DExT

Padmanabhan, Deepan Chakravarthi; Plöger, Paul G.; Arriaga, Octavio; Valdenegro-Toro, Matias

Published in:
Proceedings of the 1st World Conference on eXplainable Artificial Intelligence

Publication date:
2023

DExT: Detector Explanation Toolkit

Deepan Chakravarthi Padmanabhan[1 [0000−0003−0638−014X]], Paul G. Plöger[1 [0000−0001−5563−5458]], Octavio Arriaga[2 [0000−0002−8099−2534]], and Matias Valdenegro-Toro[3 [0000−0001−5793−9498]]

[1] Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany
deepangrad@gmail.com
[2] University of Bremen, Bremen, Germany arriagac@uni-bremen.de
[3] University of Groningen, Groningen, The Netherlands m.a.valdenegro.toro@rug.nl

Abstract. State-of-the-art object detectors are treated as black boxes due to their highly non-linear internal computations. Even with unprecedented advancements in detector performance, the inability to explain how their outputs are generated limits their use in safety-critical applications. Previous work fails to produce explanations for both bounding box and classification decisions, and generally make individual explanations for various detectors. In this paper, we propose an open-source Detector Explanation Toolkit (DExT) which implements the proposed approach to generate a holistic explanation for all detector decisions using certain gradient-based explanation methods. We suggests various multi-object visualization methods to merge the explanations of multiple objects detected in an image as well as the corresponding detections in a single image. The quantitative evaluation show that the Single Shot MultiBox Detector (SSD) is more faithfully explained compared to other detectors regardless of the explanation methods. Both quantitative and human-centric evaluations identify that SmoothGrad with Guided Backpropagation (GBP) provides more trustworthy explanations among selected methods across all detectors. We expect that DExT will motivate practitioners to evaluate object detectors from the interpretability perspective by explaining both bounding box and classification decisions.

Keywords: Object detectors · Explainability · Quantitative evaluation · Human-centric evaluation · Saliency methods

1 Introduction

Object detection is imperative in applications such as autonomous driving [15], medical imaging [5], and text detection [18]. An object detector outputs bounding boxes to localize objects and categories for objects of interest in an input image. State-of-the-art detectors are deep convolutional neural networks [54] with high accuracy and fast processing compared to traditional detectors. However, convolutional detectors are considered black boxes [37] due to over-parameterization and hierarchically non-linear internal computations. This non-intuitive decision-making process restricts the capability to debug and improve detection systems.

The user trust in model predictions has decreased and consequently using detectors in safety-critical applications is limited. In addition, the process of verifying the model and developing secure systems is challenging [12] [52]. Numerous previous studies state interpreting detectors by explaining the model decision is crucial to earning the user's trust [48] [32] [40], estimating model accountability [20], and developing secure object detector systems [12] [52].

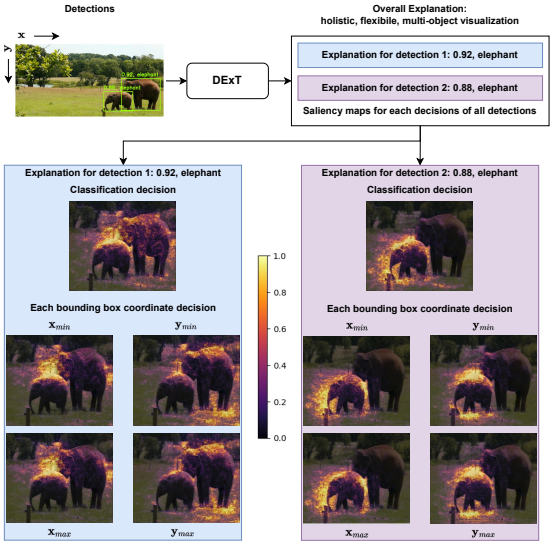


Fig. 1. A depiction of the proposed approach to interpret all object detector decisions. The corresponding explanations are provided in the same colored boxes. This breakdown of explanations offers more flexibility to analyze decisions and serves as a holistic explanation for all the detections.

With a range of users utilizing detectors for safety critical applications, providing humanly understandable explanations for the category and each bounding box coordinate predictions together is essential. In addition, as object detectors are prone to failures due to non-local effects [30], the visualization techniques for detector explanations should integrate explanations for multiple objects in a single image at the same time. Previous saliency map-based methods explaining detectors [26] [46] [17] focus on classification or localization decisions individually, not both at the same time.

In this paper, we consider three deficits in the literature: methods to explain each category and bounding box coordinate decision made by an object detector, visualizing explanations of multiple bounding boxes into the same output explanation image, and a software toolkit integrating the previously mentioned aspects.

This work concentrates on providing individual humanly understandable explanations for the bounding box and classification decisions made by an object detector for any particular detection, using gradient-based saliency maps. Figure 1 provides an illustration of the proposed solution by considering the complete output information to generate explanations for the detector decision.

Explanations for all the decisions can be summarized by merging the saliency maps to achieve a high-level analysis and increasing flexibility to analyze detector decisions, improving improving model transparency and trustworthiness. We suggest methods to combine and visualize explanations of different bounding boxes in a single output explanation image as well as an approach to analyze the detector errors using explanations.

This work contributes DExT, software toolkit, to explain each decisions (bounding box regression and object classification jointly), evaluate explanations, and identify errors made by an object detector. A simple approach to extend gradient-based explanation methods to explain bounding box and classification decisions of an object detector. An approach to identify reasons for the detector failure using explanation methods. Multi-object visualization methods to summarize explanations for all output detections in a single output explanation. And an evaluation of gradient-based saliency maps for object detector explanations, including quantitative results and a human user study.

We believe our work reveals some major conclusions about object detector explainability. Overall quantitative metrics do not indicate that a particular object detector is more interpretable, but visual inspection of explanations indicates that recent detectors like EfficientDet seem to be better explained using gradient-based methods than older detectors (like SSD or Faster R-CNN, shown in Figure 2), based on lack of artifacts on their heatmaps. Detector backbone has a large impact on explanation quality (Figure 6).

The user study (Section 4.4) reveals that humans clearly prefer the convex polygon representation, and Smooth Guided Backpropagation provides the best detector explanations, which is consistent with quantitative metrics. We believe these results are important for practitioners and researchers of object detection interpretability. The overall message is to explain both object classification and bounding box decisions and it is possible to combine all explanations into a single image using the convex polygon representation of the heatmap pixels. The appendix of this paper is available at https://arxiv.org/abs/2212.11409.

2 Related Work

Interpretability is relatively underexplored in detectors compared to classifiers. There are post hoc [26] [46] [17] and intrinsic [21] [51] detector interpretability approaches. Detector Randomized Input Sampling for Explanation (D-RISE) [26] in a model-agnostic manner generates explanations for the complete detector output. However, saliency map quality depends on the computation budget, the method is time consuming, and individual explanations for bounding boxes are not evaluated. Contrastive Relevance Propagation (CRP) [46] extends Layer-wise Relevance Propagation (LRP) [7] to explain individually the bounding box and classification decisions of Single Shot MultiBox Detector (SSD). This procedure includes propagation rules specific to SSD. Explain to fix (E2X) [17] contributes a framework to explain the SSD detections by approximating SHAP [24] feature importance values using Integrated Gradients (IG), Local Interpretable Model-agnostic Explanations (LIME), and Probability Difference Analysis (PDA) explanation methods. E2X identifies the detection failure such as false negative errors using the explanations generated. The individual explanations for bounding box decisions and classification decisions are unavailable.

The intrinsic approaches majorly focus on developing detectors that are inherently interpretable. Even though the explanations are provided for free,

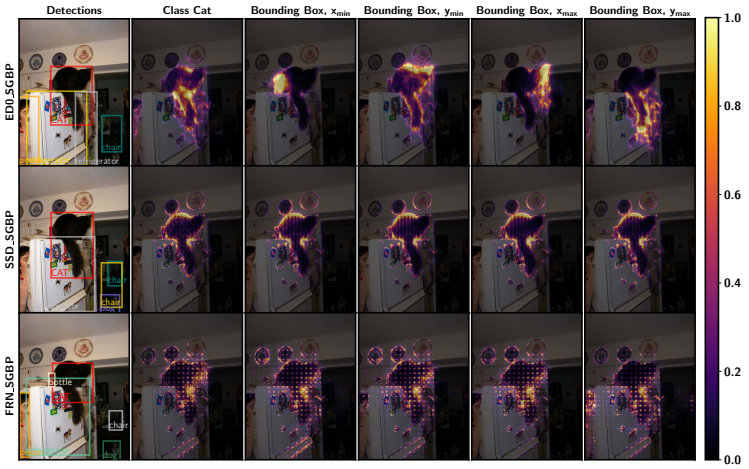


Fig. 2. Comparison of the classification and all bounding box coordinate explanations corresponding to the cat detection (red-colored box) across different detectors using SGBP is provided. The bounding box explanations from EfficientDet-D0 illustrate the visual correspondence to the respective bounding box coordinates. The explanations from Faster R-CNN illustrate a sharp checkerboard pattern.

currently, most of the methods are model-specific, do not provide any evaluations on the explanations generated, and includes complex additional designs.

Certain attention-based models such as DEtector TRansformer (DETR) [10] and detectors using non-local neural networks [49] offer attention maps improving model transparency. A few previous works with attention reveal contradicting notions of using attention for interpreting model decisions. [35] and [19] illustrate attention maps are not a reliable indicator of important input region as well as attention maps are not explanations, respectively. [8] have revealed saliency methods provide better explanations over attention modules.

We select the post hoc gradient-based explanation methods because they provide better model translucency, computational efficiency, do not affect model performance, and utilize the gradients in DNNs. Finally, saliency methods are widely studied in explaining DNN-based models [3]. A detailed evaluation of various detectors reporting robustness, accuracy, speed, inference time as well as energy consumption across multiple domains has been carried out by [4]. In this work, the authors compare detectors from the perspective of explainability.

3 Proposed Approach

3.1 Explaining Object Detectors

This work explains various detectors using gradient-based explanation methods as well as evaluate different explanations for bounding box and classification decisions. The selected detectors are: SSD512 (SSD) [23], Faster R-CNN (FRN)

[28], and EfficientDet-D0 (ED0) [43]. The short-form tags are provided in the bracket. SSD512 and Faster R-CNN are widely used single-stage and two-stage approaches, respectively. Explaining the traditional detectors will aid in extending the explanation procedure to numerous similar types of recent detectors. EfficientDet is a relatively recent state-of-the-art single-stage detector with higher accuracy and efficiency. It incorporates a multi-scale feature fusion layer called a Bi-directional Feature Pyramid Network (BiFPN). EfficientDet-D0 is selected to match the input size of SSD512. The variety of detectors selected aids in evaluating the explanation methods across different feature extractors such as VGG16 (SSD512), ResNet101 (Faster R-CNN), and EfficientNet (EfficientDet-D0). The gradient-based explanation methods selected in this work to explain detectors are: Guided Backpropagation (GBP) [41], Integrated Gradients (IG) [42], SmoothGrad [39] + GBP (SGBP), and SmoothGrad + IG (SIG). GBP produces relatively less noisy saliency maps by obstructing the backward negative gradient flow through a ReLU. For instance, an uncertainty estimate of the most important pixels influencing the model decisions is carried out using GBP and certain uncertainty estimation methods [50]. This combines uncertainty estimation and interpretability to better understand DNN model decisions. IG satisfies the implementation and sensitivity invariance axioms that are failed by various other state-of-the-art interpretation methods. SmoothGrad aids in sharpening the saliency map generated by any interpretation method and improves the explanation quality. These four explanation methods explain a particular detector decision by computing the gradient of the predicted value at the output target neuron with respect to the input image.

The object detector decisions for a particular detection are bounding box coordinates ($x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$), and class probabilities $(c_1, c_2, ..., c_k)$, where k is the total number of classes predicted by the detector. Usually these are output by heads at the last layer of the object detector. The classification head is denoted as $\text{model}_{\text{cls}}(x)$, while the bounding box regression head is $\text{model}_{\text{bbox}}(x)$. Considering that an explanation method computes a function $\text{expl}(x, \hat{y})$ of the input x and scalar output prediction $\hat{y}$ (which is one output layer neuron), then a classification explanation e_{cls} is:

$$\hat{c} = \text{model}_{\text{cls}}(x) \quad k = \arg\max_i \hat{c}_i \quad e_{\text{cls}} = \text{expl}\left(x, \hat{l}_k\right) \tag{1}$$

A bounding box explanation consists of four different explanations, one for each bounding box component $e_{x_{\min}}, e_{y_{\min}}, e_{x_{\max}}, e_{y_{\max}}$:

$$\hat{x}_{\min}, \hat{y}_{\min}, \hat{x}_{\max}, \hat{y}_{\max} = \text{model}_{\text{bbox}}(x) \tag{2}$$

$$e_{x_{\min}} = \text{expl}\left(x, \hat{x}_{\min}\right) \quad e_{y_{\min}} = \text{expl}\left(x, \hat{y}_{\min}\right) \tag{3}$$

$$e_{x_{\max}} = \text{expl}\left(x, \hat{x}_{\max}\right) \quad e_{y_{\max}} = \text{expl}\left(x, \hat{y}_{\max}\right) \tag{4}$$

In case of explaining the bounding box coordinates, the box offsets predicted by an object detectors are converted to normalized image coordinates before computing the gradient. In case of classification decisions, the logits ($\hat{l}_k$, before

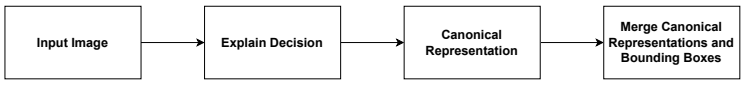
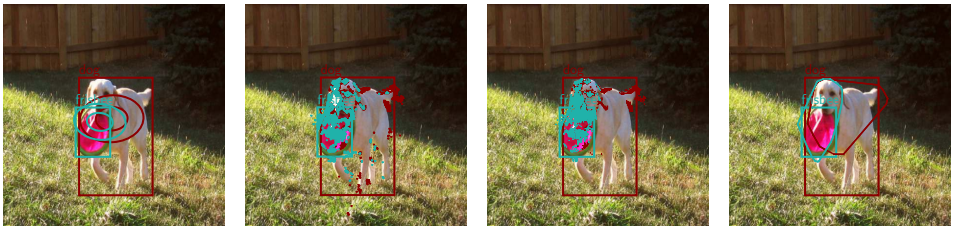


Fig. 3. Overview of the Multi-object visualizations pipeline to jointly visualize all detections.

softmax probability, $\hat{c} = \text{softmax}(\hat{l})$) are used to compute the gradient. Figure 2 illustrates the explanations generated for each decisions of the cat detection by across detectors. Saliency explanations can be computed for each bounding box of interest in the image.

3.2 Multi-object Visualization

In order to summarize the saliency maps of all detections, the individual saliency maps corresponding to each detection are represented using a canonical form. This representation illustrates the most important pixels for the decision explanation. This paper proposes four different methods for combining detection explanations into a single format: principal components, contours, density clustering, and convex polygons. Each method uses a different representation, allowing for detected bounding box, and category to be marked using same colors on the input image. The general process is described in Figure 3. An example the four multi-object visualizations are illustrated in Figure 4. Appendix F provides additional details on the multi-object visualization approaches and how different combination methods work. including explanation heatmap samples.



(a) Principal component (b) Contour (c) Density cluster (d) Convex polygon

Fig. 4. Multi-object visualizations generated to jointly visualize all detections from EfficientDet-D0 and the corresponding classification explanations generated using SIG in the same color. The combination approach is specified in sub-captions. Explanation pixels are colored same as the corresponding bounding box that is being explained.

4 Experiments

Section 4.1 visually analyzes the explanations generated for different detector and explanation method combinations. Section 4.3 provides the quantitatively evaluates different detector and explanation method combinations. Finally, Section 4.4 estimates an overall ranking for the explanation methods based on user preferences of the explanations produced for each decision. In addition, the multi-object visualization methods are ranked based on user understandability of the detections. In Section G, the procedure to analyze the failures of detector using the proposed approach is discussed.

Most of the experiments use ED0, SSD, and FRN detectors detecting common objects from COCO [22]. The additional details about these detectors are provided in Table 2. In cases requiring training a detector, different versions of SSD with various pre-trained backbones detecting marine debris provided in Table 3 are used. The marine debris detectors are trained using a train split of the Marine Debris dataset [47] and explanations are generated for the test images. These detectors are used only to study how are the explanations change across different backbones and different performance levels (epochs) in Section 4.1.

4.1 Visual Analysis

Across target decision and across detectors. The saliency maps for the classification and bounding box decisions generated using a particular explanation method for a specific object change across different detectors as shown in Figure 2. All the bounding box explanations of EfficientDet-D0 in certain scenarios provide visual correspondence to the bounding box coordinates.

Across different target objects. Figure 5 illustrate that the explanations highlight different regions corresponding to the objects explained. This behavior is consistent in most of the test set examples across the classification and bounding box explanations for all detectors.

Figure 6 illustrates the classification explanations for the wall detection across the 6 different backbones. Apart from the attribution intensity changes, the pixels highlight different input image pixels, and the saliency map texture changes. MobileNet and VGG16 illustrate thin horizontal lines and highlight other object pixels, respectively. ResNet20 highlights the wall as a thick continuous segment. Figure 18 illustrate the $y_{\min}$ and $y_{\max}$ bounding box coordinate explanations for the chain detection across different backbones. The thin horizontal lines of MobileNet are consistent with the previous example. In addition, VGG16 illustrates a visual correspondence with the $y_{\min}$ and $y_{\max}$ bounding box coordinate by highlighting the upper half and lower half of the bounding box respectively. However, this is not witnessed in other detectors. This behavior is consistent over a set of 10 randomly sampled test set images from the Marine Debris dataset.

The explanations generated using SSD model instances with ResNet20 backbone at different epochs are provided in Figure 7. The model does not provide any final detections at lower epochs. Therefore, the explanations are generated using the target neurons of the output box corresponding to the interest decision

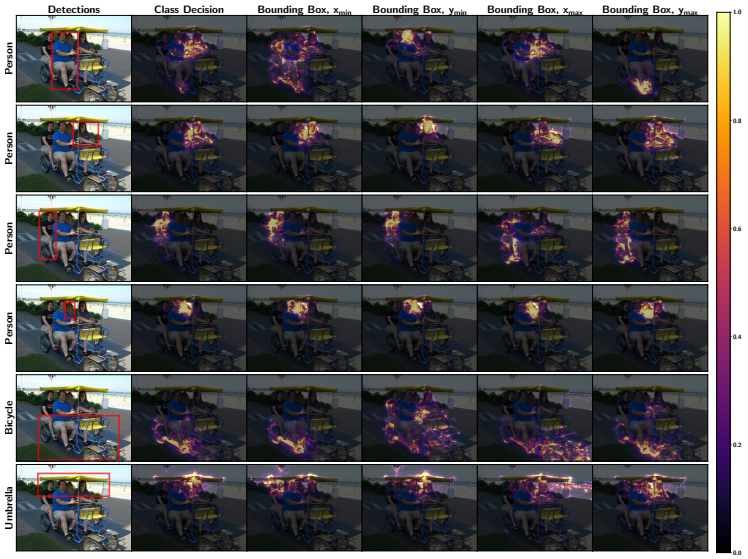


Fig. 5. Comparison of classification and bounding box explanations for all detections from EfficientDet-D0 using SIG is provided. Each row provides the detection (red-colored box) followed by the corresponding classification and all bounding box explanation heatmaps.

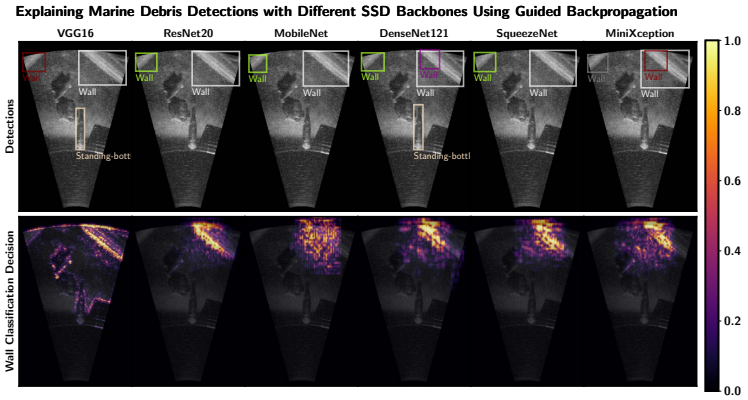


Fig. 6. Comparison of class "wall" classification explanations across different SSD backbones. The detections from each SSD backbone are provided in the first row. The explanations of the wall detection (white-colored box) vary across each backbone.

in the final detections from the trained model. Figure 7 illustrate variations in the saliency maps starting from a randomly initialized model to a completely trained model for the classification decision of the chain detection. The explanations extracted using the random model are dispersed around the features. The

explanations slowly concentrate along the chain object detected and capture the object feature to a considerable amount. This behavior is qualitatively analyzed by visualizing the explanation of 10 randomly sampled test set images from the Marine Debris dataset. In the case of the small hook explained in Figure 19, the variations between the random model and the trained model are not as considerable as the previous chain example. This illustrates the variations change with respect to each class.

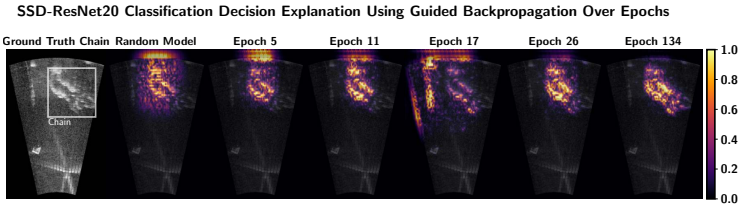


Fig. 7. Classification explanation for class "chain" across different epochs (along columns) of SSD-ResNet20 using GBP is illustrated. The first column is the chain ground truth annotation (white-colored box).

4.2 Error Analysis

The section analyzes detector errors by generating explanations using the proposed detector explanation approach. The saliency map highlighting the important regions can be used as evidence to understand the reason for the detector failure rather than assuming the possible reasons for detector failure. The failure modes of a detector are wrongly classifying an object, poorly localizing an object, or missing a detection in the image [26]. As the error analysis study requires ground truth annotations, the PASCAL VOC 2012 images are used. The PASCAL VOC images with labels mapping semantically to COCO labels are only considered as the detectors are trained using the COCO dataset. For instance, the official VOC labels such as sofa and tvmonitor are semantically mapped to couch and tv, respectively, by the model output trained on COCO.

The procedure to analyze a incorrectly classified detection is straightforward. The output bounding box information corresponding to the wrongly classified detection can be analyzed in two ways. The target neuron can be the correct class or the wrongly classified class to generate the saliency maps (Figure 8). More examples of error analysis are available in Section G in the appendix.

4.3 Quantitative Evaluation

Evaluating detector explanations quantitatively provides immense understanding on selecting the explanation method suitable for a specific detector. This section performs the quantitative evaluation of saliency explanations.

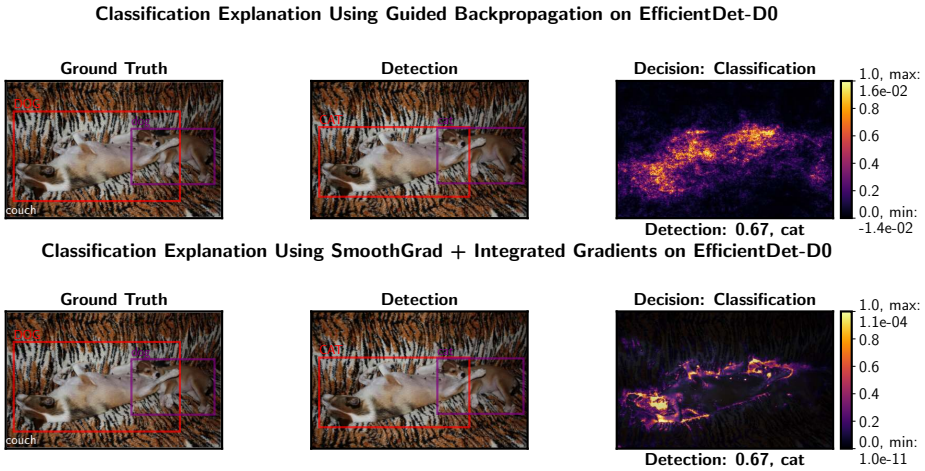


Fig. 8. Example error analysis using gradient-based explanations. EfficientDet-D0 wrongly classifies the dog (red-colored box) in ground truth as cat (red-colored box). We display two saliency explanations (GBP and SIG). In this figure, it is clear the model is imagining a long tail for the dog (GBP) and wrongly classifies the dog as a cat. The saliency map highlights certain features of the dog and the background stripes pattern along the edges of the dog body (GBP and SIG). In order to illustrate the tail clearly which is predominant in cats available in COCO dataset, the saliency map is only shown without overlaying on the input image.

Evaluation Metrics The quantitative evaluation of the explanations of a detector incorporates causal metrics to evaluate the bounding box and classification explanations. This works by causing a change to the input pixels and measuring the effect of change in model decisions. The evaluation aids in estimating the faithfulness or truthfulness of the explanation to represent the cause of the model decision. The causal metrics discussed in this work are adapted from the previous work [33] [26] [25]. The two variants of causal evaluation metrics based on the cause induced to alter the prediction are deletion and insertion metric. The deletion metric evaluates the saliency map explanation by removing the pixels from the input image and tracking the change in model output. The pixels are removed sequentially in the order of the most important pixels starting with a larger attribution value and the output probability of the predicted class is measured. The insertion metric works complementary to the deletion metric by sequentially adding the most important pixel to the image and causing the model decision to change. Using deletion metric, the explanation methods can be compared by plotting the fraction of pixels removed along x-axis and the predicted class probability along y-axis. The method with lower Area Under the Curve (AUC) illustrates a sharp drop in probability for lesser pixel removal. This signifies the explanation method can find the most important pixels that can cause a significant change in model behavior. The explanation method with less AUC is better. In the case of insertion metric, the predicted class probability increases

as the most relevant pixels are inserted. Therefore, an explanation method with a higher AUC is relatively better. [26] utilize constant gray replacing pixel values and blurred image as the start image for deletion and insertion metric calculation respectively.

Effects Tracked. The previous work evaluating the detector explanations utilize insertion and deletion metric to track the change in the bounding box Intersection over Union (IoU) and classification probability together. [26] formulate a vector representation involving the box coordinates, class, and probability. The similarity score between the non-manipulated and manipulated vectors are tracked. However, this work performs an extensive comparison of explanation methods for each decision of a detector by tracking the change in maximum probability of the predicted class, IoU, distance moved by the bounding box (in pixels), change in box height (in pixels), change in box width (in pixels), change in top-left x coordinate of the box (in pixels), and change in top-left y coordinate of the box (in pixels). The box movement is the total movement in left-top and right-bottom coordinates represented as euclidean distance in pixels. The coordinates distances are computed using the interest box corresponding to the current manipulated image and the interest box corresponding to the non-manipulated image. This extensive evaluation illustrates a few explanation methods are more suitable to explain a particular decision. As declared in the previous sections, the image origin is at the top-left corner. Therefore, a total of 7 effects are tracked for each causal evaluation metric.

Evaluation Settings. The previous section establishes the causal deletion and insertion metric along with the 7 different effects. In this section, two different settings used to evaluate the detectors using the causal metrics are discussed.

Single-box Evaluation Setting. The detector output changes drastically when manipulating the input image based on saliency values. We denote principal box to the bounding box detecting the object in the original image. In this setting, seven principal box effects are tracked across insertion and deletion of input pixels. This aids in capturing how well the explanation captures true causes of the principal box prediction. The effects measured for the single-box setting are bounded because the principal box value is always measurable. This is called a single-box setting because only the changes in the principal box are tracked.

Realistic Evaluation Setting. In this evaluation setting, all 7 effects are tracked for the complete object detector output involving all bounding boxes after the post-processing steps of a detector. In this setting, the current detection for a particular manipulated input image is matched to the interest detection by checking the same class and an IoU threshold greater than 0.9. For various manipulated input images, there is no current detection matching the interest detection. Therefore, depending on the effect tracked and to calculate AUC, a suitable value is assigned to measure the effect. For instance, if the effect tracked is the class probability for deletion metric and none of the current detection matches with the interest detection, a zero class probability is assigned. Similarly, if the effect tracked is box movement in pixels for deletion metric, the error in pixels increases to a large value.

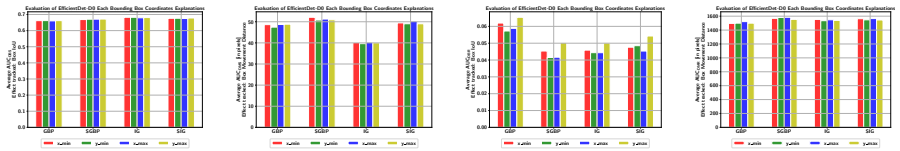
Interpretation Through Curves. Given the causes induced to change model output, effects tracked, and evaluation setting for the detector, this work uses 28 causal evaluation metrics. These correspond to causes ↓ Deletion **(D)** and ↑ Insertion **(I)**, Effects tracked Class Maximum Probability **(C)**, Box IoU **(B)**, Box Movement Distance **(M)**, Box X-top **(X)**, Box Y-top **(Y)**, Box Width **(W)**, Box Height**(H)**, and evaluation settings Single-box **(S)** and Realistic **(R)**.

To interpret a causal evaluation metric, a graph is drawn tracking the change of the effect tracked along the y-axis and the fraction of pixels manipulated along the x-axis. For instance, consider the scenario of deleting image pixels sequentially to track the maximum probability of the predicted class at single-box evaluation setting. The x-axis is the fraction of pixels deleted. The y-axis is the maximum probability of the predicted class at the output of the box tracked. In this work, the curve drawn is named after the combination of the causal evaluation metrics, effects tracked, end evaluation settings. The curves are the DCS curve, DBS curve, ICS curve. For instance, the DCS curve is the change in the maximum probability for the predicted class (C) at the single output box (S) due to removing pixels (D). The curves are the evaluation metrics used in this work and also called as DCS evaluation metric (deletion + class maximum probability + single-box setting), DBS (deletion + box IoU + single-box setting) evaluation metric, and so on.

In order to compare the performance of explanation methods to explain a single detection, as stated before, the AUC of a particular evaluation metric curve is estimated. The corresponding AUC is represented as $\text{AUC}_{<\text{evaluation\_metric}>}$. In order to estimate a global metric to compare the explanation methods explaining a particular decision of a detector, the average AUC, represented as $\text{AAUC}_{<\text{evaluation\_metric}>}$, is computed. As the explanations are provided for each detection, the evaluation set is given by the total number of detections. The total detections in the evaluation set are the sum of detections in each image of the evaluation set. The average evaluation metric curve is computed by averaging the evaluation metric curve at each fraction of pixels manipulated across all detections. AAUC of a particular evaluation metric curve is the AUC of the average evaluation metric curve.

Results Figure 9 illustrates the AAUC computed by evaluating the explanations of each bounding box coordinate is similar across different evaluation metrics curves. This similarity is consistent for all the detectors and explanation methods combinations evaluated. Therefore, the explanation methods quantitatively explain each bounding box coordinate decisions with similar performance. In this work, the AAUC for the bounding box decision is computed by averaging the AUC of all the evaluation metric curves corresponding to all the box coordinate explanations. This offers the means to evaluate the explanation methods across all the bounding box coordinate decisions.

Figure 10 and Figure 11 illustrate quantitatively complementary trends in the evaluation metric curves plotted by tracking box movement distance in pixels and box IoU. The IoU decreases and box movement distance increases as the pixels

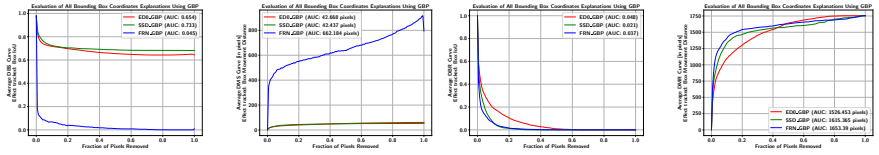


(a) Deletion - Box IoU - Single-box ↓ (b) Deletion - Box Movement - Single-box ↓ (c) Deletion - Box IoU - Realistic ↓ (d) Deletion - Box Movement - Realistic ↓

Fig. 9. The figure illustrates the average AUC, AAUC, for the evaluation metric curves obtained by tracking box IoU (a, c) and box movement distance (b, d) as the pixels are deleted sequentially. Each bar corresponds to the AAUC estimated by evaluating explanations generated for each bounding box coordinate decisions using the explanation methods specified in the x-axis of all detection made by EfficientDet-D0 in the evaluation set images. AAUC is computed by averaging the AUC of all the evaluation metric curves generated using the combination specified in the sub-captions. Lower AAUC is better in all the plots.

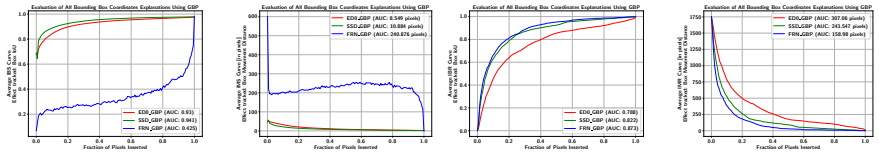
are deleted sequentially as shown in Figure 10. Similarly, Figure 11 illustrates the increase in box IoU and decrease in box movement distance as pixels are inserted to a blurred version of the image. There is a large difference in the AAUC between the single-stage and two-stage detectors. This is primarily due to the RPN in the two-stage detectors. The proposals from RPN are relatively more sensitive to the box coordinate change than the predefined anchors of the single-stage detectors. In addition, Figure 10d and Figure 11d indicates the steady change of box coordinates in the final detections of the EfficientDet-D0. However, SSD and Faster R-CNN saturate relatively sooner. In the remainder of this work, the ability of the box IoU effect is used for quantitative evaluation. This is only because the box IoU effect offers the same scale between 0 to 1 as the class maximum probability effect. In addition, both box IoU and class maximum probability effect follow the trend lower AUC is better for the deletion case. However, it is recommended to consider all the box IoU and box movement distance effects at the level of each box coordinate for a more accurate evaluation.

Figure 12 and Figure 17 aids in understanding the explanation method interpreting both the classification and bounding box decision of a particular detector more faithful than other explanation methods. Figure 12a illustrate SSD512 classification decisions are better explained by SGBP at single-box setting for deletion metrics. However, the bounding box decisions are not explained as well as the classification decisions. Figure 12b illustrate a similar scenario for SGBP with EfficientDet-D0 and Faster R-CNN at the realistic setting for deletion metrics. However, all selected explanation methods explain the bounding box and classification decisions of SSD512 relatively better at the single-box setting for insertion metrics. In general, none of the selected explanation methods explain both the classification and bounding box regression decisions substantially well compared to other methods for all detectors. This answers EQ13. Similarly,



(a) Deletion - Box IoU - Single-box ↓ (b) Deletion - Box Movement - Single-box ↓ (c) Deletion - Box IoU - Realistic ↓ (d) Deletion - Box Movement - Realistic ↓

Fig. 10. Comparison of average curves obtained by tracking box IoU (a, c) and box movement distance (b, d) as the pixels are deleted sequentially. Each average curve is the average of the evaluation curves plotted by evaluating the explanations of all bounding box coordinate decisions across all the detections by the respective detector. The explanations are generated using GBP. The evaluation metric curve is generated using the combination specified in the sub-captions.

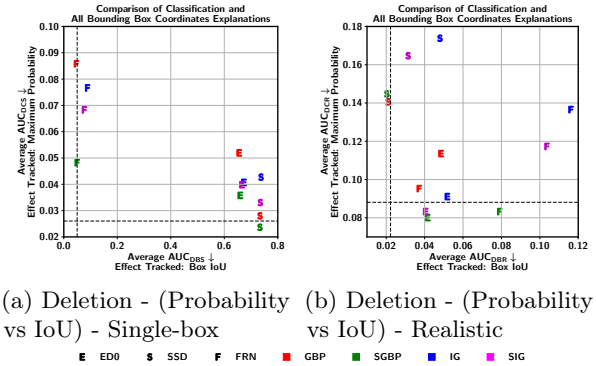


(a) Insertion - Box IoU - Single-box ↑ (b) Insertion - Box Movement - Single-box ↑ (c) Insertion - Box IoU - Realistic ↑ (d) Insertion - Box Movement - Realistic ↑

Fig. 11. Comparison of average curves obtained by tracking box IoU (a, c) and box movement distance (b, d) as the pixels are inserted sequentially. Each average curve is the average of the evaluation curves plotted by evaluating the explanations of all bounding box coordinate decisions across all the detections by the respective detector. The explanations are generated using GBP. The evaluation metric curve is generated using the combination specified in the sub-captions.

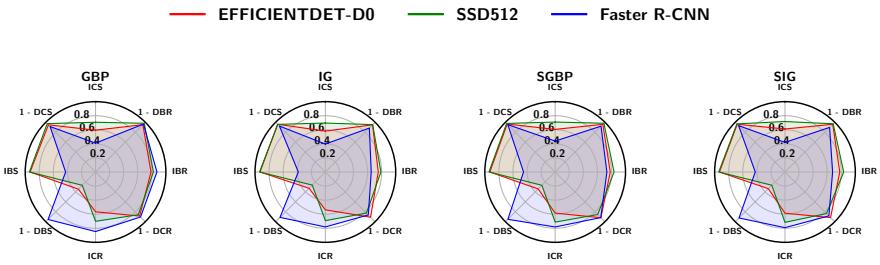
none of the detectors is explained more faithfully for both classification and bounding box decisions among the selected detectors by a single method across all evaluation metrics discussed. This is illustrated by no explanation methods (by different colors) or no detectors (by different characters) being represent in the lower left rectangle or upper right rectangle in Figure 12 and Figure 17 respectively.

Figure 14a and Figure 14c illustrate AAUC of the classification saliency maps and the saliency maps combined using different merging methods are different in certain scenarios while tracking the maximum probability. The AAUC of all the box coordinate saliency maps is provided for a baseline comparison. This denotes the effect on maximum probability by removing pixels in the order of most important depending on the all box coordinates saliency maps. Similarly, Figure 14b and Figure 14d illustrate the similarity in the AAUC of all box



(a) Deletion - (Probability vs IoU) - Single-box (b) Deletion - (Probability vs IoU) - Realistic

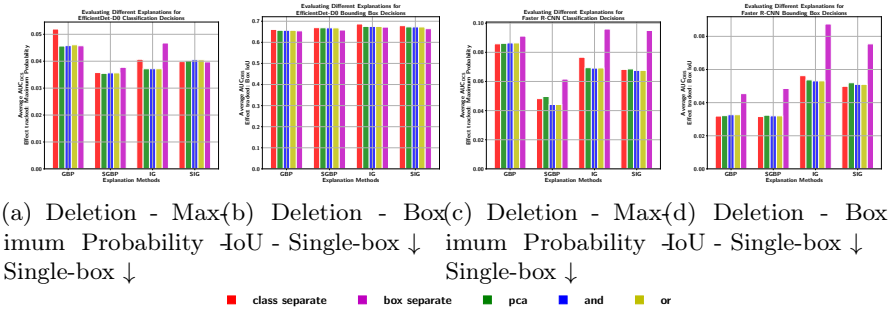
Fig. 12. Comparison between the Deletion AAUC of the evaluation metric curves for the classification and all bounding box coordinate explanations generated across the chosen explanation methods and detectors. Explanation methods (highlighted with different colors) placed at a lower value in the x-axis and y-axis perform relatively better at explaining the box coordinates and classification decisions respectively. Detectors (marked with different characters) placed at a lower value in x-axis and y-axis are relatively better explained for the box coordinates and classification decisions respectively.



(a) GBP (b) IG (c) SGBP (d) SIG

Fig. 13. Multi-metric comparison of quantitative results. According to these metrics, all methods perform similarly when considering all object detectors. The user study and visual inspection of explanation heatmaps reveal more information.

coordinate explanations and the merged saliency maps while tracking the box IoU. In Figure 14a, the evaluation of the GBP classification saliency map is less faithful than the merged saliency map. Therefore, the merged saliency map represents the classification decision more faithfully than the standalone classification explanation in the case of EfficientDet-D0. However, Figure 14a and Figure 14c illustrate in the case of SGBP explaining EfficientDet-D0 and certain cases of Faster R-CNN respectively separately classification saliency maps are more faithful in depicting the classification decision. The larger AAUC for all the box coordinate saliency maps generated using each method for Faster R-CNN indicate the box saliency maps are not faithful to the bounding box



(a) Deletion - Maximum Probability - Single-box ↓ (b) Deletion - Box IoU - Single-box ↓ (c) Deletion - Maximum Probability - Single-box ↓ (d) Deletion - Box IoU - Single-box ↓

Fig. 14. Comparison of average AUC, AAUC, for the evaluation metric curves obtained by tracking maximum probability (a, c) and box IoU (b, d) as the most important pixels based on the explanation generated using the explanation methods specified in the x-axis are deleted sequentially. All the explanations are generated for detection made by EfficientDet-D0 (left) and Faster R-CNN (right) in the evaluation set images. Lower AAUC is better in both plots.

decisions of Faster R-CNN. This is coherent with the visual analysis. Therefore, in certain scenarios merging is helpful to represent the reason for a particular decision. However, each individual saliency map provides peculiar information about the detection. For instance, the visual correspondence shown in Figure 2 to each bounding coordinate information is seen only at the level of individual box coordinate explanations.

An overall comparison of all quantitative metrics is shown in Figure 13. For the purpose of understanding, the ranking of explanation methods explaining a particular detector is provided in Table 1. SGBP performs relatively better across all selected detectors. In addition, IG is ranked least across all the selected detectors. SSD detector is better explained by all the explanation methods. One of the reasons can be SSD is a simpler architecture compared to EfficientDet-D0 and Faster R-CNN. EfficientDet-D0 and Faster R-CNN include a Bi-directional Feature Pyramid Network (BiFPN) and Region Proposal Network (RPN) respectively. However, further experiments should be conducted for validation.

4.4 Human-centric Evaluation

The human-centric evaluation ranks the explanation methods for each detector and ranks the multi-object visualization methods with a user study. All important details of the user study are presented in Appendix H.

Ranking Explanation Methods. Previous work assess the user trust in the model explanations generated by a particular explanation method [26] [34] [29]. As user trust is difficult to evaluate precisely, this work in contrast to previous works estimate the user preferability of the explanation methods. The user preferability for the methods GBP, SGBP, IG, and SIG are evaluated by comparing two explanations corresponding to a particular predictions. In this

Table 1. Ranking of all the explanation methods for a particular detector based on the quantitative evaluation metrics. A lower value is a better rank. The explanation method better explaining a particular detector is awarded a better rank. Each detector is ranked with respect to each evaluation metric considering a particular explanation method. The column names other than the last column and the first two columns represent the average AUC for the respective evaluation metric. The overall rank is computed by calculating the sum along the row and awarding the best rank to the lowest sum. OD - Object detectors, IM - Interpretation method.

OD	IM	DCS	ICS	DBS	IBS	DCR	ICR	DBR	IBR	Overall Rank
ED0	GBP	4	3	1	2	4	3	3	1	3
	SGBP	1	2	2	4	1	2	2	2	2
	IG	3	4	4	3	3	4	4	4	4
	SIG	2	1	3	1	2	1	1	3	1
SSD	GBP	2	3	2	3	1	3	2	3	3
	SGBP	1	2	1	2	2	2	1	1	1
	IG	4	4	4	4	4	4	7	4	4
	SIG	3	1	3	1	3	1	3	2	2
FRN	GBP	4	3	1	2	2	1	1	1	1
	SGBP	1	1	2	1	1	3	2	2	2
	IG	3	4	4	4	4	4	4	4	4
	SIG	2	2	3	3	3	2	3	3	3

study, the explanation methods are compared directly for a particular interest detection and interest decision across SSD, EDO, and FRN detector separately. The evaluation identifies the relatively more trusted explanation method by the users for a particular detector. The explanation methods are ranked by relatively rating the explanations generated using different explanation methods for a particular detection made by a detector. The rating serves as a measure of user preference.

A pair of explanations generated by different explanation methods using the same interest decision and same interest detection for the same detector is shown to a number of human users as shown in Figure 38. The detector, interest decision, interest detection, and explanation method used to generate explanations are randomly sampled for each question and each user. In addition, the image chosen for a particular question is randomly sampled from an evaluation set. The evaluation set is a randomly sampled set containing 50 images from the COCO test 2017. This avoids incorporating any bias into the question generation procedure. Each question is generated on the fly for each user performing the task. The explanations are named Robot A explanation and Robot B explanation to conceal the names of the explanation methods to the user. The robots are not detectors. In this study, the robots are treated as explanation methods. Robot A explanation and Robot B explanation for each question is randomly assigned with a pair of explanation method output. This is done to reduce the bias due to positioning and ordering bias of the explanations as shown to users. The task provided for the user is to rate the quality of the Robot A explanation based on the Robot B explanation. The scoring gives scores in the range $[-2, 2]$ depending if Robot A or B is better. The available options are provided in Table 5.

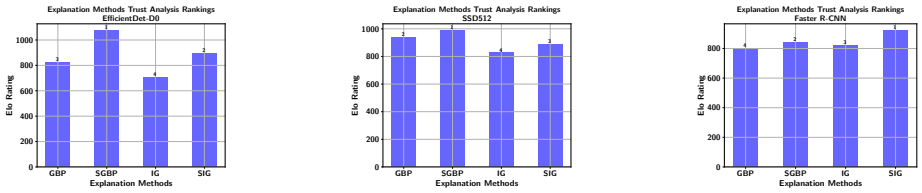


Fig. 15. Ranking obtained for the explanation methods from the user trust study for each detector selected in this work. An initial Elo rating of 1000 is used for all explanation methods. The explanation method with a higher Elo rating has gained relatively more user preferability in the random pair-wise comparisons of explanations for each detector. The rank of a particular method is provided on the top of the bar corresponding to the method.

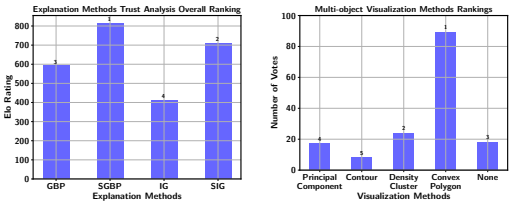
A single question in the evaluation is treated as a game between two randomly matched players. The explanation methods are the players. The game result depends on the explanation quality produced by the competing explanation methods for a particular detection decision. In case of a draw, both explanation methods receive the same score. During non-draw situations, the points won by a particular explanation method are the points lost by the other explanation method. By treating all the questions answered by numerous users as individual games, the global ranking is obtained using the Elo rating system [13]. Each explanation method is awarded an initial Elo rating of 1000.

Ranking Multi-Object Visualization Methods. The rank for multi-object visualization methods is obtained by voting for the method producing the most understandable explanation among the four methods. Each user is asked a set of questions showing the multi-object visualization generated by all four methods. The user is provided with a *None of the methods* option to chose during scenarios where all the multi-object visualizations generated are confusing and incomprehensible to the user. The methods are ranked by counting the total number of votes each method has obtained. The experiment is performed using COCO 2017 test split and the VOC 2012.

Results Each user is requested to answer 10 questions, split as 7 and 3 between Task 1 and Task 2, respectively. 52 participants have answered the user study for both task 1 and task 2. The participants range across researchers, students, deep learning engineers, office secretaries, and software engineers.

Figure 15 indicates SGBP provide relatively more reasonable explanations with higher user preferability for both single-stage detectors. Similarly, SIG is preferred for the two-stage detector. Figure 16a illustrates the top two ranks are obtained by SmoothGrad versions of the SGBP and IG for all detectors. GBP relatively performs in the middle rank in the majority of cases. SGBP achieves the first rank in both the human-centric evaluation and functional evaluation. Figure 16a illustrates the overall ranking taking into account all the bounding

box and classification explanations together. The ranking is similar in analyzing the bounding box and classification explanations separately.



(a) Explanation Methods (b) Multi-Object Visualization Methods

Fig. 16. Ranking obtained from the user study considering all user answers. The rank of a particular method is provided on the top of the bar corresponding to the method.

The ranking of multi-object visualization methods clearly illustrate that majority of the users are able to understand convex polygon-based explanations. 18 answers among the total 156 are *None of the methods* because none of the four other methods provided a legible summary of all the explanation methods and detections. The users have selected principal component-based visualization in cases involving less than 3 detections in an image. In addition, *None of the methods* is chosen in most of the cases involving more than 9 detections or more than 3 overlapping detections in an image. Among the total participants, only 89 users (57%) agree with the convex polygon-based visualization. Therefore, by considering the remaining 43% users, there is a lot of need to improve the multi-object visualization methods discussed in this work and achieve a better summary.

5 Conclusions and Future Work

Explaining convolutional object detectors is crucial given the ubiquity of detectors in autonomous driving, healthcare, and robotics. We extend post-hoc gradient-based explanation methods to explain both classification and bounding box decisions of EfficientDet-D0, SSD512, and Faster R-CNN. In order to integrate explanations and summarize saliency maps into a single output images, we propose four multi-object visualization methods: PCA, Contours, Density clustering, and Convex polygons, to merge explanations of a particular decision.

We evaluate these detectors and their explanations using a set of quantitative metrics (insertion and deletion of pixels according to saliency map importance) and with a user study to understand how useful these explanations are to humans. Insertion and deletion metrics indicate that SGBP provides more faithful explanations in the overall ranking. In general there is no detector that clearly provides better explanations, as a best depends on the criteria being used, but visual inspection indicates a weak relationship that newer detectors (like EfficientDet) have better explanations without artifacts (Figure 2), and that different backbones do have an influence on the saliency map quality (Figure 6).

The user study reveals a human preference for SGBP explanations for SSD and EfficientDet (and SIG for Faster R-CNN), which is consistent with the

quantitative evaluation, and for multi-object explanation visualizations, convex polygons are clearly preferred by humans.

We analyze certain failure modes of a detector using the formulated explanation approach and provide several examples. The overall message of our work is to always explain both object classification and bounding box decisions, and that it is possible to combine explanations into a single output image through convex polygon representation of the saliency map.

Finally, we developed an open-source toolkit, DExT, to explain decisions made by a detector using saliency maps, to generate multi-object visualizations, and to analyze failure modes. We expect that DExT and our evaluation will contribute to the development of holistic explanation methods for object detectors, considering all their output bounding boxes, and both object classification and bounding box decisions.

Limitations. Firstly, the pixel insertion/deletion metrics might be difficult to interpret [16] and more advanced metrics could be used [45]. However, the metric selected should consider the specifics of object detection and evaluate both classification and bounding box regression. Moreover, as detectors are prone to non-local effects, removing pixels from the image [30] can cause bounding boxes to appear or disappear. Therefore, special tracking of a particular box is needed. We extend the classic pixel insertion/deletion metrics [3] for object detection considering these two aspects.

The second limitation is about the user study. Given the challenges in formulating a bias-free question, we ask users to select which explanation method is better. This is a subjective human judgment and does not necessarily have to correspond with the true input feature attribution made by the explanation method. Another part of the user study is comparing multi-object visualization methods, where we believe there is a much clearer conclusion. The novelty of our work is to combine quantitative, qualitative, and a user study, to empirically evaluate saliency explanations for detectors considering object classification and bounding box regression decisions.

In general, saliency methods are prone to heavy criticisms questioning the reliability of the methods. This study extends a few gradient-based saliency methods for detectors and conducts extensive evaluation. However, we acknowledge that there are other prominent saliency methods to study.

Our work evaluates and explains real-world object detectors without any toy example. The literature has previously performed basic sanity checks on toy usecases that does not include multiple localization and classification outputs. In addition, object detectors are categorized on the basis of number of stages (single-stage [23] [43] and two-stage [28]), availability of anchors (anchor-based [23] [43] and anchor-free [27] [44]), and vision transformer based detectors [10] [9]. We explain detectors specific to certain groups (SSD512, Faster R-CNN, and EfficientDet) and leave anchor-free and transformer-based detectors for future.

Even though fully white-box interpretable models would be the best solution [31], this is not yet available at the model scale required for high object detection performance.

References

1. Abdulla, W.: Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. GitHub (2017), (Online accessed on 20 September 2021)
2. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In: 6th International Conference on Learning Representations (ICLR) Conference Track Proceedings (2018)
3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.H.: Gradient-Based Attribution Methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS, vol. 11700, pp. 169–191. Springer, Cham (2019)
4. Arani, E., Gowda, S., Mukherjee, R., Magdy, O., Kathiresan, S.S., Zonooz, B.: A comprehensive study of real-time object detection networks across multiple domains: A survey. Transactions on Machine Learning Research (2022), survey Certification
5. Araújo, T., Aresta, G., Galdran, A., Costa, P., Mendonça, A.M., Campilho, A.: UOLO - Automatic Object Detection and Segmentation in Biomedical Images. In: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T.F., Martel, A.L., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A.P., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (eds.) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2018 and ML-CDS 2018. LNCS, vol. 11045, pp. 165–173. Springer, Cham (2018)
6. Arriaga, O., Valdenegro-Toro, M., Muthuraja, M., Devaramani, S., Kirchner, F.: Perception for Autonomous Systems (PAZ). Computing Research Repository (CoRR) **abs/2010.14541** (2020)
7. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE **10**(7), 1–46 (07 2015)
8. Bastings, J., Filippova, K.: The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In: Alishahi, A., Belinkov, Y., Chrupala, G., Hupkes, D., Pinter, Y., Sajjad, H. (eds.) Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP. pp. 149–155. Association for Computational Linguistics ACL (2020)
9. Beal, J., Kim, E., Tzeng, E., Park, D.H., Zhai, A., Kislyuk, D.: Toward transformer-based object detection. CoRR **abs/2012.09958** (2020)
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision – ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer (2020)
11. Deepan Chakravarthi Padmanabhan, Paul G. Plöger, O.A., Valdenegro-Toro, M.: Sanity checks for saliency methods explaining object detectors. In: Proceedings of the 1st World Conference on eXplainable Artificial Intelligence (2023)
12. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608 (2017)
13. Elo, A.E.: The Rating of Chess Players, Past and Present. BT Batsford Limited (1978)
14. Ester, M., Kriegel, H., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Simoudis, E., Han, J., Fayyad,

U.M. (eds.) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD). pp. 226–231. AAAI Press (1996)
15. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Gläser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. IEEE Transactions on Intelligent Transportation Systems (TITS) **22**(3), 1341–1360 (2021)
16. Grabska-Barwinska, A., Rannen-Triki, A., Rivasplata, O., György, A.: Towards better visual explanations for deep image classifiers. In: eXplainable AI approaches for debugging and diagnosis. (2021)
17. Gudovskiy, D.A., Hodgkinson, A., Yamaguchi, T., Ishii, Y., Tsukizawa, S.: Explain to Fix: A Framework to Interpret and Correct DNN Object Detector Predictions. Computing Research Repository (CoRR) **abs/1811.08011** (2018)
18. He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., Li, X.: Single Shot Text Detector with Regional Attention. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3066–3074. IEEE (2017)
19. Jain, S., Wallace, B.C.: Attention is not Explanation. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) Volume 1 (Long and Short Papers). pp. 3543–3556. Association for Computational Linguistics (ACL) (2019)
20. Kim, B., Doshi-Velez, F.: Machine Learning Techniques for Accountability. AI Magazine **42**(1), 47–52 (2021)
21. Kim, J.U., Park, S., Ro, Y.M.: Towards Human-Like Interpretable Object Detection Via Spatial Relation Encoding. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 3284–3288. IEEE (2020)
22. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014)
23. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016)
24. Lundberg, S.M., Lee, S.: A Unified Approach to Interpreting Model Predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). p. 4768–4777. NIPS'17, Curran Associates, Inc. (2017)
25. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: British Machine Vision Conference (BMVC). p. 151. BMVA Press (2018)
26. Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V.I., Mehra, A., Ordonez, V., Saenko, K.: Black-box Explanation of Object Detectors via Saliency Maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11443–11452 (2021)
27. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788. IEEE (2016)
28. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI) **39**(6), 1137–1149 (2017)

29. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. Association for Computing Machinery (ACM) (2016)
30. Rosenfeld, A., Zemel, R.S., Tsotsos, J.K.: The Elephant in the Room. Computing Research Repository (CoRR) **abs/1808.03305** (2018)
31. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
32. Rudin, C., Wagstaff, K.L.: Machine learning for science and society. Machine Learning **95**(1), 1–9 (2014)
33. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE **109**(3), 247–278 (2021)
34. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision **128**(2), 336–359 (2020)
35. Serrano, S., Smith, N.A.: Is Attention Interpretable? In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL). pp. 2931–2951. Association for Computational Linguistics (ACL) (2019)
36. Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning (ICML) 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. Proceedings of Machine Learning Research (PMLR) (2017)
37. Shwartz-Ziv, R., Tishby, N.: Opening the Black Box of Deep Neural Networks via Information. Computing Research Repository (CoRR) **abs/1703.00810** (2017)
38. 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) Workshop Track Proceedings (2014)
39. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: SmoothGrad: removing noise by adding noise. Computing Research Repository (CoRR) **abs/1706.03825** (2017)
40. Spiegelhalter, D.: Should We Trust Algorithms? Harvard Data Science Review **2**(1) (2020)
41. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for Simplicity: The All Convolutional Net. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations (ICLR) Workshop Track Proceedings (2015)
42. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning (ICML) 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. Proceedings of Machine Learning Research (PMLR) (2017)
43. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and Efficient Object Detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10778–10787. IEEE (2020)

44. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 9626–9635. IEEE (2019)
45. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., Preece, A.: Sanity checks for saliency metrics. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 6021–6029 (2020)
46. Tsunakawa, H., Kameya, Y., Lee, H., Shinya, Y., Mitsumoto, N.: Contrastive Relevance Propagation for Interpreting Predictions by a Single-Shot Object Detector. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2019)
47. Valdenegro-Toro, M.: Forward-Looking Sonar Marine Debris Datasets. GitHub (2019), (Online accessed on 01 December 2021)
48. Wagstaff, K.L.: Machine Learning that Matters. In: Proceedings of the 29th International Conference on Machine Learning (ICML) 2012. icml.cc / Omnipress (2012)
49. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-Local Neural Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7794–7803. IEEE (2018)
50. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and Interpretability in Convolutional Neural Networks for Semantic Segmentation of Colorectal Polyps. Medical Image Analysis **60** (2020)
51. Wu, T., Song, X.: Towards Interpretable Object Detection by Unfolding Latent Structures. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6032–6042. IEEE (2019)
52. Zablocki, É., Ben-Younes, H., Pérez, P., Cord, M.: Explainability of vision-based autonomous driving systems: Review and challenges. Computing Research Repository (CoRR) **abs/2101.05307** (2021)
53. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer (2014)
54. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object Detection in 20 Years: A Survey. Computing Research Repository (CoRR) **abs/1905.05055** (2019)

A Broader Impact Statement

As concerns on AI safety is increasing, explainable machine learning is imperative to gain human trust and satisfy the legal requirements. Any machine learning model user for human applications should be able to explain its predictions, in order to be audited, and to decide if the predictions are useful or further human processing is needed. Similarly, such explanations are pivotal to earn user trust, increase applicability, address safety concerns for complex object detection models.

We expect that our work can improve the explainability of object detectors, by first steering the community to explain all object detector decisions (bounding box and object classification), considering to visualize all saliency explanations in a single image per detector decision, and to evaluate the non-local effect of image pixels into particular detections. We believe that saliency methods can be used to partially debug object detection models. Consequently, saliency methods are useful to explain detector and address the trustworthiness and safety concerns in critical applications using detectors.

However, additional validation of explanations is needed. We also perform sanity checks in object detectors [11] with similar conclusions and validation of saliency map quality. Additional large scale user studies could be done to evaluate how useful these explanations are for humans, instead of just asking which explanation method is better.

Even though fully white-box interpretable models would be the best solution [31], this is not yet available at the model scale required for high object detection performance.

In addition, the detectors are evaluated in various combinations with two settings: single-box and realistic. Both the former and the latter help to better understand the effects of the most relevant pixels on the predictions for the output box as well as the overall detector output respectively. From the overall ranking based on the quantitative evaluation metrics, all the explanation methods interpret SSD more faithfully in comparison to other detectors. SGBP provides more faithful explanations in the overall ranking developed from the quantitative evaluation metrics. This is coherent with the user study. Humans understand the explanations from SGBP in comparison more than the explanations generated by other shortlisted explanation methods.

Convex polygon-based multi-object visualizations are better understood and preferred by humans. However, there is substantial scope to improve the generated multi-object visualizations.

B Detectors Details

Detectors detecting common objects available in COCO dataset is provided in Table 2:

Detectors trained on Marine Debris Dataset is provided in Table 3:

Table 2. Summary of object objector implementations used in this work. The detectors are trained to detect common objects using COCO dataset. The mAP reported is at 0.5 IoU. val35k represents 35k COCO validation split images. minival is the remaining images in the validation set after sampling val35k.

Detector	Train split	Test split	mAP (%)	Weights	Code
FRN	train+val35k 2014	minival2014	54.4	[1]	[1]
SSD	train+val35k 2014	test-dev 2015	46.5	[23]	[6]
ED0	train 2017	test-dev 2017	53.0	[43]	[6]

Table 3. Details about the marine debris objector used in this work. Reported mAP is at 0.5 IoU.

SSD Backbones	mAP (%)	Input Image Size
VGG16	91.69	300 x 300
ResNet20	89.85	96 x 96
MobileNet	70.30	96 x 96
DenseNet121	73.80	96 x 96
SqueezeNet	68.37	96 x 96
MiniXception	71.62	96 x 96

C Explanation Methods

In this paper we use Guided Backpropagation (GBP), Integrated Gradients (IG), and their variations using SmoothGrad (SGBP and SIG). We describe these methods in detail below:

Guided Backpropagation. (GBP) [41] is a backpropagation-based attribution method. GBP provides information about the input image features utilized by a DNN for the particular prediction. The method calculates the loss function gradient for a specific object category with respect to image pixels. In this approach, the activations at a higher level unit under study are propagated backward to reconstruct the input image. The reconstructed input image illustrates the input image pixels strongly activating the higher-level unit. The feature map f after passing though a ReLU activation *relu* at layer l, where i denotes each feature is given in Equation 5:

$$f_i^{l+1} = relu\left(f_i^l\right) = \max\left(f_i^l, 0\right) \tag{5}$$

GBP handles backpropagation through ReLU non-linearity by combining vanilla backpropagation and DeconvNets as specified in Equation 6.

$$R_i^l = \left(f_i^l > 0\right) \cdot \left(R_i^{l+1} > 0\right) \cdot R_i^{l+1} \tag{6}$$

The reconstructed image R at any layer l is generated by the positive forward pass activations f_i^l and the positive error signal activations R_i^{l+1}. This aids

in guiding the gradient by both positive input and positive error signals. The negative gradient flow is prevented in GBP, thereby, providing more importance to the neurons increasing the activation of the higher layer unit under study. In addition, this suppresses the image aspects negatively affecting the activation. Therefore, the regenerated images are relatively less noisy compared to the Gradients and DeconvNet methods. The explanation is computed as the gradient of a particular output neuron with respect to the input image, considering the previously mentioned modified ReLU gradient:

$$\text{expl}_{\text{GBP}}(x, \hat{y}) = \frac{\partial \hat{y}}{\partial x} \tag{7}$$

Integrated Gradients. (IG) [42] achieves the implementation invariance as well as sensitivity axioms. The gradient-based attribution methods such as Gradients [38], DeconvNet [53], GBP [41], LRP [7], and DeepLIFT [36] fail either of two rules or both. The sensitivity rule states for a baseline and input image differing by a single feature and resulting in different predictions, the differing feature must be assigned a non-zero attribution. In addition, a zero attribution should be assigned to constant variables in the trained function. The implementation invariance rule signifies that the attribution method result should not be dependent on the network implementation. The functionally equivalent models should have identical attributions. Furthermore, IG satisfies the completeness axiom by balancing out the difference in the model output for the input image and baseline to the sum of all feature attribution. IG integrates along the local gradient for a particular image pixel over a linear path from the baseline x' to input image x pixels. IG for feature i in the input image is calculated using Equation 8 [2].

$$\text{IntegratedGradients}_i(x, F) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha \tag{8}$$

α is the interpolation constant for peturbing the features along the straight path between baseline and input image. $F(x)$ is the model function mapping input image to output prediction. The solution is obtained using numerical approximation because calculating definite integral for Equation 8 is difficult. The full integrated gradients calculation is done over all input features and is:

$$\text{expl}_{\text{IG}}(x, \hat{y}) = [\text{IntegratedGradients}_i(x, \hat{y}) \, \forall i \in 0 \ldots \dim(x)] \tag{9}$$

SmoothGrad. [39] is an approach to sharpen the saliency maps generated by any gradient-based explanation method. The idea is to estimate a saliency map by averaging all the saliency maps generated for different image samples by adding a small random noise. Given $\text{expl}_{\text{M}}(x, \hat{y})$ is the unsmoothed saliency map explaining the decision for predicting class c with any previous saliency method. The final saliency map $\text{expl}_{\text{SM}}(x, \hat{y})$ for the input image x is given by Equation 10. N is the total number of image samples generated by adding Gaussian noise $\mathcal{N}(0, \sigma^2)$ with standard deviation σ.

$$\text{expl}_{\text{SM}}(x, \hat{y}) = N^{-1} \sum_{1}^{n} \text{expl}_{\text{M}}(x + \epsilon, \hat{y}) \tag{10}$$

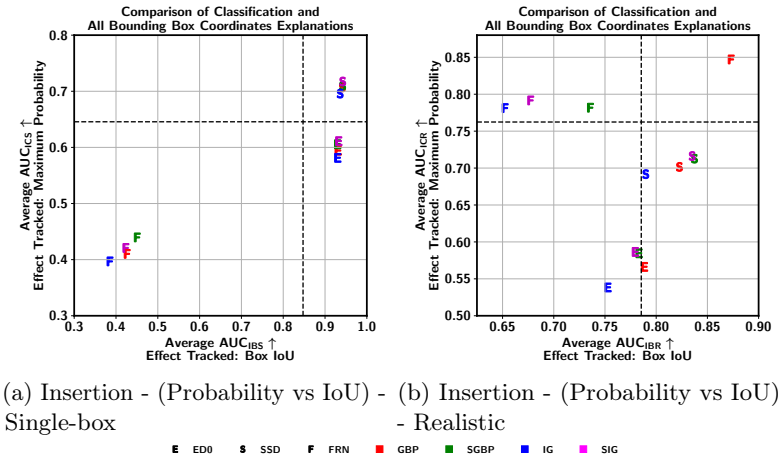
With $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ being samples from a Gaussian distribution. The hyperparameters are the sample size to average the saliency maps N and standard deviation or noise level σ. [39] suggests a noise level between 10-20% balances the saliency map sharpness and captures the object structure. This is followed by averaging the saliency maps obtained for different noise levels to generate a final smoothed saliency map.

We combine SmoothGrad with Guided Backpropagation to produced Smooth Guided Backpropagatio (SGBP), and SmoothGrad with Integrated Gradients to produce Smooth Integrated Gradients (SIG).

D Additional Comparison of Quantitative Metrics

Table 4. Ranking of all detectors for a particular explanation method based on the quantitative evaluation metrics. A lower value is a better rank. The detector better explained by a particular explanation method is awarded a better rank. Each detector is ranked with respect to each evaluation metric considering a particular explanation method. The column names other than the last column and the first two columns represent the AAUC for the respective evaluation metric. The overall rank is computed by calculating the sum along the row and awarding the best rank to the lowest sum. OD - Object detectors, IM - Interpretation method.

IM	OD	DCS	ICS	DBS	IBS	DCR	ICR	DBR	IBR	Overall Rank
GBP	ED0	2	2	2	2	2	3	3	3	3
	SSD	1	1	3	1	3	2	1	2	1
	FRN	3	3	1	3	1	1	2	1	2
SGBP	ED0	2	2	2	2	1	3	2	2	2
	SSD	1	1	3	1	3	2	1	1	1
	FRN	3	3	1	3	2	1	3	3	3
IG	ED0	1	2	2	2	1	3	2	2	2
	SSD	2	1	3	1	3	2	1	1	1
	FRN	3	3	1	3	2	1	3	3	3
SIG	ED0	2	2	2	2	1	3	2	2	2
	SSD	1	1	3	1	3	2	1	1	1
	FRN	3	3	1	3	2	1	3	3	3



(a) Insertion - (Probability vs IoU) - Single-box

(b) Insertion - (Probability vs IoU) - Realistic

Fig. 17. Comparison between the insertion AAUC of the evaluation metric curves for the classification and all bounding box coordinate explanations generated using different explanation methods across all detectors. This offers a means to understand the explanation method generating more faithful explanations for both classification explanations and all bounding box coordinates. As the curves to compute the respective AUC are computed using insertion metric, higher values in both axis are better. The explanation methods (highlighted with different colors) placed at a higher value in x-axis and y-axis perform relatively better at explaining the box coordinates and classification decisions respectively. The detectors (marked with different characters) placed at a higher value in x-axis and y-axis are relatively better explained for the box coordinates and classification decisions respectively.

E Visual Analysis

Figure 18 and Figure 19 illustrates the change in explanations across different backbones and performance levels.

F Multi-object Visualization

In order to summarize the explanations for a particular decision across all objects in an image, four multi-object visualization methods are proposed in Section 3.2. This procedure is concisely presented in Figure 20, Figure 21, Figure 22, and Figure 23. Figure 24 and Figure 25 illustrates the summarized visualizations for all objects predicted using all the proposed methods. The principal component-based method represents the maximum and minimum data spread of the saliency map pixel intensities as ellipses centered at the center of mass. The contour-based method draws the contour map with two levels as depicting a heatmap and the output detection with the same color is difficult. The density cluster-based method performs density clustering using DBSCAN [14]. The hyperparameters

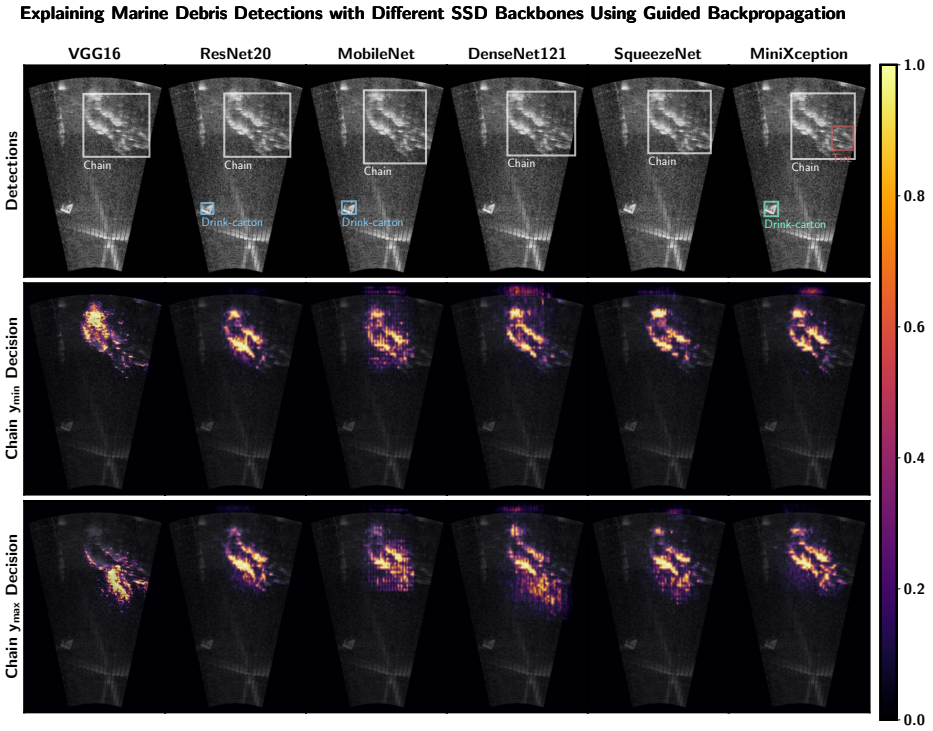


Fig. 18. An illustration of the chain $y_{\min}$ and $y_{\max}$ explanations across different SSD backbones is provided. The detections from each SSD backbone are provided in the first row. The chain detection explained is marked using a white-colored box. The explanations vary across each backbone. SSD-VGG16 $y_{\min}$ and $y_{\max}$ explanations highlight the upper half and lower half of the chain respectively, corresponding to the bounding box coordinate locations.

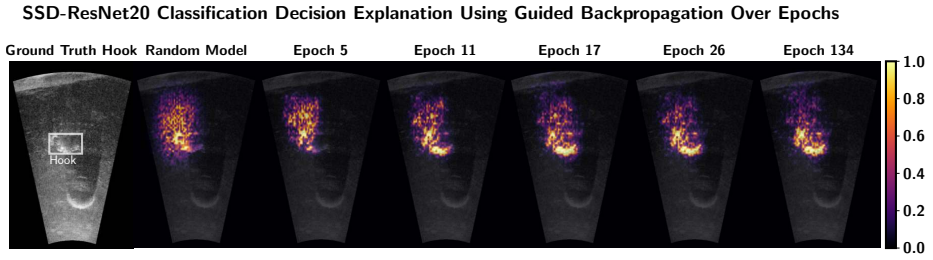


Fig. 19. The hook classification explanation across different epochs (along columns) of SSD-ResNet20 using GBP is illustrated. The first column is the hook ground truth annotation (white-colored box).

of DBSCAN are tuned using the method stated in [14]. Finally, the convex polygon-based method draws a convex polygon over the density clustered saliency

map pixels. This method provides a legible representation as the convex polygon resemble an irregularly shaped bounding box.

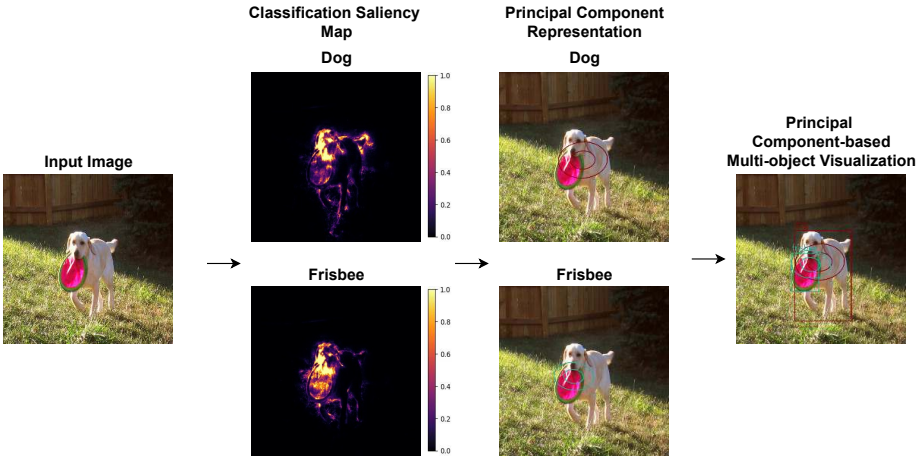


Fig. 20. The detector predicts a dog and a frisbee in the input image. The saliency map for the corresponding classification decisions are converted into a canonical form represented as elliptical principal components. The final multi-object visualization is generated by combining the ellipses, bounding boxes, and class predictions into a single image with a particular color for each object.

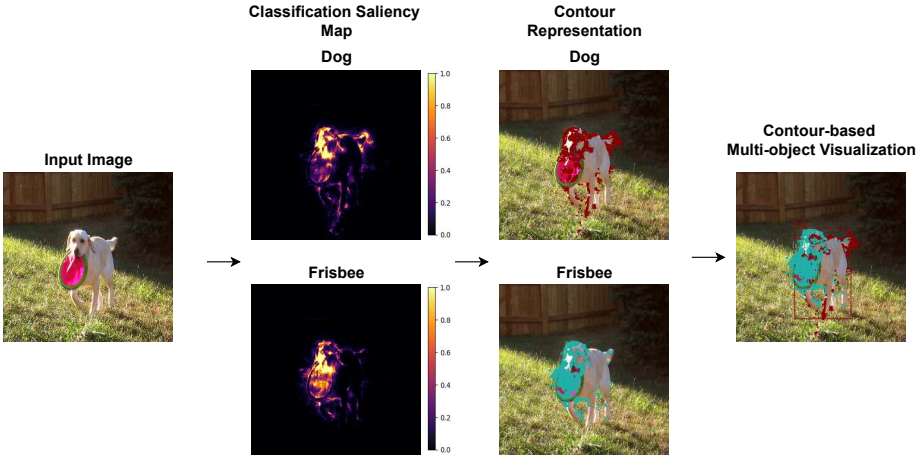


Fig. 21. The detector predicts a dog and a frisbee in the input image. The saliency map for the corresponding classification decisions are converted into a canonical form represented as contours based on importance for the decision. The final multi-object visualization is generated by combining the contours, bounding boxes, and class predictions into a single image with a particular color for each object.

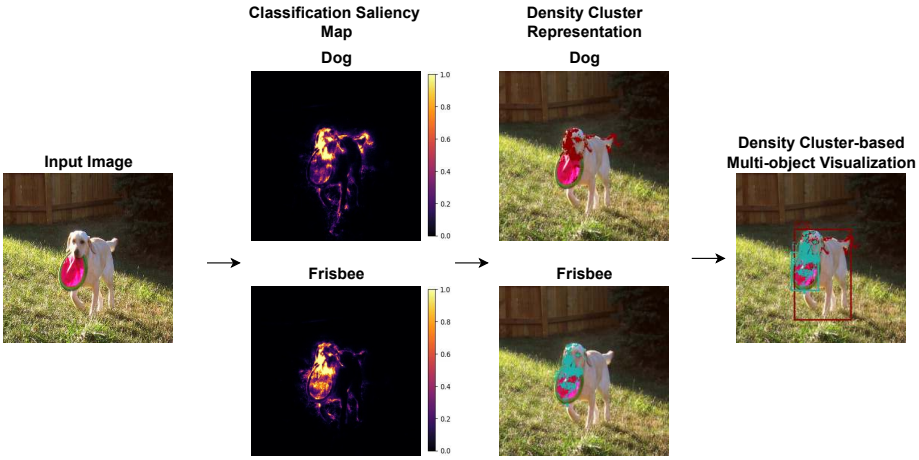


Fig. 22. The detector predicts a dog and a frisbee in the input image. The saliency map for the corresponding classification decisions are converted into a canonical form represented as density clusters based on importance for the decision. The final multi-object visualization is generated by combining the density clusters, bounding boxes, and class predictions into a single image with a particular color for each object.

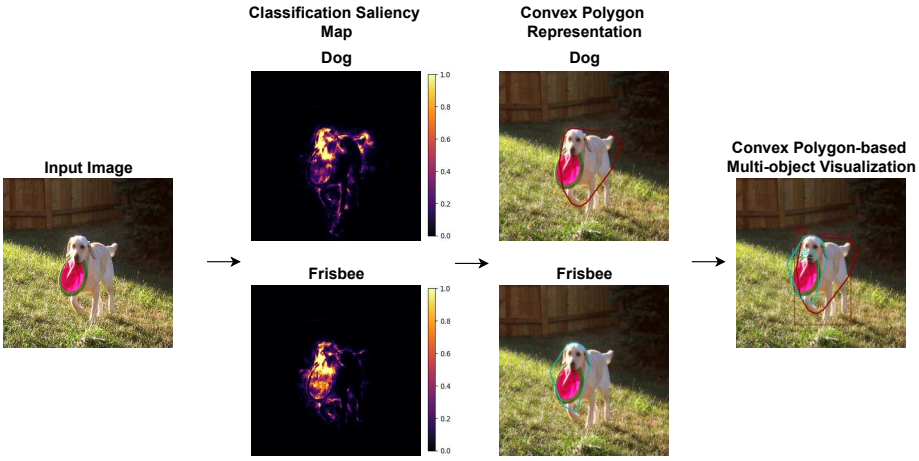
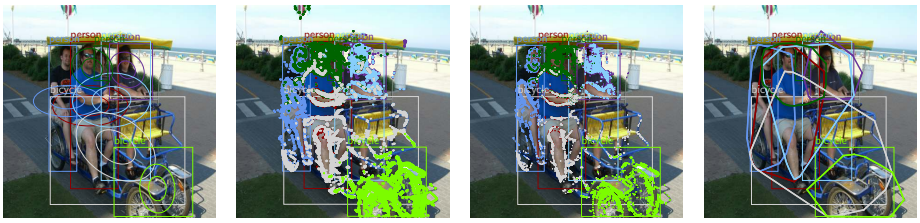
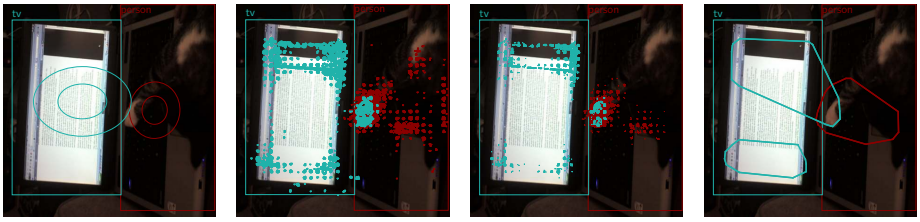


Fig. 23. The detector predicts a dog and a frisbee in the input image. The saliency map for the corresponding classification decisions are converted into a canonical form represented as convex polygon. The final multi-object visualization is generated by combining the polygons, bounding boxes, and class predictions into a single image with a particular color for each object.



(a) Principal component (b) Contour (c) Density cluster (d) Convex polygon

Fig. 24. Multi-object visualizations generated to visualize together all the detections from SSD512 and the corresponding classification explanations generated using SGBP in the same color. The multi-object visualization approach is specified in the sub-captions. The important pixels responsible for the decision explained in the case of the principal component-based and convex polygon-based are the pixels inside the ellipses and irregular polygons respectively, marked in the same color as the corresponding detection. The important pixels responsible for the decision explained in the case of contour-based and density-based are the pixels highlighted in the same color as the corresponding detection.



(a) Principal component (b) Contour (c) Density cluster (d) Convex polygon

Fig. 25. Multi-object visualizations generated to visualize together all the detections from Faster R-CNN and the corresponding classification explanations generated using SGBP in the same color. The multi-object visualization approach is specified in the sub-captions. The important pixels responsible for the decision explained in the case of the principal component-based and convex polygon-based are the pixels inside the ellipses and irregular polygons respectively, marked in the same color as the corresponding detection. The important pixels responsible for the decision explained in the case of contour-based and density-based are the pixels highlighted in the same color as the corresponding detection.

G Additional Examples of Error Analysis

We provide six additional examples, two are about poor localization (Figures 26 and 28), and six about misclassification or confusion with background (Figures 29, 30, 27, and 31).

The saliency maps for each bounding box coordinates can provide visual evidence for poor localization (Figure 26). Finally, by generating saliency maps for the bounding box coordinate or classification decisions of the adjusted prior box close to the missed ground truth detection, the reason for missing detections can be studied (Figure 29).

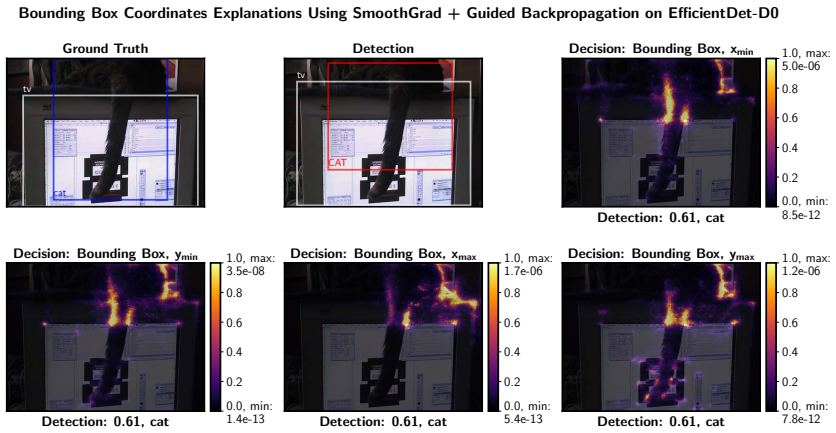


Fig. 26. Example error analysis using gradient-based explanations. EfficientDet-D0 localizes the cat detection (red-colored box) poorly (IoU: 0.69) in the detection subplot. It is evident from the saliency map of y_max bounding box that the detector is looking at the end part of the tail. However, the detector misses the tail of the cat because of other nearby features from the monitor display.

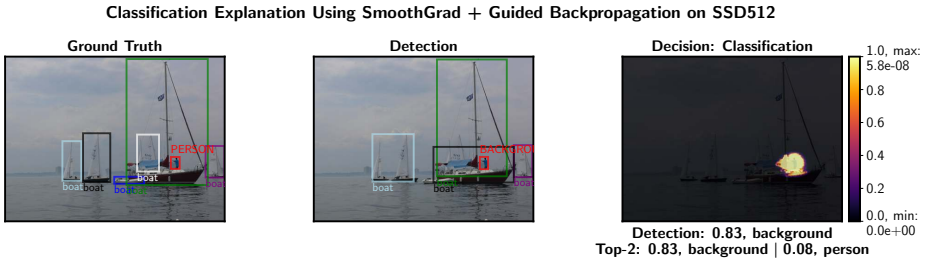


Fig. 27. Example error analysis using gradient-based explanations. SSD512 misses the person (red-colored box) in the ground truth subplot using the proposed approach. The red-colored box in the detection subplot is the closest output box to the ground truth. The saliency map highlights the the entire person and part of the boat, possibly indicating that the person feature is not prominent in that region. The detector classifies the box as background. However, the second dominant class of the box is person.

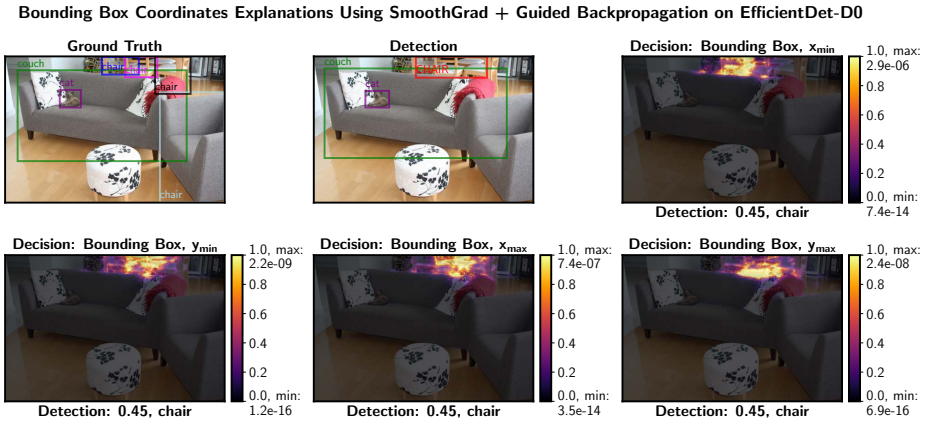


Fig. 28. Example error analysis using gradient-based explanations. EfficientDet-D0 localizes only a single chair in the back (red-colored box) It is evident from the localization saliency maps that the detector is localizing all the nearby chairs together as a single instance, and the saliency indicates this clearly. The bounding box saliency should focus in a single chair instead of multiple ones.

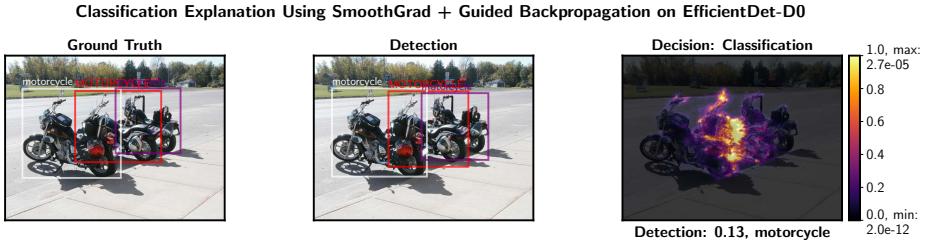


Fig. 29. Example error analysis using gradient-based explanations. EfficientDet-D0 misses the motorcycle (red-colored box) in the ground truth subplot. The red-colored box in the detection subplot is the closest output box to the ground truth. The motorcycle tank, right throttle, and certain other surfaces of the missed motorcycle are clearly highlighted. However, the detector does not have sufficient evidence to accept the classification result due to lower confidence for the motorcycle class (0.13) than the confidence threshold (0.5) for acceptable detections.

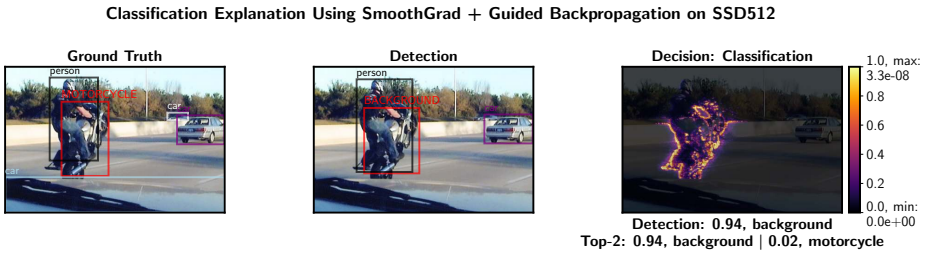


Fig. 30. Example error analysis using gradient-based explanations. SSD512 misses the motorcycle (red-colored box) in the ground truth subplot using the proposed approach. The red-colored box in the detection subplot is the closest output box to the ground truth. The saliency map highlights the entire motorcycle, person, and edges of the lane divider. The detector classifies the box as background. However, the second dominant class of the box is motorcycle. This is probably due to the person occluding part of the motorcycle.

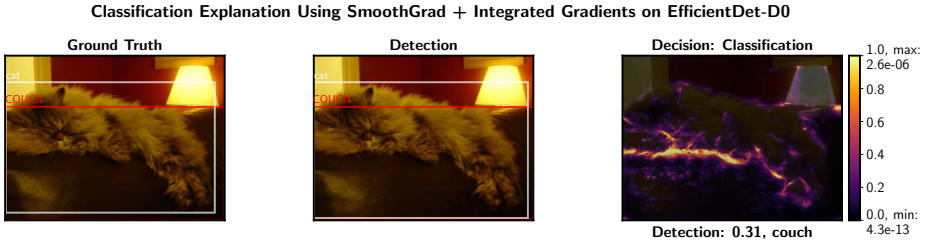


Fig. 31. Example error analysis using gradient-based explanations. EfficientDet-D0 misses the motorcycle (red-colored box) in the ground truth subplot. The red-colored box in the detection subplot is the closest output box to the ground truth. The couch surface and the context of a cat lying on the surface of couch are clearly highlighted. However, the detector does not have sufficient evidence to accept the classification result due to lower confidence for the couch class (0.31) than the confidence threshold (0.5) for acceptable detections. This is likely due to the cat occluding part of the couch.

H Details of User Study

This section provides the task description given to the users and screenshots of the application developed to perform the user study and human-centered analysis.

Table 5. User study options and scores awarded to respective explanations.

Options	A Score	B Score
Robot A explanation is much better	2	-2
Robot A explanation is slightly better	1	-1
Both explanations are same	0	0
Robot A explanation is slightly worse	-1	1
Robot A explanation is much worse	-2	2

H.1 Task Description

Firstly, thank you for your time. I assure you that I will use the answers solely for research purposes without disclosing any user identity. The evaluation includes two tasks. Task I: Questions 1-7. Task II: Questions 8-10.

Task I: Which Robot's explanation is better?

- An artificial intelligence (AI) agent performing the task of localizing and classifying all the objects in an image is called an object detector.
- The output from an object detector to detect a single object includes the **bounding box** representing the maximum rectangular area occupied by the object and the **class name** representing the category of the object inside the bounding box. The output is called detection.
- (x_left_top, y_left_top) and (x_right_bottom, y_right_bottom) are the two coordinate points to represent a bounding box. The class name of the object is represented as a text label near the bounding box as shown in Figure 32.
- Therefore, each detection is made of two decisions (predictions), namely, bounding box coordinates decision and classification decision.
- In this study, the reason for a particular decision, say bounding box coordinates or class prediction, in a single detection, is shown.
- This reason behind the decision-making process is given by the explanation. In this task, the explanation is generated by two different robots, **Robot A** and **Robot B**. The explanation images are provided for classification and bounding box decisions separately.

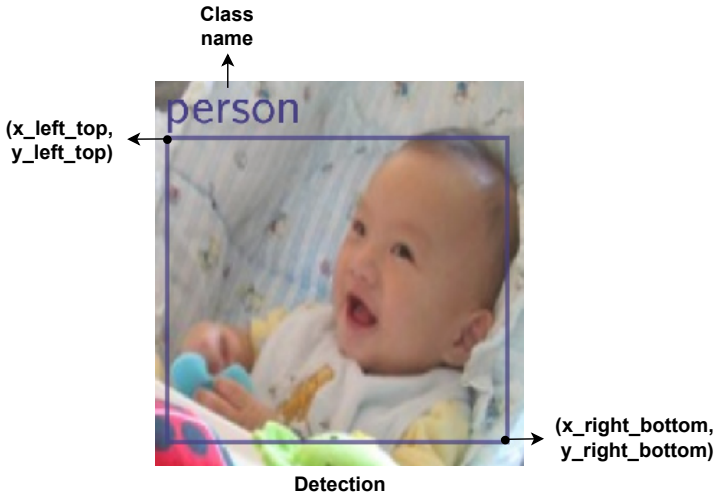


Fig. 32. An illustration of a detection output from a detector.

– The explanation for a particular decision is provided by highlighting the pixels important for the decision-making process. The color bar provided in Figure 33 on the right of the explanation image indicates the pixel importance value.

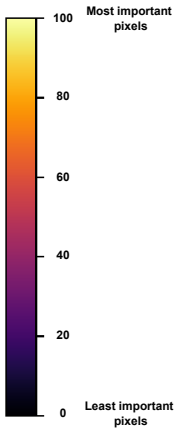


Fig. 33. A heatmap representing the importance of pixels for a particular decision.

– In task 1, the author requests you to rate Robot A's explanation by comparing it against Robot B's explanation in terms of understandability and meaningfulness of the explanation.

– A few classification decision explanations, Figure 34 and Figure 35, are provided below:

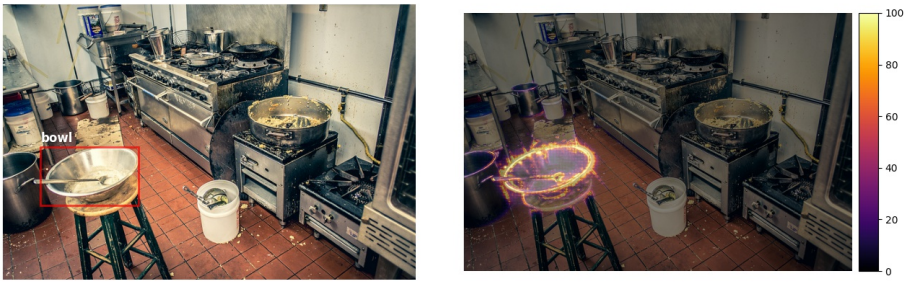


Fig. 34. A bowl detection made by the detector (shown in the left). An explanation for the bowl classification decision (right). The most of the important pixels highlight the object detected. The pixel importance values of objects other than the detected object are very less and negligible.

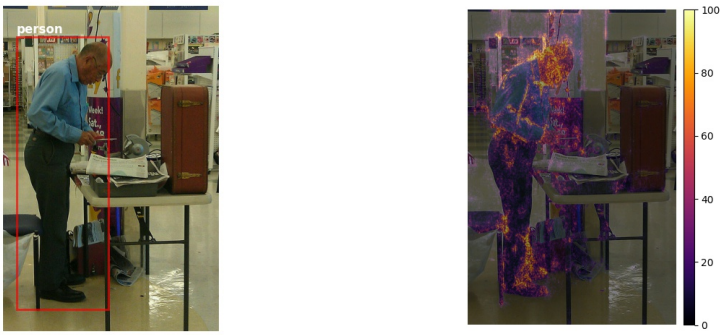


Fig. 35. A person detection made by the detector (left). An explanation for the person classification decision (right). The most of the important pixels highlight the object detected. However, the explanations highlight pixels other than the detected object and is highly noisy.

– A few bounding box coordinate explanations, Figure 36 and Figure 37, are provided below:

Task II: Which method is better to summarize all detections and corresponding explanations?

Fig. 36. A person detection made by the detector (left). An explanation for the person y_left_top coordinate prediction (right). The most of the important pixels highlight the object detected. In addition, the explanation is coherent with the bounding box coordinate as the explanation highlights region near the y_left_top.

Fig. 37. A person detection made by the detector (left). An explanation for the person x_right_bottom coordinate prediction (right). The important pixels highlight the object detected. However, the explanation highlights numerous pixels outside the the detected object and is highly noisy.

- Each image shown in this task includes all the detection made by the detector. Similar to the previous task, each detection is represented as shown in Figure 32.
- In addition to all detections, each image illustrates the explanation for a particular decision, say bounding box coordinate or classification result, for all objects detected by the detector.
- In order to map the detection and the respective explanation, the same colors are used.
- The explanations are represented using 4 different methods. However, visually, across the 4 methods, the important pixels responsible for a particular decision are highlighted using either dots, ellipses, or irregular polygon.
- For ellipses and irregular polygon, the pixels inside the ellipse and irregular polygon are the important pixels responsible for the decision-making process.
- One of the options is *None of the methods*. This option can be selected when the detection and corresponding explanation of multiple objects illustrated in all 4 images are confusing and illegible to coherently understand the detection and the corresponding explanation.

H.2 Application Screenshots

This section provides the snapshots of the user study application. Figure 38 and Figure 39 shows a sample Task 1 and Task 2 question. Figure 40 illustrates the additional questions asked to understand the background of the user.

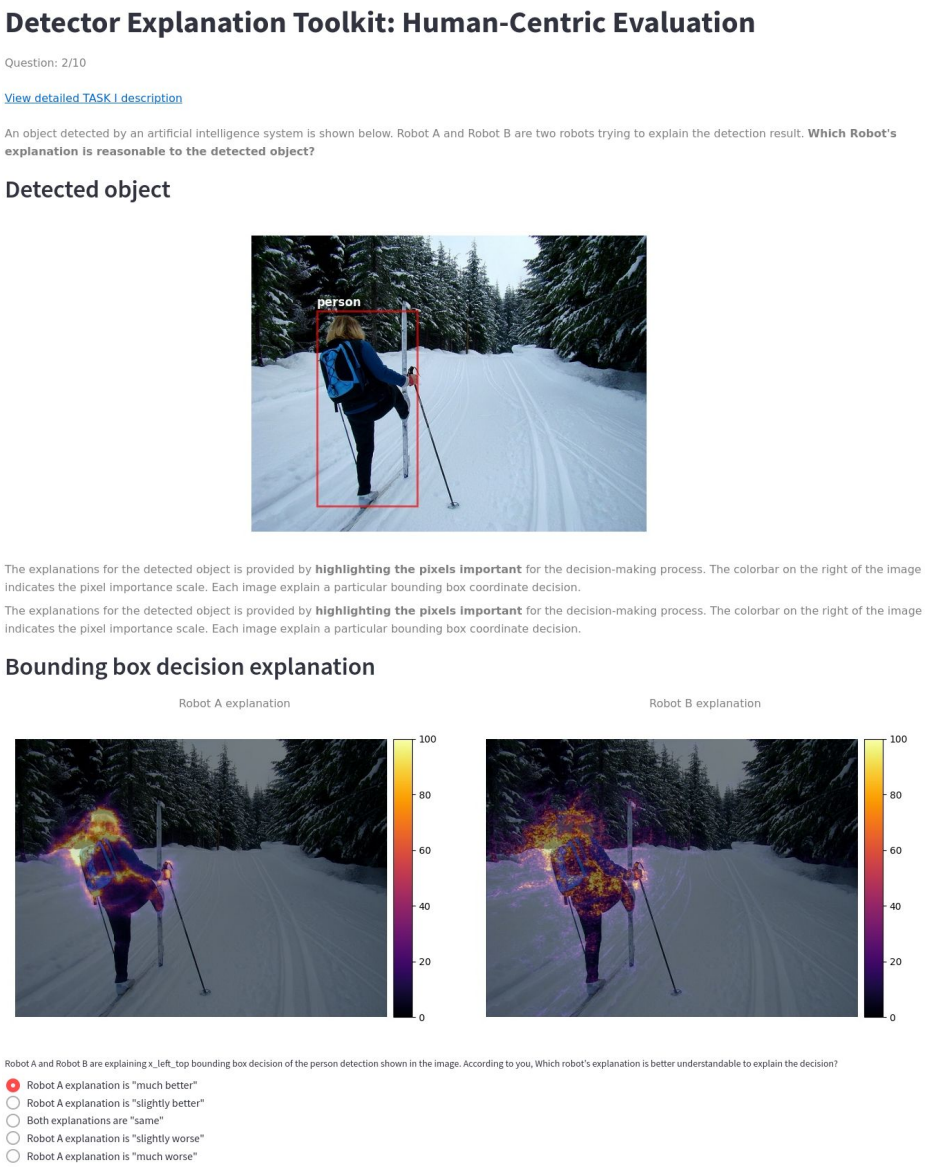


Fig. 38. Sample Task 1 question asked to rank explanation methods based on the user trust in the explanations for a particular detector decision. The figure is best viewed in digital form.

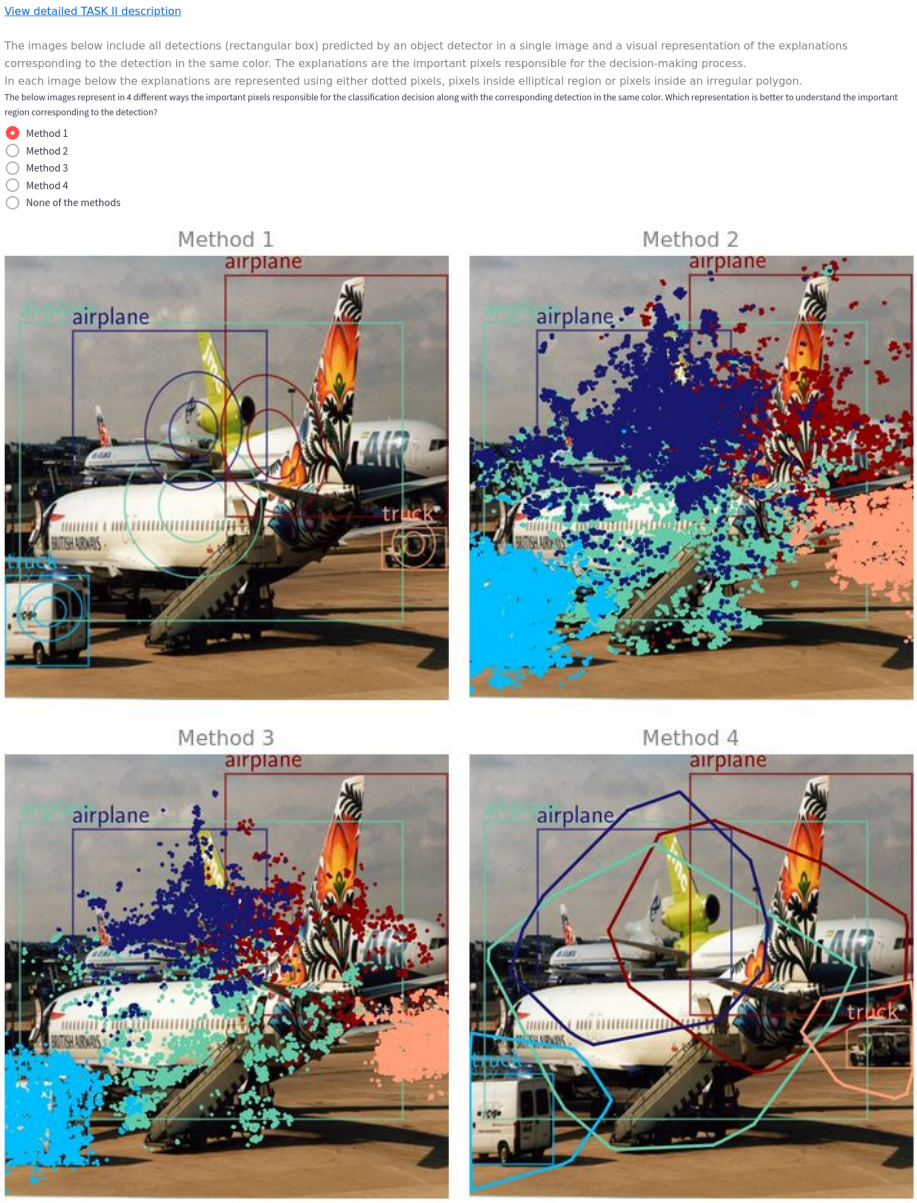


Fig. 39. Sample Task 2 question asked to rank the multi-object visualization methods depending on the user understandability. The figure is best viewed in digital form.

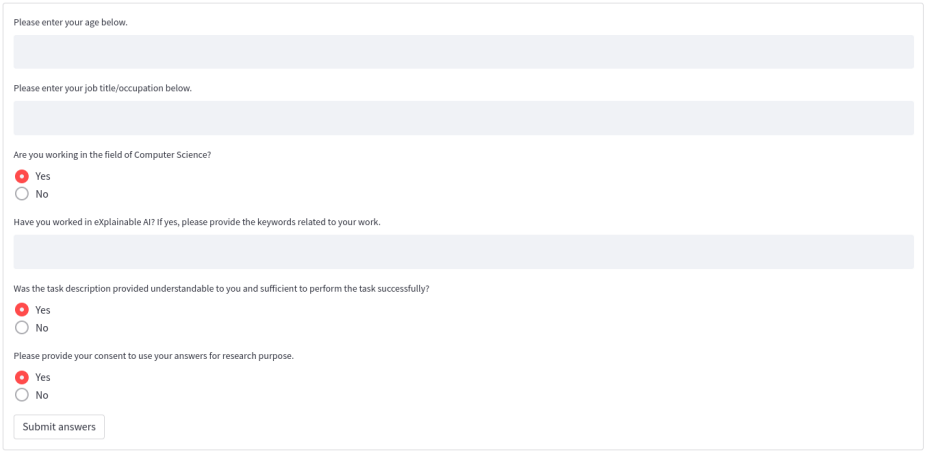


Fig. 40. Additional questions asked to understand the user background. The figure is best viewed in digital form.

I Screenshots of DExT

This chapter provides the screenshots of the DExT interactive application which is available online at: https://share.streamlit.io/deepanchakravarthipadmanabhan/dext/app.py.

The code to launch the application locally along with the DExT python-based package is available at https://github.com/DeepanChakravarthiPadmanabhan/dext.

Figures 41, 42, 43, and 44 shows the sequential process involved in analyzing an input image. Figure 45 illustrates the user interface provided to interactively generate explanations and evaluate the explanations for different detections across various explanation method and detector combinations.

Fig. 41. Illustration of the input image user uploaded by the user (left) and detections (right) made by the SSD512, the detector selected by the user, in the input image. The detectors available for off the shelf analysis are EfficientDet-D[0-7, 7x], SSD512, and Faster R-CNN.

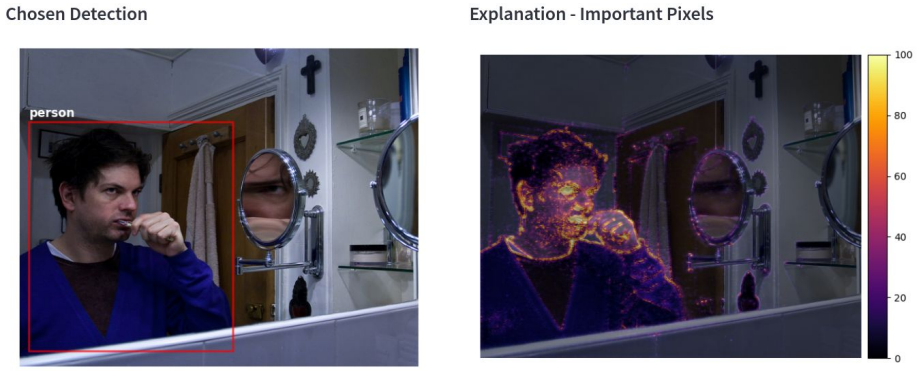


Fig. 42. Illustration of the interest detection (left) selected by user to generate explanation and saliency map (right) generated using GBP. The explanation interprets the classification decision for the interest detection. The interpretation methods available are GBP, SGBP, IG, and SIG. The interest detections are integer choices depending on the total detections in the image. The interest decisions are classification decision for the detected class and bounding box coordinates.

Fig. 43. Illustration of the single-box (left) and realistic (right) evaluation setting is provided as shown in DExT interactive application. Left: When the input image is the manipulated image by removing 80% of the most important pixels, the prior box detected as the output box for the original input image is shown. Right: The output detections for the manipulated input image. There are no output detections after removing 80% of the most important pixels.

Fig. 44. illustration of the manipulated image after removing 80% of the most important pixels depending on the generated saliency map.

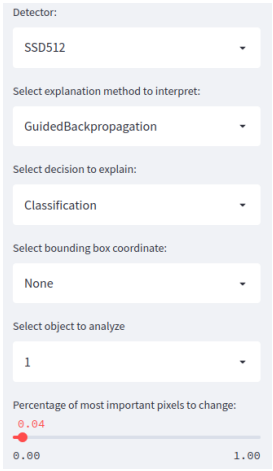


Fig. 45. User interface of the DExT interactive application. The user can select any detector, interpretation method, interest decision, and interest detection. In addition, a slider to control the fraction of input image pixels deleted is provided.