Revealing Hidden Context Bias in Segmentation and Object Detection through Concept-specific Explanations

Maximilian Dreyer[1], Reduan Achtibat[1], Thomas Wiegand[1,2,3],
Wojciech Samek[1,2,3,†], Sebastian Lapuschkin[1,†]
[1] Fraunhofer Heinrich-Hertz-Institute, [2] Technical University of Berlin,
[3] BIFOLD – Berlin Institute for the Foundations of Learning and Data
[†]corresponding authors: {wojciech.samek | sebastian.lapuschkin}@hhi.fraunhofer.de

Abstract

Applying traditional post-hoc attribution methods to segmentation or object detection predictors offers only limited insights, as the obtained feature attribution maps at input level typically resemble the models' predicted segmentation mask or bounding box. In this work, we address the need for more informative explanations for these predictors by proposing the post-hoc eXplainable Artificial Intelligence method L-CRP to generate explanations that automatically identify and visualize relevant concepts learned, recognized and used by the model during inference as well as precisely locate them in input space. Our method therefore goes beyond singular input-level attribution maps and, as an approach based on the Concept Relevance Propagation technique, is efficiently applicable to state-of-the-art black-box architectures in segmentation and object detection, such as DeepLabV3+ and YOLOv6, among others. We verify the faithfulness of our proposed technique by quantitatively comparing different concept attribution methods, and discuss the effect on explanation complexity on popular datasets such as CityScapes, Pascal VOC and MS COCO 2017. The ability to precisely locate and communicate concepts is used to reveal and verify the use of background features, thereby highlighting possible biases of the model. Code is available on https://github.com/maxdreyer/L-CRP.

1. Introduction

Deep Neural Networks (DNNs) have proven to be successful in providing accurate predictions in several critical object localization tasks, including autonomous driving [4] or medical screening [34]. However, the reasoning of these highly complex and non-linear models is generally not transparent [41, 43], and as such, their decisions may be biased towards unintended or undesired features [3, 29, 49]. In particular, object localization models sometimes base their decisions on features that lie outside of the segmentation mask or bounding box [22, 39], as shown in Figure 1, where the context of a hurdle is used for the segmentation of a horse. Understanding such contextual usage by DNNs is crucial to meet the requirements set in governmental regulatory frameworks and guidelines [12, 19].

In order to increase our understanding of DNN predictions, the field of eXplainable Artificial Intelligence (XAI) has proposed several techniques that can be characterized as local, global or a combination of both.

In literature predominantly addressed are *local* XAI methods for explaining single predictions. These methods compute attribution scores of input features, which can be visualized in form of heatmaps highlighting important input regions. However, in object localization or segmentation tasks, such attribution maps often resemble the prediction, *e.g.* the segmentation mask itself, and hence add little value for gaining new insights into the decision process of a model, *e.g.* when no context is used (see Figure 2 (*right*)).

Alternatively, *global* XAI aims to visualize which features or concepts have been learned by a model or play an important role in a model's reasoning in general. Nonetheless, it is not clear which features were actually used for a particular prediction or how they interact. For individual samples, latent activation maps of features can be visualized, that however are of low resolution and give no indication whether the feature was actually used or merely present in the input, as illustrated in Figure 1 (*right*).

Glocal concept-based explanations, on the other hand, offer a new dimension of model understanding compared to traditional heatmap-based approaches by not only indicating *where* the model pays attention to, but also informing about *what* it sees in the relevant input regions [1, 37, 58].

Our *glocal* methodology is hereby based on Concept Relevance Propagation (CRP) [1], an extension to local attribution methods based on the (modified) model gradi-

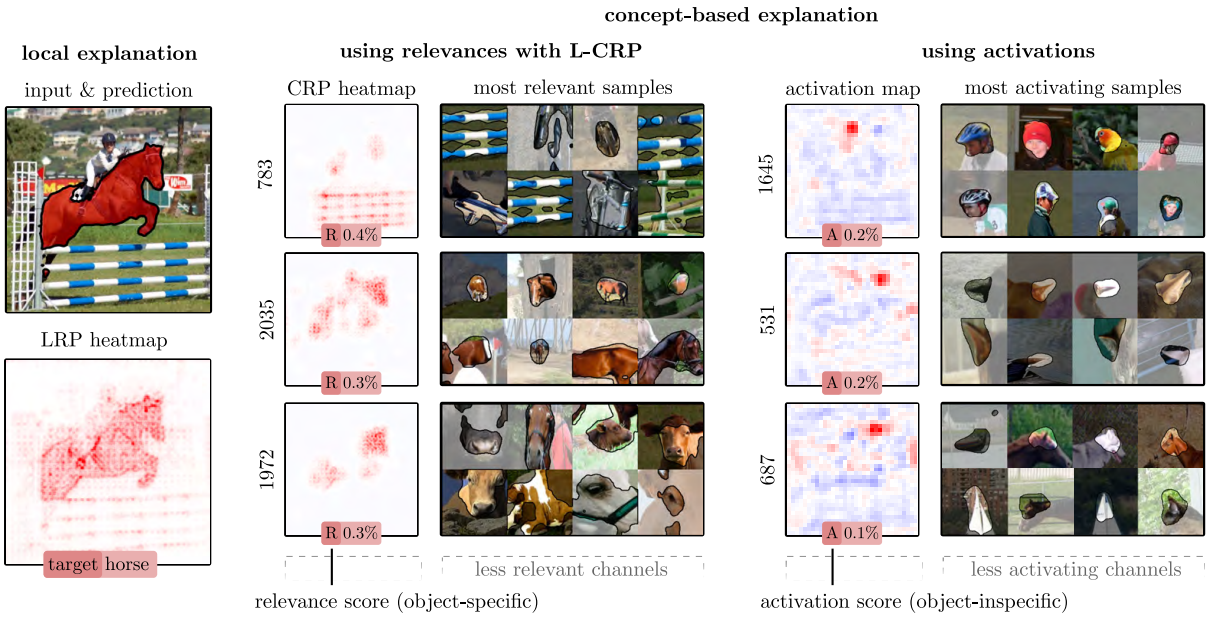

Figure 1. Our concept-based eXplainable Artificial Intelligence (XAI) approach (*center*) goes beyond traditional local heatmaps (*left*) or latent activation analysis (*right*), by communicating *which* latent concepts contributed to a particular detection, *to what degree* in terms of relevance (R) and *where* concepts are precisely located in input space. Whereas the shown heatmap from Layer-wise Relevance Propagation (LRP) only indicates the relevance of pixels overall, our method disentangles the latent space and makes it possible to understand individual concepts forming the prediction outcome. Especially in the multi-object context, it is crucial to attain object-class-specific explanations, which is not possible by the analysis of latent activations: Concepts with the highest activation values can refer to *any* class that is present in the image, such as the horse's rider, or none at all, since activations do not indicate whether a feature is actually used for inference.

ent, such as Layer-wise Relevance Propagation (LRP) [5]. Specifically, CRP allows to disentangle local attributions by computing relevance scores and heatmaps for individual latent features, and thus to localize concepts precisely in input space. Concept localization allows us in Section 4.2 to identify context biases and the use of background features, such as the "vertical bar" concept targeting the hurdle in Figure 1.

Contributions In this work, we address the limited explainability insight of heatmap-based local XAI methods for object localization tasks. We therefore present a novel method called CRP for Localization Models (L-CRP) for *glocal* concept-based explanations based on CRP for state-of-the-art architectures for semantic segmentation and object detection tasks, including UNet [17], DeepLabV3+ [11], YOLOv5 [14] and YOLOv6 [30] trained on the public datasets of CityScapes [13], Pascal VOC 2012 [16] and MS COCO 2017 [31]. Concretely,

1. We demonstrate how to explain state-of-the-art architectures for segmentation as well as object detection with LRP locally and by adapting CRP glocally on a concept level.

2. We evaluate our method in terms of faithfulness and explanation complexity using different concept attribution methods, including activation [56], gradient, GradCAM [48] and several LRP [5] parameterizations.

3. By localizing concepts in input space via concept-specific heatmaps and having object masks or bounding boxes available, we compute context scores for concepts indicating to which degree a concept is used for encoding background or object features and how the model makes use of the respective information during inference. We show how these insights can be used to detect biases in the model and verify our findings by interacting with the model.

2. Related Work

In the following, the current landscape of XAI methods for segmentation and object detection models is presented, followed by a description of concept-based XAI techniques.

2.1. Explainable AI for Segmentation

The literature of XAI for segmentation predominantly focuses on local techniques. While some works apply backpropagation-based methods such as Grad-CAM [15, 33, 48, 52] or LRP [2] to compute attribution heatmaps, others propose to apply perturbation-based methods [22, 53]. Contrary to classification tasks, local attributions are not

computed w.r.t. a single output class neuron, but w.r.t. the whole output feature map that forms the segmentation mask representing the class. Several such local XAI techniques have been compiled into the Neuroscope toolbox [45].

Alternatively, Losch *et al*. inspect latent features of segmentation models and introduce Semantic BottleNecks [32] modules to increase the latent space's human interpretability. They visualize and investigate intermediate filter activation maps, which, however, are of limited insight for understanding particular predictions due to not being class- and outcome-specific. Further, the fidelity of activation maps is limited to a convolutional channel's spatial resolution, as shown in Figure 1, where the attributions from L-CRP offer high resolution, as well as object- and outcome-specificity.

Alternatively, another group of works investigates inherently interpretable architectures, such as U-Noise [28], SegNBDT [53], MSGA-Net [24], and ProtoSeg [42] offering prototypical explanations. However, a large number of models applied in industry and research are not designed to be human-interpretable in the first place and thus require post-hoc methods for interpretation, such as ours.

2.2. Explainable AI for Object Detection

Similar to image segmentation, several local XAI methods have been presented for object detection. These techniques can be grouped into methods based on the (modified) gradient such as Gradient-SHAP [25], LRP [23, 50], Spatial Sensitive Grad-CAM (SS-GradCAM) [54] and EX2 [20], or input-perturbation techniques such as LIME [20] and masking [39, 44, 55]. Methods based on the gradient hereby explain the output class logit of a chosen bounding box analogously to the classification case. It is to note, that perturbation-based attribution methods require a high number of prediction re-evaluations by the model, resulting in run times in the order of minutes per data point, as *e.g.* for [39]. Our method is based on the (modified) gradient and, therefore, can be computed in the order of seconds [1].

2.3. Concept-based Explanations

In recent years, a multitude of methods emerged to visualize in a human-interpretable way concepts in the latent space learned by a model. A line of work [8, 9] assumes that individual neurons encode distinct concepts, others view concepts as directions described by a superposition of neurons [26, 51]. Similar to contemporary literature [1, 8, 9], we treat each neuron as an independent concept to achieve the highest granularity in explanations, while the method presented in this paper can be, in principle, also extended trivially to concept directions in the latent space.

In the image domain, contemporary work relies on activation maximization for visualizing concepts [18, 36, 40, 57], where in its simplest form, input images are sought that give rise to the highest activation value of a specific concept unit. However, high activation does not necessitate that corresponding input features are representative of a neuron's function, as adversarial examples illustrate. In this work, we make use of Relevance Maximization (RelMax) [1] that mitigates the aforementioned issues by choosing reference images from the original training distribution based on maximal relevance instead of activation.

Glocal XAI methods try to bridge the gap between the visualization of concepts on a global scale and attribution of their role during per-sample model inference. Shrouff *et al*. [47] combine TCAV [26] with Integrated Gradients to enable local attributions of neuron vectors, however, without offering localization of latent features in input space. NetDissect [7] assigns concepts to individual channels by computing the overlap between spatial activations with predefined segmentation masks. However, the fidelity of localization capabilities are limited to a convolutional channel's spatial resolution, as shown in Figure 1. Achtibat *et al*. [1] propose the idea of CRP, an extension of latent feature attribution methods based on the (modified) gradient, where a concept-specific heatmap can be computed by restricting the backward propagation of attributions through the network. This allows to attain precise concept-conditional heatmaps in input space, which we will use in Section 4.2 to investigate the use of background features by the model.

3. Methods

Our *glocal* concept-based method L-CRP is based on the principle of CRP, with LRP as the feature attribution method of our choice. Therefore, we first introduce LRP and CRP to attain concept-based explanations for individual predictions. Thereafter, we describe how concept-based explanations with L-CRP for segmentation and object detection can be obtained.

3.1. Layerwise-Relevance-Propagation

Layer-wise Relevance Propagation [5] is an attribution method based on the conservation of flows and proportional decomposition. For a model $f(x) = f_n \circ \cdots \circ f_1(x)$ with n layers, LRP first calculates all activations during the forward pass starting with f_1 until the output layer f_n is reached. Thereafter, the prediction score $f(x)$ of any chosen model output class is redistributed as an initial quantity of relevance R_n back towards the input layer after layer.

For a layer's output neuron j, the distribution of its assigned relevance score R_j towards its lower layer input neurons i is given by applying the basic decomposition rule

$$R_{i \leftarrow j} = \frac{z_{ij}}{z_j} R_j \,, \tag{1}$$

with z_{ij} describing the contribution of neuron i to the activation of neuron j. The aggregated pre-activations z_{ij} at output neuron j are represented by z_j with $z_j = \sum_i z_{ij}$.

The relevance of a neuron i is then simply an aggregation of all incoming relevance quantities

$$R_i = \sum_j R_{i \leftarrow j} \, . \tag{2}$$

In order to ensure robust decompositions and stable heatmaps, several purposed LRP rules have been introduced in literature [35, 43]. During the experiments in Section 4, we compare the rules of LRP-ε, LRP-γ and LRP-z$^+$, leading to different explanations in terms of faithfulness and complexity as shown in Section 4.1. Please refer to Appendix A.1 for a detailed description of used LRP-rules.

3.2. Concept-Relevance-Propagation

With CRP, the authors of [1] combine global concept visualization techniques with the local feature attribution method of LRP. A first step to the unification of local and global XAI is the realization that during the LRP backward pass, intermediate relevance scores are readily available, as computed in Equation (2). In order to also achieve concept-conditional heatmaps in input space, CRP firstly proposes to restrict the relevance propagation process via conditions.

Concretely, a condition c_l can be specified for one or multiple neurons j corresponding to concepts of interest in a layer l. Multiple such conditions are combined to a condition set θ. To disentangle attributions for latent representations, the relevance decomposition formula in Equation (2) is therefore extended with a "filtering" functionality:

$$R_{i \leftarrow j}^{(l-1,l)}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{ij}}{z_j} \cdot \sum_{c_l \in \theta_l} \delta_{jc_l} \cdot R_j^l(\mathbf{x}|\theta) \tag{3}$$

where δ_{jc_l} "selects" the relevance quantity R_j^l of layer l and neuron j for further propagation, if j meets the (identity) condition(s) c_l tied to concepts we are interested in. For layers without conditions, no filtering is applied. A concept-conditional heatmap can thus be computed by conditioning the modified backward pass of LRP via such a condition c_l for the concept's corresponding neuron or filter.

For the visualization of concepts, we adhere to the proposition of [1] and collect reference input images for which a latent neuron is most relevant, *i.e.*, useful during inference. Thereafter, by computing a conditional heatmap for the neuron of interest and reference sample, the relevant input part is further cropped out and masked to increase the focus of the given explanation on the core features encoded by the investigated neurons, as detailed in [1].

3.3. Extending Attributions to Object Localization

In order to obtain L-CRP for the generation of local explanations for segmentation and object detection models with CRP, the task-specific output vectors and maps have to be handled accordingly.

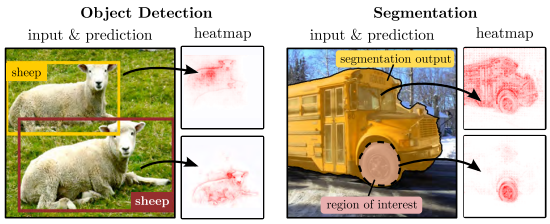

Figure 2. Local explanations of object detection (*left*) and semantic segmentation (*right*) with the LRP-z$^+$ rule. In the object detection case, we can explain each predicted object. For segmentation, an object's segmentation map or a part of it can be explained. Explanations with L-CRP can be found in the Appendix A.2

Semantic Segmentation Image segmentation models follow an encoder-decoder architecture in which the output mirrors (a scaled version of) the input dimensions in width and height. In contrast to the 1-dimensional output in classification tasks, the segmentation output consists of a 2d-map for each learned object category.

For simplicity, we assume a decision function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ for binary segmentation with a one dimensional input $\mathbf{x} \in \mathbb{R}^n$ and output $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^n$ of length n in the following. This is not necessarily a restriction to the input and output size, as any input can be flattened to a single dimension.

Here, any input feature x_i (*e.g.* pixel) might contribute to all the output values $f_j(\mathbf{x})$. Therefore, we can assign to feature x_i a relevance score for each output j. In fact, explaining a segmentation prediction can also be viewed as performing an explanation for each output pixel separately, and eventually adding the resulting attribution scores via, *e.g.*, a weighted sum. Here, a specific Region of Interest (ROI) can be selected to be explained by setting weights to zero for pixels outside the ROI. An example of a local explanation using two kinds of ROIs is shown in Figure 2 (*right*), where both the whole segmentation output of a bus or the bus wheels are explained via binary masking.

For modified backpropagation-based attribution methods such as LRP, the output tensor used as the starting point for relevance propagation can be adapted directly in order to control the meaning of an explanation. Please refer to Appendix A.3 for a comparison of different initialization schemes, including uniform, softmax or logit initialization.

For the initialization of relevance propagation at the output, we adhere to the in-literature often practiced approach by focusing on output scores corresponding to the highest class prediction. The relevance propagation for explaining class y is initialized in the last output layer L as

$$R_{(p,q,c)}^L(\mathbf{x}|\theta) = \delta_{cy} f_{(p,q,c)}(\mathbf{x}) \mathbf{1}_{(p,q)}(\mathbf{x}|y) \tag{4}$$

with an indicator function

$$\mathbf{1}_{(p,q)}(\mathbf{x}|y) = \begin{cases} 1, & \text{if } y = \text{argmax}_k f_{(p,q,k)}(\mathbf{x}) \\ 0, & \text{otherwise} \, . \end{cases} \tag{5}$$

The tensor indices p and q refer here to the spatial dimension $w \times h$ and channel (class) dimension c. Thus, not whole channel output maps are explained, but only the output values, which correspond to the selected class.

Object Detection Object detection networks often consist of a decoder part and a prediction module. The output per bounding box then includes class scores and information about the position of the bounding box. Local attributions are then similarly computed to the classification task w.r.t. the class score of a chosen bounding box.

For a bounding box predictor $f : \mathbb{R}^n \to \mathbb{R}^{N \times (n_c+4)}$ with an output of N bounding boxes for n_c object classes with four coordinates, the relevance $R^L_{(b,c)}(\mathbf{x}|\theta)$ of feature x_i for bounding box k of class y is then given by initializing the relevance propagation at the output as

$$R^L_{(b,c)}(\mathbf{x}|\theta) = \delta_{bk}\delta_{cy} f_{(b,c)}(\mathbf{x}) \tag{6}$$

with b representing the bounding box axis. Thus, an individual explanation can be computed for each bounding box, as shown in Figure 2 (*left*).

4. Experiments

The experimental section is divided into two parts, beginning with the evaluation of our concept-based XAI method using different feature attribution methods, thereby showing the superiority of computing latent relevance scores instead of activations – especially in the multi-object case. In the second part, we leverage the fact that we can localize the object through ground truth masks or bounding boxes and thus are able to measure how much the background instead of the actual object is used. We show how this can be used to detect possible biases in the model corresponding to specific feature encodings. We apply our method to a UNet on the CityScapes dataset, a DeepLabV3+ on the Pascal VOC 2012 dataset, a YOLOv5 and YOLOv6 model on the MS COCO 2017 dataset. Please refer to Appendix A.1 for further details on the models.

4.1. Evaluation of Concept-based Explanations

In recent years, several methods have been proposed to evaluate local explanations. Following the authors of [10, 21], we evaluate our presented method w.r.t. faithfulness and complexity. Faithfulness measures whether an attribution truly represents features utilized by the model during inference, while complexity measures how concise explanations are, which is of interest in context of *e.g.* human interpretation. Since the literature for evaluating local *concept-based* explanations is limited, we propose two experiments to test for faithfulness and complexity.

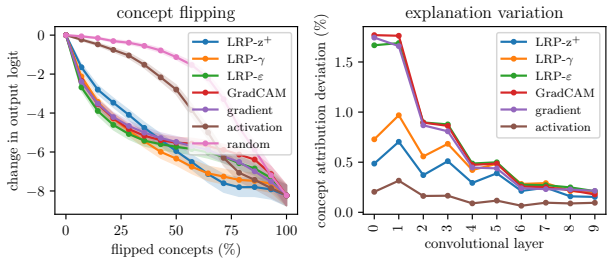

Figure 3. Measuring the faithfulness (*left*) and complexity (*right*) of concept attributions for the UNet architecture. (*Left*): Activations of the most relevant (or activating) concepts are set to zero successively, and the output difference measured. The Standard Error of Mean (SEM) over 100 samples is shown in semi-transparent color. (*Right*): The variation of attributions is lowest using activations and highest using gradient, GradCAM or LRP-ε.

4.1.1 Faithfulness of Concept Relevances

In order to assess the faithfulness of our concept-based explanations, we measure the impact on the decision outcome if a set of concepts is perturbed. This idea is analogous to the pixel flipping experiment in [5], only to use latent concepts instead of input features. We compare different LRP-rules, GradCAM, gradient (used by *e.g.* TCAV [26]) and activation (used by *e.g.* NetDissect [56]).

Concretely, we begin by computing the relevance scores of all concepts in a layer for a given object prediction. Please note, that since we conceptualize each convolutional channel to correspond to a distinct concept, the concept relevance of a channel is acquired via spatially sum-aggregation of intermediate relevance scores. Then, we successively deactivate the most relevant channels first in descending order by setting their activations to zero, re-evaluate the model output and measure output differences. Whereas the change in the class logit of a predicted bounding box is computed for object detection, for segmentation the mean change of the output mask's logits is calculated.

Alternatively, we also perform concept flipping backwards, by initializing all filters with zero activation and successively "unflipping" the most relevant concepts in descending order, *i.e.*, restoring their original activation values. This technique is designated as "concept insertion".

A faithful explanation is hereby characterized by a strong decline (incline) in performance if concepts are flipped (inserted). A typical experimental outcome is shown in Figure 3 (*left*) for layer `features.10` of the UNet model. As shown, perturbation of the most activating concepts is significantly less faithful than using relevance-based scores. This is expected, since relevance is object-specific and thereby filters out other activating concepts irrelevant for the current detection. Regarding concept relevances, the scores of LRP-ε, gradient and GradCAM are often char-

Table 1. Experimental results for evaluating various concept attribution approaches in terms of faithfulness (higher is better (↑)) and complexity (lower is better (↓)). For each approach, the evaluated scores for all models are displayed (UNet|DeepLabV3+|YOLOv5|YOLOv6).

	faithfulness (↑)		**complexity** (↓)	
	concept flipping	concept insertion	explanation variation	concepts for 80 % of attr. (%)
LRP-z^+	4.24 | 2.52 | 1.89 | 1.23	4.64 | 2.62 | 2.49 | 1.66	0.48 | 0.19 | 0.27 | 0.40	26.9 | 42.8 | 52.4 | 22.4
LRP-γ	4.51 | 2.78 | 2.07 | 1.34	5.16 | 2.75 | 2.63 | **1.72**	0.72 | 0.28 | 0.48 | 0.57	**22.2** | 34.0 | 41.0 | 18.1
LRP-ε	**4.54** | 3.25 | **2.43** | **1.42**	**5.38** | 3.29 | **2.79** | 1.42	0.84 | 0.65 | 0.73 | 0.94	26.5 | 37.8 | 34.9 | 23.1
GradCAM	4.41 | **3.79** | 1.81 | 1.11	5.27 | **3.91** | 2.37 | 1.25	0.81 | 0.60 | 0.89 | 0.96	27.1 | **28.3** | **25.0** | **13.8**
gradient	4.45 | 3.66 | 1.81 | 1.10	5.25 | 3.77 | 2.36 | 1.23	0.72 | 0.56 | 0.87 | 0.97	33.2 | 42.7 | 36.7 | 28.4
activation	2.83 | 2.10 | 1.49 | 0.82	3.49 | 2.36 | 2.17 | 1.14	**0.28** | **0.09** | **0.15** | **0.23**	61.4 | 63.2 | 68.3 | 45.9

acterized by the strongest decline/incline, as the gradient of the model faithfully measures the local sensitivity of the model to changes. However, as more features are perturbed, LRP-γ and LRP-z^+ often perform better, as they better represent the important features in a more global manner by filtering out noisy attributions [5]. In order to receive a score for a whole model, the faithfulness tests are performed in various layers throughout the models on 100 randomly chosen predictions, and the area under or over the curve measured per layer and mean-aggregated to form a final faithfulness score. As can be seen in Table 1, the results depicted in Figure 3 are reflected throughout all tested models, mostly showing the best scores for LRP-ε. Layer-wise faithfulness scores for all models can be found in the Appendix A.4.

4.1.2 Explanation Complexity and Interpretation Workload

While high faithfulness suggests that the concept attributions represent the model behavior correctly, they can still be noisy and not human-interpretable [27]. This effect is due to highly non-linear decision boundaries in DNNs [6]. In order to measure the complexity of explanations and the workload a stakeholder has to put in for understanding the explanations, two different measures are computed.

First, the standard deviation of latent concept attributions per class is measured, indicating the amount of noise. A low variation suggests, that explanations of the same class are similar, resulting in a lower amount of complexity [21]. As a second measure, the amount of concepts necessary to study in order to comprehend 80 % of all attributions is computed. The more relevance is focused on a small number of concepts, the fewer concepts need to be analyzed.

An example for measuring explanation complexity is shown in Figure 3 (*right*), where the explanation variation in the first ten convolutional layers of the UNet architecture is shown. Activations exhibit the smallest variation because filters activate on average more often if the feature is present in the image even though not used for inference. Regarding relevance-based concept attributions, gradient, GradCAM and LRP-ε show a high deviation especially in lower-level layers, indicating noisy attributions.

Whereas activation-based approaches lead to a low explanation variation, the distribution of concept attributions is rather uniform, leading to unconcise explanations and a large interpretation workload. Here, relevance-based approaches, which generate object-specific attributions result in smaller relevant concept sets. Plots for all architectures are illustrated in Appendix A.5.

The results of measuring explanation complexity for every model using all predictions in the test datasets are given in Table 1, confirming the previous observations in Figure 3.

Taking into account the results of the faithfulness tests, it is apparent, that relevance-based concept attributions show higher faithfulness than activation, but are not necessarily easier to interpret in terms of explanation complexity alone. Here, LRP-γ attributions show a good compromise between faithfulness and complexity in most experiments.

4.2. Concept Context Scores for Bias Detection

Ideally, a DNN trained with enough variety in training data learns abstract and generalized features. However, several works have shown that DNNs can develop biases or Clever Hans features, caused by spurious correlations in the data [3, 38, 46, 49]. Also in our experimental datasets, objects are displayed together with objects of other classes. A person is often pictured together with a kite, sitting on a horse, walking on the beach, or standing on a surfing board in splashing water. In fact, we can identify concepts for exactly the above-mentioned use cases, as shown in Figure 4.

4.2.1 Measuring the Context of Concepts

In order to automatically identify latent encodings used by a model to perceive an object's background (possibly corresponding to a bias), we propose to compute context scores. We compare latent activation maps, latent relevance maps (LRP, GradCAM and its spatial sensitive variant SS-GradCAM), and input attributions (Guided GradCAM [48] and L-CRP). We define the context score C of concept i as

Figure 4. Examples of background concepts for class "person" of the YOLOv6 model in layer `ERBlock_5.0.rbr_dense` with 512 concepts. For each concept, the mean relevance (R), context score (C) and context interpretation (white box) is given.

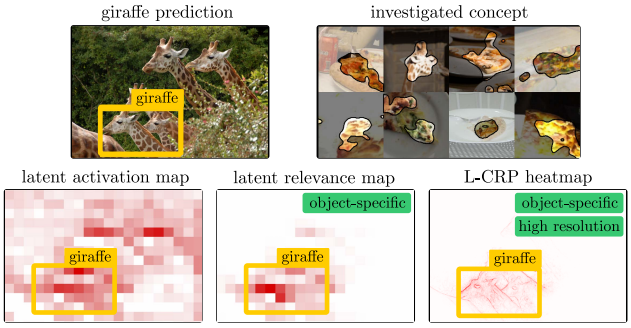

Figure 5. Measuring the context score of a concept encoding for the spotted texture of the giraffe skin (shown are most relevant concept samples) using the latent activation and LRP relevance map as well as L-CRP heatmap. L-CRP results in the most precise localization and is object-specific at the same time.

the fraction of positive attribution a^+ outside of the object bounding box compared to the overall sum, as

$$C_i = \sum_j^n \frac{1}{n} \frac{\sum_{p,q} a^+_{(p,q,i)}(\mathbf{x}_j)\bar{m}_{(p,q)}(\mathbf{x}_j)}{\sum_{p,q} a^+_{(p,q,i)}(\mathbf{x}_j)} \in [0,1] \qquad (7)$$

for n samples $\mathbf{x}_j$ and mask $\bar{m} \in \{0,1\}$ marking all background values with a value of one and zero else. Indices p and q refer to the spatial dimension of the attribution maps. In case of segmentation, an object is localized via pixel-accurate ground-truth masks, whereas for object detection bounding boxes are available. Masks or boxes are resized correspondingly to measure context scores in latent space.

Note, that latent feature maps can only be assessed for convolutional layers, but not for dense layers contained in most classification networks. L-CRP heatmaps, however, can be computed meaningfully for both layer types.

An example of computing context scores is shown in Figure 5 for the YOLOv6 model and concept 28 of layer `ERBlock_5.0.rbr_dense`, which corresponds to the spotted skin pattern of giraffes. As can be seen, latent activation maps are not object- and outcome-specific, leading

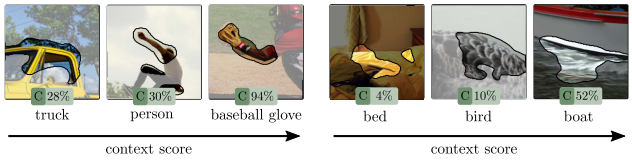

Figure 6. Context scores (C) of concepts vary between classes, as shown for concepts 2 (*left*) and 106 (*right*) of the YOLOv6 model.

to a broad activation on all giraffes in the image, and thus, to over-estimated context scores. For high-level layers, the high resolution of L-CRP maps leads to a more accurate context estimation than low-resolution latent maps.

Computing context scores, we found that concepts are used differently throughout classes (shown in Figure 6). The texture of waves, *e.g.*, can be found in bed blankets, feathers of birds, or in water surrounding boats, illustrating the significance of measuring context scores for *each* class.

4.2.2 Evaluating Context Scores

For evaluating estimated context scores, we propose to measure the model's sensitivity of concepts on the background of objects. We therefore inspect the influence on the concepts' relevances when the background is perturbed. Concretely, we define background sensitivity S of concept i as

$$S_i = \sum_j^n w_j \frac{|R_i(\mathbf{x}_j) - R_i(\tilde{\mathbf{x}}_j)|}{\max\{|R_i(\mathbf{x}_j)|, |R_i(\tilde{\mathbf{x}}_j)|\}} \in [0,1]\,, \qquad (8)$$

with concept relevances $R_i(\mathbf{x}_j)$, weights $w_j = \frac{|R_i(\mathbf{x}_j)|}{\sum_k |R_i(\mathbf{x}_k)|}$, and object samples $\tilde{\mathbf{x}}_j$ and $\mathbf{x}_j$ with and without perturbed backgrounds, respectively. Specifically, we apply gray-scale random noise, random noise, gray, and random color perturbation. Each perturbation is further performed with 100 % and 50 % alpha-blending, totaling 8 perturbations on 60 random detections. All context scores are computed for the 50 most relevant concepts of a class and the corresponding 15 most relevant detections, according to L-CRP attributions with the LRP-z$^+$-rule on the test data.

Ideally, concepts with a high context score C also have high background sensitivity S. As summarized in Table 2, L-CRP results in both the highest correlation and lowest Root Mean Square Deviation (RMSD) values between context and sensitivity scores. Here, we evaluate three layers of each model. Using activations (as *e.g.* NetDissect) leads to relatively high correlations, but large RMSDs as context scores are over-estimated. GradCAM effectively rescales latent activations, therefore also suffering from over-estimation. Specifically introduced for localization models, SS-GradCAM improves on GradCAM by using the spatial gradient information [54]. However, being reliant on the unmodified gradient, all GradCAM variants' localization capabilities are limited by noisy attributions caused

Table 2. Comparing computed context scores with measured background sensitivity. Ideal is a high correlation and low Root Mean Square Deviation (RMSD). RMSD values are given for all models (UNet | DeepLabV3+ | YOLOv5 | YOLOv6).

	RMSD (%)	correlation (%)
L-CRP (ours)	**19.8** | **18.0** | **15.7** | **16.7**	**69.4**
LRP	21.2 | 19.4 | 21.9 | 16.9	66.8
Guided GradCAM	27.2 | 19.3 | 25.5 | 25.7	43.0
SS-GradCAM	25.9 | 19.9 | 28.9 | 23.9	40.1
GradCAM	27.8 | 24.1 | 25.9 | 23.1	24.7
activation	47.6 | 35.5 | 37.7 | 41.4	65.1

through gradient shattering [6]. Further implementation details and discussions are given in Appendix A.6.

4.2.3 Context-based Interaction with the Model

The availability of context scores in combination with our method allows to detect the use of single background features, and thus to precisely interact with the model.

To probe the reaction of a model, an object's background can be manipulated. As an example, we noticed that frisbee detections rely on background concepts corresponding to dog features as shown in Figure 7. Removing the dog from the image via in-painting leads then to a missed detection.

Alternatively, as we are able to pinpoint background concepts in latent space, we can flip the corresponding concepts and measure the effect on predictions. With a dog present, concept flipping decreases the predicted output logit by about 6 %, as shown in Figure 7, whereas predictions without any dog present are not significantly influenced.

We further perform latent background concept flipping for frisbee or surfing board detections with "person" context features as well as person detections with "surfing board" concepts, visualized in Figure 7 (*bottom*). Here, the removal of person concepts affects the surfing board prediction more strongly than the other way around. In some examples shown in Appendix A.6, flipping three background concepts even leads to missed surfing board predictions. This is expected, as surfing boards are more likely to be depicted with a person (97 % co-occurrence) than a person with a surfing board (5 % co-occurrence) in the training data, favoring the likeliness of the model to use context features. The variety of "person" contexts is much higher, allowing the model to become more generalized in concept utility [29]. All flipped background concepts are visualized in Appendix A.6.

5. Conclusion

We propose L-CRP as an extension of the CRP method, enabling local concept-based understanding of segmentation and object detection models. By visualizing and localizing a model's internal representations utilizing L-CRP, the insight that a stakeholder gains can be significantly improved compared to traditional local XAI methods, which tend to merely resemble the object localization output. We evaluate several concept attribution methods, showing a trade-off between explanation faithfulness and complexity, with relevance-based attributions providing the best compromise. Finally, we apply our method to measure to what degree concepts refer to an object's context (background). Compared to other XAI methods, context scores derived from L-CRP most faithfully represent a model's use of concepts, as concept localizations are precise and object-specific. In experiments, we reveal potentially harmful context bias through context scores, enabling us to verify the model behavior by probing the corresponding encodings.

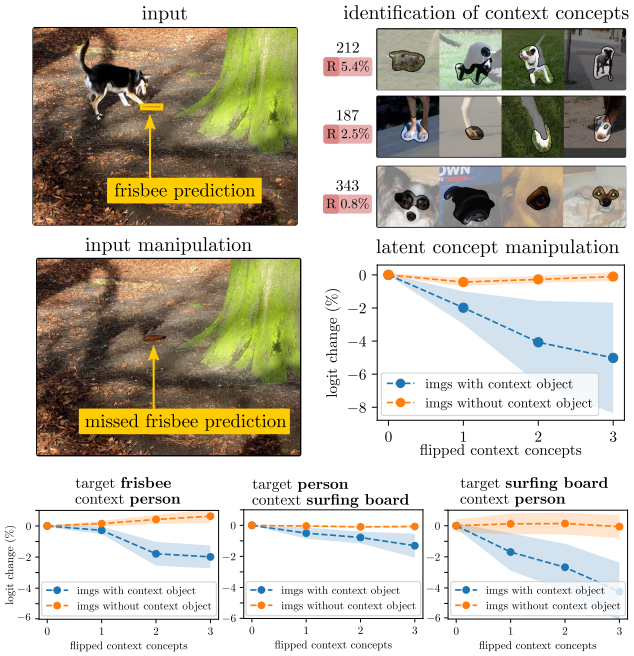

Figure 7. Dog concepts are relevant for a frisbee detection (*top*), leading to a missed prediction by the YOLOv6 model when the dog is removed from the image. We alternatively flip the dog concepts and measure a significant decrease in the output of the frisbee class for samples with both frisbee and dog present. We performed latent concept flipping also for other cases (*bottom*). The SEM is visualized in semi-transparent color.

Acknowledgements

This work was partly supported by the German Ministry for Education and Research under grants [BIFOLD (01IS18025A, 01IS18037I)], the European Union's Horizon 2020 research and innovation programme as grant [iToBoS (965221)], the German Research Foundation (ref. DFG KI-FOR 5363) and the state of Berlin within the innovation support program ProFIT as grant [BerDiBa (10174498)].

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From "where" to "what": Towards human-understandable explanations through concept relevance propagation. *arXiv preprint arXiv:2206.03208*, 2022. 1, 3, 4

[2] Awadelrahman MA Ahmed and Leen AM Ali. Explainable medical image segmentation via generative adversarial networks and layer-wise relevance propagation. *Nordic Machine Intelligence*, 1(1):20–22, 2021. 2

[3] Christopher J Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295, 2022. 1, 6

[4] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019. 1

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10(7):e0130140, 2015. 2, 3, 5, 6

[6] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *34th International Conference on Machine Learning (ICML)*, volume 70, pages 342–350, 2017. 6, 8

[7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 3

[8] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Àgata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proc. Natl. Acad. Sci. USA*, 117(48):30071–30078, 2020. 3

[9] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 5(3):e24, 2020. 3

[10] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pages 1383–1391. PMLR, 2020. 5

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[12] Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions. Communication: Building trust in human centric artificial intelligence. *COM*, 168, 2019. 1

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[14] Glenn Jocher et. al. YOLOv5n Nano models, Roboflow integration, TensorFlow export, OpenCV DNN support. *Zenodo*, 2021. 2

[15] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, 133:281–296, 2022. 2

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2

[17] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019. 2

[18] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 3

[19] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017. 1

[20] Denis Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, Yasunori Ishii, and Sotaro Tsukizawa. Explain to fix: A framework to interpret and correct dnn object detector predictions. *arXiv preprint arXiv:1811.08011*, 2018. 3

[21] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. 5, 6

[22] Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, and Volker Fischer. Grid saliency for context explanations of semantic segmentation. *Advances in neural information processing systems*, 32, 2019. 1, 2

[23] Apostolos Karasmanoglou, Marios Antonakakis, and Michalis Zervakis. Heatmap-based explanation of yolov5 object detection with layer-wise relevance propagation. In *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE, 2022. 3

[24] Meghana Karri, Chandra Sekhara Rao Annavarapu, and U Rajendra Acharya. Explainable multi-module semantic guided attention based network for medical image segmentation. *Computers in Biology and Medicine*, page 106231, 2022. 3

[25] Hiroki Kawauchi and Takashi Fuse. Shap-based interpretable object detection method for satellite imagery. *Remote Sensing*, 14(9):1970, 2022. 3

[26] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *35th International Conference on Machine Learning (ICML)*, pages 2668–2677, 2018. 3, 5

[27] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 6

[28] Teddy Koker, Fatemehsadat Mireshghallah, Tom Titcombe, and Georgios Kaissis. U-noise: Learnable noise masks for interpretable image segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 394–398. IEEE, 2021. 3

[29] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019. 1, 8

[30] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: a single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[32] Max Losch, Mario Fritz, and Bernt Schiele. Semantic bottlenecks: Quantifying and improving inspectability of deep representations. *International Journal of Computer Vision*, 129(11):3136–3153, 2021. 3

[33] Harsh Mankodiya, Dhairya Jadav, Rajesh Gupta, Sudeep Tanwar, Wei-Chiang Hong, and Ravi Sharma. Od-xai: Explainable ai-based semantic object detection for autonomous vehicles. *Applied Sciences*, 12(11):5310, 2022. 2

[34] Yoav Mintz and Ronit Brodie. Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies*, 28(2):73–81, 2019. 1

[35] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-Wise Relevance Propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 193–209. Springer, Cham, 2019. 4

[36] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 3

[37] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018. 1

[38] Frederik Pahde, Maximilian Dreyer, Wojciech Samek, and Sebastian Lapuschkin. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models. *arXiv preprint arXiv:2303.12641*, 2023. 6

[39] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452, 2021. 1, 3

[40] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017. 3

[41] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[42] Mikołaj Sacha, Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protoseg: Interpretable semantic segmentation with prototypical parts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1481–1492, 2023. 3

[43] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. 1, 4

[44] David Schinagl, Georg Krispel, Horst Possegger, Peter M Roth, and Horst Bischof. Occam's laser: Occlusion-based attribution maps for 3d object detectors on lidar data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1141–1150, 2022. 3

[45] Christian Schorr, Payman Goodarzi, Fei Chen, and Tim Dahmen. Neuroscope: An explainable ai toolbox for semantic segmentation and image classification of convolutional neural nets. *Applied Sciences*, 11(5):2199, 2021. 3

[46] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020. 6

[47] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*, 2021. 3

[48] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 6

[49] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018. 1, 6

[50] Hideomi Tsunakawa, Yoshitaka Kameya, Hanju Lee, Yosuke Shinya, and Naoki Mitsumoto. Contrastive relevance propagation for interpreting predictions by a single-shot object detector. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019. 3

[51] Johanna Vielhaben, Stefan Blücher, and Nils Strodthoff. Sparse subspace clustering for concept discovery (SSCCD). *arXiv preprint arXiv:2203.06043*, 2022. 3

[52] Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13943–13944, 2020. 2

[53] Alvin Wan, Daniel Ho, Younjin Song, Henk Tillman, Sarah Adel Bargal, and Joseph E Gonzalez. SegNBDT: Visual decision rules for segmentation. *arXiv preprint arXiv:2006.06868*, 2020. 2, 3

[54] Toshinori Yamauchi and Masayoshi Ishikawa. Spatial sensitive grad-cam: Visual explanations for object detection by incorporating spatial sensitivity. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 256–260. IEEE, 2022. 3, 7

[55] Yicheng Yan, Xianfeng Li, Ying Zhan, Lianpeng Sun, and Jinjun Zhu. Gsm-hm: Generation of saliency maps for black-box object detection model based on hierarchical masking. *IEEE Access*, 10:98268–98277, 2022. 3

[56] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018. 2, 5

[57] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. In *3rd International Conference on Learning Representations (ICLR)*, 2015. 3

[58] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 1