

Interactivity x Explainability: Toward Understanding How Interactivity Can Improve Computer Vision Explanations

Indu Panigrahi
Princeton University
Princeton, New Jersey, USA
indup@princeton.edu

Sunnie S. Y. Kim*
Princeton University
Princeton, New Jersey, USA
sunniesuhyoung@princeton.edu

Amna Liaqat*
Princeton University
Princeton, New Jersey, USA
al0910@princeton.edu

Rohan Jinturkar
Princeton University
Princeton, New Jersey, USA
rohanj@alumni.princeton.edu

Olga Russakovsky
Princeton University
Princeton, New Jersey, USA
olgarus@princeton.edu

Ruth Fong
Princeton University
Princeton, New Jersey, USA
ruthfong@princeton.edu

Parastoo Abtahi
Princeton University
Princeton, New Jersey, USA
parastoo@princeton.edu

Abstract

Explanations for computer vision models are important tools for interpreting how the underlying models work. However, they are often presented in static formats, which pose challenges for users, including information overload, a gap between semantic and pixel-level information, and limited opportunities for exploration. We investigate interactivity as a mechanism for tackling these issues in three common explanation types: heatmap-based, concept-based, and prototype-based explanations. We conducted a study (N=24), using a bird identification task, involving participants with diverse technical and domain expertise. We found that while interactivity enhances user control, facilitates rapid convergence to relevant information, and allows users to expand their understanding of the model and explanation, it also introduces new challenges. To address these, we provide design recommendations for interactive computer vision explanations, including carefully selected default views, independent input controls, and constrained output spaces.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Computer vision**.

Keywords

Explanations, Interpretability, Human-Centered AI, Explainable AI (XAI), Computer Vision

*Authors contributed equally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI EA '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1395-8/2025/04

<https://doi.org/10.1145/3706599.3719730>

ACM Reference Format:

Indu Panigrahi, Sunnie S. Y. Kim, Amna Liaqat, Rohan Jinturkar, Olga Russakovsky, Ruth Fong, and Parastoo Abtahi. 2025. Interactivity x Explainability: Toward Understanding How Interactivity Can Improve Computer Vision Explanations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3706599.3719730>

1 Introduction

AI explanations help users understand and diagnose complex AI models, such as modern computer vision (CV) models processing thousands of pixels [1, 5]. While the explainable AI (XAI) community has developed many explanation methods, they often present information in a static form which can be challenging for users to interpret [10, 38, 42, 47, 69, 91]. Moreover, a single static explanation cannot meet the diverse needs of users with varying expertise [49].

Most XAI methods produce static CV explanations [14, 16, 27, 45, 67, 70, 71, 73, 74, 76, 77, 93, 95, 97]. However, recent studies have identified several challenges with static explanations [10, 38, 42, 47, 69, 91] which we categorize into three groups:

- (1) **Information overload:** Users feel that there is too much information to process at once [10, 38, 42, 47, 69, 91].
- (2) **Semantic-pixel gap:** Users find it difficult to connect image pixels to objects, attributes, and abstract concepts [10, 42, 69].
- (3) **Limited exploration:** Users lack the necessary tools to probe the explanation to deepen their understanding [42].

Prior work in Information Visualization has shown that interacting with data helps with interpretation [30, 36, 44, 92], and consequently, several works have called for interactivity in XAI applications [30, 48, 65, 80, 83, 84]. However, interactivity has mostly been explored at the dataset level [7, 11, 22, 26, 30, 89, 92], for visualizing model weights [15, 22, 31, 35, 90], and for non-visual data [8, 9, 50, 80]. In this work, we explore interactivity *at the explanation level* for images. Specifically, we investigate two research questions: **RQ1:** How do end-users **leverage interactivity** to help understand information conveyed by CV explanations? **RQ2:** How do end-users **perceive interactive CV explanations**?

To answer these questions, we conducted a within-subjects study using explanations for a bird identification model [88]. We selected bird identification over common object or scene classification use cases [20, 96] to examine the effects of domain knowledge. We focused on three widely used explanation types (heatmap-based, concept-based, and prototype-based explanations) and investigated three interactive mechanisms: **Filtering** to control the amount of information, **Overlays** to connect pixel-level landmarks to semantic labels, and **Counterfactuals** to allow users to edit images and observe a change in the explanation. We recruited 24 participants with varying levels of machine learning (ML) and birding expertise for an ~ 90-minute study. Each participant completed a set of tasks, alongside a questionnaire and interview.

We found that interactivity addresses some limitations of static CV explanations by providing users with tools to bridge the semantic-pixel gap, rapidly pinpoint the information they seek, explore the CV model beyond the immediate task, and gain clarity around the presentation of the static explanation. However, some interactive mechanisms were at times overwhelming for users and introduced new challenges, which informed the design recommendations we present in this paper. Our exploration and study findings contribute to our understanding of interactive mechanisms for XAI and bring us closer to the design of effective interactive CV explanations.

2 Related Work

Given the increasing ubiquity of AI [32, 33, 53, 57, 79], there exists a growing need for end-users to understand a model's behavior so they can appropriately trust and use the model. To address this need, many XAI methods have been developed over the past decade across various subfields of AI, including CV and robotics [17, 51, 58, 68, 71], and across different domains, such as healthcare and climate science [2, 3, 19, 52, 54–56, 78, 81]. Historically, much work has focused on *developing methods* to explain a model's outputs (i.e., generating an answer for “*why did the model make this prediction?*”). Recently, research has emerged at the intersection of XAI and Human-Computer Interaction (HCI) on *evaluating how useful* explanation methods are for users with varying machine learning (ML) expertise [34, 40, 62, 64], which we build on for CV explanations. In this section, we highlight literature most related to our focus on interactive CV explanations.

2.1 Computer Vision Explanations

In this work, we focus on *attribution-based* explanations, which compute a measure of importance by identifying the parts of an input that are critical for a model's decision [28, 59] and have been the most developed in XAI research thus far. We discuss three types of attribution-based explanations. *Heatmap-based explanations* typically generate a temperature heatmap that indicates the importance of each pixel or image region for the model's prediction [14, 27, 67, 71, 73, 74, 77, 93, 95]. *Concept-based explanations* explain a model's output by assigning a numerical importance scores to semantic concepts (e.g., +2.7 for “wings”) [45, 70, 76, 97]. *Prototype-based explanations* learn a set of important image regions (i.e., prototypes) from training images [16, 21, 37, 63, 66, 72].

Computer vision explanations have primarily been static. However, several studies have highlighted three main issues with static

explanations. The first is *information overload*. Prior work has found that the amount of information in explanations can be overwhelming. With heatmaps, users often want to view the “extremes” of a heatmap and find intermediate colors distracting [10, 47, 91]. For concepts, users report feeling overwhelmed by the number of items [42, 69], often the number of annotated objects in a dataset (e.g., 1197 in the Broden dataset [6]). Similarly, prototypes can require many regions to explain a model well enough [38]. The second is a *semantic-pixel gap*. Prior work suggests users often find it difficult to connect image regions to the represented object and vice versa. For concepts, users have difficulty identifying the concepts in the images [69]. With heatmaps and prototypes, users struggle to discern what objects are in the highlighted regions [10, 42]. The third is a *limited means for user-driven exploration*. Although attribution-based explanations show *what* regions are important, they do not explain *why* they are important [23, 59, 87]. Indeed, Kim et al. [42] found that users wanted to know why certain regions were deemed important by exploring the underlying causal relationships.

2.2 Interactivity and Explainable AI

While the problems presented could be partially mitigated by adjusting the amount of information in the static explanation on a case-by-case basis, interactivity has been identified as an effective [44, 82–84] and customizable [82, 84] way for users to interact with and understand data [11, 92]. Much of the existing work in interactivity for XAI has largely implemented interfaces that allow users to explore images at the dataset level by filtering through data points based on model predictions or their calculated similarity [7, 11, 22, 26, 30, 89]. Other work provides tools to examine the inner workings of models (e.g., model weights and learned features) [15, 22, 24, 29, 31, 35, 90]. There is limited research on interactivity at the explanation level for a single image. Some related work includes interactivity for textual or tabular data [8, 9, 50, 80]. For example, Liu et al. [50] explore interactive explanations that rely on direct manipulation, such as sliders and drop-down menus. However, they focus on comparing how human-AI teams with interactive explanations perform compared to AI-only agents when identifying out-of-distribution points. Within computer vision, Nguyen et al. [65] develop an interface for human-AI teams where users provide feedback on explanations. In contrast, our work seeks to investigate how users leverage and perceive interactive CV explanations and to provide design recommendations.

3 Explorative User Study

In this section, we describe the details of our IRB-approved study.

3.1 Study Method

3.1.1 Terminology. We tested three *explanation types*: heatmap-based, concept-based, and prototype-based (i.e., the rows in fig. 1). We created mock-up explanations, allowing us to carefully control the *presentation* of explanations across participants. For each explanation type, we created four *presentation types*: Static, Filtering, Overlays, and Counterfactuals (i.e., the columns in fig. 1). The latter three were interactive mechanisms inspired by the problems outlined in Section 2.1. An *explanation* is a pairing of one explanation type and one presentation type (i.e., one cell in fig. 1).

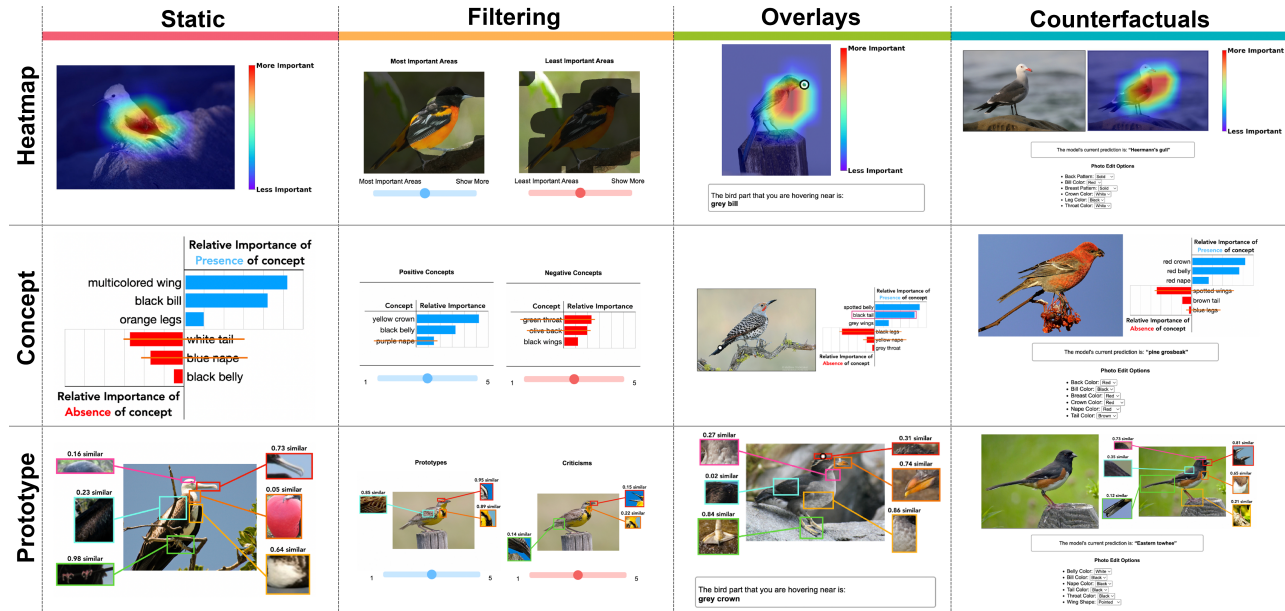


Figure 1: 12 explanation mock-ups for 3 explanation types (rows) and 4 presentation types (columns). All bird images were from the Caltech-UCSD Birds-200-2011 dataset [88].

3.1.2 *Setup.* For the **Static** baseline (fig. 1, 1st col), we based our mock-ups on prior works [16, 42, 73]. We then adapted them so that they shared some uniform characteristics across presentation types. First, we wanted all explanations to have a visual component, so we used bar graphs instead of numerical scores in the concept version (heatmap and prototype versions already had visual components), similar to TCAV [39]. This design decision was also in line with prior work, which reported that users dislike or feel overwhelmed by the numerical coefficients [42]. Second, we wanted explanations to describe an individual image (i.e., local explanation), not an entire output class (i.e., global explanation) which in our case was a bird species. Concept-based explanations are typically global, as concepts may not appear in every image. Thus, we made local concept-based explanations by crossing out concepts not present via strikethrough [42].

Filtering (fig. 1, 2nd col) is a type of interactivity that allows the user to conditionally remove information [25, 36, 75, 92]. We utilized sliders that hide or show components of the explanation, conditioned on their assigned importance [50]. In the concept and prototype versions, the sliders remove or add concepts and prototypes. In the heatmap version, we replaced the color gradient with a binary mask that grows or shrinks as the slider is moved. We chose this design because layering the mask on top of the color gradient makes the underlying image difficult to see [42].

Overlays (fig. 1, 3rd col), described by Shneiderman as “details-on-demand” [75], are extra layers of information [46] that appear only when the user indicates that they want to view them [86]. We incorporate a hover-over that supplies the pixel locations or the semantic labels of bird parts in the explanation, depending on which is missing. When users hover over a heatmap or prototype version, a dot appears on the closest bird part and a dialog box below the explanation shows a textual label for that part. For the

concept version, when users hover over a concept in the bar chart that is visible in the image (i.e., not crossed out), a dot representing the concept’s location appears on the image.

Counterfactuals (fig. 1, 4th col) explain model predictions by allowing the user to explore how changes to model inputs affect model outputs [60]. Typically, static counterfactual explanations show an edited image with the smallest modification needed to change the model’s output prediction from one class to another (e.g., how does an image need to change so that a model predicts “Cardinal” instead of “Blue jay”?) [28, 85]. We utilize counterfactuals as a form of interactivity by allowing users to edit an image and see how the explanation and prediction subsequently change. Users are provided with six drop-down menus that we refer to as “edit options”. Each menu corresponds to a bird part and provides two options for the attribute of that bird part: one is the original and the other is a randomly selected alternative. For example, if the bird has a red wing and the selected alternative is a blue wing, clicking the menu for “wing” will display the options “red” and “blue”. If a user selects the alternative attribute, the image updates to show the edit, and the explanation updates when the edited image changes it. Additionally, we include a dialog box that displays the model’s prediction for the current image; this also updates if the edited image changes the model’s prediction.

For each explanation, we chose a different image to mitigate learning effects. We used a series of preprocessing steps and consulted with two birding experts to select 12 images of similar difficulty in terms of bird identification and image quality. We performed these checks to ensure some level of consistency across images in terms of lighting, camera angle, or other factors. Additionally, two CV experts (two authors) checked that the explanations were of similar difficulty.

3.1.3 Participants. We designed a within-subjects study and recruited 24 participants with varying levels of ML and domain expertise (see Appendix A). We advertised our study through various platforms, including several institutions' email lists, birding labs, online conservation and birding forums, X, and Mastodon. Due to our snowball sampling process, all participants held an academic or research affiliation. Studies were conducted over Zoom video calls and lasted 90–100 minutes. All participants received a \$25 gift card.

3.1.4 Procedure. Each participant tried all 12 explanations in Figure 1. We formed a balanced Latin square across the 3 explanation types and randomized the presentation type ordering so that we had 4 participants per ordering, one from each expertise category. For each explanation, the participants explored and interacted with the explanation mock-up in a web application and then provided written answers to the survey questions (see Appendix B, images of the interface are included in our supplementary material.). Specifically, participants were first asked to read the instructions and then view and interact with the explanation. The instructions described the explanation type and how to use the interactive mechanism when it was present. The participant could choose when to proceed to the survey, though we enforced a time limit of roughly two minutes. The survey questions asked participants to identify the most and least important parts of a bird based on the explanation; these tasks required participants to actively engage with each explanation. All task questions included a reference to a bird part guide that we created to assist with birding terminology. At the end of the study, participants rated their level of agreement with a set of statements and ranked the presentation types in terms of their general preference and preference in the context of learning about a new bird species, allowing for ties.

3.1.5 Analysis. We collected qualitative data from the study recordings and quantitative data from the survey questions. For our qualitative data, we created a codebook from the audio and video recordings of the studies and performed a Reflexive Thematic Analysis [12, 13] to extract findings and outline recommendations. Some codes directly came from verbal responses (e.g., “User prefers counterfactuals for learning because exploration”), while other codes described participant interactions (e.g., “User interacted with each of the edit options individually”). Two authors created an initial codebook based on four studies; one of those authors coded the remaining studies. All authors iterated on the codebook. Since one author coded all the studies and refined the codebook in agreement with the other authors, we did not calculate inter-rater reliability.

3.2 Study Results

We present the quantitative results from the surveys followed by the themes from our qualitative analysis of the interviews, summarize the takeaways, and offer design recommendations.

3.2.1 Questionnaire Results. After viewing all 12 explanations, participants rated a set of statements (see Appendix B) on a 5-point Likert scale and ranked the four presentation types. Lower numbers indicate better ratings (fig. 2). For the analysis, we used a two-sided Wilcoxon signed-rank test with Holm correction; we used Holm to avoid assuming independence. Participants rated Counterfactuals as having significantly more information compared to Static,

Filtering, and Overlays ($p < .01$). Participants rated Filtering as having significantly more information than Static ($p = .015$). 4 participants noted that while Filtering provided more information than Static, the additional information was not overwhelming. For connecting pixel-level information to semantic-level information, participants rated Overlays as significantly less difficult than Static, Filtering, and Counterfactuals ($p < .01$). There were no significant differences in ratings for the model understanding statement. In general, participants preferred Overlays significantly more than Static ($p < .01$) and Counterfactuals ($p = .02$). For learning about bird species, participants preferred Overlays significantly more than Static presentations ($p < .01$).

3.2.2 Qualitative Results. Here, we discuss the qualitative themes we developed. Parentheses indicate the number of participants.

Participants appreciate interactive mechanisms that augment the explanation without changing the underlying explanation. Participants appreciated Filtering and Overlays as tools for adjusting the amount of detail in the explanation without changing the explanation itself. Specifically, participants mentioned that Filtering made explanations less overwhelming by giving them control over the amount of visible information (12) and this control in turn made explanations easier to digest (4). Participants also used this interactive mechanism to reduce the amount of information (8) or gradually add information (7) during the task. Similarly, participants liked Overlays (18) despite the additional information presented. For example, P2 said: “[Overlays have] more information than [Static], but I feel like the way they’re presented makes it so that you can take it in easier, and it doesn’t feel like more information”. On the other hand, participants found Counterfactuals overwhelming (12), as image edits altered the underlying explanations.

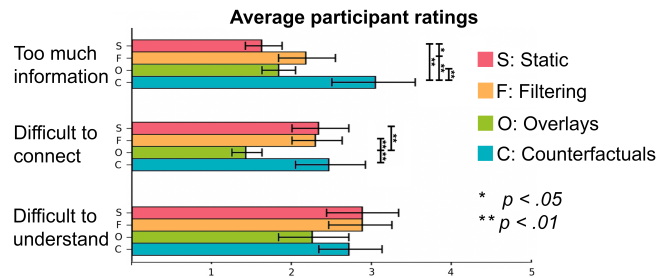


Figure 2: Average participant ratings for survey statements. Lower is better. Error bars are 95% confidence intervals.

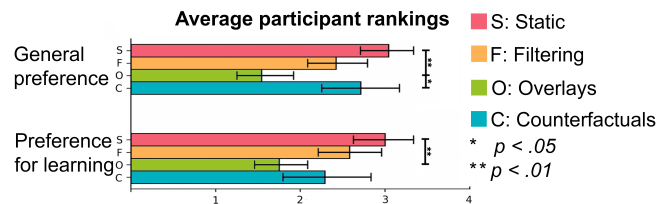


Figure 3: Average participant rankings for general preference and preference for learning bird species. Ties are allowed. Lower is better. Error bars are 95% confidence intervals.

Participants find interactive mechanisms that alter the underlying explanation overwhelming. Participants found that Counterfactuals were overwhelming because there could be simultaneous changes in the explanation when the image was modified. Additionally, they struggled to untangle the causal relationships because the inputs were interdependent (11). P6 commented that “*In my mind, it’s hard to reason about the combination of influences*”, and P2 stated for the prototypes version that they “*think there’s too much going on with the photo editing and the [prototypes] popping out*”. In fact, participants often performed one edit at a time, making sure to revert each edit option to the default choice before trying another (61 out of 72 Counterfactuals trials). Moreover, participants had to explore a large number of possible explanations that resulted from the 64 possible edited images (9). P7 remarked: “*There were so many things to look at—it was confusing*”. Overall, participants found Counterfactuals confusing and hard to interpret (12), particularly those with low ML and domain expertise (5), and needed more time to explore the space of possible inputs and the resulting outputs (6).

Although participants find Counterfactuals overwhelming, they utilize them to resolve confusion around static presentations by inducing systematic changes in model predictions and explanations. Participants leveraged Counterfactuals to clarify aspects of static explanations they found confusing and to confirm their understanding of the explanations. For example, many participants were confused about static concept-based explanations (16), and some used Counterfactuals to edit the image and observe the resulting changes in the model’s prediction and explanation to clarify their understanding, which ultimately led to a more accurate interpretation of the static explanation (6). P17 noted: “*It was helpful to be able to change the color of [a bird part] and see how that affected the bar graph...that sort of helped me understand what the [strikethrough] is for*”. Additionally, the visual changes helped participants identify the bird parts (15). Hohman et al. observed similar behaviors where “*...participants used Counterfactuals often throughout their exploration, both as a direct task and as a sanity check for feature sensitivity*” in their study with ML practitioners [30].

Participants leverage Counterfactuals to explore a range of explanations to better understand the AI model more broadly and beyond the task at hand. Many participants (20) found Counterfactuals useful for understanding the AI model’s reasoning at a high level. Some participants expressed a desire to use Counterfactuals to explore other birds outside of the one image given and to build a macroscopic understanding of the AI model. Specifically, participants appreciated the ability to change the bird species by editing the image (6). This ability to explore allowed participants to develop an understanding of the model more broadly. For example, P15 thought it helped them build rules about how changes in the model’s input affect its output, and P18 noted that this mechanism would allow them to identify the decision points of the model. These uses were beyond the scope of the tasks which only asked for the most and least important bird parts in the given image. This suggests that participants saw Counterfactuals as a source of information to explore and expand their understanding of the model, even if they were not asked to do so.

Participants felt that Filtering and Overlays allowed them to quickly focus on information of interest. Some participants

appreciated that Filtering automatically sorted the bird parts by importance (e.g., similarity scores sorted by magnitude in prototype-based explanations) because it obviated the need for manual comparison (6). For instance, P6 liked the prototype version of Filtering because “*it produces a ranking of the similarity scores that I don’t need to think about myself*”. Similarly, participants found Overlays useful for easily and quickly identifying the names of the bird parts, even with the bird part guide provided (8). For example, P21 remarked that Overlays were “*faster because I could hover over it and it could tell me what the parts were*”. Thus, participants felt that the automatic sorting and labeling offered by Filtering and Overlays respectively enabled them to efficiently hone in on the information that they were seeking. Participants noted that the lack of these features in Static presentations was inconvenient (5).

3.2.3 Summary. Participants found that interactive features augmenting explanations, such as Filtering and Overlays, helped them understand information without being overwhelming. While Counterfactuals introduced significant visual changes by altering the underlying explanations, participants used them to clarify confusing aspects of static explanations, reaffirm their understanding, and explore the AI model more broadly. With Filtering and Overlays, participants felt they could quickly focus on relevant information, and they appreciated the flexibility that these mechanisms offered when interpreting explanations.

3.2.4 Design Recommendations. Participants expressed that interactive mechanisms were at times confusing or inconvenient. Based on these findings, we recommend the following design considerations for future work on interactive CV explanations.

Avoid interdependent input controls. Participants found Counterfactuals confusing because they could not disentangle the effects of multiple simultaneous edits on the prediction or the explanation. A more appropriate design might involve preventing users from performing more than one edit at a time, thereby clarifying the causal relationship between each input and the corresponding changes.

Constrain the input and output space. With Counterfactuals, many participants felt overwhelmed by the number of possible image edits and the extent of changes to the explanation. We suggest limiting the input and output space to create a manageable range of information, tailored to the application, allowing participants to easily explore all.

Design an optimal static default view. We recommend that users should have the option to choose whether or not to engage with interactive mechanisms. Therefore, the default static presentation should provide enough information for the initial interpretation of the explanation without overwhelming users [69].

4 Limitations

Our study is a preliminary step toward evaluating interactive CV explanations and thus has several limitations. For example, our study sessions were relatively short, with participants only having two minutes to explore each explanation. Some participants expressed a desire for more time, suggesting the need for future studies to explore deeper engagement with interactive mechanisms for CV explanations. Additionally, longitudinal studies are needed

to understand how users interact across repeated sessions and the role of interactive mechanisms that maintain user history.

We chose a simple bird identification task, which was a reasonable starting point, because it was consistent with prior work [42, 43, 61] and allowed us to recruit participants with low and high domain expertise. However, it is unclear to what extent our findings can be generalized to more complex, high-risk, and high-impact applications such as medical diagnostics. Moreover, our sample size was relatively small, underscoring the need for larger-scale studies. Lastly, our interactive mechanisms were not directly comparable to the static explanations. For example, the Filtering mechanism for heatmap-based explanations lacked the color gradient present in its static counterpart. Carefully controlled experiments are needed for quantitative analysis of performance, trust, and confidence in tasks using various interactive computer vision explanations.

5 Future Work

Future research should explore a broader range of interactive mechanisms for CV explanations beyond the three examined in this work. Additionally, some participants suggested integrating information beyond the image, such as geographical context or general facts, to improve the utility of explanations. Further work is needed to define the boundaries of what information should be included in interactive explanations across different application domains. Collaborating with domain experts and stakeholders through participatory design will be critical for advancing these directions.

Finally, prior work has shown that explanations may lead to misplaced trust in AI systems [4, 18, 40, 41, 94]. Interactivity in explanations could amplify these risks, further misleading users and enabling confirmation bias. For example, users might iteratively interact with Counterfactuals until the explanations appear to align with their preconceptions. More research is needed to carefully examine these possibilities and mitigate potential negative outcomes.

6 Conclusion

Most computer vision explanations are static, limiting user interaction. Drawing from prior XAI research, we identified three key issues with these explanations: information overload, a semantic-pixel gap, and limited means for user-driven exploration. Given the potential of interactivity in XAI, we explored how users engage with and perceive interactive explanations by comparing three mechanisms across three explanation types. We conducted a study with 24 participants of varying ML and domain expertise. Our findings showed that interactivity helps address the challenges of static explanations by offering users greater control and quicker access to relevant information. While some mechanisms were overwhelming, they allowed users to expand their understanding of the model and refine their grasp of the static explanation itself. Based on these insights, we offered design recommendations for interactive computer vision explanations, including optimized default views, independent input controls, and constrained interaction spaces.

Acknowledgments

Firstly, we thank our participants for taking the time to participate in our study. We also thank Patrick Newcombe and Fengyi Guo for their help in identifying bird species and images of similar difficulty

for the study. Lastly, we thank members of Princeton HCI and the Princeton Visual AI Lab, especially Vikram Ramaswamy, for providing feedback and resources throughout this project.

This material is based upon work supported by the National Science Foundation under grant No. 2145198 (OR) and the Graduate Research Fellowship (SK), as well as the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), the Princeton SEAS Innovation Fund (RF), and Open Philanthropy (RF, OR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052
- [2] Anna Markella Antoniadou, Yuhang Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. 2021. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences* 11, 11 (2021). doi:10.3390/app11115088
- [3] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2021. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *CoRR abs/2112.11561* (2021). arXiv:2112.11561 <https://arxiv.org/abs/2112.11561>
- [4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717
- [5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénézet, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, C (jun 2020), 82–115. doi:10.1016/j.inffus.2019.12.012
- [6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [7] Alex Bäuerle, Ángel Alexander Cabrera, Fred Hohman, Megan Maher, David Koski, Xavier Suau, Titus Barik, and Dominik Moritz. 2022. Symphony: Composing Interactive Interfaces for Machine Learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 210, 14 pages. doi:10.1145/3491102.3502102
- [8] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 204–219. doi:10.1145/3581641.3584075
- [9] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 314, 27 pages. doi:10.1145/3613904.3642106
- [10] Pieter Blijnaut. 2010. Visual span and other parameters for the generation of heatmaps. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (Austin, Texas) (ETRA '10). Association for Computing Machinery, New York, NY, USA, 125–128. doi:10.1145/1743666.1743697
- [11] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. 2022. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 746–766. doi:10.1145/3490099.3511122
- [12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa
- [13] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.

- 1080/2159676X.2019.1628806
- [14] Wieland Brendel and Matthias Bethge. 2019. Approximating CNNs with Bag-of-local-Features Models Works Surprisingly Well on ImageNet. In *International Conference on Learning Representations (ICLR)*.
 - [15] Shuo Chang, Peng Dai, Lichan Hong, Cheng Sheng, Tianjiao Zhang, and Ed H. Chi. 2016. AppGrouper: Knowledge-based Interactive Clustering Tool for App Search Results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) (IUI '16). Association for Computing Machinery, New York, NY, USA, 348–358. doi:10.1145/2856767.2856783
 - [16] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Neural Information Processing Systems (NeurIPS)*.
 - [17] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Kam-Fai Wong, Kevin Knight, and Hua Wu (Eds.). Association for Computational Linguistics, Suzhou, China, 447–459. <https://aclanthology.org/2020.aacl-main.46>
 - [18] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. doi:10.1145/3544548.3580672
 - [19] Alex J DeGrave, Zhuo Ran Cai, Joseph D Janizek, Roxana Daneshjoo, and Su-In Lee. 2023. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nat. Biomed. Eng.* (Dec. 2023).
 - [20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848
 - [21] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. 2022. Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
 - [22] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. doi:10.1145/3185517
 - [23] Nadia El Bekri, Jasmin Kling, and Marco F. Huber. 2020. A Study on Trust in Black Box Models and Post-hoc Explanations. In *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, Francisco Martínez Álvarez, Alicia Troncoso Lora, José António Sáez Muñoz, Héctor Quintián, and Emilio Corchado (Eds.). Springer International Publishing, Cham, 35–46.
 - [24] K. J. Kevin Feng and David W. McDonald. 2023. Addressing UX Practitioners' Challenges in Designing ML Applications: an Interactive Machine Learning Approach. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 337–352. doi:10.1145/3581641.3584064
 - [25] Ana Figueiras. 2015. Towards the Understanding of Interaction in Information Visualization. In *2015 19th International Conference on Information Visualisation*. 140–147. doi:10.1109/iV.2015.34
 - [26] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlick: interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) (CHI '08). Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/1357054.1357061
 - [27] Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *International Conference on Computer Vision (ICCV)*.
 - [28] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2376–2384. <https://proceedings.mlr.press/v97/goyal19a.html>
 - [29] Lijie Guo, Elizabeth M. Daly, Ozgur Alkan, Massimiliano Mattetti, Owen Corne, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 537–548. doi:10.1145/3490099.3511111
 - [30] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300809
 - [31] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Chau. 2020. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2020). doi:10.1109/TVCG.2019.2934659
 - [32] Z. Huang, F. Bianchi, M. Yuksekogonul, T. J. Montine, and J. Zou. 2023. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med* 29, 9 (Sep 2023), 2307–2316.
 - [33] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. 2020. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends® in Computer Graphics and Vision* 12, 1–3 (2020), 1–308. doi:10.1561/06000000079
 - [34] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2022. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 11945–11953. doi:10.1609/aaai.v36i11.21452
 - [35] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 1343–1352. doi:10.1145/1753326.1753529
 - [36] D.A. Keim. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (2002), 1–8. doi:10.1109/2945.981847
 - [37] Eoin M. Kenny, Mycal Tucker, and Julie Shah. 2023. Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=hWwY_Jq0xsN
 - [38] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf
 - [39] 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 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2668–2677. <https://proceedings.mlr.press/v80/kim18d.html>
 - [40] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *European Conference on Computer Vision (ECCV)*.
 - [41] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *ACM Conference on Human Factors in Computing Systems (CHI)*.
 - [42] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages. doi:10.1145/3544548.3581001
 - [43] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
 - [44] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 1375–1386. doi:10.1145/3025453.3025592
 - [45] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *International Conference on Machine Learning (ICML)*.
 - [46] Nicholas Kong and Maneesh Agrawala. 2012. Graphical Overlays: Using Layered Elements to Aid Chart Reading. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2631–2638. doi:10.1109/TVCG.2012.229
 - [47] Simone Kriglstein, Günter Wallner, and Margit Pohl. 2014. A user study of different gameplay visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (CHI '14). Association for Computing Machinery, New York, NY, USA, 361–370. doi:10.1145/2556288.2557317
 - [48] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) (IUI '15). Association for Computing Machinery, New York, NY, USA, 126–137. doi:10.1145/2678025.2701399
 - [49] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020*

- CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590
- [50] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 408 (oct 2021), 45 pages. doi:10.1145/3479552
- [51] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.
- [52] Jörn Lötsch, Dario Kringsel, and Alfred Ultsch. 2022. Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics* 2, 1 (2022), 1–17. doi:10.3390/biomedinformatics2010001
- [53] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment Anything in Medical Images. *Nature Communications* 15 (2024), 1–9.
- [54] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI* 9 (2022), 100169. doi:10.1016/j.egyai.2022.100169
- [55] Antonios Mamelakis, Imme Ebert-Uphoff, and Elizabeth A. Barnes. 2022. *Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science*. Springer International Publishing, Cham, 315–339. doi:10.1007/978-3-031-04083-2_16
- [56] Quentin Meteier, Marine Capallera, Leonardo Angelini, Elena Mugellini, Omar Abou Khaled, Stefano Carrino, Emmanuel De Salis, Stéphane Galland, and Susanne Boll. 2019. Workshop on Explainable AI in Automated Driving: A User-Centered Interaction Approach. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings (Utrecht, Netherlands) (AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 32–37. doi:10.1145/3349263.3350762
- [57] Senne Michielssen, Adam Maloof, Joe Haumacher, Alexander Dreger, Kyle Bonicki, and Karl Hallgren. 2024. Using large language models to generate baseball spray charts in the absence of numerical data. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology* 0 (07/2024 2024), 1–8. doi:10.1177/17543371241257734
- [58] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2024. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Comput. Surv.* 56, 7, Article 168 (apr 2024), 36 pages. doi:10.1145/3616864
- [59] Tim Miller. 2021. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review* 36 (2021), e14. doi:10.1017/S0269888921000102
- [60] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. doi:10.1145/3593013.3594001
- [61] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (April 2024), 39 pages. doi:10.1145/3641022
- [62] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s, Article 295 (jul 2023), 42 pages. doi:10.1145/3583558
- [63] Meike Nauta, Ron van Bree, and Christin Seifert. 2021. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [64] Mobeen Nazar, Muhammad Mansoor Alam, Eiad Yafi, and Mazliham Mohd Su'ud. 2021. A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access* 9 (2021), 153316–153348. doi:10.1109/ACCESS.2021.3127881
- [65] Giang Nguyen, Mohammad Reza Taesiri, Sunnie S. Y. Kim, and Anh Nguyen. 2024. Allowing Humans to Interactively Guide Machines Where to Look Does Not Always Improve Human-AI Team's Classification Accuracy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 8169–8175.
- [66] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. In *Neural Information Processing Systems (NeurIPS)*.
- [67] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*.
- [68] Kun Qian, Marina Danilevsky, Yannis Katsis, Ban Kaws, Erick Oduor, Lucian Popa, and Yunyao Li. 2021. XNLP: A Living Survey for XAI Research in Natural Language Processing. In *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces (College Station, TX, USA) (IUI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 78–80. doi:10.1145/3397482.3450728
- [69] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. 2023. Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10932–10941.
- [70] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, and Olga Russakovsky. 2022. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. doi:10.48550/ARXIV.2206.07690
- [71] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778
- [72] Dawid Rymarczyk, Lukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor, and Bartosz Zieliński. 2022. Interpretable Image Classification with Differentiable Prototypes Assignment. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII (Tel Aviv, Israel)*. Springer-Verlag, Berlin, Heidelberg, 351–368. doi:10.1007/978-3-031-19775-8_21
- [73] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. doi:10.1109/ICCV.2017.74
- [74] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Neural Information Processing Systems (NeurIPS)*.
- [75] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. doi:10.1109/VL.1996.545307
- [76] Pushkar Shukla, Sushil Bharati, and Matthew Turk. 2023. CAVLI - Using Image Associations To Produce Local Concept-Based Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3750–3755.
- [77] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR) Workshops*.
- [78] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. 2020. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging* 6 (2020). doi:10.3390/jimaging6060052
- [79] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfoh, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkumar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (Aug. 2023), 172–180.
- [80] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* 5, 8 (2023), 873–883.
- [81] Helen Smith. 2021. Clinical AI: opacity, accountability, responsibility and liability. *AI & SOCIETY* 36, 2 (2021), 535–545. doi:10.1007/s00146-020-01019-6
- [82] Arjun Srinivasan, Steven M. Drucker, Alex Ender, and John Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 672–681. doi:10.1109/TVCG.2018.2865145
- [83] S. Shyam Sundar, Jeeyun Oh, Saraswathi Bellur, Haiyan Jia, and Hyang-Sook Kim. 2012. Interactivity as self-expression: a field experiment with customization and blogging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Austin, Texas, USA) (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 395–404. doi:10.1145/2207676.2207731
- [84] S. Shyam Sundar, Qian Xu, and Saraswathi Bellur. 2010. Designing interactivity in media interfaces: a communications perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Atlanta, Georgia, USA) (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2247–2256. doi:10.1145/1753326.1753666
- [85] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepeti Ghadiyaram. 2022. Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals. In *European Conference on Computer Vision (ECCV)*.
- [86] Laton Vermette, Shruti Dembla, April Y. Wang, Joanna McGrenere, and Parmit K. Chilana. 2017. Social CheatSheet: An Interactive Community-Curated Information Overlay for Web Applications. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 102 (dec 2017), 19 pages. doi:10.1145/3134737
- [87] S Wachter, B Mittelstadt, and C Russell. 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology* 31, 2 (2018), 841–887.
- [88] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001. California

- Institute of Technology.
- [89] Zijie J. Wang, Fred Hohman, and Duen Horng Chau. 2023. WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings. In *Demo, Annual Meeting of the Association for Computational Linguistics (ACL)*. <https://poloclub.github.io/wizmap/>
- [90] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. 2021. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1396–1406. doi:10.1109/TVCG.2020.3030418
- [91] Jinyue Xia and David C. Wilson. 2018. Instructor Perspectives on Comparative Heatmap Visualizations of Student Engagement with Lecture Video. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) (*SIGCSE '18*). Association for Computing Machinery, New York, NY, USA, 251–256. doi:10.1145/3159450.3159487
- [92] Ji Soo Yi, Youn ah Kang, John Stasko, and J.A. Jacko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1224–1231. doi:10.1109/TVCG.2007.70515
- [93] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*.
- [94] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.
- [95] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [96] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).
- [97] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision (ECCV)*.

Appendix

A Expertise Criteria and Distribution

We recruited 6 participants from each of the following 4 categories: high-ML and high-domain (HM-HD), high-ML and low-domain (HM-LD), low-ML and high-domain (LM-HD), and low-ML and low-domain (LM-LD). Participants rated their familiarity with ML and the task domain of birding on a 5-point scale, and were considered to have high or low expertise based on their responses:

- LM: “1: I don’t know anything about ML”, “2: I have heard about a few ML concepts or applications”, or “3: I know the basics of ML and can hold a short conversation about it.”
- HM: “4: I have taken a course on ML and/or have experience working with an ML system” or “5: I often use and study ML.”
- LD: “1: I don’t know anything about birding”, “2: I have heard about a few birding concepts”, or “3: I know the basics of birding and can hold a short conversation about it.”
- HD: “4: I have taken a course on birding and/or have experience in birding” or “5: I often conduct bird-watching and study birding.”

Specifically, the distribution of participants was as follows:

- LM-HD included P11, P12, P15, P16, P19, and P21.
- HM-HD included P10, P14, P17, P18, P23, and P24.
- HM-LD included P3, P4, P6, P8, P13, and P20.
- LM-LD included P1, P2, P5, P7, P9, and P22.

B Survey Questions

Q1. Rate your agreement with the following statements on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree):

- (1) It was difficult to read the explanation.

- (2) There was too much information provided.
- (3) It was difficult to connect the explanation to parts of the photo.

Q2. Based on the labels in the table, rank your preference of the following four groups of explanation designs (1 being most preferred and 4 being least preferred) and verbally explain your thought process. The labels were W, X, Y, and Z with each letter representing Static, Filtering, Overlays, and Counterfactuals respectively.

Q3. Let’s say that you needed to learn about a new bird species. Rank how useful the four groups of explanation designs would be for doing so (1 being most useful and 4 being least useful) and verbally explain your thought process. As with Q2, the labels were W, X, Y, and Z with each letter representing Static, Filtering, Overlays, and Counterfactuals respectively.