Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild

Yu Xiang
University of Michigan
yuxiang@umich.edu

Roozbeh Mottaghi
Stanford University
roozbeh@cs.stanford.edu

Silvio Savarese
Stanford University
ssilvio@stanford.edu

Abstract

3D object detection and pose estimation methods have become popular in recent years since they can handle ambiguities in 2D images and also provide a richer description for objects compared to 2D object detectors. However, most of the datasets for 3D recognition are limited to a small amount of images per category or are captured in controlled environments. In this paper, we contribute PASCAL3D+ dataset, which is a novel and challenging dataset for 3D object detection and pose estimation. PASCAL3D+ augments 12 rigid categories of the PASCAL VOC 2012 [4] with 3D annotations. Furthermore, more images are added for each category from ImageNet [3]. PASCAL3D+ images exhibit much more variability compared to the existing 3D datasets, and on average there are more than 3,000 object instances per category. We believe this dataset will provide a rich testbed to study 3D detection and pose estimation and will help to significantly push forward research in this area. We provide the results of variations of DPM [6] on our new dataset for object detection and viewpoint estimation in different scenarios, which can be used as baselines for the community. Our benchmark is available online at http://cvgl.stanford.edu/projects/pascal3d

1. Introduction

In the past decade, several datasets have been introduced for classification, detection and segmentation. These datasets provide different levels of annotation for images ranging from object category labels [5, 3] to object bounding box [7, 4, 3] to pixel-level annotations [23, 4, 28]. Although these datasets have had a significant impact on advancing image understanding methods, they have some major limitations. In many applications, a bounding box or segmentation is not enough to describe an object, and we require a richer description for objects in terms of their 3D pose. Since these datasets only provide 2D annotations, they are not suitable for training or evaluating methods that reason about 3D pose of objects, occlusion or depth.

To overcome the limitations of the 2D datasets, 3D datasets have been introduced [22, 20, 25, 8, 19]. However, the current 3D datasets have a number of drawbacks as well. One drawback is that the background clutter is often limited and therefore methods trained on these datasets cannot generalize well to real-world scenarios, where the variability in the background is large. Another drawback is that some of these datasets do not include occluded or truncated objects, which again limits the generalization power of the relevant learnt models. Moreover, the existing datasets typically only provide 3D annotation for a few object classes and the number of images or object instances per category is usually small, which prevents the recognition systems from learning robust models for handling intra-class variations. Finally and most critically, most of these datasets supply annotations for a small number of viewpoints. So they are not suitable for object detection methods aiming at estimating continuous 3D pose, which is a key component in various scene understanding or robotics applications. In summary, it is necessary and important to have a challenging 3D benchmark which overcomes the above limitations.

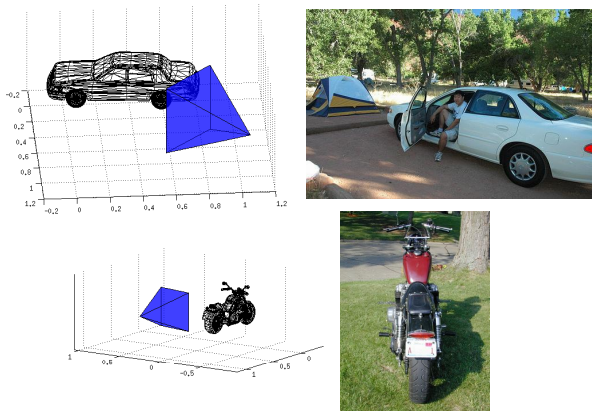


Figure 1. Example of annotations in our dataset. The annotators select a 3D CAD model from a pool of models and align it to the object in the image. Based on the 3D geometry of the model and the annotated 2D locations of a set of landmarks, we automatically compute the azimuth, elevation and distance of the camera (shown in blue) with respect to the object. Images are uncalibrated, so the camera can be at any arbitrary location.

	PASCAL3D+ (ours)	ETH-80 [13]	[26]	3DObject [22]	EPFL Car [20]	[27]	KITTI [8]	NYU Depth [24]	NYC3DCars [19]	IKEA [15]
# of Categories	12	8	2	10	1	4	2	894	1	11
Avg. # Instances per Category	∼3000	10	∼140	10	20	∼660	80,000	39	3,787	∼73
Indoor(I) / Outdoor(O)	Both	I	Both	Both	I	Both	O	I	O	I
Cluttered Background	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓
Non-centered Objects	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓
Occlusion Label	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗
Orientation Label	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Dense Viewpoint	✓	✗	✗	✗	✗	✓	✓	✗	✓	✓

Table 1. Comparison of our PASCAL3D+ dataset with some of the other 3D datasets.

Our contribution in this work is a new dataset, *PASCAL3D+*. Our goal is to overcome the shortcomings of the existing datasets and provide a challenging benchmark for 3D object detection and pose estimation. In PASCAL3D+, we augment the 12 rigid categories in the PASCAL VOC 2012 dataset [4] with 3D annotations. Specifically, for each category, we first download a set of CAD models from Google 3D Warehouse [1], which are selected in such a way that they cover the intra-class variability. Then each object instance in the category is associated with the closest CAD model in term of 3D geometry. Besides, several landmarks of these CAD models are identified in 3D, and the 2D locations of the landmarks are labeled by annotators. Finally, using the 3D-2D correspondences of the landmarks, we compute an accurate continuous 3D pose for each object in the dataset. As a result, the annotation of each object consists of the associated CAD model, 2D landmarks and 3D continuous pose. In order to make our dataset large scale, we add more images from ImageNet [3] for each category. In total, more than 20,000 additional images with 3D annotations are included. Figure 1 shows some examples in our dataset. We also provide baseline results for object detection and pose estimation on our new dataset. The results show that there is still a large room for improvement, and this dataset can serve as a challenging benchmark for future visual recognition systems.

There are several advantages of our dataset: i) PASCAL images exhibit a great amount of variability and better mimic the real-world scenarios. Therefore, our dataset is less biased compared to datasets which are collected in controlled settings (*e.g.*, [22, 20]). ii) Our dataset includes *dense* and *continuous* viewpoint annotations. The existing 3D datasets typically discretize the viewpoint into multiple bins (*e.g.*, [13, 22]). iii) On average, there are more than 3,000 object instances per category. Hence, detectors trained on our dataset can have more generalization power. iv) Our dataset contains occluded and truncated objects, which are usually ignored in the current 3D datasets. v) Finally, PASCAL is the main benchmark for 2D object detection. We hope our efforts on providing 3D annotations to PASCAL can benchmark 2D and 3D object detection methods with a common dataset.

The next section describes the related work and other 3D datasets in the literature. Section 3 provides dataset statistics such as viewpoint distribution and variations in degree of occlusion . Section 4 describes the annotation tool and the challenges for annotating 3D information in an unconstrained setting. Section 5 explains the details of our baseline experiments, and Section 6 concludes the paper.

2. Related Work

We review a number of commonly used datasets for 3D object detection and pose estimation. ETH-80 dataset [13] provides a multi-view dataset of 8 categories (*e.g.*, fruits and animals), where each category contains 10 objects observed from 41 views, spaced equally over the viewing hemisphere. The background is almost constant for all of the images, and the objects are centered in the image. [26] introduces another multi-view dataset that includes *motorbike* and *sport shoe* categories in more challenging real-world scenarios. There are 179 images and 101 images corresponding to each category respectively. On average a motorbike is imaged from 11 views. For shoes, there are about 16 views around each instance taken at 2 different elevations. 3DObject dataset [22] provides 3D annotations for 10 everyday object classes such as *car*, *iron*, and *stapler*. Each category includes 10 instances observed from different viewpoints. EPFL Car dataset [20] consists of 2,299 images of 20 car instances at multiple azimuth angles. The elevation and distance is almost the same for all of these instances. Table-Top-Pose dataset [25] contains 480 images of 3 categories (*mouse*, *mug*, and *stapler*), where each consists of 10 instances under 16 different poses.

These datasets exhibit some major limitations. Firstly, most of them have more or less clean background. Therefore, methods trained on them will not be able to handle cluttered background, which is common in real-world scenarios. Secondly, these datasets only include a limited number of instances, which makes it difficult for recognition methods to learn intra-class variations. To overcome these issues, more challenging datasets have been proposed. ICARO [16] contains viewpoint annotations for 26 object categories. However, the viewpoints are sparse and not densely annotated. [27] provides 3D pose annotations for a subset of 4 categories of the ImageNet dataset [3]: *bed* (400 images), *chair* (770 images), *sofa* (800 images) and *table* (670 images). Since the ImageNet dataset is mainly

designed for the classification task, the objects in the dataset are usually not occluded and they are roughly centered. The KITTI dataset [8] provides 3D labeling for two categories (*car* and *pedestrian*), where there are 80K instances per category. The images of this dataset are limited to street scenes, and all of the images have been obtained by cameras mounted on top of a car. This may pose some issues concerning the ability to generalize to other scene types. More recently, NYC3DCars dataset [19] has been introduced, which contains information such as 3D vehicle annotations, road segmentation and direction of movement. Although the imagery is unconstrained for this dataset in terms of camera type or location, the images are constrained to street scenes of New York. Also, the dataset contains only one category. [15] provides dense 3D annotations for some of the IKEA objects. Their dataset is also limited to indoor images and the number of instances per category is small.

Simultaneous use of 2D information and 3D depth makes the recognition systems more powerful. Therefore, various datasets have been collected by RGB-D sensors (such as Kinect). RGB-D Object Dataset [12] contains 300 physically distinct objects organized into 51 categories. The images are captured in a controlled setting and have a clean background. Berkeley 3-D Object Dataset [11] provides annotation for 849 images of over 50 classes in real office environments. NYU Depth [24] includes 1,449 densely labeled pairs of aligned RGB and depth images. The dataset includes 35,064 distinct instances, which are divided into 894 classes. SUN3D [29] is another dataset of this type, which provides annotations for scenes and objects. There are three limitations for these types of datasets that make them undesirable for 3D object pose estimation: i) They are limited to indoor scenes as the current common RGB-D sensors have a limited range. ii) They do not provide the orientation for objects (they just provide the depth). iii) Their average number of images per category is small.

Our goal for providing a novel dataset is to eliminate the mentioned shortcomings of other datasets, and enhance 3D object detection and pose estimation methods by training and evaluating them on a challenging and real world benchmark. Table 1 shows a comparison between our dataset and some of the most relevant datasets mentioned above.

3. Dataset Details and Statistics

We describe the details of our PASCAL3D+ dataset and provide some statistics. We annotated the 3D pose densely for all of the object instances in the `trainval` subset of PASCAL VOC 2012 detection challenge images (including instances labeled as 'difficult'). We consider the 12 rigid categories of PASCAL VOC, since estimating the pose of the deformable categories is still an open problem. These categories are *aeroplane*, *bicycle*, *boat*, *bottle*, *bus*, *car*, *chair*, *diningtable*, *motorbike*, *sofa*, *train* and *tvmonitor*. In total, there are 13,898 object instances that appear in 8,505 PASCAL images. Furthermore, we downloaded 22,394 images from ImageNet [3] for the 12 categories. For the ImageNet images, the objects are usually centered without occlusion and truncation. On average, there are more than 3,000 instances per category in our PASCAL3D+ dataset.

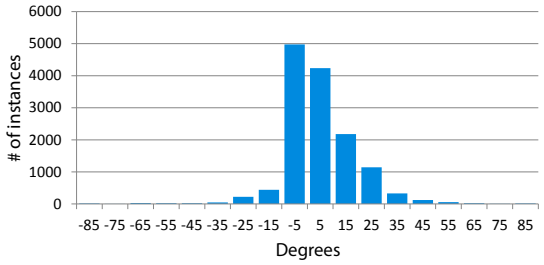


Figure 3. **Elevation distribution.** The distribution of elevation among the PASCAL images across all the categories.

The annotation of an object contains the azimuth, elevation and distance of the camera pose in 3D (we explain how the annotation is obtained in the next section). Moreover, we assign a visibility state to landmarks that we identify for each category: 1) **visible**: the landmark is visible in the image. 2) **self-occluded**: the landmark is not visible due to the 3D geometry and the pose of the object. 3) **occluded-by**: the landmark is occluded by an external object. 4) **truncated**: the landmark appears outside the image area. 5) **unknown**: none of the above four states. To ensure high quality labeling, we hired annotators for the annotation instead of posting the task on crowd-sourcing platforms.

Figure 2 shows the distribution of azimuth among the PASCAL images for the 12 categories, where azimuth 0° corresponds to the frontal view of the object. As expected, the distribution of viewpoints is biased. For example, very few images are taken from the back view (azimuth 180°) of *sofa* since the back of sofa is usually against a wall. For *tvmonitor*, there is also a high bias towards the frontal view. Since *bottles* are usually symmetric, the distribution is dominated by azimuth angles around zero. The distribution of elevation among the PASCAL images across all categories is shown in Figure 3. It is evident that there is large variability in the elevation as well. These statistics show that our dataset has a fairly good distribution in pose variation.

We also analyze the object instances based on their degree of occlusion. The statistics in Figure 4 show that PASCAL3D+ is quite challenging as it includes object instances with different degrees of occlusion. The main goal of most previous 3D datasets was to provide a benchmark for object pose estimation. So they usually ignored occluded or truncated objects. However, handling occlusion and truncation is important for real world applications. Therefore, a challenging dataset like ours can be useful. In Figure 4, we divide the object instances into three classes based on the

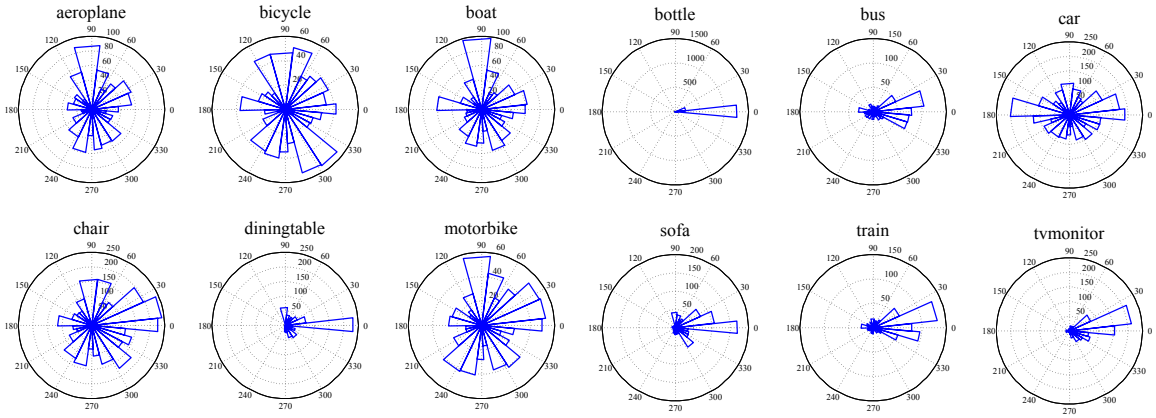


Figure 2. **Azimuth distribution.** Polar histograms show the distribution of azimuth among the PASCAL images for each object category.

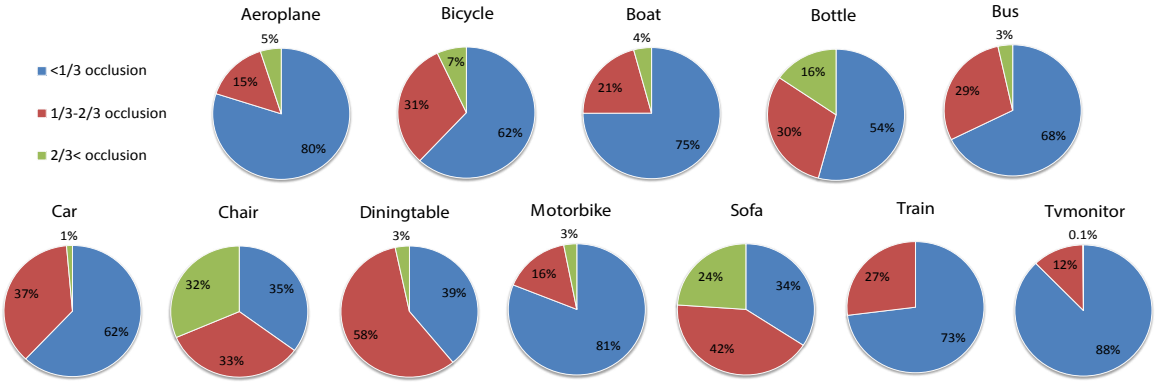


Figure 4. **Occlusion distribution.** The distribution of object instances based on the degree of occlusion in the PASCAL images.

ratio of their externally occluded or truncated landmarks to all landmarks (0 to 1/3, 1/3 to 2/3 and above 2/3). The instances of some categories such as *chair* or *diningtable* are highly occluded, which poses a big challenge to the existing object detection and pose estimation methods.

4. 3D Annotation

Providing 3D annotations for unconstrained images is not trivial since only a single image of a scene is available and the camera parameters are unknown. In this section, we explain the details of our annotation tool and the procedure for 3D annotation labeling.

For each category, we downloaded 3D CAD models from Google 3D Warehouse [1], which is a public repository for 3D CAD models. We select the CAD models in such a way that they represent intra-class variations of a particular category. For example, we select *SUV*, *sedan*, *hatchback*, etc., for the car category. For the aeroplane category, we choose *airliner*, *fighter*, *propeller*, and so on. The 3D CAD models for two example categories are shown in Figure 5. For a sub-category (e.g., propeller aeroplane), more than one CAD model can be selected to better capture the variations in the sub-category.

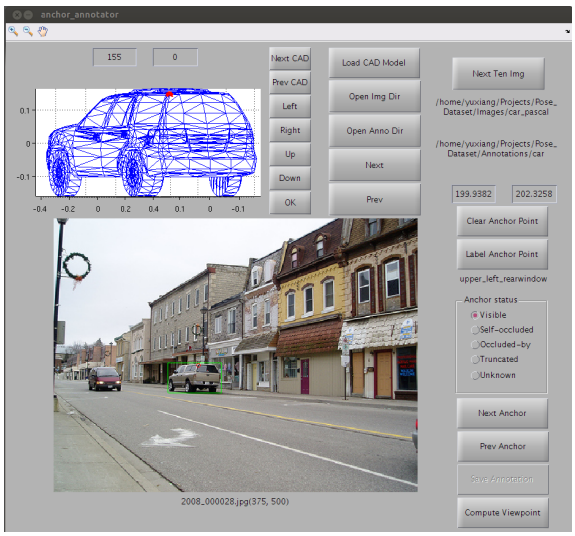
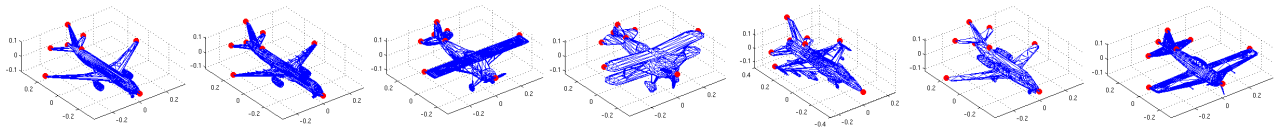
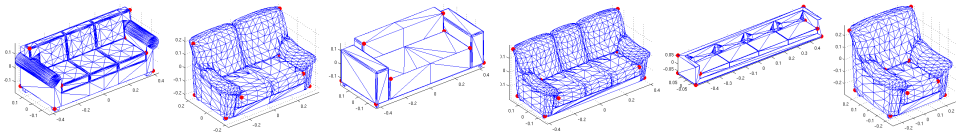


Figure 6. A snapshot of our annotation tool. The blue mesh is the 3D CAD model chosen by the annotator, and the red circle corresponds to one of the landmarks.

For each CAD model, we identify a set of landmarks,



(a) Aeroplane



(b) Sofa

Figure 5. Examples of 3D CAD models used for annotation. To better capture intra-class variability of object categories, different types of CAD models are chosen. The red points represent the identified landmarks.

which are shown with red circles in Figure 5. The landmarks are chosen such that they are shared among instances in a category and can be identified easily in the images. Most of the landmarks correspond to the corners in the CAD models. The task of annotators is to select the closest CAD model for an object instance in terms of 3D geometry and label the landmarks of the CAD model on the 2D image. Then we use these 2D annotations of the landmarks and their corresponding locations on the 3D CAD models to find the azimuth, elevation and distance of the camera pose in 3D for each object instance. A visualization of our annotation tool is shown in Figure 6. The annotator first selects the 3D CAD model that best resembles the object instance. Then, he/she rotates the 3D CAD model until it is aligned with the object instance visually. The alignment provides us with rough azimuth and elevation angles, which are used as initialization in computing the continuous pose. Based on the 3D geometry and the rough pose of the CAD model (after alignment), we compute the visibility of the landmarks. After this step, we show the visible (not self-occluded) landmarks on the 3D CAD model one by one and ask the annotator to mark their corresponding 2D location in the image. For occluded or truncated landmarks, the annotator provides its visibility status as explained in Section 3.

As the result of the annotation, for each object instance in the dataset, we obtain the correspondences between 3D landmarks $\mathbf{X}$ on the CAD model and their 2D projection $\mathbf{x}$ on the image. By using a pinhole camera model, the relationship between the 2D and 3D points is given by: $\mathbf{x}_i = K[R|\mathbf{t}]\mathbf{X}_i$, where K is the intrinsic camera matrix, and R and $\mathbf{t}$ are the rotation matrix and the translation vector respectively. We use a virtual intrinsic camera matrix K, where the focal length is assumed to be 1, the skew is 0 and the aspect ratio is 1. We assume a simplified camera model, where the world coordinate is defined on the 3D CAD model and the camera is facing the origin of the world coordinate system. In this case, R and $\mathbf{t}$ are determined by the azimuth, elevation and distance of the camera pose in 3D. So we can minimize the re-projection error of the 3D landmarks to obtain the continuous pose of the object:

$$\min_{R,\mathbf{t}} \sum_{i=1}^{L} ||\mathbf{x}_i - \tilde{\mathbf{x}}_i||_2, \tag{1}$$

where L is the number of visible landmarks and $\tilde{\mathbf{x}}_i$ is the annotated landmark location in the image. By solving the minimization problem (1), we can find the rotation matrix R and the translation vector $\mathbf{t}$, which provide the azimuth, elevation and distance of the object pose. This is the well-studied Perspective-n-Points (PnP) problem for which various solutions (*e.g.*, [18, 2, 14]) exist. We use the constrained nonlinear optimization implementation of MATLAB to solve (1). For degenerate cases, where there are not enough landmarks visible to compute the pose (less than 2 landmarks), we use the rough azimuth and elevation specified by the annotator instead.

5. Baseline Experiments

In this section, we provide baseline results in terms of object detection, viewpoint estimation and segmentation. We also show that how well the baseline method can handle different degrees of occlusion. For all the experiments below, we use the `train` subset of PASCAL VOC 2012 (detection challenge) for training and the `val` subset for evaluation. We adapt DPM [6] (`voc-release4.01`) to joint object detection and viewpoint estimation.

5.1. Detection and Viewpoint Estimation

The original DPM method uses different mixture components to capture pose and appearance variations of objects. The object instances are assigned to these mixture components based on their aspect ratios. Since the aspect ratio does not necessarily correspond to the viewpoint, viewpoint estimation with the original DPM is impractical. Therefore, we modify DPM similar to [17] such that each mixture component represents a different azimuth section. We refer to this modified version as Viewpoint-DPM (VDPM). In the

	aeroplane	bicycle	boat	bottle	bus	car	chair	diningtable	motorbike	sofa	train	tvmonitor	Avg.
DPM [6]	42.2 / –	49.6 / –	6.0 / –	20.0 / –	54.1 / –	38.3 / –	15.0 / –	9.0 / –	33.1 / –	18.9 / –	36.4 / –	33.2 / –	29.6 / –
VDPM - 4V	40.0 / 34.6	45.2 / 41.7	3.0 / 1.5	– / –	49.3 / 26.1	37.2 / 20.2	11.1 / 6.8	7.2 / 3.1	33.0 / 30.4	6.8 / 5.1	26.4 / 10.7	35.9 / 34.7	26.8 / 19.5
VDPM - 8V	39.8 / 23.4	47.3 / 36.5	5.8 / 1.0	– / –	50.2 / 35.5	37.3 / 23.5	11.4 / 5.8	10.2 / 3.6	36.6 / 25.1	16.0 / 12.5	28.7 / 10.9	36.3 / 27.4	29.9 / 18.7
VDPM - 16V	43.6 / 15.4	46.5 / 18.4	6.2 / 0.5	– / –	54.6 / 46.9	36.6 / 18.1	12.8 / 6.0	7.6 / 2.2	38.5 / 16.1	16.2 / 10.0	31.5 / 22.1	35.6 / 16.3	30.0 / 15.6
VDPM - 24V	42.2 / 8.0	44.4 / 14.3	6.0 / 0.3	– / –	53.7 / 39.2	36.3 / 13.7	12.6 / 4.4	11.1 / 3.6	35.5 / 10.1	17.0 / 8.2	32.6 / 20.0	33.6 / 11.2	29.5 / 12.1
DPM-VOC+VP [21] - 4V	41.5 / 37.4	46.9 / 43.9	0.5 / 0.3	– / –	51.5 / 48.6	45.6 / 36.9	8.7 / 6.1	5.7 / 2.1	34.3 / 31.8	13.3 / 11.8	16.4 / 11.1	32.4 / 32.2	27.0 / 23.8
DPM-VOC+VP [21] - 8V	40.5 / 28.6	48.1 / 40.3	0.5 / 0.2	– / –	51.9 / 38.0	47.6 / 36.6	11.3 / 9.4	5.3 / 2.6	38.3 / 32.0	13.5 / 11.0	21.3 / 9.8	33.1 / 28.6	28.3 / 21.5
DPM-VOC+VP [21] - 16V	38.0 / 15.9	45.6 / 22.9	0.7 / 0.3	– / –	55.3 / 49.0	46.0 / 29.6	10.2 / 6.1	6.2 / 2.3	38.1 / 16.7	11.8 / 7.1	28.5 / 20.2	30.7 / 19.9	28.3 / 17.3
DPM-VOC+VP [21] - 24V	36.0 / 9.7	45.9 / 16.7	5.3 / 2.2	– / –	53.9 / 42.1	42.1 / 24.6	8.0 / 4.2	5.4 / 2.1	34.8 / 10.5	11.0 / 4.1	28.2 / 20.7	27.3 / 12.9	27.1 / 13.6

Table 2. The results of DPM, VDPM and DPM-VOC+VP are shown. The first number indicates the Average Precision (AP) for detection and the second number shows the Average Viewpoint Precision (AVP) for joint object detection and pose estimation.

original DPM, half of the mixture components are mirrored versions of the other half. So the training images are mirrored and assigned to the mirror mixture components. Similarly, we mirror the training images and assign them to the mirrored viewpoint components in VDPM. Another way to perform joint object detection and pose estimation is to treat it as a structure labeling problem. In Pepik et al. [21], they utilize structural SVM to predict the object bounding box and pose jointly, where the model is called DPM-VOC+VP. In our baseline experiments, we divide the azimuth angles into 4, 8, 16 and 24 sections and train VDPM and DPM-VOC+VP models for each case.

To evaluate object detection, we use Average Precision (AP) as the metric and use the standard 50% overlap criteria of PASCAL VOC [4]. For viewpoint estimation, the commonly used metric is the average over the diagonal of the viewpoint confusion matrix [22]. However, this metric only considers the viewpoint accuracy among the correctly detected objects, which makes it non-comparable for two detectors with different sets of detected objects. Since viewpoint estimation is closely related to detection, we need a metric for joint detection and pose estimation. We propose a novel metric called Average Viewpoint Precision (AVP) for this propose similar to AP in object detection. In computing AVP, an output from the detector is considered to be correct if and only if the bounding box overlap is larger than 50% *AND* the viewpoint is correct (i.e., the two viewpoint labels are the same in discrete viewpoint space or the distance between the two viewpoints is smaller than some threshold in continuous viewpoint space). Then we can draw a Viewpoint Precision-Recall (VPR) curve similar to the PR curve. Average viewpoint precision is defined as the area under the VPR curve. Therefore, AVP is the metric for joint detection and pose estimation. Note that detection PR curve is always an upper bound of the VPR curve. Small gap between AVP and AP indicates high viewpoint accuracy among the correctly detected objects.

The results of the original DPM with 6 mixture components, VDPM and DPM-VOC+VP [21] for different azimuth sections are shown in Table 2. Since the instances of the *bottle* category are often symmetric across different azimuth angles, it is ignored in VDPM and DPM-VOC+VP.

	0–1/3	1/3–2/3	2/3–max
aeroplane	57.2	11.5	16.2
bicycle	70.6	30.4	8.7
boat	13.1	0.7	0.9
bus	77.4	35.7	4.1
car	55.3	12.3	3.4
chair	22.0	7.5	0.9
diningtable	33.3	19.9	7.8
motorbike	56.5	12.6	0.1
sofa	35.3	34.2	15.8
train	50.2	35.2	15.3
tvmonitor	58.0	8.1	2.2
Avg.	48.1	18.9	6.8

Table 3. The Normalized Average Precisions from VDPM with 8 views for object detection at different degrees of occlusion.

The detection performance of VDPM is on par with DPM. Compared with VDPM, DPM-VOC+VP achieves better viewpoint estimation in a tradeoff of slightly lower detection performance. For most categories, as we increase the number of viewpoints, the viewpoint estimation task becomes harder and the AVP reduces, which is not surprising. We can see from Table 2 that there is still a large room for improvement both in detection and pose estimation on our dataset. Hence, our 3D annotations can be valuable for developing new 3D object detection methods.

5.2. Sensitivity of Detection to Occlusion

Since our dataset provides occlusion labels for landmarks, we can analyze the performance of detection at different degrees of occlusion. The occlusion of landmarks does not directly determine the degree of occlusion of the object, but it has a strong correlation with it. For example, all landmarks can be occluded while most of the object can be observed, but such a case does not happen in reality. Therefore, we use the ratio of externally occluded or truncated landmarks to all landmarks as a measure for the degree of occlusion. We refer to it as the "occlusion ratio". In this experiment, we analyze the detection performance of VDPM with 8 views in terms of different degree of occlusion. We partition the instances into three occlusion sets, i.e., the set with occlusion ratio between 0 and 1/3, the set with occlusion ratio between 1/3 and 2/3, and the set with occlusion ratio larger than 2/3. Since the number of instances in each occlusion set is different, we report Normalized Average Precision in Table 3 as suggested by [10].

	aeroplane	bicycle	boat	bottle	bus	car	chair	diningtable	motorbike	sofa	train	tvmonitor	Avg.
GT CAD	43.8	28.7	43.0	66.0	78.4	67.3	41.8	28.0	60.0	40.3	59.2	72.3	52.4
Random CAD	32.8± 0.3	29.2± 0.5	28.7± 1.1	62.5± 1.0	67.2± 0.8	61.8± 0.5	35.8± 0.8	21.3± 0.6	54.6± 0.3	34.7± 0.5	53.8± 0.6	60.5± 2.8	45.2
VDPM - 4 views	22.6	16.1	23.4	–	50.7	51.2	25.7	12.4	34.4	27.3	35.1	56.6	32.3
VDPM - 8 views	24.1	16.6	23.5	–	52.7	51.2	27.6	10.8	35.7	29.4	40.2	55.0	33.3
VDPM - 16 views	24.7	16.6	23.5	–	57.8	51.9	26.5	10.1	37.9	29.5	40.2	55.9	34.1
VDPM - 24 views	24.5	16.9	20.5	–	57.1	50.9	27.2	11.5	37.3	27.6	39.8	54.7	33.5

Table 4. Segmentation accuracy obtained by projecting the 3D CAD models onto the images. Please refer to the text for more details.

It is evident that the detectors have difficulty in handling highly occluded objects. In order to achieve good performance in detection and pose estimation on our dataset, it is important to handle the occluded and truncated objects. Our dataset enables evaluation of occlusion reasoning as well.

5.3. Segmentation using 3D Pose

We show that estimating the viewpoint with the corresponding CAD model for an object enables object segmentation. To find the upper bound for segmentation in this way, we project the ground truth CAD model (the one that the annotator selected for the object instance) onto the image using the ground truth azimuth, elevation and distance. To evaluate the segmentation, we use the annotations provided by [9]. The first row of Table 4 shows the segmentation accuracy using the ground truth poses, where we use the standard PASCAL evaluation metric for segmentation. The accuracy is not 100% due to several reasons. First, we do not consider occlusion reasoning in the projection, and the ground truth mask from [9] is just for the visible part of the object. Second, due to the simplified camera model in computing the continuous pose and the limited number of CAD models in our dataset, the projection matrix we use is an approximation to the real one. So we also include the re-projection error in our 3D annotation, which can be considered to be a measure for the quality of the annotation. Figure 7 shows segmentation examples for each category in our dataset using the ground truth pose. As an example of the re-projection error, the predicted legs of the diningtable are not precisely aligned with the object in the image, which results in a large penalty in the computing the segmentation accuracy. For the chairs, a large penalty is introduced due to occlusion. Occlusion reasoning is also important for segmentation.

To show the importance of using the right CAD model for annotation, instead of projecting the ground truth CAD model, we project a randomly chosen model (from the set of CAD models for a particular category) and evaluate the segmentation performance. As shown in the second row of Table 4, the average accuracy drops by about 7%. The shown accuracy is the average over 5 different random selections. Note that the performance for *bicycle* with random models is higher than the case with the ground truth models. This is due to the inaccuracy in 2D segmentation annotation of bicycle. In most cases, the areas that correspond to the background are labeled as bicycle (e.g., around the spokes).

We also evaluate how well the automatic approaches can perform segmentation. In this experiment, we infer the azimuth automatically from VDPMs, but use the ground truth elevation, distance and CAD model in the projection. More specifically, for each detected object, we project the CAD model to the image. We consider an object as detected if there is a bounding box with more than 50% intersection over union overlap associated with it. The performance drops significantly for the automatic approach. Note that the segmentation performance becomes better as we use finer discretization of azimuth (with the exception of 24 viewpoints). The low performance with 24 views might be due to the low performance of VDPM in viewpoint estimation for 24 views as shown in Table 2.

6. Conclusion

To further improve the development of 3D object detection and pose estimation methods, we provide a large scale benchmark PASCAL3D+ with 3D annotations of objects. PASCAL3D+ overcomes the limitations of the existing 3D datasets and better matches real-world scenarios. We developed an algorithm and annotation tool to provide the continuous 3D viewpoint annotations in unconstrained settings, where the camera parameters are unknown and only a single image of object instances is available. We also provide baseline results for object detection, viewpoint estimation and segmentation on our PASCAL3D+ dataset. The results illustrate that there is still a large room for improvement in all these tasks. We hope our dataset can push forward the research in 3D object detection and pose estimation.

Acknowledgments

We acknowledge the support of ONR grant N00014-13-1-0761 and NSF CAREER grant #1054127. We thank Taewon Kim, Yawei Wang and Jino Kim for their valuable help in building this benchmark. We thank Bojan Pepik for his help in conducting the experiments with DPM-VOC+VP.

References

[1] Google 3D Warehouse. `http://sketchup.google.com/3dwarehouse`.

[2] A. Ansar and K. Daniilidis. Linear pose estimation from points or lines. In *ECCV*, 2002.

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

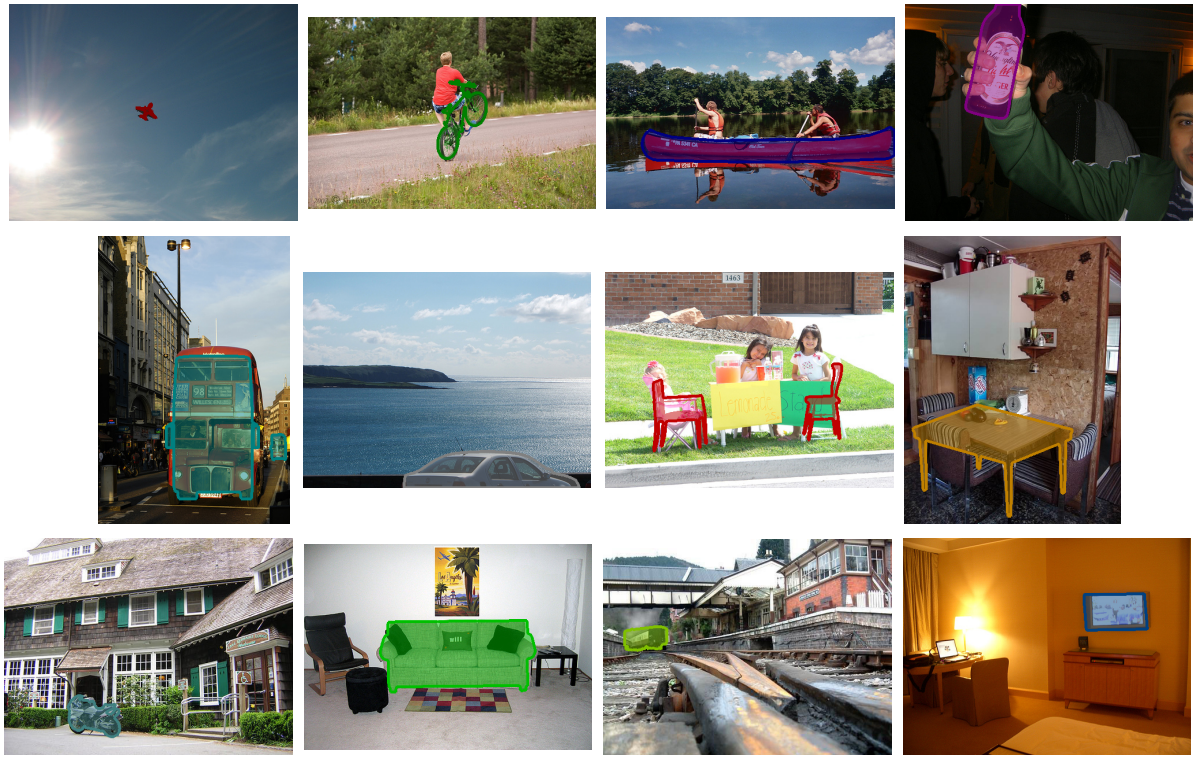


Figure 7. Segmentation results obtained by projecting the 3D CAD models to the images. Each figure shows an example for one of the 12 categories in our dataset.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*, 2004.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[7] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *IJCV*, 2009.

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[9] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[10] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012.

[11] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *ICCV Workshop on Consumer Depth Cameras in Computer Vision*, 2011.

[12] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011.

[13] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*, 2003.

[14] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *IJCV*, 2009.

[15] J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013.

[16] R. J. Lopez-Sastre, C. Redondo-Cabrera, P. Gil-Jimenez, and S. Maldonado-Bascon. ICARO: Image Collection of Annotated Real-world Objects. `http://agamenon.tsc.uah.es/Personales/rlopez/data/icaro`, 2010.

[17] R. J. Lopez-Sastre, T. Tuytelaars, and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *ICCV Workshop on Challenges and Opportunities in Robot Perception*, 2011.

[18] C. P. Lu, G. D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *PAMI*, 2000.

[19] K. Matzen and N. Snavely. Nyc3dcars: A dataset of 3d vehicles in geographic context. In *ICCV*, 2013.

[20] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[21] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[22] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, 2007.

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[25] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese. Depth-encoded hough voting for coherent object detection, pose estimation, and shape recovery. In *ECCV*, 2010.

[26] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool. Towards multi-view object class detection. In *CVPR*, 2006.

[27] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012.

[28] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[29] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013.