Supplementary Material
Ground-truth or DAER: Selective Re-query of Secondary Information

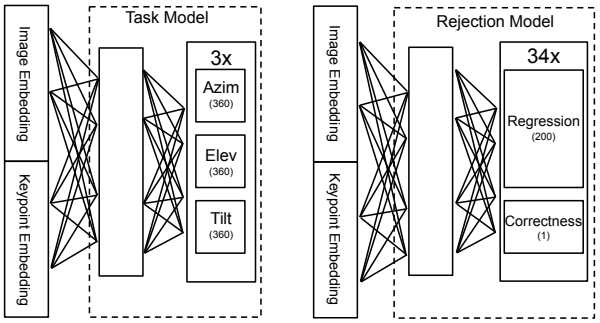

Figure 1: Prediction layers for the KCVE task (left) and rejection (right) models, which accept the keypoint and image embeddings from [3].

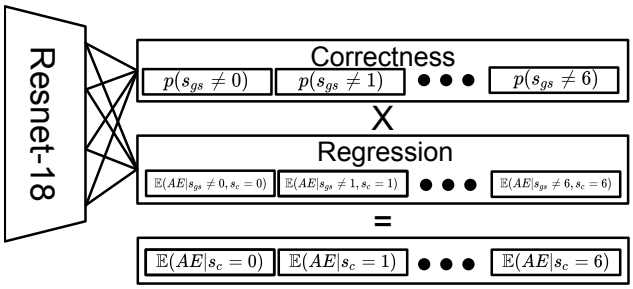

Figure 2: Our rejection model architecture and output format for the HSC task. Each potential candidate seed is given two outputs, which are multiplied to estimate the expected additional error.

1. Architecture and Training Details

1.1. Keypoint-Conditioned Viewpoint Estimation

Architecture The Click-Here CNN (CH-CNN) architecture consists of separate, mostly independent, branches to process the image and keypoint. The features produced by these branches are concatenated and passed through two linear layers to produce the desired output. As this architecture has been proven capable of integrating keypoint and image data, we use it not only for the task model, but also as the backbone of the rejection model with the output layers shown in Figure 1. Further information on the base architecture is available in the original work [3].

The output of the task model (Figure 1-left) is of size 3x3x360, consisting of three vehicle classes (car, bus, motorbike), three angles (azimuth, elevation, tilt) and 360 potential angle values. The output of the rejection model (Figure 1-right) is of size 34x(200+1), consisting of 34 potential keypoint classes, 200 binned outputs per keypoint class to regress the additional error, and one output per keypoint class to estimate the correctness.

Training The rejection model is trained in two phases. In the first phase, it is trained on a combination of rendered [3] and real [4] data. Candidate seeds are generated by randomly selecting an x-y location on the image. An Adam optimizer [2] is used with learning rate $1e^{-4}$ and early stopping is performed on the validation loss with a patience of 5 epochs.

In the second phase, the rejection model is trained exclusively on the PASCAL3D+ dataset [4]. The same optimizer settings are used, however the one-hot additional error target is softened by convolving with a Gaussian kernel with standard deviation 3. Early stopping is performed on validation loss with a patience of 100 epochs.

Regression and correctness ablations use the same training procedure, where back propagation is only performed on the appropriate loss. For the no seed ablation, a tensor of zeros is given to the rejection model in place of the keypoint map, and no further modifications are made to architecture or training.

1.2. Hierarchical Scene Classification

The architecture and output layers used for the hierarchical scene classification task are shown in Figure 2. As a backbone, we use a ResNet-18 which has been pretrained on ImageNet [1], and truncate the output to 2 elements per seed class (14 total). Seven of these outputs—the correctness outputs—are trained using a cross-entropy loss to determine

$$p(s_{gs} \neq \text{class}|x) = 1 - p(s_{gs} = \text{class}|x) \ , \qquad (1)$$

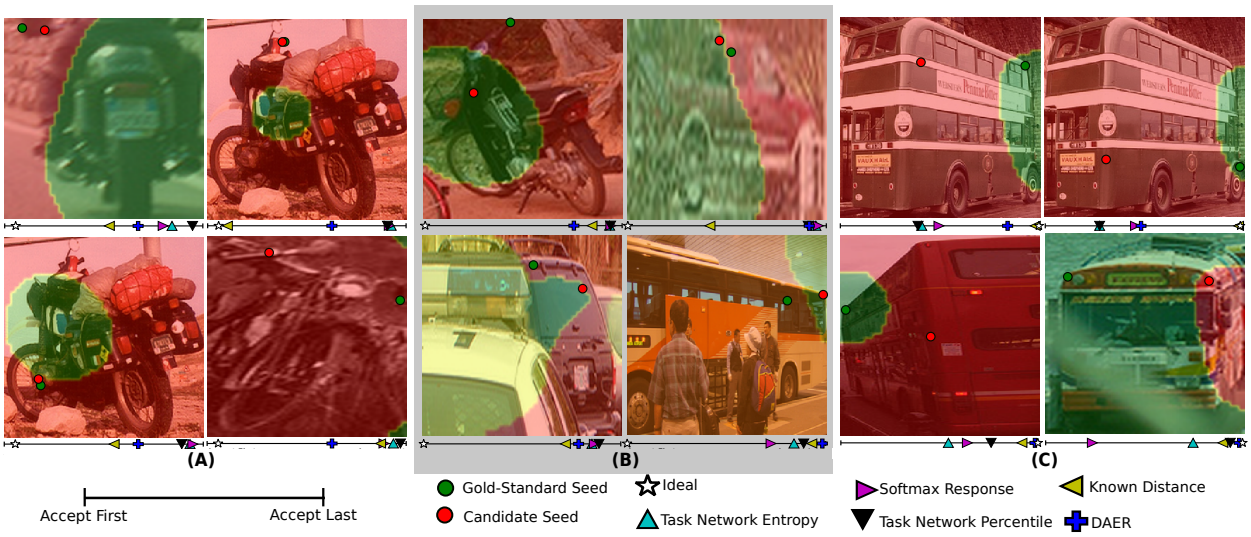

Figure 3: Quantitatively chosen KCVE heatmaps. The gold-standard seed is shown in green, the candidate seed is shown in red, and a red-yellow-green heatmap gives the additional error for that keypoint click. Methods closer to the white "ideal" star are better for that example. (A) The four cases where the gold-standard seed provided the worst absolute performance. (B) The four cases where the candidate seed improved upon the gold-standard seed the most. (C) The four cases with the highest additional error.

while the other seven are trained using a binary cross entropy to find

$$\mathbb{E}(\text{AE}\,|x, s_{gs} \neq \text{class}, s_c = \text{class}) \quad . \tag{2}$$

The model is trained for 50 epochs with learning rate 1e^{-5} and the model with the best validation AMAE is used for evaluation. The correctness-only rejection models, regression-only rejection models, and seeding models, are trained identically using only the appropriate outputs, except the learning rate is increased to $1e^{-4}$ and accuracy is used in place of AMAE to select the best seeding model.

Since the individual outputs in this architecture correspond to different seeds, the blind ablation was performed by reducing the output to a single value that regresses the additional error regardless of the input seed.

2. Quantitatively Chosen KCVE Examples

In the main text, we use qualitatively chosen examples to illustrate characteristics of the task and rejection models. Here, we show additional examples that were selected using quantitative criteria on our crowdsourced keypoints: the four cases where the gold-standard seed resulted in the highest geodesic error (Figure 3-A), the four cases where the candidate seed improved upon the gold-standard seed the most (Figure 3-B), and the four cases with the highest additional error (Figure 3-C).

Figure 3-A shows four cases where the performance of the candidate seed is poor but should be accepted, as we would expect a worker to continue returning seeds that are near the "target" gold-standard seed even though it won't result in better performance. We see that DAER outperforms baselines that do not have prior knowledge of the gold-standard seed location in all four cases by accepting these instances earlier.

In Figure 3-B, where the candidate seed improves upon the gold-standard, the rejection model must understand that despite returning a different answer than the gold-standard, the candidate seed does not make performance worse. Given the depth to which the rejection model must be able to understand the task model to make this distinction, it is unsurprising that no method clearly outperforms the others on these four samples.

Figure 3-C answers the most intuitive question of seed rejection: how well does a rejection model reject seeds with high additional error? We see that while using an oracle measure of distance is best in some instances, DAER outperforms all baselines that do not have prior knowledge of the correct answer in all four cases by accepting these instances last.

3. Crowdsourcing Keypoint Clicks

Keypoint annotations are collected from US-based annotators using the interface shown in Figure 4. The worker is shown an image containing one or more vehicles, and is asked to click all instances of a specific keypoint class. If an annotator responded that the keypoint class wasn't present, we provided the query to another annotator up to two ad-

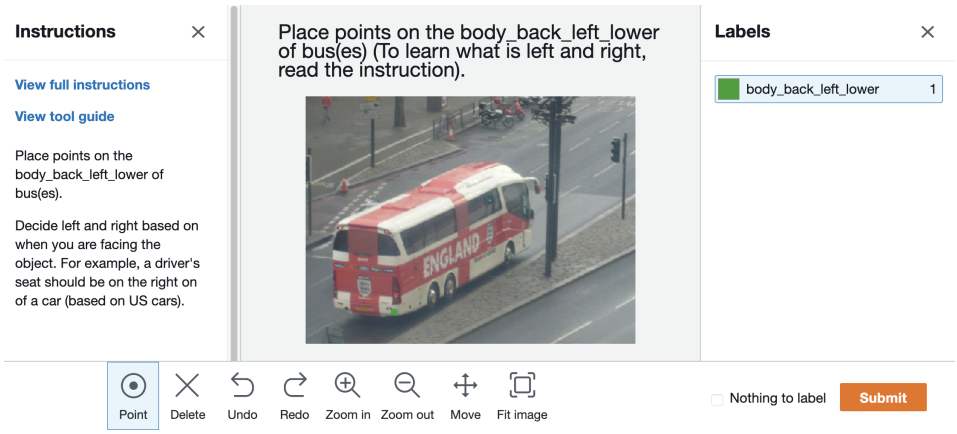

Figure 4: The interface provided to crowd workers for crowdsourcing keypoint clicks.

ditional times. If all three annotators responded that the keypoint class wasn't present, we assumed the gold standard was incorrect or too difficult, and removed it from the evaluation.

To match the annotated keypoints with the corresponding verified gold-standard keypoint from PASCAL3D+, we use a three-step process: first, we associate all keypoints to vehicle crops which contain them. Next, we match these keypoints to the gold-standard keypoint of the same class in that vehicle crop. Last, if a vehicle crop contains multiple candidate keypoints of the same class, we select the one that is nearest to the gold-standard keypoint. Using this process, we receive annotations matching 6,042 of the 6,593 gold-standard keypoints.

Analyzing the distribution of matched keypoints, we found that 40% of keypoints were within 5 pixels of the matching gold-standard and 57% were within 10 pixels of the matched gold-standard keypoint. We further found that 6.3% (381) of keypoints cause additional error, while 1.3% (81) cause more than 5° additional error, and 0.5% (30) cause more than 150° additional error.

Percentile	AMAE
70^{th}	0.3472
80^{th}	0.3092
90^{th}	0.3303

Table 1: AMAE at various sampling percentiles.

4. Evaluation of Sampling Method by Percentile

For the KCVE task, we consider a sampling-based baseline in which 10,000 samples are taken from the CH-CNN output distribution, and the sample at the n^{th} percentile distance from the mean is used as the scoring function. We present the results for the 70^{th}, 80^{th}, and 90^{th} percentile in Table 1, justifying our choice of the 80^{th} percentile as our baseline.

5. Per-Fold KCVE Results

Table 2 shows the per-fold AMAE of the various rejection models on the KCVE task. We see that no single method performs best across all folds, but DAER is the most consistent: DAER does not perform worse than 25.3% above its mean on any fold, while the corresponding number for the best baseline (sampling percentile) is 80.4%.

As each baseline only addresses one of the described subgoals (e.g., distance only finds the cause of error and sampler only understands model response), this suggests that some folds contain more instances of one source of error, and again highlights the importance of the subgoals described in the main paper. While focusing solely one subgoal allows baselines to perform well on folds where that source of error is more frequent, DAER's understanding of both subgoals leads to more consistent and overall better performance.

	Fold					
Method	1	2	3	4	5	Mean
Softmax Response	0.561	0.999	0.167	1.430	1.496	0.9306
Distance	0.419	**0.254**	0.147	0.757	0.405	0.3964
Entropy	0.325	0.556	0.112	0.421	0.353	0.3534
Sampler	**0.292**	0.558	0.125	0.370	**0.201**	0.3092
Correctness	0.312	0.343	0.118	0.414	0.282	0.2937
Regression	2.342	0.274	1.102	0.992	1.107	1.1633
No Seed	0.852	0.778	0.480	1.003	0.889	0.8002
DAER	0.322	0.307	**0.109**	**0.335**	0.359	**0.2864**

Table 2: Per-Fold AMAE on the KCVE task. Baselines are above the thick line, ablations and DAER are below. The best performer per-fold is shown in bold (lower is better).

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, Miami, FL, 2009. IEEE. 1

[2] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 2015 International Conference on Learning Representations*, pages 1–15, San Diego, California, Jan. 2017. 1

[3] Ryan Szeto and Jason J. Corso. Click Here: Human-Localized Keypoints as Guidance for Viewpoint Estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1604–1613, Venice, Oct. 2017. IEEE. 1

[4] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, Steamboat Springs, CO, USA, Mar. 2014. IEEE. 1