# Supplementary Material: PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation

Sida Peng[1*] Yuan Liu[1*] Qixing Huang[2] Xiaowei Zhou[1†] Hujun Bao[1†]
[1]Zhejiang University [2]University of Texas at Austin

In the supplementary material, we provide details on how to generate the synthetic images and results on all objects of the YCB-Video dataset [5]. To show the robustness of our approach against occlusion and truncation, we also present qualitative results on the Occlusion LINEMOD [1] and Truncation LINEMOD dataset. In addition, we provide a video to show the results on the YCB-Video dataset.

## 1. Details about Training data

As described in Section 4.1, we add synthetic images to the training set to prevent overfitting. For the LINEMOD [3] and YCB-Video [5] datasets, we render 10000 images for each object. The viewpoints of objects are uniformly sampled, while the in-plane rotations and 3D translations are randomly generated using the distribution computed from the training data. To deal with different scenes, we randomly choose the backgrounds of synthetic images from the SUN397 dataset [6]. Figure 1 shows some exemplar synthetic images of objects.

Inspired by [2], we use their "Cut and Paste" strategy to further synthesize 10000 images. Specifically, we first select one image from each object's training images respectively. Then, the object regions are cut from these images and then are pasted into one image. This strategy not only introduces the real domain information but also simulates occlusions among objects, as shown in Figure 1.

## 2. Detailed Results on the YCB-Video dataset

Table 1 shows our detailed results on all 21 objects in the YCB-Video dataset [5], which considers 024_bowl, 036_wood_block, 051_large_clamp, 052_extra_large_clamp and 061_foam_brick as symmetric objects.

## 3. Qualitative results

Figure 2 and Figure 3 show qualitative results on the Occlusion LINEMOD [1] and Truncation LINEMOD, respectively. These results demonstrate the robustness of our approach to occlusion and truncation.

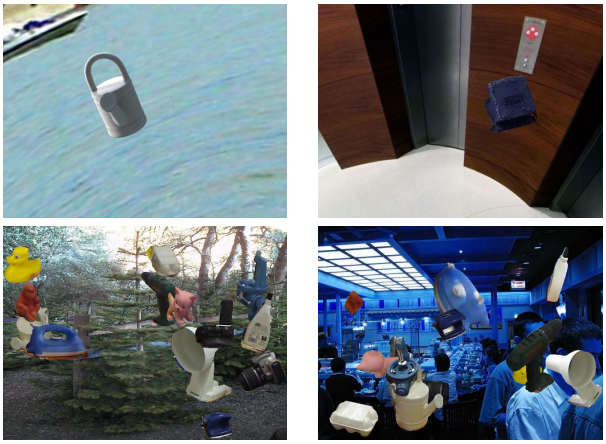

Figure 1. Exemplar images of synthetic training data. Rendered images are shown in the first row. The second row shows the synthetic images generated by the "Cut and Paste" strategy [2].

| | 2D Projection | | | ADD(-S) AUC | | |
|---|---|---|---|---|---|---|
| object | PoseCNN [5] | Oberweger [4] | OURS | PoseCNN [5] | Oberweger [4] | OURS |
| 002_master_chef_can | 74.2 | 0.09 | 29.7 | **55.17** | 49.1 | **81.6** |
| 003_cracker_box | 0.12 | **64.7** | 50.35 | 52.9 | **83.6** | 80.5 |
| 004_sugar_box | 7.11 | **72.2** | 61.25 | 68.3 | 82.0 | **84.9** |
| 005_tomato_soup_can | 5.21 | 39.8 | **60.69** | 66.1 | **79.7** | 78.2 |
| 006_mustard_bottle | 6.44 | **87.7** | 82.35 | 80.8 | **91.4** | 88.3 |
| 007_tuna_fish_can | 2.96 | 38.9 | **45.21** | 70.6 | 49.2 | **62.2** |
| 008_pudding_box | 5.14 | **78.0** | 52.80 | 62.2 | **90.1** | 85.2 |
| 009_gelatin_box | 15.80 | 94.8 | **94.85** | 74.8 | **93.6** | 88.7 |
| 010_potted_meat_can | 23.10 | 41.2 | **62.92** | 59.5 | **79.0** | 65.1 |
| 011_banana | 0.26 | **10.3** | 8.18 | 72.1 | **51.9** | 51.8 |
| 019_pitcher_base | 0.00 | 5.4 | **79.30** | 53.1 | 69.4 | **91.2** |
| 021_bleach_cleanser | 1.16 | 23.2 | **37.51** | 50.2 | **76.1** | 74.8 |
| 024_bowl | 4.43 | 26.1 | **33.99** | 69.8 | 76.9 | **89.0** |
| 025_mug | 0.78 | 29.2 | **52.98** | 58.4 | 53.7 | **81.5** |
| 035_power_drill | 3.31 | 69.5 | **74.74** | 55.2 | 82.7 | **83.4** |
| 036_wood_block | 0.00 | **2.1** | 2.06 | 61.8 | 55.0 | **71.5** |
| 037_scissors | 0.00 | 12.1 | **56.35** | 35.3 | **65.9** | 54.8 |
| 040_large_marker | 1.38 | 1.9 | **6.80** | 58.1 | **56.4** | 35.8 |
| 051_large_clamp | 0.28 | 24.2 | **44.94** | 50.1 | **67.5** | 66.3 |
| 052_extra_large_clamp | 0.58 | 1.3 | **7.77** | 46.5 | 53.9 | **53.9** |
| 061_foam_brick | 0.00 | **75.0** | 25.00 | 85.9 | **89.0** | 80.6 |
| average | 3.72 | 39.4 | **47.39** | 61.0 | 72.8 | **73.4** |

Table 1. The accuracies of our method and the baseline methods on the **YCB-Video** dataset in terms of the **2D projection** and **ADD(-S) AUC** metrics. We consider 024_bowl, 036_wood_block, 051_large_clamp, 052_extra_large_clamp and 061_foam_brick as symmetric objects, as suggested by [5].

*The first two authors contributed equally. The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG and the ZJU-SenseTime Joint Lab of 3D Vision.
†Corresponding authors: Hujun Bao and Xiaowei Zhou.

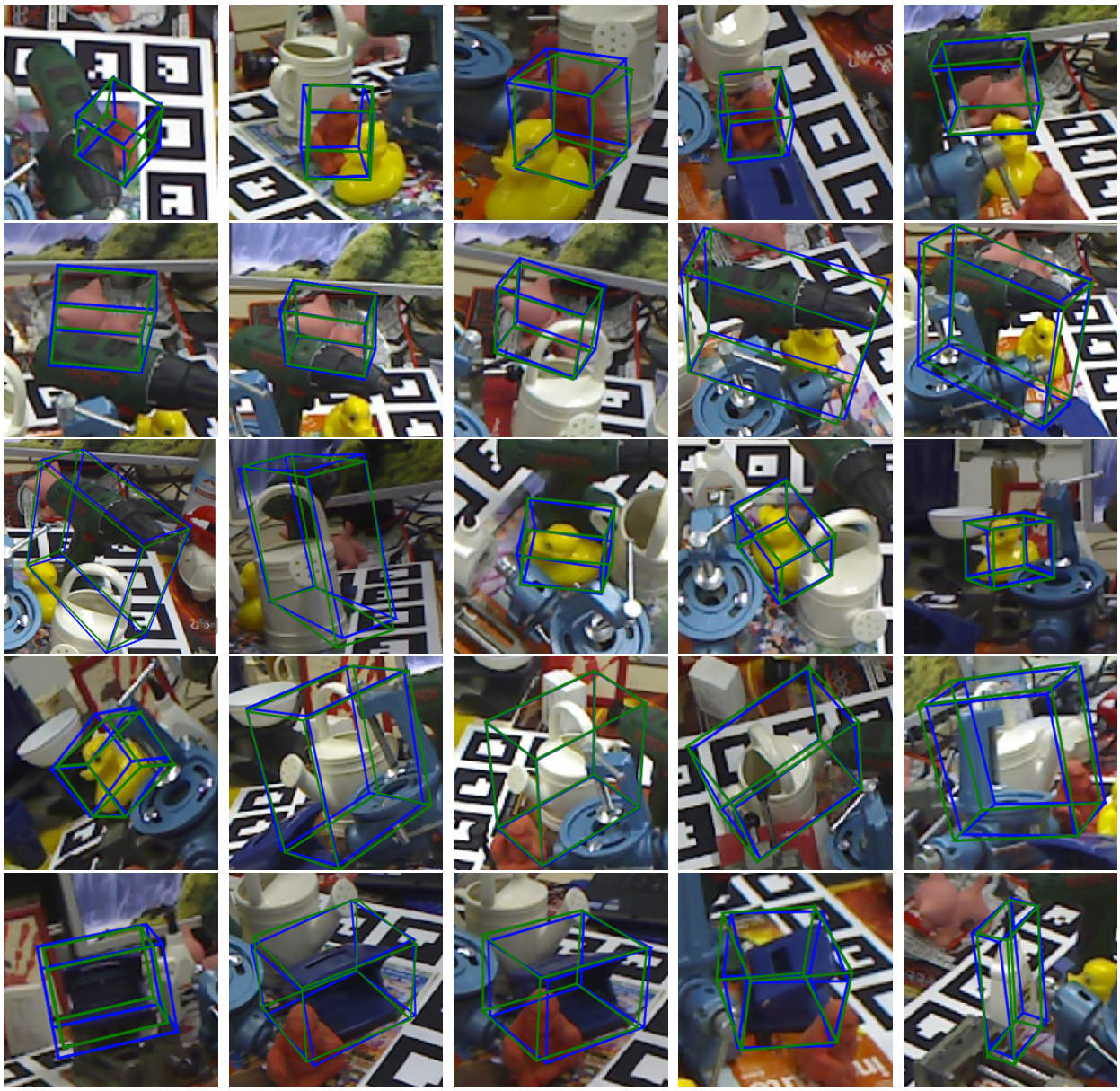

Figure 2. Results on the Occlusion LINEMOD dataset. Green 3D bounding boxes represent the ground truth poses, and blue 3D bounding boxes represent our predictions. It can be seen that our method is quite robust to severe occlusions and cluttered scenes, which demonstrates that the keypoint localization based on dense predictions can handle well occluded objects.

## References


[1] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014.

[2] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*, 2017.

[3] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2012.

[4] M. Oberweger, M. Rad, and V. Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *ECCV*, 2018.

[5] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018.

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

Figure 3. Results on the Truncation LINEMOD dataset. Green 3D bounding boxes represent the ground truth poses, and blue 3D bounding boxes represent our predictions. Although the objects are severely truncated, our method robustly infers the poses based on the visible parts, which indicates that the direction-field representation well deals with truncation.