3D Pose Estimation and 3D Model Retrieval for Objects in the Wild Supplementary Material

Alexander Grabner[1] Peter M. Roth[1] Vincent Lepetit[2,1]

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria

[2]Laboratoire Bordelais de Recherche en Informatique, University of Bordeaux, France

{alexander.grabner,pmroth,lepetit}@icg.tugraz.at

In the following, we provide additional qualitative results for our 3D model retrieval approach in Sec. 1, which complement those presented in the paper. Furthermore, we analyze failure cases for both 3D model retrieval and the underlying 3D pose estimation in Sec. 2. Finally, in Sec. 3 we discuss implementation details, parameter choices, and other relevant settings.

1. 3D Model Retrieval

Fig. 2 shows additional qualitative results for 3D model retrieval from ShapeNet [1] given previously unseen images from Pascal3D+ [9] validation data for all twelve categories. Our approach predicts accurate 3D poses and 3D models for objects of different categories.

Fig. 1 presents further 3D model alignment results for object detections which are not fully accurate. We significantly improve the alignment between the object in the image and an RGB rendering of our retrieved 3D model by taking advantage of our predicted 6-DoF pose and 3-DoF dimensions compared to just using a 3-DoF viewpoint.

2. Failure Modes

Most failure cases of our 3D pose estimation on Pascal3D+ relate to low-resolution or ambiguous objects.

Fig. 3 shows 3D pose estimation results on low-resolution image windows from Pascal3D+ validation data. After re-scaling, the over-smoothed input RGB images lack details and sharp discontinuities, which results in incorrect pose predictions. In fact, even for a human it is difficult to identify the correct object poses in these examples.

Fig. 4 shows additional failure cases, observing that heavy occlusions, bad illumination conditions and difficult object poses, which are far from the poses seen during training, result in incorrect pose predictions.

As shown in Fig. 5, some objects from Pascal3D+ are symmetrical, which makes their poses not well defined. For example, it is impossible to differentiate between the front and back of a symmetric unmanned boat. This issue is even more apparent for tables: Many tables are ambiguous with respect to an azimuth rotation of π, $\frac{\pi}{2}$ or even have an axis of symmetry, such as a round table. When our approach predicts one of the possible poses that is not the annotated ground truth pose, this is considered as a mistake by the commonly used evaluation protocol [8].

Fig. 6 shows that visual distortions due to wide-angle lenses (*i.e.*, fish-eye effects), deformed and demolished objects and heavy occlusions can disturb the model retrieval step, even if the pose estimation was successful.

3. Implementation Details

In the following, we provide implementation details and other parameters used in our work:

Intrinsic camera parameters: In Pascal3D+, the ground truth poses were computed from 2D-3D correspondences assuming the same intrinsic parameters for all images. We employ the same parameters in our approach.

Data augmentation: Like others [4, 5, 7, 8], we perform data augmentation by jittering ground truth detections and exclude detections marked as occluded or truncated from the evaluation. Additionally, we augment samples for which the longer edge of the ground truth image window is greater than 224 pixel by applying Gaussian blurring with various kernel sizes and σ. We randomly sample negative example 3D models from the available data. All augmentation parameters are randomized after each training epoch.

Meta parameters: We normalize the projections so that the image pixel range is mapped to the interval [0,1] and use the same Huber loss ($\delta = 0.01$) for all 19 estimated values. Experimentally, we found $\alpha = 1$, $\beta = 1e^{-5}$ and $\gamma = 1e^{-3}$ to work well and set $m = 1$.

Network parameters: We use a batch size of 50, train our networks for 100 epochs and decrease the initial learning rate of $1e^{-4}$ by one order of magnitude after 50 and 90 epochs, and employ the Adam optimization algorithm.

3D dimensions: For both Pascal3D+ and ShapeNet, 3D models are normalized to fit within a unit cube centered at the origin. Thus, we estimate 3D dimensions in model space in the range [0,1]. Since these dimensions tend to be consistent within a category, estimating them is not a major issue. Table 1 shows quantitative results for 3D dimension estimation. We achieve high accuracy across all categories.

	x	y	z
Median Absolute Error	0.022	0.015	0.014

Table 1: 3D dimension estimation errors on Pascal3D+. We report the mean performance across all categories.

Figure 1: We use our predicted 6-DoF pose and 3-DoF dimensions to refine the alignment between the object and a rendering. Left: A detected object, which is not centered on the image window. Middle: A rendering which just uses our predicted 3-DoF rotation. Right: A rendering which uses our predicted 6-DoF pose and 3-DoF dimensions.

References

[1] A. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An Information-Rich 3D Model Repository. Technical report, Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[3] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*, pages 630–645, 2016. 4

[4] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3D Bounding Box Estimation Using Deep Learning and Geometry. In *Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 1

[5] G. Pavlakos, X. Zhou, A. Chan, K. Derpanis, and K. Daniilidis. 6-DoF Object Pose from Semantic Keypoints. In *International Conference on Robotics and Automation*, pages 2011–2018, 2017. 1

[6] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2014. 4

[7] H. Su, C. Qi, Y. Li, and L. Guibas. Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views. In *International Conference on Computer Vision*, pages 2686–2694, 2015. 1

[8] S. Tulsiani and J. Malik. Viewpoints and Keypoints. In *Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 1

[9] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond Pascal: A Benchmark for 3D Object Detection in the Wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, 2014. 1

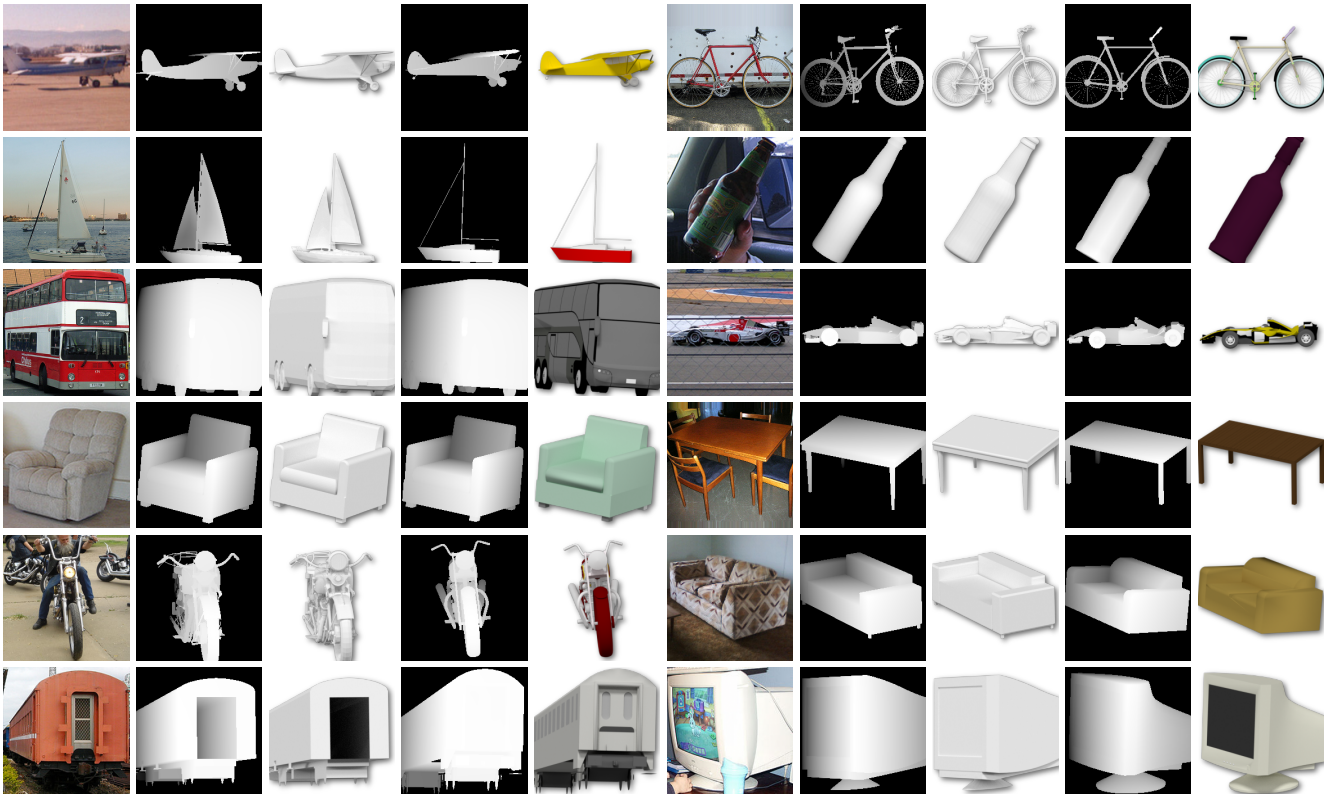

Figure 2: Qualitative results for 3D pose estimation and 3D model retrieval from ShapeNet given images from Pascal3D+ for all twelve categories. For each category, we show: the query RGB image; the depth image and RGB rendering of the ground truth 3D model from Pascal3D+ under the ground truth pose from Pascal3D+; the depth image and RGB rendering of our retrieved 3D model from ShapeNet under our predicted pose.

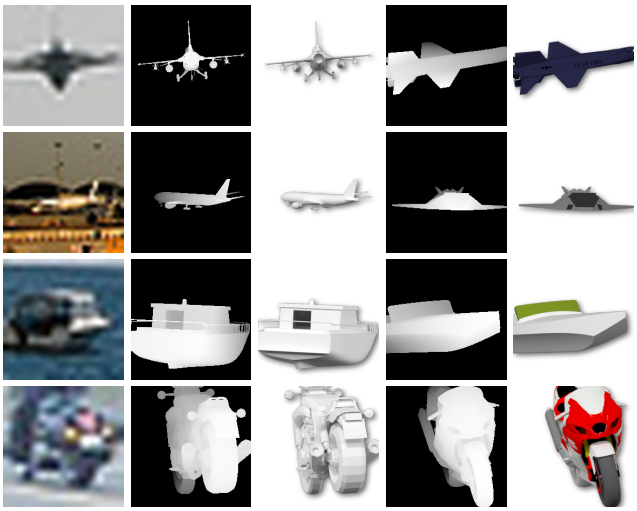

Figure 3: 3D pose estimation fails due to low-resolution image windows (same image arrangement as in Fig. 2). In fact, for more than 55% of Pascal3D+ validation detections the longer edge of the 2D image window is smaller than 224 pixel, which is the fixed spatial input size of pre-trained CNNs like VGG [6] or ResNet [2, 3]. If the resolution is too low, we cannot predict an accurate 3D pose.

Figure 4: 3D pose estimation fails in difficult situations (same image arrangement as in Fig. 2). We observe that heavy occlusions (first row), bad illumination conditions (second row) and difficult object poses (third and fourth row), which are far from the poses seen during training, result in incorrect pose predictions. In the last row, we see that not even the annotated ground truth pose is correct.

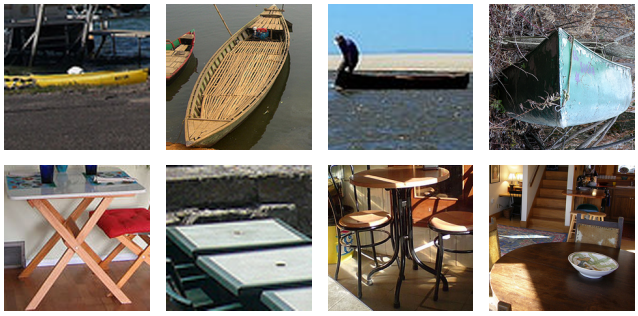

Figure 5: Objects with ambiguous poses from Pascal3D+ validation data. First row: It is impossible to differentiate between the front and back of symmetric boats. Second row: Tables which are ambiguous with respect to an azimuth rotation of π (first image), $\frac{\pi}{2}$ (second and third image) or even have an axis of symmetry (fourth image).

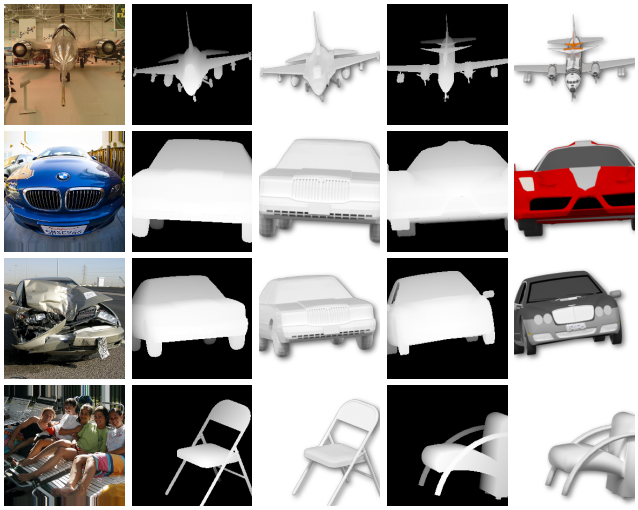

Figure 6: 3D model retrieval results for challenging cases where pose estimation was successful (same image arrangement as in Fig. 2). The test images can exhibit fish-eye effects due to wide-angle lenses (first and second row), contain deformed or demolished objects (third row), or objects under heavy occlusions (fourth row), which disturb object retrieval. Note however that the ground truth 3D models are not accurate.