Aligning 3D Models to RGB-D Images of Cluttered Scenes

Saurabh Gupta[1], Pablo Arbeláez [2], Ross Girshick[3], Jitendra Malik[1]
[1]UC Berkeley. [2]Universidad de los Andes, Colombia. [3]Microsoft Research.

Truly understanding a scene involves reasoning not just about what is visible but also about what is not visible. Consider for example the images in Figure 1. After we recognize an object as a chair, we have a pretty good sense of how far it extends in depth and what it might look like from another viewpoint. One way of achieving this kind of understanding in a computer vision system would be by 'replacing in-place' the chair pixels by the rendering of a 3D CAD model of the chair [1, 4]. This explicit correspondence to a 3D CAD model leads to a richer representation than output from traditional computer vision algorithms like object detection, semantic or instance segmentation, fine-grained categorization and pose estimation. Our proposed system starts from a single RGB-D image of a cluttered indoor scene and produces the output visualized in Figure 1. Our approach is able to successfully retrieve relevant models and align them with the data. We believe such an output representation will enable the use of perception in fields like robotics.

Figure 2 describes our approach. We use the output of the detection and segmentation system from [3], and first infer the pose of each detected object using a convolutional neural network. We train this CNN on synthetic data using surface normal images instead of depth images as input. We show that this CNN trained on synthetic data works better than one trained on real data. We then use this coarse pose estimate to initialize a search over a small set of 3D models, their scales and exact placements, to finally output a set of 3D model that have been aligned to the objects present in the image. In doing so we only used 2D annotations on the image for training all our models and at test time, are able to generate a rich 3D representation of the scene. We also observe a 48% relative improvement in performance at the task of 3D detection over the current state-of-the-art ('Sliding Shapes') [5], while being an order of magnitude faster at the same time.

CNN for Coarse Pose Estimation: Our first contribution is a shallow three-layer CNN for estimating 3D pose of objects in an RGB-D image. Here, we adopt the geocentric constraint and train the CNN to predict the azimuth of the object (for example for the chair category, which direction the chair is facing in the top view). Additionally, this CNN is trained on surface normal images of synthetic CAD models rendered in different poses, and when we compare this model to a comparable model that was trained on real data, we observe that it works equally well or better than the model trained on real data. This experimental result demonstrate how we can train models for pose estimation without having to manually annotate 3D pose for objects in images.

Model Alignment: We then use the inferred segmentation mask and coarse pose estimate to initialize an iterative closest point based search for the most similar 3D model, and its exact placement in the scene. Here we iterate between the following two steps: a) render the model using the current estimate of translation t, rotation R and scale s and establish correspondence between the observed object and the visible extent of the rendered model, and b) re-estimate the model translation and rotation to minimize the distance between the corresponding points. We find that, when initialized properly, this works well even when working at the level of object categories rather than exact instances, for which ICP has traditionally been used. We use this procedure to align a small number of CAD models to the data and pick the one that fits the data the best.

Results: The final output of our method is a set of 3D CAD model aligned to objects in a given RGB-D image. Figure 1 shows example results. A trivial side product of our algorithm is a 3D bounding box around the object. When we compare these 3D bounding boxes from our algorithm to those from the current state-of-the art [5] at the task of 3D detection we observe large improvements (mean average precision of 58.5% as compared to 39.6% for [5]), while being an order of magnitude faster. We also outperform [2].

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

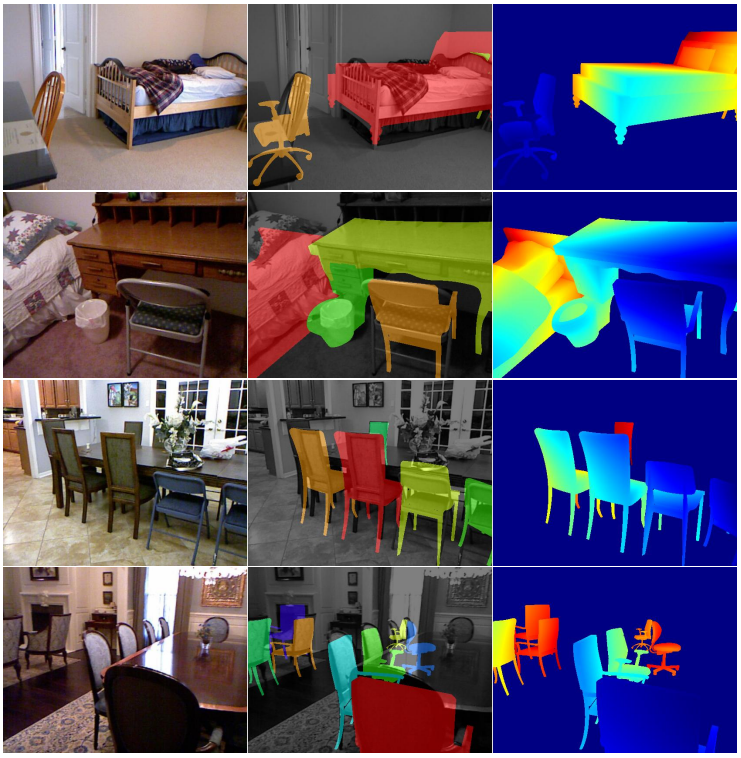

Figure 1: **Output of our system**: We use a single RGB-D image as input and output 3D models associated with objects in the scene.

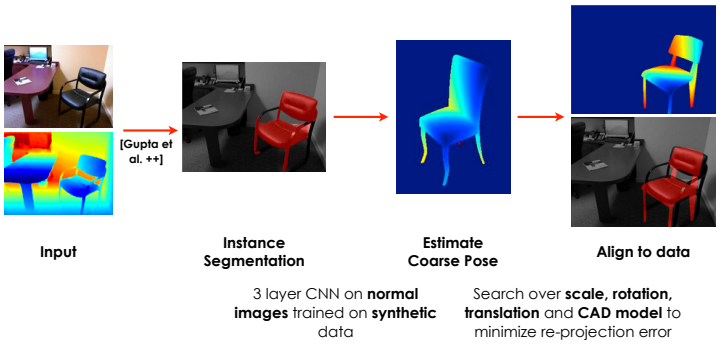

Figure 2: **Overview of approach**: We start with object detection and instance segmentation output from our previous work [3], and first infer the pose of the object using a convolutional neural network, and then search for the best fitting model that explains the data.

[1] Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[2] Ruiqi Guo and Derek Hoiem. *Scene understanding with complete scenes and structured representations*. PhD thesis, UIUC, 2014.

[3] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, 2014.

[4] Joseph J Lim, Aditya Khosla, and Antonio Torralba. Fpm: Fine pose parts-based model with 3d cad models. In *ECCV*, 2014.

[5] Shuran Song and Jianxiong Xiao. Sliding shapes for 3D object detection in depth images. In *ECCV*, 2014.