Supplemental Materials

Anonymous submission

1 Methods

1.1 Obtaining AI Estimates

For objects of size L, the relationship between the number of segmented objects n_i and ground truth y_i is shown in Figure 1. Figure 2 shows how the transformation approximates this relationship.

2 Evaluation

2.1 Prior Distributions

For parameters, we have $b_{hi}, v_{hi}, b_{mi}, v_{mi} \sim \mathcal{N}(0, 10)$. For ground truth y_i, it only appears in at most two data-generating processes: one for the human estimate and one for the AI estimate. Since the amount of data that can be used to compute its posterior is small, we choose a uniform prior $f(y_i) \propto 1$.

2.2 Additional Images for the AI

We generated a separate set of images to train the AI models so all images given to humans would be new to it. There is one potential issue with this newly generated training data set: the learned models may not generalize well if the images tend to have certain features. For example, if most images in the training data set contain almost empty jars, then it can be hard for the trained AI to process a new image where the jar is full. We now explain how this issue is addressed.

All images were created by simulating dropping objects into a jar and this image generation process contains randomness. We cannot control how many objects will end up in the jar, so we generated a large amount of images and then only kept some of them as the final training data. First, we confirm that the generated images have the aforementioned issue. For every shape, we plotted the histogram of the number of objects in the jar, shown in Figure 3. We can see that the empirical distribution for the images is not uniform. For example, if we consider disk (L), it is more likely for the AI to encounter a jar with a large number of disks in the training data set. Our solution is to discard some of those images that contain a lot of disks so the histogram for the remaining images would become uniform. We repeated this step for all shapes. The remaining 1408 images were used to train the AI transformation model and data-generating process. There are 223 images for cylinder (S), 235 for cylinder (L), 250 for disk (S), 250 for disk (L), 250 for sphere (S), and 200 for sphere (L). The number of objects in the jar is between 54 and 170 for cylinder (S), 8 and 50 for cylinder (L), 5 and 40 for disk (S), 5 and 25 for disk (L), 55 and 1062 for sphere (S), and 50 and 857 for sphere (L).

3 Additional Results

Relative improvement (RI) results compared to various baselines are presented in Figures 4 to 6. There are some common patterns across all figures and we use Figure 4 as an example. We see horizontal lines for disks and vertical lines for spheres. When humans have a 0° view of disks, they produce accurate estimates and greatly help the AI, so there is a large improvement regardless of which angle the AI has. For spheres, the AI performs much better than humans. No matter which angle the human has, the AI mostly relies on its own estimate, so the relative improvement does not change much as long as the AI has the same viewing angle. For cylinders, the pattern exists but is not as clear.

4 Closed-form Solutions

4.1 Predictions

We first derive the closed-form solution for human predictions. Recall that for an image i, the ground truth is y_i number of objects, the human estimate is y_{hi}, their viewing angle is $a_{hi} \in [0, 90]$, and we define bias $b(a_{hi}; b_h)$ and variance $v(a_{hi}; v_h)$ as functions of the viewing angle parameterized by b_h and v_h. We have the following data-generating process for human estimates:

$$y_{hi} \mid y_i, b_h, v_h \sim \mathcal{N}(b(a_{hi}; b_h) + y_i, v(a_{hi}; v_h)) \quad (1)$$

Using this, we can compute the posterior distribution of the ground truth $f(y_i \mid y_{hi}, b_h, v_h)$, and then take the posterior mode as the prediction. With a uniform prior on y_i, its posterior is proportional to the likelihood. We then recognize that its posterior kernel leads to a normal distribution on y_i:

$$f(y_i \mid y_{hi}, b_h, v_h) \propto \exp\left(-\frac{(y_{hi} - b(a_{hi}; b_h) - y_i)^2}{2v(a_{hi}; v_h)}\right) \quad (2)$$

$$y_i \mid y_{hi}, b_h, v_h \sim \mathcal{N}(y_{hi} - b(a_{hi}; b_h), v(a_{hi}; v_h)) \quad (3)$$

The posterior mode is the mean of this normal distribution. AI predictions can be obtained similarly.

Next we show how to compute the posterior of y_i given estimates from two agents. The posterior mode will be the combined prediction. The idea is reducing the problem to sensor fusion with normally distributed noise (Becker 2015). From the data-generating process, we have:

$$y_{hi} - b(a_{hi}; b_h) \mid y_i, b_h, v_h \sim \mathcal{N}(y_i, v(a_{hi}; v_h)) \quad (4)$$

$$y_{mi} - b(a_{mi}; b_m) \mid y_i, b_m, v_m \sim \mathcal{N}(y_i, v(a_{mi}; v_m)) \quad (5)$$

The interpretation is that after we remove the bias from the two estimates, we get noisy measurements of the same ground truth. We further simplify the notation following this insight: $y_1 = y_{hi} - b(a_{hi}; b_h), v_1 = v(a_{hi}; v_h), y_2 = y_{mi} - b(a_{mi}; b_m), v_2 = v(a_{mi}; v_m)$. When computing the prediction, we fix the parameters to their posterior mode and put a uniform prior on the truth, so we effectively have a frequentist setting. We can now apply the formula in (Becker 2015) to obtain:

$$y_i \mid y_{hi}, b_h, v_h, y_{mi}, b_m, v_m \sim \mathcal{N}\left(\frac{y_1 v_2 + y_2 v_1}{v_1 + v_2}, \frac{v_1 v_2}{v_1 + v_2}\right) \quad (6)$$

4.2 Bias and Variance

In our model, bias and variance are both functions of the angle. Once we learn these functions, we can make inferences for new angles beyond the five angles available in our data set. Alternatively, if we are willing to sacrifice this ability to generalize, we can also obtain closed-form solutions for bias and variance for any fixed angle.

From now on, we consider a subset of the data set that only contains images for a certain shape viewed from a fixed angle, such as all cylinder (S) images viewed from 0°. Suppose there are n images in this subset. We use the human estimates as an example to show the derivation. Let $b_1 = b(a_{hi}; b_h)$. The data-generating process is:

$$y_{hi} \mid y_i, b_h, v_h \sim \mathcal{N}(b_1 + y_i, v_1) \quad (7)$$

$$f(y_{hi} \mid y_i, b_h, v_h) = \frac{1}{\sqrt{2\pi v_1}} \exp\left(-\frac{(y_{hi} - (b_1 + y_i))^2}{2v_1}\right) \quad (8)$$

$$= \frac{1}{\sqrt{2\pi v_1}} \exp\left(-\frac{((y_{hi} - y_i) - b_1)^2}{2v_1}\right) \quad (9)$$

Hypothetically, if we had samples $y_{hi} - y_i \sim \mathcal{N}(b_1, v_1)$ in the frequentist setting, we would obtain the same probability density function as above, and taking the product over i would lead to the same likelihood function. Furthermore, if we put uniform priors on b_1 and v_1 in the Bayesian setting, the posterior will equal the likelihood, and the maximum a posteriori estimate (MAP) will equal the maximum likelihood estimate (MLE). So, this can be reduced to a frequentist problem where we have samples generated from a normal distribution and want the maximum likelihood estimate (MLE) of the mean and variance (Casella and Berger 2002). The closed-form formulas are as follows:

$$\hat{b_1} = \frac{\sum_i (y_{hi} - y_i)}{n} \quad (10)$$

$$\hat{v_1} = \frac{\sum_i (y_{hi} - y_i)^2}{n} \quad (11)$$

We can verify that these frequentist solutions are close to our solutions and the results are in Figure 7 and Figure 8. We also see that the closed-form solutions sometimes deviate from our solutions. Our solution is constrained since bias and variance have to be a cubic function of the angle. But in some cases, the relationship between the closed-form bias or variance and the angle cannot be captured by a cubic function. Another possible reason is the use of priors. The closed-form solutions assume uniform priors, while we use normal priors in our evaluation.

4.3 Weighted Average Baseline

We show how to obtain the closed-form solution for the optimal weight used in the weighted average baseline. Recall that for an image i, $\hat{y}_{hi}$ is the human prediction and $\hat{y}_{mi}$ is the AI prediction. Let w be the weight given to the human prediction. The weighted average of these two predictions is $\hat{y}_i = w\hat{y}_{hi} + (1-w)\hat{y}_{mi}$. Since MSE is a quadratic function of w, we set the first derivative to 0:

$$\text{MSE} = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 \quad (12)$$

$$= \frac{1}{n}\sum_i (y_i - (w\hat{y}_{hi} + (1-w)\hat{y}_{mi})^2 \quad (13)$$

$$\frac{d\text{MSE}}{dw} = \frac{2}{n}\sum_i (y_i - (w\hat{y}_{hi} + (1-w)\hat{y}_{mi}) = 0 \quad (14)$$

$$\hat{w} = \frac{\sum_i (y_i - \hat{y}_{mi})(\hat{y}_{hi} - \hat{y}_{mi})}{\sum_i (\hat{y}_{hi} - \hat{y}_{mi})^2} \quad (15)$$

References

Becker, A. 2015. Sensor Fusion with Normally Distributed Noise.

Casella, G.; and Berger, R. L. 2002. *Statistical inference*. Pacific Grove, Calif: Duxbury, 2. ed edition. ISBN 978-0-534-24312-8.

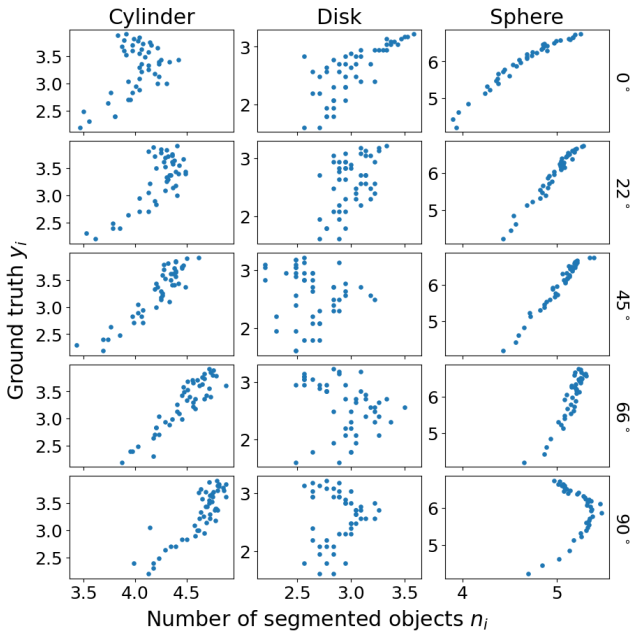

Figure 1: The relationship between the number of segmented objects n_i and ground truth y_i in a log-log plot. Each row corresponds to a viewing angle shown on the right and each column corresponds to one shape. Only cylinders, disks, and spheres of size L are plotted here. For a fixed shape, y_i can be approximated by polynomials of n_i and the angle a_{mi}. This trend is learned in the transformation stage.

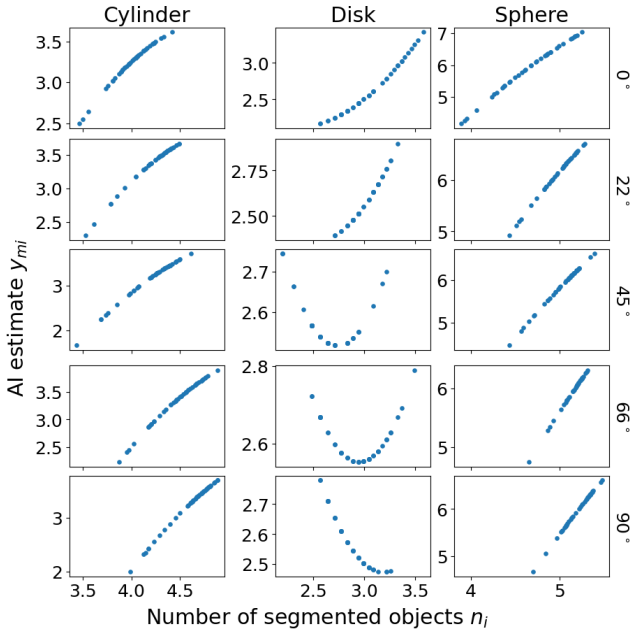

Figure 2: How the number of segmented objects n_i is transformed into the AI estimate y_{mi}. Only cylinders, disks, and spheres of size L are plotted here. For each shape, we learn a family of quadratic functions involving n_i and the angle a_{mi} to capture the relationship between n_i and the ground truth y_i.

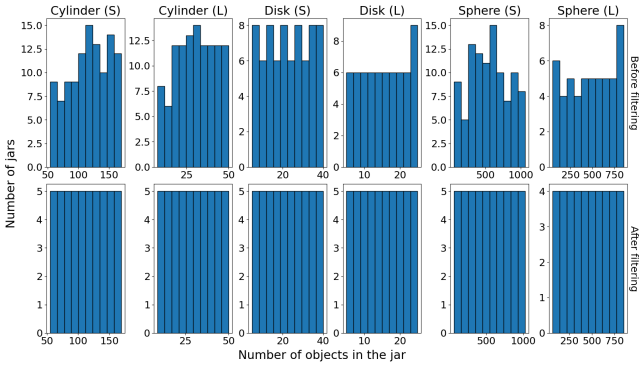

Figure 3: Histograms of the number of objects in the jar. Each column corresponds to one shape. The top row is the distribution for the newly generated images, and the bottom row is the distribution after discarding some images. Each jar can be viewed from five angles but the number of objects in it will remain the same, so only images with a 90° view are used for the plot.

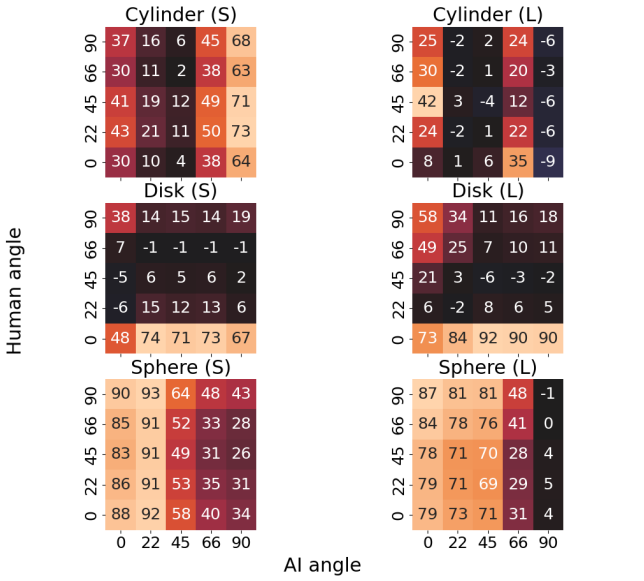

Figure 4: Relative improvement (RI) in percentage over average baseline. S = small; L = large. Each subplot is the result for a certain type of object in the jar. Diagonal entries correspond to when the human and AI have the same viewing angle. Off-diagonal entries correspond to when the human and AI have different viewing angles.

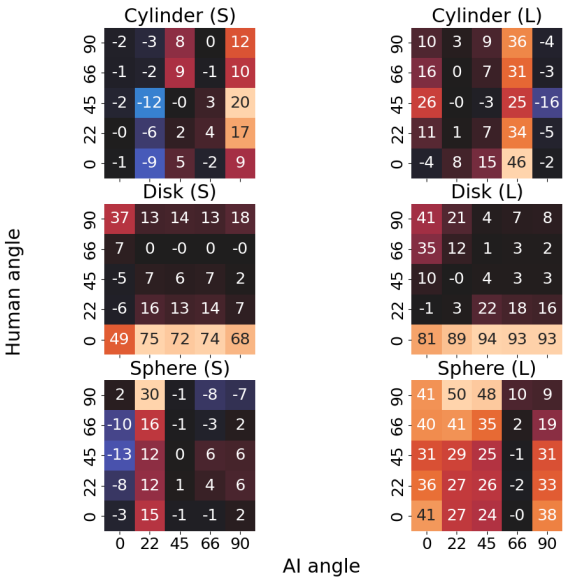

Figure 5: Relative improvement (RI) in percentage over weighted average baseline.

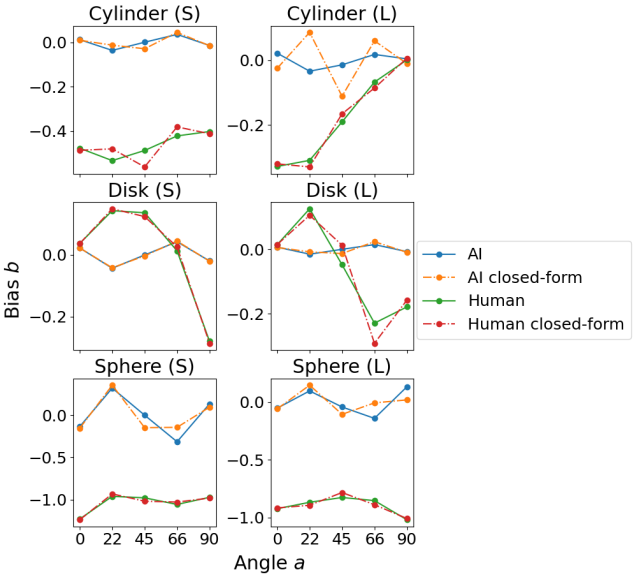

Figure 7: Bias learned by our model and bias computed using closed-form formulas. The results are broken down by shape. Solid lines are the values in our model. Dash-dotted lines are the results of closed-form formulas.

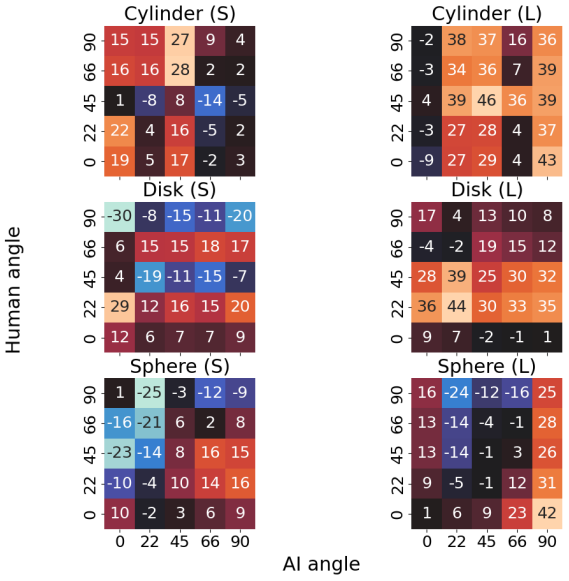

Figure 6: Relative improvement (RI) in percentage over pick the best baseline.

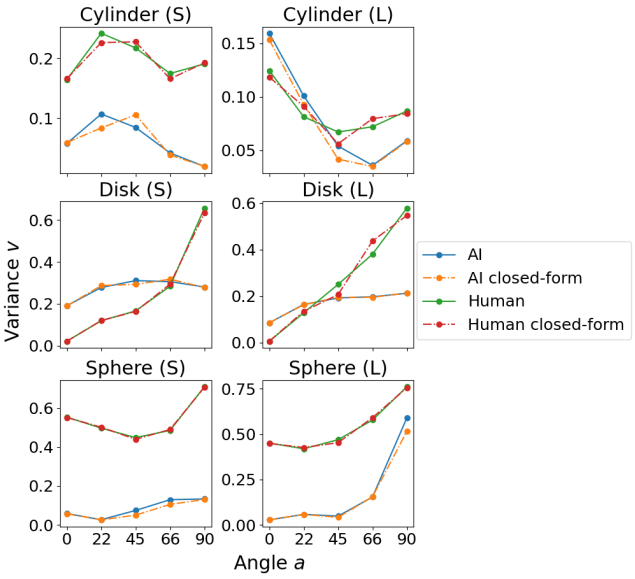

Figure 8: Variance learned by our model and variance computed using closed-form formulas. The results are broken down by shape. Solid lines are the values in our model. Dash-dotted lines are the results of closed-form formulas.