Pose from Shape: Deep Pose Estimation for Arbitrary 3D Objects

Yang Xiao
https://youngxiao13.github.io
Xuchong Qiu
http://imagine.enpc.fr/~qiux/
Pierre-Alain Langlois
http://imagine.enpc.fr/~langloip/
Mathieu Aubry
http://imagine.enpc.fr/~aubrym/
Renaud Marlet
http://imagine.enpc.fr/~marletr/

LIGM (UMR 8049)
Ecole des Ponts ParisTech, UPE
Champs-sur-Marne, France

A Datasets

Pascal3D+ [12] provides images with 3D annotations for 12 object categories. The images are selected from the training and validation set of PASCAL VOC 2012 [4] and ImageNet [3], with 2k to 4k images in the wild per category. An approximate 3D CAD model is provided for each object as well as its 3D orientation in the image. Following the protocol of [5, 8, 11], we use the ImageNet-trainval and Pascal-train images as training data, and the 2,113 non-occluded and non-truncated objects of the Pascal-val images as testing data. As in [11], we use the metric $Acc_{\frac{\pi}{6}}$, which measures the percentage of test samples having a pose prediction error smaller than $\frac{\pi}{6}$: $\Delta(R_{\text{pred}}, R_{\text{gt}}) = \|\log(R_{\text{pred}}^T R_{\text{gt}})\|_{\mathcal{F}}/\sqrt{2} < \frac{\pi}{6}$.

ObjectNet3D [13] is a large-scale 3D dataset similar to Pascal3D+ but with 100 categories, which provide a wider variety of shapes. To verify the generalization power of our method for unknown categories, we follow the protocol of StarMap [15]: we evenly hold out 20 categories (every 5 categories sorted in the alphabetical order) from the training data and only used them for testing. For a fair comparison, we actually use the same subset of training data as in [13] (also containing keypoint annotations) and evaluate on the non-occluded and non-truncated images of the 20 categories, using the same $Acc_{\frac{\pi}{6}}$ metric.

Pix3D [10] is a recent dataset containing 5,711 non-occluded and non-truncated images of 395 CAD shapes among 9 categories. It mainly features furniture, with a strong bias towards chairs. But contrary to Pascal3D+ and ObjectNet3D, that only feature approximate models and rough alignments, Pix3D provides exact models and pixel-level accurate poses. Similar to the training paradigm of [9, 10], we train on ShapeNetCore [2] with input images made of

rendered views on random SUN397 backgrounds [14] using random texture maps included in ShapeNetCore, and test on Pix3D real images and shapes.

ShapeNetCore is a subset of ShapeNet [2] containing 51k single clean 3D models, covering 55 common object categories of man-made artifacts. We exclude the categories containing mostly objects with rotational symmetry or small and narrow objects, which results in 30 remaining categories: *airplane, bag, bathtub, bed, birdhouse, bookshelf, bus, cabinet, camera, car, chair, clock, dishwasher, display, faucet, lamp, laptop, speaker, mailbox, microwave, motorcycle, piano, pistol, printer, rifle, sofa, table, train, watercraft* and *washer*. We randomly choose 200 models from each category and use Blender to render each model under 20 random views with various textures included in ShapeNetCore.

LINEMOD [6] has become a standard benchmark for 6D pose estimation of textureless objects in cluttered scenes. It consists of 15 sequences featuring one object instance for each sequence to detect with ground truth 6D pose and object class. As other authors, we left out categories bowl and cup, that have a rotational symmetry, and consider only 13 classes. The common evaluation measure with LINEMOD is the ADD-0.1d metric [6]: a pose is considered correct if the average of the 3D distances between transformed object vertices by the ground truth transformation and the ones by estimated transformation is less than 10% of the object's diameter. For the objects with ambiguous poses due to symmetries, [6] replaces this measure by ADD-S which is specially tailored for symmetric objects. We choose ADD-0.1d and ADD-S-0.1d as our evaluation metrics.

B Evaluation Metrics

For results on LINEMOD, the ADD [6] metric is used to compute the averaged distance between points transformed using the estimated pose and the ground truth pose:

$$\textbf{ADD} = \frac{1}{m} \sum_{\mathbf{x} \in \mathcal{M}} ||(\mathbf{R}\mathbf{x} + \mathbf{t}) - (\hat{\mathbf{R}}\mathbf{x} + \hat{\mathbf{t}})|| \tag{1}$$

where m is the number of points on the 3D object model, $\mathcal{M}$ is the set of all 3D points of this model, $\mathbf{p} = [\mathbf{R}|\mathbf{t}]$ is the ground truth pose and $\hat{\mathbf{p}} = [\hat{\mathbf{R}}|\hat{\mathbf{t}}]$ is the estimated pose. Following [1], we compute the model diameter d as the maximum distance between all pairs of points from the model. With this metric, a pose estimation is considered to be correct if the computed averaged distance is within 10% of the model diameter d.

For the objects with ambiguous poses due to symmetries, [6] replaces this measure by ADD-S, which uses the closet point distance in computing the average distance for 6D pose evaluation as in:

$$\textbf{ADD-S} = \frac{1}{m} \sum_{\mathbf{x}_1 \in \mathcal{M}} \min_{\mathbf{x}_2 \in \mathcal{M}} ||(\mathbf{R}\mathbf{x}_1 + \mathbf{t}) - (\hat{\mathbf{R}}\mathbf{x}_2 + \hat{\mathbf{t}})|| \tag{2}$$

C Ablation and parameter study

Ablation and parameter study on the number of rendered images. Table 1 shows the experimental results of pose estimation on 20 novel categories of ObjectNet3D for different numbers and layouts of rendered images. The viewpoints are sampled evenly at N_{azi}

$N_{azi} \times N_{ele}$	0	1×1	6×1	3×2	2×3	12×1	6×2	4×3	18×1	9×2	6×3
$Acc\frac{\pi}{6} \uparrow$	50	56	59	60	58	59	**62**	58	58	60	59
$MedErr \downarrow$	50	45	44	44	51	46	**40**	46	51	43	45

Table 1: Ablation and parameter study on ObjectNet3D of the number and layout of rendering images at the input of the network when using multiple views to represent shape. Performance depending on the number of azimuthal and elevation samples.

Randomization Range	$[-0°, 0°]$	$[-45°, 45°]$	$[-90°, 90°]$	$[-180°, 180°]$
$Acc\frac{\pi}{6} \uparrow$	56	**62**	60	55
$MedErr \downarrow$	47	**40**	43	52

Table 2: Parameter study of azimuthal randomization used as a specific data augmentation of our approach. Performance depending on the range of azimuthal variation during training.

azimuths and elevated at N_{ele} different elevations. $N_{ele} = 1, 2, 3$ represents respectively elevations at $(30°)$, $(0°, 30°)$, $(0°, 30°, 60°)$. The $Acc\frac{\pi}{6}$ metric measures the percentage of testing samples with a angular error smaller than $\frac{\pi}{6}$ and $MedErr$ is the median angular error (°) over all testing samples.

The table shows that using shape information encoded from rendered images (when $N_{azi} \times N_{ele} > 0$) can indeed help pose estimation on novel categories, i.e., that are not included in the training data. In the first column (0 rendered images) we show the performance of our baseline without using the 3D shape of the object, compared to this result, the network trained with only one rendered image has a clearly boosted accuracy.

The table also shows that more rendered images in the network input does not necessarily mean a better performance. In the table, the network trained with 12 rendered images elevated at 0°and 30°gives the best result. This may be because the ObjectNet3D dataset is highly biased towards low elevations on the hemisphere, which can be well represented without using the rendered image captured at high elevation such as 60°.

Parameter study on the azimuthal randomization strategy. Table 2 summarizes the parameter study on the range of azimuthal jittering applied to input shapes during network training. The poor results obtained for $[-0°, 0°]$ and $[-180°, 180°]$ are due the objects with symmetries, typically at 90°or 180°.

D Qualitative Results on LINEMOD

Some qualitative results for 13 LINEMOD objects are shown in Figure 1. Given object image and its shape, our approach gives a coarse pose estimate which is then refined by pose refinement method given by DeepIM [7].

References

[1] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Figure 1: Visual results of object pose estimation on LINEMOD [6]. For each sample, the four columns from left to right represent: the input image, the correct shape and orientation, our initial estimate and the final estimate after refining our initialization with DeepIM [7].

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University – Princeton University – Toyota Technological Institute at Chicago, 2015.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[5] Alexander Grabner, Peter M. Roth, and Vincent Lepetit. 3D pose estimation and 3D model retrieval for objects in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision (ACCV)*, 2012.

[7] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *European Conference on Computer Vision (ECCV)*, 2018.

[8] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3D bounding box estimation using deep learning and geometry. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *International Conference on Computer Vision (ICCV)*, 2015.

[10] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3d shape modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[13] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. ObjectNet3D: A large scale database for 3D object recognition. In *European Conference Computer Vision (ECCV)*, 2016.

[14] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[15] Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *European Conference on Computer Vision (ECCV)*, 2018.