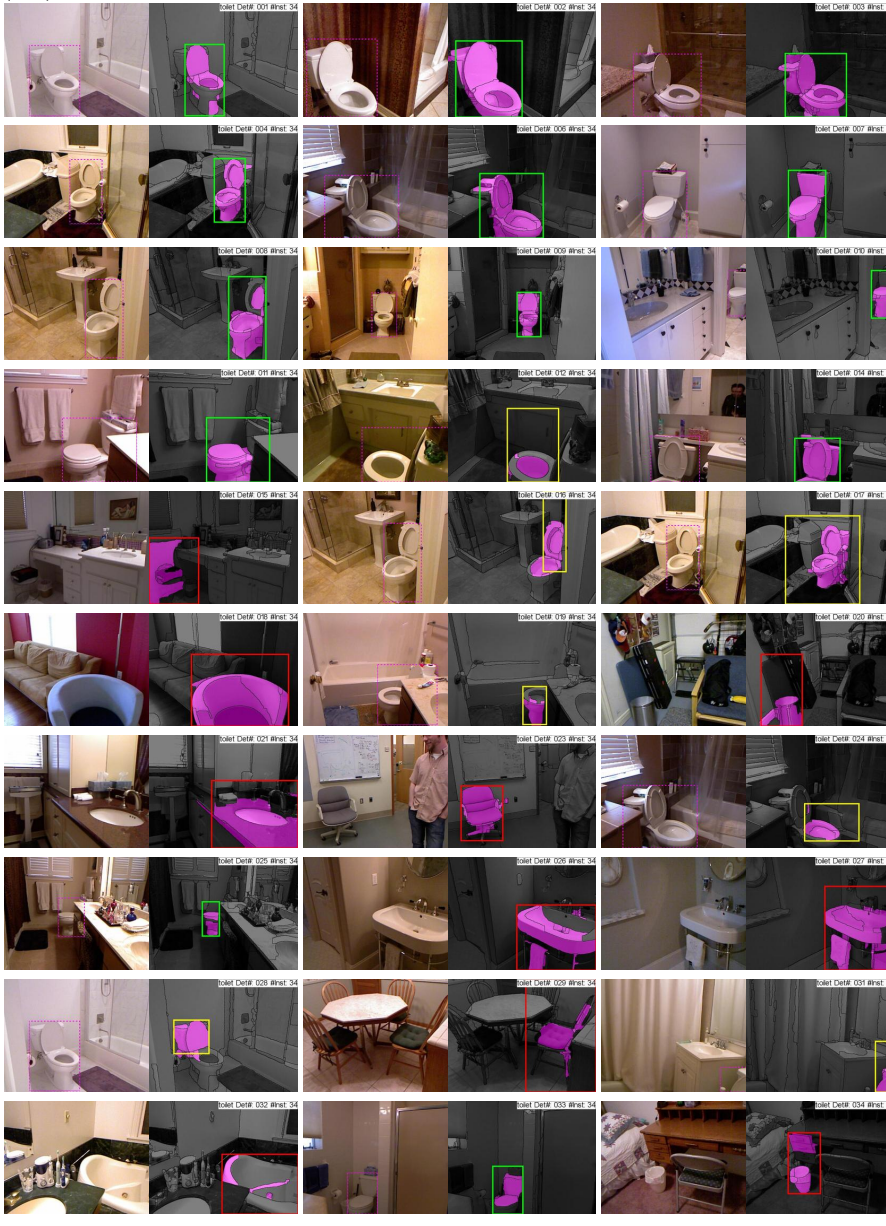


Fig. 8. Output of our lamp detector: We visualize 30 detections uniformly sampled from the first $numInsts$ detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

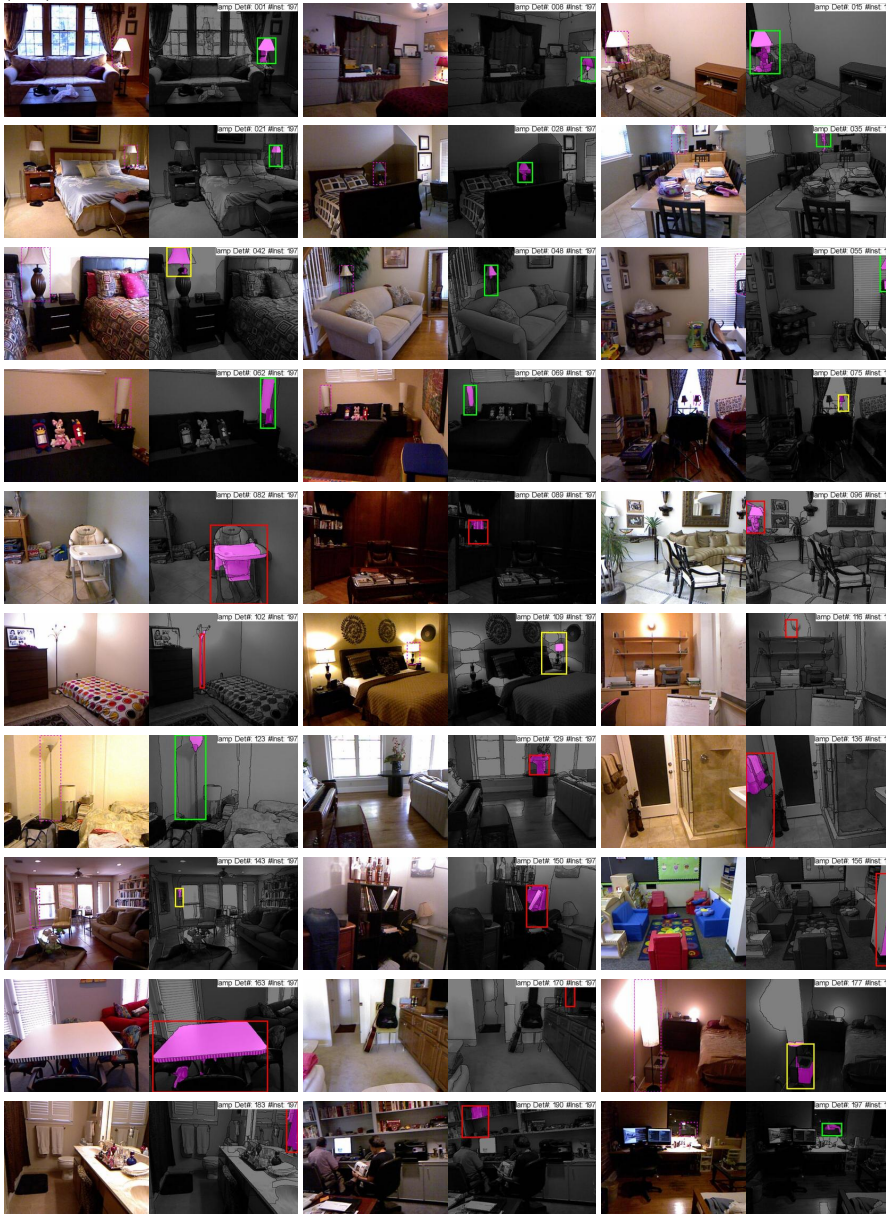


Fig. 9. Output of our pillow detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

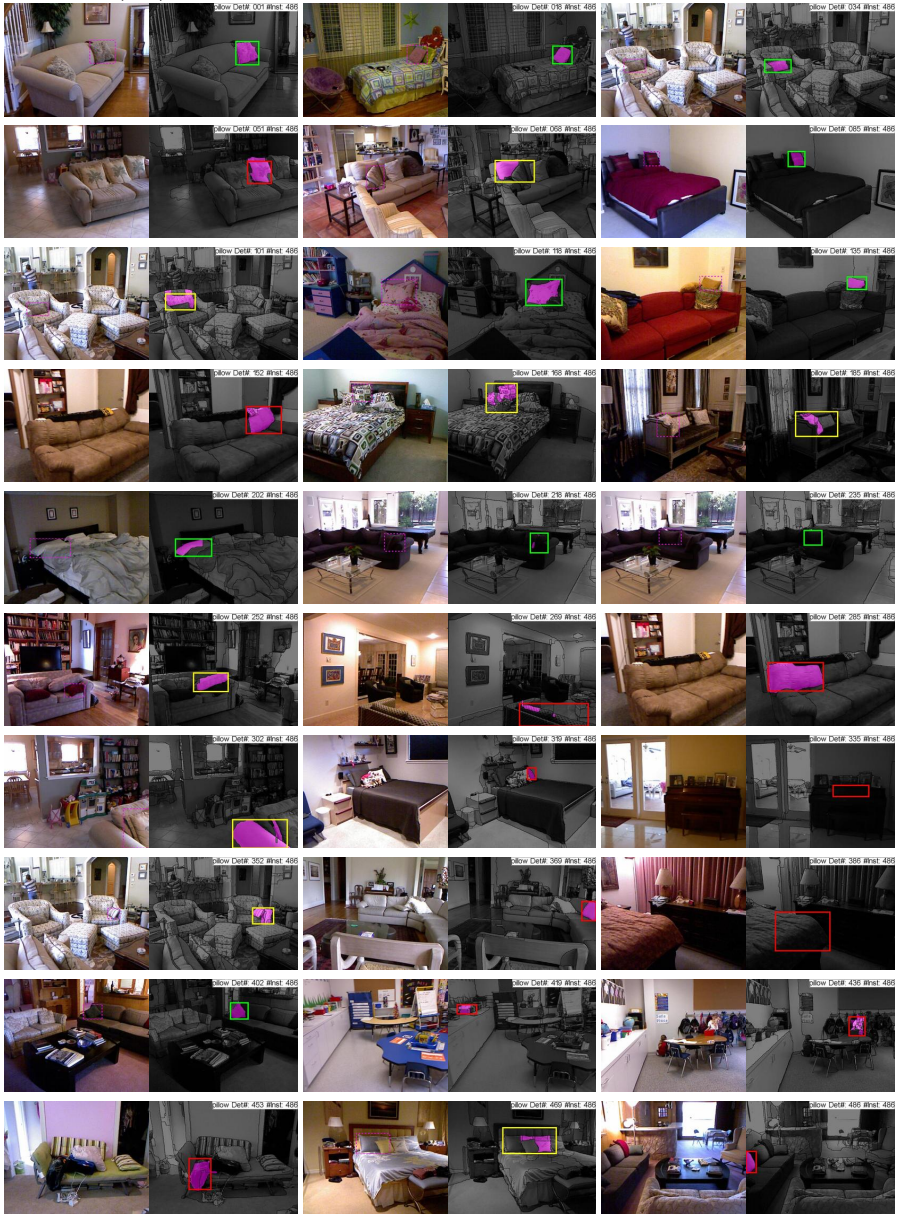


Fig. 10. Output of our counter detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

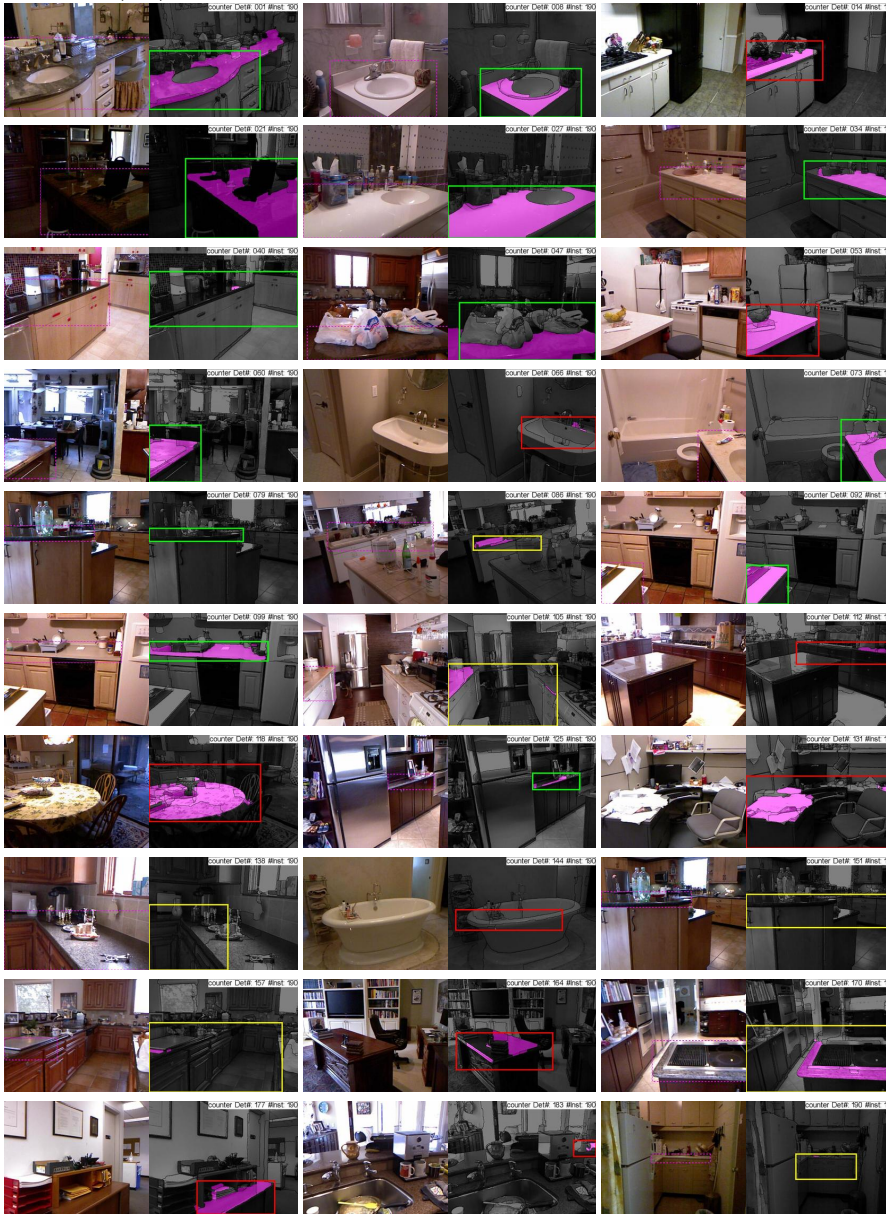


Fig. 11. Output of our night-stand detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

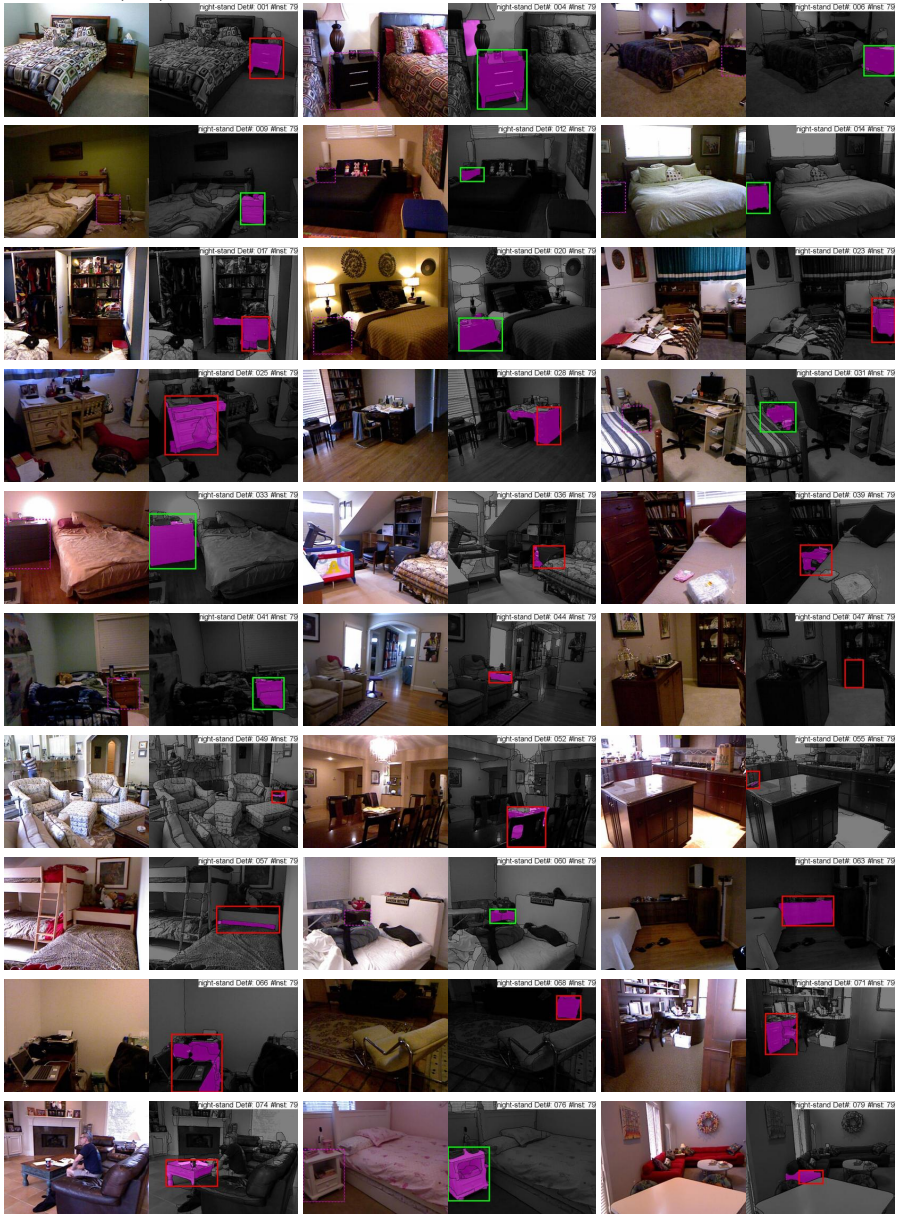


Fig. 12. Output of our television detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

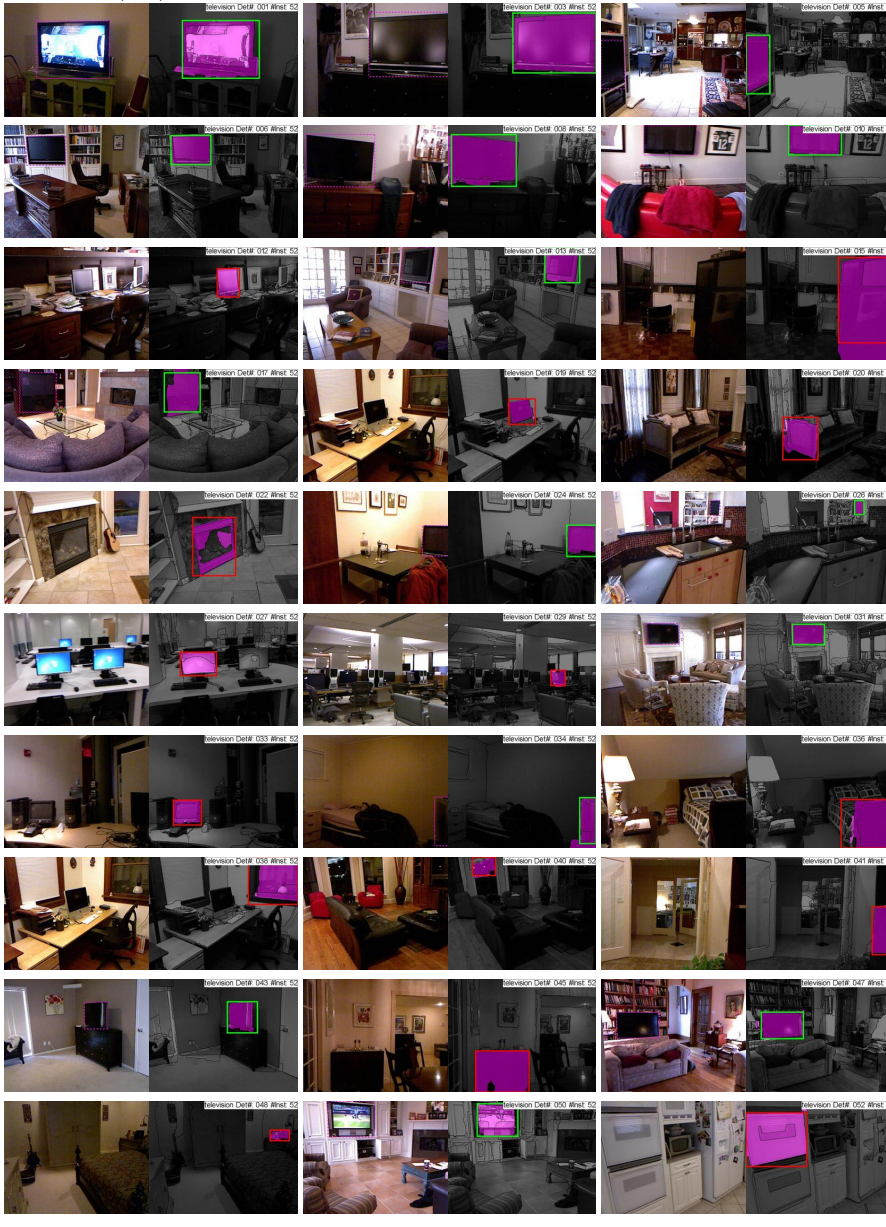


Fig. 13. Output of our monitor detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

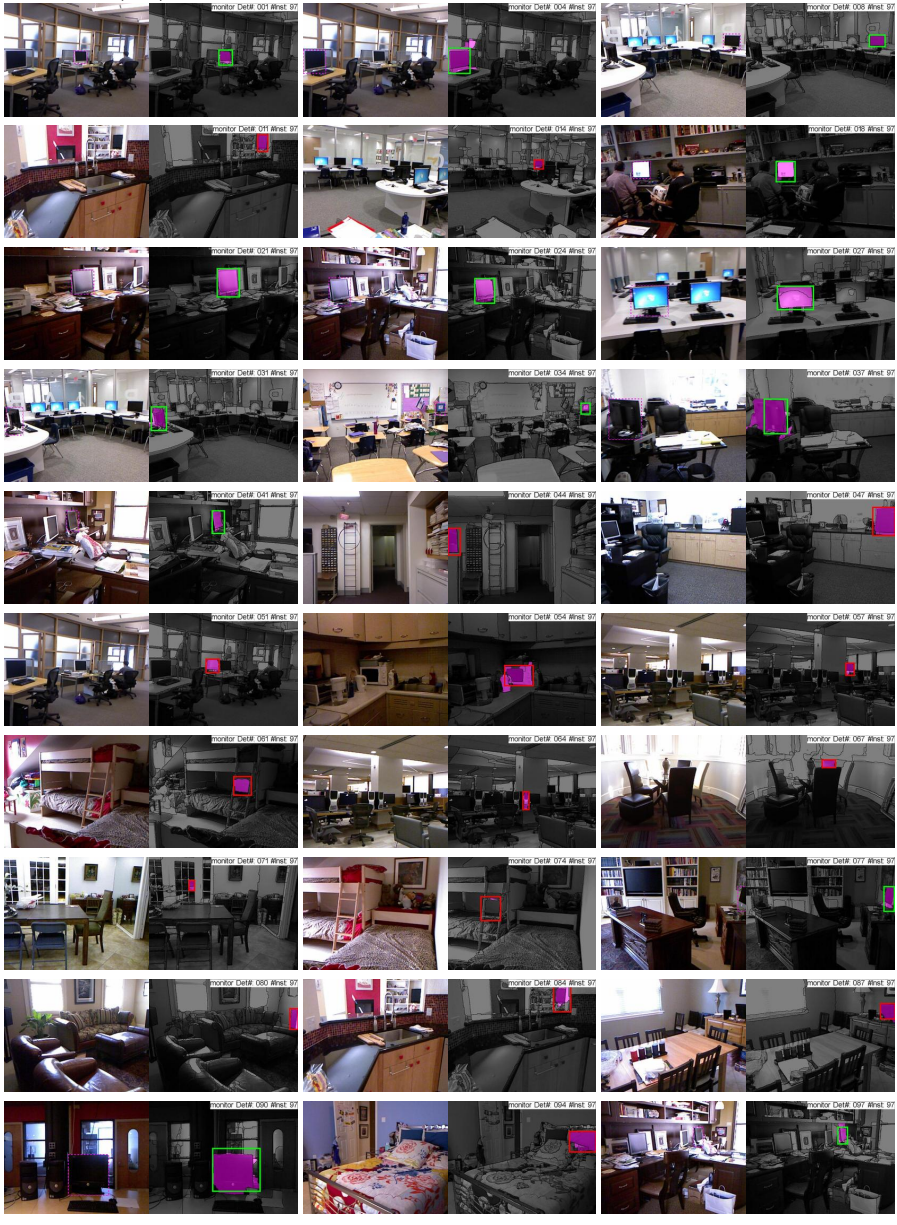


Fig. 14. Output of our garbage-bin detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

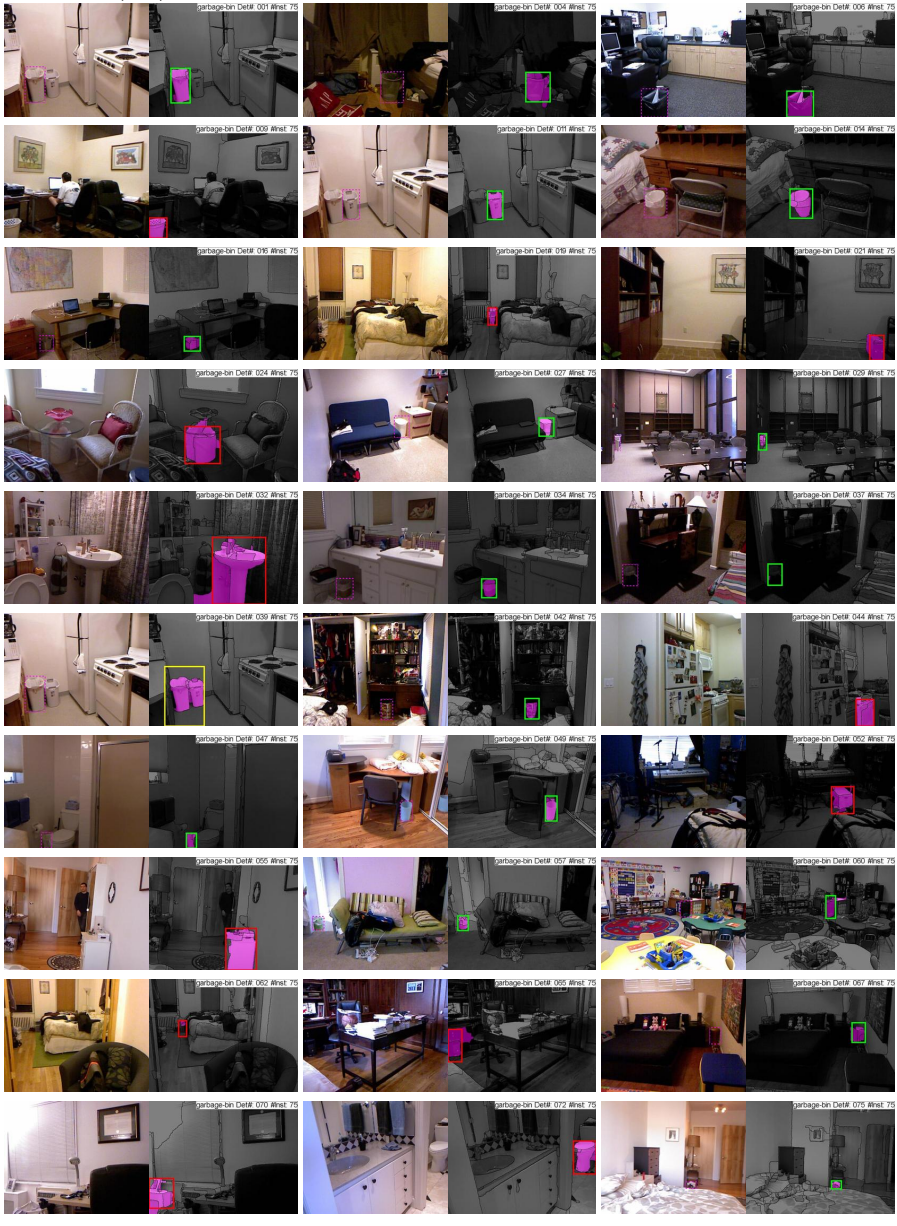


Fig. 15. Output of our door detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

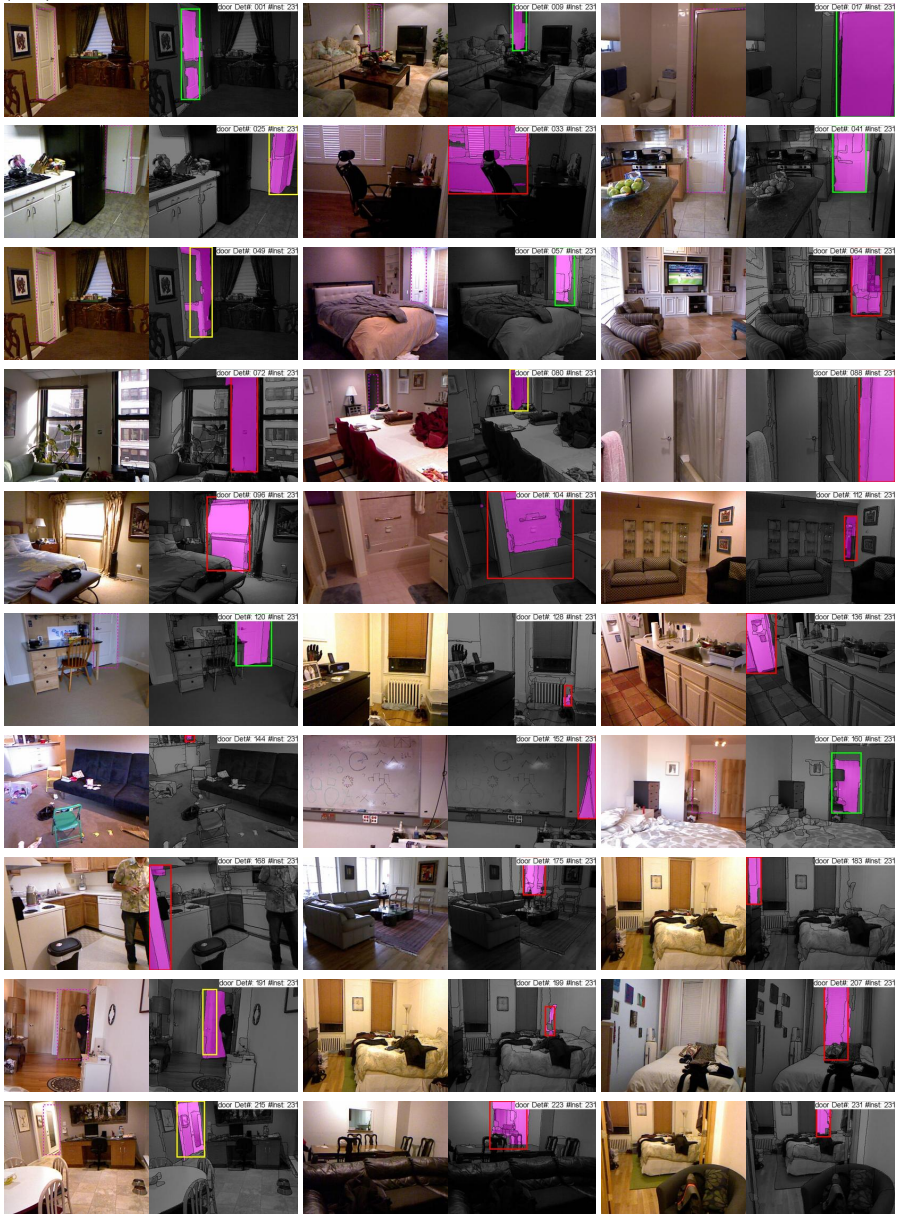


Fig. 16. Output of our desk detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

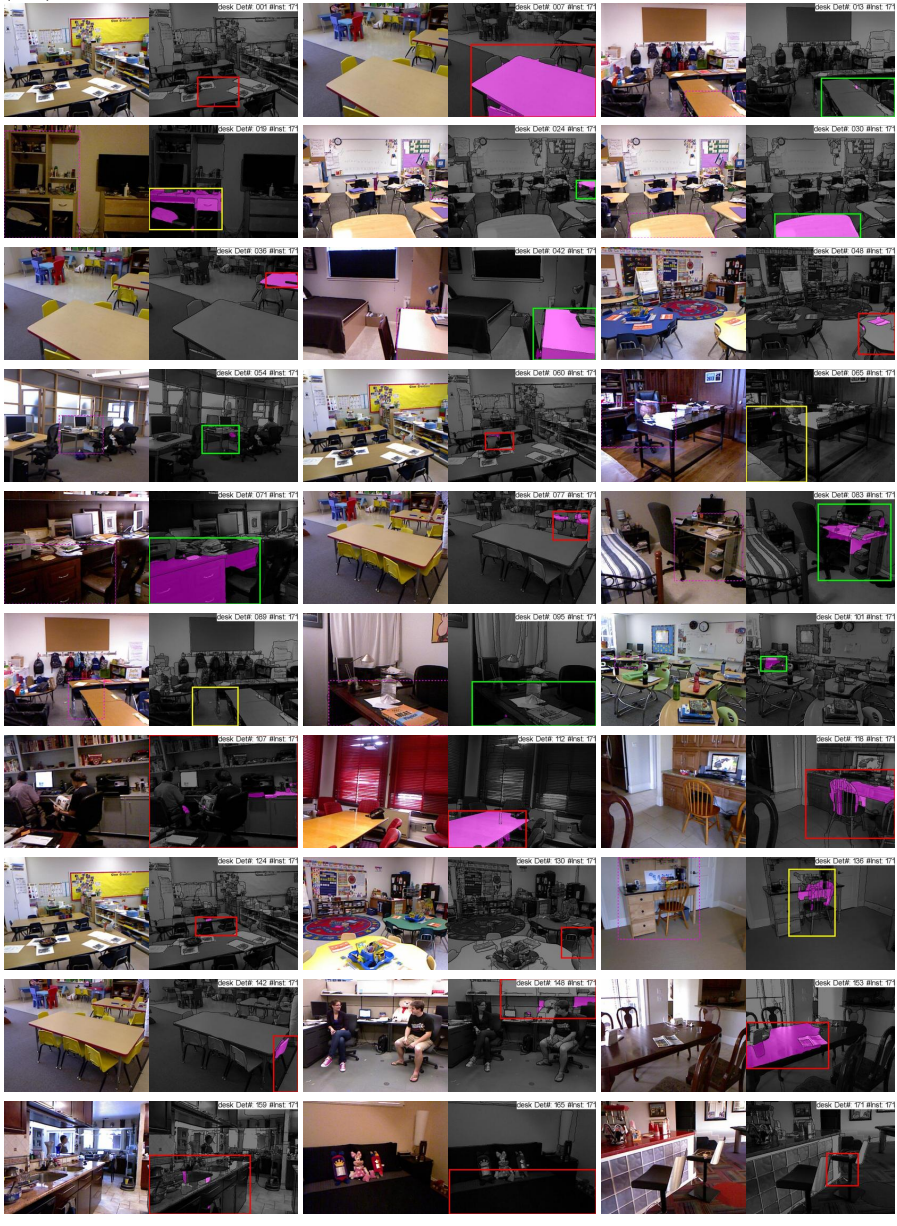


Fig. 17. Output of our bookshelf detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

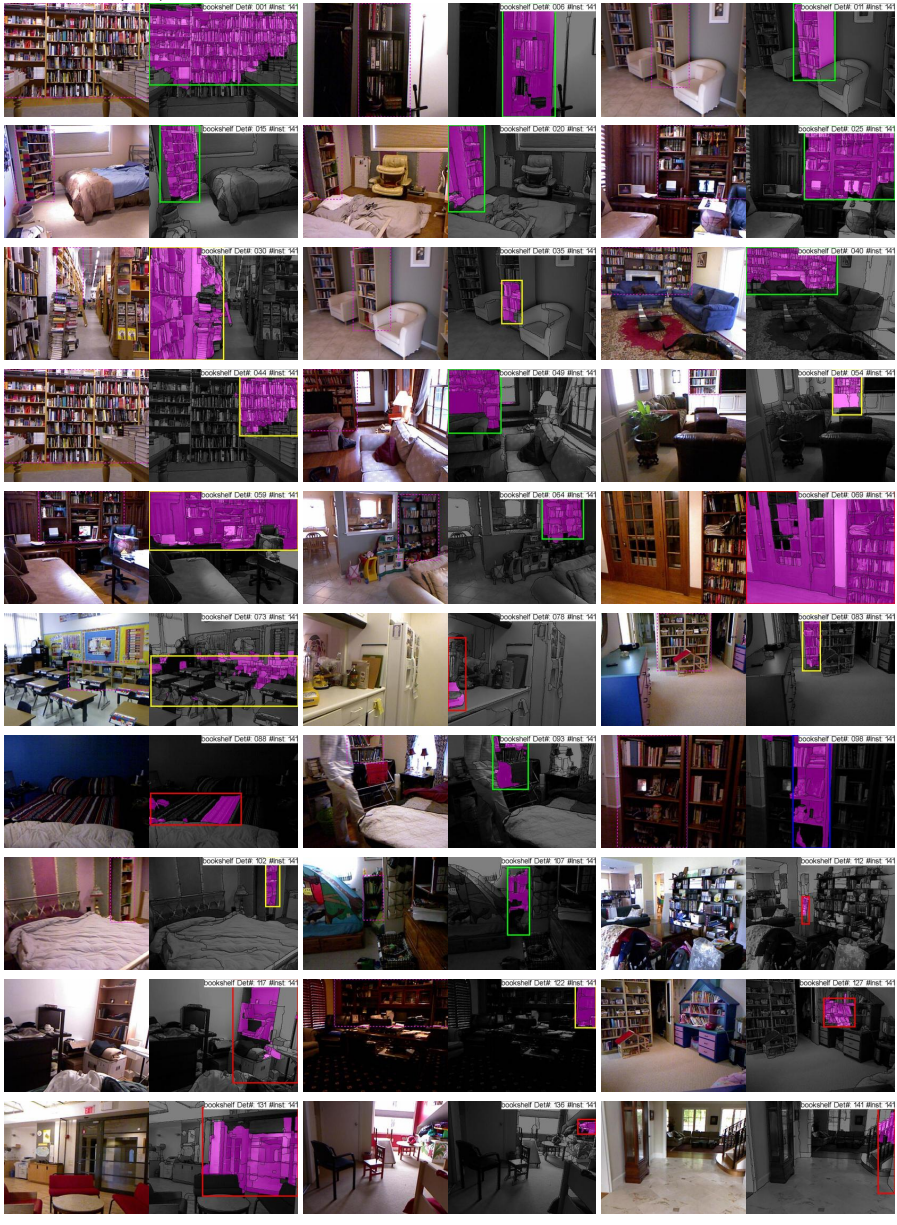


Fig. 18. Output of our dresser detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

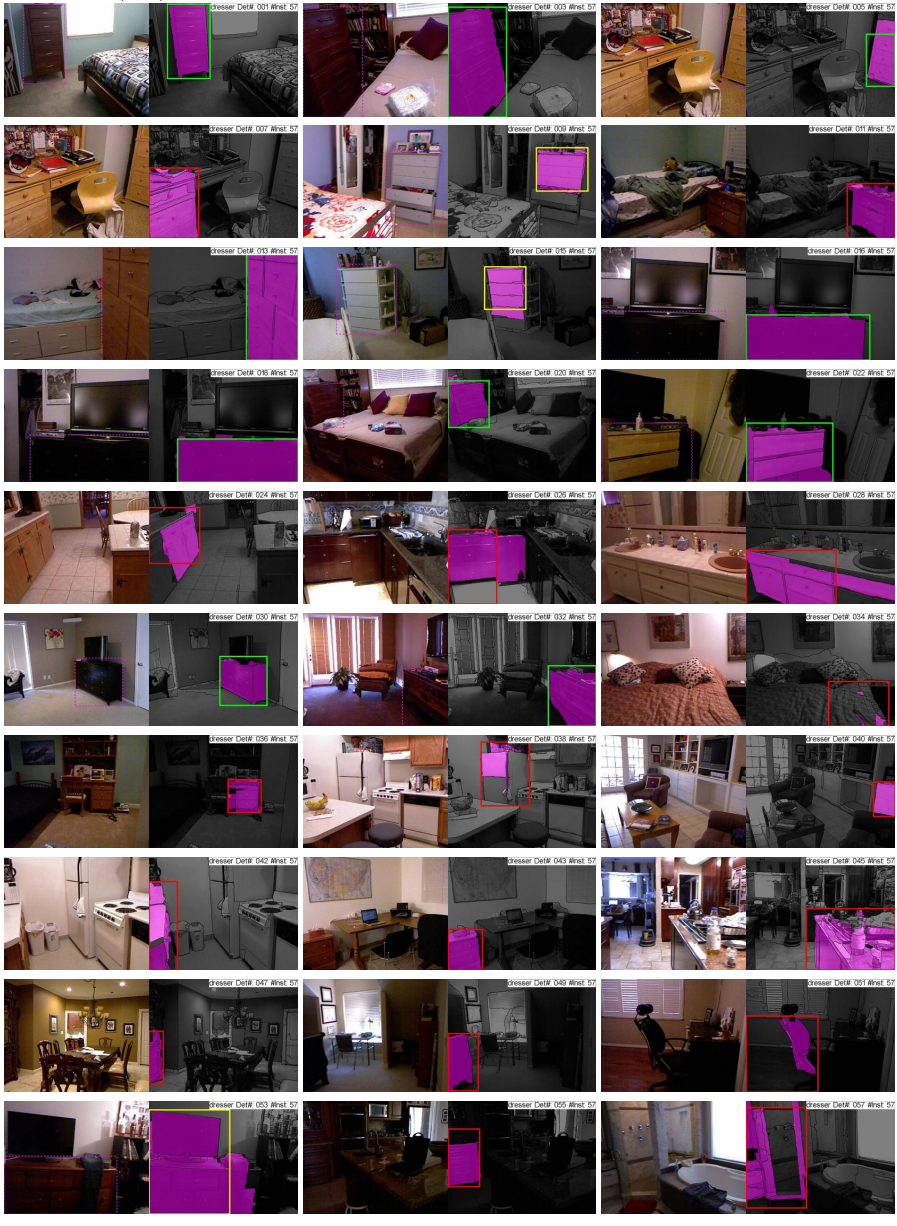


Fig. 19. Output of our box detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

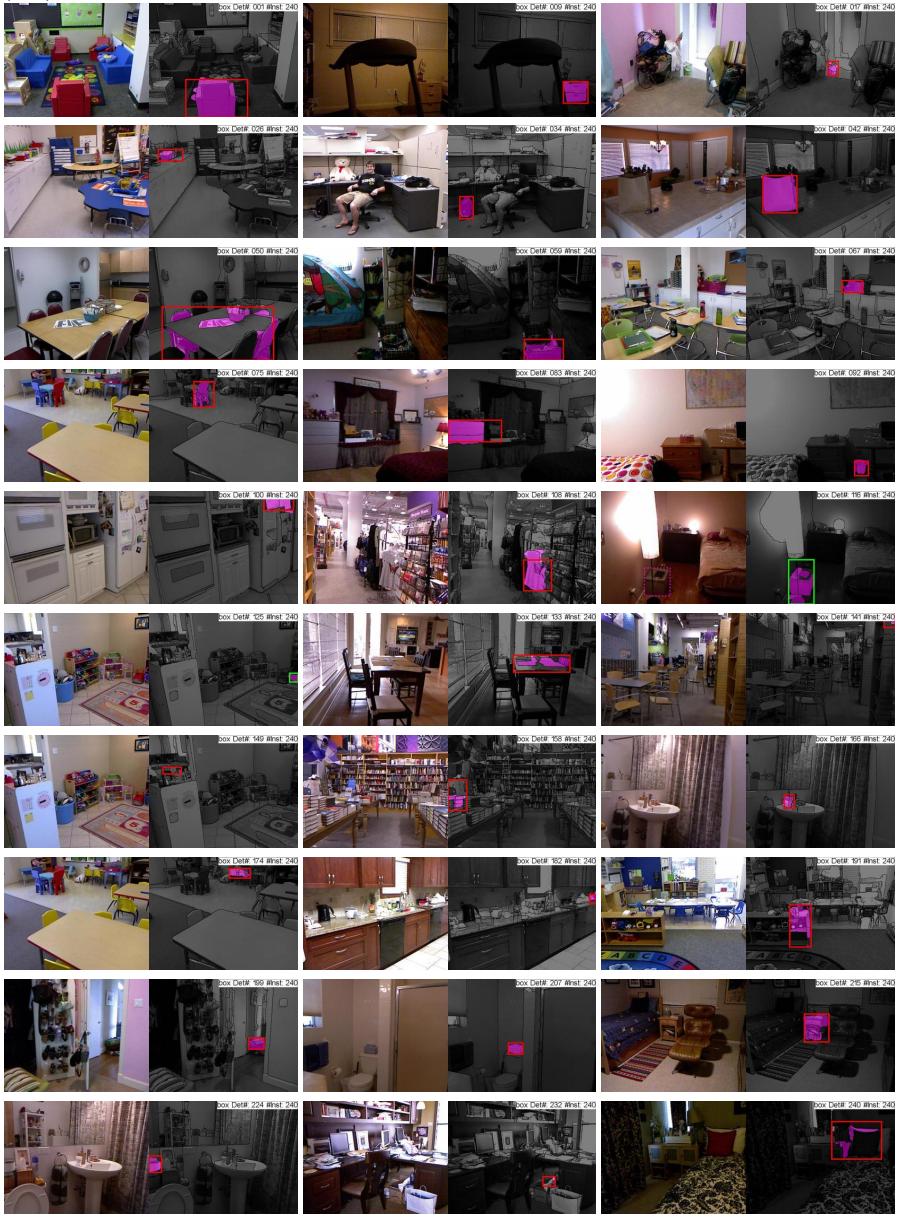


Fig. 20. Output of our bathtub detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.

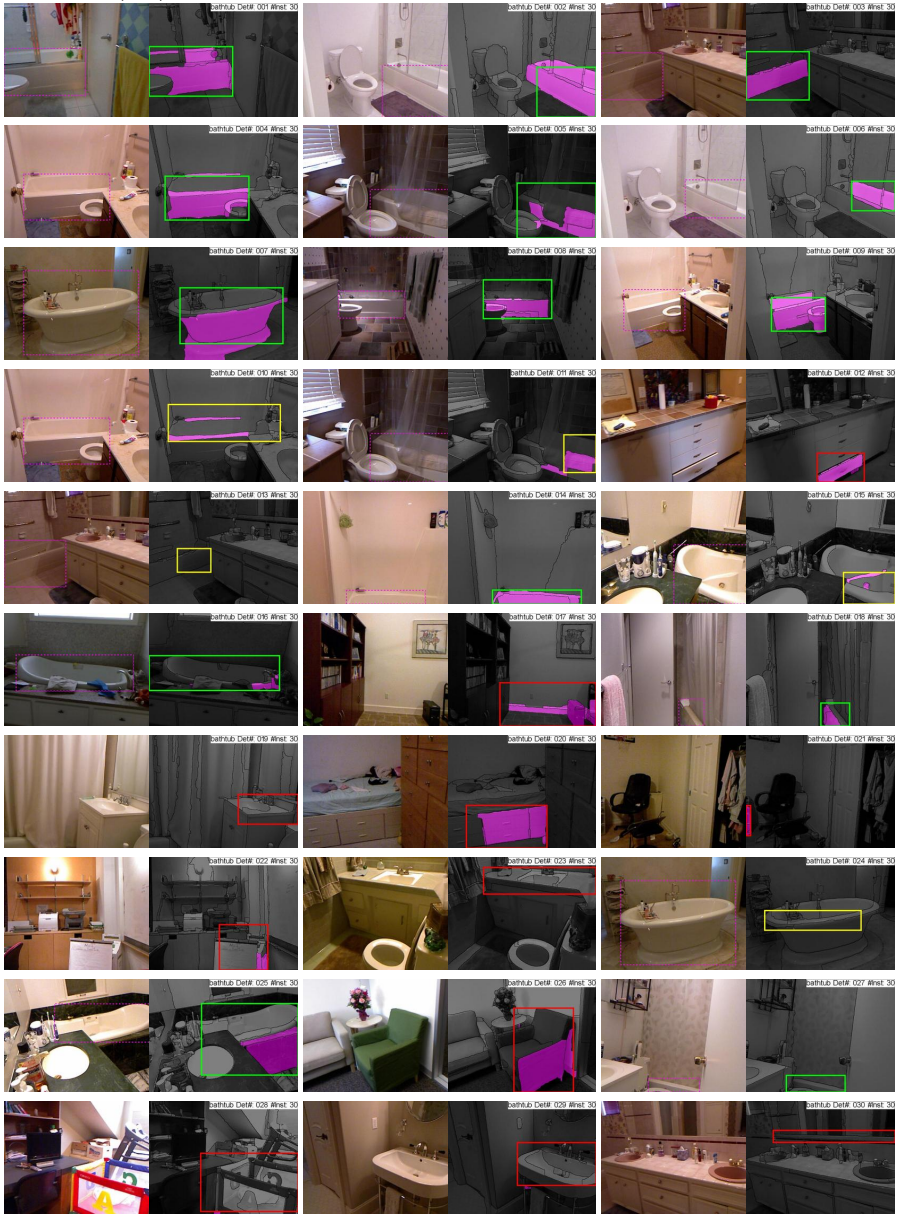
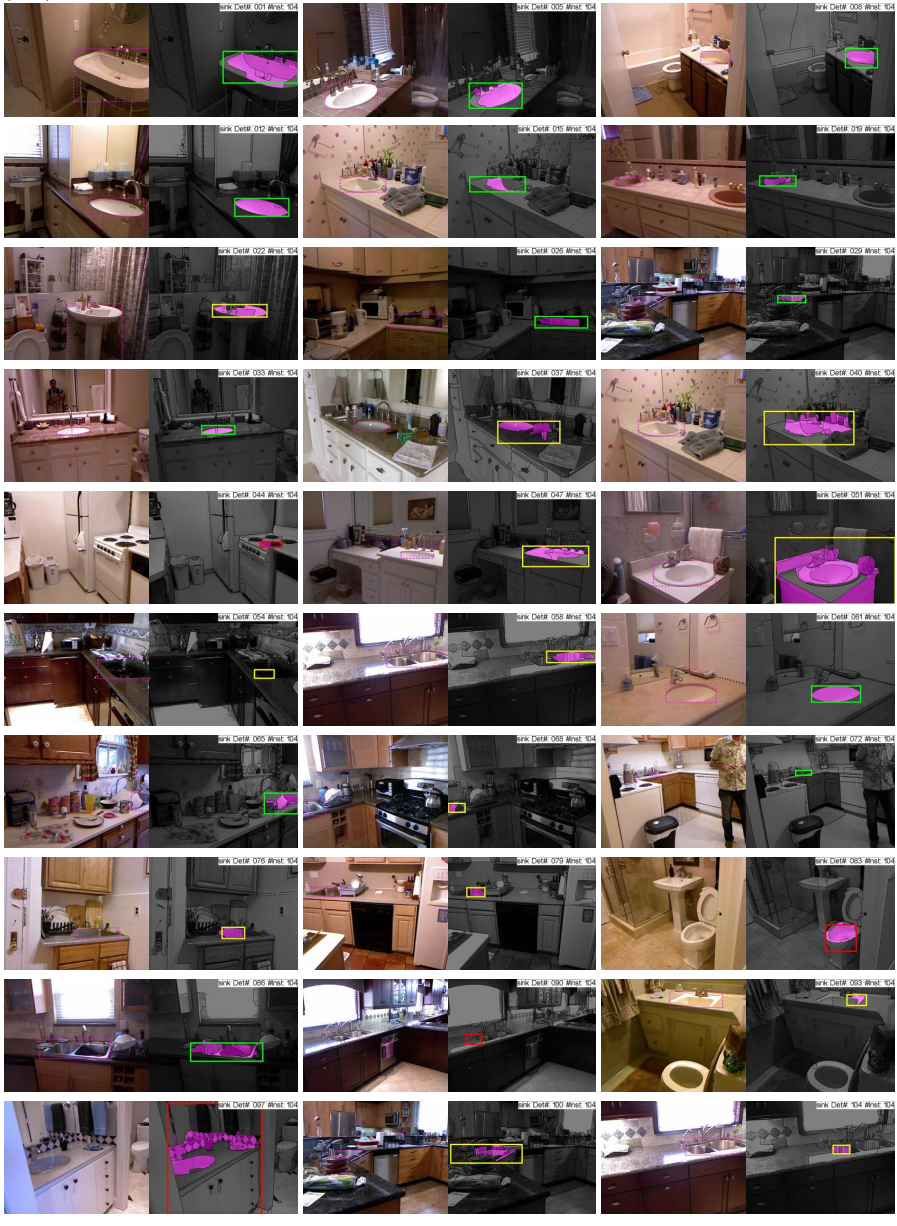


Fig. 21. Output of our sink detector: We visualize 30 detections uniformly sampled from the first *numInsts* detections. The examples are laid out row-wise with 3 examples per row. The first image has the ground truth bounding box and the second image shows the output from the object detector color coded as follows: true positive (green), duplicate detection (blue), mis-localization (yellow) and confusion with other categories (red). We also mark the mask for each detection in magenta.



References

1. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI (2010)
2. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
3. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3d object dataset: Putting the kinect to work. In: Consumer Depth Cameras for Computer Vision (2013)
4. soo Kim, B., Xu, S., Savarese, S.: Accurate localization of 3d objects from RGB-D data using segmentation hypotheses. In: CVPR (2013)
5. Lin, D., Fidler, S., Urtasun, R.: Holistic scene understanding for 3D object detection with RGBD cameras. In: ICCV (2013)