HIVE: Evaluating the Human Interpretability of Visual Explanations

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky

Princeton University, Princeton NJ 08544, USA
{sunniesuhyoung, nmeister, vr23, ruthfong, olgarus}@princeton.edu

Abstract. As AI technology is increasingly applied to high-impact, high-risk domains, there have been a number of new methods aimed at making AI models more human interpretable. Despite the recent growth of interpretability work, there is a lack of systematic evaluation of proposed techniques. In this work, we introduce HIVE (Human Interpretability of Visual Explanations), a novel human evaluation framework that assesses the utility of explanations to human users in AI-assisted decision making scenarios, and enables falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation of visual interpretability methods. To the best of our knowledge, this is the first work of its kind. Using HIVE, we conduct IRB-approved human studies with nearly 1000 participants and evaluate four methods that represent the diversity of computer vision interpretability works: GradCAM, BagNet, ProtoPNet, and ProtoTree. Our results suggest that explanations engender human trust, even for incorrect predictions, yet are not distinct enough for users to distinguish between correct and incorrect predictions. We open-source HIVE to enable future studies and encourage more human-centered approaches to interpretability research. HIVE can be found at https://princetonvisualai.github.io/HIVE.

Keywords: Interpretability, Explainable AI (XAI), Human studies, Evaluation framework, Human-centered AI

1 Introduction

With the growing adoption of AI in high-impact, high-risk domains, there have been a surge of efforts aimed at making AI models more interpretable. Motivations for interpretability include allowing human users to trace through a model's reasoning process (accountability, transparency), verify that the model is basing its predictions on the right reasons (fairness, ethics), and assess their level of confidence in the model (trustworthiness). The *interpretability* research field tackles these questions and is comprised of diverse works, including those that provide explanations of the behavior and inner workings of complex AI models [6,7,25,27,50,61,64,73,77], those that design inherently interpretable models [10,13,14,15,17,18,38,48,53], and those that seek to understand what is easy and difficult for these models [3,68,75] to make their behavior more interpretable.

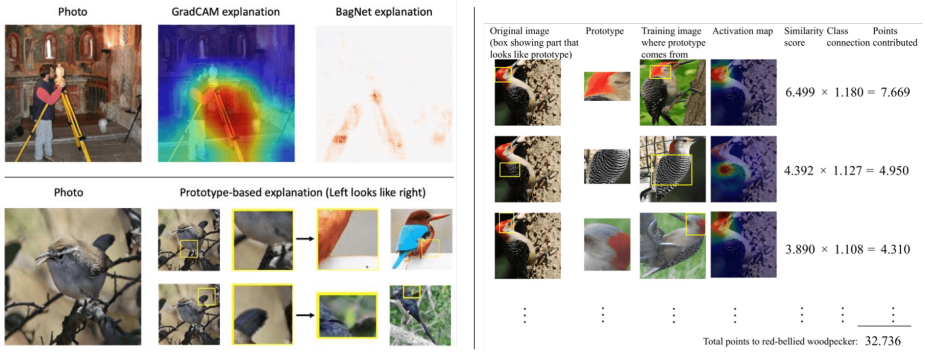

Fig. 1. Different forms of explanation. (Top left) Heatmap explanations (GradCAM [61], BagNet [10]) highlight decision-relevant image regions. (Bottom left) Prototype-based explanations (ProtoPNet [15], ProtoTree [48]) match image regions to prototypical parts learned during training. This schematic is much simpler than actual explanations. (Right) Actual ProtoPNet explanation example from the original paper. **While existing evaluation methods typically apply to only one explanation form, HIVE evaluates and compares diverse interpretability methods.**

Despite much methods development, there is a relative lack of standardized evaluation methods for proposed techniques. Existing evaluation methods for computer vision interpretability methods are focused on feature attribution heatmaps that highlight "important" image regions for a model's prediction. Since we lack ground-truth knowledge about which regions are *actually* responsible for the prediction, different evaluation metrics use different proxy tasks for verifying these important regions (e.g., measuring the impact of deleting regions or the overlap between ground-truth objects and highlighted regions) [26,34,50,51,70,74]. However, these automatic evaluation metrics are disconnected from downstream use cases of explanations; they don't capture how useful end-users find heatmaps in their decision making. Further, these metrics don't apply to other forms of explanations, such as prototype-based explanations produced by some of the recent interpretable-by-design models [15,17,48].

In part due to these challenges, the interpretability of a proposed method is often argued through a few exemplar explanations that highlight how a method is more interpretable than a baseline model. However, recent works suggest that some methods are not as interpretable as originally imagined and may engender over-trust in automated systems [1,19,32,33,46,47,49,62]. They caution against an over-reliance on intuition-based justifications and raise awareness for the need of falsifiable hypotheses [44] and proper evaluation in interpretability research.

Our contributions. As more diverse interpretability methods are being proposed, it is more important than ever to have a standardized and rigorous evaluation framework that allows for falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation. To this end, we develop HIVE (Human Interpretability of Visual Explanations). HIVE evaluates diverse vi-

sual interpretability methods by evaluating all methods on a common task. We carefully design the tasks to reduce the effect of confirmation bias and human prior knowledge in interpretability evaluation, and assess the utility of explanations in AI-assisted decision making scenarios. HIVE also examines how well interpretable-by-design models' reasoning process aligns with that of humans, and how human users tradeoff interpretability and accuracy.

To demonstrate the extensibility and applicability of HIVE, we conduct IRB-approved human studies with nearly 1000 participants and evaluate four existing methods that represent different streams of interpretability work (e.g., post-hoc explanations, interpretable-by-design models, heatmaps, and prototype-based explanations): GradCAM [61], BagNet [10], ProtoPNet [15], ProtoTree [48]. To the best of our knowledge, we are the first to compare interpretability methods with different explanation forms (see Fig. 1) and the first to conduct human studies of the evaluated interpretable-by-design models [10,15,48].

We obtain a number of insights through our studies:

- When provided explanations, participants tend to believe that the model predictions are correct, revealing an issue of *confirmation bias*. For example, our participants found 60% of the explanations for *incorrect* model predictions convincing. Prior work has made similar observations for non-visual interpretability methods [52]; we substantiate them for visual explanations and demonstrate a need for rigorous evaluation of proposed methods.
- When given multiple model predictions and explanations, participants struggle to distinguish between correct and incorrect predictions based on the explanations (e.g., achieving only 40% accuracy on a multiple-choice task with four options). This result suggests that interpretability methods need to be improved to be reliably useful for AI-assisted decision making.
- There exists a gap between the similarity judgments of humans and prototype-based models [15,48] which can hurt the quality of their interpretability.
- Participants prefer to use a model with explanations over a baseline model without explanations. To switch their preference, they require the baseline model to have +6.2% to +10.9% higher accuracy.

As interpretability is fundamentally a human-centric concept, it needs to be evaluated in a human-centric way. We hope our work helps pave the way towards human evaluation becoming commonplace, by presenting and analyzing a human study design, demonstrating its effectiveness and informativeness for interpretability evaluation, and open-sourcing the code to enable future work.

2 Related work

Interpretability landscape in computer vision. Interpretability research can be described along several axes: first, whether a method is post-hoc or interpretable-by-design; second, whether it is global or local; and third, the form of an explanation (see [4,11,16,24,28,30,57,59] for surveys). *Post-hoc explanations* focus on explaining predictions made by already-trained models, whereas *interpretable-by-design (IBD)* models are intentionally designed to possess a

more explicitly interpretable decision-making process [10,13,14,15,17,18,38,48,53]. Furthermore, explanations can either be *local explanations* of a single input-output example or *global explanations* of a network (or its component parts). Local, post-hoc methods include heatmap [25,50,61,63,64,73,77], counterfactual explanation [29,65,69], approximation [56], and sample importance [37,71] methods. In contrast, global, post-hoc methods aim to understand global properties of CNNs, often by treating them as an object of scientific study [6,7,27,36] or by generating class-level explanations [55,78]. Because we focus on evaluating the utility of explanations in AI-assisted decision making, we do not evaluate global, post-hoc methods. *IBD* models can provide local and/or global explanations, depending on the model type. Lastly, explanations can take a variety of forms: two more popular ones we study are *heatmaps* highlighting important image regions and *prototypes* (i.e., image patches) from the training set that form interpretable decisions. In our work, we investigate four popular methods that span these types of interpretability work: GradCAM [61] (post-hoc, heatmap), BagNet [10] (IBD, heatmap), ProtoPNet [15] (IBD, prototypes), and ProtoTree [48] (IBD, prototypes). See Fig. 1 for examples of their explanations.

Evaluating heatmaps. Heatmap methods are arguably the most-studied class of interpretability work. Several automatic evaluation metrics have been proposed [5,26,34,50,51,70,74], however, there is a lack of consensus on how to evaluate these methods. Further, the authors of [1,2] and BAM [70] highlight how several methods fail basic "sanity checks" and call for more comprehensive metrics. Complementing these works, we use HIVE to study how useful heatmaps are to human users in AI-assisted decision making scenarios and demonstrate insights that cannot be gained from automatic evaluation metrics.

Evaluating interpretable-by-design models. In contrast, there has been relatively little work on assessing interpretable-by-design models. Quantitative evaluations of these methods typically focus on demonstrating their competitive performance with a baseline CNN, while the quality of their interpretability is often demonstrated through qualitative examples. Recently, a few works revisited several methods' interpretability claims. Hoffmann et al. [33] highlight that prototype similarity of ProtoPNet [15] does not correspond to semantic similarity and that this disconnect can be exploited. Margeloiu et al. [47] analyze concept bottleneck models [38] and demonstrate that learned concepts fail to correspond to real-world, semantic concepts. In this work, we conduct the first human study of three popular interpretable-by-design models [10,15,48] and quantify prior work's [33,48] anecdotal observation on the misalignment between prototype-based models [15,48] and humans' similarity judgment.

Evaluating interpretability with human studies. Outside the computer vision field, human studies are commonly conducted for models trained on tabular datasets [40,41,43,52,76]; however, these do not scale to the complexity of modern vision models. Early human studies for visual explanations have been limited in scope: They typically ask participants which explanation they find more reasonable or which model they find more trustworthy based on explanations [35,61]. Recently, more diverse human studies have been conducted [8,9,23,49,62,63,80].

Closest to our work are [23,49,62]. Shen and Huang [62] ask users to select incorrectly predicted labels with or without showing explanations; Nguyen et al. [49] ask users to decide whether model predictions are correct based on explanations; Fel et al. [23] ask users to predict model outputs in a concurrent work. Regarding [49,62], our *distinction* task also investigates how useful explanations are in distinguishing correct and incorrect predictions. However, different from these works, we ask users to select the correct prediction out of multiple predictions to reduce the effect of confirmation bias and don't show class labels to prevent users from relying their prior knowledge. Regarding [23], we also ask users to predict model outputs, but mainly as a supplement to our *distinction* task. Further, we ask users to identify the model output out of multiple predictions based on the explanations, whereas [23] first trains users to be a meta-predictor of the model by showing example model predictions and explanations, and then at test time asks users to predict the model output for a given image without showing any explanation. Most importantly, different from [23,49,62], we evaluate interpretability methods beyond heatmaps and conduct cross-method comparison. Our work is similar in spirit to work by Zhou et al. [79] on evaluating generative models with human perception. For general guidance on running human studies in computer vision, refer to work by Bylinskii et al. [12].

3 HIVE design principles

In this work, we focus on AI-assisted decision making scenarios, in particular those that involve an image classification model. For a given input image, a user is shown a model's prediction along with an associated explanation, and is asked to make a decision about whether the model's prediction is correct or more generally about whether to use the model. In such a scenario, explanations are provided with several goals in mind: help the user identify if the model is making an error, arrive at a more accurate prediction, understand the model's reasoning process, decide how much to trust the model, etc.

To study whether and to what extent different visual interpretability methods are useful for AI-assisted decision making, we develop a novel human evaluation framework named HIVE (Human Interpretability of Visual Explanations). In particular, we design HIVE to allow for *falsifiable hypothesis testing* regarding the usefulness of explanations for identifying model errors, *cross-method comparison* between different explanation approaches, and *human-centered evaluation* for understanding the practical effectiveness of interpretability.

3.1 Falsifiable hypothesis testing

We join a growing body of work that cautions against intuition-based justification and subjective self-reported ratings in interpretability evaluation [1,44,39,60] and calls for objective assessment with behavior indicators [42,52,72,76]. To this end, we design two evaluation tasks, the *agreement* and *distinction* tasks, that enable *falsifiable hypothesis testing* about the evaluated interpretability method.

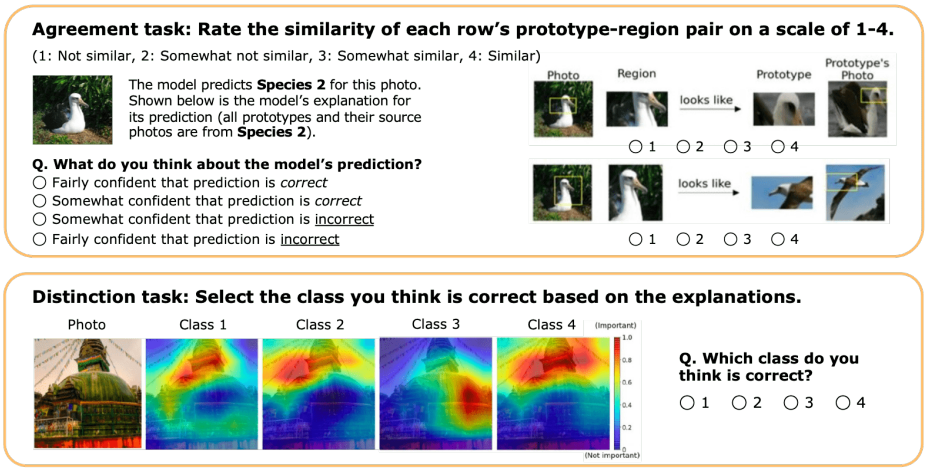

Fig. 2. Study user interfaces (UIs). We show simplified UIs for evaluating ProtoPNet [15] on the *agreement* task (top) and GradCAM [61] on the *distinction* task (bottom). Full UI snapshots are in supp. mat. See Sec. 3 for description of the tasks.

In the *agreement* task, we present participants with one prediction-explanation pair at a time and ask how confident they are in the model's prediction based on the explanation. We evaluate methods on this task in part because it is closer to existing interpretability evaluation schemes that consider a model's top-1 prediction and its explanation [61], and also because it allows us to quantify the degree to which participants believe in model predictions based on explanations.

The *agreement* task measures the amount of *confirmation bias* that arises for a given interpretability method. However, it doesn't measure the utility of explanations in distinguishing correct and incorrect predictions, a crucial functionality of explanations in AI-assisted decision making. Hence, we design and use the *distinction* task as our main evaluation task. Here we simultaneously show four predictions and their associated explanations for a given input image and ask users to identify the correct prediction based on the provided explanations. The *distinction* task also mitigates the effect of confirmation bias in interpretability evaluation, as participants now have to reason about multiple explanations at once. See Fig. 2 for the evaluation task UIs.

One concern with this setup is ensuring that participants use the provided explanations rather than their knowledge to complete the task. We take two measures to remove the effect of *human prior knowledge* in our evaluations. First, we evaluate all interpretability methods in the context of fine-grained bird species classification [66], which is a challenging task for non-bird experts. Second, as a more general measure, we omit the semantic class labels of the predictions. This measure is particularly important when evaluating interpretability methods in easier contexts, e.g., coarse-grained object classification with ImageNet [58], because the task becomes too easy otherwise (i.e., participants can select the cor-

rect prediction based on the class labels instead of using the explanations). Note that ground-truth class labels are also omitted to simulate a realistic decision making scenario where users do not have access to the ground truth.

3.2 Cross-method comparison

Existing evaluation methods typically apply to only one explanation form (e.g., heatmaps are compared against each other). In contrast, HIVE enables *cross-method comparison* between different explanation forms by focusing on downstream uses of explanations and evaluating all methods on a common task.

However, there remains a number of practical roadblocks. First, different methods may have been developed for different scenarios (e.g., fine-grained vs. coarse-grained classification), requiring us to carefully analyze the effect of the particular setting during evaluation. Second, different methods may be more or less digestible to the users. While this is an inherent part of what we are trying to evaluate, we also want to ensure that the evaluation task is doable by study participants with limited machine learning background, given most human studies in the field are run through Amazon Mechanical Turk. Hence, we actualize a specific evaluation setup for each interpretability method by creating an individual evaluation UI that respects the method's characteristics (e.g., its explanation form, dataset used for model training). We briefly describe the four methods we evaluate in this work (see Fig. 1 for example explanations) and their evaluation setups. When making any adaptations, we tried to present each method in as favorable of a way as possible. More details are in supp. mat.

GradCAM [61]. GradCAM is a post-hoc method that produces a heatmap that highlights important regions in an input image that contribute to a model's prediction. We evaluate GradCAM on ImageNet [58], which it was originally developed for, as well as on CUB [66], for which we train a standard CNN model to use as the underlying model for generating GradCAM heatmaps.

BagNet [10]. In contrast, BagNet is an interpretable-by-design model that collects evidence for a class from small regions of an image. For each class, BagNet creates a heatmap where higher values (i.e., darker red in our visualizations) imply stronger evidence for the class. BagNet then sums the values in each heatmap and predicts the class with the highest sum. We evaluate BagNet on ImageNet, for which it was originally designed, as well as on CUB, for which we train a new BagNet model using the authors' code.

ProtoPNet [15]. The next two methods reason with *prototypes*, which are small image patches from the training set that these models deem as representative for a certain class. At test time, ProtoPNet compares a given image to the set of prototypes it learned during training and finds regions in the image that are the most similar to each prototype. It computes a similarity score between each prototype-region pair, then predicts the class with the highest weighted sum of the similarity scores. The ProtoPNet model for CUB learns 10 prototypes for each of the 200 bird species (2,000 total) and produces one of the most complex explanations. Its explanation for a single prediction consists of 10 prototypes and

their source images, heatmaps that convey the similarity between matched image regions and prototypes, continuous and unnormalized similarity scores, and weights multiplied to the scores (see Fig. 1 right). In our evaluation, we abstract away most technical details based on our pilot studies, and focus on showing the most crucial component of ProtoPNet's reasoning process: the prototype-image region matches. We also ask participants to rate the similarity of each match (see Fig. 2 top) to assess how well the model's similarity judgment aligns with that of humans. See supp. mat. for the task and explanation modification details.

ProtoTree [48]. Finally, the ProtoTree model learns a tree structure along with the prototypes. Each node in the tree contains a prototype from a training image. At each node, the model compares a given test image to the node's prototype and produces a similarity score. If the score is above some threshold, the model judges that the prototype is present in the image and absent if not. The model then proceeds to the next node and repeats this process until it reaches a leaf node, which corresponds to a class. The ProtoTree model for CUB trained by the authors has 511 decision nodes and up to 10 decision steps, and our pilot studies revealed that is too overwhelming for participants. Thus in our evaluation, we significantly simplify the decision process. Participants are shown the model's decisions until the penultimate decision node, and then are asked to make decisions for only the final two nodes of the tree by judging whether the prototype in each node is present or absent in the image. This leads the participants to select one of the four (2^2) classes as the final prediction. One additional challenge is that participants may not be familiar with decision trees and thus may have trouble following the explanation. To help understanding, we introduce a simple decision tree model with two levels, walk through an example, and present two warm up exercises so that participants can get familiar with decision trees before encountering ProtoTree. See supp. mat. for more information.

3.3 Human-centered evaluation

HIVE complements existing algorithmic evaluation methods by bringing humans back into the picture and taking a *human-centered* approach to interpretability evaluation. The design of HIVE, particularly the inclusion/exclusion of class labels in Sec. 3.1 and careful actualization of the evaluation setup in Sec. 3.2, is focused on making this evaluation tractable for the participants and as fair as possible with respect to different interpretability methods. We also went through multiple iterations of UI design to present visual explanations in digestible bits so as to not overwhelm participants with their complexity. Despite the challenges, there is a very important payoff from human studies. We are able to evaluate different interpretability methods through participants' 1) ability to *distinguish* between correct and incorrect predictions based on the provided explanations, simulating a more realistic AI-assisted decision-making setting, and 2) level of *alignment* with the model's intermediate reasoning process in the case of prototype-based, interpretable-by-design models. We also gain a number of valuable insights that can only be obtained through human studies.

3.4 Generalizability & Scalability

In closing we discuss two common concerns about human studies: generalizability and scalability. We have shown HIVE's *generalizability* by using it to evaluate a variety of methods (post-hoc explanations, interpretable-by-design models, heatmaps, prototype-based explanations) in two different settings (coarse-grained object recognition with ImageNet, fine-grained bird recognition with CUB). Further, a recent work by Ramaswamy et al. [54] uses HIVE to set up new human studies, for evaluating example-based explanations and finding the ideal complexity of concept-based explanations, demonstrating that HIVE can be easily generalized to new methods and tasks. Regarding *scalability*, human study costs are not exorbitant contrary to popular belief and can be budgeted for like we budget for compute. For example, our GradCAM distinction study cost $70 with 50 participants compensated at $12/hr. The real obstacles are typically the time, effort, and expertise required for study design and UI development; with HIVE open-sourced, these costs are substantially mitigated.

4 HIVE study design

In this section, we describe our IRB-approved study design. See supp. mat. and `https://princetonvisualai.github.io/HIVE` for UI snapshots and code.

Introduction. For each participant, we first introduce the study and receive their informed consent. We also request optional demographic data regarding gender identity, race and ethnicity, and ask about the participant's experience with machine learning; however, no personally identifiable information was collected. Next we explain the evaluated interpretability method in simple terms by avoiding technical jargon (i.e., replacing terms like "image" and "training set" to "photo" and "previously-seen photos"). We then show a preview of the evaluation task and provide example explanations for one correct and one incorrect prediction made by the model to give the participant appropriate references. The participant can access the method description at any time during the task.

Objective evaluation tasks. Next we evaluate the interpretability method on a behavioral task (*distinction* or *agreement*) introduced in Sec. 3.1 and Fig. 2. Detailed task descriptions are available in supp. mat.

Subjective evaluation questions. While the core of HIVE is in the objective evaluation tasks, we also ask subjective evaluation questions to make the most out of the human studies. Specifically, we ask the participant to self-rate their level of understanding of the evaluated method before and after completing the task, to investigate if the participant's self-rated level of understanding undergoes any changes during the task. After the task completion, we disclose the participant's performance on the task and ask the question one last time.

Interpretability-accuracy tradeoff questions. While interpretability methods offer useful insights into a model's decision, some explanations come at the cost of lower model accuracy. Hence in the final part of the study, we investigate the *interpretability-accuracy tradeoff* participants are willing to make when comparing an interpretable method against a baseline model that doesn't come with

any explanation. In high-risk scenarios a user may prefer to maximize model performance over interpretability. However, another user may prefer to prioritize interpretability in such settings so that there would be mechanisms for examining the model's predictions. To gain insight into the tradeoff users are willing to make, we present three scenarios: low-risk (e.g., bird species recognition for scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for autonomous driving). For each scenario, we then ask the participant to input the minimum accuracy of the baseline model that would convince them to use it over the model with explanations and also describe the reason for their choices.

5 Experiments

5.1 Experimental details

Datasets & Models. We evaluate all interpretability methods on classification tasks and use images from the CUB [66] test set and the ImageNet [58] validation set to generate model predictions and explanations. On CUB, we evaluate all four methods: GradCAM [61], BagNet [10], ProtoPNet [15], ProtoTree [48]. On ImageNet, we evaluate GradCAM and BagNet. See supp. mat. for details.
Human studies. For each study, i.e., an evaluation of one interpretability method on one task (*distinction* or *agreement*), we recruited 50 participants through Amazon Mechanical Turk (AMT). In total, we conducted 19 studies with 950 participants; see supp. mat. for the full list. The self-reported machine learning experience of the participants was 2:5 1:0, between "2: have heard about..." and "3: know the basics..." The mean study duration was 6.9 minutes for GradCAM, 6.6 for BagNet, 13.6 for ProtoPNet, and 10.4 for ProtoTree. Participants were compensated based on the state-level minimum wage of $12/hr.
Statistical analysis. For each study, we report the mean task accuracy and standard deviation of the participants' performance which captures the variability between individual participants' performance. We also compare the study result to random chance and compute the p-value from a 1-sample t-test.[1] When comparing results between two groups, we compute the p-value from a 2-sample t-test. Results are deemed statistically significant under $p < 0:05$ conditions.

5.2 The issue of confirmation bias

Let us first examine how the four methods perform on the *agreement* task, where we present participants with one prediction-explanation pair at a time and ask how confident they are in the model's prediction. Results are summarized in Tab. 1. On CUB, participants found 72.4% of correct predictions convincing for

[1] We compare our results to chance performance instead of a baseline without explanations because we omit semantic class labels to remove the effect of human prior knowledge (see Sec. 3.1); so such a baseline would contain no relevant information.

Table 1. Agreement task results. 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*. See Sec. 5.2 for a discussion.

CUB	GradCAM [61]	BagNet [10]	ProtoPNet [15]	ProtoTree [48]
Correct	72.4% ± 21.5 (2.9)	**75.6% ± 23.4 (3.0)**	73.2% ± 24.9 (3.0)	66.0% ± 33.8 (2.8)
Incorrect	32.8% ± 24.3 (2.8)	***42.4% ± 28.7 (2.7)***	***46.4% ± 35.9 (2.4)***	37.2% ± 34.4 (2.7)
ImageNet	GradCAM [61]	BagNet [10]	-	-
Correct	**70.8% ± 26.6 (2.9)**	66.0% ± 27.2 (2.8)	-	-
Incorrect	***44.8% ± 31.6 (2.7)***	35.6% ± 26.9 (2.7)	-	-

GradCAM, 75.6% for BagNet, 73.2% for ProtoPNet, and 66.0% ProtoTree. However, they also thought 67.2% of incorrect predictions were correct for GradCAM, 57.6% for BagNet, 53.6% for ProtoPNet, and 62.8% for ProtoTree. Similarly on ImageNet, participants found 70.8% of correct predictions convincing for GradCAM and 66.0% for BagNet, yet also believed in 55.2% and 64.4% of incorrect predictions, respectively. These results reveal an issue of *confirmation bias*: When given explanations, participants tend to believe model predictions are correct, even if they are wrong. Still, the confidence ratings are overall higher for correct predictions than incorrect predictions, suggesting there is some difference between their explanations. More results and discussion are in supp. mat.

5.3 Objective assessment of interpretability

Next we discuss findings from our main evaluation task, the *distinction* task, where we ask participants to select the correct prediction out of four options based on the provided explanations. Results are summarized in Tab. 2.

Participants perform better on correctly predicted samples. On correctly predicted samples from CUB, the mean task accuracies are 71.2% on GradCAM, 45.6% on BagNet, 54.5% on ProtoPNet and 33.8% on ProtoTree, all above the 25% chance baseline. That is, participants can identify which of the four explanations correspond to the ground-truth class correctly predicted by the model. On incorrect predictions, however, the accuracies drop from 71.2% to 26.4% for GradCAM and from 45.6% to 32.0% for BagNet, and we observe a similar trend in the ImageNet studies. These results suggest that explanations for correct predictions may be more coherent and convincing than those for incorrect predictions. Even so, all accuracies are far from 100%, indicating that the evaluated methods are not yet reliably useful for AI-assisted decision making.

Participants struggle to identify the model's prediction. For GradCAM and BagNet, we ask participants to select the class they think the model predicts (*output prediction*) in addition to the class they think is correct (*distinction*). For

Table 2. Distinction and output prediction task results. For each study, we report the mean accuracy and standard deviation of the participants' performance. *Italics* denotes methods that do not statistically significantly outperform 25% random chance ($p > 0.05$); **bold** denotes the highest performing method in each group. In the top half, we show the results of all four methods on CUB. In the bottom half, we show GradCAM and BagNet results on ImageNet, without vs. with ground-truth class labels. **Overall, participants struggle to identify the correct prediction or the model output based on explanations.** See Sec. 5.3 for a discussion.

CUB		GradCAM [61]	BagNet [10]	ProtoPNet [15]	ProtoTree [48]
Distinction	Correct	**71.2% ± 33.3**	45.6% ± 28.0	54.5% ± 30.3	33.8% ± 15.9
	Incorrect	*26.4% ± 19.8*	**32.0% ± 20.8**	-	-
Output prediction	Correct	**69.2% ± 32.3**	50.4% ± 32.8	-	-
	Incorrect	**53.6% ± 27.0**	*30.0% ± 24.1*	-	-
ImageNet		GradCAM [61]	with labels	BagNet [10]	with labels
Distinction	Correct	**51.2% ± 24.7**	49.2% ± 30.8	38.4% ± 28.0	34.8% ± 27.7
	Incorrect	*30.0% ± 22.4*	*27.2% ± 20.3*	*26.0% ± 18.4*	*27.2% ± 18.7*
Output prediction	Correct	**48.0% ± 28.3**	**48.0% ± 35.6**	46.8% ± 29.0	42.8% ± 27.4
	Incorrect	**35.6% ± 24.1**	33.2% ± 25.2	34.0% ± 24.1	32.8% ± 25.5

BagNet, this is a straightforward task where participants just need to identify the most activated (most red, least blue) heatmap among the four options, as BagNet by design predicts the class with the most activated heatmap. However, accuracy is not very high, only marginally above the *distinction* task accuracy. This result suggests that BagNet heatmaps for the top-4 (or top-3 plus ground-truth) classes look similar to the human eye, and may not be suitable for assisting humans with tasks that involve distinguishing one class from another. For GradCAM, participants also struggle on this task but to a lesser degree.

Showing ground-truth labels hurts performance. For GradCAM and BagNet, we also investigate the effect of showing ground-truth class labels for the presented images. We have not been showing them to simulate a realistic decision making scenario where users don't have access to the ground truth. However, since the task may be ambiguous for datasets like ImageNet whose images may contain several objects, we run a second version of the ImageNet studies showing ground-truth class labels on the same set of images and compare results. Somewhat surprisingly, we find that accuracy decreases, albeit by a small amount, with class labels. One possible explanation is that class labels implicitly bias participants to value heatmaps with better localization properties, which could be a suboptimal signal for the *distinction* and *output prediction* tasks.

Automatic evaluation metrics correlate poorly with human study results. We also analyze GradCAM results using three automatic metrics that evaluate the localization quality of post-hoc attribution maps: pointing game [74], energy-based pointing game [67], and intersection-over-union [77]. In the *agreement* studies, we find near-zero correlation between participants' confidence in the model prediction and localization quality of heatmaps. In the *distinction* studies, we also do not see meaningful relationships between the participants' choices and these automatic metrics. These observations are consistent with the

findings of [49,23], i.e., automatic metrics poorly correlate with human performance in post-hoc attribution heatmap evaluation. See supp. mat. for details.

5.4 A closer examination of prototype-based models

We are the first to conduct human studies of ProtoPNet and ProtoTree which produce some of the most complex visual explanations. As such, we take a closer look at their results to better understand how human users perceive them.

A gap exists between similarity ratings of ProtoPNet & ProtoTree and those of humans. We quantify prior work's [33,48] anecdotal observation that there exists a gap between model and human similarity judgment. For ProtoTree, the Pearson correlation coefficient between the participants' similarity ratings and the model similarity scores is 0.06, suggesting little to no relationship. For ProtoPNet, whose similarity scores are not normalized across images, we compute the Spearman's rank correlation coefficient (= 0:25; p = 0:49 for *distinction* and = 0:52, p = 0:12 for *agreement*). There is no significant negative correlation between the two, indicating a gap in similarity judgment that may hurt the models' interpretability. See supp. mat. for more discussion.

Participants perform relatively poorly on ProtoTree, but they understand how a decision tree works. Since the previously described ProtoTree *agreement* study does not take into account the model's inherent tree structure, we run another version of the study where, instead of asking participants to rate each prototype's similarity, we ask them to select the first step they disagree with in the model's explanation. The result of this study (52:8% 19:9%) is similar to that of the original study (53:6% 15:2%); in both cases, we cannot conclude that participants outperform 50% random chance (p = 0:33, p = 0:10). To ensure participants understand how decision trees work, we provided a simple decision tree example and subsequent questions asking participants if the decision tree example makes a correct or incorrect prediction. 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 supp. mat. for details.

5.5 Subjective evaluation of interpretability

To complement the objective evaluation tasks, we asked participants to self-rate their level of method understanding three times. The average ratings are 3:7 0:9 after the method explanation, 3:8 0:9 after the task, and 3:5 1:0 after seeing their task performance, which all lie between the fair (3) and good (4) ratings. Interestingly, the rating tends to *decrease* after participants see their task performance (p < 0:05). Several participants indicated that their performance was lower than what they expected, whereas no one suggested the opposite, suggesting that participants might have been disappointed in their task performance, which in turn led them to lower their self-rated level of method understanding.

5.6 Interpretability-accuracy tradeoff

In the final part of our studies, we asked participants for the minimum accuracy of a baseline model they would require to use it over the evaluated interpretable model with explanations for its predictions. Across all studies, participants require the baseline model to have a higher accuracy than the model that comes with explanations, and by a greater margin for higher-risk settings. On average, participants require the baseline model to have +6.2% higher accuracy for low-risk, +8.2% for medium-risk, and +10.9% for high-risk settings. See supp. mat. for the full results and the participants' reasons for their choices.

6 Conclusion

In short, we introduce and open-source HIVE, a novel human evaluation framework for evaluating diverse visual interpretability methods, and use it to evaluate four existing methods: GradCAM, BagNet, ProtoPNet, and ProtoTree.

There are a few limitations of our work: First, we use a relatively small sample size of 50 participants for each study due to our desire to evaluate four methods, some under multiple conditions. Second, while HIVE takes a step towards use case driven evaluation, our evaluation setup is still far from real-world uses of interpretability methods. An ideal evaluation would be contextually situated and conducted with domain experts and/or end-users of a real-world application (e.g., how would bird experts choose to use one method over another when given multiple interpretability methods for a bird species recognition model).

Nonetheless, we believe our work will facilitate more user studies and encourage human-centered interpretability research [20,21,22,45], as our human evaluation reveals several key insights about the field. In particular, we find that participants generally believe model predictions are correct when given explanations for them. Humans are naturally susceptible to confirmation bias; thus, interpretable explanations will likely engender trust from humans, even if they are incorrect. Our findings underscore the need for evaluation methods that fairly and rigorously assess the usefulness and effect of explanations. We hope our work helps shift the field's objective from focusing on method development to also prioritizing the development of high-quality evaluation methods.

Acknowledgments. This material is based upon work partially supported by the National Science Foundation (NSF) under Grant No. 1763642. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), Princeton SEAS Project X Fund (RF, OR), Open Philanthropy (RF, OR), and Princeton SEAS and ECE Senior Thesis Funding (NM). We thank the authors of [10,15,31,33,48,61] for open-sourcing their code and/or trained models. We also thank the AMT workers who participated in our studies, anonymous reviewers who provided thoughtful feedback, and Princeton Visual AI Lab members (especially Dora Zhao, Kaiyu Yang, and Angelina Wang) who tested our user interface and provided helpful suggestions.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS (2018)
2. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: NeurIPS (2020)
3. Agarwal, C., D'souza, D., Hooker, S.: Estimating example difficulty using variance of gradients. In: CVPR (2022)
4. Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion (2020)
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one (2015)
6. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: CVPR (2017)
7. Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A.: Seeing what a GAN cannot generate. In: ICCV (2019)
8. Biessmann, F., Refiano, D.I.: A psychophysics approach for quantitative comparison of interpretable computer vision models (2019)
9. Borowski, J., Zimmermann, R.S., Schepers, J., Geirhos, R., Wallis, T.S.A., Bethge, M., Brendel, W.: Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. In: ICLR (2021)
10. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In: ICLR (2019)
11. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G.K., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P.W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J.B., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S.Ó., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T.K., Dyer, L., Khan, S., Bengio, Y., Anderljung, M.: Toward trustworthy AI development: Mechanisms for supporting verifiable claims (2020)
12. Bylinskii, Z., Herman, L., Hertzmann, A., Hutka, S., Zhang, Y.: Towards better user studies in computer graphics and vision. arXiv (2022)
13. Böhle, M., Fritz, M., Schiele, B.: Convolutional dynamic alignment networks for interpretable classifications. In: CVPR (2021)
14. Böhle, M., Fritz, M., Schiele, B.: B-cos networks: Alignment is all we need for interpretability. In: CVPR (2022)
15. 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)
16. Chen, V., Li, J., Kim, J.S., Plumb, G., Talwalkar, A.: Towards connecting use cases and methods in interpretable machine learning. In: ICML Workshop on Human Interpretability in Machine Learning (2021)
17. Donnelly, J., Barnett, A.J., Chen, C.: Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. In: CVPR (2022)

18. Dubey, A., Radenovic, F., Mahajan, D.: Scalable interpretability via polynomials. arXiv (2022)
19. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. In: IJHCS (2003)
20. Ehsan, U., Riedl, M.O.: Human-centered explainable AI: Towards a reflective sociotechnical approach. In: HCII Late Breaking Papers (2020)
21. Ehsan, U., Wintersberger, P., Liao, Q.V., Mara, M., Streit, M., Wachter, S., Riener, A., Riedl, M.O.: Operationalizing human-centered perspectives in explainable AI. CHI Extended Abstracts (2021)
22. Ehsan, U., Wintersberger, P., Liao, Q.V., Watkins, E.A., Manger, C., III, H.D., Riener, A., Riedl, M.O.: Human-Centered Explainable AI (HCXAI): Beyond opening the black-box of AI. CHI Extended Abstracts (2022)
23. 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)
24. Fong, R.: Understanding convolutional neural networks. Ph.D. thesis, University of Oxford (2020)
25. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV (2019)
26. Fong, R., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017)
27. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: CVPR (2018)
28. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: DSAA (2018)
29. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: ICML (2019)
30. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Magazine (2019)
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
32. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: CSCW (2000)
33. 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)
34. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: NeurIPS (2019)
35. Jeyakumar, J.V., Noor, J., Cheng, Y.H., Garcia, L., Srivastava, M.: How can I explain this to you? An empirical study of deep neural network explanation methods. In: NeurIPS (2020)
36. Kim, B., Reif, E., Wattenberg, M., Bengio, S., Mozer, M.C.: Neural networks trained on natural scenes exhibit gestalt closure. Computational Brain & Behavior (2021)
37. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: ICML (2017)
38. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML (2020)
39. Kunkel, J., Donkers, T., Michael, L., Barbu, C.M., Ziegler, J.: Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In: CHI (2019)

40. Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S.J., Doshi-Velez, F.: Human evaluation of models built for interpretability. In: HCOMP (2019)
41. Lage, I., Ross, A.S., Kim, B., Gershman, S.J., Doshi-Velez, F.: Human-in-the-loop interpretability prior. In: NeurIPS (2018)
42. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: FAccT (2019)
43. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: A joint framework for description and prediction. In: KDD (2016)
44. Leavitt, M.L., Morcos, A.S.: Towards falsifiable interpretability research. In: NeurIPS Workshop on ML Retrospectives, Surveys & Meta-Analyses (2020)
45. Liao, Q.V., Varshney, K.R.: Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv (2021)
46. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue (2018)
47. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A.: Do concept bottleneck models learn as intended? In: ICLR Workshop on Responsible AI (2021)
48. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: CVPR (2021)
49. Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In: NeurIPS (2021)
50. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: BMVC (2018)
51. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In: CVPR Workshop on Responsible Computer Vision (2021)
52. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability. In: CHI (2021)
53. Radenovic, F., Dubey, A., Mahajan, D.: Neural basis models for interpretability. arXiv (2022)
54. Ramaswamy, V.V., Kim, S.S.Y., Fong, R., Russakovsky, O.: Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. arXiv (2022)
55. Ramaswamy, V.V., Kim, S.S.Y., Meister, N., Fong, R., Russakovsky, O.: ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. arXiv (2022)
56. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: KDD (2016)
57. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. In: Statistics Surveys (2021)
58. 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)
59. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer (2019)
60. Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., Höllerer, T.: I can do better than your AI: Expertise and explanations. In: IUI (2019)
61. 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)

62. Shen, H., Huang, T.H.K.: How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. In: HCOMP (2020)
63. Shitole, V., Li, F., Kahng, M., Tadepalli, P., Fern, A.: One explanation is not enough: Structured attention graphs for image classification. In: NeurIPS (2021)
64. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR Workshops (2014)
65. Vandenhende, S., Mahajan, D., Radenovic, F., Ghadiyaram, D.: Making heads or tails: Towards semantically consistent visual counterfactuals. In: ECCV (2022)
66. 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)
67. 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)
68. Wang, P., Vasconcelos, N.: Towards realistic predictors. In: ECCV (2018)
69. Wang, P., Vasconcelos, N.: SCOUT: Self-aware discriminant counterfactual explanations. In: CVPR (2020)
70. Yang, M., Kim, B.: Benchmarking attribution methods with relative feature importance (2019)
71. Yeh, C.K., Kim, J., Yen, I.E.H., Ravikumar, P.K.: Representer point selection for explaining deep neural networks. In: NeurIPS (2018)
72. Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: CHI (2019)
73. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
74. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: ECCV (2016)
75. Zhang, P., Wang, J., Farhadi, A., Hebert, M., Parikh, D.: Predicting failures of vision systems. In: CVPR (2014)
76. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect on confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: FAccT (2020)
77. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
78. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: ECCV (2018)
79. Zhou, S., Gordon, M.L., Krishna, R., Narcomey, A., Fei-Fei, L., Bernstein, M.S.: HYPE: A benchmark for human eye perceptual evaluation of generative models. In: NeurIPS (2019)
80. Zimmermann, R.S., Borowski, J., Geirhos, R., Bethge, M., Wallis, T.S.A., Brendel, W.: How well do feature visualizations support causal understanding of CNN activations? In: NeurIPS (2021)