Supplemental Material For: *Aria Digital Twin: A New Benchmark Dataset for Egocentric 3D Machine Perception*

1. Introduction

In this supplemental material document, we dive deep into the implementation of the system accuracy measurement and more detailed results of it. We perform more qualitative and quantitative analyses on the 2D object detection, image segmentation and 3D object detection tasks. Furthermore, we introduce another important use case of the ADT dataset that can quantitatively evaluate a manual 3D bounding box annotation pipeline before it is applied to large-scale egocentric data.

2. System Accuracy

We provide additional information and figures in this section to better describe the methodology. We also provide additional tables with results for the reader to better understand the data statistics and how the accuracy of the system depends on different factors.

Figures 1 and 2 illustrate the system accuracy analysis on an exemplar frame. Figure 1 shows a portion of a zoomed in RGB image where a wooden spoon is being moved in by an Aria wearer. As described in Section 3.3, we take this image and manually label the centers of each marker. The system accuracy estimation pipeline then estimated the object pose relative to the image which best aligns the projection of the 3D markers to the hand labels. Figure 2 shows the final results after the optimization described in Section 3.3. The green crosses are the manual labels; the red crosses are the marker reprojections onto the image plane given all system measurements at the capture time for this frame; and the blue crosses are the reprojections of markers after applying the optimized object pose using Eqn.5 in Section 3.3. The misalignment between the green crosses and the red crosses indicates the error of the object pose. The alignment between the green crosses and the blue crosses confirms that the estimation of the true object poses is correct.

Table 2 shows the system accuracy statistics for each of the two scenes. The accuracy in the office is slightly better than the accuracy in the Apartment. We expect the root cause to be the higher ceilings in the apartment, where the motion capture cameras are installed, yielding a slightly worse tracking accuracy. Table 1 shows the system accuracy measurement of 32 dynamic objects averaged on a per-object basis. The total system error comes from the 3D object reconstruction, motion capture system, Aria device poses and Aria device calibration.

Figure 1: Cropped version of example Aria image used for system accuracy tests.

3. Performance Analysis on 2D Object Detection and Image Segmentation

The performance of the state-of-the-art models, namely FPN and VIT-Det, for 2D object detection and image segmentation tasks on the ADT dataset is significantly lower than their performance on the COCO dataset. We expect this discrepancy is largely due to the domain difference between these two datasets, which is consistent with the findings of [1]. Despite the rectification of the Aria fisheye RGB images to bring ADT closer to the distribution of COCO, the egocentric nature of the data still remains a challenge for these algorithms. Table 3 shows the per-category mAP. As can be seen from the table, large furniture, appliances cate-

Object Name	Measurement Count	Translation Error [mm]	Rotation Error [deg]	Reprojection Error[pixel]
BlackCeramicBowl	10	3.05	0.66	5.05
Donut_B	11	3.61	1.06	4.84
MuffinPan	10	3.64	0.59	5.45
RedClock	10	3.72	1.03	4.19
DecorativeBoxHexLarge	12	3.77	1.05	5.05
CoffeeCan_2	10	4.06	0.66	5.43
Mortar	11	4.19	0.74	6.45
ChoppingBoard	10	4.25	0.49	5.23
BlackCeramicDishLarge	10	4.31	0.71	5.26
WoodenFork	13	4.53	1.65	6.71
BirdhouseToy_2	17	4.77	1.11	4.55
BambooPlate	10	4.82	0.67	7.34
BirdHouseToy	12	5.08	0.72	7.53
Orange_A	14	5.22	2.28	8.19
ToothBrushHolder	12	5.32	1.66	7.24
CakeMocha_A	15	5.62	0.69	6.14
WoodenSpoon	10	5.85	2.02	6.42
WoodenBowl	10	5.85	0.74	6.53
BlackPictureFrame	13	6.00	1.16	8.73
BlackTablet	7	6.19	1.11	6.69
BlackCeramicMug	10	6.53	1.69	6.59
BookDeepLearning	11	6.56	0.96	10.31
WoodenBoxSmall	12	6.73	1.28	8.83
Flask	14	7.17	1.49	5.71
GreenDecorationTall	10	8.02	1.37	8.81
BlackRoundTable	11	8.43	0.65	5.72
Cracker	10	8.49	2.25	7.20
BlackKitchenChair	9	12.24	0.79	5.66
WhiteChair	6	12.35	0.77	6.83
Jam	14	12.57	1.52	7.32
Cereal	9	16.29	2.18	11.82
DinoToy	10	25.39	4.65	7.25

Table 1: Mean system accuracy results for select objects ranked by the translation error.

Error	Apartment	Office
Object translation [mm]	6.94	4.48
Object rotation [deg]	1.3	1.04
Reprojection Measured [pixels]	6.9	4.18
Reprojection Optimized [pixels]	0.56	0.47

Table 2: Mean system accuracy results, split by scene location.

gories such as couch, chair, refrigerator are typically easier for the detectors to detect in these videos while their performance is poor on object categories such as potted plant, mouse, remote etc. Though this can be attributed to the scale of the objects present in the videos, it also highlights the challenges of building a real world index of everyday objects from in the wild recordings. Furthermore, in a qualitative analysis, Figure 3 show the performance of both detectors along with the ground truth. FPN shows better performance detecting large objects and objects under viewpoint variance. Although VIT-Det seems to be better at detecting small objects compared to FPN, its overall inferior performance to FPN suggests a possible mismatch between the training scale and the sizes of the ADT images at the inference stage.

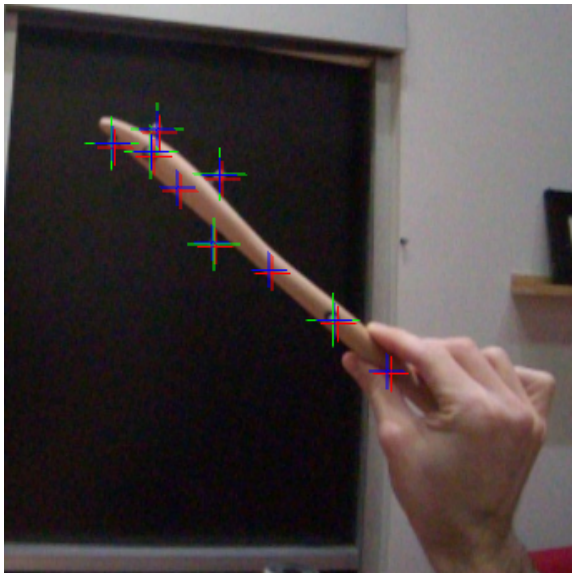


Figure 2: Cropped version of example Aria image used for system accuracy tests with results. Red: system's estimate of where the markers should project. Green: hand labels of where the markers are located in the image. Blue: system estimate of where the markers should be after optimizing for the true object relative pose.

4. Performance Analysis of 3D Object Detection

The 3D object detection performance of Cube-RCNN and Total3d is significantly lower on the ADT dataset. We therefore conduct more analyses on the failure cases to enlighten the challenges of 3D object detection research. Our observations include two major failure cases: 1) 2D object detection failure, 2) 3D pose prediction failure. Since we analyse 2D object detection failures in Section 3, we will focus on 3D pose prediction failures in this section. Figure 4a shows a typical failure case of 3D pose prediction. Cube R-CNN roughly localizes the 3D position of eight chairs but fails in predicting 3D poses accurately enough to pass the IoU threshold of 0.25.

Additionally, we observed frequent failure cases with the depth estimation which is a fundamental limitation of 3D detection models based on single image inputs, since 3D data is challenging to infer from a single 2D image. Figure 4b and Figure 4c show two failure examples for Total3D and Cube R-CNN, respectively. The reprojected 3D bounding boxes fit well on the 2D images. However as evident from the 3D visualizations, the predicted poses are significantly erroneous when compared to the ground truth. This problem can be potentially solved by a more advanced 3D object detector using multi-camera sensors from Aria.

Category	FPN Box	FPN Seg	VIT-Det Box	VIT-Det Seg
Frisbee	18.55	21.10	7.51	6.80
Bottle	2.91	3.03	1.28	1.32
Cup	5.67	5.64	4.56	4.83
Fork	8.12	2.85	4.25	1.13
Knife	14.50	10.58	10.82	7.93
Spoon	14.20	6.24	7.07	3.78
Bowl	17.81	17.41	7.23	7.53
Banana	16.87	12.73	8.25	6.32
Apple	21.64	24.03	12.31	14.03
Sandwich	14.15	10.94	8.88	11.41
Orange	19.84	21.80	9.87	10.80
Carrot	37.08	53.02	38.84	29.75
Donut	3.93	4.57	2.29	2.54
Cake	10.25	12.52	9.21	10.84
Chair	34.38	17.44	20.80	9.58
Couch	49.77	49.87	27.82	32.20
Potted Plant	0.51	0.48	0.40	0.38
Bed	7.29	2.42	6.34	3.61
Dining Table	25.02	7.63	2.37	0.75
TV	24.73	29.65	19.10	23.76
Laptop	12.66	12.78	2.30	2.61
Mouse	1.11	0.98	0.20	0.17
Remote	1.47	0.30	1.82	0.54
Keyboard	4.01	3.31	0.44	0.30
Oven	0.05	0.01	0.61	0.37
Toaster	0.09	0.11	2.22	2.54
Refrigerator	48.47	48.45	42.89	43.63
Book	10.12	9.23	3.40	2.83
Clock	34.33	34.97	32.21	33.37
Vase	0.34	0.28	0.22	0.12
Scissors	7.52	0.14	10.92	0.33

Table 3: Per-category 2D detection and segmentation mean mAP computed across all videos in the dataset. Large furniture and appliances are easier to detect for the detectors than the smaller objects like remotes. This indicates the challenges in the constructing real world index of everyday objects.

5. Comparison with Manual 3D Bounding Box Annotation

Accurate 3D bounding boxes in the ADT ground truth dataset can be leveraged to benchmark the accuracy of a video-based manual annotation pipeline. To set up the evaluation, we select 20 randomly sampled videos (10% of the total videos) from the dataset for manual annotation of 3D bounding boxes using objects from 10 categories. Figure 5 shows examples of the manual annotations. We evaluate each manual bounding box annotation of an object by com-

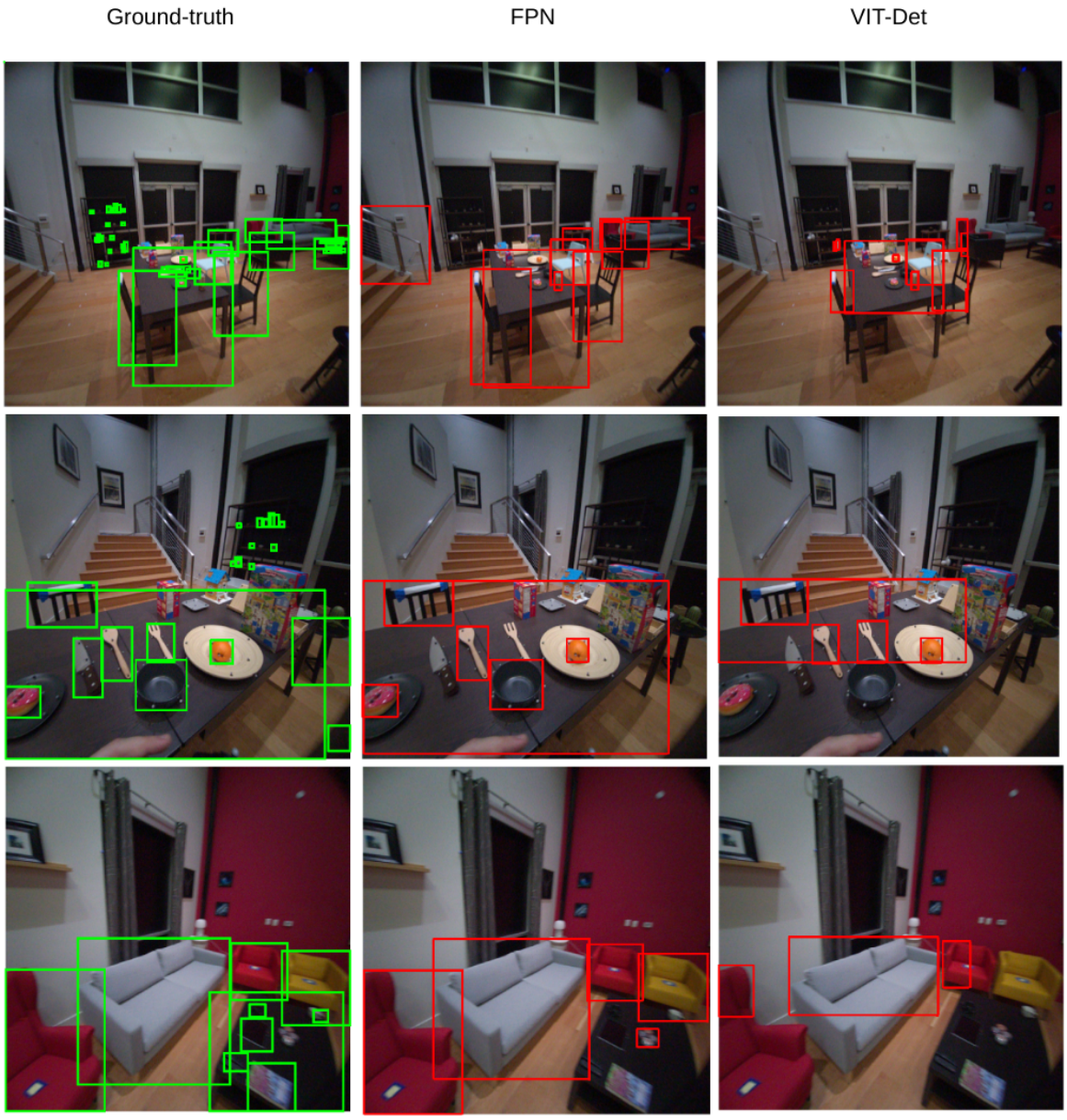
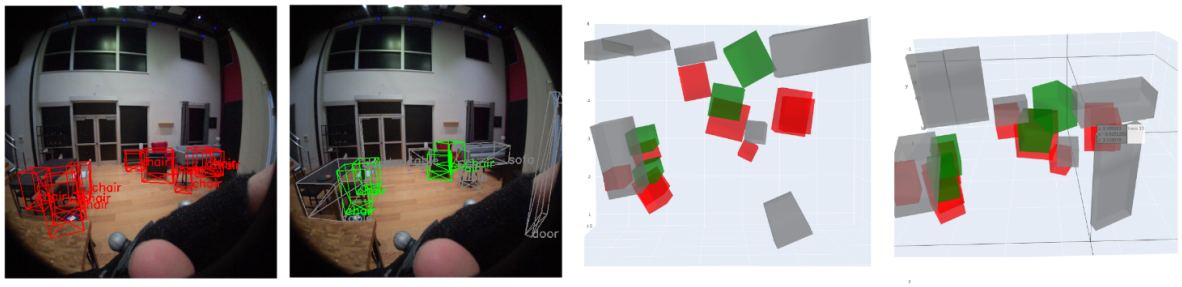


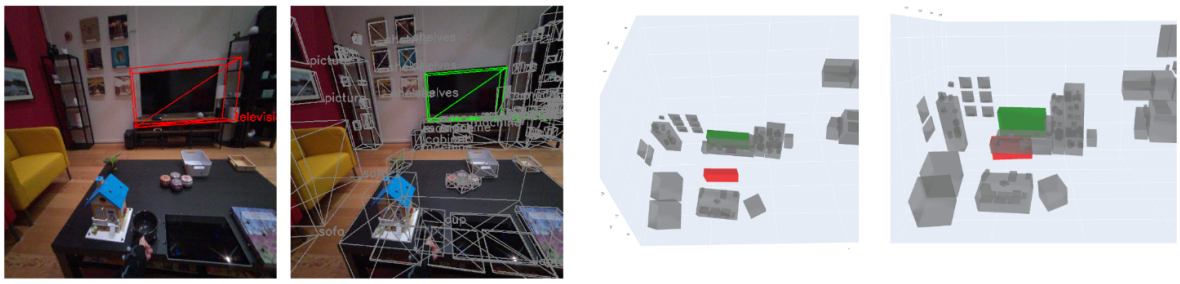
Figure 3: Each row is an example of the comparison among the ground-truth, FPN 2D detection result and VIT-Det 2D detection result. All three examples shows that FPN tends to detect larger objects better than that of VIT-Det, such as the dining table in the first and second example, and the sofa and armchairs in the third example. FPN also shows promising robustness results under view point variance such as the dining table in the second example and the leftmost armchair in the third example. In contrast, VIT-Det seems to be better at detecting smaller objects such as the bottles on the shelf behind the dining table in the first example and the fork in the second example.

puting the difference from the 6DoF ground truth pose in ADT, including translation, rotation and scale errors. The mean translation error is 0.329 meters; the mean rotation error is 4.29 deg and the mean relative scale error is 0.32. We show the evaluation results on three example categories in Table 4.

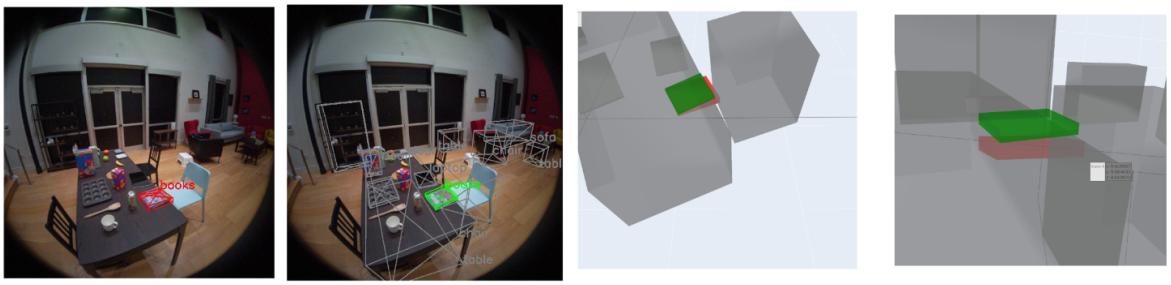
The experiment above introduces a distinct advantage for testing a semi-automatic annotation pipeline and for training annotators with continuous, quantified and visualized feedback before creating large-scale tasks. Visualizations



(a) A failure example of Cube R-CNN on predicting 3D poses of chairs.



(b) A failure example of Total3d on predicting the 3D pose of a TV object.



(c) A failure example of Total3d on predicting the 3D pose of a book object.

Figure 4: From left to right: 3D object detection in red bounding boxes; ground truth bounding boxes in green for the target object and in gray for other objects; predicted 3D bounding boxes from a top down view; predicted 3D bounding boxes from a side view.

	Sofa	Photo Frame	Chair
Center Prediction (m)	0.296	0.162	0.041
Rotation (deg)	3.869	1.952	1.553
Relative Scale	0.15	0.27	0.10

Table 4: Benchmarking of the manual annotations. It shows error in manually annotated objects measured against the accurate ground truth provided by the ADT. Smaller objects are difficult to annotate with accuracy as can be seen from the higher relative scale error of the photo frames.

such as those shown in Figure 5 can act as a quick reference for educating annotation teams on the common failure modes and patterns.

References

[1] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. *International Journal of Computer Vision (IJCV)*, 2022.

Figure 5: Examples of the manual annotation. Small and thin objects are typically more difficult to manually annotate compared to large and bulky objects. The error margin for annotating a photo frame is much smaller as compared to annotating bigger furniture objects such as the sofa and bed. Typically annotating the depth becomes a challenging task and is often the main cause of the error. The ADT dataset allows for an accurate estimate of these errors as shown in table 4