Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior

Angie Boggust
aboggust@mit.edu
MIT CSAIL
Cambridge, Massachusetts, USA

Benjamin Hoover
IBM Research
Cambridge, Massachusetts, USA

Arvind Satyanarayan
MIT CSAIL
Cambridge, Massachusetts, USA

Hendrik Strobelt
IBM Research
Cambridge, Massachusetts, USA

ABSTRACT

Saliency methods — techniques to identify the importance of input features on a model's output — are a common step in understanding neural network behavior. However, interpreting saliency requires tedious manual inspection to identify and aggregate patterns in model behavior, resulting in ad hoc or cherry-picked analysis. To address these concerns, we present Shared Interest: metrics for comparing model reasoning (via saliency) to human reasoning (via ground truth annotations). By providing quantitative descriptors, Shared Interest enables ranking, sorting, and aggregating inputs, thereby facilitating large-scale systematic analysis of model behavior. We use Shared Interest to identify eight recurring patterns in model behavior, such as cases where contextual features or a subset of ground truth features are most important to the model. Working with representative real-world users, we show how Shared Interest can be used to decide if a model is trustworthy, uncover issues missed in manual analyses, and enable interactive probing.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

human-computer interaction, interpretability, machine learning, saliency methods

ACM Reference Format:
Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3491102.3501965

1 INTRODUCTION

As machine learning continues to be deployed in real-world applications, it is increasingly important to understand the reasoning behind model decisions. A common first step for doing so is to compute the model's *saliency*. In this setting, saliency is the output of any function that, given an input instance (e.g., an image), computes a score representing the importance of each input feature (e.g., pixel) to the model's output. Example saliency methods range from Vanilla Gradients [36], where scores represent the amount a small change in an input feature would have on the model's output, to black-box methods like LIME [34] that use interpretable surrogate models trained to mimic the original model's decision boundary. By analyzing saliencies, users can identify features important to the model's decision and determine how aligned these features are with human decision-making.

While saliency methods provide the much-needed ability to inspect model behavior, making sense of their output can still present analysts with a non-trivial burden. In particular, saliencies are often visualized as solitary heatmaps, which do not provide any additional structure or higher-level visual abstractions to aid analysts in interpretation. As a result, analysts must rely solely on their visual perception and priors to generate hypotheses about model behavior. Similarly, saliency methods operate on individual instances, making it difficult to conduct large-scale analyses of model behavior and uncover recurring patterns. As a result, analysts must choose between time-consuming (often infeasible) manual analysis of all instances or ad hoc (often biased) selection of meaningful subsets of instances.

In response, we introduce Shared Interest: a method for comparing model saliencies with human-generated ground truth annotations. Shared Interest quantifies the alignment between these two components by measuring three types of coverage: Ground Truth Coverage (GTC), or the proportion of ground truth features identified by the saliency method; Saliency Coverage (SC), or the proportion of saliency features that are also ground truth features; and IoU Coverage (IoU), the similarity between the saliency and ground truth feature sets. These coverage metrics enable a richer and more structured interactive analysis process by allowing analysts to sort, rank, and aggregate input instances based on model behavior. The metrics are agnostic to model architecture, input modality, and saliency method, and they can also be composed together (e.g., high SC and low GTC) to identify recurring patterns in the alignment of model and human decision-making.

We demonstrate how Shared Interest enables structured large-scale analysis of model behavior across multiple domains and saliency methods. By applying Shared Interest to computer vision

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3501965

and natural language classification and regression tasks and using a variety of common saliency methods, we identify 8 recurring patterns of interesting model behaviors: HUMAN ALIGNED, SUFFICIENT SUBSET, SUFFICIENT CONTEXT, CONTEXT DEPENDENT, CONFUSER, INSUFFICIENT SUBSET, DISTRACTOR, and CONTEXT CONFUSION. These patterns range from cases where the model's decision and explanation tightly align with human reasoning (HUMAN-ALIGNED) to cases where contextual features are most important to the model's incorrect prediction (DISTRACTOR). Through representative case studies of real-world model interpretability workflows, we explore how Shared Interest helps a dermatologist and a machine learning researcher conduct more systematic analyses of model behavior. Users find that, unlike their prior exploration that was tedious and ad hoc, Shared Interest rapidly surfaces reasons to question a model's reliability, opportunities to learn from the model's representations, and issues missed during previous manual analysis.

We further demonstrate that Shared Interest is not only valuable to understanding a model's predictive performance but can also be used to *query* model behavior. Leveraging the Shared Interest metrics alongside interactive human annotation enables a question-and-answer process where analysts probe input features and Shared Interest identifies the model's decisions whose saliency feature sets are most aligned. In an example human annotation workflow with an image classification task, we show how Shared Interest can reveal insights about the input features most salient to particular predictions and the model's understanding of secondary objects or background features.

Shared Interest is publicly available, with source code at https://github.com/mitvis/shared-interest and live demos at http://shared-interest.csail.mit.edu/.

2 RELATED WORK

Machine learning systems are increasingly designed for high-stakes tasks such as cancer diagnosis, and, as these systems achieve human-caliber or super-human accuracy [13], the temptation to deploy them correspondingly increases. In tandem, a body of work has identified dangerous pitfalls in commonly used models and their underlying training data [5, 6]. To protect against the repercussions of deploying biased or ungeneralizable models, a growing effort focuses on understanding model decisions [11, 32] and characterizing model errors [28]. In this paper, we focus on post hoc saliency methods, also known as feature attribution methods [40], that allow us to observe model reasoning [7, 12, 26, 34, 35, 37, 38, 41].

Saliency methods explain deep learning model decisions on the instance-level. Providing one interpretation at a time may be sufficient to answer questions about model behavior for a small collection of instances. However, it does not scale to answering questions about global model behavior or dataset characteristics. Moreover, the output saliency maps require careful visual assessment to determine if the model used human-salient features to make its decision. Together, these drawbacks often result in the tedious inspection of only a few examples that are cherry-picked or selected ad hoc. By quantifying instances based on the agreement between model and human reasoning, Shared Interest offers a more comprehensive overview of model behavior across all instances and enables systematic evaluation of model behavior.

A recent body of work has questioned whether saliency methods are a reliable instrument for interpreting deep learning models [1, 2, 22, 42, 49]. These papers propose saliency "tests" to measure each method's ability to faithfully represent model behavior. While confirming the fidelity of saliency methods is a critical area of research, it is an orthogonal issue to the focus of our paper as even the most accurate saliency method will still exhibit instance-wise limitations.

Similar to our contribution are Olah et al. [30] and Kim et al. [21] who argue feature-level saliency is not semantically meaningful enough and we should use higher levels of abstraction (e.g., hidden layer representations or concepts) instead. To combat the scalability limitations of instance-wise interpretation, they suggest decomposing activations through matrix factorization [30], activation atlases [8], or concept activation vectors [21]. Shared Interest shares its underlying motivation with this work — a lack of semantically-meaningful structure in saliency methods and supporting scalable interpretability — but offers an alternate way forward. In particular, although we compute attribution back to input features, we do so to compare salient features to human-provided ground truth. In doing so, Shared Interest brings structure and scale to the task of reading model saliencies and more directly expresses the alignment between human and model reasoning.

Aside from saliency methods, a growing number of techniques help users visually interpret models [16, 46]; however, these tools often focus on understanding patterns learned by intermediate nodes [3, 17, 50] or are architecture-specific [18, 20, 39]. In contrast, Shared Interest is agnostic to model architecture, saliency method, and dataset modality, and it can be incorporated into existing model interpretation workflows.

3 THE SHARED INTEREST METHOD

Shared Interest is a method for computing the alignment between model and human decision-making. To do so, we introduce three metrics that measure the relationship between saliency and ground truth annotations. We utilize these metrics to understand model behavior across computer vision (CV) and natural language processing (NLP) tasks.

3.1 Metric Definitions

Mathematically, we use S to represent the set of input features important for a model's decision as determined by a saliency method and G to represent the set of input features important to a human's decision as indicated by a ground truth annotation. For example, in a CV classification task, G might represent the pixels within an object-level bounding box, and S might represent the set of pixels salient to the model's decision as determined by a saliency method. Similarly, in an NLP sentiment classification task, G might be the set of input tokens annotated as indicative of sentiment, and S is the set of tokens determined to be important to the model's prediction.

We compute three metrics: IoU Coverage (IoU), Ground Truth Coverage (GTC), and Saliency Coverage (SC). Each metric takes G

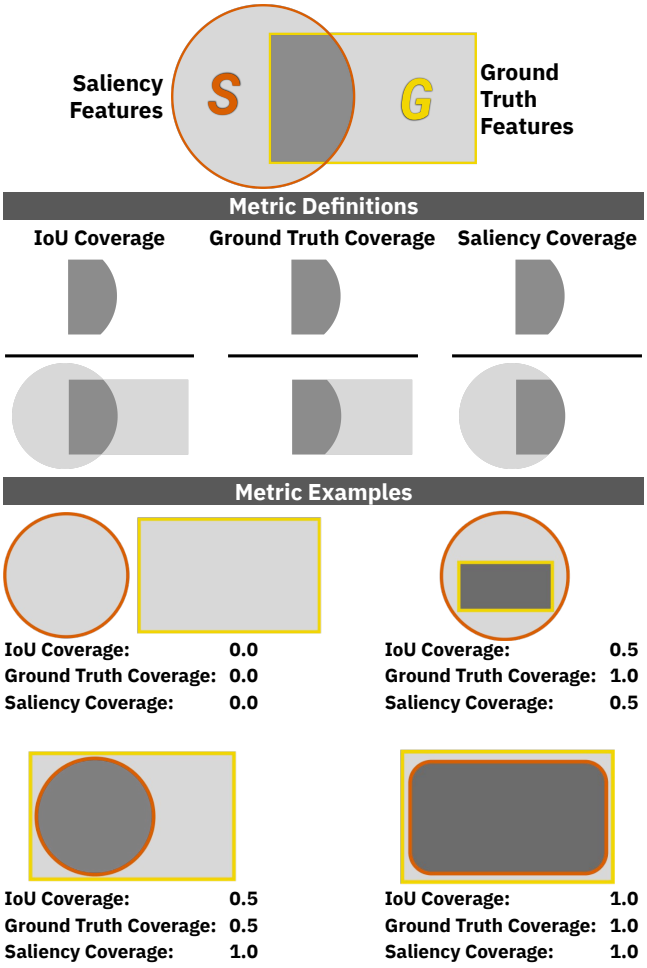

Figure 1: Shared Interest takes a set of saliency features S and a set of ground truth features G and outputs three metrics for identifying instances of interest: IoU, GTC, SC. IoU represents the alignment of model-salient and human-salient features. GTC represents the proportion of human-salient features used by the model. SC represents the proportion of model-salient features used by a human.

and S as inputs and outputs a score between 0 and 1, inclusive.

$$\text{IoU} = \frac{|G \cap S|}{|G \cup S|} \tag{1}$$

$$\text{GTC} = \frac{|G \cap S|}{|G|} \tag{2}$$

$$\text{SC} = \frac{|G \cap S|}{|S|} \tag{3}$$

IoU (Eq. 1) is the strictest metric, and it represents the similarity between the ground truth and saliency feature sets. It is the number of features in both the ground truth and saliency sets divided by the number of features in at least one of the ground truth and saliency sets. In machine learning terms, it is the Jaccard index. GTC (Eq. 2) measures how strictly the model relies on *all* ground truth

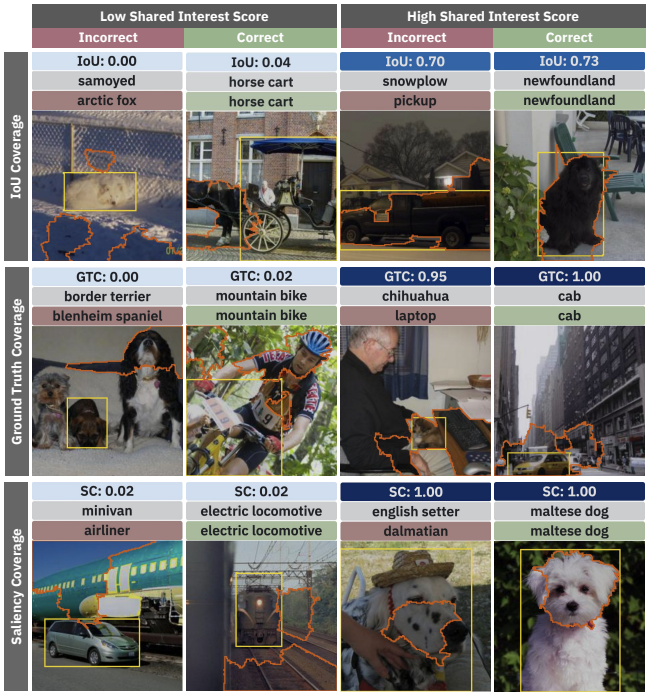

Figure 2: The Shared Interest metrics uncover interesting instances of model behavior. Here, we show an example ImageNet images with high and low scores for each Shared Interest metric. Each image is annotated with its label (grey), prediction (green if correct, red otherwise), ground truth features (yellow), and LIME saliency features (orange). A score of zero under all three metrics indicates the ground truth set (G) and saliency feature set (S) are disjoint. High scores can indicate the model is relying on the ground truth features (IoU), a subset of the ground truth features (SC), or a superset of the ground truth features (GTC).

features — the proportion of the ground truth feature set, G, that is also part of the saliency feature set, S. It is analogous to concepts of recall or sensitivity in machine learning: the proportion of true positives (saliency features that are also ground truth features) successfully identified among all positives (ground truth features). SC (Eq. 3) measures how strictly the model relies on *only* ground truth features — the proportion of the saliency feature set, S, that is also part of the ground truth feature set, G. In machine learning terms, it is analogous to precision: the fraction of true positives (saliency features that are also ground truth features) successfully identified among all detected positives (saliency features).

A score of zero under all three metrics means that an instance's saliency and ground truth feature sets are disjoint, which often indicates that background information is important to the model's prediction. In Figure 2, we show example scenarios using an ImageNet [10] image classification task and LIME [34] saliency maps (see Section 3.2 for details). When a correctly classified instance has a low score, it often indicates there is contextual information in the background that is important to the model, such as the train tracks surrounding the *electric locomotive*. When an instance has a

low score and is incorrectly classified, it can indicate the model is focusing on a secondary object (e.g., the wrong dog) or incorrectly relying on background context (e.g., using snow to predict *arctic fox*).

A high IoU score indicates the explanation and ground truth feature sets are very similar (IoU = 1 $\implies$ $S = G$), meaning the features that are critical to human reasoning are also important to the model's decision. Correctly classified instances with high IoU scores indicate the model was correct in ways that tightly align with human reasoning. Incorrectly classified instances with high IoU scores, on the other hand, are often challenging for the model, such as the image of a snowplowing truck that is labeled as *snowplow* but predicted as *pickup*.

High GTC signals that the ground truth features are the most relevant to the model's decision (GTC = 1 $\implies$ $G \subseteq S$). When a correctly classified instance has high GTC it indicates that the model relies on the object and relevant background features (e.g., the cab and the street) to make a correct prediction. Incorrectly classified instances with high GTC are examples where the model overly relies on local contextual information such as using the keyboard and person's lap to predict *laptop*.

High SC indicates the model relies almost exclusively on ground truth features to make its prediction (SC = 1 $\implies$ $S \subseteq G$). Filtering for correctly classified instances with high SC can surface instances where a subset of the object, such as the dog's face, was important to the model's prediction. Incorrectly classified instances with high SC suggests that an insufficient portion of the object is salient to the prediction (e.g., a small region of black and white spots to predict *dalmatian*).

Shared Interest metrics can also be combined to yield exciting insights. For example, instances with high SC and low GTC indicate the model is focused on a subset of the ground truth region, whereas high GTC and low SC indicate the model is relying on the ground truth and contextual features to make its prediction.

3.2 Experimental Setup

In subsequent sections of this paper, we apply Shared Interest to CV and NLP tasks, including multi-class image classification, binary classification of medical images, and sentiment regression on text reviews. Shared Interest surfaces interesting results across a variety of saliency methods, including gradient-based methods like Vanilla Gradients [36] and Integrated Gradients [41], as well as model-agnostic methods like LIME [34] and SIS [7]. We explore additional saliency methods in Section A. Since S is a discrete feature set, Shared Interest can be straightforwardly applied to saliency methods like SIS [7] that output feature sets. However, to apply Shared Interest to methods that output a continuous score (e.g., Integrated Gradients [41]), we compute S by discretizing these scores. We demonstrate that Shared Interest is robust to discretization procedure by employing score-based and model-based thresholding. Score-based thresholding, used in the CV examples, creates discrete feature sets using only the saliency. For example, we threshold Vanilla Gradients at one standard deviation above the mean to allow for variance in the number and value of salient features across instances, and we select LIME's top n positively contributing features to demonstrate that even naive thresholding can be effective. Model-based thresholding, used in the NLP examples, creates discrete feature sets containing features directly correlated with the model's prediction. In these examples, features positively correlated with the model's prediction are iteratively selected until the model can confidently predict the correct class using only those features. Section B contains additional examples of discretization techniques.

ImageNet Image Classification. In the ImageNet image classification examples (Sections 3.1, 4, and 5.3), we use two subsets of the original ImageNet dataset: the dog and vehicle subsets from ImageNet-9 [48]. Since ImageNet only provides bounding box annotations for a subset of images, we further subset these sets to only contain images with annotations. We use features in the bounding box regions as G. We use a pretrained ResNet50 [15] provided by PyTorch [31] trained on 1000-way classification on ImageNet. In Sections 3.1 and 4, we use LIME [34] explanations as S. To compute LIME, we use the author's implementation (https://github.com/marcotcr/lime) with 1000 samples per image, a Ridge Regression linear model, cosine distance function, and an exponential kernel. We create the saliency feature set using the top 5 features that had a positive impact on the model's prediction, where features are super-pixels defined by QuickShift [45]. In Section 5.3, we compute Vanilla Gradients [12] using Captum [24] for all 1000 ImageNet classes. We take the absolute value of the gradients and discretize by thresholding each saliency map at one standard deviation above the mean.

Melanoma Classification. In the melanoma classification example (Section 5.1), we use lesion images and segmentations from the ISIC 2016 Challenge [14]. Each image is classified as malignant or benign and contains a lesion segmentation that we use to represent G. We trained a ResNet50 [15] model from scratch for 4 epochs using Cross-Entropy loss, Adam [23] optimization, a learning rate of 0.1, a batch size of 128, and class-weighted sampling. Since the test set is not public, we evaluate on the validation set and achieve 0.822 balanced class accuracy. We use LIME [34] explanations as S. To compute LIME, we use the author's implementation (https://github.com/marcotcr/lime) with 1000 samples per image, a Ridge Regression linear model, cosine distance function, and an exponential kernel. We create the saliency feature set using the top 5 features that had a positive impact on the model's prediction, where features are super-pixels defined by QuickShift [45].

BeerAdvocate Sentiment Regression. In the beer review regression examples (Sections 4 and 5.2), we use beer reviews from the BeerAdvocate dataset processed by Lei et al. [25]. Each review is annotated with scores ranging from 0 to 1 in 0.1 increments representing 0 to 5 star reviews in half-star increments. Each review has a score and sentence level annotation for each aspect (aroma, appearance, palette, taste). To directly compare Shared Interest to prior saliency method analysis (Section 5.2), we use the recurrent neural networks (RNNs), SIS rationales, and integrated gradient explanations from Carter et al. [7] available at: https://github.com/b-carter/SufficientInputSubsets. The RNNs were trained on each aspect of the dataset. The SIS procedure selected the sufficient input subsets using an 85% model confidence threshold. For direct comparison, the Integrated Gradients were also iteratively selected from highest to lowest impact on the predicted class

until the model made the original prediction with 85% confidence. To apply the Shared Interest definitions to this regression task, we define correctness as whether the model's output is within a half-star ($\pm\Delta = 0.05$) of the actual value.

4 RECURRING PATTERNS IN MODEL BEHAVIOR

To study how Shared Interest could aid in understanding model behavior, we conducted iterative rounds of qualitative analysis. We applied our metrics to models across a range of domains (computer vision and natural language processing), tasks (regression or classification), saliency methods (gradient-based or model-agnostic), and model architectures (convolutional or recurrent neural networks). Using Shared Interest to sort, explore, and characterize individual instances, several patterns in model decision-making emerged. We validated these patterns and refined their definitions through further iterative analysis (i.e., applying Shared Interest to additional datasets and models and confirming that these patterns continued to surface). Ultimately, we identified eight cases of model behavior defined in terms of Shared Interest metrics and model correctness. In Figure 3, we show an example of each case in a computer vision setting (ImageNet classification with LIME) and a natural language processing setting (BeerAdvocate aroma sentiment prediction [7, 25, 27] with Integrated Gradients [7, 41]).

4.1 Human Aligned

Instances that fall into the HUMAN ALIGNED category are predicted correctly and have high IoU, thus indicating that the model is making a correct prediction and its rationale for that prediction aligns with a human's. For example, in the CV setting of Figure 3a, almost every pixel in the ground truth bounding box was important for the model to make the correct prediction of *trailer truck*. In the NLP example, the saliency method shows that the model uses almost every word related to the beer's aroma to make its correct prediction of 0.9 (strong positive sentiment). HUMAN ALIGNED instances indicate cases when the model is faithful to human decisions, and ideally, all instances would fall into this category.

4.2 Sufficient Subset

The SUFFICIENT SUBSET category contains instances with high SC and low GTC, revealing where a subset of the human-annotated features are important for the model to make a correct prediction. For example, in Fig. 3(b), the human-annotated ground truth in the CV example covers the entire tractor. However, the saliency method indicates that the tractor's tire was most important for the model to make its correct prediction. In the NLP case, the salient regions indicate the model considers the words "complex", "aroma", "chocolate", and "vanilla" important to predict strong positive sentiment. The SUFFICIENT SUBSET category may indicate that the human reasoning annotation includes extraneous information (e.g., the stop words in the aroma review) or that the model relies on an adequate but incomplete set of features that may not generalize (e.g., only the tractor tire).

4.3 Sufficient Context

The SUFFICIENT CONTEXT category includes correctly predicted instances with low IoU, indicating there is information in non-ground truth features correlated with the correct prediction. Analyzing instances in this category can validate if contextual features are indeed meaningful and not spurious correlation. The CV example in Figure 3c shows the model uses the helmet to predict *snowmobile*. While the presence of a snowmobile helmet correlates with the existence of a snowmobile, this model would not be robust to real-life scenarios. This result might inspire further exploration of instances with snowmobiles and snowmobile helmets to confirm there is not a rigid dependence between the two objects. In the NLP example, the saliency method indicates that the word "zilch" is important to the model's decision. However, "zilch" corresponds to the taste aspect instead of aroma. This discrepancy suggests a correlation between reviewer scores for aroma and other aspects and that the model may overfit to features related to other aspects. Instances in these cases may be of interest to an analyst to identify correlated features that exist in the training data but would not generalize to the real world.

4.4 Context Dependent

The CONTEXT DEPENDENT category identifies correctly classified instances with high GTC and low SC, meaning the model relies on ground truth and contextual features to make a correct prediction. In Figure 3d, the CV example shows the model relies not only on the streetcar (the labeled class) but also on the train tracks to predict *streetcar*. In the NLP example, the model uses the ground truth words "rich" and "smells" along with words like "nutty", "cocoa", and "almond" to make its positive sentiment prediction. While the context is semantically correlated with the ground truth in both examples, these instances may indicate nongeneralizable correlations and require further exploration to uncover whether it is reasonable for the model to use context in its prediction.

4.5 Confuser

Confusers are instances where the model relies on human-salient features but still makes an incorrect prediction. In Shared Interest terms, members of the CONFUSER case are incorrectly classified instances with high IoU. In the CV example in Figure 3e, the CONFUSER case identifies an ambiguous label — the image is labeled as *moped*, but the model predicts *motor scooter* — a known problem with ImageNet [4, 44]. Using Shared Interest, instances of this failure case immediately rise to the forefront of the analysis process via the CONFUSER class. In the NLP example, the saliency method finds the model relies on almost all the ground truth words to make a strong positive sentiment prediction. While the ground truth sentence has a positive sentiment, the reviewer only gave the aroma a 0.6 (weakly positive). In both domains, the CONFUSER case helps immediately identify instances with imprecise dataset labels. Discovering such instances might encourage an analyst to conduct further exploratory analysis on the dataset or perform additional preprocessing to resolve ambiguities.

Figure 3: Using Shared Interest we identify eight recurring patterns in model behavior across two domains and saliency methods: an ImageNet [10] classification model with LIME [34] and a BeerAdvocate sentiment regression model [7, 25, 27] with Integrated Gradients [7, 41]. Each example includes the IoU, GTC, and SC scores, true label (grey), and model prediction (green if correct, red otherwise). Ground truth features are highlighted in yellow and saliency features are highlighted in orange.

4.6 Insufficient Subset

Insufficient subset identifies incorrectly classified instances with high SC and low GTC, meaning a subset of the ground truth features is important to the model, but it makes an incorrect prediction. In Fig. 3(f), with the CV example, the model predicts the dog breed *whippet* on an image of two whippets in a shopping cart. The image is labeled as *shopping cart*, and the saliency method indicates the model relied on the faces of the dogs as opposed to the pixels of the shopping cart. In the NLP example, the saliency method indicates the model relies upon the words "vague" and "odor" in the aroma sentence to predict negative sentiment (0.3). However, other words in the sentence not indicated as salient to the model do contain positive sentiment (e.g., "caramel malt" and "noble hops") and likely contributed to the actual label of 0.6. In general, insufficient subset cases can signal to an analyst that the model is overly reliant on a small set of features and warrants further exploration.

4.7 Distractor

Distractors are cases where the model does not rely on ground truth features (low IoU) and makes an incorrect prediction. In

Fig. 3(g), the CV example shows an instance labeled *moped*, but the model predicts *church* as the saliency covers pixels related to the church in the background. In this case, the image contains multiple objects but only has a single label, which is a known flaw of ImageNet [4, 44]. In the NLP example, the saliency method indicates the model relies on the words "metal", "corn", and "nothing awesome" to predict negative sentiment. While the known aroma words ("the smell is sweet malty lagery") have positive sentiment, the model is distracted by negative sentiment words elsewhere in the review. DISTRACTOR instances may indicate that the model is overfitting to the overall sentiment of the review rather than the specific sentiment associated with the aroma.

4.8 Context Confusion

The CONTEXT CONFUSION case contains instances where the model is using ground truth features but is confused by other features and, thus, makes an incorrect prediction. In Shared Interest terms, these instances have high GTC and low SC. For example, in the CV setting in Figure 3h, the saliency indicates the presence of the field next to the trailer truck is important for the model to predict *harvester*. In the NLP example, the model relies on words in the aroma sentence as well those surrounding the sentence to predict strong positive sentiment (0.9) as opposed to weakly positive sentiment (0.6). In this instance, many of the surrounding words contain positive sentiment (e.g., "impressive" and "extremely"), which may have caused the model to predict more positively than the actual class.

5 INTERACTIVE INTERPRETABILITY WORKFLOWS

We demonstrate how Shared Interest can be used for real-world analysis through case studies of three interactive interpretability workflows of deep learning models. The first case study follows a domain expert (a dermatologist) using Shared Interest to determine the trustworthiness of a melanoma prediction model. The second case study follows a machine learning expert analyzing the faithfulness of their model and saliency method. The final case study examines how Shared Interest can analyze model behavior even without pre-existing ground truth annotations.

We developed visual prototypes for each case study to make the Shared Interest method explorable and accessible to all users, regardless of machine learning background. The computer vision and natural language processing prototypes (Figure 4 and Figure 5) focus on sorting and ranking input instances so users can examine model behavior. Each input instance (image or review) is annotated with its ground truth features (highlighted in yellow) and its saliency features (highlighted in orange) and is shown alongside its Shared Interest scores, label, and prediction. The interface enables sorting and filtering based on Shared Interest score, Shared Interest case, label, and prediction. The human annotation interface (Figure 6) is designed for interactive probing. The interface enables users to select and annotate an image with a ground truth region and returns the top classes with the highest Shared Interest scores for that ground truth region. Code for the prototypes is available at https://github.com/mitvis/shared-interest, and live demos of each prototype are available at http://shared-interest.csail.mit.edu/.

5.1 Model Analysis by a Domain Expert

Our first case study follows the use case of a domain expert, a dermatologist, who wishes to evaluate the trustworthiness of a machine learning model that could assist them in diagnosing melanoma. Accurate and early melanoma diagnosis is a critical task that can significantly impact patient outcomes, and machine learning could assist dermatologists in making more accurate decisions. In order to do so, however, our participant noted it would be imperative for dermatologists to be able to evaluate how the model operates personally.

We evaluate Shared Interest in this context to understand how its ability to convey model behavior may help a domain expert determine whether or not they should trust a model. To do so, we applied Shared Interest to a Melanoma Classification task (see Section 3.2 for details). We used a Convolutional Neural Network trained on the ISIC Melanoma dataset [9, 43] to classify images of lesions as either *malignant* (cancerous) or *benign*. We used lesion segmentations from the dataset as the ground truth feature sets and the output of LIME [34] towards the predicted class as the saliency feature sets. Using a prototype visual interface (Figure 4) designed to enable interactive analysis of Shared Interest, we explored examples with the dermatologist for 30 minutes. We used the Shared Interest cases to outline the conversation by showing the dermatologist examples from each case. However, the dermatologist guided the analysis by discussing the insights that excited them and suggesting what to investigate next. Throughout the conversation, we asked the dermatologist open-ended questions (e.g., *"How do you feel about the model after seeing these examples?"*) to understand how they would evaluate a model and how Shared Interest could aid in evaluation.

Using the HUMAN ALIGNED, CONTEXT DEPENDENT, and SUFFICIENT SUBSET categories, the dermatologist surfaced insight into cases where the model was trustworthy. Analyzing *malignant* lesions in the HUMAN ALIGNED case surfaced examples where the model correctly classified cancerous lesions by relying on features of the lesion. The dermatologist agreed with the model on these images and began to build trust with the model, noting *"obviously it does a pretty good job."* CONTEXT DEPENDENT images identified cases where the model relied not only on the lesion but also on surrounding skin. While there was potential for the dermatologist to distrust the model, they actually found these instances especially interesting because cancerous cells can lie beyond the pigmented lesion boundary. Thus, the dermatologist wondered if *"there are really subtle changes that we are not picking up that [the model] is able to."* Images in the SUFFICIENT SUBSET case showed cases where the model only relied on a subset of the lesion. While the dermatologist agreed with the model, they expressed some concern that it was not using the complete lesion, especially when there were meaningful cancerous features in the unused regions.

Shared Interest was also able to quickly reveal cases where the model was not trustworthy. The SUFFICIENT CONTEXT and DISTRACTOR cases showed images where the model relied on contextual features such as peripheral skin regions or the presence of artifacts (see Figure 4). While the dermatologist was tolerant to a few instances where the model relied on non-salient features, seeing the

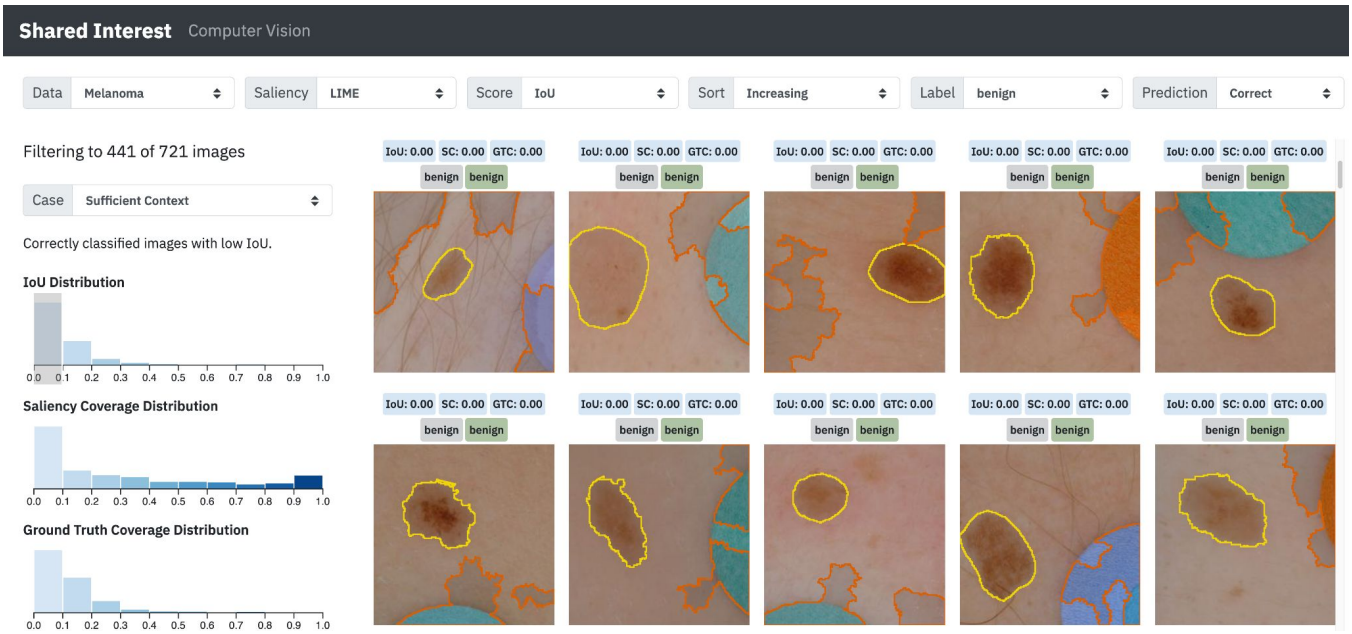

Figure 4: Shared Interest can help domain experts decide the degree to which they trust a particular model. We use Shared Interest with a dermatologist to analyze a model trained to predict melanoma. The computer vision prototype displays lesion images with segmentations (yellow), LIME explanations (orange), actual (grey) and predicted classifications (green if correct, red otherwise), and all three Shared Interest scores. It enables efficient visual analysis, even by non-expert users, by filtering and sorting based on score, case, label, and prediction. The SUFFICIENT CONTEXT case, shown here, surfaces images where the model has latched onto artifacts to make a *benign* prediction. Since these artifacts only occur in *benign* dataset images, they are sufficient to make a prediction; however, this model would not generalize to clinical cases where the artifacts may also occur in *malignant* images. This demo is at: http://shared-interest.csail.mit.edu/computer-vision/

number of images in these cases led the dermatologist to distrust the model in all cases, stating *"I would discard the model."*

By classifying inputs into cases where the model was or, more importantly, was not aligned with human reasoning, Shared Interest enabled the dermatologist to rapidly and confidently decide whether or not to trust the model. If the dermatologist had evaluated the model by randomly selecting images, they might not have identified that the model repeatedly made decisions based on background information, and they would not have known how frequently that case occurred. As the dermatologist said, Shared Interest is *"helpful [as a way to] see how the computer is thinking and allow me to understand if I should trust it."*

5.2 Saliency Method Analysis by a Machine Learning Researcher

Our second case study is representative of use cases where a machine learning expert wants to analyze a model or saliency method they are developing. To evaluate Shared Interest's value in the development pipeline, we worked with an author of the Sufficient Input Subset (SIS) interpretability method [7] whose goal is to understand how well SIS explains model decisions. During development of the SIS method, one of the ways the researchers analyzed the method was by applying it to the BeerAdvocate dataset and comparing the SIS saliencies (called "rationales" by the researchers) to the ground truth annotations. This process enabled them to evaluate whether the rationales *"fell within the ground truth"* and represented a *"compact set"* of meaningful features.

To recreate the researcher's original workflow, we applied Shared Interest to the BeerAdvocate reviews annotated on the appearance aspect, trained Recurrent Neural Networks, and SIS rationales from Carter et al. [7] (see Section 3.2 for details). We populated the visual prototype with the results (Figure 5) and used it to explore the Shared Interest cases with the researcher. Throughout the 45 minute conversation, we asked the researcher open-ended questions to understand how they evaluate a saliency method and how Shared Interest might aid in their evaluation. While we used the Shared Interest cases as a guide, the researcher led the analysis, and insights from the cases often inspired them to examine additional settings.

Using Shared Interest, the researcher surfaced numerous insights that inspired confidence in the SIS algorithm. For example, the researcher immediately identified that most reviews have high SC, indicating most of the SIS rationales were contained almost entirely within the ground truth. Since *"ideally, the model is learning the right set of features and thus the rationales live within the correct set of features"*, the researcher found the distribution of scores indicative that the SIS procedure was capturing meaningful information. In their original analysis, the researchers had even computed a metric

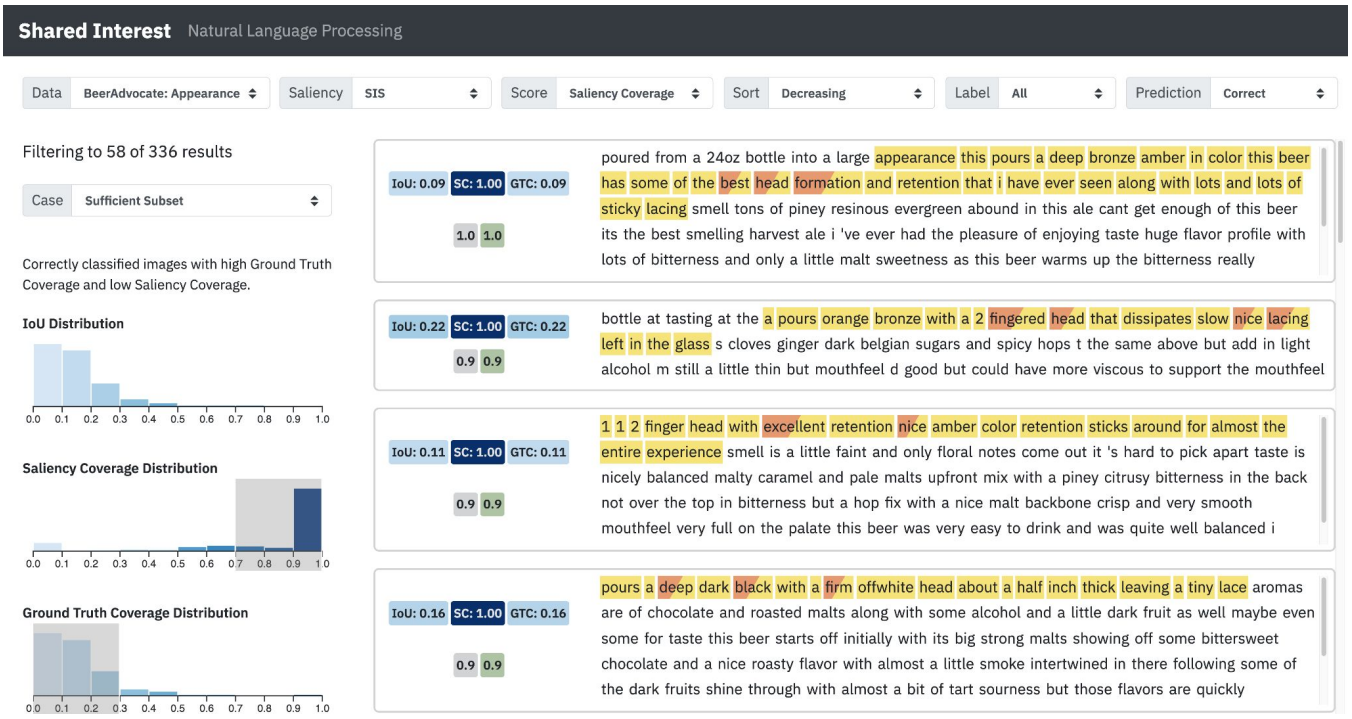

Figure 5: Shared Interest can assist machine learning experts by enabling efficient large-scale analyses of their methods. We evaluate Shared Interest with a saliency method researcher by assessing the method's performance on a model trained to predict sentiment on text reviews. The NLP prototype displays reviews from the BeerAdvocate dataset with ground truth features (yellow) and SIS saliency features (orange) highlighted. Each review is annotated with its Shared Interest scores, label (grey), and prediction (green if correct, red otherwise). The SUFFICIENT SUBSET case, shown here, identifies reviews where the saliency method indicates the model relied on meaningful features, such as "best head formation", and where it overfits to general positive sentiment words such as "excellent" and "nice". This demo is at: http://shared-interest.csail.mit.edu/nlp/

equivalent to SC as a quantitative way to analyze their method. So, seeing the same metric populated by Shared Interest validated the use of Shared Interest and the SUFFICIENT SUBSET and INSUFFICIENT SUBSET categories. The researcher found it reassuring to find HUMAN ALIGNED and SUFFICIENT SUBSET instances that matched their expectations, such as rationales that contained appearance-specific words (e.g., "red", "copper", and "head") but did not contain uninformative ground truth words like stop words. The SUFFICIENT SUBSET category was significant to the researcher since it aligned with SIS's goal to find minimal rationales. Seeing all of these examples at once helped the researcher identify cases where the rationale was indeed a meaningful sufficient subset of words such as "lovely looking".

Shared Interest also helped the researcher uncover previously unknown pitfalls in the model and the data. Looking at instances in the SUFFICIENT SUBSET category, the researcher identified common cases of model overfitting, such as a correct prediction using only the word "beautiful". As the researcher put it, *"These are positive words, so it makes sense they are correlated with positive appearance, but I don't think they should be sufficient for separating [the appearance] aspect from others."* Looking at reviews in the INSUFFICIENT CONTEXT case exposed instances where the model was again overfitting to positive sentiment. However, now the researcher was even more concerned since it caused an incorrect prediction. Although the researcher had *"previously observed that the model had associated single tokens that were general positive sentiment words with predicting high sentiment"*, they did not *"as quickly notice particular words like 'beautiful' that were immediately surfaced [via Shared Interest]."* Finally, looking at SUFFICIENT CONTEXT reviews, where SIS rationales are disjoint from the ground truth annotation, the researcher uncovered reviews with incomplete or incorrect annotations. Until using Shared Interest, the researcher had previously never identified an incorrectly annotated review, saying *"In the past, I did not note any cases where I thought the annotations might have been incomplete. I think that's a pretty interesting insight."*

Overall, the researcher found that grouping and aggregating via Shared Interest helped them *"see all of the [reviews] grouped together by the various cases"* which categorized and *"clearly described what the various patterns are"*. In the researchers' original analysis, they had *"skimmed through a big file of [reviews] not sorted in any way"*, and, while they *"were noting patterns, it was harder to keep track of these different cases."* If they would have had access to Shared Interest at the time of their original analysis, this researcher thought it would have *"more quickly exposed some of the patterns and behaviors that we identified and also led to additional discoveries."*

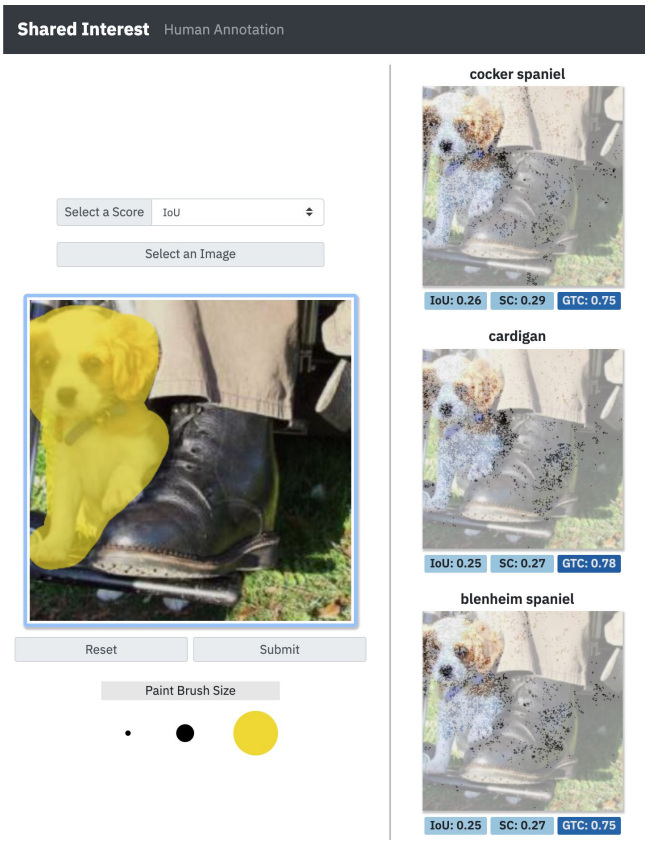

Figure 6: Shared Interest permits visual workflows where users can interactively probe and explore model behavior. We use the prototype to evaluate an ImageNet classification model. By comparing the user's annotation (shown in yellow) to the saliency feature set (shown via saturation) for every ImageNet class, Shared Interest surfaces the classes most related to the annotated features. In this example, the model relates the dog breed classes *cocker spaniel*, *cardigan*, and *blenheim spaniel* to the features of the highlighted dog. This demo is at: http://shared-interest.csail.mit.edu/human-annotation/

5.3 Interactive Probing of Model Behavior

For our final case study, we demonstrate a workflow where Shared Interest can be used as a mechanism to *query* model behavior. For any input instance, rather than computing the saliency for only the predicted class, we do so for all possible classes. Moreover, users can interactively annotate the instance to designate a "ground truth" region of interest instead of relying on a single pre-existing ground truth. By calculating all three Shared Interest metrics between these two sets of features and returning classes with the highest Shared Interest scores, we enable users to engage in a style of "what if" reasoning. Users can interactively probe the model to understand what input features are important to trigger a particular prediction.

In Figure 7, we show an example of this style of "what if" analysis on an ImageNet classification task (see Section 3.2) for details). By interactively re-specifying the "ground truth" on a single image, we repeatedly probe the model and surface insights about its behavior. Since the model was trained to predict the *otterhound* in the image, we can use Shared Interest to validate that the model has indeed learned the salient features of the dog. By selecting the pixels associated with the dog's face and body (Figure 7a), we find that, although none of the top three returned classes are *otterhound*, they are all dog breeds, and the salient feature sets are focused primarily on the dog. This result may suggest that the model has learned generalizable features associated with dogs — a positive characteristic if we plan to deploy this model.

Since the model learned to associate the entire dog region with dog classes, this prompts a follow-up question: how much of the dog do we have to annotate before the model no longer associates it with dog breed classes? Brushing over just the dog's head (Figure 7b) or even just the dog's snout (Figure 7c) still returns dog breeds as the top classes. This result suggests the model has learned to correlate even small characteristic features (e.g., black noses) with dogs.

This style of analysis also enables us to ask questions about other objects in the image. Although the model was trained to classify this image as *otterhound*, it was also trained to classify 1,000 ImageNet objects. Thus, the model may know salient information about other objects in the image as well. In Figure 7d we validate this claim by brushing the person's hat and observing the top returned classes are types of hats: *sombrero*, *cowboy hat*, and *bonnet*. Similarly, we select the person's hand (Figure 7e) and, as *hand* is not a class our model was trained to detect, observe classes associated with hands such as *cleaver*, *notebook*, and *space bar*. This result is intriguing because a hand is often, but not always, present in images of these objects. Thus, further analysis is warranted to determine if the model is overly-reliant on the presence of hands to make predictions for these classes.

We can also probe the model to see if it has learned anything about image backgrounds or textures, despite only being trained on foreground objects. In Figure 7f, we select a region of the stone wall. Interestingly, the model returns classes associated with rocks such as *cliff*, suggesting that training on images with foreground labels may still impart information to the model about background scenes.

As we have seen, Shared Interest allows us to probe model behavior in new ways, enabling exploration into what the model has learned and where it might fail. Users can identify subsets of features important to classification, explore how well a model can identify secondary objects, and even study the extent to which a model has learned about objects it has never classified. Using this procedure can help a user test hypotheses about what the model has learned and identify information that could help them improve model behavior.

6 DISCUSSION

This paper presents Shared Interest, a method for large-scale analysis of machine learning model behavior via metrics that quantify instances based on the model's alignment with human reasoning.

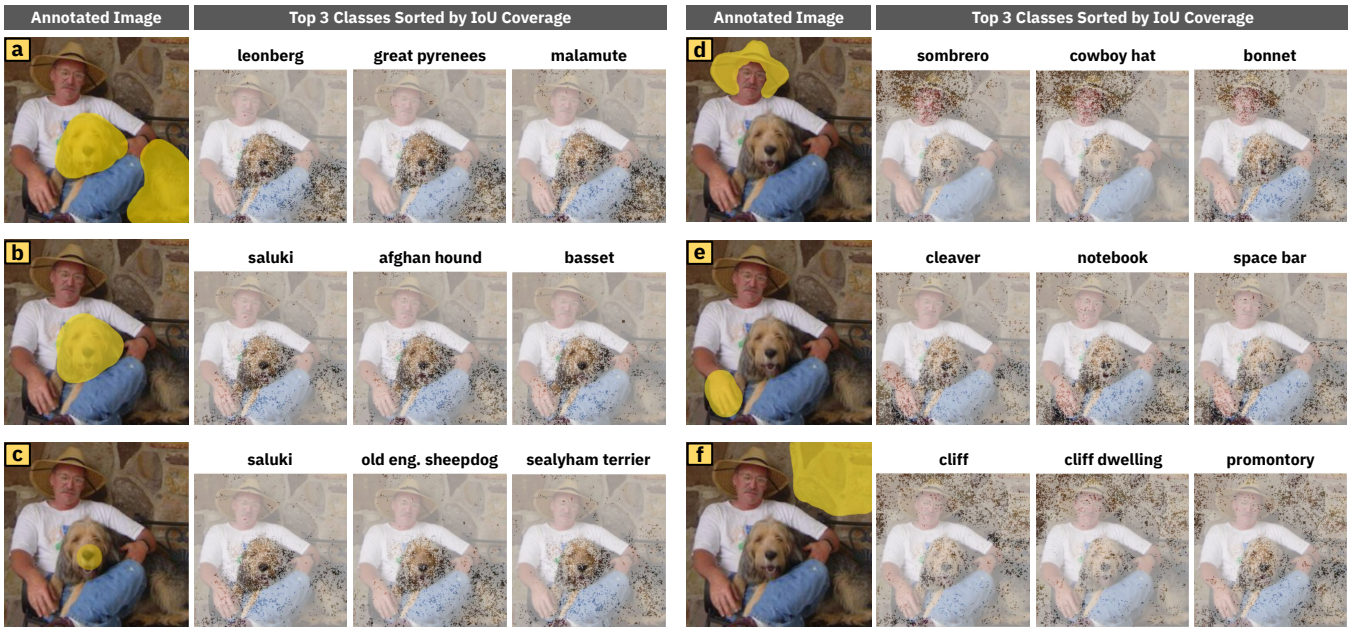

Figure 7: Shared Interest enables users to probe the model with different ground truth features to understand model behavior. By comparing the user's annotation (shown in yellow) to the saliency feature set (shown via saturation) for every ImageNet class, Shared Interest surfaces the classes most related to the annotated features. Probing with smaller and smaller sets of features (a-c) shows the model has learned characteristic features of dogs. Probing with secondary objects (d) demonstrates that the model learns about objects other than the labeled object in the image. Probing the model with features it has not learned to classify (e) indicates the model learns to relate these features to associated objects (e.g., hand and cleaver). Finally, probing with background features (f) demonstrates that the model has learned related features despite only being trained on foreground classification.

Shared Interest enables instances to be sorted, ranked, and aggregated based on this alignment. Using Shared Interest, we identified eight patterns in model behavior that recur across multiple domains (computer vision and natural language processing), model architectures (convolutional and recurrent neural networks), and saliency methods (gradient-based and model-agnostic). These patterns range from cases where the ground truth features are important to the model's incorrect prediction (CONFUSER) to cases where the ground truth features are not important to the model's correct classification (SUFFICIENT CONTEXT). We evaluate Shared Interest's usefulness through representative case studies of real-world interactive visual analysis workflows. Working with a dermatologist and a machine learning researcher revealed that although analysts want to explore model behavior, they find current methods are tedious to use and require too much ad hoc inspection to feel entirely confident in the results. In contrast, with Shared Interest, both types of users could systematically explore model behavior to identify reasons to question the model's reliability and validate novel saliency methods. In a final case study, we demonstrate that Shared Interest is not restricted to merely understanding a model's predictive performance but can also support interactive "what if" analysis to determine the input features most important to particular predictions.

6.1 Limitations

While the Shared Interest methodology can help users efficiently and comprehensively understand model behavior, it requires data paired with ground truth annotations. Research datasets, such as those used in this paper may include such annotations, but real-world data rarely do due to the time and effort required in the collection process. While this issue can limit Shared Interest's applicability, we believe that understanding model behavior is critical enough to warrant the collection of human annotations. Collection may range from annotating a few instances (e.g., via the probing interface) for general research analysis to annotating entire datasets when deciding to deploy a model on a critical task.

Additionally, existing ground truth annotations often highlight the features associated with the label, such as the pixels corresponding to the dog in an image. However, human decision-making may not perfectly align with those features. A human may only need to look at a subset of the features like the dog's face to know the image contains a dog. Alternatively, a human may need additional features like a hockey player or ice rink to know that a black round object is a hockey puck. Thus, as more work focuses on understanding human decision-making and annotating datasets in a corresponding rich fashion, Shared Interest metrics will more precisely communicate human-AI alignment.

Finally, throughout this paper, we rely on saliency methods as proxies for model reasoning. However, researchers have demonstrated cases where saliency methods do not accurately reflect the model [1, 22]. While saliency methods are valuable tools that can give insight into model behavior, Shared Interest can inherit their limitations. For instance, if a saliency method returns an explanation that does not accurately reflect the model's decision-making, Shared Interest will not be able to quantify human-AI alignment accurately. Nonetheless, we designed Shared Interest to be agnostic to the saliency method; so, as methods evolve, Shared Interest's ability to communicate model-human alignment will also improve.

6.2 Future Work

Shared Interest opens the door to several promising directions for future work. One straightforward path is applying Shared Interest to tabular data — a standard format used to train models, particularly in healthcare applications. Tabular data is often more semantically complex than image or text data and thus allows us to bring further nuance to the recurring behavior patterns we have identified in this paper. For instance, fields in tabular data may correlate in more specific and fine-grained ways (e.g., as proxy variables [29]) than the foreground/background context we have distinguished in this paper. As tabular data for these uses cases often contains sensitive personal information (e.g., health data), one could imagine using a version of our interactive probing prototype to systematically analyze how a particular model may perpetuate or amplify bias in the data.

Another avenue for future work is using Shared Interest to compare the fidelity of different saliency methods. Previous work has conducted experiments to determine how faithful saliency methods are to an underlying model [1] based on cascading randomization of internal model layers and its effect, or lack thereof, on the resulting saliency map. These prior studies compute quantitative metrics to identify divergences between perturbations to the model and the effect on the saliency. However, these metrics operate only over individual pixels. By rerunning these studies and quantifying results in terms of Shared Interest metrics, we can increase the level of abstraction of the results. For instance, rather than defining input invariance [22] over individual pixels, we could define it over GTC, SC, or IoU and distinguish whether saliency map sensitivity represents a semantically meaningful signal.

Finally, an exciting direction might consider Shared Interest during model training. Currently, model developers have limited introspection into this process. Typically, they rely on curves that visualize a model's loss function or performance on the training and validation sets. While visual analytics research has helped, most existing work has focused on depicting model architecture or performance [19, 33, 47]. Shared Interest, however, allows us to evaluate training in terms of the model's reasoning. By comparing how saliency features change over epochs, Shared Interest could identify the instances a model gets right immediately and the instances that take several more updates to classify correctly. These insights could inform future training procedures or augment the dataset with more informative examples.

ACKNOWLEDGMENTS

This work is supported by a grant from the MIT-IBM Watson AI Lab. Research was also sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Canada, 9525–9536.

[2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging Tests for Model Explanations. *arXiv preprint arXiv:2011.05429* (2020).

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, USA, 3319–3327.

[4] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with ImageNet? arXiv:2006.07159 [cs.CV]

[5] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Waikoloa, USA, 1536–1546.

[6] Brandon Carter, Siddhartha Jain, Jonas Mueller, and David Gifford. 2021. Overinterpretation reveals image classification model pathologies. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Virtual Event.

[7] Brandon Carter, Jonas Mueller, Siddhartha Jain, and David K. Gifford. 2019. What made you do this? Understanding black-box decisions with sufficient input subsets. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, Naha, Japan, 567–576.

[8] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. 2019. Activation Atlas. *Distill* (2019). https://doi.org/10.23915/distill.00015 https://distill.pub/2019/activation-atlas.

[9] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1902.03368 [cs.CV]

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Miami, USA, 248–255.

[11] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML]

[12] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing Higher-Layer Features of a Deep Network. *Technical Report, Univeristé de Montréal* (2009).

[13] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118.

[14] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. 2016. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1605.01397 [cs.CV]

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, USA, 770–778.

[16] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2019. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* 25, 8 (2019), 2674–2693.

[17] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng (Polo) Chau. 2020. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 1096–1106.

[18] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer

Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*. ACL, Virtual Event, 187–196.

[19] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng (Polo) Chau. 2018. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 88–97.

[20] Minsuk Kahng, Nikhil Thorat, Duen Horng (Polo) Chau, Fernanda B. Viégas, and Martin Wattenberg. 2019. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 310–320.

[21] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, Stockholm, Sweden, 2668–2677.

[22] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (Un)reliability of Saliency Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700. Springer, 267–280.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. San Diego, USA.

[24] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896 [cs.LG]

[25] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Austin, USA, 107–117.

[26] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*. Long Beach, USA, 4765–4774.

[27] Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning Attitudes and Attributes from Multi-aspect Reviews. In *Proceedings of the International Conference on Data Mining (ICDM)*. IEEE, Brussels, Belgium, 1020–1025.

[28] Christopher Meek. 2016. A Characterization of Prediction Errors. arXiv:1611.05955 [cs.LG]

[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 115:1–115:35.

[30] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. *Distill* (2018). https://doi.org/10.23915/distill.00010 https://distill.pub/2018/building-blocks.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 8024–8035.

[32] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (2020), 137–141.

[33] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D. Williams. 2017. Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 61–70.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, San Francisco, USA, 1135–1144.

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 618–626.

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*. Banff, Canada.

[37] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. arXiv:1706.03825 [cs.LG]

[38] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*, Yoshua Bengio and Yann LeCun (Eds.). San Diego, USA.

[39] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. 2018. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 667–676.

[40] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. 2020. Visualizing the Impact of Feature Attribution Baselines. *Distill* (2020). https://doi.org/10.23915/distill.00022 https://distill.pub/2020/attribution-baselines.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, Sydney, Australia, 3319–3328.

[42] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun D. Preece. 2020. Sanity Checks for Saliency Metrics. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, New York, USA, 6021–6029.

[43] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5, 1 (2018).

[44] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 119. PMLR, Virtual Event, 9625–9635.

[45] Andrea Vedaldi and Stefano Soatto. 2008. Quick Shift and Kernel Methods for Mode Seeking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman (Eds.), Vol. 5305. Springer, Marseille, France, 705–718.

[46] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda B. Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65.

[47] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mané, Doug Fritz, Dilip Krishnan, Fernanda B. Viégas, and Martin Wattenberg. 2018. Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 1–12.

[48] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2021. Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, Virtual Event.

[49] Mengjiao Yang and Been Kim. 2019. Benchmarking Attribution Methods with Relative Feature Importance. arXiv:1907.09701 [cs.LG]

[50] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 8689. Springer, Zurich, Switzerland, 818–833.

A ADDITIONAL SALIENCY METHODS

The Shared Interest metrics represent model decision-making using a saliency feature set computed via a saliency method. Throughout the paper, we show that Shared Interest can be used with a variety of saliency methods, including Vanilla Gradients [36], Integrated Gradients [41], LIME [34], and SIS [7].

Here we explore additional saliency methods: SmoothGrad [37], Guided Backpropagation [38], Gradient SHAP [26], and Grad-Cam [35]. For each saliency method we show examples of the Shared Interest cases (Figure A) and the distribution of Shared Interest scores (Figure C) across vehicle images from ImageNet-9 [44]. Implementations used to compute the saliency methods are available at https://github.com/mitvis/shared-interest. The SmoothGrad implementation is based on the Google PAIR implementation: https://github.com/PAIR-code/saliency, and all other methods are implemented using Captum [24]. In each example, we compute attribution with respect to the predicted class.

We find that the Shared Interest cases occur regardless of saliency method, indicating Shared Interest can be successfully used with various saliency methods. While each saliency method highlights the features important to the model's decision, each method computes importance differently, and the outputted saliency features vary across saliency methods. For example, Vanilla Gradients represents importance as the impact a slight change in each feature would have on the model's output, and it often results in sparse and noisy feature subsets. On the other hand, GradCAM computes the gradients with respect to the last convolutional layer, which results in continuous feature regions. The Shared Interest scores are often higher for continuous feature regions than sparse feature sets (see Figure C). Thus, Shared Interest scores should only be compared within a single method because each method's score distribution will vary slightly due to variations in the methods. Further, the high and low values selected for each Shared Interest case will also vary based on the saliency method.

B ADDITIONAL DISCRETIZATION TECHNIQUES

To utilize Shared Interest's set-based metrics, the saliencies from methods that output continuous-valued scores must be discretized. In this paper, we show that Shared Interest surfaces insights into model behavior using a variety of discretization techniques, including score-based thresholding and model-based thresholding (see Section 3.2). In Figure B we show additional score-based discretization techniques, and in Figure C, we show the Shared Interest score distributions for each technique. These techniques threshold values based on the saliency scores. Some methods take features whose scores are above a particular value (i.e., mean, one standard deviation above the mean, and two standard deviations above the mean). Other methods take the top n features, such as the top 5% - 75% of features or the same number of saliency features as ground truth features.

Many discretization techniques can successfully be used in Shared Interest, but each technique makes its own assumptions about the model and saliency method. The Shared Interest metrics depend on the thresholding technique, and the distribution of Shared Interest scores will vary based on the threshold. For example, assuming ground truth features are also the most salient features, stricter thresholds that result in fewer saliency features will increase SC and decrease GTC and IoU. Thus, Shared Interest scores should only be compared within a single discretization technique. The "high" and "low" Shared Interest values used to compute the Shared Interest cases also depend on the discretization procedure and should be chosen based on the score distribution. Finally, size-based thresholds (e.g., features with the top 25% of saliency values) artificially determine the number of saliency features and cause Shared Interest metrics to vary across instances depending on the number of ground truth features. For example, if the method results in a small saliency feature set, instances with large ground truth feature sets will have low IoU and GTC scores. The metrics will be comparable as long as the ground truth feature sets are similar in size across instances (e.g., normalized medical images). However, if the number of ground truth features varies significantly across the dataset, a thresholding technique based on the saliency value or the model's behavior is preferred.

Figure A: Shared Interest is agnostic to saliency method, and the Shared Interest cases occur across a variety of saliency methods. Here we show an example of each of the eight Shared Interest cases — HUMAN ALIGNED (HA), SUFFICIENT SUBSET (SS), SUFFICIENT CONTEXT (SC), CONTEXT DEPENDENT (CD), CONFUSER (C), INSUFFICIENT SUBSET (IS), DISTRACTOR (D), and CONTEXT CONFUSION (CC) — across saliency methods — Vanilla Gradients (VG) [36], SmoothGrad (SG) [37], Guided Backpropagation (GBP) [38], Integrated Gradients (IG) [41], Gradient SHAP (SHAP) [26], and Grad-CAM [35]. Images are vehicle images in ImageNet-9 [10] and the saliency methods are thresholded at one standard deviation above the mean. Each image is annoated with the label (grey), prediction (green if correct, red otherwise), and Shared Interest scores. The ground truth features are shown in yellow and the saliency features are shown via saturation.

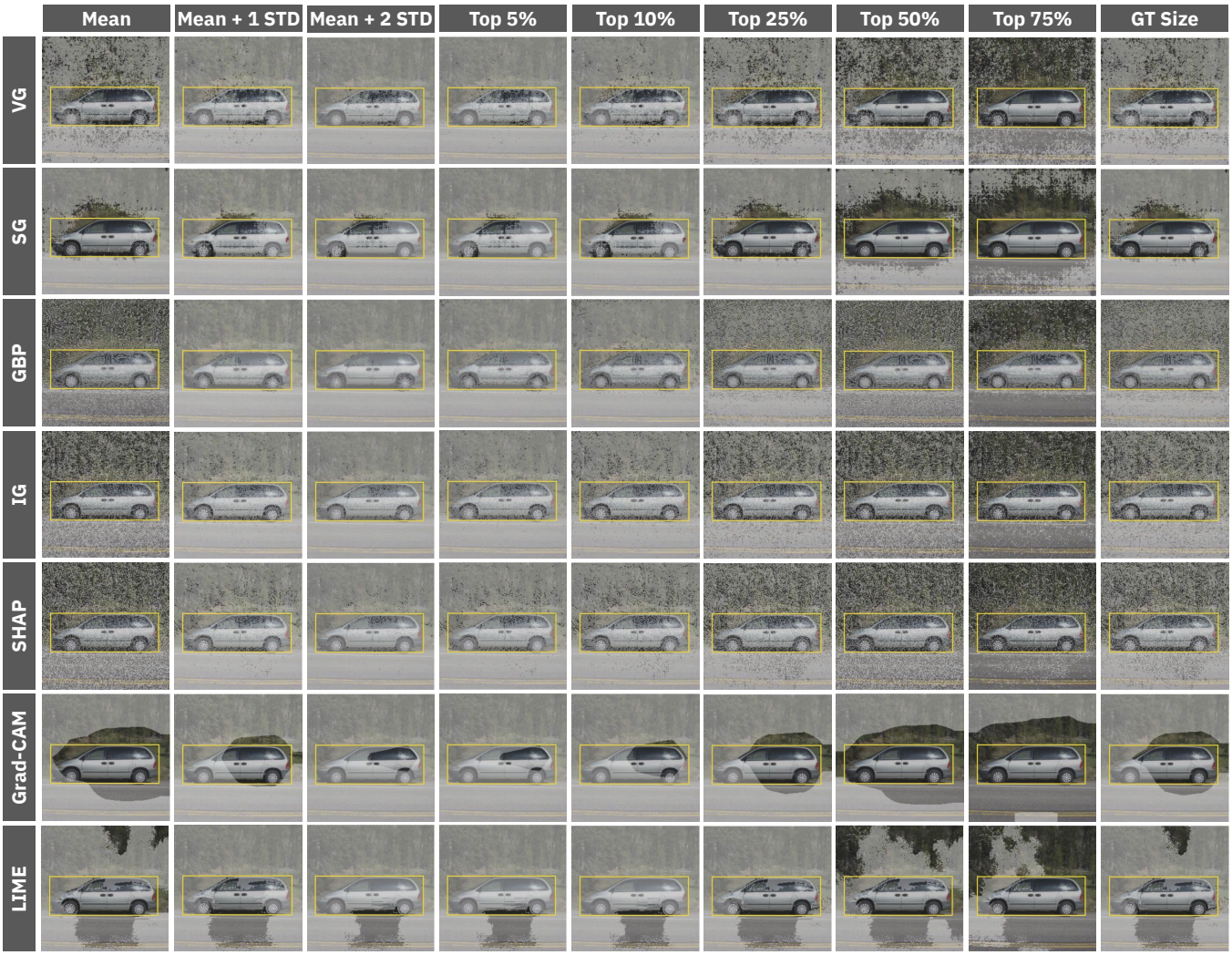

Figure B: Many discretization techniques can be successfully used with Shared Interest. Discretization can be score-based (dependent on the saliency scores) or model-based (dependent on the model's output). Here we show score-based thresholding techniques across different saliency methods on an ImageNet [10] *minivan* image. We compare Vanilla Gradients (VG) [36], SmoothGrad (SG) [37], Guided Backpropagation (GBP) [38], Integrated Gradients (IG) [41], Gradient SHAP (SHAP) [26], Grad-CAM [35], and LIME [34]. Score-based thresholding can be performed on the saliency values (e.g., all features with a saliency value greater than the mean saliency value) or based on the number of features (e.g., all features whose saliency is in the top 10% of saliency values). In this example, we compare thresholds of the mean, one standard deviation above the mean, two standard deviations above the mean, top 5%-75% of saliency values, and the same number of features as ground truth features (GT Size). As the thresholding technique relaxes, more features are added to the saliency feature set (shown via saturation), which changes the relationship to the ground truth features (shown in yellow). The Shared Interest scores depend on the discretization, and analysis should be performed with the discretization procedure's assumptions in mind.

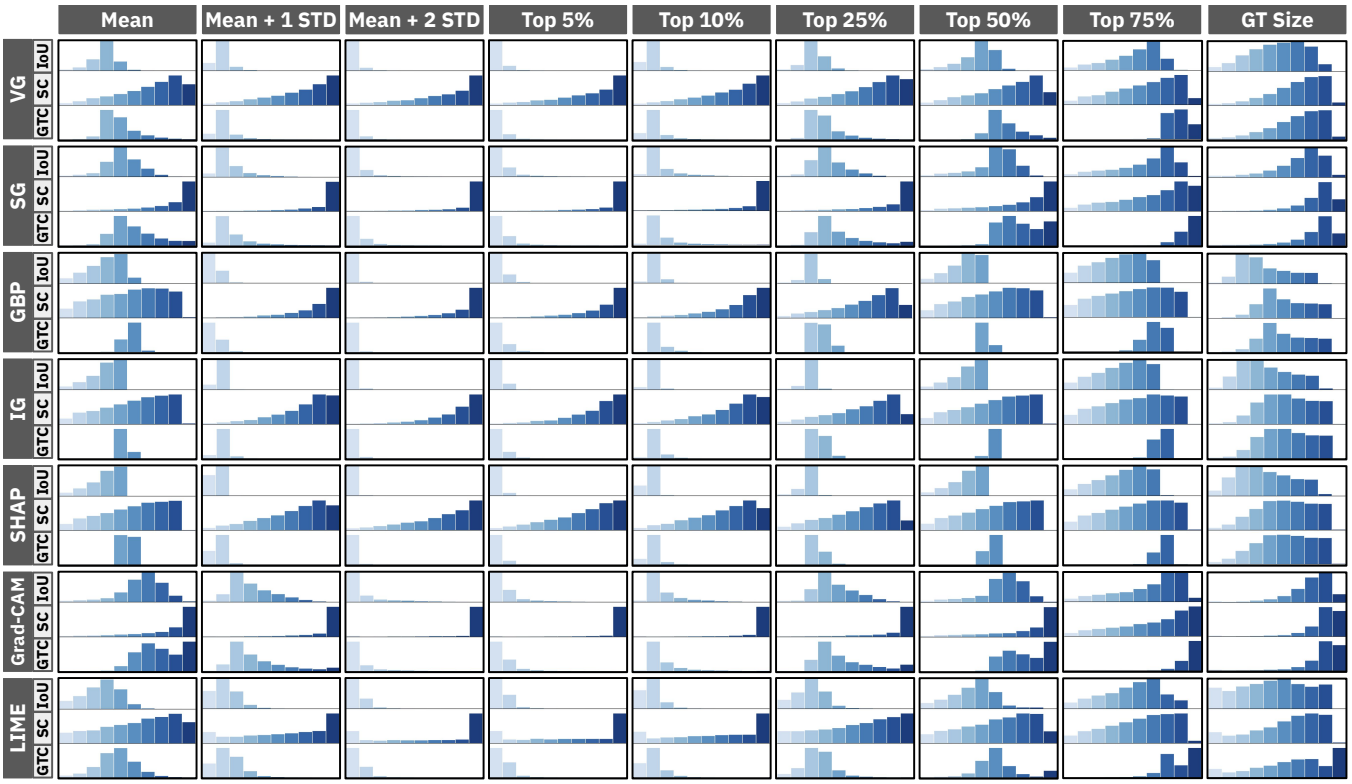

Figure C: Discretization techniques directly impact the Shared Interest scores. Here we compare the IoU, SC, and GTC Shared Interest score distributions on vehicle images from the ImageNet-9 dataset [44]. We compare distributions across saliency methods — Vanilla Gradients (VG) [36], SmoothGrad (SG) [37], Guided Backpropagation (GBP) [38], Integrated Gradients (IG) [41], Gradient SHAP (SHAP) [26], Grad-CAM [35], and LIME [34] — and thresholding technique — mean, one standard deviation above the mean, two standard deviations above the mean, top 5%-75% of saliency values, and the same number of features as ground truth features (GT Size). As the thresholding technique becomes stricter, IoU and GTC decrease and SC increases.