StereOBJ-1M: Large-scale Stereo Image Dataset for 6D Object Pose Estimation Supplementary Material

Xingyu Liu Shun Iwase Kris M. Kitani
Carnegie Mellon University

A. Overview

In this document, we provide additional details on StereOBJ-1M dataset as presented in the main paper. We present additional baseline results on instance-level pose detection for `centrifuge_tube` class in Section B. In Section C, we provide details on the hardware of data capturing. In Section D, we provide more details on viewpoint distribution of each object class. Lastly, in Section E, we visualize more data samples from our dataset.

B. Multi-instance Pose Detection Results

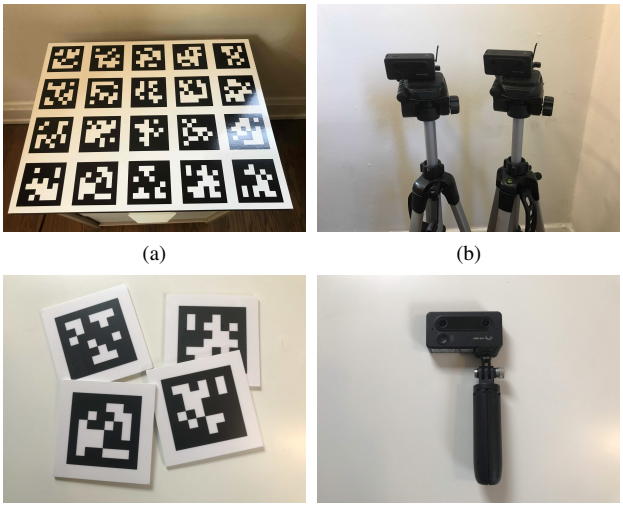
In the main paper, we report the results of two baselines on **single-object pose estimation** of 17 out of 18 objects on the test set where there is at most one object instance from a category in a scene. However, for `centrifuge_tube`, there are usually multiple instances recorded in a scene. Therefore, `centrifuge_tube` is used in **multi-object pose detection** task. In this task, the framework is supposed to perform instance-level detection and pose estimation simultaneously.

To adapt to instance-level pose detection, we modify the baseline formulation by introducing additional 2D object detection before pose estimation. Given a detected rough 2D bounding box of an object instance, we crop the image patch and send it to pose estimation baselines, i.e. PVNet [4] and KeyPose [2], to estimate the 2D keypoint locations and therefore 6D pose of that object instance. The 2D object detector we used is Faster-RCNN [5].

We use Average Precision (AP) as the evaluation metrics of multi-instance pose detection. When calculating AP in 2D object detection, a detection result is considered correct if the IoU between the detected bounding box and a ground truth bounding box is larger than a threshold. Different from 2D object detection, we consider a pose detection result to be correct if the ADD(-S) distance between the detected 6D pose and a ground truth pose is smaller than a threshold. We use 10% of the object diameter as the threshold of ADD(-S). We report the pose detection results with single-RGB image as input in Table 1. We notice that the above baseline suffers when two or multiple instances object overlap in the image and are included in the same image patch. In this case, the pose estimation framework cannot distinguish different instances and fails in keypoint prediction.

Table 1: AP results of **object pose detection** with single RGB image as input.

method	PVNet [4]	KeyPose [2]
`centrifuge_tube`	15.19	17.64



(a) (b) (c) (d)

Figure 1: **Hardware for data collection.** (a) large fiducial marker board; (b) static cameras with tripods; (c) small fiducial markers; (d) moving stereo camera.

C. Data Capturing Hardware

We present the hardware used for capturing the data in Figure 1, including a large fiducial marker board, several small fiducial markers, two static cameras with two tripods, and one moving stereo camera. We used the same Weewiew stereo camera [1] for all three cameras, though the two stereo cameras can be monocular. Weewiew stereo camera has a stereo baseline of approximately 4.5cm which is

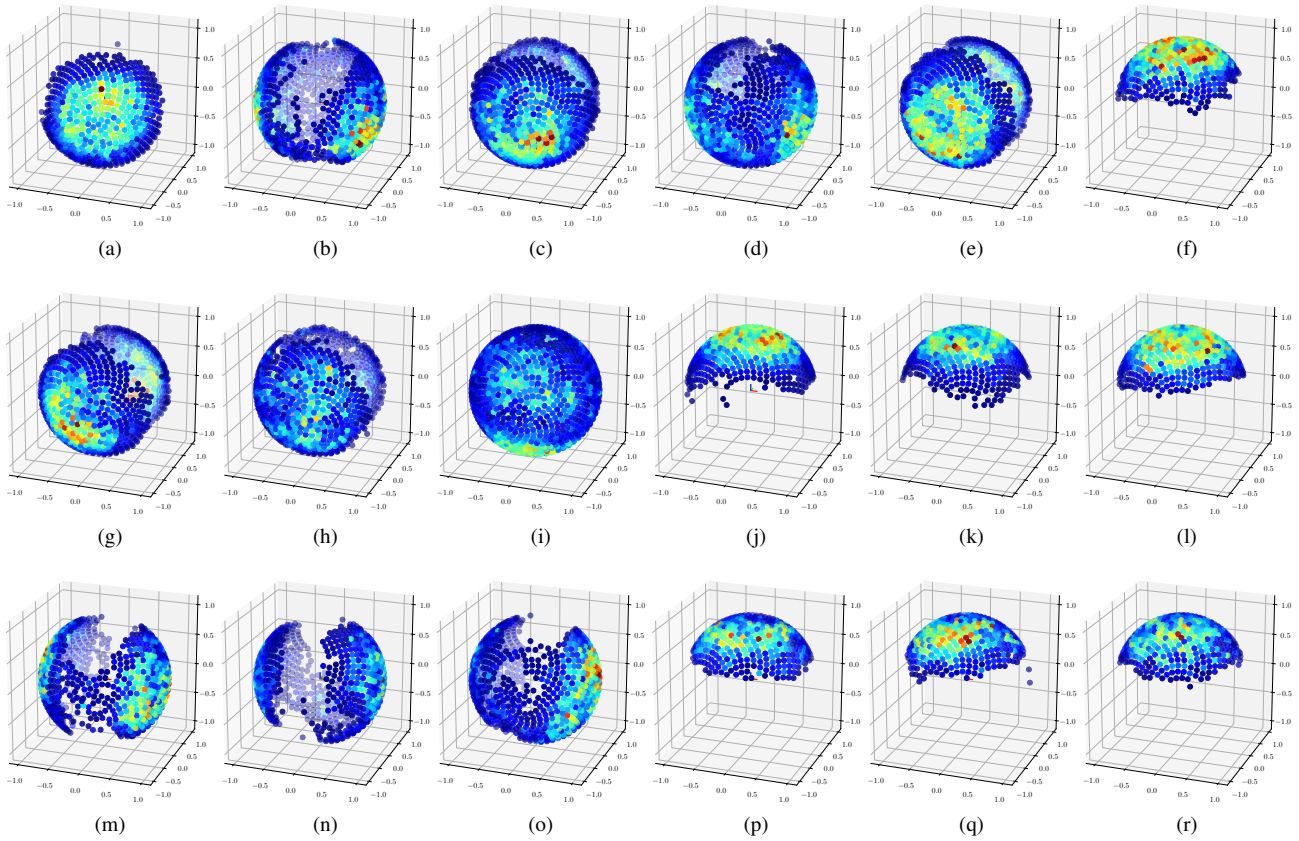


Figure 2: **Viewpoint distribution of the 18 objects in our dataset.** (a) `blade_razor`; (b) `hammer`; (c) `needle_nose_pliers` (d) `screwdriver`; (e) `side_cutters`; (f) `tape_measure`; (g) `wire_stripper`; (h) `wrench`; (i) `centrifuge_tube`; (j) `microplate`; (k) `tube_rack_2`; (l) `tube_rack_50`; (m) `pipette_0.5_10`; (n) `pipette_10_100`; (o) `pipette_100_1000`; (p) `sterile_rack_10`; (q) `sterile_rack_200`; (r) `sterile_rack_1000`.

close to the distance between the two human eyes. All three cameras are calibrated.

The fiducial markers are the first 20 AprilTags [7]. The large fiducial marker board is printed on a 20in $\times$ 16in plastic picture frame. Though the large fiducial marker board needs to be accurately measured by its physical dimensions with a vernier caliper, the small fiducial markers do not.

D. Viewpoint Coverage Distribution

Viewpoint coverage percentage is illustrated in Section 4.2 and Figure 5 of the main paper. We illustrate a more detailed viewpoint distribution for each object in Figure 2. The viewpoints are drawn as 3D points on the unit sphere centered at the object center. Their positions on the unit sphere are determined by the Azimuth and Elevation of the viewpoint. Their density on the sphere is shown by heatmap color. Notice that for objects such as `microplate` and `tape_measure`, there is no viewpoint distributed on the $-z$ space, because there is only one possible side up when being put on a desktop.

E. More Visualizations of Data Samples

We provide more visualizations of data samples from our dataset. As illustrated in Figure 3, the data annotation has high precision.

References

[1] WEEVIEW INC. weeview: 3d camera. https://www.weeview.co/. 1

[2] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *CVPR*, 2020. 1

[3] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *ICCV*, 2019. 3

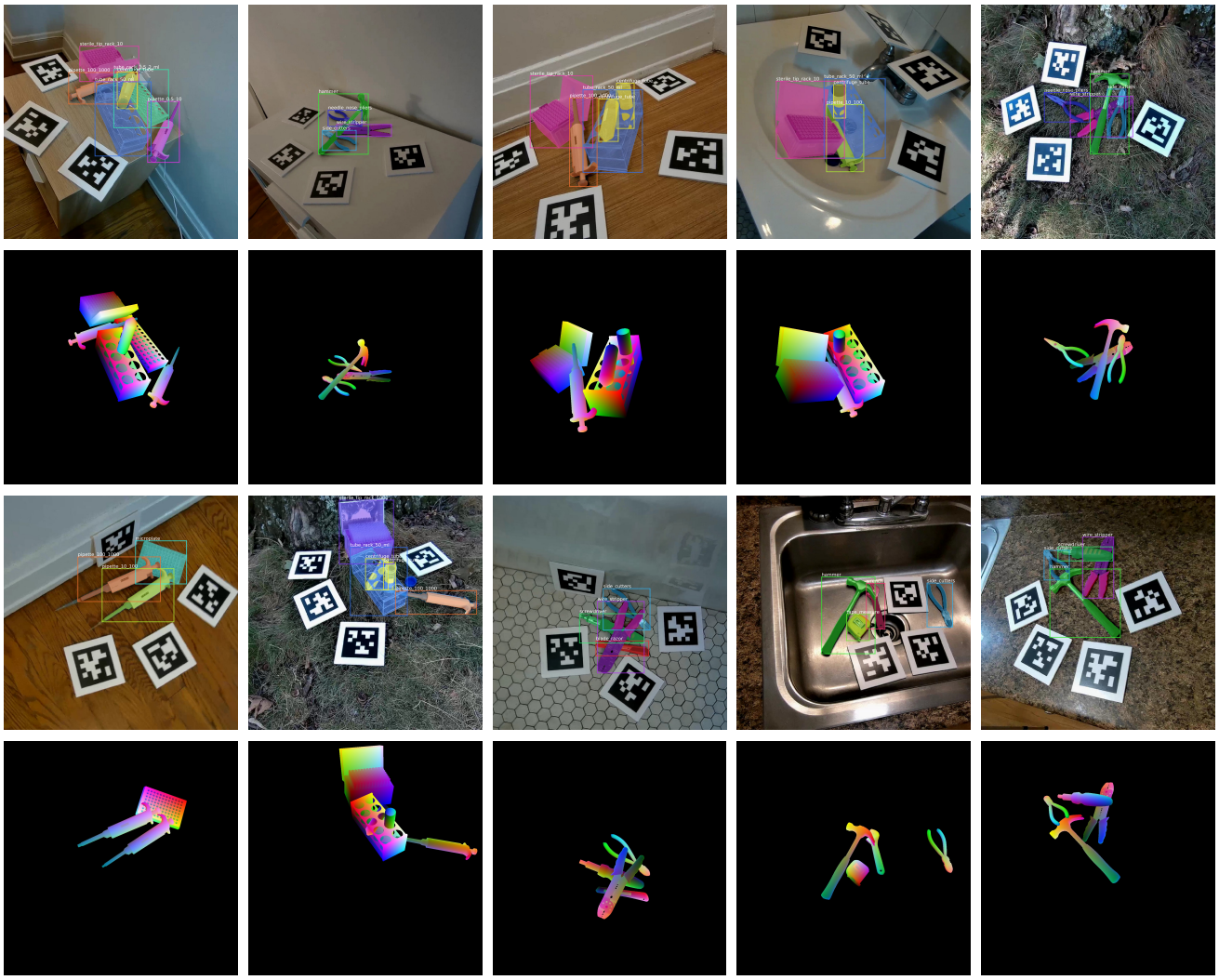


Figure 3: **Visualization of data samples from StereOBJ-1M dataset.** The first row is left stero images with semantic masks and bounding boxes superimposed. In the second row, we use normalized coordinate map [3, 6] to illustrate the 6D poses of the corresponding objects, where the coordinates of the object surface points are normalized to $[0, 1]^3$ and converted to RGB values in $[0, 255]^3$ at projected pixels.

[4] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 1

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[6] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, 2019. 3

[7] John Wang and Edwin Olson. Apriltag 2: Efficient and robust fiducial detection. In *IROS*, 2016. 2