Explainable, Interactive Content-Based Image Retrieval

Bhavan Vasu | Brian Hu | Bo Dong | Roddy Collins | Anthony Hoogs

Kitware, Inc., New York, USA

Correspondence
*Please address all correspondence to Brian Hu (Email: brian.hu@kitware.com) or Anthony Hoogs (Email: anthony.hoogs@kitware.com)

Summary

Quantifying the value of explanations in a human-in-the-loop (HITL) system is difficult. Previous methods either measure explanation-specific values that do not correspond to user tasks and needs or poll users on how useful they find the explanations to be. In this work, we quantify how much explanations help the user through a utility-based paradigm that measures change in task performance when using explanations versus not. Our chosen task is content-based image retrieval (CBIR), which has well-established baselines and performance metrics independent of explainability. We extend an existing HITL image retrieval system that incorporates user feedback with similarity-based saliency maps (SBSM) that indicate to the user which parts of the retrieved images are most similar to the query image. The system helps the user understand what it is paying attention to through saliency maps, and the user helps the system understand their goal through saliency-guided relevance feedback. Using the MS-COCO dataset, a standard object detection and segmentation dataset, we conducted extensive, crowd-sourced experiments validating that SBSM improves interactive image retrieval. Although the performance increase is modest in the general case, in more difficult cases such as cluttered scenes, using explanations yields an 6.5% increase in accuracy. To the best of our knowledge, this is the first large-scale user study showing that visual saliency map explanations improve performance on a real-world, interactive task. Our utility-based evaluation paradigm is general and potentially applicable to any task for which explainability can be incorporated.

KEYWORDS:
explainable AI (XAI), image retrieval, user study, saliency

1 | INTRODUCTION

Deep-learning based systems can perform at or above human levels on some tasks (e.g. image recognition [1]), but the realization that such systems are vulnerable to visually-imperceptible adversarial attacks [2] has created a "trust gap" between these systems and their potential users [3,4]. One approach to bridging this gap is the concept of explainable AI (XAI) [5,6] in which systems provide not just an answer but an explanation of *why* the system has chosen that answer. The question arises as to how to evaluate an explanation; [7] argues for evaluating explanations along a number of axes which place the explanation in an ecosystem including the system, the user, the explanation, and (crucially) *the task the user is trying to perform.* An explanation's utility exists independently from the system's accuracy; a user may prefer a less-accurate system if they understand the circumstances for when the system may be wrong to a more-accurate system whose errors are incomprehensible.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ail2.41

In this work, we present a utility-based evaluation for explanations in an image retrieval system (a.k.a reverse image search). Building upon the method proposed in [8], our base system is an open-source, human-in-the-loop (HITL) image retrieval system which takes a single query image as input, returns a set of results, and then solicits relevance feedback from the user. The user indicates which results are or are not relevant to the query; this feedback is then used to train a lightweight classifier which re-ranks the results so that higher-ranked results should be more relevant. In the base system, the user has no way of knowing why any particular result is returned and considered relevant. In the XAI-enabled variant, the user is presented with saliency maps highlighting which regions in a result match most closely with the query; our hypothesis is that these explanations allow the user to more efficiently guide the system towards the object of interest and thus achieve a higher retrieval accuracy than without XAI (Figure 1). Saliency maps display additional information to the user, allowing them to provide more informed relevance feedback to the system. For example, a user may choose to not provide positive feedback on an example in which the saliency map is not well localized on the object of interest. This should help create a more generalizable image retrieval system which is trained with better feedback examples. We utilize *similarity-based saliency maps* (SBSM) [8] to visualize which areas in an image the content-based image retrieval (CBIR) system uses when retrieving and ranking results; the SBSM thus serves to "explain" the CBIR system's decisions to the user. We have implemented SBSMs in our open-source Social Media Query Toolkit (SMQTK) [9], and have conducted user studies to demonstrate that SBSMs allow the user to more efficiently retrieve images.

Our contributions include:

- To the best of our knowledge, the first large-scale, quantifiable improvement from using visual saliency maps on a real-world task. We conducted a crowd-sourced human study involving 476 Amazon Mechanical Turk subjects, each performing two rounds of image retrieval from a pool of 160 queries drawn from 24 object classes against an archive of approximately 123K MS-COCO images. Although the performance increase is modest in the general case, in more difficult cases such as cluttered scenes, using explanations yields an 6.5% increase in accuracy.
- A user study protocol to quantify the improvement in image retrieval performance with the help of explanations on a high-level machine learning task. The most common methods for evaluating visual explanations compare saliency maps with ground truth annotations, which does not actually measure how explanations improve task performance. We believe this utility-based paradigm is general and could easily be adapted to other kinds of explanations and tasks. For example, this approach could be extended to other image understanding tasks, such as image classification or object detection.
- Quantification of target data distribution parameters that lead to the greatest improvement in performance.
- Direct quantitative measurement of human trust in an explanation system based on a series of post-task questionnaires.

Our paper is structured as follows: first, we discuss related work (Section 2). Next, we review the similarity-based saliency map algorithm (Section 3) and provide a detailed description of the proposed human study evaluation protocol and quantitative metrics (Section 4). Finally, we describe our findings obtained after analyzing results from our user study (Section 5).

2 | RELATED WORK

Interpretable machine learning is a central topic of research, with interpretability playing an important role in high-stakes situations especially. However, research in this area has been largely limited to either better estimates of the 'goodness' of an explanation as a standalone component or algorithms to generate better explanations. In contrast, the focus of our work lies in developing a novel framework for measuring the effectiveness of explanations. [10] makes several important arguments regarding the interpretability of a model and conclude by highlighting the need for a fixed objective evaluation to term a model truly interpretable. [11] asked equally important questions and discussed material regarding the trade-offs present in successfully explaining a model's decision to an end-user. The work in [12], while acknowledging the importance of the understandability of explanations, also highlights the lack of a strong evaluation setup to evaluate the interpretability of models. Prior conceptual work on interpretability [10,11,12] concludes that explanations need to agree with human intuition and there is a lack of a commonly accepted quantitative evaluation standard.

Interpretability of models can be categorized into either white-box or black-box approaches. White-box approaches are limited to a narrow set of models and require access to internal elements of the model, while black-box approaches require only knowledge of the model inputs and outputs, and thus can explain any model. [13] introduced interpretability for linear models

while[14], [15], [16] and [17] did the same for deep networks. Local Interpretable Model Agnostic Explanations (LIME)[18] proposes to draw random samples around the instance for an explanation by fitting an approximate linear decision model. Random Input Sampling for Explanations (RISE)[19] was proposed as an improvement over LIME for generating black-box model explanations. More recently, Iterative and Adaptive Sampling[20] introduced an iterative method to sample around important regions in the image. With substantial research quantifying the effects of using one method over another[21], the scope of the present paper is not to decide which method is better, but rather to arrive at an evaluation strategy that is fair across all methods.

Evaluation protocols such as[22,23] are qualitative or task-specific, limiting reproducibility. Previous works for visual saliency have used metrics such as computational efficiency and Intersection over Union (IoU) to measure the alignment of saliency maps with object ground truth annotations, even though these are not directly related to interpretability.[24,25] proposed alternative solutions to the interpretability problem include asking the participants to guess the model decision based on the explanation for an instance and checking the agreement between the predicted and actual model decision;[26] measured the speed and accuracy with which the human can predict the model's decisions. These studies are quantitative but lack reproducibility due to the absence of a controlled environment and access to skilled participants with subject-specific knowledge. More recently,[27] proposed a study that operates under a crowdsourcing scenario with participants that are not task experts. The contribution of our work is to quantitatively and directly measure the effects of using explanations in an HITL system, in which the user interacts with the system to provide relevance feedback, yielding higher performance than a standard image retrieval model. An additional challenge in our work relative to[27] is that our system's underlying retrieval model dynamically evolves in response to the user's feedback. Our work also differs because we tie interpretability of a model to task performance (to wit, retrieval accuracy), human-AI interaction and finally, computational efficiency. For a given method of explanation generation, our evaluation setup can identify the subsets of data that will benefit the most from having visual explanations as well.

3 | SIMILARITY-BASED SALIENCY MAPS

Similarity-based saliency maps (SBSM) are a variant of pertubation-based saliency maps, which can be used to indicate the importance of image regions against some criteria [28,29,30,31]. In the context of content-based image retrieval (CBIR), a saliency map should indicate how a particular region on the retrieved image impacts visual similarity with a query image. However, classification-based saliency maps indicate how a particular image region impacts classification probability, which is not applicable to image similarity. Our SBSM instead measures how image regions in a result image (when compared to a query image) contribute to the distance metric used by the CBIR system when computing image similarity. Most similiar to our work are representer point selection[32] and explainable additive models[33], which have been used for image classification tasks but have not yet been applied to image retrieval. Representer point selection[33] produces a set of positive and negative images for a given prediction, which can be used for interpretability purposes. In our study, positive and negative images are actually used to train a classifier in a human-in-the-loop, interactive image retrieval paradigm. Explainable additive models[32] start with a base network and train an interpretable additive explainer via model distillation. In contrast, the SBSM algorithm proposed here is purely black-box and only requires access to the inputs and outputs of the base network.

We perturb a retrieved image by applying a binary mask to block out a region of interest. Inside the binary mask, the region of interest has value 0; all other pixels have value 1. In general, the region of interest can be of any shape. In our setup, we simply use a b x b square block. By sliding the square block over the retrieval image with a stride of s, we are able to show the importance of the blocked areas on image similarity. In order to leverage parallel computing resources (GPUs), we generate a set of binary masks $\boldsymbol{M}$, each of which represents a state of sliding the square block.

Mathematically, given a query image Q, a retrieval image A and a binary mask $m_i \in \boldsymbol{M}$, the importance of the region blocked out by m_i is estimated as follows:

$$K(Q, A, m_i) = \max(D' - D, 0)(\mathbb{I} - m_i), \tag{1}$$

$$D' = \|f(Q), f(A \odot m_i)\|, \tag{2}$$

$$D = \|(f(Q), f(A)\|, \tag{3}$$

where, $f : I \rightarrow \mathbb{R}^n$ is a black-box model, which maps a input image I to a n dimensional vector; $\odot$ denotes element-wise multiplication; $\|\vec{v}_1, \vec{v}_2\|$ is the similarity between the two vectors $\vec{v}_1$ and $\vec{v}_2$ based on a user-defined distance metric (e.g. L_2 distance); $\mathbb{I}$ is a matrix with all entries are 1 and the same shape as m_i. Given a binary mask set $\boldsymbol{M}$ with N binary masks, the

SBSM is calculated as:

$$\text{SBSM}(Q, A, M) = \sum_{i}^{N} K(Q, A, m_i) \odot \frac{1}{\sum_{i}^{N} (\mathbb{I} - m_i)}. \quad (4)$$

Intuitively, Eq. 4 says that when a region is overlapped by multiple masks, the mean value is used to express the importance of the region. An overview of the proposed SBSM approach is illustrated in Figure 2.

4 | EVALUATION PROTOCOL

Identifying conditions under which an image retrieval system excels requires the use of a wide variety of images from a large distribution of natural image to ensure stable and reproducible outcomes. In this paper, we consider the MS-COCO[34] dataset because it has a large range of possible queries per class and a high number of samples per class.

User evaluation protocol. The archive consisted of approximately 123K images from the 2014 and 2017 editions of the MS-COCO[34] dataset. Our query set is 24 object types or classes represented by a total of 160 images. The query classes were selected based on pilot evaluation results indicating greater XAI benefit in images with higher category diversity. The 160 images were sampled into 148 "task pair" image ID tuples; the first image was the without-XAI image; the second was the with-XAI image. Task pairs were generated such that the same class was never presented as both *with-XAI* and *without-XAI* for a given user. We structured our AMT jobs such that each experiment (or HIT, "Human Intelligence Task") was performed by a different AMT user ID. In the end, we were able to loop over each task pair at least three times, yielding a total of 476 data points. Human studies conducted on AMT often mandate additional rigor to ensure the quality of the data. To uphold trust in the data, we used the following criteria: tasks were only assigned to residents of either US or Canada, users had to have completed at least 1000 HITs prior to the study, and user had to have had an average hit approval rate above 97%.

The HIT for a task pair was: for each type (with- and without-XAI), find at least 12 instances of the class. To ensure proper counterbalancing, the order of the *with-XAI* and *without-XAI* conditions was random. The specific procedure was: read a tutorial; perform the first image search task; answer a sub-task questionnaire; perform the second image search task; answer a sub-task questionnaire, then answer a short final overall questionnaire. The sub-task questionnaires asked specific questions based on the with- or without-XAI condition; the final overall questionnaire asked comparative questions. By comparing results obtained based on the with-XAI and without-XAI conditions, we hope to identify query classes that benefit the most from explanations in terms of performance and trust.

The basic task of our user evaluation is: *Given a query image containing an object of a designated type, find 12 additional instances of that object type in the archive.* Note that the type label is provided for the user's reference only; our underlying image retrieval system that makes use of SMQTK remains wholly image-based. The **with-XAI** condition supplies an SBSM map with each result; the **without-XAI** condition has no SBSM map. The with-XAI state is shown in Figure 3. The specific hypothesis is that **the SBSM map "explains" the retrieval ranking by highlighting result regions which the retrieval system pays attention to and that this explanation will help the user provide more efficient relevance feedback.** The main performance metric is the number of images containing the target type that have been retrieved at the end of the experiment.

Quantifiable Metrics. The user study conducted in this paper collects both user and query dependent metrics during image retrieval. Under user-specific metrics, we ask users a wide variety of questions that try to identify the effectiveness of the system and explanations. The purpose of these questions is to monitor human trust in the system's retrieval ability across different classes, queries, and instance sizes. Additionally, we also collect other system-level parameters like the number of refinements taken to complete a task and finally the number of images adjudicated by different users on the same task.

The improvement in performance is calculated by comparing the number of positive samples found using the with-XAI and without-XAI versions of the system. Similarly, we also aggregate task-specific parameters like the number of adjudications and user trust to test whether there are improvements with XAI. We hypothesize that a higher level of interaction from the user in the form of adjudications signifies a more significant deviation in model and user interpretation of the image. This kind of analysis helps us to identify conditions under which the user and system benefited from explanations, conditions that did not benefit from explanations, but helped the user understand the system better and finally conditions that are negatively impacted by explanations.

To compute statistical significance of the reported results, we used a permutation or randomization test. More specifically, to test the null hypothesis that the number of positive samples found is the same in the with-XAI and without-XAI conditions, we randomly shuffled the labels associated with XAI condition and reassigned each data point to its new label. We did this while

preserving the overall query distribution such that the number of data points for each query class was preserved (i.e. sampling without replacement). For each permutation, we computed a new resampled XAI gain. Doing this multiple times builds a distribution of resampled XAI gains, from which we can compute a p-value for the statistical significance of our observed XAI gain. We calculated the permutation test using a total of N=10,000 randomizations of the original input data. We used an alpha level of 0.1 for all statistical tests.

5 | RESULTS

5.1 | Improvement in human-AI performance

In this work, the benefit of explanations in the form of saliency maps is measured through a utility-based paradigm, e.g. by computing the number of positive examples found in the with- and without-XAI conditions. If explanations help, we expect that the number of positive examples found will be higher in the with- versus the without-XAI condition. We believe that saliency maps can provide users additional information upon which they can base their relevance feedback decisions, which results in a better and more generalizable image retrieval system that can be quantified by different metrics.

In brief, our results show that **XAI benefits classes with high image class diversity or clutter**. Our results are visualized in Figure 4 for the 24 query classes. The X-axis is the *ratio* of the number of images found with-XAI to those found without-XAI. The Y-axis is the average number of unique classes appearing in the archive for a given type; higher numbers indicate a more "busy" or "confused" image. The diameter of the circles represents the average size of class instances in the archive; dining tables (the largest) are about 40 times larger than spoons (the smallest.) The insight from Figure 4 is that when XAI benefits an object class, that is, when the class appears to the right of X==1.0, *these classes tend to be in more diverse images* than those which do not benefit from XAI. In other words, *SBSM XAI helps find objects in busy images by highlighting the object of interest*. We also observed a slight XAI benefit for smaller objects. In contrast, our results also show that XAI does not benefit (and may even hurt performance) on classes with very little clutter. In this case, XAI may not be required as objects are more easily localizable even without the aid of saliency maps (and saliency maps incur additional information overhead that may not be beneficial). We leave a more complete quantification of conditions under which XAI shows benefit for future work.

To facilitate analysis, we define a set of attributes based on the "Confuser" (Y-axis) and "Area" (bubble diameter) attributes in Figure 4: A class is a member of **C**lutter if it is in the top 12 classes for "Confuser", else it is **U**ncluttered; likewise, a class is a member of **L**arge if it is in the top 12 classes for area, else it is in **S**mall. Using these attributes, the 24 query classes are partitioned among the categories (CL, CS, UL, US) as follows:

- **CL** (*Cluttered, Large*): dining_table, microwave, oven, sink, toaster, vase
- **CS** (*Cluttered, Small*): book, chair, fork, knife, spoon, wine_glass
- **UL** (*Uncluttered, Large*): cake, keyboard, laptop, sandwich, tv, umbrella
- **US** (*Uncluttered, Small*): apple, backpack, carrot, clock, handbag, orange

The conclusion of Figure 4 can be restated as *SBSM XAI benefits members of the **C**luttered classes, and to a lesser degree, the Small classes*. For the **C**luttered classes, users were able to find on average 1.2 more positive examples with XAI (18.9 vs. 17.7 positive examples per query image with and without XAI, respectively). For the **S**mall classes, users were able to find on average 0.7 more positive examples with XAI (18.6 vs. 17.9 positive examples per query image with and without XAI, respectively). This translates into an observed 6.5% gain from XAI for the **C**luttered class groups (p=0.09, permutation test) and a smaller 3.9% increase in XAI benefit for the **S**mall classes (p=0.18, permutation test). To ensure that the quantitative improvement is not due to simply having more user study query examples in a given group, Table 1, for the **qD** attribute, shows the distribution of queries according to these attributes for XAI benefit across with- and without-XAI case. The **ratio** attribute shows the relative prevalence of the **CL**, **CS**, **UL**, and **US** groups between the with- and without-XAI conditions; we normalize by these ratios to compensate for having more user study query examples for one group compared to another.

5.2 | Level of human-AI interaction

The human-AI interaction is in the form of positive and negative adjudications to the system that help influence the retrieval results. In this section, we compare the number of adjudications required across both the with-XAI and without-XAI system for

the same query image for different users. Table 2, attribute **nA**, shows the distribution of human adjudications for the conditions. For the **S**mall classes, users made on average 1.7 more adjudications with XAI (32.9 vs. 31.2 adjudications per query image with and without XAI, respectively). For the **C**luttered classes, users made on average 1.7 more adjudications with XAI (39.6 vs. 37.9 adjudications per query image with and without XAI, respectively). This translates into **S**mall classes showing a 4.4% increase in adjudications and the **C**luttered classes showing adjudication gains of 5.5%. Finally, for the **C**luttered classes, users took on average 148.2 seconds longer with XAI (589.1 vs. 440.9 seconds per query image with and without XAI, respectively) and for the **S**mall classes, users took on average 134.4 seconds longer with XAI (625.7 vs. 491.2 seconds per query image with and without XAI, respectively). As a result, we observed a 33.6% and 27.3% increase in time taken in seconds (**T**) to complete the task for the cluttered and small classes, respectively. These results suggest that while XAI helped to increase the overall number of positive examples found, it also slightly increased the number of adjudications required and the time taken to complete the task.

5.3 | Human trust in system

To determine human trust in the system, we asked the users several questions about their experience using the system. We chose to use a self-reporting questionnaire due to ease of training and compatibility with our workflow which allowed us to run the user study remotely on AWS. We acknowledge that data collected through such questionnaires could potentially be biased, and suggest that task-based evaluations could be used as an alternative measure of the ease-of-use or understandability of explanations (which is left as an area for future work).

Questionnaire responses. Users were presented with questions at three points during the experiment: questionnaires tailored to the with- and without-XAI conditions after the individual tasks, and a short final questionnaire. Questions were on a 6-point Likert scale, 1 =“strongly disagree", 6=“strongly agree". We made the following observations:

- **XAI helps users give feedback.** 60% agreed at some level that “Overall, I feel saliency maps helped me give better feedback."
- **XAI helps users understand.** 83% agreed at some level, with 56% agreeing or strongly agreeing, that “Saliency map helped me understand how the system “thinks."
- **XAI improves ease-of-use.** 62% agreed at some level, with 38% agreeing or strongly agreeing, that “Saliency maps made the system easier to use."
- **No clear signal on preference for saliency maps.** 58% agreed at some level that they would “prefer to do [the task] with saliency maps rather than without"; however, 60% **also** agreed at some level with the opposite question that they would “prefer to do [the task] without saliency maps rather than with."
- **Responder confidence.** 95% agreed or strongly agreed that they understood the questions.

Figure 5 shows a detailed breakdown of user responses for their preference towards an XAI and non-XAI system to perform image retrieval. More directly, we collate answers to the statement, “If given the option to redo the task, I would prefer to do it with saliency maps rather than without saliency maps". We did not observe any clear trends when comparing the qualitative survey results shown in Figure 5 with the quantitative results reported in Table 2. For example, on the cluttered image classes we did not necessarily observe a stronger preference for using saliency maps. Please note the data points shown in Figure 5 for comparing preference to the with- and without-XAI conditions are obtained from different sub-populations of our user study group assuming they are equal and balanced in all other aspects. In other words, the same user does not see both the XAI and the no-XAI condition for the same query class (to avoid potential bias due to pre-exposure to the same query class), therefore we present user preferences across both with and without-XAI.

6 | CONCLUSION

We evaluated, both quantitatively and qualitatively, how augmenting an image retrieval system with XAI in the form of saliency maps improves the human-in-the-loop task of finding images containing a given object type from the archive. We tasked 476 users to search for different objects, once with the XAI condition and once without. Queries were drawn from a pool of 24 object types; the archive contained approximately 123K images from MS-COCO [34]. We found that the user was able to find 6.5% more

instances in the case when the query object occurs in cluttered images, and 3.9% more instances in the case when the query object is relatively small. This suggests a potential benefit to XAI in cluttered image scenes. Users also provided 5.5% more feedback in the cluttered-image case, and 4.4% more feedback in the small-object case. The XAI user feedback via questionnaires indicates that the users feel XAI helps them give more relevant feedback and increases ease-of-use and understanding of how the system operates. We believe that the utility-based paradigm for evaluating the effectiveness of explanations introduced here is also broadly applicable to other human-machine teaming tasks where explainability can be incorporated, which we leave as an area of future work.

ACKNOWLEDGMENTS

This material is based on research sponsored by Air Force Research Laboratory and DARPA under Cooperative Agreement number N66001-17-2-4028. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and DARPA or the U.S. Government. Distribution Statement 'A' (Approved for Public Release, Distribution Unlimited). We would also like to thank Benjamin Pikus for helpful comments.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: ; 2016: 770–778.
2. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* 2013.
3. Lengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change* 2016; 105: 105–120.
4. Jiang H, Kim B, Guan M, Gupta M. To trust or not to trust a classifier. In: ; 2018: 5541–5552.
5. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* 2017.
6. Gunning D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2017; 2.
7. Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* 2018.
8. Dong B, Collins R, Hoogs A. Explainability for Content-Based Image Retrieval. In: ; 2019.
9. Kitware . Social Media Query Toolkit (SMQTK). https://github.com/Kitware/SMQTK; .
10. Lipton ZC. The Mythos of Model Interpretability. arXiv, cs. 2016.
11. Herman B. The Promise and Peril of Human Evaluation for Model Interpretability. *ArXiv* 2017; abs/1711.07414.

12. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* 2017.

13. Haufe S, Meinecke F, Görgen K, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 2014; 87: 96–110.

14. Zeiler M, Fergus R. Visualizing and understanding convolutional networks. In: . 8689 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag; 2014: 818–833

15. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: Bengio Y, LeCun Y., eds. *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*; 2014.

16. Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. In: ; 2019: 8928–8939.

17. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 2017; 65: 211–222.

18. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. In: ; 2016: 1135–1144.

19. Petsiuk V, Das A, Saenko K. RISE: Randomized Input Sampling for Explanation of Black-box Models. In: ; 2018.

20. Vasu B, Long C. Iterative and Adaptive Sampling with Spatial Attention for Black-Box Model Explanations. In: ; 2020.

21. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019; 1(5): 206–215.

22. Kononenko I, others . An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research* 2010; 11(Jan): 1–18.

23. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: ; 2017: 4765–4774.

24. Ribeiro MT, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. In: ; 2018.

25. Lakkaraju H, Bach SH, Leskovec J. Interpretable decision sets: A joint framework for description and prediction. In: ; 2016: 1675–1684.

26. Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 2011; 51(1): 141–154.

27. Schmidt P, Biessmann F. Quantifying Interpretability and Trust in Machine Learning Systems. *arXiv preprint arXiv:1901.08558* 2019.

28. Fong R, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *CoRR* 2017; abs/1704.03296.

29. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* 2016; abs/1602.04938.

30. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. *CoRR* 2013; abs/1311.2901.

31. Petsiuk V, Das A, Saenko K. RISE: Randomized Input Sampling for Explanation of Black-box Models. *ArXiv e-prints* 2018.

32. Chen R, Chen H, Ren J, Huang G, Zhang Q. Explaining neural networks semantically and quantitatively. In: ; 2019: 9187–9196.

33. Yeh CK, Kim JS, Yen IE, Ravikumar P. Representer point selection for explaining deep neural networks. In: ; 2018: 9311–9321.

34. Lin TY, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. In: Springer. ; 2014: 740–755.

35. Authors . The frobnicatable foo filter. 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.

36. Authors . Frobnication tutorial. 2014. Supplied as additional material `tr.pdf`.

37. Alpher A. Frobnication. *Journal of Foo* 2002; 12(1): 234–778.

38. Alpher A, Fotheringham-Smythe JPN. Frobnication revisited. *Journal of Foo* 2003; 13(1): 234–778.

39. Alpher A, Fotheringham-Smythe JPN, Gamow G. Can a machine frobnicate?. *Journal of Foo* 2004; 14(1): 234–778.

40. Lin T, Maire M, Belongie SJ, et al. Microsoft COCO: Common Objects in Context. *CoRR* 2014; abs/1405.0312.

41. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. In: ; 2017.

42. Swain MJ, Ballard DH. Color Indexing. *Int. J. Comput. Vision* 1991; 7(1): 11–32. doi: 10.1007/BF00130487

43. Barla A, Odone F, Verri A. Histogram intersection kernel for image classification.. In: ; 2003: 513-516.

44. Gordo A, Larlus D. Beyond Instance-Level Image Retrieval: Leveraging Captions to Learn a Global Visual Representation for Semantic Retrieval. In: ; 2017: 5272–5281

45. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* 2016; abs/1610.02391.

46. Mai L, Jin H, Lin ZL, Fang C, Brandt J, Liu F. Spatial-Semantic Image Search by Visual Feature Synthesis. In: ; 2017: 1121–1130

47. Yang F, Li J, Wei S, Zheng Q, Liu T, Zhao Y. Two-stream Attentive CNNs for Image Retrieval. In: MM '17. ACM; 2017; New York, NY, USA: 1513–1521

48. Rahman MM, Desai BC, Bhattacharya P. Visual Keyword-based Image Retrieval Using Latent Semantic Indexing, Correlation-enhanced Similarity Matching and Query Expansion in Inverted Index. In: IDEAS '06. IEEE Computer Society; 2006; Washington, DC, USA: 201–208

49. Frankel C, Swain MJ, Athitsos V. WebSeer: An Image Search Engine for the World Wide Web. tech. rep., Chicago, IL, USA: 1996.

50. Babenko A, Lempitsky VS. Aggregating Deep Convolutional Features for Image Retrieval. *CoRR* 2015; abs/1510.07493.

51. Paulin M, Mairal J, Douze M, Harchaoui Z, Perronnin F, Schmid C. Convolutional Patch Representations for Image Retrieval: an Unsupervised Approach. *CoRR* 2016; abs/1603.00438.

52. Li Y, Su H, Qi CR, Fish N, Cohen-Or D, Guibas LJ. Joint Embeddings of Shapes and Images via CNN Image Purification. *ACM Trans. Graph.* 2015; 34(6): 234:1–234:12. doi: 10.1145/2816795.2818071

53. Radenovic F, Tolias G, Chum O. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. *CoRR* 2016; abs/1604.02426.

54. Yu W, Yang K, Yao H, Sun X, Xu P. Exploiting the Complementary Strengths of Multi-layer CNN Features for Image Retrieval. *Neurocomput.* 2017; 237(C): 235–241. doi: 10.1016/j.neucom.2016.12.002

55. Frome A, Corrado GS, Shlens J, et al. DeViSE: A Deep Visual-Semantic Embedding Model. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ., eds. *Advances in Neural Information Processing Systems 26*Curran Associates, Inc. 2013 (pp. 2121–2129).

56. Li X, Liao S, Lan W, Du X, Yang G. Zero-shot Image Tagging by Hierarchical Semantic Embedding. In: SIGIR '15. ACM; 2015; New York, NY, USA: 879–882

57. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ., eds. *Advances in Neural Information Processing Systems 26*Curran Associates, Inc. 2013 (pp. 3111–3119).

58. Norouzi M, Mikolov T, Bengio S, et al. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In: ; 2014.

59. Wang F, Kang L, Li Y. Sketch-based 3D shape retrieval using Convolutional Neural Networks.. In: IEEE Computer Society; 2015: 1875-1883.

60. Yang Y, Hospedales TM. Deep Neural Networks for Sketch Recognition. *CoRR* 2015; abs/1501.07873.

61. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA. Striving for Simplicity: The All Convolutional Net. *CoRR* 2014; abs/1412.6806.

62. Gan C, Wang N, Yang Y, Yeung DY, Hauptmann AG. DevNet: A Deep Event Network for multimedia event detection and evidence recounting. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2015: 2568-2577.

63. Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. *CoRR* 2015; abs/1512.04150.

64. Zhou X, Huang T. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems* 2003; 8(6): 536–544. doi: 10.1007/s00530-002-0070-3

65. Rui Y, Huang TS, Ortega M, Mehrotra S. Relevance Feedback: A Power Tool for Interactive Content-based Image Retrieval. *IEEE Trans. Cir. and Sys. for Video Technol.* 1998; 8(5): 644–655. doi: 10.1109/76.718510

List of Figures

1 Pipeline of the human-in-the-loop image retrieval system that searches for relevant images in an archive given a query image. An initial set of ranked results is retrieved based on distance to the query image features extracted from a convolutional neural network. The user then provides positive and negative feedback on the relevance of the returned results. Explanations in the form of saliency maps indicating image regions used by the system are also be provided to the user in the with-XAI condition. This feedback is used to train a lightweight classifier that re-ranks the results over multiple iterations. 12

2 Similarity-based saliency map (SBSM) generation. Result image R is masked by binary mask m_i and run through feature extractor f; the distance between this new feature vector and the original query/result distance is used to compute how relevant masked region m_i is in the final SBSM. 13

3 Proposed user interface for evaluating the effectiveness of visual explanations. The user performs image retrieval once with XAI and once without. This figure shows the with-XAI condition for retrieving cows; the without-XAI condition would have a different query class and no explanations. Top, task instructions; left; the user has already found 9 cows; right, results marked as “not cows”; center: unadjudicated results list with corresponding saliency maps. 14

4 Class-wise image retrieval performance improvement with XAI. The X-axis is the ratio of number of images containing the class found with-XAI vs. without-XAI; values greater than 1.0 indicate an XAI benefit. The Y-axis is the average number of classes appearing in images in the archive; higher numbers indicate the given class appears with more classes. The relative diameter of each circle indicates the average size of the class in images in the archive. 15

5 Results for improvement in human trust with XAI; the statement asked was **"If given the option to redo the task, I would prefer to do it with saliency maps rather than without saliency maps."** Survey responses are normalized from a scale of 0 to 1: strongly diagree (0.0), disagree (0.2), slightly disagree (0.4), slightly agree (0.6), agree (0.8), and strongly agree (1.0). 16

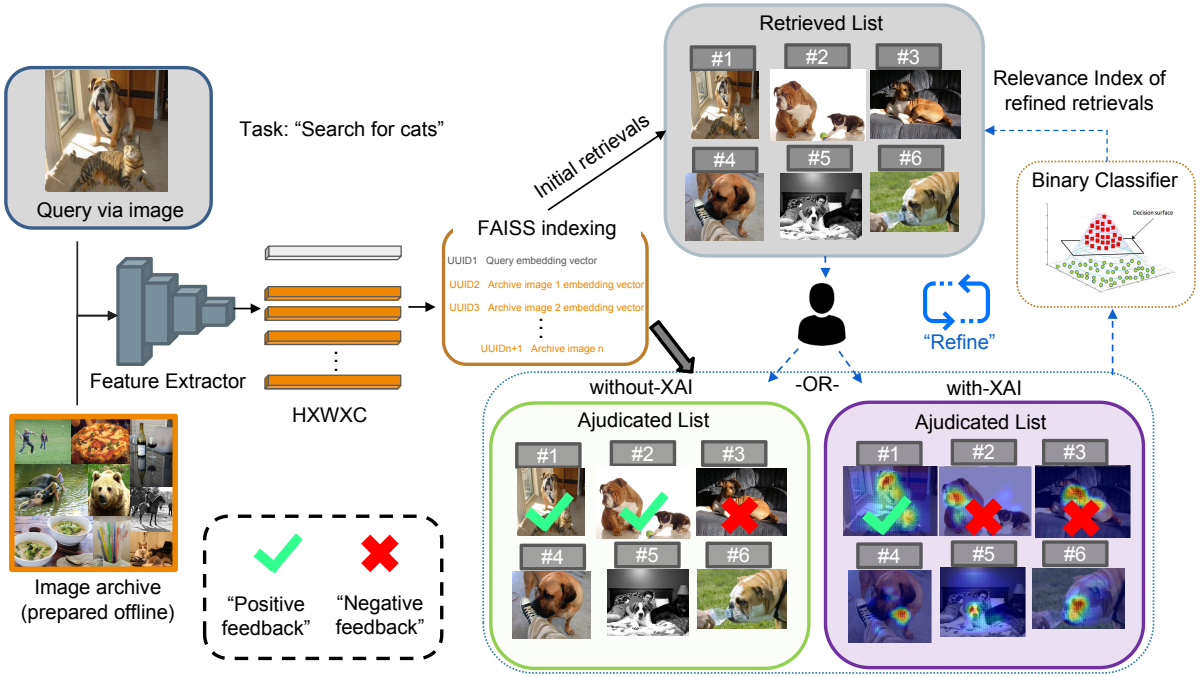

FIGURE 1 Pipeline of the human-in-the-loop image retrieval system that searches for relevant images in an archive given a query image. An initial set of ranked results is retrieved based on distance to the query image features extracted from a convolutional neural network. The user then provides positive and negative feedback on the relevance of the returned results. Explanations in the form of saliency maps indicating image regions used by the system are also be provided to the user in the with-XAI condition. This feedback is used to train a lightweight classifier that re-ranks the results over multiple iterations.

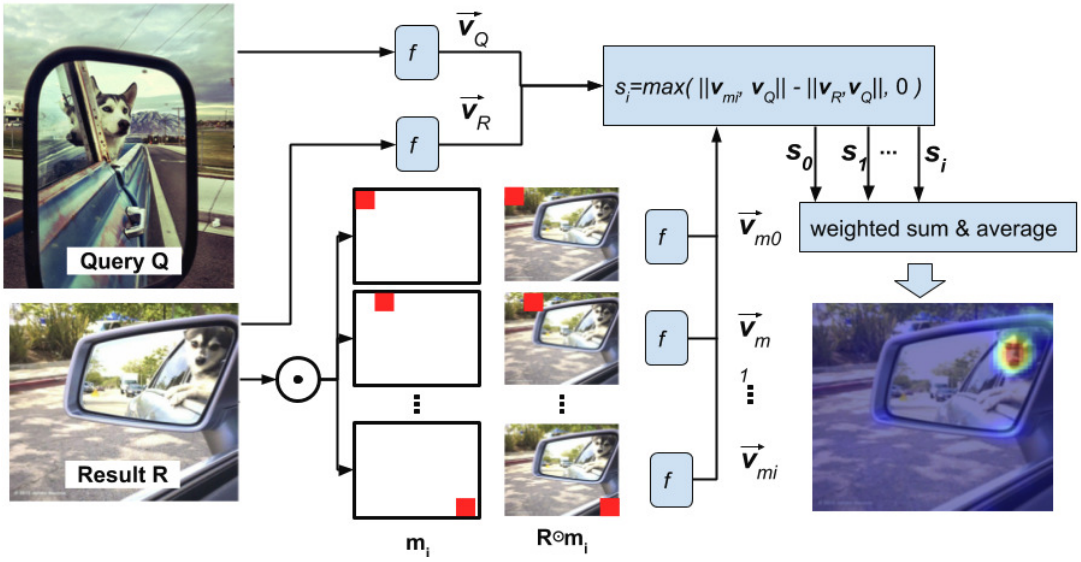

FIGURE 2 Similarity-based saliency map (SBSM) generation. Result image R is masked by binary mask m_i and run through feature extractor f; the distance between this new feature vector and the original query/result distance is used to compute how relevant masked region m_i is in the final SBSM.

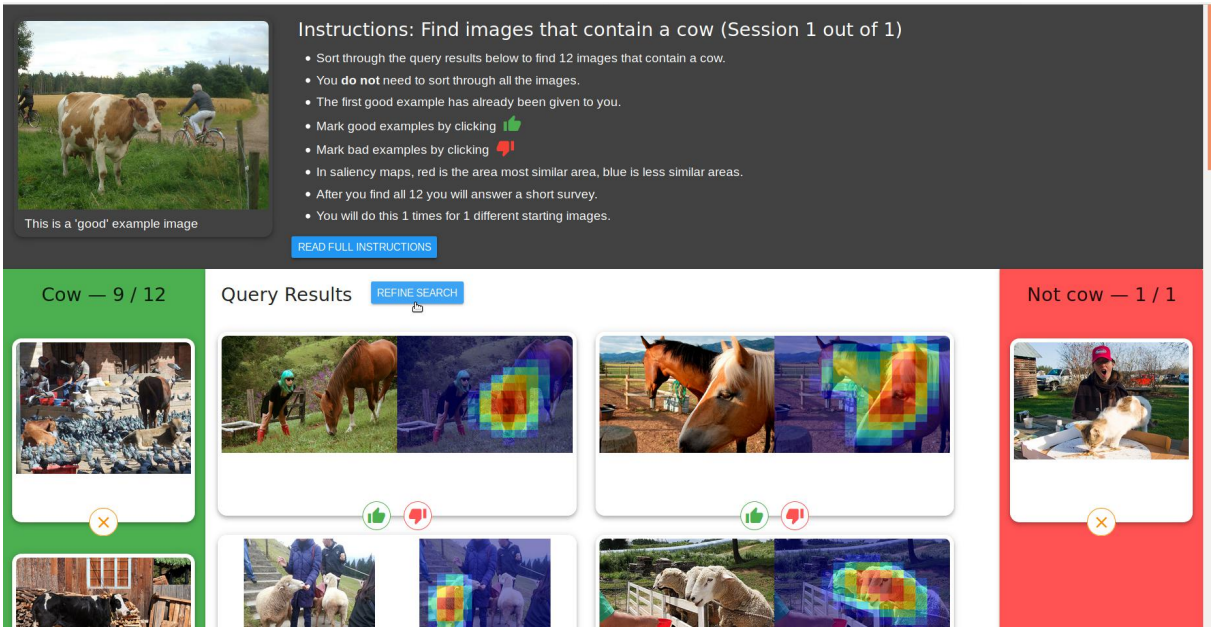

FIGURE 3 Proposed user interface for evaluating the effectiveness of visual explanations. The user performs image retrieval once with XAI and once without. This figure shows the with-XAI condition for retrieving cows; the without-XAI condition would have a different query class and no explanations. Top, task instructions; left; the user has already found 9 cows; right, results marked as "not cows"; center: unadjudicated results list with corresponding saliency maps.

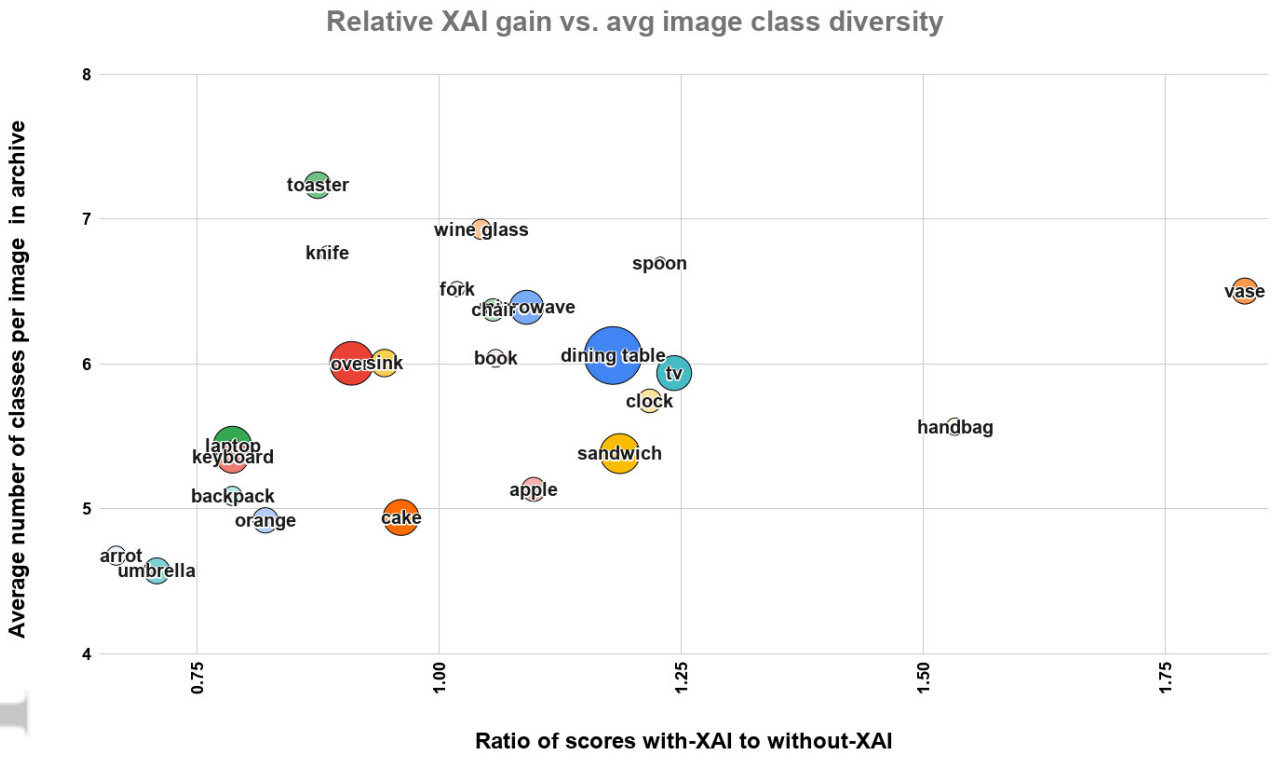

FIGURE 4 Class-wise image retrieval performance improvement with XAI. The X-axis is the ratio of number of images containing the class found with-XAI vs. without-XAI; values greater than 1.0 indicate an XAI benefit. The Y-axis is the average number of classes appearing in images in the archive; higher numbers indicate the given class appears with more classes. The relative diameter of each circle indicates the average size of the class in images in the archive.

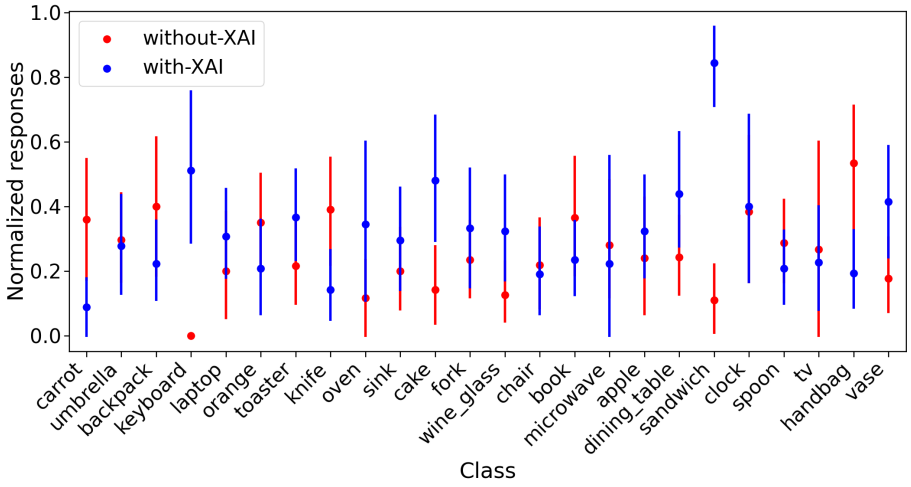

FIGURE 5 Results for improvement in human trust with XAI; the statement asked was **"If given the option to redo the task, I would prefer to do it with saliency maps rather than without saliency maps."** Survey responses are normalized from a scale of 0 to 1: strongly diagree (0.0), disagree (0.2), slightly disagree (0.4), slightly agree (0.6), agree (0.8), and strongly agree (1.0).

List of Tables

1 Query distribution for the **CS**, **CL**, **US**, **UL** superclasses as raw counts (**qD-raw**) and fractions (**qD**). **ratio** shows the relative distribution of with-XAI to no-XAI. 18

2 Results of with / without-XAI conditions for the **CS**, **CL**, **US**, **UL** superclasses across the attributes of number of positives retrieved (**nP**), number of adjudications (**nA**), and time taken in seconds (**T**). Ratios are normalized using the **ratio** attribute for the class from Table 1. The percent XAI gain for the cluttered and small query classes is shown in the last two columns, respectively. 19

		with-XAI		no-XAI	
attr	clutter	size(**L**)	size(**S**)	size(**L**)	size(**S**)
qD-raw	**C**	111	141	138	150
	U	101	123	89	99
qD	**C**	0.23	0.30	0.29	0.32
	U	0.21	0.26	0.19	0.21
ratio	**C**	0.80	0.94		
	U	1.13	1.24		

TABLE 1 Query distribution for the **CS**, **CL**, **US**, **UL** superclasses as raw counts (**qD-raw**) and fractions (**qD**). **ratio** shows the relative distribution of with-XAI to no-XAI.

		ratio XAI/no-XAI		XAI gain for	
attr	clutter	size(**L**)	size(**S**)	**Cluttered (C)**	**Small (S)**
nP	**C**	1.01	1.09	6.5%	3.9%
	U	1.00	0.99		
nA	**C**	1.27	0.91	5.5%	4.4%
	U	1.05	1.16		
T	**C**	1.48	1.22	33.6%	27.3%
	U	1.33	1.30		

TABLE 2 Results of with / without-XAI conditions for the **CS**, **CL**, **US**, **UL** superclasses across the attributes of number of positives retrieved (**nP**), number of adjudications (**nA**), and time taken in seconds (**T**). Ratios are normalized using the **ratio** attribute for the class from Table 1. The percent XAI gain for the cluttered and small query classes is shown in the last two columns, respectively.