Supplementary Material

Here we provide the model architecture and training details for our CEN models along with additional experimental results.

A. Model Parameters

For both datasets, the worker and context encoders are fully connected neural networks consisting of two hidden layers each with 200 neurons with ReLU activations. The image encoder has one hidden layer with 200 units and outputs a K dimensional embedding vector. For the CELEBA dataset, the embedding dimension K is set to four since we provide the workers with four different attributes to cluster on. For the RETINA dataset, we set $K = 10$ since we do not know *a priori* how many different attributes the workers will use. We jointly train the three encoders with a mini-batch size of 100 using ADAM with $\alpha = 0.001, \beta_1 = 0.9$, and $\beta_2 = 0.999$. We experimented with various learning rates $\alpha \in \{0.00001, 0.0001, 0.001, 0.01\}$ and found that the CEN performance was robust to these variations. The regularization constants are set to $\lambda_1 = 5\text{E}-6$ and $\lambda_2 = 0.001$. We experimented with $\lambda_1 \in \{1\text{E}-6, 5\text{E}-6, 1\text{E}-5, 5\text{E}-5\}, \lambda_2 \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$ and saw that the prediction accuracy decreases when $\lambda_1 > 5\text{E}-6, \lambda_2 > 0.001$, but relatively stable otherwise. Hence, we choose the largest possible learning rate to reduce training time.

The positive margin, negative margin, and positive similarity weight are each set to $\xi_p = 1, \xi_n = 6$ and $\gamma = 5$, respectively. The prediction accuracy decreased when $\xi_n/\xi_p < 4$. We experimented with varying positive similarity weights $\gamma \in \{1, 2, ..., 10\}$ and found that $\gamma = \{4, 5, 6\}$ achieves similar best prediction accuracy when trained on the full dataset. In Table S1 we show the impact of γ on the label prediction accuracy for both datasets. The optimum value of γ should be expected to change depending on the variance in the level of detail workers cluster grids. Models were trained for 20 epochs which we determined to be sufficient for learning interpretable embeddings. When utilizing all the data from 620 HITs, the training time was on average 2.5 minutes for both datasets running on CPUs (Macbook Pro 13-inch, Late 2012, 2.5 GHz Intel Core i5, 8GB RAM, Apple, CA, USA). Upon publication we will make the code for our GUI and CEN model available.

Table S1. **Impact of positive similarity weight** γ on label prediction accuracy. $\gamma = 6$ was used for all results presented in the main paper.

	$\gamma = 1$	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	$\gamma = 10$
CELEBA	68.5%	69.8%	69.8%	69.3%	69.2%
RETINA	68.1%	69.3%	69.4%	69.1%	68.7%

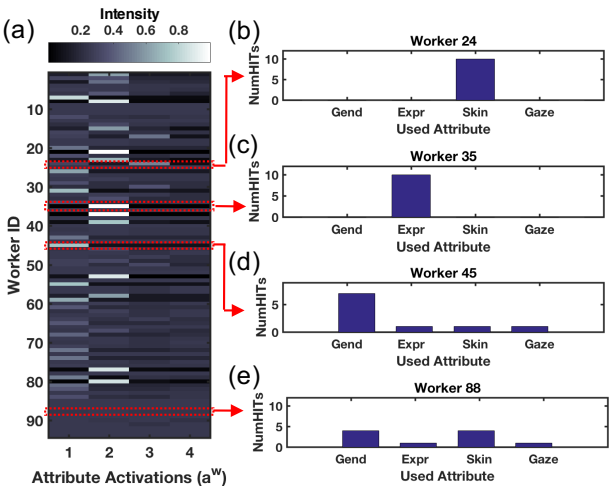

Figure S1. **Visualizing the worker model.** On the left we see the predicted attribute activation vectors for each worker from the CELEBA dataset. Attribute dimension labels were inferred from worker annotations. Brighter colors indicate a stronger preference for a given attribute. On the right we show the actual attributes used by a set of representative workers inferred from their text annotations. We can see that our worker attribute predictions are consistent with the actual attributes used by the workers.

B. Interpretation of the Worker Model

We explore the learned attribute activation vectors a^w for each worker to examine if their prior biases were captured by the worker encoder. The output attribute activation vectors for each of the 94 workers are shown in Fig. S1(a) as a stacked heatmap. On the right side of Fig. S1, we show the distribution of attributes that four representative workers have used over the course of performing ten HITs, inferred from their text annotations. Fig. S1(a) shows that our model predicts a high activation in a^w_3 for worker 24. In Fig. S1(b) we can see that this worker consistently used the skin color attribute for all ten HITs they performed. This indicates that worker 24 had a strong prior bias towards grouping based on skin color and was unaffected by the different contexts formed by the grid. Note that it is highly unlikely that all ten randomly generated grids shown to worker 24 highlighted the skin color attribute. Fig. S1(c) shows that worker 35 relied mainly on the expression attribute. Similarly for worker 45 we observe a strong bias towards the gender attribute as the worker encoder outputs a high activation for a^w_1. Worker 88 used a variety of attributes suggesting that they are more sensitive to the context provided by the grid. We observe a near uniform attribution activation vector for this worker.

C. Comparison of the Joint Embeddings

We compare the quality of the embeddings for the CELEBA dataset produced by the CEN-mixture model with those learned by baseline approaches that do not model context. We project the four dimensional joint embedding vec-

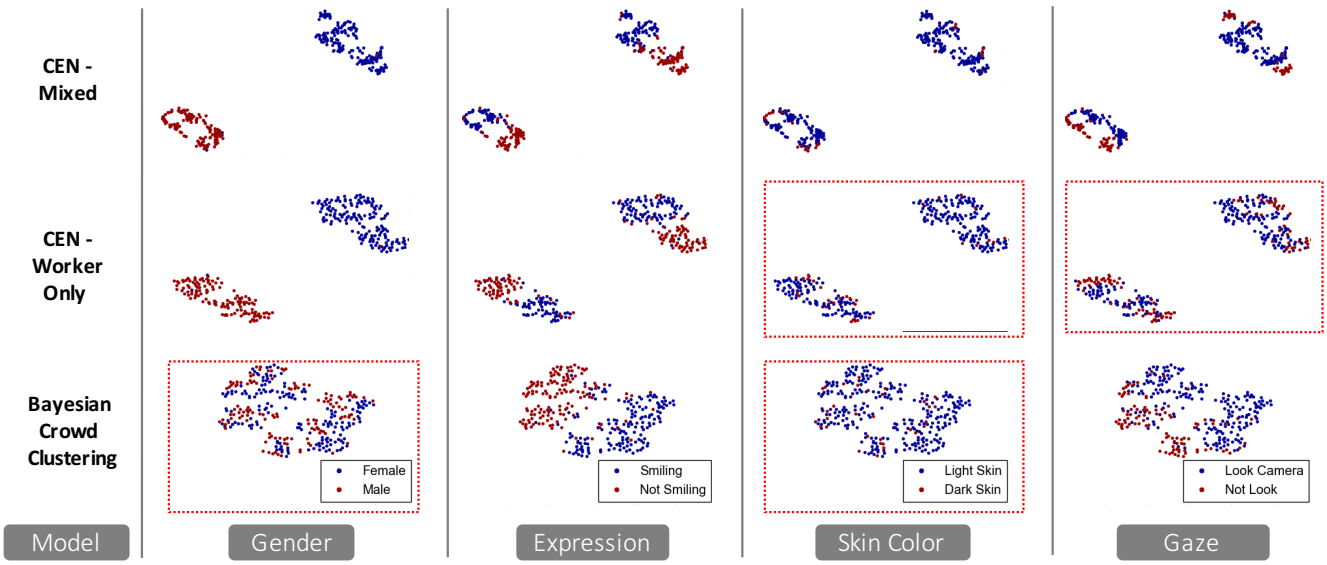

Figure S2. **Comparing embedding quality.** 2D t-SNE projections of the four dimensional joint embedding space for three embedding models on the CELEBA dataset. Colors denote binarized ground truth categories for each of the four attributes: gender, expression, skin color, and gaze direction. Dotted red boxes highlight attributes for which the CEN-mixture model produces more compact embeddings compared to the baselines.

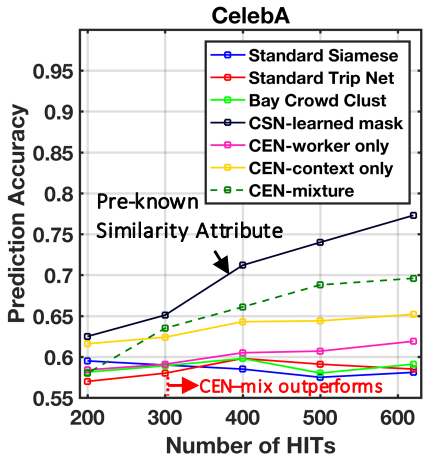

Figure S3. **Held-out Label Prediction on CELEBA**. Prediction accuracy on held out labels for the CELEBA dataset plotted against the amount of available data during training.

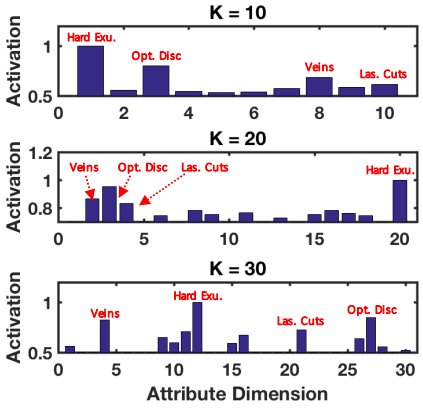

Figure S4. **Varying the embedding dimension for the RETINA dataset**. We plot the average activations for each dimension of a^m produced by the CEN-mixture model for three different values of embedding dimension K. Red labels were inferred from the text annotations.

tors x_i down to two dimensions using t-SNE [22] and color code each point according to its ground truth attribute. For each attribute, the ground truth categories were binarized for simplicity, i.e. smiling vs not smiling. In Fig. S2 we show the low dimensional embeddings learned by the CEN-mixture model, CEN-worker only model, and the Bayesian Crowd Clustering baseline. The CEN-mixture model better separates the ground truth categories in the embedding space. This shows the positive impact of modeling context. The worker encoder only model finds well separated embeddings along the gender and expression attribute (which are relatively easy to distinguish) but does not perform well on the skin color and gaze attributes (which are attributes that workers more often disagree on). The Bayesian Crowd Clustering baseline has difficulty separating the gender and skin color attributes.

D. Heldout Label Predictions on CELEBA

Fig. S3(a) shows the pairwise prediction accuracy for each model plotted against a varying number of training samples for the CELEBA dataset. Standard Siamese Networks and Triplet Networks fail to capture the multiple attributes used to cluster the images and have the lowest prediction accuracy of 58.1% and 58.5%. The Bayesian Crowd Clustering method slightly improves on that with an accuracy 59.1%. The worker only variant of our model achieves a prediction accuracy of 62.1%. This is superior to Siamese Networks and Bayesian Crowd Clustering but still fails to

capture the tendency of workers to shift their clustering criterion based on the context highlighted by images in the grid. The context only model variant performs substantially better with a prediction accuracy of 65.2%. This indicates that the context information is indeed influencing the worker's decisions. Finally, the CEN-mixture outperforms all previous baselines with a prediction accuracy of 69.8% (75.1% when trained on noiseless labels). The CSN model with learned masks obtains the highest accuracy of 77.3%, but it is important to note that this model was trained on triplets pre-labeled with the true similarity attributes used to cluster them. The CEN-mixture model achieves strong predictive performance without any prior knowledge of the similarity attributes.

E. Prior Number of Attributes

For the RETINA dataset, we do not know the number of number of attributes the workers will use. Hence, we set $K = 10$ which serves as our prior guess of an upperbound on the number of attributes the workers are going to use. Although the attribute vector dimension was set to $K = 10$, we observed that four dimensions were consistently highly activated across different values of K. In Fig. S4 we see that the attribute dimensions we selected are the four most highly activated dimensions of a^m for $K = 10, 20$, and 30.

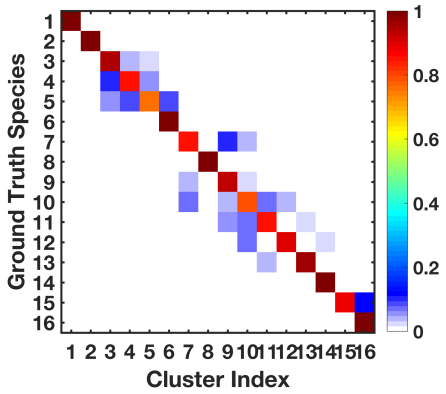

Figure S5. **Confusion plot for the BIRDS dataset**.

F. Clustering Learned Embeddings

For the BIRDS dataset, we perform K-means clustering on the learned 4 dimensional embedding space with $K = 16$ and compare the ground truth bird species of an image with its assigned cluster. To quantify the agreement between the ground truth species and the learned clusters, we use the multi-class version of Matthew's Correlation Coefficient (MCC) [8], where MCC $= 1$ indicates perfect prediction, and a value between -1 and 0 denotes total disagreement depending on the true distribution.

$$\text{MCC} = \frac{\sum_{k,l,m=1}^{N} C_{kk}C_{ml} - C_{lk}C_{km}}{\sqrt{\sum_{k=1}^{N}\sum_{l=1}^{N} C_{lk} \sum_{\substack{f,g=1 \\ f\neq k}}^{N} C_{gf}} \sqrt{\sum_{k=1}^{N}\sum_{l=1}^{N} C_{kl} \sum_{\substack{f,g=1 \\ f\neq k}}^{N} C_{fg}}} \tag{9}$$

The confusion plot in Fig. S5 reveals high correlation (MCC $= 0.914$) between the ground truth species and the learned clusters, suggesting that the CEN is able to make fine-grained distinctions amongst bird species despite highly noisy training data (25.6% for the BIRDS dataset).

To show that the learned embeddings are useful for fine-grained classification tasks, we trained a CNN with the cluster assignments as the image category labels and compared the resulting accuracy when training on the ground truth labels (900 images for training and 100 for testing). Our 'embedding label' CNN resulted in a test accuracy of 68.1%, while the ground truth CNN produced an accuracy of 76.2%.

G. Limitations

If workers do not have a diverse set of abilities it will be challenging to learn embeddings that capture all subtle variations in a given dataset. However, in our experiments we observed that MTurkers discovered small distinctions in challenging domains e.g. retina images and bird species (see Fig. 7, 8)

Our context model assumes that the ordering of the images in the grid does not effect clustering behavior. In practice this may have some effect on the workers. To produce disentangled attribute vectors we assume that a majority of the grids are clustered along a single attribute. However, from our experiments we observe that this only has to be very weakly satisfied as many workers used a mixture of attributes.