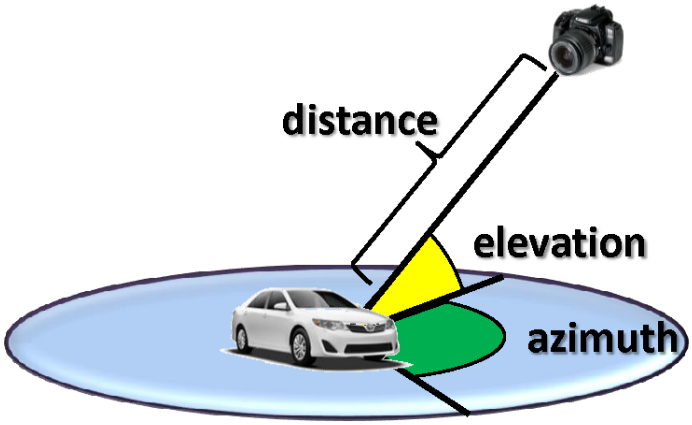
Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild

INTRODUCTION



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 [1] with 3D annotations. Furthermore, more images are added for each category from ImageNet [2]. 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 [3] on our new dataset for object detection and viewpoint estimation in different scenarios, which can be used as baselines for the community.

PUBLICATION

- Yu Xiang, Roozbeh Mottaghi and Silvio Savarese. Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild. In IEEE Winter Conference on Applications of Computer Vision (WACV), 2014. bibtex, pdf

PASCAL3D+

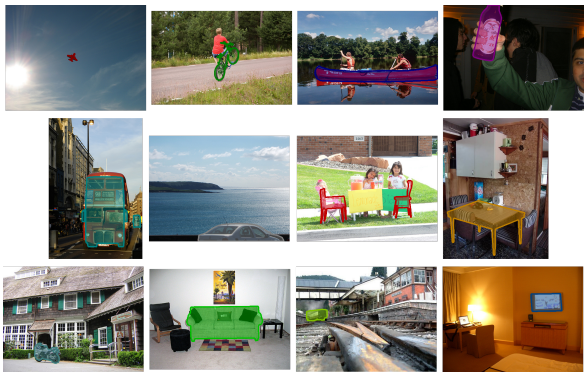
- release1.1 ~ 7.5GB (PASCAL VOC 2012 and ImageNet images and annotations, 3D CAD models, annotation tool, VDPM code, and segmentation code)
- release1.0 ~ 1GB (PASCAL VOC 2012 train and validation images and annotations, 3D CAD models, and annotation tool)

NOTE ON 3D OBJECT RECONSTRUCTION

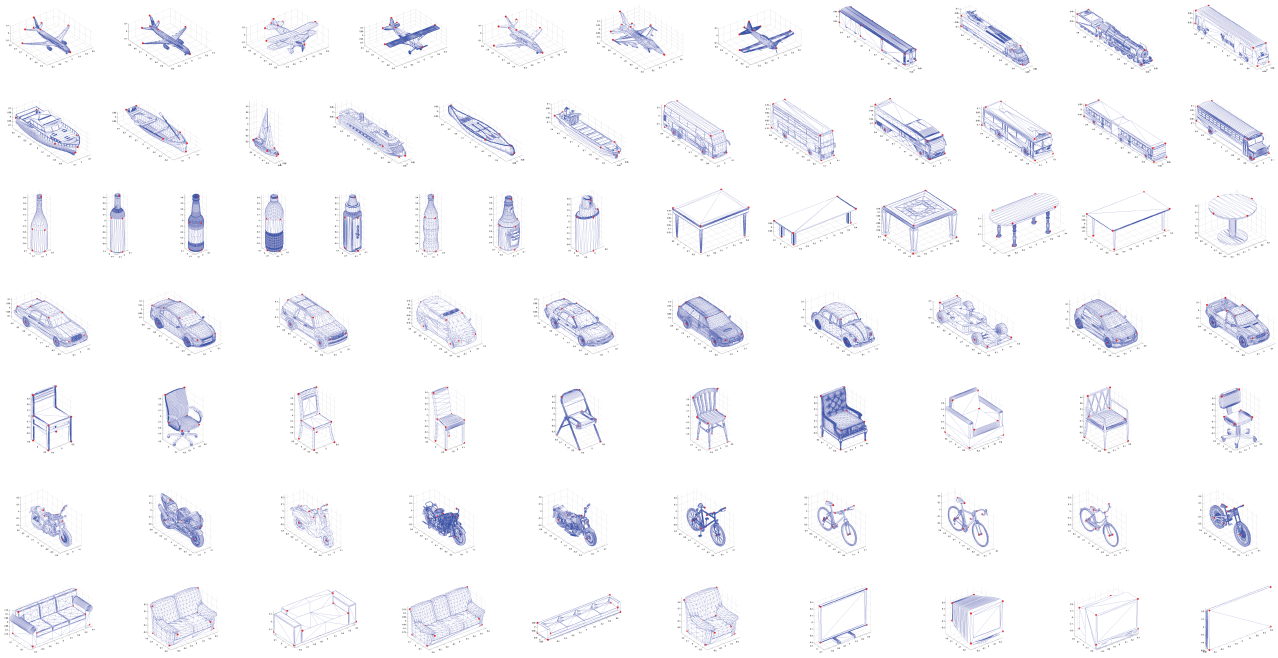
- When PASCAL3D+ is used for benchmarking 3D object reconstruction, we do NOT suggest using the 3D CAD models in PASCAL3D+ for training, since the same set of 3D CAD models is used to annotate the test set. Using the 3D CAD models in both training and testing for 3D reconstruction will be biased. Please see more detailed discussion in the Appendix of [9].

ANNOTATION PROCESS

CAD alignments



CAD MODELS



OBJECT DETECTION EVALUATION

- We use Average Precision (AP) as the metric to evalute object detection, where the standard 50% overlap criteria of PASCAL VOC [1] is applied.
- VDPM: modified version of DPM, where mixture components correspond to discretized viewpoints. VDPM is trained on the PASCAL VOC 2012 train set, and tested on the PASCAL VOC 2012 val set.
- DPM-VOC+VP [4] is trained on the PASCAL VOC 2012 train set, and tested on the PASCAL VOC 2012 val set.
- RCNN [5] is trained on the PASCAL VOC 2012 train set by fine-tuning a pre-trained CNN on ImageNet images, and tested on the PASCAL VOC 2012 val set.

Method	aeroplane	bicycle	boat	bottle	bus	car	chair	diningtable	motorbike	sofa	train	tvmonitor	Average
DPM [3]	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 - 4 Views	40.0	45.2	3.0	--	49.3	37.2	11.1	7.2	33.0	6.8	26.4	35.9	26.8
VDPM - 8 Views	39.8	47.3	5.8	--	50.2	37.3	11.4	10.2	36.6	16.0	28.7	36.3	29.9
VDPM - 16 Views	43.6	46.5	6.2	--	54.6	36.6	12.8	7.6	38.5	16.2	31.5	35.6	30.0
VDPM - 24 Views	42.2	44.4	6.0	--	53.7	36.3	12.6	11.1	35.5	17.0	32.6	33.6	29.5
DPM-VOC+VP[4] - 4 Views	41.5	46.9	0.5	--	51.5	45.6	8.7	5.7	34.3	13.3	16.4	32.4	27.0
DPM-VOC+VP[4] - 8 Views	40.5	48.1	0.5	--	51.9	47.6	11.3	5.3	38.3	13.5	21.3	33.1	28.3
DPM-VOC+VP[4] - 16 Views	38.0	45.6	0.7	--	55.3	46.0	10.2	6.2	38.1	11.8	28.5	30.7	28.3

DPM-VOC+VP[4] - 24 Views	36.0	45.9	5.3	--	53.9	42.1	8.0	5.4	34.8	11.0	28.2	27.3	27.1
RCNN [5]	72.4	68.7	34.0	--	73.0	62.3	33.0	35.2	70.7	49.6	70.1	57.2	56.9
[7] - 4 Views	78.1	72.4	51.2	--	78.0	63.9	26.2	45.8	76.9	51.7	77.1	65.4	62.4
[7] - 8 Views	76.8	72.7	52.1	--	79.0	65.5	24.7	45.4	76.2	52.5	76.3	66.1	62.5
[7] - 16 Views	76.0	71.2	51.6	--	77.8	63.4	24.2	44.6	75.6	49.4	74.8	63.0	61.0
[7] - 24 Views	77.1	70.4	51.0	--	77.4	63.0	24.7	44.6	76.9	51.9	76.2	64.6	61.6
[8] - 4 Views	77.6	71.8	47.6	--	75.9	60.9	17.2	54.9	75.9	48.7	77.2	63.6	61.0
[8] - 8 Views	79.3	69.3	43.7	--	76.7	57.7	17.9	54.8	73.8	51.6	78.0	61.1	60.4
[8] - 16 Views	75.3	71.9	44.4	--	76.2	59.9	15.9	51.9	75.5	50.0	76.9	62.2	60.0
[8] - 24 Views	76.6	67.7	42.7	--	76.1	59.7	15.5	51.7	73.6	50.6	77.7	60.7	59.3

OBJECT DETECTION AND POSE ESTIMATION EVALUATION

- We propose a new metric called Average Viewpoint Precision (AVP) to evalute object detection and pose estimation joinly similar to AP in object detection. In computing the precision of 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). The recall is the same for AP and AVP. As a result, AP is always an upper bound of AVP.
- VDPM: modified version of DPM, where mixture components correspond to discretized viewpoints. VDPM is trained on the PASCAL VOC 2012 train set, and tested on the PASCAL VOC 2012 val set.
- DPM-VOC+VP [4] is trained on the PASCAL VOC 2012 train set, and tested on the PASCAL VOC 2012 val set.
- Viewpoints & Keypoints [6] is trained on the PASCAL VOC 2012 train set and the ImageNet images in PASCAL3D+, and tested on the PASCAL VOC 2012 val set with object detections from RCNN [5].

Method	aeroplane	bicycle	boat	bottle	bus	car	chair	diningtable	motorbike	sofa	train	tvmonitor	Average
VDPM - 4 Views	34.6	41.7	1.5	--	26.1	20.2	6.8	3.1	30.4	5.1	10.7	34.7	19.5
VDPM - 8 Views	23.4	36.5	1.0	--	35.5	23.5	5.8	3.6	25.1	12.5	10.9	27.4	18.7
VDPM - 16 Views	15.4	18.4	0.5	--	46.9	18.1	6.0	2.2	16.1	10.0	22.1	16.3	15.6
VDPM - 24 Views	8.0	14.3	0.3	--	39.2	13.7	4.4	3.6	10.1	8.2	20.0	11.2	12.1
DPM-VOC+VP[4] - 4 Views	37.4	43.9	0.3	--	48.6	36.9	6.1	2.1	31.8	11.8	11.1	32.2	23.8
DPM-VOC+VP[4] - 8 Views	28.6	40.3	0.2	--	38.0	36.6	9.4	2.6	32.0	11.0	9.8	28.6	21.5
DPM-VOC+VP[4] - 16 Views	15.9	22.9	0.3	--	49.0	29.6	6.1	2.3	16.7	7.1	20.2	19.9	17.3
DPM-VOC+VP[4] - 24 Views	9.7	16.7	2.2	--	42.1	24.6	4.2	2.1	10.5	4.1	20.7	12.9	13.6
Viewpoints&Keypoints [6] - 4 Views	63.1	59.4	23.0	--	69.8	55.2	25.1	24.3	61.1	43.8	59.4	55.4	49.1
Viewpoints&Keypoints [6] - 8 Views	57.5	54.8	18.9	--	59.4	51.5	24.7	20.5	59.5	43.7	53.3	45.6	44.5
Viewpoints&Keypoints [6] - 16 Views	46.6	42.0	12.7	--	64.6	42.7	20.8	18.5	38.8	33.5	42.5	32.9	36.0
Viewpoints&Keypoints [6] - 24 Views	37.0	33.4	10.0	--	54.1	40.0	17.5	19.9	34.3	28.9	43.9	22.7	31.1
[7] - 4 Views	70.3	67.0	36.7	--	75.4	58.3	21.4	34.5	71.5	46.0	64.3	63.4	55.4
[7] - 8 Views	66.0	62.5	31.1	--	68.7	55.7	19.2	31.9	64.0	44.7	61.8	58.0	51.3
[7] - 16 Views	51.4	43.0	23.6	--	68.9	46.3	15.2	29.3	49.4	35.6	47.0	37.3	40.6
[7] - 24 Views	43.2	39.4	16.8	--	61.0	44.2	13.5	29.4	37.5	33.5	46.6	32.5	36.1
[8] - 4 Views	64.6	62.1	26.8	--	70.0	51.4	11.3	40.7	62.7	40.6	65.9	61.3	50.7
[8] - 8 Views	58.7	56.4	19.9	--	62.4	45.2	10.6	34.7	58.6	38.8	61.2	49.7	45.1
[8] - 16 Views	46.1	39.6	13.6	--	56.0	36.8	6.4	23.5	41.8	27.0	38.8	36.4	33.3
[8] - 24 Views	33.4	29.4	9.2	--	54.7	35.7	5.5	23.0	30.3	27.6	44.1	34.3	28.8

ADD YOUR RESULTS

- Send your detection and pose estimation results to < yuxiang at cs dot stanford dot edu >.

- Ideal format: for each test image, a set of detected bounding boxes with detection scores and viewpoints.

RELATED DATASETS

- 3D Object Dataset: a benchmark for object detection and pose estimation (10 categories with 10 object instances for each category).
- EPFL Car Dataset: a multi-view car dataset for pose estimation (20 car instances).
- KITTI Detection Dataset: a street scene dataset for object detection and pose estimation (3 categories: car, pedestrian and cyclist).
- PASCAL VOC Detection Dataset: a benchmark for 2D object detection (20 categories).
- SUN3D: a database of big spaces reconstructed using SfM and object labels.
- LabelMe3D: a database of 3D scenes from user annotations.
- NYC3DCars: a database of 3D vehicles in geographic context.

ACKNOWLEDGEMENTS

- 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. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
2. J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.
3. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. PAMI, 2010.
4. B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In CVPR, 2012.
5. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
6. S. Tulsiani and J. Malik. Viewpoints and keypoints. In CVPR, 2015.
7. F. Massa, R. Marlet and M. Aubry. Crafting a multi-task CNN for viewpoint estimation. In BMVC, 2016.
8. P. Poirson, P. Ammirato, C.Y. Fu, W. Liu, J. Kosecka and A. Berg. Fast Single Shot Detection and Pose Estimation. In 3DV, 2016.
9. S. Tulsiani, T. Zhou, A.A. Efros and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In CVPR, 2017.

Contact : yuxiang at cs dot stanford dot edu

Last update : 6/29/2017