SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite

Shuran Song, Samuel P. Lichtenberg, Jianxiong Xiao
Princeton University

Scene understanding is one of the most important and challenging tasks in computer vision. Although remarkable progress has been made in the past few decades, the performance of general-purpose scene understanding is still far from satisfactory. The recent arrival of affordable depth sensors in consumer markets enables us to obtain reliable depth maps at a very low cost, greatly simplifying some common challenges in computer vision and enabling breakthroughs for several tasks, such as body pose estimation [5], intrinsic image [1], and 3D modeling [2].

RGB-D sensors have also enabled rapid progress for scene understanding. However, although we can download color images from the Internet easily, it is hard to obtain RGB-D data online. Therefore the existing RGB-D recognition benchmarks, such as NYU Depth v2 [4], are an order-of-magnitude smaller than modern recognition datasets (e.g. PASCAL VOC) for color images. Although these small datasets successfully bootstrapped initial research and enabled early progress in the past few years, the size limit is now the common bottleneck in advancing research to the next level. It causes easy overfitting of the algorithm during evaluation, and it cannot support learning for data-hungry algorithms that achieve state-of-the-art performance in color-based recognition tasks. Furthermore, although the RGB-D images these datasets provide contain depth maps, the annotation and evaluation metrics are mostly in the 2D image domain but not directly in 3D, as shown in Figure 1. However, scene understanding will be much more useful in the real 3D space. Hence we advocate that the community should use the depth map to reasoning about scenes and evaluating algorithms in 3D. We introduce SUN RGB-D, a dataset containing 10,355 RGB-D images with dense annotations in both 2D and 3D, for both objects and rooms.

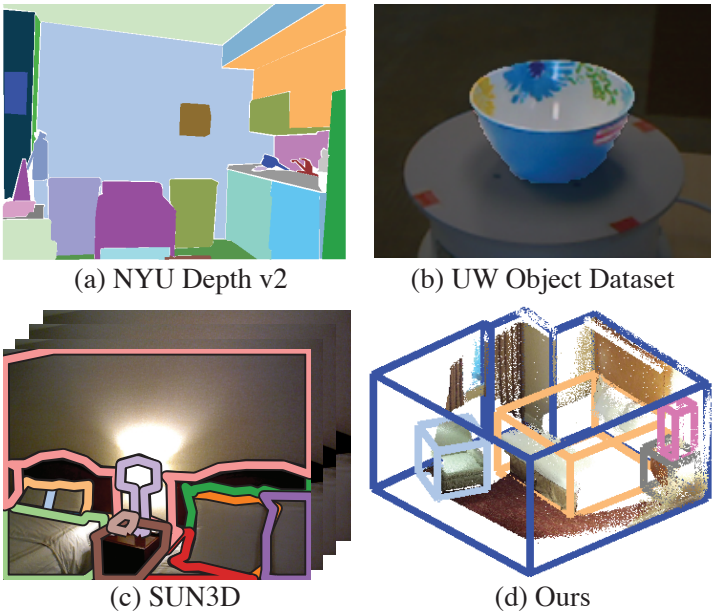
The goal of our dataset construction is to obtain an image dataset captured by various RGB-D sensors (Intel RealSense, Asus Xtion LIVE PRO, Microsoft Kinect versions 1 and 2) at a similar scale as the PASCAL VOC object detection benchmark. We capture 3,784 images using Kinect v2 and 1,159 images using Intel RealSense. We included the 1,449 images from the NYU Depth V2 captured by Kinect v1 and We also choosed 554 realistic scene images from the Berkeley B3DO Dataset captured by Kinect v1, manually selected 3,389 distinguished frames without significant motion blur from the SUN3D videos captured by Asus Xtion. In total, we obtain 10,335 RGB-D images. To improve the depth map quality, we take short videos and use our SIFT+ICP algorithm to do the refinement.

For each image, we annotate the objects with both 2D polygons and 3D bounding boxes and the room layout with 3D polygons. For the 10,335 RGB-D images, we have 146,617 2D polygons and 64,595 3D bounding boxes (with accurate orientations for objects) annotated. Therefore, there are 14.2 objects in each image on average. In total, there are 47 scene categories and about 800 object categories.

To evaluate all major tasks that work towards total scene understanding. we focus on six important recognition tasks towards total scene understanding that integrates objects, room layout and scene class, includes scene categorization, semantic segmentation, object detection, object orientation, room layout estimation, as well as a final total scene understanding task that integrates everything. The final task for our scene understanding benchmark is to estimate the whole scene including objects and room layout in 3D [3]. This task is also referred to "Basic Level Scene Understanding" [6]. We propose this benchmark task as the final goal to integrate both object detection and room layout estimation to obtain a total scene understanding, recognizing and localizing all the objects and the room structure.

We choose state-of-the-art algorithms to evaluate each task. For the tasks without existing algorithm or implementation, we adapt popular algorithms from other tasks. For each task, whenever possible, we try to evaluate algorithms using color, depth, as well as RGB-D images to study the relative importance of color and depth, and gauge to what extent the information from both is complementary. Various evaluation results show that we can apply standard techniques that invented design for color (e.g. hand craft features, deep learning features, detector, sift flow label transfer) to depth domain and it can achieve comparable performance for various tasks.In most of cases, when we combining these two source of information the performance get improved.

By constructing a PASCAL-scale dataset using various sensors and defining a benchmark for all major scene understanding tasks with 3D evaluation metrics, we hope to lay the foundation for advancing RGB-D scene understanding in the coming years.



(a) NYU Depth v2 (b) UW Object Dataset (c) SUN3D (d) Ours

Figure 1: **Comparison of RGB-D recognition benchmarks.** Apart from 2D annotation, our benchmark provided high quality 3D annotation for both objects and room layout.

[1] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. *CVPR*, 2013.

[2] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *UIST*, 2011.

[3] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with RGBD cameras. In *ICCV*, 2013.

[4] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[5] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.

[6] Jianxiong Xiao, James Hays, Bryan C. Russell, Genevieve Patterson, Krista Ehinger, Antonio Torralba, and Aude Oliva. Basic level scene understanding: Categories, attributes and structures. *Frontiers in Psychology*, 4(506), 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00506.

This is an extended abstract. The full paper is available at the http://rgbd.cs.princeton.edu.