

## Supplementary material for “HIVE: Evaluating the Human Interpretability of Visual Explanations”

In this document, we provide additional details on some sections of the main paper. Code is available at <https://princetonvisualai.github.io/HIVE>.

**Section A:** We provide more information on the evaluation tasks.

**Section B:** We provide more information on the four evaluated methods and the modifications we made to their original explanation form.

**Section C:** We provide more information about our human studies.

**Section D:** We report additional results and analyses.

- **Section D.1:** We supplement Section 5.2 of the main paper and discuss the *agreement* study results with vs. without examples from the predicted class.
- **Section D.2:** We supplement Section 5.3 of the main paper and provide more details on our analysis with automatic evaluation metrics.
- **Section D.3:** We supplement Section 5.4 of the main paper and discuss the participants’ similarity ratings and decisions. We also provide a plot of the participant vs. ProtoPNet prototype similarity ratings.
- **Section D.4:** We supplement Section 5.5 of the main paper and provide the full results of our subjective evaluation.
- **Section D.5:** We supplement Section 5.6 of the main paper and provide the full results of our interpretability-accuracy tradeoff study.

**Section E:** We show the simple decision tree model for fruit classification we used to introduce ProtoTree.

**Section F:** We show snapshots of our full user interface.

## A Details on the evaluation tasks

**Agreement task.** For each image, we show one model prediction-explanation pair and ask the participants how confident they are in the model’s prediction. We show 10 images in total (5 correct, 5 incorrect predictions in random order). Participants rate their confidence in the given prediction on a 4-point scale (1: confident prediction is incorrect, 2: somewhat confident prediction is incorrect, 3: somewhat confident prediction is correct, 4: confident prediction is correct).

**Distinction task.** For each image, we show four model prediction-explanation pairs for it (in random order) and ask the participants to identify the correct prediction based on the explanations. For GradCAM [10] and BagNet [1], participants are tasked with 10 sample images (5 correct and 5 incorrect predictions), each of which is shown with four heatmaps. On correctly predicted samples, the four heatmaps correspond to the top-4 predicted classes. On incorrectly predicted ones, we show heatmaps for the top-3 predicted classes and the heatmap of the ground-truth class. For ProtoPNet [2], we show four correctly predicted samples in total. Each sample is presented with four explanations corresponding to the top-4 predicted classes. We reduce the total number of samples and focus on correctly predicted samples due to the complexity of the ProtoPNet

explanations; even with this change, the ProtoPNet study duration is twice as long as that of GradCAM and BagNet. For ProtoTree [7], we show 10 correctly predicted samples in total and ask participants to select the correct decisions on the two final nodes which lead to four ( $2^2$ ) different predictions.

Additionally for ProtoPNet [2] and ProtoTree [7], we ask participants to rate the similarity of prototype-region pairs in both tasks using a 4-point Likert scale (1: not similar, 2: somewhat not similar, 3: somewhat similar, 4: similar).

## B Details on the evaluated interpretability methods

**GradCAM [10].** For our ImageNet [9] studies, we generate GradCAM explanations for the ResNet50 [5] model in the `torchvision` library which achieves 76.1% classification accuracy. For our CUB studies, we generate GradCAM explanations for a ResNet50 [5] model we trained on the CUB [11] training set. This model achieves 81.0% accuracy on the CUB test set. We used the code by Gildenblat et al. [4] to generate GradCAM visualizations.<sup>1</sup> For the *agreement* task, we generate the GradCAM heatmap for the model prediction and normalize it into the  $[0, 1]$  range. For the *distinction* task, we generate four GradCAM heatmaps for each image: for correct predictions, we generate heatmaps for the top-4 predictions; for incorrect predictions, we generate heatmaps for the top-3 predictions and for the ground-truth class. We identify the local minimum and maximum of the four heatmaps, and then normalize the heatmaps into the  $[0, 1]$  range. This way, we preserve the intensity difference between heatmaps for different predictions. See Fig. A1 for an example set of GradCAM explanations.

**BagNet [1].** For our ImageNet studies, we use the BagNet33 model trained by the original authors which achieves 66.7% accuracy on ImageNet classification. For our CUB studies, we train a BagNet33 model on the CUB training set. This model achieves 74.2% accuracy on the CUB test set. For the *agreement* task, we use the authors’ code as is and normalize each heatmap individually by clipping the values above the 99th percentile.<sup>2</sup> On the other hand, for the *distinction* task, we normalize the four heatmaps together so that we preserve the intensity difference. See Fig. A2 for an example set of BagNet explanations.

<sup>1</sup> <https://github.com/jacobgil/pytorch-grad-cam>

<sup>2</sup> <https://github.com/wielandbrendel/bag-of-local-features-models>

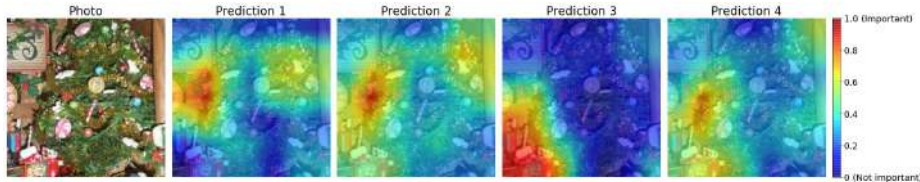


Fig. A1. GradCAM explanations shown in the *distinction* task.

**ProtoPNet [2].** For ProtoPNet, we used the ResNet34-based model trained by Hoffmann et al. [6]. We pruned 331 prototypes from this model to improve interpretability. The resulting model has 1669 prototypes and achieves 79.9% accuracy on the CUB [11] test set. For generating explanations, we used the code by the original authors with some modifications which we describe below.<sup>3</sup> In our studies, given an explanation, participants are asked to rate the similarity of each prototype-region pair, then either rate the level of confidence in the prediction’s correctness (*agreement*) or select the correct class (*distinction*). To make ProtoPNet’s explanations more suitable for these tasks, we made the following modifications to the original explanation form.

- The ProtoPNet model calculates evidence for all classes using the learned prototypes, then predicts the class with the highest evidence. However, we deemed it is unrealistic to ask users to review explanations for all 200 bird classes in CUB. Hence, we only present explanations for one (*agreement*) or four (*distinction*) classes and ask users to examine them.
- The original explanation (Fig. A3 left) shows activation maps, similarity scores, class connection weights, and the total class evidence. In our version (Fig. A3 right) we remove them as we seek to investigate what participants rate as similar and not.
- In the original explanation, prototypes are presented in the order of highest to lowest similarity. In ours, we randomly shuffle the order of prototypes because we don’t want to skew the participants’ region-prototype similarity ratings.

**ProtoTree [7].** For ProtoTree, we used the model trained by the original authors which achieves 81.7% accuracy on the CUB [11] test set. This model is a pruned tree of depth 10 and 511 nodes. We used the authors’ code to generate explanations with some modifications we describe below.<sup>4</sup>

- Same as what we did for ProtoPNet explanations, we removed the similarity scores as we seek to investigate what participants rate as similar and not.
- For the local explanation, we converted the horizontal explanation (Fig. A4) into a vertical one (Fig. A5). A vertical explanation is a better representation of how the model reasons, as the model starts from the root node and proceeds down the tree until it reaches one of the bottom leaves. Further, it is easier for the participants to examine the explanation by scrolling up and down.

---

<sup>3</sup> <https://github.com/cfchen-duke/ProtoPNet>

<sup>4</sup> <https://github.com/M-Nauta/ProtoTree>

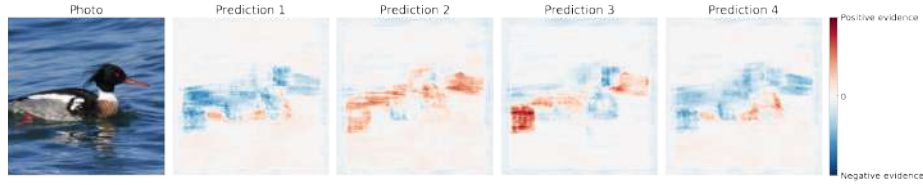


Fig. A2. BagNet explanations shown in the *distinction* task.

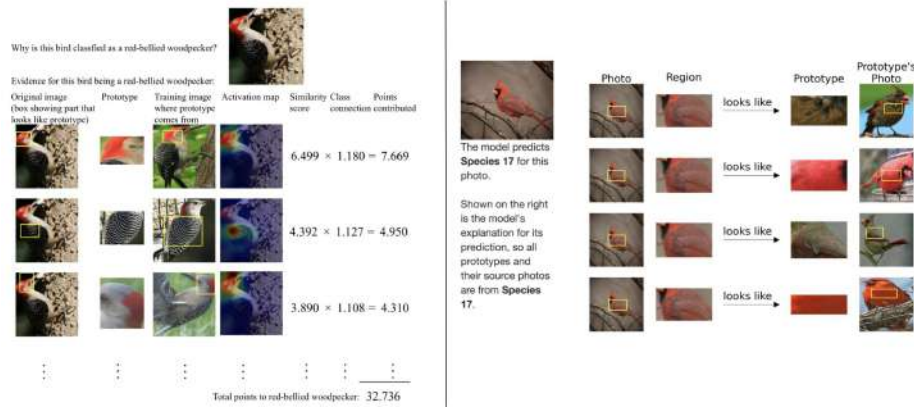
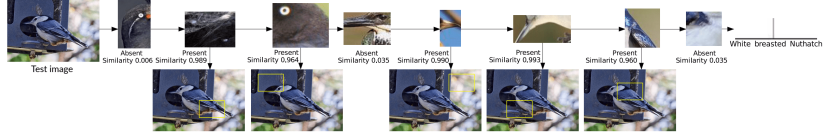
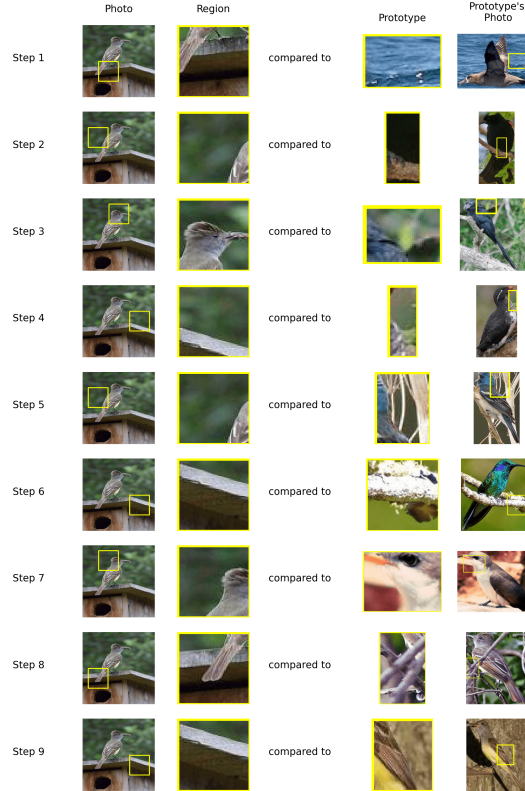


Fig. A3. ProtoPNet original and modified explanations. The original explanation (left) taken from the original paper [2] contains details such as activation maps, similarity scores, and class connection weights. In our version (right), we remove these to abstract away the complexities and have the participants focus on examining the similarity between prototypes and their matched image regions.



**Fig. A4. ProtoTree original explanation.** We show the original explanation displayed in Fig. 9 of the original paper [7]. See Fig. A5 for our modified explanation.



**Fig. A5. ProtoTree modified explanation.** See Fig. A4 for the original explanation.

## C Details on the human studies

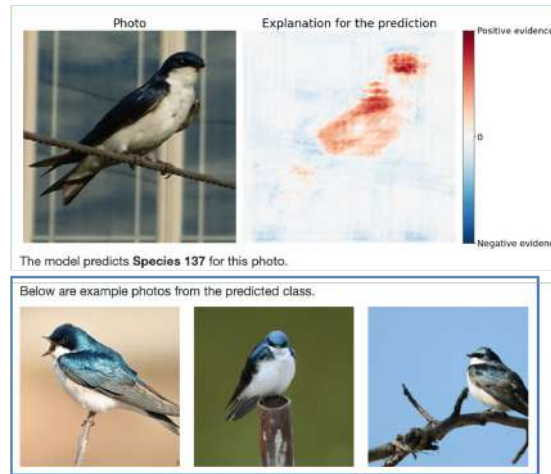
We ran our study through Human Intelligence Tasks (HITs) deployed on Amazon Mechanical Turk (AMT). We recruited participants who are US-based, have done over 1000 HITs, and have prior approval rate of at least 98%. For each study, we deployed 10 HITs, each with a different set of input images and explanations. To reduce the variance with respect to the input, we had 5 participants complete each HIT, so each study had 50 participants. Participants were compensated based on the state-level minimum wage of \$12/hr.

The demographic distribution was: man 60%, woman 38%, non-binary 1%, no gender reported 1%; White 74%, Black/African American 9%, Asian 7%, no race/ethnicity reported 7%, Hispanic/Latino/Spanish Origin of any race 2%, American Indian/Alaska Native 1%, Native Hawaiian/Other Pacific Islander 0%. The self-reported machine learning experience was  $2.5 \pm 1.0$ , between “2: have heard about...” and “3: know the basics...” The average study duration was  $6.9 \pm 3.5$  minutes for GradCAM,  $6.6 \pm 3.5$  for BagNet,  $13.6 \pm 6.2$  for ProtoPNet, and  $10.4 \pm 3.1$  for ProtoTree.

## D Additional results and analyses

### D.1 Agreement study results with vs. without example images

In Section 5.2 of the main paper, we described the results of our *agreement* study. Here we provide additional results.



**Fig. A6. BagNet *agreement* study input with example images.** For the study version with example images, we additionally show three example images from the predicted class (highlighted in the blue box).

For GradCAM and BagNet, we run another version of the *agreement* study where we show three example images from the predicted class, in addition to the test image, prediction, and heatmap (see Fig. A6). Since ProtoPNet and ProtoTree explanations consist of source images of the learned prototypes, we take this measure to provide similar supplementary information for GradCAM and BagNet. As expected, participants improve on the task when they see example images from the predicted class (5.3% overall improvement for GradCAM, 7.1% for BagNet). However, even with the help of example images, participants tend to believe in incorrect predictions, which suggests that incorrect top-1 predictions from high-performance models such as ResNet50 and BagNet are oftentimes convincing. Between CUB and ImageNet, task accuracy is overall higher on CUB, but both yield similar insights. See Tab. A1 for full results.

**Table A1. Agreement study results with vs. without examples.** For each study, we show mean accuracy, standard deviation of the participants’ performance, and mean confidence rating in parentheses. *Italics* denotes methods with accuracy not statistically significantly different from 50% random chance ( $p > 0.05$ ); **bold** denotes the highest performing method in each group. **In all studies, participants leaned towards believing that model predictions are correct when provided explanations, regardless of if they are actually correct.** For example, for GradCAM on CUB, participants thought 72.4% of correct predictions were correct and  $100 - 32.8 = 67.2\%$  of incorrect predictions were correct. These results reveal an issue of *confirmation bias*. **Comparing results with vs. without example images from the predicted class, participants improve on the task when they see examples, but still tend to believe in incorrect predictions.** See Appendix D.1 for a discussion.

CUB	GradCAM [10]	w/ examples	BagNet [1]	w/ examples
Correct	72.4% $\pm$ 21.5 (2.9)	83.2% $\pm$ 15.7 (3.3)	75.6% $\pm$ 23.4 (3.0)	<b>83.6% <math>\pm</math> 17.3 (3.3)</b>
Incorrect	32.8% $\pm$ 24.3 (2.8)	36.8% $\pm$ 22.8 (2.8)	<i>42.4% <math>\pm</math> 28.7 (2.7)</i>	<b>44.4% <math>\pm</math> 30.5 (2.6)</b>
ImageNet	GradCAM [10]	w/ examples	BagNet [1]	w/ examples
Correct	70.8% $\pm$ 26.6 (2.9)	<b>78.4% <math>\pm</math> 25.6 (3.2)</b>	66.0% $\pm$ 27.2 (2.8)	77.2% $\pm$ 23.3 (3.2)
Incorrect	<b>44.8% <math>\pm</math> 31.6 (2.7)</b>	<i>43.6% <math>\pm</math> 32.4 (2.6)</i>	35.6% $\pm$ 26.9 (2.7)	<i>42.8% <math>\pm</math> 32.7 (2.6)</i>

## D.2 Analysis with automatic evaluation metrics

In Section 5.3 of the main paper, we briefly summarized our analysis with automatic evaluation metrics. Here we discuss the results in more detail.

We further analyzed GradCAM heatmaps set using three automatic evaluation metrics: pointing game [13], energy-based pointing game (energy game) [12], and intersection-over-union (IoU) [14]. Pointing game considers a heatmap correct when its highest-intensity point lies inside the segmentation/bounding-box annotation. Energy game calculates how much energy in a heatmap falls inside the segmentation/bounding-box annotation. IoU captures the amount of overlap between a binarized heatmap (according to some threshold) and the segmentation/bounding-box annotation. For all three metrics, higher values indicate better localization quality.

We evaluate up to three GradCAM explanations per image, all using the same segmentation/bounding-box annotation for the ground-truth class: heatmaps for the ground-truth class, predicted class, and class with the second-highest score. Results are summarized in Tab. A2. For CUB heatmaps, we calculate the three metrics on the entire test set (top table). For ImageNet heatmaps, we calculate the metrics on 5,000 randomly sampled validation images. Since ImageNet images sometimes have multiple bounding box annotations, we report results evaluated with one bounding box that yields the best result (middle table) and results evaluated with the union of bounding boxes (bottom table). We find that all three metrics are highest on the ground-truth/predicted class heatmaps for correctly predicted samples. However, we find that these metrics are also high for other heatmaps, even when they are for wrong classes.

Next, we calculate these metrics on images/heatmaps we showed the participants and analyze our human study results. In the *agreement* study, we find near-zero correlation between participants’ confidence in the model prediction and localization quality of heatmaps. In the *distinction* study, we also do not see meaningful relationships between participants’ choices and these automatic metrics, possibly because all four heatmaps have similar localization quality. These observations are consistent with the findings of [8,3], i.e., automatic metrics poorly correlate with human performance in post-hoc attribution heatmap evaluation. Overall, our analysis reveals a limitation of automatic metrics.

### D.3 Similarity judgment of humans vs. prototype-based models

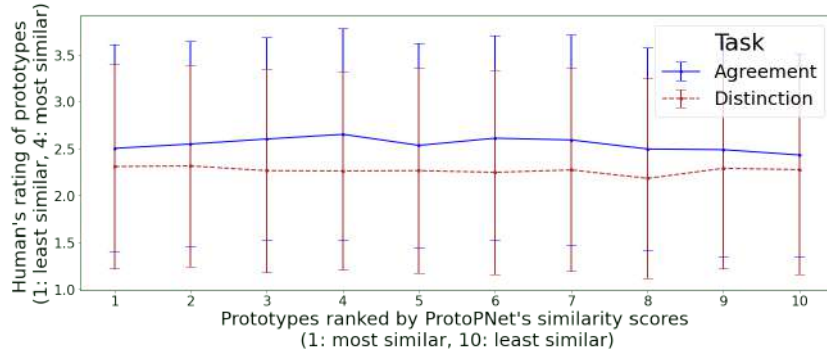
In Section 5.4 of the main paper, we quantified the gap between prototype-based models and human users’ notion of similarity. Here we show a plot of participant vs. ProtoPNet prototype similarity rating (Fig. A7). There is no significant negative correlation between the two. This result suggests a gap between ProtoPNet and human judgments of similarity.

Nonetheless, we find that participants are consistent in their similarity ratings and decisions. When examining ProtoPNet and ProtoTree explanations, on average participants assign higher similarity ratings to prototypes of the class they select to be correct (2.9 out of 4 for both ProtoPNet *agreement* and *distinction* tasks, 2.4 for ProtoTree *agreement*) and lower similarity ratings to prototypes of the class they select to be incorrect (2.0 and 2.1 for ProtoPNet *agreement* and *distinction*, 2.0 for ProtoTree *agreement*). The similarity ratings between the two groups are statistically significantly different in all studies. This suggests that participants understand how the model reasons (i.e., they predict the bird class whose prototypes appear most similar to the given photo).



**Table A2. Evaluation of GradCAM heatmaps using automatic metrics.** We report the mean and standard deviation of three automatic evaluation metrics calculated on heatmaps for the ground-truth class, predicted class, and class with the second-highest score. **All three metrics are highest on the ground-truth/predicted class heatmaps for correctly predicted samples. However, these metrics are also high for other heatmaps, even when they are for wrong classes.**

CUB [11] heatmaps evaluated with the segmentation mask				
Prediction	Class	Pointing game [13]	Energy game [12]	IoU [14]
Correct	GT/Predicted	$0.92 \pm 0.27$	$0.12 \pm 0.07$	$0.38 \pm 0.15$
	Second-highest	$0.74 \pm 0.44$	$0.09 \pm 0.06$	$0.24 \pm 0.15$
Incorrect	GT	$0.73 \pm 0.45$	$0.08 \pm 0.06$	$0.23 \pm 0.16$
	Predicted	$0.83 \pm 0.37$	$0.09 \pm 0.06$	$0.29 \pm 0.15$
	Second-highest	$0.80 \pm 0.40$	$0.09 \pm 0.06$	$0.26 \pm 0.15$
ImageNet [9] heatmaps evaluated with the bounding box that yields the best result				
Prediction	Class	Pointing game [13]	Energy game [12]	IoU [14]
Correct	GT/Predicted	$0.95 \pm 0.22$	$0.27 \pm 0.13$	$0.60 \pm 0.28$
	Second-highest	$0.93 \pm 0.26$	$0.26 \pm 0.13$	$0.60 \pm 0.27$
Incorrect	GT	$0.91 \pm 0.29$	$0.23 \pm 0.14$	$0.52 \pm 0.31$
	Predicted	$0.82 \pm 0.38$	$0.22 \pm 0.15$	$0.52 \pm 0.33$
	Second-highest	$0.84 \pm 0.37$	$0.23 \pm 0.15$	$0.52 \pm 0.33$
ImageNet [9] heatmaps evaluated with the union of the bounding boxes				
Prediction	Class	Pointing game [13]	Energy game [12]	IoU [14]
Correct	GT/Predicted	$0.95 \pm 0.22$	$0.29 \pm 0.13$	$0.65 \pm 0.26$
	Second-highest	$0.93 \pm 0.26$	$0.28 \pm 0.13$	$0.64 \pm 0.26$
Incorrect	GT	$0.91 \pm 0.29$	$0.24 \pm 0.14$	$0.56 \pm 0.30$
	Predicted	$0.82 \pm 0.38$	$0.24 \pm 0.15$	$0.56 \pm 0.32$
	Second-highest	$0.84 \pm 0.37$	$0.24 \pm 0.15$	$0.56 \pm 0.32$



**Fig. A7. Participant vs. ProtoPNet prototype similarity rating.** There exists a gap between ProtoPNet’s similarity scores and human judgments of similarity (Spearman’s  $\rho = -0.25$ ,  $p = 0.49$  for *distinction*;  $\rho = -0.52$ ,  $p = 0.12$  for *agreement*).

#### D.4 Subjective evaluation results

In Section 5.5 of the main paper, we summarized our subjective evaluation results. Here we provide the full results.

In Tab. A3, we report the participants’ self-rated level of understanding of the given model’s reasoning process. Overall, the participants rated their level of understanding between 3 (fair) and 4 (good). Interestingly, we find that the rating tends to decrease after the participants see their task performance. Several participants indicated that their performance was lower than what they expected: “I thought I would do a bit better!”, “my score wasn’t as high as I would have liked”, “I was surprised that my score was not very much higher than random guessing. I thought I had a good idea of the model, especially making judgements about the amount of positive and negative evidence, but it seems I did not.” No one suggested the opposite. This trend suggests that participants might have been disappointment in their task performance, which in turn led them to lower their self-rated level of method understanding.

**Table A3. Participants’ self-rated level of method understanding.** We report the mean and standard deviation of the participants’ self-rating of their method understanding. Participants provide ratings three times: after reading about the method (post-intro), after completing the task (post-task), and after learning about their task performance (post-results). The rating tends to *decrease* after the participants see their task performance ( $p < 0.05$ ).

Dataset	Method	Study	Post-intro	Post-task	Post-results
CUB [11]	GradCAM [10]	Agreement	$3.7 \pm 0.9$	$3.8 \pm 0.9$	$3.3 \pm 1.1$
		Agreement w/ examples	$3.7 \pm 1.0$	$3.9 \pm 0.7$	$3.4 \pm 1.0$
		Distinction	$3.4 \pm 1.0$	$3.5 \pm 1.2$	$3.6 \pm 0.8$
	BagNet [1]	Agreement	$3.5 \pm 1.0$	$3.7 \pm 0.8$	$3.3 \pm 1.1$
		Agreement w/ examples	$3.7 \pm 0.8$	$3.9 \pm 0.8$	$3.6 \pm 1.0$
		Distinction	$3.8 \pm 0.7$	$4.0 \pm 0.8$	$3.9 \pm 0.8$
	ProtoPNet [2]	Agreement	$3.9 \pm 0.8$	$4.0 \pm 0.8$	$3.7 \pm 0.8$
		Distinction	$4.1 \pm 0.8$	$3.9 \pm 0.8$	$3.7 \pm 1.1$
	ProtoTree [7]	Agreement	$3.7 \pm 0.8$	$3.7 \pm 1.0$	$3.4 \pm 0.8$
		Agreement (tree)	$3.7 \pm 0.7$	$3.5 \pm 0.9$	$3.2 \pm 1.1$
		Distinction	$3.4 \pm 1.0$	$3.6 \pm 1.1$	$3.3 \pm 1.2$
ImageNet [9]	GradCAM [10]	Agreement	$3.7 \pm 0.9$	$3.9 \pm 0.9$	$3.0 \pm 1.0$
		Agreement w/ examples	$3.4 \pm 0.8$	$3.7 \pm 0.8$	$3.5 \pm 0.9$
		Distinction	$3.9 \pm 0.9$	$3.7 \pm 1.0$	$3.7 \pm 1.0$
		Distinction w/ labels	$3.9 \pm 0.9$	$3.8 \pm 1.0$	$3.8 \pm 0.9$
	BagNet [1]	Agreement	$3.7 \pm 0.8$	$3.9 \pm 0.7$	$3.4 \pm 1.0$
		Agreement w/ examples	$3.8 \pm 0.9$	$3.9 \pm 0.9$	$3.3 \pm 1.0$
		Distinction	$3.9 \pm 0.8$	$3.9 \pm 0.8$	$3.8 \pm 1.0$
		Distinction w/ labels	$3.8 \pm 0.9$	$4.0 \pm 0.8$	$3.8 \pm 0.8$
Mean across all studies			$3.7 \pm 0.9$	$3.8 \pm 0.9$	$3.5 \pm 1.0$

## D.5 Interpretability-accuracy tradeoff results

In Section 5.6 of the main paper, we summarized our interpretability-accuracy tradeoff study results. Here we provide more details.

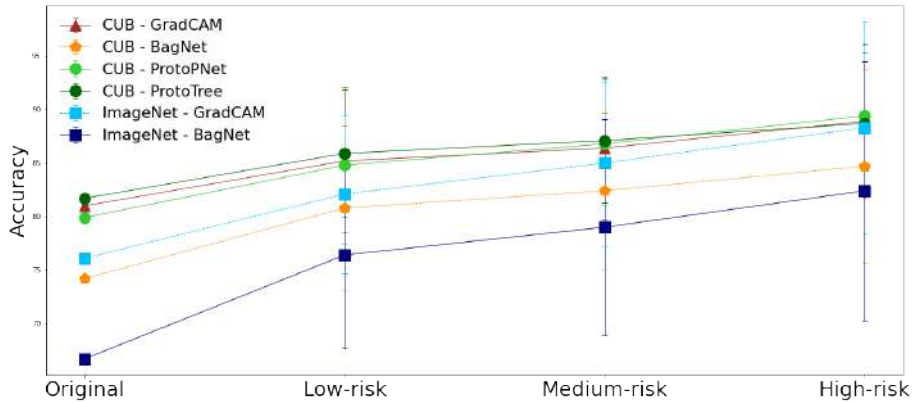
In Tab. A4 and Fig. A8, we show the full results of our interpretability-accuracy tradeoff study. We report the accuracy of the evaluated interpretable model and the minimum accuracy of a baseline model that participants require in order to use it over the model with explanations under different risk settings. Across all studies, we find that participants require the baseline model to have higher accuracy than the evaluated interpretable model, and input a higher accuracy requirement for higher-risk settings. On average, participants require the baseline model to have +6.2% higher accuracy for low-risk (e.g., bird species recognition for scientific or educational purposes), +8.2% for medium-risk (e.g., object recognition for automatic grocery checkout), and +10.9% for high-risk (e.g., scene understanding for autonomous driving) settings.

We observe this trend in the participants’ written responses as well. Most participants write that they would use the baseline model only when it has higher accuracy than the evaluated interpretable model: “I would need the black box model to give me a nice boost in accuracy, or I would just stick to the bagnet model, since it is pretty accurate.” On the contrary, participants exhibit different levels of desire for interpretability. Some deem interpretability as important: “Understanding how a prediction works is important. For me to accept a model with no explanations, the level of accuracy needs to be higher”, “I prefer to understand how models work, so the black box model has to be significantly better than the other model for me to use it. As the stakes become higher, I want its accuracy to be higher because there’s no way for me to question or check its progress if it’s wrong.” Other participants willingly tradeoff interpretability for accuracy: “I don’t need to know how it works. So, as long as it’s marginally better, it should be used”, “I don’t care about not having an explanation, so if the accuracy of a different model has just a 1% improvement in performance then I would choose the better performing model.”

Nonetheless most participants express a need for higher-accuracy models in higher-risk settings: “The higher the risk, the more accurate I need it to be in order to feel confident using it”, “If I were to choose to use a model that did not provide reasoning for me to utilize in evaluating how the decision was made I would need to know that the model would give me significantly better results, especially in a high-risk scenario as described above, but even in the medium risk setting, being able to assess the reasoning of the model is an invaluable tool and I would only be willing to give it up for significant increases in accuracy.”

**Table A4. Interpretability-accuracy tradeoff results.** We report the mean and standard deviation of the additional accuracy participants require for the baseline model, to use it over the model with explanations. For example in the GradCAM *agreement* study with CUB, participants require the baseline model to have +5.6% accuracy beyond the model that comes with GradCAM explanations and achieves 81.0% accuracy, in the low-risk setting. See Fig. A8 for a visualization of the results.

Dataset	Method	Study	Orig	Low-risk	Med-risk	High-risk
CUB [11]	GradCAM [10]	Agreement	81.0	+5.6 (±6.9)	+6.2 (±5.7)	+7.7 (±7.0)
		Agreement w/ examples		+4.2 (±6.1)	+5.7 (±5.7)	+7.7 (±7.5)
		Distinction		+2.9 (±6.9)	+4.5 (±5.2)	+8.1 (±6.9)
	BagNet [1]	Agreement	74.2	+6.8 (±7.9)	+7.8 (±8.1)	+12.3 (±9.2)
		Agreement w/ examples		+6.1 (±7.1)	+8.1 (±6.3)	+10.7 (±9.2)
		Distinction		+7.0 (±8.1)	+8.8 (±7.4)	+8.4 (±8.4)
	ProtoPNet [2]	Agreement	79.9	+5.8 (±6.6)	+7.8 (±4.9)	+9.4 (±6.6)
		Distinction		+4.1 (±7.9)	+6.1 (±6.4)	+9.7 (±7.1)
	ProtoTree [7]	Agreement	81.7	+3.8 (±6.5)	+4.2 (±6.3)	+5.1 (±6.5)
Agreement (tree)		+3.7 (±5.5)		+5.8 (±5.1)	+6.7 (±6.6)	
Distinction		+5.1 (±5.7)		+6.4 (±5.8)	+9.2 (±6.2)	
ImageNet [9]	GradCAM [10]	Agreement	76.1	+6.3 (±7.6)	+8.1 (±8.6)	+11.8 (±10.7)
		Agreement w/ examples		+4.8 (±6.8)	+8.6 (±7.6)	+11.4 (±10.8)
		Distinction		+5.3 (±7.3)	+9.8 (±6.7)	+12.4 (±8.6)
		Distinction w/ labels		+7.6 (±7.7)	+9.3 (±7.9)	+13.2 (±9.0)
	BagNet [1]	Agreement	66.7	+9.9 (±7.5)	+14.1 (±9.5)	+17.5 (±11.1)
		Agreement w/ examples		+9.7 (±8.5)	+13.2 (±10.4)	+17.6 (±13.0)
		Distinction		+7.9 (±9.3)	+9.6 (±9.2)	+11.2 (±11.2)
		Distinction w/ labels		+11.4 (±9.2)	+12.4 (±10.4)	+16.6 (±11.6)
Mean across all studies				+6.2 (±7.7)	+8.2 (±7.9)	+10.9 (±9.7)



**Fig. A8. Visualization of the interpretability-accuracy tradeoff results.** This plot shows that participants desire higher accuracies for the baseline model, especially in higher-risk settings. See Tab. A4 for the full results.

## E Simple decision tree used for explaining ProtoTree

One additional challenge of evaluating the ProtoTree model is that participants may not be familiar with decision trees. To mitigate this challenge, we introduce a simple decision tree model for fruit classification before introducing ProtoTree. This simple decision tree model takes in an input image and makes an output classification (Class A, B, C, D, E) based on three decision nodes. We first walk through the participants through an example. We then present two warm-up exercises so that the participants can become more familiar with decision trees. When the participants submit their answers, we also provide the correct answer and the reason for it. Participants achieved 86.5% performance on this task, implying that the low task accuracy for ProtoTree is not due to a lack of comprehension of decision trees. See Fig. A9 for the UI.


**Warming up**  
This page is meant to give you a sense of the model and the task we will introduce in this study.

Here we have an example model that classifies fruit photos into Fruit A, B, C, D, E based on a series of decisions regarding Color, Count, and Shape. Specifically, this model makes decisions in a tree structure as you can see below.

---

**Example**  
For this example photo, the model reasons in the following way: The model first judges that the photo's fruit has Color "Yellow." Based on this decision, it moves onto the next step and judges that the fruit has Shape "Skinny." After these two decisions, the model arrives at its prediction: **Fruit E**.

Example Photo



Model predicts **Fruit E**

Model's decision process

```

graph TD
    Color[Color] -->|Green| Count[Count]
    Color -->|Red| C[C]
    Color -->|Yellow| Shape[Shape]
    Count -->|Less than 3| A((A))
    Count -->|Greater than 3| B((B))
    C --> Round((Round))
    C --> Skinny((Skinny))
    Shape --> Round
    Shape --> Skinny
    Round --> D((D))
    Skinny --> E((E))
          
```

**Q. Is the model's prediction correct or incorrect?**  
☒ Correct ☐ Incorrect


**Q. What do you think about the model's prediction?**  
☒ Fairly confident that prediction is correct  
☐ Somewhat confident that prediction is correct  
☐ Somewhat confident that prediction is incorrect  
☐ Fairly confident that prediction is incorrect

**Q. Choose the first step where you think the model makes a mistake. If you think the model makes no mistakes, select "None."**  
☒ None ☐ Color ☐ Count ☐ Shape

---

**Your Turn!**  
After examining the photo and the model's decision process, please answer the three questions as above, then click "Submit."

Photo 1



Model predicts **Fruit B**

Model's decision process

```

graph TD
    Color[Color] -->|Green| Count[Count]
    Color -->|Red| C[C]
    Color -->|Yellow| Shape[Shape]
    Count -->|Less than 3| A((A))
    Count -->|Greater than 3| B((B))
    C --> Round((Round))
    C --> Skinny((Skinny))
    Shape --> Round
    Shape --> Skinny
    Round --> D((D))
    Skinny --> E((E))
          
```

**Q. Is the model's prediction correct or incorrect?**  
☐ Correct ☒ Incorrect

**Q. What do you think about the model's prediction?**  
☐ Fairly confident that prediction is correct  
☐ Somewhat confident that prediction is correct  
☐ Somewhat confident that prediction is incorrect  
☒ Fairly confident that prediction is incorrect


**Q. Choose the first step where you think the model makes a mistake. If you think the model makes no mistakes, select "None."**  
☐ None ☒ Color ☐ Count ☐ Shape

**Submit**

**Correct!** You have successfully identified that the model's prediction is incorrect and that the model made a mistake on the Color decision.

---

Photo 2



Model predicts **Fruit A**

Model's decision process

```

graph TD
    Color[Color] -->|Green| Count[Count]
    Color -->|Red| C[C]
    Color -->|Yellow| Shape[Shape]
    Count -->|Less than 3| A((A))
    Count -->|Greater than 3| B((B))
    C --> Round((Round))
    C --> Skinny((Skinny))
    Shape --> Round
    Shape --> Skinny
    Round --> D((D))
    Skinny --> E((E))
          
```

**Q. Is the model's prediction correct or incorrect?**  
☒ Correct ☐ Incorrect

**Q. How confident are you in the model's decision?**  
☒ Fairly confident that prediction is correct  
☐ Somewhat confident that prediction is correct  
☐ Somewhat confident that prediction is incorrect  
☐ Fairly confident that prediction is incorrect

**Q. Choose the first step where you think the model makes a mistake. If you think the model makes no mistakes, select "None."**  
☒ None ☐ Color ☐ Count ☐ Shape

**Submit**

**Fig. A9. A simple decision example.** We use this model to introduce participants to decision trees before explaining the more complex ProtoTree. See Appendix E for details.

## F UI snapshots

In Section 4 of the main paper, we outlined our study design. Here we provide snapshots of our study UIs in the following order.

**1. Study introduction.** For each participant, we first briefly introduce the study and receive their informed consent. The consent form was approved by the IRB and acknowledges that participation is voluntary, refusal to participate will involve no penalty or loss of benefits, etc. See Fig. A10.

**2. Demographics and background.** To help future researchers calibrate our results and do proper comparison, we request optional demographic data regarding gender identity, race and ethnicity. We also ask the participant’s experience with machine learning. See Fig. A11.

**3. Method introduction.** We introduce each interpretability method/model in simple terms. See Fig. A12.

**4. Task preview and first subjective evaluation.** To encourage participants to carefully read the method explanation, we show a preview of the task they will complete along with a correct and incorrect prediction. Participants then answer their first subjective evaluation question. In Fig. A13 we shown an example from the ProtoPNet *agreement* study.

**5. Task.** Participants then proceed onto the main task. We show the UI for the following 8 studies:

- GradCAM *distinction* (Fig. A14)
- GradCAM *agreement* (Fig. A15)
- BagNet *distinction* (Fig. A16)
- BagNet *agreement* (Fig. A17)
- ProtoPNet *distinction* (Fig. A18)
- ProtoPNet *agreement* (Fig. A19)
- ProtoTree *distinction* (Fig. A20)
- ProtoTree *agreement* (Fig. A21)

**6. Second and third subjective evaluation.** After the task, participants complete their second subjective evaluation question. We then disclose their task performance and ask the third subjective evaluation question. These questions allow us to investigate if the participants’ self-rated level of method understanding undergoes any changes throughout the study. See Fig. A22.

**7. Interpretability-accuracy tradeoff.** Finally, we investigate the tradeoff participants are willing to make when comparing the evaluated interpretable model against a baseline model that doesn’t come with any explanation. We present three scenarios to the participants: low-risk (e.g., scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for self-driving cars). We then ask them to input the minimum accuracy of a baseline model that would convince them to use the baseline model over the model that comes with explanations and briefly describe their reasoning. See Fig. A23.

### Study introduction

In this study, we aim to evaluate the interpretability of computer vision models. We will provide explanations of how a model makes its prediction and ask you to evaluate how interpretable it is through several questions and tasks. The expected duration of the study is 5-15 minutes.

### Consent

Please read the consent form. If you understand and consent to these terms, click "I Accept" to continue.

☒ I Accept

[Next Page](#)

Fig. A10. 1. Study introduction.

### Demographics and background

#### Q. Demographics (Optional)

##### Gender identity

- ☐ Man  
☐ Non-binary  
☐ Woman  
☐ Prefer to self-describe below

##### Race and ethnicity (select one or more)

- ☐ American Indian or Alaska Native  
☐ Asian  
☐ Black or African American  
☐ Native Hawaiian or Other Pacific Islander  
☐ White  
☐ Hispanic or Latino or Spanish Origin of any race

#### Q. How much experience do you have with machine learning (ML)?

- ☐ I don't know anything about ML  
☐ I have heard about a few ML concepts or applications  
☐ I know the basics of ML and can hold a short conversation about it  
☐ I have taken a course on ML and/or have experience working with a ML system  
☐ I often use and study ML in my life

[Next Page](#)

Fig. A11. 2. Demographics and background.

## Task preview

Here's a preview of the task you will complete.


### Examine model predictions

Open a test photo. The model will make predictions for the species name on a preselected 5 box spanner that contains your photo. Specifically, for each prediction, the model identifies a region in the photo that it believes contains the bird. It also provides a confidence score for each prediction. Judge whether you agree with the model's identified region for testing the region and providing a label. Below is a photo of a red-tailed black eagle, with your confidence in the model's prediction.


Open the test photo again, compare your prediction to one of the model's five high confidence test sets. When making the prediction, the model provides more information on each test set region.

**Task:** Rate the similarity of each test set prediction region pair on a scale of 1-4.

1: Not Similar 2: Somewhat Not Similar 3: Somewhat Similar 4: Similar




**Region**




Region 1: Red-tailed black eagle

**Region**




Region 2: Red-tailed black eagle

**Region**




Region 3: Red-tailed black eagle

**Region**




Region 4: Red-tailed black eagle

**Region**




Region 5: Red-tailed black eagle

**Region**




Region 6: Red-tailed black eagle

**Region**




Region 7: Red-tailed black eagle

**Region**




Region 8: Red-tailed black eagle

**Region**




Region 9: Red-tailed black eagle

**Region**




Region 10: Red-tailed black eagle

**Region**




Region 11: Red-tailed black eagle

**Region**




Region 12: Red-tailed black eagle

**Region**




Region 13: Red-tailed black eagle

**Region**




Region 14: Red-tailed black eagle

**Region**




Region 15: Red-tailed black eagle

**Region**




Region 16: Red-tailed black eagle

**Region**




Region 17: Red-tailed black eagle

**Region**




Region 18: Red-tailed black eagle

**Region**



Region 19: Red-tailed black eagle

**Region**



Region 20: Red-tailed black eagle

**Region**



Region 21: Red-tailed black eagle

**Region**



Region 22: Red-tailed black eagle

**Region**



Region 23: Red-tailed black eagle

**Region**



Region 24: Red-tailed black eagle

**Region**



Region 25: Red-tailed black eagle

**Region**



Region 26: Red-tailed black eagle

**Region**



Region 27: Red-tailed black eagle

**Region**



Region 28: Red-tailed black eagle

**Region**



Region 29: Red-tailed black eagle

**Region**



Region 30: Red-tailed black eagle

**Region**



Region 31: Red-tailed black eagle

**Region**



Region 32: Red-tailed black eagle

**Region**



Region 33: Red-tailed black eagle

**Region**



Region 34: Red-tailed black eagle

**Region**



Region 35: Red-tailed black eagle

**Region**



Region 36: Red-tailed black eagle

**Region**



Region 37: Red-tailed black eagle

**Region**



Region 38: Red-tailed black eagle

**Region**



Region 39: Red-tailed black eagle

**Region**



Region 40: Red-tailed black eagle

**Region**



Region 41: Red-tailed black eagle

**Region**



Region 42: Red-tailed black eagle

**Region**



Region 43: Red-tailed black eagle

**Region**



Region 44: Red-tailed black eagle

**Region**



Region 45: Red-tailed black eagle

**Region**



Region 46: Red-tailed black eagle

**Region**



Region 47: Red-tailed black eagle

**Region**



Region 48: Red-tailed black eagle

**Region**



Region 49: Red-tailed black eagle

**Region**



Region 50: Red-tailed black eagle

**Region**



Region 51: Red-tailed black eagle

**Region**




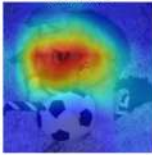
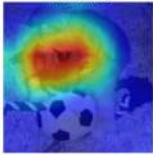
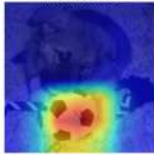
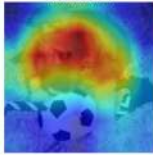
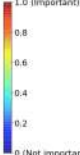
### Examine model predictions

For each photo, we show explanations for the model's 4 predictions.

First, select the class you think the model predicts (i.e. gives the highest score). Second, select the class you think is correct. The two classes can be different because the model makes incorrect predictions on some photos.

For either question, random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.

This is a photo of **Norwegian elkhound, elkhound**.

Photo	Prediction 1	Prediction 2	Prediction 3	Prediction 4	
					
<b>Q. Which class do you think the model predicts?</b> <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4		<b>Q. Which class do you think is correct?</b> <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4			
<b>Q. How confident are you in your answer?</b> <input type="radio"/> Not confident at all <input type="radio"/> Slightly confident <input type="radio"/> Somewhat confident <input type="radio"/> Fairly confident <input type="radio"/> Completely confident		<b>Q. How confident are you in your answer?</b> <input type="radio"/> Not confident at all <input type="radio"/> Slightly confident <input type="radio"/> Somewhat confident <input type="radio"/> Fairly confident <input type="radio"/> Completely confident			

Click "Next Photo" after answering all questions.

1 / 10

Next Photo

Click on "Next Page" after selecting answers for all 10 photos.

Next Page

Click "Method Description" to open or close method description.

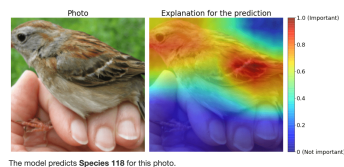
Method Description

Fig. A14. 5. Task: GradCAM *distinction*.

### Examine model predictions

For each photo, the model predicts which of the 200 species the bird in the photo belongs to. Next to the photo, we show an explanation of the model prediction.

After examining the explanation, rate your confidence in the model's prediction.



The model predicts Species 118 for this photo.

**Q. What do you think about the model's prediction?**

- ☐ Fairly confident that prediction is correct  
☐ Somewhat confident that prediction is correct  
☐ Somewhat confident that prediction is incorrect  
☐ Fairly confident that prediction is incorrect

Click "Next Photo" after answering all questions.

1 / 10

Next Photo

Click on "Next Page" after selecting answers for all 10 photos.

Next Page

Click "Method Description" to open or close method description.

Method Description

Fig. A15. 5. Task: GradCAM *agreement*.

### Examine model predictions

For each photo, we show explanations for the model's 4 predictions.

First, select the class you think the model predicts (i.e. gives the highest score). Second, select the class you think is correct. The two classes can be different because the model makes incorrect predictions on some photos.

For either question, random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.

The figure displays a bird perched on a branch, followed by four heatmaps representing model predictions. The heatmaps are labeled 'Prediction 1' through 'Prediction 4'. A vertical color bar on the right side of the heatmaps indicates the level of evidence, with red representing 'Positive evidence' and blue representing 'Negative evidence'. The heatmaps show varying degrees of red and blue areas, indicating the model's confidence in its predictions. Prediction 3 shows the most intense red evidence, particularly on the bird's body.

Fig. A16. 5. Task: BagNet *distinction*.

### Examine model predictions

For each photo, the model predicts which of the 1000 classes the photo belong to (e.g., hornbill, panther, television, strawberry). Next to the photo, we show an explanation of the model prediction.

After examining the explanation, rate your confidence in the model's prediction.

Photo

Explanation for the prediction

Positive evidence

0

Negative evidence

The model predicts **Class 739** for this photo.

Q. What do you think about the model's prediction?

- ☐ Fairly confident that prediction is correct
- ☐ Somewhat confident that prediction is correct
- ☐ Somewhat confident that prediction is incorrect
- ☐ Fairly confident that prediction is incorrect

Click "Next Photo" after answering all questions.

4 / 10

Next Photo

Click on "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Method Description" to open or close method description.

Method Description

Fig. A17. 5. Task: BagNet *agreement*.


## Simulate the model

Given a bird photo, the ProtoPNet model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity.

For a given photo, we show explanations of how the model reasons for 4 bird species. For each bird species, rate how similar each prototype is to the photo region. Note that the (region, prototype) pairs are presented in random order. At the end, choose the bird species you think is correct.

Random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.

Photo



Prototypes and their source photos are from the specified species.





































That is, Species 1 explanation only contains the prototypes and the prototype's photos from Species 1.

**Task: Rate the similarity of each prototype-region pair on a scale of 1-4.**

1: Not Similar  
2: Somewhat Not Similar  
3: Somewhat Similar  
4: Similar

Click on "Species 1", "Species 2", "Species 3" and "Species 4" to move between species.  
For your HIT to be approved, you have to rate all prototypes in all 4 species.

Species 1
Species 2
Species 3
Species 4

Photo	Region	looks like	Prototype	Prototype's Photo	
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		→			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4

**Q. Choose the bird species you think is correct, then click "Next Photo."**

☐ Species 1 ☐ Species 2 ☐ Species 3 ☐ Species 4

**Q. How confident are you in your answer?**

☐ Not confident at all  
☐ Slightly confident  
☐ Somewhat confident  
☐ Fairly confident  
☐ Completely confident

**1 / 4**

Next Photo

If you can't click "Next Photo" after rating all prototypes and answering both questions, try clicking on a different answer and then click on your desired answer.

---

Click "Next Page" after selecting answers for all 4 photos.

Next Page

Click "Model Description" to open or close model description.

Model Description

Fig. A18. 5. Task: ProtoPNet *distinction*.


### Examine model predictions

Given a bird photo, the ProtoPNet model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity.

Judge whether you agree with the model's identified region by rating the region and prototype similar from on a scale of 1-4. At the end, rate your confidence in the model's prediction.

Note that the (photo region, prototype) pairs are presented in order of similarity, from high similarity to low. When making its prediction, the model places more importance on pairs with higher similarity.

**Task: Rate the similarity of each row's prototype-region pair on a scale of 1-4.**  
(1: Not Similar, 2: Somewhat Not Similar, 3: Somewhat Similar, 4: Similar)



The model predicts **Species 90** for this photo.

Shown on the right is the model's explanation for its prediction, so all prototypes and their source photos are from **Species 90**.

Photo	Region	looks like	Prototype	Prototype's Photo	Rating
		looks like			<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4
		looks like			<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4
		looks like			<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		looks like			<input checked="" type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		looks like			<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		looks like			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4
		looks like			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4
		looks like			<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4
		looks like			<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4
		looks like			<input checked="" type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4

**Q. What do you think about the model's prediction?**

☐ Fairly confident that prediction is correct  
☐ Somewhat confident that prediction is correct  
☐ Somewhat confident that prediction is incorrect  
☒ Fairly confident that prediction is incorrect

Click "Next Photo" after selecting the rows and answering the question.

**1 / 10**

[Next Photo](#)

Click "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

Click "Model Description" to open or close model description.

[Model Description](#)

Fig. A19. 5. Task: ProtoPNet *agreement*.

**Predict the bird species**


Given a bird photo, the ProtoTree model predicts the species based on prototypes it has learned from previously seen photos. Specifically at each step, the model identifies a region in the photo that looks the most similar to the step's prototype and judges whether the prototype is absent or present in the photo.

For each photo, we show the model's decisions for the first several steps. For the remaining final two steps, you will decide whether the prototypes are absent or present in the photo, which will lead to a bird species prediction.

Random guessing will get you 25% accuracy. You will receive a reward based on your performance beyond this 25% random chance.

---

**Photo**



**Model's decisions for the first several steps**

	Photo	Region	Prototype	Prototype's Photo	Decision
Step 1					Absent
Step 2					Absent
Step 3					Present
Step 4					Present
Step 5					Present
Step 6					Absent
Step 7					Absent

---

**Possible decisions for the final two steps**

**Step 8**

Photo: Region: compared to Prototype: Prototype's Photo:

Q1. Do you think this prototype absent or present in the photo?  
☐ Absent ☐ Present

---

**If you selected "Absent" in Q1:**

**Step 9**

Photo: Region: compared to Prototype: Prototype's Photo:

Q2-1. Do you think this prototype is absent or present in the photo?  
☐ Absent ☐ Present

---

**If you selected "Present" in Q1:**

**Step 9**

Photo: Region: compared to Prototype: Prototype's Photo:

Q2-2. Do you think this prototype is absent or present in the photo?  
☐ Absent ☐ Present

---

**Predicted bird species**

If you choose Absent in Q1 and Absent in Q2, you will arrive at the prediction Species 5.  
 If you choose Absent in Q1 and Present in Q2, you will arrive at the prediction Species 2.  
 If you choose Present in Q1 and Absent in Q2, you will arrive at the prediction Species 3.  
 If you choose Present in Q1 and Present in Q2, you will arrive at the prediction Species 4.

Q3. How confident are you in your answer?  
☐ Not confident at all  
☐ Slightly confident  
☐ Somewhat confident  
☐ Fairly confident  
☐ Completely confident

---

Click "Next Photo" after answering both questions.

1 / 10

[Next Photo](#)

---

Click "Next Page" after selecting answers for all 10 photos.

[Next Page](#)

---

Click "Model Description" to open or close model description.




















[Model Description](#)

Fig. A20. 5. Task: ProtoTree *distinction*.

### Examine model predictions

For each photo, examine the model's decision for each prototype and select the first step you disagree with the model's decision. Then rate your confidence in the model's prediction.

We ask you to select the first step you disagree with because the steps below your selected step is considered to be part of a wrong path. Since the ProtoTree model has a tree structure, once it makes an incorrect decision it goes on a wrong path and cannot reach the correct bird species.

Photo	Photo	Region	Prototype	Prototype's Photo	Similarity	Decision
 <p>Model predicts Species 106</p>	Step 1		compared to		0.00	Absent
	Step 2		compared to		0.00	Absent
	Step 3		compared to		1.00	Present
	Step 4		compared to		0.02	Absent
	Step 5		compared to		0.01	Absent
	Step 6		compared to		0.11	Absent
	Step 7		compared to		0.04	Absent
	Step 8		compared to		0.03	Absent
	Step 9		compared to		0.04	Absent

Q. Select the first step you disagree with the model's decision. If you agree with all steps, select "Agree with All."

☐ Step 1 ☐ Step 2 ☐ Step 3 ☐ Step 4 ☐ Step 5 ☐ Step 6 ☐ Step 7 ☐ Step 8 ☐ Step 9 ☐ Agree with All

Q. What do you think about the model's prediction?

☐ Fairly confident that prediction is correct  
☐ Somewhat confident that prediction is correct  
☐ Somewhat confident that prediction is incorrect  
☐ Fairly confident that prediction is incorrect

Click "Next Photo" after answering both questions.

1 / 10

Next Photo

Click "Next Page" after selecting answers for all 10 photos.

Next Page

Click "Model Description" to open or close model description.

Model Description

Fig. A21. 5. Task: ProtoTree *agreement*.



### Post-task evaluation

Q. How well do you think you understand the model's reasoning process?

☐ Very Poor ☐ Poor ☐ Fair ☐ Good ☐ Very Good

Next Page

### Your performance

In the previous task, 5 of 10 photos were correct predictions and the remaining 5 were incorrect predictions.

If we assign the 5 predictions with your highest "confident that prediction is correct" rating to correct and the rest as incorrect, you identified 3 out of 5 correct predictions and 3 out of 5 incorrect predictions.

Here are the individual answers you selected.

For the 5 correct predictions, you responded:

1. Fairly confident that prediction is correct
2. Fairly confident that prediction is correct
3. Somewhat confident that prediction is incorrect
4. Fairly confident that prediction is correct
5. Somewhat confident that prediction is incorrect

For the 5 incorrect predictions, you responded:

1. Somewhat confident that prediction is correct
2. Somewhat confident that prediction is correct
3. Fairly confident that prediction is incorrect
4. Fairly confident that prediction is incorrect
5. Fairly confident that prediction is incorrect

Q. How well do you think you understand the model's reasoning process?

☐ Very Poor ☐ Poor ☐ Fair ☒ Good ☐ Very Good

Next Page

Fig. A22. 6. Second and third subjective evaluation.

### Choose which model to use

The ProtoNet model achieves an overall accuracy of **79.9%** in 200 bird species recognition.

In the previous task, 5 of 10 photos were correct predictions and the remaining 5 were incorrect predictions. If we assign the 5 predictions with your highest "confident that prediction is correct" rating to correct and the rest as incorrect, you identified out of correct predictions and out of incorrect predictions. (When there are ties, we randomly assigned some to correct and some to incorrect.)

Alternatively, you can use a **Black-box** model that doesn't come with an explanation of its prediction.

Q. What is the minimum accuracy of the Black-box model that would convince you to use the Black-box model over the ProtoNet model?

[Low-risk setting] Scientific or educational purposes. E.g. You have a stack of bird images and want to know their species in a lab and/or a classroom.

Recall that the ProtoNet model achieves 79.9% accuracy.

70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100  
Selected Black-box model accuracy: 73%

[Medium-risk setting] Biodiversity and ecosystem monitoring. E.g. You want to collect large amounts of bird images and automatically label them.

Recall that the ProtoNet model achieves 79.9% accuracy.

70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100  
Selected Black-box model accuracy: 80%

[High-risk setting] Veterinary science or medical diagnosis. E.g. You have a sick bird and want to identify its species so that it can receive proper treatment and diagnosis.

Recall that the ProtoNet model achieves 79.9% accuracy.

70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100  
Selected Black-box model accuracy: 95%

Briefly describe the reason for your choices.

Next Page

Fig. A23. 7. Interpretability-accuracy tradeoff.

## References

1. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In: ICLR (2019)
2. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: Deep learning for interpretable image recognition. In: NeurIPS (2019)
3. Fel, T., Colin, J., Cadène, R., Serre, T.: What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods (2021)
4. Gildenblat, J., contributors: PyTorch library for CAM methods. <https://github.com/jacobgil/pytorch-grad-cam> (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
6. Hoffmann, A., Fanconi, C., Rade, R., Kohler, J.: This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. In: ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI (2021)
7. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: CVPR (2021)
8. Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In: NeurIPS (2021)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
11. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
12. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: CVPR Workshops (2020)
13. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: ECCV (2016)
14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)