Towards Co-Evaluation of Cameras, HDR, and Algorithms for Industrial-Grade 6DoF Pose Estimation Supplementary Material

Agastya Kalra Guy Stoppi Dmitrii Marin Vage Taamazyan Aarrushi Shandilya
Rishav Agarwal Anton Boykov Tze Hao Chong Michael Stark
Intrinsic Innovation LLC

The supplementary material contains additional experiments and figures that we did not include in the main paper. In Appendix 1, we give an additional ablation on the impact of viewpoint diversity on the quality of the proposed robot consistency method. In Appendix 2, we provide additional figures showing the dataset from the paper. In Appendix 3, we provide the detailed evaluation table which was used to calculate the performance of cameras on representative parts in the paper.

1. Robot Consistency Viewpoint Ablation

In this section we provide an additional ablation study on the viewpoint diversity required for robot consistency to produce accurate estimates of model performance. In our proof in section 3.2 of the paper, we assumed that the camera poses were spherically symmetric - i.e. created a full sphere around the object, however in practice this is not feasible. The goal of this experiment is to determine the minimum camera coverage required for robot consistency to accurately estimate model performance.

Testing Setup. A sphere can be decomposed into an azimuth and zenith angle as shown in Figure 1. In our dataset, the azimuth is a full 360 degrees because our robot can rotate 360 degrees in the yaw-axis. Our goal is to find the appropriate range of zenith angles for the camera w.r.t. the robot. This corresponds to the range of the pitch and roll of the robot. To do this, we follow the synthetic procedure described in the main paper. We generate 4 cameras, similar in location, FOV and resolution to the Basler-HR cameras shown in Figure 3. We then synthetically generate sets of 30 scenes for 10 different parts and using different zenith angle ranges. For instance, if the zenith angle range is 35 degrees, we place the camera directly on top of the objects, then add a uniformly sampled jitter of ±35 degrees in the zenith angle and move the camera accordingly.

For each part, we train 20 different keypoint models for pose estimation M_i by varying different hyperparameters

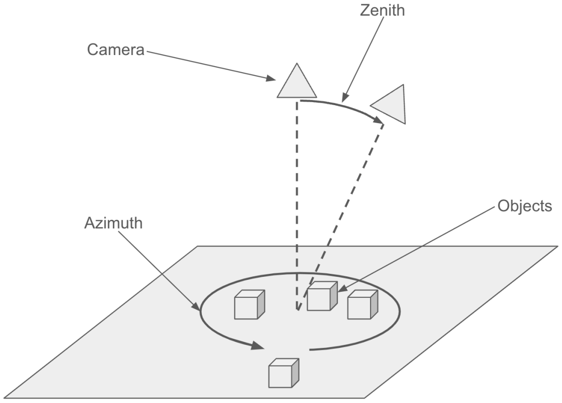

Figure 1. **Visualizing the azimuth and zenith angles of a camera with respect to objects in a scene.**

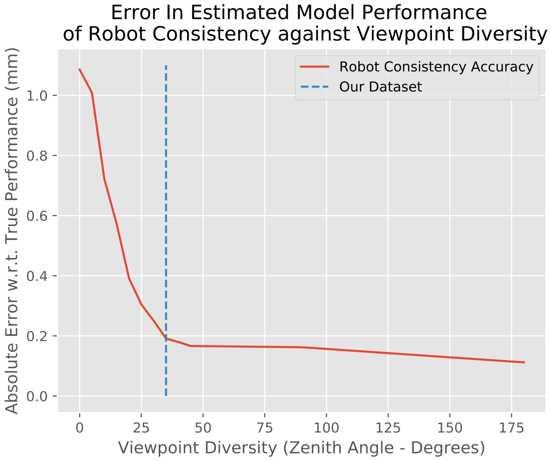

Figure 2. **As viewpoint diversity increases, the accuracy of model performance estimated by robot consistency improves**. The above plot shows the improvement in absolute error w.r.t. true performance against different zenith angles. We find that at zenith angles $\approx 30 - 40$ degrees, the robot consistency metric starts to plateau. Since this is close to the 33 degree zenith coverage in our dataset, the model performance estimates obtained using robot consistency on our camera pose distribution are fairly accurate.

such as input resolution, backbone etc. We evaluate all 20 of these models on all 10 parts using both the ground truth pose ${}^{gt}T_{CO}$ available in the synthetic world, and the robot consistency method. We take the absolute difference between the model performance $E_{rc}(M)$ estimated via robot consistency and compare it to perfect evaluation $E_{gt}(M)$ obtained via ground truth:

$$\frac{1}{N}\sum_{i}^{N}|E_{rc}(M_i) - E_{gt}(M_i)| \tag{1}$$

.

The above equation is called the *Absolute Error w.r.t. True Performance* because it represents the error between the model performance estimate from robot consistency and the true model performance evaluated using the perfect synthetic ground truth. We compute this error for multiple zenith angles ranging from 0 to 180 degrees and plot this result in Figure 2.

Discussion. Figure 2 shows that using a very small range of zenith angles (less than 10 degrees) leads to poor evaluation accuracy for robot consistency. However as the zenith angle range goes beyond 30 degrees, the accuracy seems to plateau. This works out for our dataset which has a ±33 degree zenith angle range i.e. the model performance estimates obtained using robot consistency on our dataset are fairly accurate.

2. Additional Samples from Dataset

Here, we include a set of figures that highlights some of the diversity in our dataset. Figure 3 visualizes our multi-camera setup. Figure 4 shows sample images from all the 13 different viewpoints captured of a single scene, along with the modalities of AOLP/DOLP obtained through the FLIR-monoP camera and the depth map processed from Photoneo's point cloud. We also include figure 5 to showcase some of the difficult part types in our dataset. Figures 6 to 8 show all 30 scenes captured in a given sequence for evaluating robot consistency using different camera modalities.

3. Part-wise Results

In this section we provide the full results of the abbreviated Table 2b from in main paper in Table 1. We find the results are consistent with the paper and Basler-HR performs best. However with further research perhaps polarization or structured light may perform better.

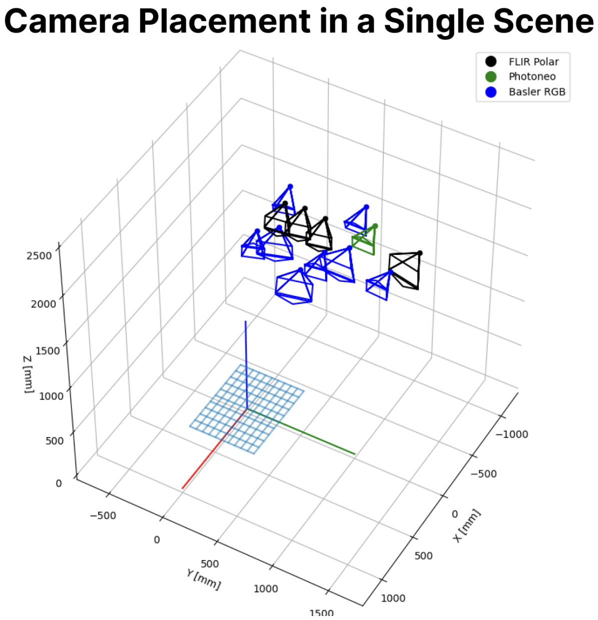

Figure 3. **Our dataset uses 13 cameras placed at a distance of 2.5m above the robot base**. We visualize each camera's pose relative to the robot base at origin (0,0,0). In general, the robot presents the objects to the camera at 0.5m above the base, so most images are at a working distance of 1.5m to 2.2m. The cameras are spread out in an XY plane of size 0.5m x 1.3m, leading to a variety of baselines, including some very large ones.

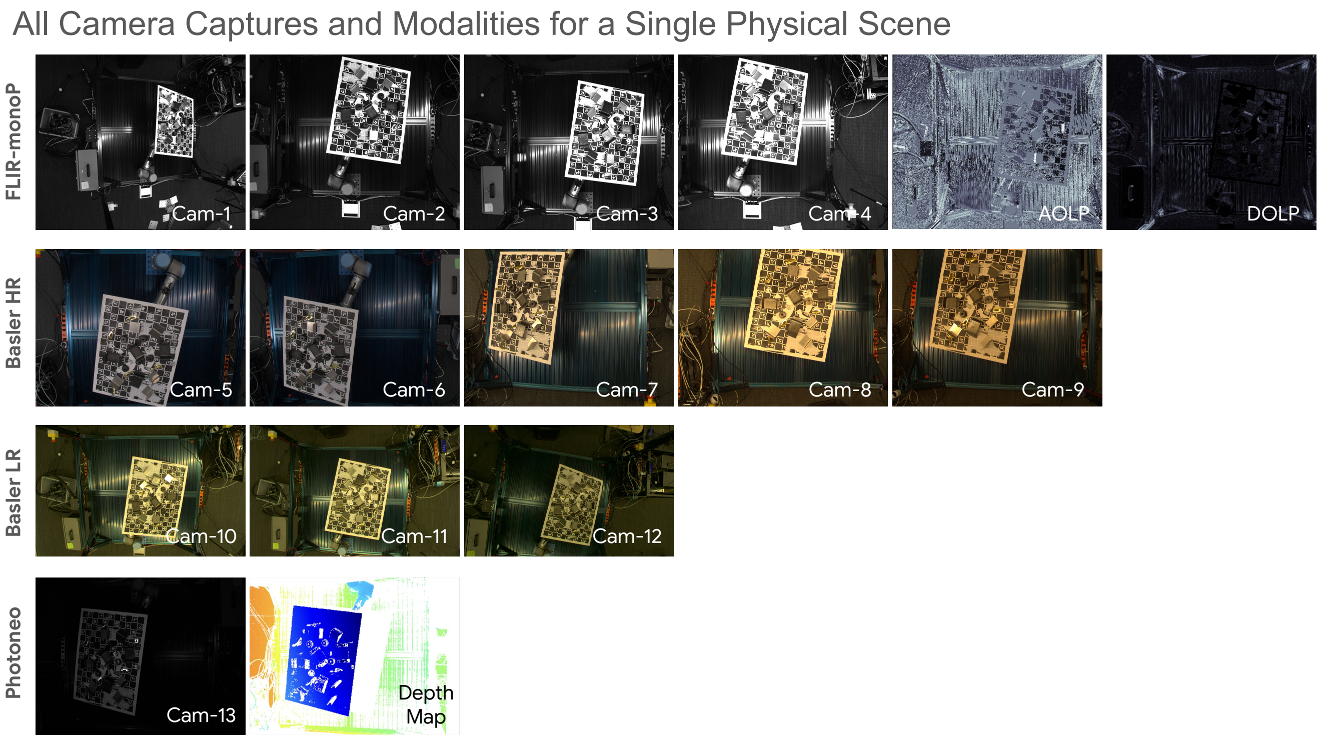

Figure 4. **Our dataset contains wide baselines with diverse viewpoints**. The above figure shows each image collected from every camera per scene. We use 13 cameras at baselines up to 1m. The captured images include RGB and grayscale, along with AOLP/DOLP Polarization, and Depth Map.

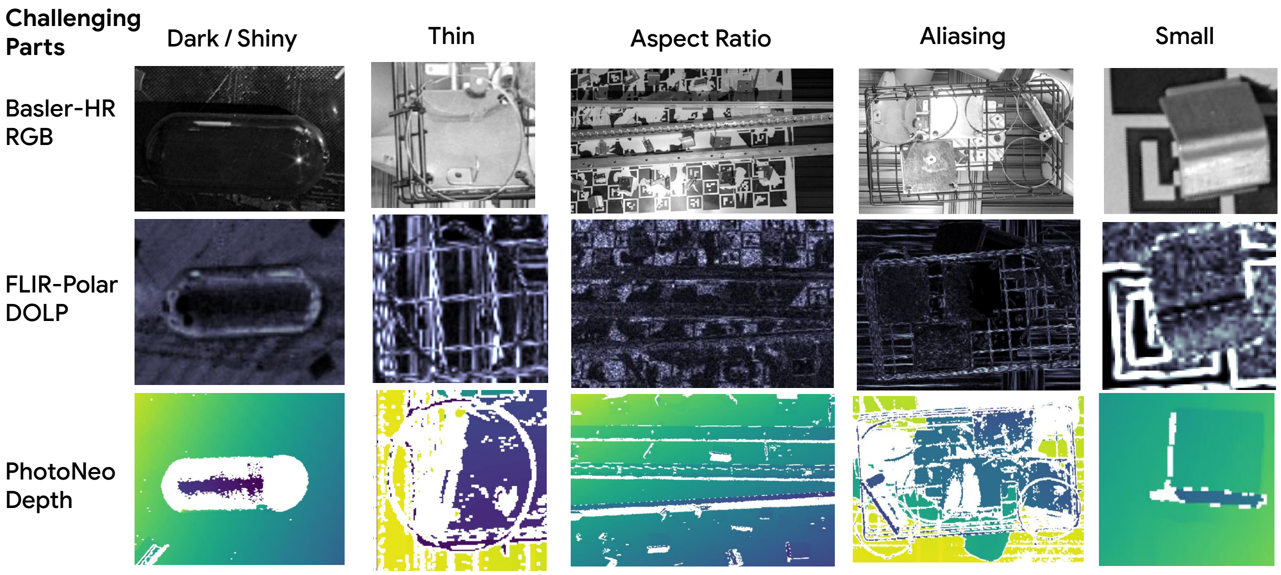

Figure 5. **Some of the challenging parts available in our dataset.** From left to right: (1) Dark / Reflective parts cause issues in Photoneo's structured light point cloud, although they are bright in the degree of linear polarization. (2) Thin object span only a few pixels, making them difficult to detect. (3) Long parts with non-uniform aspect ratio e.g. 25:1 makes it difficult for detection algorithms. (4) Thin basket wires create a repeating pattern, making it difficult to determine the pose. (5) Corner Bracket 1 is only 1.5x1.5cm at a distance of 150-250cm making it difficult to distinguish.

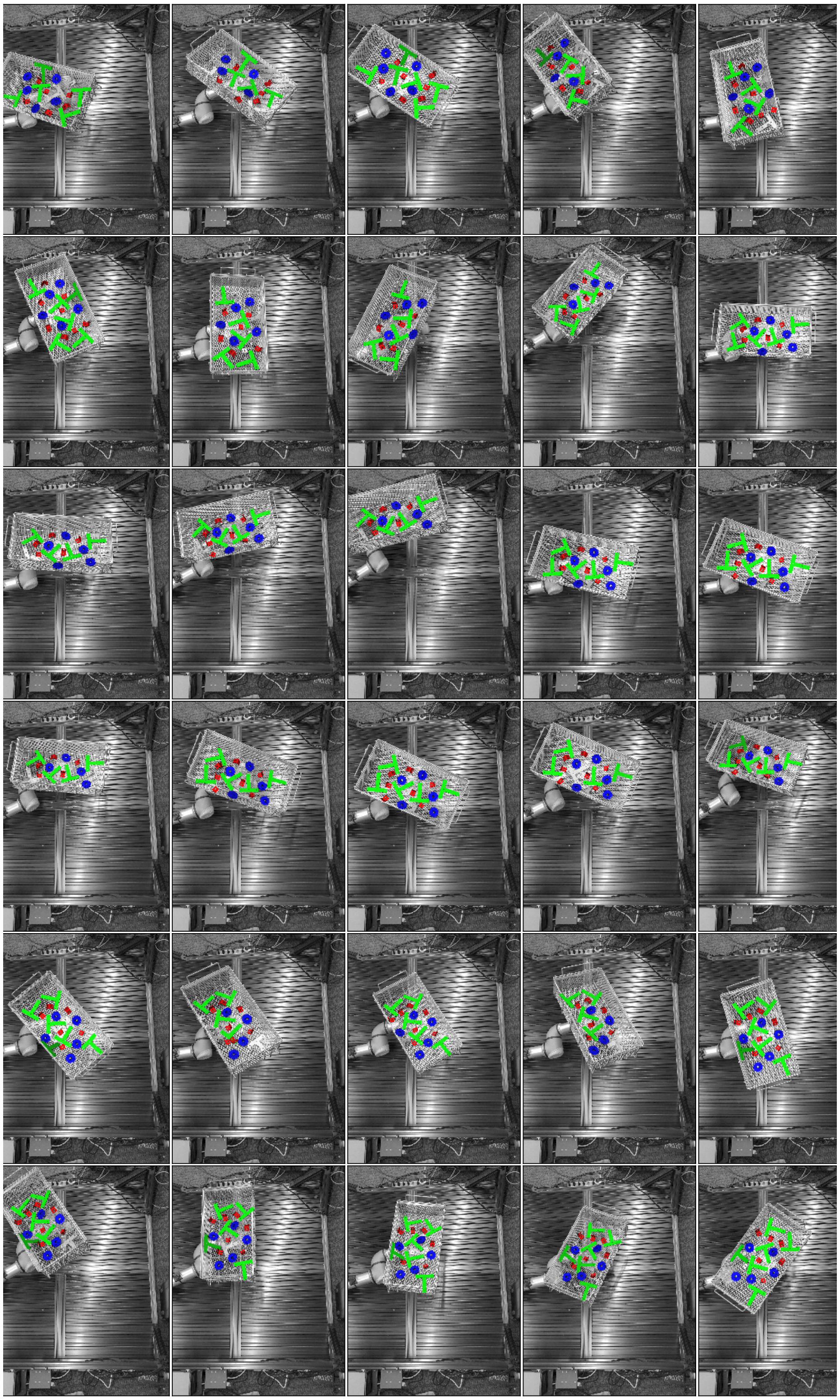

Figure 6. **The predictions from 30 scenes captured by Basler-HR that are used for robot consistency calculation.** The above shows each of the 30 captures from the perspective of 1 Basler-HR camera with pose predictions rendered on top. The poses are calculated using all 5 cameras, however here we only show one camera's viewpoint.

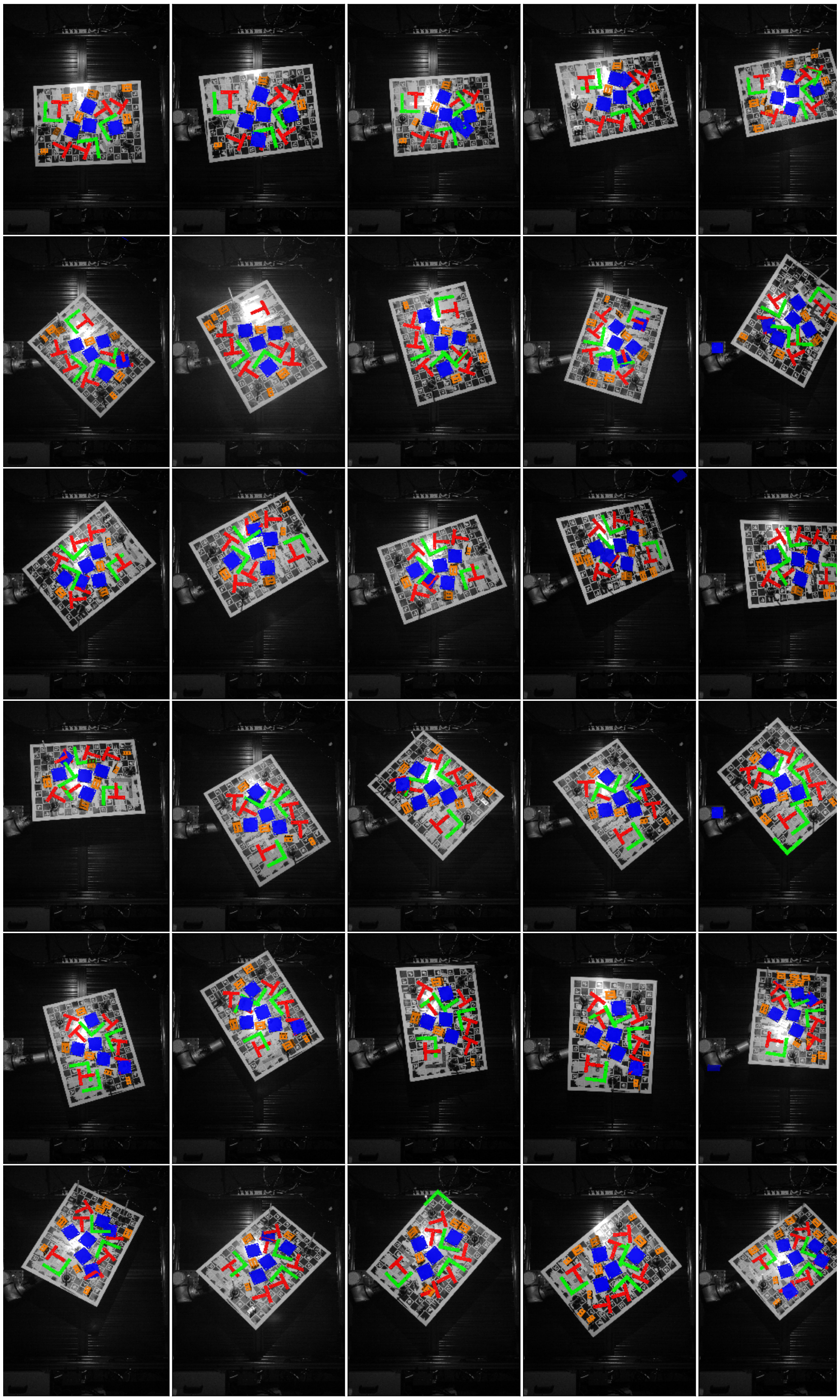

Figure 7. **The predictions from 30 scenes captured by Photoneo that are used for robot consistency calculation.** The above shows each of the 30 captures from the perspective of a Photoneo camera with pose predictions rendered on top. Here, the pose prediction incorporated the depth map processed from Photoneo's point cloud.

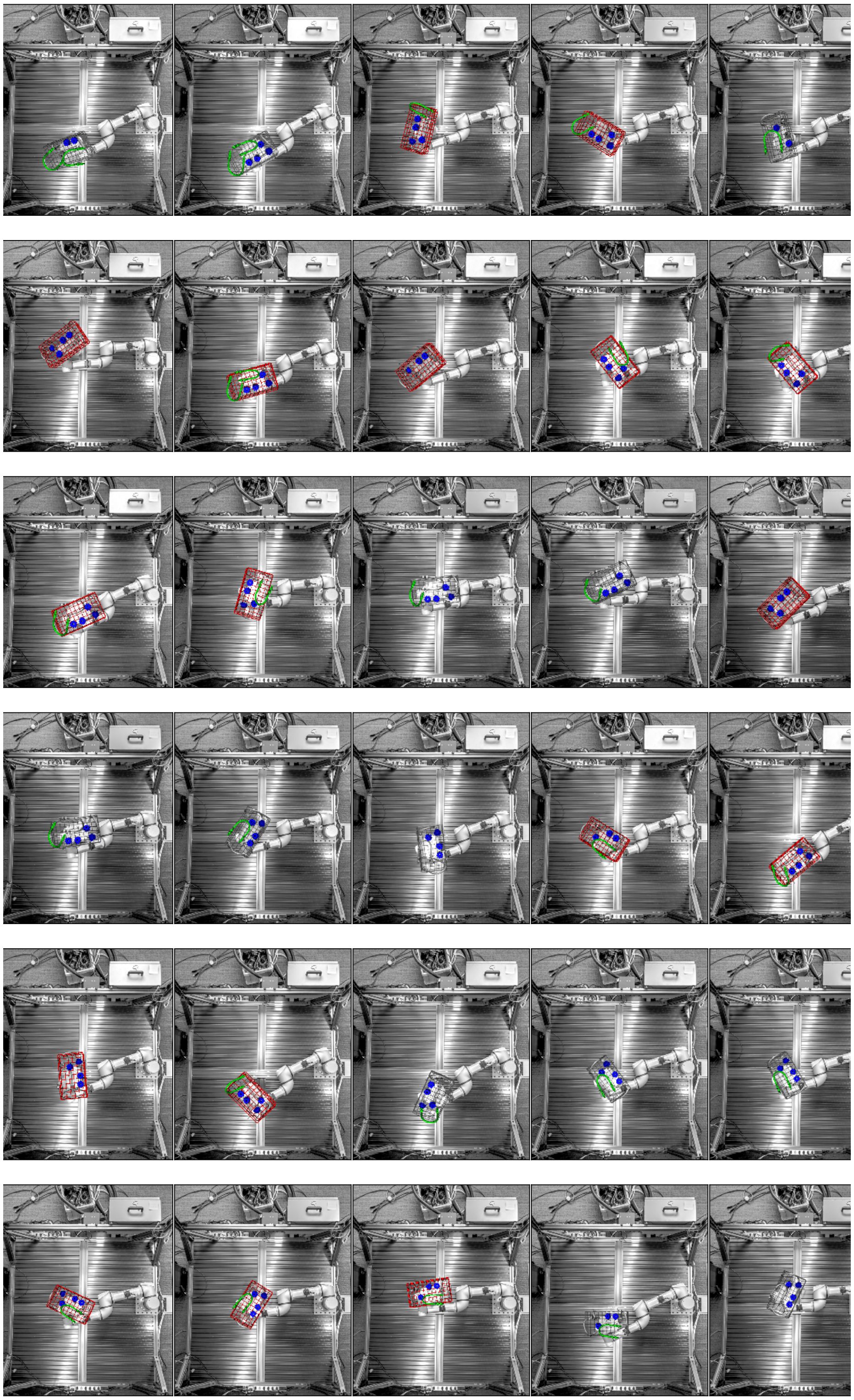

Figure 8. **The predictions from 30 scenes captured by multi-view FLIR that are used for robot consistency calculation.** The above shows each of the 30 captures from the perspective of a FLIR grayscale camera with pose predictions rendered on top. The poses were calculated with all 4 cameras, including edge refinement using AOLP/DOLP.

Camera	Basler LR			Basler HR			FLIR-monoP			Photoneo		
Part	MVD	Recall	Precision	MVD	Recall	Precision	MVD	Recall	Precision	MVD	Recall	Precision
Corner bkt	2.068	0.64	0.98	1.218	0.77	1.00	2.677	0.72	0.99	3.275	0.63	0.66
Corner bkt 0	3.096	0.54	0.88	2.256	0.68	0.98	3.256	0.40	0.95	3.206	0.53	0.58
Corner bkt 1	6.982	0.47	0.88	2.356	0.87	0.93	8.005	0.61	0.97	10.801	0.74	0.72
Corner bkt 2	11.064	0.47	0.77	4.176	0.62	0.87	10.625	0.29	0.85	11.924	0.57	0.63
Corner bkt 3	3.573	0.52	0.87	1.813	0.75	0.97	3.054	0.42	0.95	5.344	0.34	0.51
Corner bkt 4	2.275	0.80	0.95	1.263	0.85	0.96	2.909	0.80	0.95	3.315	0.78	0.75
Corner bkt 6	3.185	0.65	0.70	1.491	0.90	0.84	4.118	0.78	0.87	16.032	0.59	0.29
Gear 1	2.941	0.85	0.92	1.553	0.99	1.00	2.831	0.98	1.00	6.931	0.72	0.67
Gear 2	2.580	0.99	0.83	1.652	1.00	1.00	2.521	0.98	1.00	6.521	0.93	0.63
L bkt	2.476	0.85	0.98	1.476	0.81	0.98	3.124	0.82	0.99	2.912	0.86	0.95
Handrail bkt	5.546	0.31	0.95	2.087	0.72	0.99	4.481	0.49	0.96	8.700	0.36	0.75
Hex manifold	2.359	0.84	0.99	1.883	0.87	1.00	2.713	0.81	1.00	1.633	0.57	0.98
Oblong float	4.574	0.97	0.99	3.141	0.95	1.00	4.952	0.85	0.97	8.128	0.06	0.29
Pegboard basket	6.280	0.27	0.86	2.435	0.42	0.97	3.681	0.21	0.97	10.201	0.14	0.74
Pipe fitting	12.612	0.90	0.97	7.784	0.92	0.98	13.184	0.72	0.98	6.217	0.15	0.25
Single pinch clamp	6.476	0.72	0.77	4.312	0.89	1.00	8.101	0.78	1.00	4.333	0.83	0.64
Square bkt	4.411	0.78	0.98	2.622	0.88	0.99	4.310	0.79	0.99	2.978	0.84	0.77
T bkt	2.773	0.97	0.98	1.683	0.96	1.00	3.254	0.94	1.00	3.187	0.95	0.98
U bolt	6.018	0.80	0.98	5.497	0.75	0.99	6.557	0.72	1.00	5.033	0.77	0.99
Wraparound bkt	2.430	0.83	0.96	1.683	0.85	1.00	2.917	0.85	1.00	6.476	0.71	0.83

Table 1. **Performance on different parts across cameras**. We included the aggregated version of this table in Table 2b of the main paper.