Classification Metrics for Image Explanations

Towards Building Reliable XAI-Evaluations

Benjamin Fresz*
Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Stuttgart, Germany,
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart
Stuttgart, Germany
benjamin.fresz@ipa.fraunhofer.de

Lena Loercher*
Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Stuttgart, Germany
lena.loercher@ipa.fraunhofer.de

Marco F. Huber
Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Stuttgart, Germany,
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart
Stuttgart, Germany
marco.huber@ipa.fraunhofer.de

ABSTRACT

Decision processes of computer vision models—especially deep neural networks—are opaque in nature, meaning that these decisions cannot be understood by humans. Thus, over the last years, many methods to provide human-understandable explanations have been proposed. For image classification, the most common group are saliency methods, which provide (super-)pixelwise feature attribution scores for input images. But their evaluation still poses a problem, as their results cannot be simply compared to the unknown ground truth. To overcome this, a slew of different proxy metrics have been defined, which are—as the explainability methods themselves—often built on intuition and thus, are possibly unreliable. In this paper, new evaluation metrics for saliency methods are developed and common saliency methods are benchmarked on ImageNet. In addition, a scheme for reliability evaluation of such metrics is proposed that is based on concepts from psychometric testing.

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**; *Machine learning*; *Computer vision*; • **Human-centered computing** → Human computer interaction (HCI); • **General and reference** → **Metrics**.

KEYWORDS

eXplainable AI, XAI, saliency maps, saliency metrics, heatmaps, quantitative evaluation, psychometric testing, validity, reliability, objective XAI evaluation

ACM Reference Format:
Benjamin Fresz, Lena Loercher, and Marco F. Huber. 2024. Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3630106.3658537

*Both authors contributed equally to this research.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658537

1 INTRODUCTION

In recent years, eXplainable Artificial Intelligence (XAI) has gained significant attention as a means to address the black-box nature of many Machine Learning (ML) models. XAI methods aim to provide transparency and interpretability, allowing users to understand the decision-making process of ML models. While various XAI techniques have been developed, their evaluation remains challenging, particularly in computer vision tasks. A common approach of explaining image classification and object detection decisions are so-called *saliency maps* that highlight image regions deemed particularly important for the prediction. The evaluation of such methods for image classification is essential to assess their effectiveness and compare different approaches. However, this is still an open problem despite various approaches to assess the properties of the saliency method, mainly due to the subjective nature of evaluations [28], the fallibility of user studies [13], and the different concepts used to evaluate such metrics [39]. It is particularly difficult to compare and assess saliency explanations beyond anecdotal evidence, as by definition they only provide local explanations, i.e., explanations for individual data points. A remedy for this can be ways to evaluate local explanations over entire datasets as in [6], resulting in a global assessment of explanation properties.

For XAI methods in general, the lack of ground-truth explanations complicates their robust assessment, sometimes attempted to be solved via creating specific datasets with ground-truth explanations [4, 7]. Additional to using such datasets, concepts from other disciplines with similar problems—lack of ground truth and tests with possibly differing underlying concepts—such as psychometric testing can be used [36].

The aim of this work is to further develop the ideas of Arias-Duart et al. [5, 6], which provide a set of metrics for saliency methods. This set is extended to a comprehensive list of metrics that mimic common metrics for classification evaluation based on the definition of correct and incorrect feature importance (FI) in images. Additionally, it is shown how the reliability (as part of validity) of the proposed metrics can be assessed, based on [36]. As such, our contributions are:

- We extend the list of saliency metrics. While [5] introduces some of them, others are overlooked. The additional metrics in particular provide interesting additional information, as shown in Section 5.
- We show how such metrics can be assessed regarding reliability (as a precursor to validity) in order to test them for their practical use.
- With the full set of saliency metrics and by adding B-cos networks [10] and the popular SHAP [21], we provide a more complete benchmark of XAI methods.
- We provide an in-depth discussion of the saliency metrics and show which properties of XAI methods (as given by Nauta et al. [24]) they address.

In the following chapter, the Focus score of Arias-Duart et al. [6] is briefly introduced, the saliency methods used in this paper and further works related to XAI-evaluation are discussed, with the proposed methodology being presented in Section 3, including the definition of (in)correct FI and the psychometric evaluation approach used here. The experiment setup and specifically created datasets are described in Section 4, followed by a selection of the results for the proposed metrics in Section 5, accompanied by a discussion of the limitations of the approach in Section 6. The paper is closed with a summary in Section 7.

2 RELATED WORK

This work builds on the idea of Arias-Duart et al. [6], where a metric for evaluating XAI methods is proposed. Since there is no ground truth for individual image pixels as to which class they can be assigned to, the authors have proposed a different way of evaluating explanations of saliency methods: they create mosaics from four images of various classes and assume that the evidence towards a class is more prevalent in images labelled with that class. The explanations, which are provided as FI by the examined XAI methods, are then evaluated by comparing positive feature attribution on images belonging to the correct class with positive feature attribution on the entire mosaic. Positive feature attribution here refers to the summed up feature attribution values of each pixel in the respective part of the mosaic.

In Section 2.1 the saliency methods considered in the original paper [6] as well as some additional methods examined in this paper are briefly presented. Section 2.2 addresses the state of the art with regard to XAI evaluation.

2.1 Saliency Methods

The saliency methods for which the Focus score was determined and analyzed in [6] will be described shortly in the following. Linear Interpretable Model-Agnostic Explanations (LIME) was one of the first methods to provide model-agnostic explanations in the form of feature attribution. The feature attribution is calculated by sampling around a data-point and fitting a simpler linear model to the weighted samples [27]. For images, LIME can create class-specific explanations by highlighting image regions—so-called superpixels—that are deemed especially relevant for the target class. Layer-wise Relevance Propagation (LRP), on the other hand, uses first-order Taylor expansions for local renormalization layers to generate saliency maps [8]. Integrated Gradients (IntGrad), which was proposed in [34], calculates the feature importance of an image by forming the gradient of the model output with respect to the model input and integrating this gradient over a baseline image ("neutral" input, e.g., grey image). Gradient-weighted Class Activation Mapping (Grad-CAM) produces class-specific saliency maps by computing gradient information in the last convolutional layer of a neural network [29]. A modification of Grad-CAM, namely Grad-CAM++, uses a weighted combination of the positive partial derivatives in the last convolutional layer, improving the performance of Grad-CAM for multiple objects of the same class in a single image and object localization [12]. Lastly, SmoothGrad describes the exchange of the often noisy gradient-based explanations by a weighted local average, thus possibly improving the visual quality and informativeness [31].

In this work, the range of examined saliency methods is extended to also include SHapley Additive exPlanations (SHAP) and B-cos. SHAP is not an explanation method itself but a unifying framework for feature attribution methods, especially Shapley Regression, Shapley Sampling, Quantitative Input Influence Feature Attributions, LIME, DeepLIFT, and LRP [21]. The game-theoretic interpretation of these methods, which are used to approximate Shapley Values (given certain hyperparameter choices), provides the possibility of receiving feature attributions with three desired criteria: Local accuracy, missingness, and consistency. KernelSHAP was chosen to approximate Shapley Values here, due to its comparatively low runtime. In contrast to all the post-hoc explanation methods described before, B-cos networks [10] generate model-inherent saliency maps by changing the activation functions of neural networks. This forces the network weights to align with the network-input and requires the networks to be trained with the B-cos transform as activation functions.

2.2 XAI Evaluation

In recent years, several approaches have been proposed for evaluating XAI methods. An overview of the methods published by the end of 2020 can be found in [24]. The authors list twelve properties of XAI methods that can be tested, often with various automated checks assessing (part of) a specific property. They group these properties into user-, presentation-, and content-properties, with six of them belonging to the last class and three to each of the previous ones. The content-properties are the most likely ones to be objectively measurable, although the authors of this paper expect that single metrics will most likely only assess a small subset of the available properties at once, as most of them are quite disjoint in their interpretation, e.g., covariate complexity—denoting how complex the (interactions of) features in the explanation are—and consistency—denoting how deterministic and implementation-invariant an explanation is—probably require quite different assessment methods. The metrics in this paper are therefore limited to assessing two of these properties: The *contrastivity* of saliency explanations, denoting how strongly an explanation discriminates between different outcomes of the ML model. An explanation that does not discriminate well would probably highlight general information such as edges in the mosaic images, thus resulting in bad saliency metrics. Additionally, with the assumption fulfilled that the used models are able to distinguish between the relevant

classes and evidence can mainly be found within images of the corresponding class, the *correctness* of saliency methods can be assessed (as described in the beginning of this chapter).

The explanation type and thus, the evaluation usually depends on the type of input data for which predictions or models need to be explained. Because of this, available toolboxes are limited to certain data types while still providing multiple evaluation metrics, e.g., [3] for tabular explanations and [17] for image explanations. Doshi-Velez and Kim [15] proposed three different stages of XAI evaluation, each with increasing effort and cost: functionally-grounded, human-grounded, and application-grounded. As part of the functionally-grounded evaluation and thus, early on in the development and implementation of XAI methods, metrics such as the one presented here can be used.

User studies are often viewed as the gold-standard of XAI evaluation, although their results have to be taken with caution as users tend to overestimate their understanding of the ML model [13, 37], which distorts the study results. In addition to user studies and metrics that can be evaluated on a specific use-case, more general documentation approaches have been suggested, for instance Explanation Fact Sheets [32], which contain information on relevant aspects of XAI methods. A similar approach, although more anecdotal and use-case specific, can be found in [9], which aims at providing a standardized format to assess and discuss trade-offs when evaluating saliency methods.

In the absence of established ways to compare XAI methods on a non-task-specific basis, so-called *sanity checks* can be used. These can test saliency methods for image classification for desiderata such as model-invariance and input-invariance [2, 19]. Even though a successful check does not provide enough information to fully trust a model, an unsuccessful one does show problematic behavior. Such sanity checks can also be formulated for object detection models [25], although the idea of general sanity checks and the ones which are not task-specific can be criticized due to possibly introducing a selection bias [40]. Other sanity checks involve the creation of "ground-truth" saliency maps that are compared with the generated explanations [18].

Rao et al. [26] propose a metric similar to Focus [6], which also only uses positive feature attributions, but limit all classes to appear in the mosaic at most once. The authors guarantee the basic assumption of class-specific features occurring exclusively in the target class by constraining the classifier to use the information from one part of the mosaic only. This is done by building one separate model head for every image in the mosaic. This ensures that no visual information between images is exchanged, limiting the classifier in its decision to rely on single images. The authors then define their metric based on whether the saliency methods still highlight other parts of the mosaic. Moreover, they evaluate the mosaics visually by humans via a systematic assessment approach that entails clustering them with their previously defined scores.

In [38] a benchmark of common saliency methods and evaluation metrics is provided. The work concludes that the evaluation results are inconclusive and the metrics in part contradict each other.

Finally, Tomsett et al. [36] investigate XAI evaluation metrics and present an approach from psychometric testing to assess them. Apart from [36], however, there is little research that addresses the topic of XAI metrics evaluation.

Table 1: XAI evaluation metrics proposed in this work. All of them can be calculated with the true positives, true negatives, false positives, and false negatives defined in Section 3.1.2.

Evaluation Metrics	Formula
Precision (Focus score)	$\frac{tp}{tp+fp}$
Sensitivity (Recall)	$\frac{tp}{tp+fn}$
Specificity	$\frac{tn}{tn+fp}$
False-Negative-Rate	$\frac{fn}{tp+fn}$
False-Positive-Rate	$\frac{fp}{tn+fp}$
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$
F1-Score	$\frac{2 \cdot \text{precision} \cdot \text{sensitivity}}{\text{precision}+\text{sensitivity}}$

3 METHODOLOGY

This section extends the Focus score for evaluating XAI methods from Arias-Duart et al. [6] by incorporating negative FI. First, the construction of mosaics is explained, followed by the definition of true positives and negatives, and false positives and negatives with respect to FI in the mosaics. These are then used to define additional evaluation metrics. In the second part, an approach from psychometric testing is introduced to examine the suitability of these metrics for evaluating XAI methods.

3.1 Proposed Metrics

To calculate the saliency metrics, so-called *mosaics* are used. They consist of a 2×2 grid of images of different classes from the original dataset. The idea behind them is that—given a model that is able to distinguish between the relevant classes—FI for a given class C should be attributed to the part of the mosaic that belongs to class C (as denoted by the labels in the original dataset). This then allows to calculated metrics akin to classical metrics for classification tasks, as described in the following.

3.1.1 Mosaics. The proposed approach adapts the procedure from the original Focus paper [6]. The mosaics used to test and evaluate various saliency methods are constructed of four images: two from the assigned target class and two from different classes within the same dataset. All images are selected randomly from their specific classes. The images are arranged in random positions in the 2×2 grid without overlap. To maintain consistency of visual patterns between mosaics and the training data, the individual images are scaled to a uniform size of 224×224 pixels. Accordingly, the mosaics have a resolution of 448×448 pixels. Because the individual images of the mosaics are part of the training data, the noise introduced by them is ensured to fall within the distribution of the training data. Figure 1 shows an example mosaic for each dataset considered in this paper. The datasets used for mosaic construction for the experiments are described in detail in Section 4.

3.1.2 True and False Feature Importance. For a more holistic evaluation of XAI methods, further metrics are defined in addition to the Focus score—the precision for FI—by considering negative FI. In general, pixels with a positive feature attribution value contribute to the prediction of the target class and pixels with a negative

Figure 1: One sample mosaic for each of the regarded datasets (cf. Section 4). On the left the mosaic comprises the ImageNet classes "tabby" and "sports car", in the middle "Bernese Mountain Dog" and "Greater Swiss Mountain Dog", and on the right the classes "lorikeet", "mashed potato", and "American chameleon".

feature attribution value contribute to the prediction of the other classes. Here, both are taken into account for the specification of true positives and negatives as well as false positives and negatives related to the FI on the mosaics. However, true and false FI can only be approximated, because the pixel-wise ground truth is unknown. This is done as follows:

- true positive (tp): positive FI on the images of the target class

$$tp = \sum_{x,y} \mathrm{FI}(c_{\mathrm{img}}(x,y) = c_{\mathrm{target}} \wedge \mathrm{FI} > 0)$$

  where $x, y \in \{0, 1\}$ are the image coordinates, i.e., $(0, 0)$ is the image on the bottom left, $(0, 1)$ is bottom right, $(1, 0)$ is top left, and $(1, 1)$ is top right. FI is the feature importance in the entire mosaic, c_{img} is the class label at position (x, y), based on the four images used to create the mosaic and c_{target} is the target class for the FI. The computation of the FI depends on the saliency method under investigation.
- false positive (fp): positive FI on the images that do not belong to the target class

$$fp = \sum_{x,y} \mathrm{FI}(c_{\mathrm{img}}(x,y) = \neg c_{\mathrm{target}} \wedge \mathrm{FI} > 0)$$

- false negative (fn): negative FI on the images of the target class

$$fn = \sum_{x,y} \mathrm{FI}(c_{\mathrm{img}}(x,y) = c_{\mathrm{target}} \wedge \mathrm{FI} < 0)$$

- true negative (tn): negative FI on the images of other classes

$$tn = \sum_{x,y} \mathrm{FI}(c_{\mathrm{img}}(x,y) = \neg c_{\mathrm{target}} \wedge \mathrm{FI} < 0)$$

Thus, by also considering negative FI, true and false negatives can be calculated in addition to true and false positives, so that a full confusion matrix can be defined. This enables the computation of metrics commonly applied in classification tasks, specifically for assessing saliency methods (cf. Table 1). With the additional metrics, XAI methods can therefore be evaluated more comprehensively.

Please note that not all XAI methods provide negative FI. Accordingly, the additional metrics can only be calculated for B-cos, IntGrad, LRP and SHAP. The other XAI methods can only be evaluated using the precision metric. An approach to examine the suitability of the metrics for evaluating the XAI methods is explained in the next section.

3.2 Evaluation Approach

Despite the absence of ground truth for evaluating the explanation methods and, consequently, the saliency metrics, the evaluation of certain properties of such metrics can still be conducted. Given the analogous challenges of lacking ground truth in psychometric approaches, corresponding evaluation procedures can be adapted to saliency maps, as proposed in [36]. Two fundamental concepts are *validity*, i.e., the extent to which a test or variable measures what it is intended to measure, and *reliability*, i.e, the consistency of results a test produces. While a reliable test does not guarantee validity, reliability is a necessary condition for validity [22]. In psychometric testing, the scenario is usually described by raters (e.g., psychologists) administering tests to a patient, where different types of reliability can be evaluated to assess whether a metric produces reliable and thus, possibly valid results. In this paper, two adapted reliability tests from [36] are considered.

3.2.1 Inter-rater Reliability. For the purpose of selecting a metric to choose between various saliency methods, this metric ideally yields the same ranking of saliency methods across all images of a dataset. When the ranking of saliency methods remains consistent across all (or most) test images, it is highly likely that the ranking for new images will be the same as well. This makes it easier to identify the best performing saliency method for future tasks. This paradigm can be compared to *inter-rater reliability*, where the images can be regarded as different raters administering a battery of tests to be scored by the saliency methods [36]. Intuitively, each image (rater) produces a ranking of saliency methods via the respective metric. This ranking can then be checked for agreement over all images (raters) across a dataset. Krippendorff's $\alpha \in [-1, 1]$

is a common statistic used to assess agreement between raters [20]. It is calculated as $\alpha = 1 - \frac{D_o}{D_e}$, where D_o denotes the disagreement observed and D_e denotes the disagreement by chance. A value of $\alpha = 1$ signifies perfect agreement in the ranking of saliency methods, α close to 0 indicates random rankings, and $\alpha < 0$ indicates systematic disagreement. The implementation used is the one by Castro [11].

3.2.2 Inter-method Reliability. The saliency precision can also be used to identify images or classes that are particularly challenging to classify within a given dataset [6]. In order to utilize an evaluation metric for this objective, difficult classes and images should be found consistently. This consistency should be independent of the saliency method employed, as the model used remains the same across all methods. Such a desideratum can be compared to *inter-method reliability*, which can be quantified using Spearman's ρ [36], which measures whether the relation of two variables X, Y can be described via a monotonic function (i.e., an increase in X also results in an increase in Y [33]). Spearman's ρ can be calculated as the Pearson correlation p between the ranks of X and Y, resulting in $\rho = p_{R(X),R(Y)} = \frac{\mathrm{cov}(R(X),R(Y))}{\sigma_{R(X)} \cdot \sigma_{R(Y)}}$, where $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ denote the standard deviations of the rank variables $R(X)$ and $R(Y)$, respectively, and $\mathrm{cov}(R(X), R(Y))$ the covariance of the rank variables [23]. A high value of ρ indicates that the saliency methods exhibit agreement in their variations across different images. Consequently, images with high (low) saliency metric scores for one method will consistently receive similarly high (low) scores across all saliency methods.

4 DATASETS

In the following, general information about the experiment setup is given, before the used mosaic datasets are described in more detail.

To compute the metric scores, the relative magnitude of the FI is used. For visualisation sake, the saliency maps are normalized via max-scaling, mapping them to the interval $[-1, 1]$. This is the straightforward extension of the normalization used by Arias-Duart et al. [6] to also work for saliency methods with negative FI. This normalization preserves the 0-point of the saliency maps and leaves the proposed metrics unchanged.

We evaluate the metrics on two different neural network architectures, with the comparatively small VGG architecture with a depth of 11 layers [30] and the larger ResNet architecture with a depth of 50 layers [16]. Since the VGG architecture contains batch normalization, its implementation differs slightly between the B-cos-version and the conventional one. To remove all bias terms in their networks, the authors of B-cos change the batch normalization to not contain a centering operation, resulting in a so-called "uncentered" batch normalization [10].

In general, it is difficult to disentangle the performance of the model and the performance of the explanation methods. Specifically, incorrectly distributed FI and thus, a low saliency metric score, could indicate either a high performing model with a low-fidelity explanation method (with feature attribution distributed evenly across a mosaic) or a bad model and a high-fidelity explanation method. To distinguish between these cases, the saliency scores and the model performance must always be considered in combination.

To test the reliability over different datasets and models and the amount of information provided by the saliency metrics, the scenarios described in the following subsections are tested. The hyperparameters used for the XAI methods can be found in Appendix B.1.

4.1 Corner Cases with Small Datasets

To establish the overall behavior of the proposed metrics, their performance and coherence with expectation can be tested in simple corner cases, for which clear expectations can be formulated. As corner cases, for which the metric behavior can be predicted, two datasets are used with two classes and 100 mosaics per class each. For these datasets, the classes are chosen from the ones represented in ImageNet, with images for the mosaics chosen from all class images at random. As all models are pretrained on ImageNet and a benchmark of their performance is most informative without any changes, the models are not fine-tuned on the specific datasets. This evaluation approach limits options regarding datasets, as only ones built from ImageNet (or at least with the same classes) can be used. Otherwise, fine-tuning of models or relabeling of classes would be necessary. This is contrary to Arias-Duart et al. [6], who adapt the network architecture to the number of classes in the dataset under consideration and thus, need to fine-tune their models. Our approach provides an unbiased assessment of popular models but also complicates the reporting of accuracy, as models trained on ImageNet might have learned high-level features like the difference between cats and dogs but not the specific difference between certain dog breeds, resulting in a low top-1 accuracy but in a high top-k accuracy for k > 1. Thus, both top-1 and top-5 accuracy need to be considered to make sure that the used models have learned relevant features to classify the datasets correctly.

A further difference to [6] is the creation of the mosaics: For the corner cases in Sections 4.1.1 and 4.1.2, samples from the training set of the ImageNet subset [1] are used for the mosaic construction, instead of only test data samples (where 'training' and 'test' refer to the corresponding partitions of this dataset). Since the aim of this work is to test the proposed metrics for the evaluation of different XAI methods (in contrast to e.g. performance evaluation of the networks), no relevant effects of leakage are expected. This assumption was confirmed in experiments with unseen datapoints of the same dataset.

4.1.1 Easy to Distinguish Classes. In the first corner case, the mosaics for the saliency metrics are created with two ImageNet-classes, which are expected to consist of very dissimilar features and thus, should be easily distinguishable by the pre-trained models. The target classes "tabby" and "sports car" are used and the dataset is referred to as the Cars/Cats dataset in the following. Since there should be (nearly) no overlap between the relevant features for both classes, next to perfect saliency metric scores are to be expected. One sample mosaic for this dataset can be seen in Figure 1 on the left.

4.1.2 Difficult to Distinguish Classes. The second dataset consists of mosaics built from two classes that look similar to laypeople and have strongly overlapping features. The classes chosen for this dataset are "Greater Swiss Mountain Dog" and "Bernese Mountain

Dog". The dataset is referred to as the Mountain Dogs dataset. For these mosaics, it is expected that the models will not be able to separate these classes, resulting in near-random performance and saliency metrics. One sample mosaic for this dataset can be seen in Figure 1 in the middle.

4.2 ImageNet

After testing the saliency metric behavior for corner cases, the third dataset uses all classes of the ImageNet dataset [14] as described in [6]. The mosaics constructed by Arias-Duart et al. [6] are available online and were used here. Compared to the other datasets, this dataset better represents most real-world computer vision tasks. For all of the 1,000 ImageNet-classes, mosaics are created, but due to hardware and runtime constraints, only ten mosaics per target class are feasible, resulting in 10,000 mosaics overall. The hardware used for the experiments and resulting runtime for this dataset is described in Appendix A.1. For the ImageNet dataset, the mosaics once again contain two target class images, with the other two images being chosen at random from all other possible classes, as can be seen on the right side of Figure 1.

5 RESULTS

In this chapter, the results and findings of the proposed saliency metrics are discussed. At first, the results for inter-rater and inter-method reliability are summarized, followed by some general findings with the saliency methods and metrics in Section 5.3. To give a better intuition for these results, Figure 2 provides an example of how the saliency maps differ between the saliency methods for ResNet50. The same can be seen for VGG11 in Figure 5 in the appendix. A more detailed view and discussion of the results can be found in the Appendix B.

5.1 Inter-rater Reliability

Krippendorff's α can be used as a metric for inter-rater reliability, indicating whether the ranking of XAI methods by a saliency metric is stable over all (or most) of the images in a dataset. Detailed results can be found in appendix B.4. In these, some tendencies emerge: The consistency of the saliency method ranking depends on the model type. In the experiments, ResNet50 almost always receives higher α-values than the VGG11-model.

SmoothGrad, LIME, Grad-CAM, and Grad-CAM++ provide only positive FI, thus only the precision-reliability (the reliability of the original Focus metric [6]) can be evaluated for all used saliency methods. For the datasets, the easier the models can distinguish between classes, the more reliable the precision ranking becomes, with the highest values for ResNet50 for the Cars/Cats dataset with $\alpha = 0.88$ and for ImageNet with $\alpha = 0.71$. The highest α-values for VGG11 are below 0.6, thus underpinning that the saliency method performance (and metric reliability) depends on the model type.

When the precision-reliability is evaluated for only B-cos, LRP, IntGrad and SHAP, the results are similar, but for these methods, additional saliency metrics can be calculated with negative FI. They follow similar trends: The easier the dataset, the higher the reliability, and in general higher reliability for ResNet50 than for VGG11, except for the false-positive-rate and specificity on ImageNet. Although not perfect, the inter-rater reliability for all classes of ImageNet (with values between 0.49 and 0.85) shows that the saliency metrics produce consistent results, thus enabling a user to choose between different saliency methods. Overall, the reliability of sensitivity, false-negative-rate, false-positive-rate, specificity, accuracy, and F1-score is higher than for precision, showing the added benefits of these metrics.

5.2 Inter-method Reliability

Spearman's ρ correlation between the results of the metrics for different saliency methods can be used to examine whether mosaics are consistently difficult to explain correctly for all methods. For ρ, some dependencies emerge: The correlation values for one metric differ between models on the same dataset, between different datasets for the same model and between the different metrics. For more detailed results, see Section B.2 and Figure 11 in the appendix. While most correlation values are rather low (< 0.8), in some cases for certain methods, correlations close to 1 can be seen, especially for the datasets for which the used classes were expected to be difficult to distinguish (especially on the Mountain Dogs dataset for Grad-CAM, Grad-CAM++ and SHAP). On these datasets, all XAI methods do not perform well based on the saliency metrics, thus possibly indicating a joint failure of certain saliency methods. For the other datasets, no clear correlation pattern emerges, with correlations < 0.8. Overall, the performances of the XAI methods can be highly correlated between some of them, given that the model is not able to distinguish well between different classes, while for more diverse datasets, the performances of the XAI methods are not strongly correlated. This could be paraphrased as: "*The saliency methods tend to work individually but some of them fail jointly.*"

5.3 General Findings

Additional to the more specific findings above, some general tendencies for the saliency methods can be identified.

B-cos highlights the upper left corner of images, possibly because the bias was removed in the network architecture, forcing the network to "create its own bias" via mostly irrelevant but stable features like image edges [10] (see Figure 2 for an example).

IntGrad does seem to yield mostly random performances in the metrics. Together with a good balance between positive and negative FI, this results in metrics close to 0.5 (see Figures 3c and 3d). As can be seen in Figure 3a, the precision for the Cars/Cats dataset for IntGrad is above 0.5, showing a performance better than random guessing as indicated by the other metrics. Visually, IntGrad explanations do mainly look like noise that seems to be stronger on the relevant image parts.

While some methods in theory provide negative FI, the magnitude of their positive importance is higher than the negative one, thus yielding misleading interpretations for some of the metrics when inspected on their own. This is illustrated, for example, by the precision and specificity in Figure 3: B-cos provides a high precision, but a low specificity, because it barely provides any negative FI-values and is tailored towards the correct attribution of positive FI. This imbalance towards positive FI is especially prevalent when the classes in the mosaics are more difficult to distinguish (see Figures 3c and 3d).

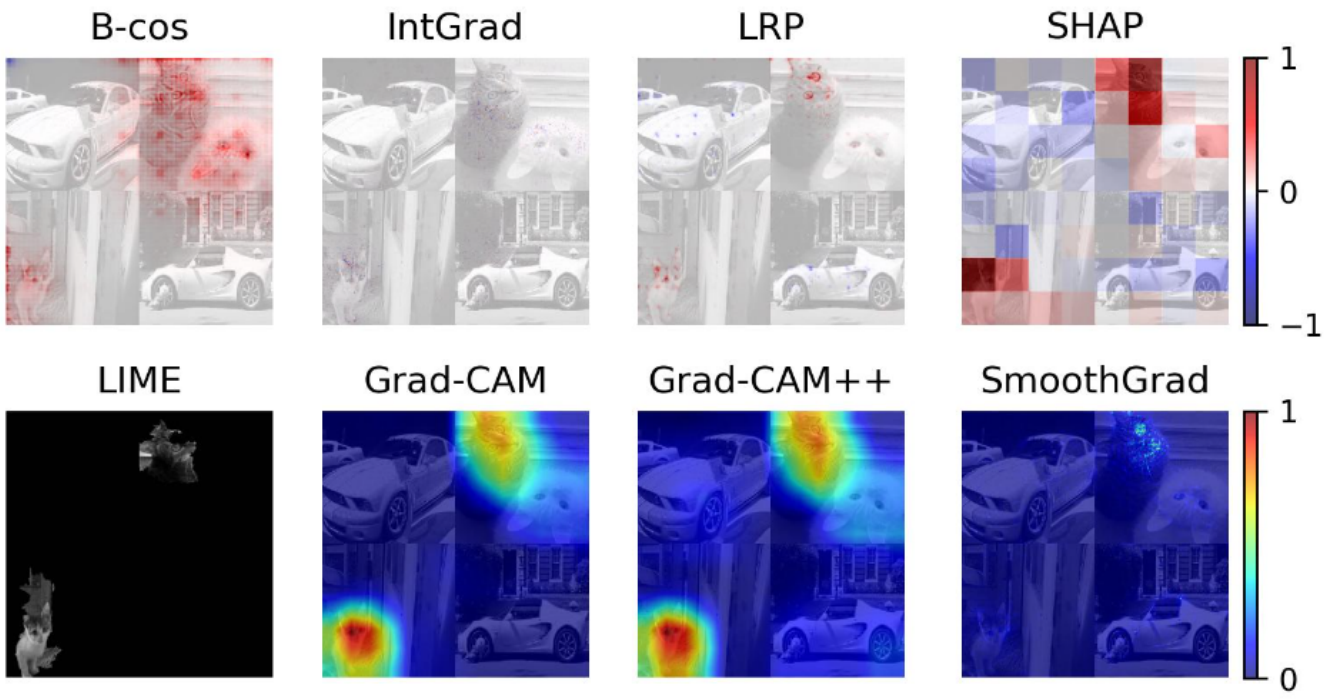


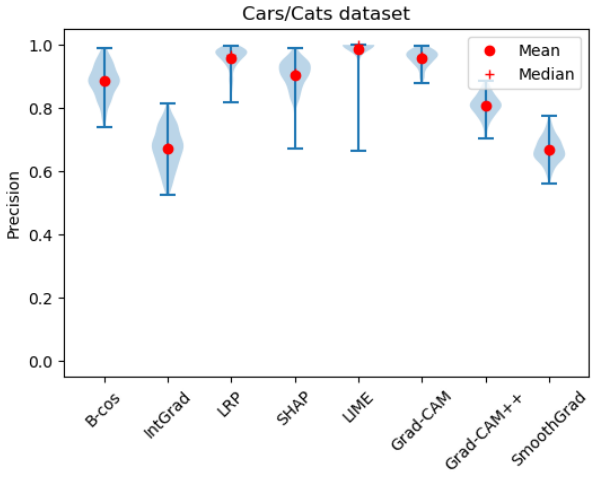
Figure 2: One sample heatmap by each saliency method for the first mosaic shown in Figure 1 of the Cars/Cats dataset. The explanations are created for ResNet50 for the target class "tabby". The upper row shows heatmaps for methods providing positive and negative FI, the lower one for methods with only positive FI. LIME uses a binary mask to highlight relevant image pieces, thus a binary masking of the original image is shown here. Similar results for VGG11 are presented in Figure 5 in the appendix.

Underpinning the initial intuition in creating the datasets in Section 4, the precision is significantly lower for Mountain Dogs than for Cars/Cats (Figure 3), with the precision for ImageNet somewhere in between.

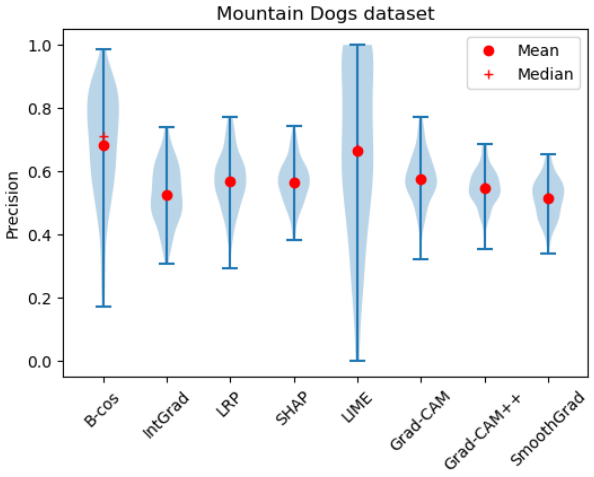
None of the tested methods provide good results in all of the metrics over different datasets, despite a sufficiently high classification accuracy for all datasets, showing that the models have learned relevant features (see Appendix A.2). While B-cos for example fared well in mean and median-performance for precision for all datasets, its specificity consistently produced values close to 0 (due to the higher prevalence of positive FI, see above and Figure 3a compared to Figure 3c). Here, it is important to note that the inter-rater reliability only measures the agreement over the saliency method ranking within a given metric but does not indicate that the different metrics lead to the same ranking of saliency methods. Complementary to the "eye-check" for B-cos and other methods as above, this aspect could be explored via Krippendorff's α between different metrics on the same dataset and model. Since the "eye-check" of the saliency method ranking between the metrics already showed that metrics usually produce different rankings and no saliency method performs well in all of them, this aspect was not explored further (Appendix B.3). An additional aspect of reliability-evaluation could be the agreement over mean and median performances of the saliency methods over different datasets, as this would show whether some methods consistently produce better performances on different datasets. As an analysis of this type of reliability just underpinned the previous results of ResNet50 providing more reliable results and method rankings mostly differing between datasets, a detailed discussion is ommitted here. For all of the datasets, large variances in the metrics can be discerned, thus indicating that some images produced almost perfect scores while others received scores towards the other end of the scale.

In classical literature, it has been long known that a single metric is not sufficient and multiple metrics are necessary to obtain a reliable assessment of a method, especially when some sort of unbalanced dataset is used [35]. In this paper, the existence of such an imbalance in the saliency methods was shown.

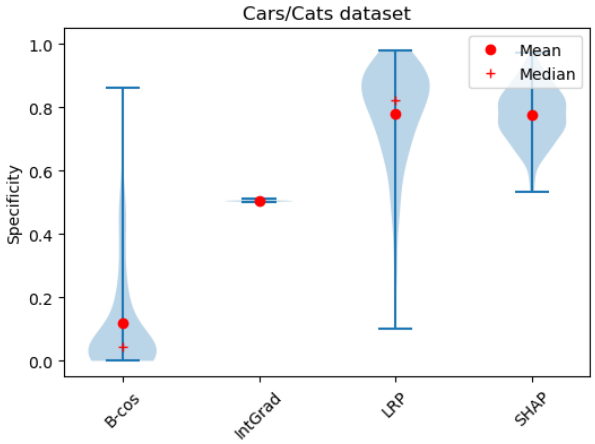
For ImageNet, the explanation methods recommended—at least for the properties of correctness and contrastivity—are LRP and SHAP, although both show clear weaknesses (see appendix B.3). LRP performs slightly better in some cases, but overall the performance of SHAP is more consistent for ResNet50 compared to LRP, especially for specificity and false-positive-rate. For VGG11, the variance of LRP is lower than for ResNet50, thus rendering LRP the best explanation method for this model for the ImageNet-mosaics. While a recommendation for saliency methods for a specific model and use-case can be made with the proposed metrics, the metric values and variances also show that no method performs to complete satisfaction (as, for instance, the highest mean-specificity for



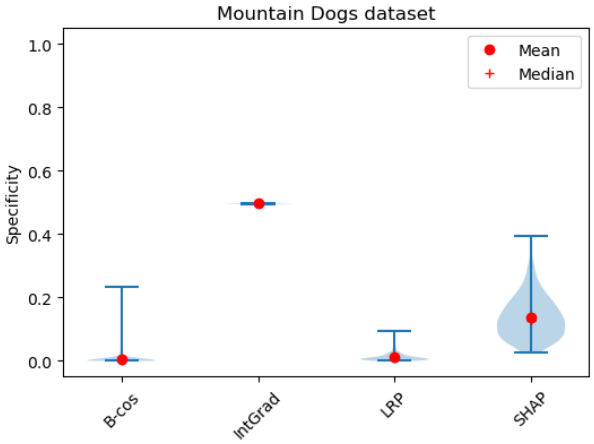
(a) Precision for all saliency methods on the Cars/Cats dataset.



(b) Precision for all saliency methods on the Mountain Dogs dataset.



(c) Specificity for saliency methods with negative FI on the Cars/Cats dataset.



(d) Specificity for saliency methods with negative FI on the Mountain Dogs dataset.

Figure 3: Exemplary results for precision and specificity for ResNet50 on the datasets with easier and more difficult to distinguish classes. Higher values are better. Note that specificity can only be calculated for methods which provide negative FI.

ResNet50 on the ImageNet-dataset is below 0.6), prompting more research for improved saliency methods.

Overall, the XAI methods perform differently in different scenarios, possibly because their underlying concepts of what constitutes important features differ [39]. The reliability assessment of the saliency metrics shows that the proposed metrics can help to choose between saliency methods, although it should be noted that a single saliency metric does not yield sufficient information for this choice for a given use-case. Instead, the combination of dataset, model, and XAI method needs to be evaluated to receive a meaningful assessment of the properties of saliency methods.

6 DISCUSSION

This paper examined the question whether the relevant features (i.e., the positive feature importance) for a class are actually located on the images of this class and if this fact can be utilized to define sensible metrics for evaluating saliency methods. Overall, this assumption holds, however, the introduced metrics are not exactly intuitive. They range between 0 and 1, where 1 can usually not be reached and 0.5 corresponds to random guessing. Additionally, if images of classes with very similar features are present in the mosaic (cf. Mountain Dogs dataset), the assumption is likely to be violated.

Another challenge for the metrics arises when, for example, all images of one class have the same background and this background is only present in this class in the dataset under consideration. In such a case, the metrics provide high (resp. low) scores for the generated explanations, but the explanations would show some sort of bias within the model when inspected by humans. It is important to note that the saliency metrics do not contain information about

the visual quality of explanations and rather correspond to a sanity check. For the saliency methods examined, higher (resp. lower) scores are always considered better, nevertheless, the metrics could be outsmarted: for instance, an explanation method that only attributes relevance to a single image pixel would lead to perfect scores but does not provide helpful information at all.

There are also some limitations of the methodology that need to be addressed: the random choice of images used for creating mosaics may introduce bias, e.g., by selecting images that are too (dis)similar and especially easy or difficult to distinguish. To mitigate this, the experiments, including mosaic creation, can be repeated multiple times or the number of generated mosaics can be increased. However, with the given number of mosaics, such effects are expected to balance out within the datasets used in this paper without exceeding runtime limits. Additionally, it is worth emphasizing that the B-cos method uses different model weights and activation functions than the other XAI methods, which could raise concerns about the direct comparability of explanation results.

Viewed from the outside, there is also the meta-level problem: saliency methods are used to understand and evaluate black box ML models. Saliency metrics are then used to evaluate the saliency methods and these metrics are then checked for reliability, etc. From a practical viewpoint, low-level information about which XAI methods to choose needs to be available without excessive amounts of work for evaluating different XAI approaches. But since there is no ground truth that can be used to verify statements at any level, the entire framework remains shaky. On the other hand, as there can probably not be a full ground truth explanation of a black-box-model that is different from just the model itself, it is necessary to employ the methods at hand to illuminate the underlying complexities at least to some extent. Therefore, it is crucial to always explicitly state the main assumptions of XAI methods and possible bias that may occur when using them, as these aspects are fundamental for selecting a suitable method for the respective model and dataset.

For future research, it would be interesting to extend the list of metrics to address further XAI properties listed in [24]. In addition, the proposed metrics could be applied to XAI methods on specific benchmark datasets to analyze and evaluate the resulting explanations.

7 SUMMARY

In this paper, new objective evaluation metrics for saliency methods were developed based on the definition of true (false) positive and negative FI in image mosaics. This definition required the assumption that evidence towards a certain class would be more prevalent in images of this class than in others, enabling the saliency metrics to use image mosaics as the basis of their calculation, mimicking common classification evaluation metrics. To test these metrics, datasets with mosaic images were created, small ones to evaluate corner cases with especially easy or difficult to distinguish classes and a larger one based on all classes of ImageNet.

For the practical use of a measurement, its validity—with its necessary condition of reliability—is crucial. Via inter-rater and inter-method reliability, the proposed metrics were established to be reliable in most cases, with the overall results showing that the performance of common saliency methods depends on the ML model and used dataset. As no clear correlation between the different saliency method results could be found, it seems that the saliency method performance also depends on the specific image being explained and goes beyond just single images being easy or difficult to explain. Due to their high inter-rater reliability, the proposed saliency metrics can be used to choose between different saliency methods for a specific use-case, although, due to the method's focus on positive FI, more than a single metric needs to be taken into account where possible. As these methods only assess the contrastivity and correctness of saliency metrics, we look forward to proposals of objective, reliable, and valid evaluation metrics for other properties of XAI methods and further reliability evaluations of other saliency metrics.

ACKNOWLEDGMENTS

Parts of this paper were created with the help of a company-specific implementation of ChatGPT (3.5 turbo). It was used to create LaTeX-code for tables and figures and to refine the drafts of some sections.

We thank the reviewers of the FAccT 2024 conference, who helped to improve this paper with their valuable feedback. This paper is funded in parts by the German Federal Ministry for Economic Affairs and Climate Action under grant no. 19A21040B (project "veoPipe") and by the Fraunhofer Gesellschaft under grant no. PREPARE 40-02702 (project "ML4Safety").

REFERENCES

[1] Wendy Kan Addison Howard, Eunbyung Park. 2018. ImageNet Object Localization Challenge. https://kaggle.com/competitions/imagenet-object-localization-challenge

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 9525–9536.

[3] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2023. OpenXAI: Towards a Transparent Evaluation of Model Explanations. arXiv:2206.11104 [cs.LG]

[4] Shideh Shams Amiri, Rosina O. Weber, Prateek Goel, Owen Brooks, Archer Gandley, Brian Kitchell, and Aaron Zehm. 2020. Data Representing Ground-Truth Explanations to Evaluate XAI Methods. arXiv:2011.09892 [cs.LG]

[5] Anna Arias-Duart, Ettore Mariotti, Darío García-Gasulla, and Jose Maria Alonso-Moral. 2023. A Confusion Matrix for Evaluating Feature Attribution Methods. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2023), 3709–3714. https://api.semanticscholar.org/CorpusID:260910875

[6] Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Giménez-Ábalos. 2022. Focus! Rating XAI Methods and Finding Biases. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (Padua, Italy). IEEE Press, 1–8. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882821

[7] Leila Arras, Ahmed Osman, and Wojciech Samek. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion* 81 (2022), 14–40. https://doi.org/10.1016/j.inffus.2021.11.008

[8] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In *Artificial Neural Networks and Machine Learning – ICANN 2016*, Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero (Eds.). Springer International Publishing, Cham, 63–71.

[9] Angie Boggust, Harini Suresh, Hendrik Strobelt, John Guttag, and Arvind Satyanarayan. 2023. Saliency Cards: A Framework to Characterize and Compare Saliency Methods. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 285–296. https://doi.org/10.1145/3593013.3593997

[10] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. 2023. B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers. arXiv:2306.10898 [cs.CV]

[11] Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

[12] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. https://doi.org/10.1109/wacv.2018.00097

[13] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

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

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[17] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* 24, 34 (2023), 1–11. http://jmlr.org/papers/v24/22-0142.html

[18] Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. 2021. Sanity Simulations for Saliency Methods. *CoRR* abs/2105.06506 (2021). arXiv:2105.06506

[19] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. The (Un)reliability of saliency methods. arXiv:1711.00867 [stat.ML]

[20] Klaus Krippendorff. 2004. Reliability in Content Analysis. *Human Communication Research* 30, 3 (2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[21] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874

[22] Kevin Murphy and Charles Davidshofer. 2004. *Psychological testing: Principles and applications (6th ed.)*. Pearson.

[23] Jerome L. Myers, Arnold D. Well, and Robert F. Lorch Jr. 2013. *Research Design and Statistical Analysis*. Routledge. https://doi.org/10.4324/9780203726631

[24] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* (feb 2023). https://doi.org/10.1145/3583558

[25] Deepan Chakravarthi Padmanabhan, Paul G. Plöger, Octavio Arriaga, and Matias Valdenegro-Toro. 2023. Sanity Checks for Saliency Methods Explaining Object Detectors. arXiv:2306.02424 [cs.CV]

[26] Sukrut Rao, Moritz Böhle, and Bernt Schiele. 2022. Towards Better Understanding Attribution Methods. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10213–10222. https://doi.org/10.1109/CVPR52688.2022.00998

[27] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938

[28] Yao Rong, Tobias Leemann, Thai trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards Human-centered Explainable AI: A Survey of User Studies for Model Explanations. arXiv:2210.11584 [cs.AI]

[29] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* abs/1610.02391 (2016). arXiv:1610.02391

[30] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[31] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. *CoRR* abs/1706.03825 (2017). arXiv:1706.03825

[32] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372870

[33] C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72. https://doi.org/10.2307/1412159

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. *CoRR* abs/1703.01365 (2017). arXiv:1703.01365

[35] Alaa Tharwat. 2021. Classification assessment methods. *Applied Computing and Informatics* 17, 1 (2021), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

[36] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. 2019. Sanity Checks for Saliency Metrics. arXiv:1912.01451 [cs.LG]

[37] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

[38] Benjamin Vandersmissen and Jose Oramas. 2023. On The Coherence of Quantitative Evaluation of Visual Explanations. arXiv:2302.10764 [cs.CV]

[39] David Watson. 2020. Conceptual Challenges for Interpretable Machine Learning. *SSRN Electronic Journal* (01 2020), 508. https://doi.org/10.2139/ssrn.3668444

[40] Gal Yona and Daniel Greenfeld. 2021. Revisiting Sanity Checks for Saliency Maps. *CoRR* abs/2110.14297 (2021). arXiv:2110.14297

A MODEL ANALYSIS

A.1 Runtime

A DGX A100 system with 40 GB of RAM-Memory was used to carry out the experiments described in Section 4. The code for generating saliency maps with varying saliency methods was executed on a 20 GB MIG slice of an NVIDIA A100 40 GB GPU. The runtimes for generating heatmaps for the ImageNet dataset from Subsection 4.2 with respect to different saliency methods can be seen in Figure 4. These runtimes highlight the performance benefits of gradient-based methods compared to the sample-based methods of LIME and SHAP.

A.2 Accuracy

For the model accuracy for mosaic datasets, see Table 2 for ResNet50 and Table 3 for VGG11. The accuracy is calculated for each class separately, with top-1 accuracy denoting whether class 1 or class 2 of a mosaic is predicted as the most likely class, top-5 accuracy denoting whether class 1 or class 2 are predicted in the five most likely classes. As the performance of these models is measured on the mosaic datasets, no true negatives nor false positives are to be expected, as all mosaics do contain images of the relevant classes. Because of this, only the accuracy can be calculated as a meaningful performance measure. These results show that both models predict the relevant classes for the mosaics often enough to expect them to have learned the relevant features for these classes. This aids in the assessment whether the assumption is fulfilled that models should highlight the images in a mosaic that correspond to the target of an explanation. The accuracy for the Mountain Dogs dataset is higher than for the others, possibly because all of the images in these mosaics contribute to the same target classes. This explanation is underpinned by the imbalance between positive and negative FI, as described in Section 5.3.

B EXPERIMENTS

B.1 Hyperparameters

Table 4 lists the hyperparameters used when executing the XAI methods in the experiments. This is done to ensure reproducibility of the results. Hyperparameters that equal the default value and methods that were used only with default hyperparameters are not included in the listing.

B.2 Results

For additional saliency maps for ImageNet mosaics see Figure 6 and for difficult to distinguish classes see Figure 7. Note that in the second case, the saliency is more evenly distributed across the different images in the mosaics. For both datasets, differences between the explanations for the model types can be seen. Compared to Figure 2, the distinction between classes is less clear in the saliency maps for ImageNet and even less for the Mountain Dogs dataset, resulting in worse saliency metric performance.

B.3 Metrics for ImageNet

In the following, a discussion of the results for the saliency metrics on ImageNet can be found. Figure 8 shows most metrics for ResNet50 for B-cos, IntGrad, LRP and SHAP, Figure 9 displays the same metrics for VGG11 and Figure 10 shows the F1-score for these methods and the precision for all considered saliency methods. For both models, some similarities can be observed: The saliency metric results of IntGrad are close to 0.5 with a low variance for all metrics but precision and F1-score. Overall, the range of saliency metric values is wide (sometimes spanning from 0 to 1), although with the values usually concentrated on a smaller range. This can be explained by some mosaics being easier and some more difficult to explain for each of the methods, which is expected when using random images from a dataset as diverse as ImageNet. While the B-cos models do perform well in some metrics (precision, sensitivity, false-negative-rate, accuracy and F1-score), they consistently perform bad in specificity and false-positive-rate (with values close to 0 and 1 respectively), showing their failure to attribute negative FI correctly and a strong bias towards positive FI as discussed in Section 5.3. Overall, based on the median performances, LRP and SHAP seem to be the best methods, with SHAP beating out LRP regarding the variance of specificity and false-positive-rate for ResNet50, while the distribution of saliency results for those metrics is better for LRP than for SHAP with VGG11. This behaviour does not show in the F1-score, which is a harmonized mean of precision and recall. But due to the higher magnitudes for positive FI, the F1-score mainly shows how well the positive FI is distributed. This is not surprising given that the F1-score can be rewritten to $\frac{2tp}{2tp+fp+fn}$. This shows that—if correct distribution of negative FI matters for a use-case—specificity and false-positive-rate should be considered along with one of the other metrics.

For the methods with only positive FI, solely the precision can be calculated. The results in Figure 10 show that the ranking of saliency methods differs between the two models, an effect especially prominent for LIME, which provides the best mean precision for ResNet50, but only the fifth-best for VGG11. Based on the precision, LIME seems to be the best-performing method for ResNet50 (although with a higher variance than the other methods) and Grad-CAM for VGG11. For both datasets, they are closely followed by B-cos, LRP and SHAP.

B.4 Inter-rater Reliability

For detailed results for inter-rater reliability for ResNet50, see Table 5, for VGG11, see Table 6. Note that some of the used saliency methods only provide positive FI, thus only the precision reliability can be calculated for them. These results can be found in Table 7. The findings for the inter-rater reliability are discussed in Section 5.1.

B.5 Inter-method Reliability

Detailed results for Spearman's ρ correlation can be found in Figure 11 for ResNet50 on the Cars/Cats dataset and on the Mountain Dogs dataset. These correlation values differ between the two datasets, as for the more difficult to distinguish classes (Mountain Dogs, Figure 11a), some methods yield highly correlated precision values, showing that these methods tend to perform similarly on the same images. This could be due to most images being difficult to distinguish and some showing clear differences (or none at all) between the dog breeds, thus effectively producing good (bad) performances on the same images. For the easier to distinguish classes

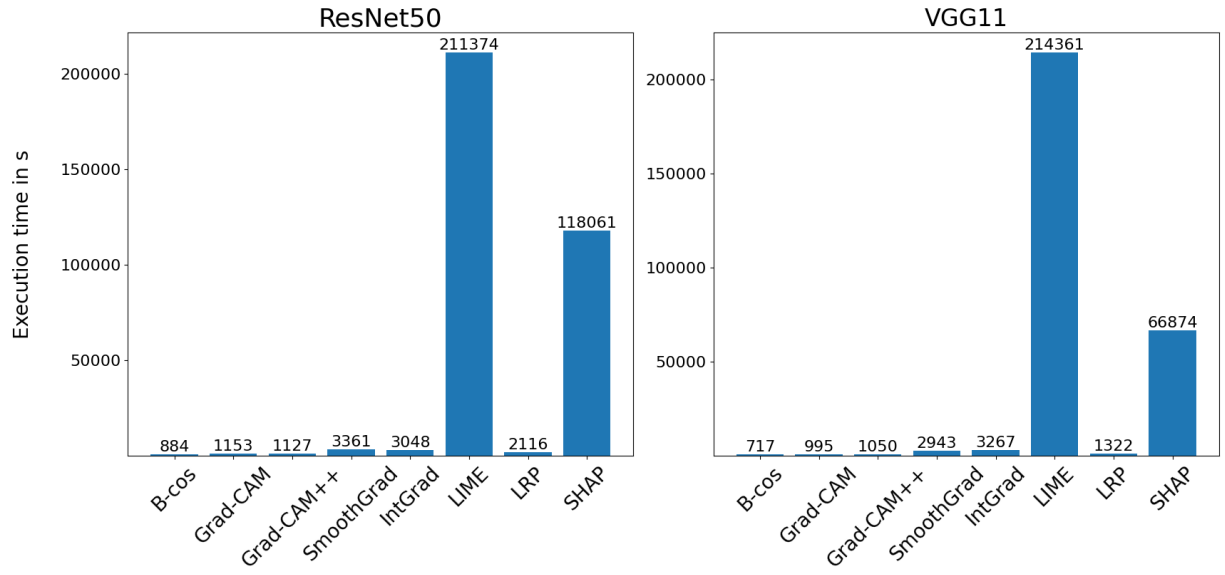


Figure 4: Visualization of the execution time of the different saliency methods in seconds. The time required to generate the saliency maps of every mosaic in the ImageNet dataset (cf. Subsection 4.2) was measured.

Table 2: Top-1 and Top-5 Accuracy for ResNet50 on the mosaic datasets, for the standard and for the B-cos model. Note that only the ImageNet-mosaics contain more than two classes.

ResNet50	top-1, class 1	top-5, class 1	top-1, class 2	top-5, class 2	top-1, class 3	top-5, class 3
Cars/Cats dataset	0.27	0.73	0.175	0.765	-	-
Cars/Cats dataset, B-cos	0.435	0.875	0.385	0.845	-	-
Mountain Dogs dataset	0.42	0.995	0.49	0.995	-	-
Mountain Dogs dataset, B-cos	0.345	1.0	0.41	0.995	-	-
ImageNet	0.6256	0.8683	0.0423	0.1971	0.0447	0.2148
ImageNet, B-cos	0.476	0.8095	0.1548	0.4727	0.1701	0.496

Table 3: Top-1 and Top-5 Accuracy for VGG11 on the datasets for each of the classes represented in the mosaics, for the standard and for the B-cos model. Note that only the ImageNet-mosaics contain more than two classes.

VGG11	top-1, class 1	top-5, class 1	top-1, class 2	top-5, class 2	top-1, class 3	top-5, class 3
Cars/Cats dataset	0.32	0.52	0.28	0.535	-	-
Cars/Cats dataset, B-cos	0.375	0.805	0.435	0.86	-	-
Mountain Dogs dataset	0.405	0.995	0.42	0.985	-	-
Mountain Dogs dataset, B-cos	0.31	0.99	0.325	0.98	-	-
ImageNet	0.3451	0.5933	0.0468	0.133	0.0447	0.1384
ImageNet, B-cos	0.3996	0.685	0.1265	0.328	0.1386	0.3408

Table 4: List of hyperparameters used when executing the XAI-methods during the experiments.

Saliency Method	Hyperparameters
LIME	$num\_samples = 1000$
SHAP	$num\_samples = 1500$, $super\_pixel\_size = 56$ (equals $8 \times 8 = 64$ superpixels per mosaic)

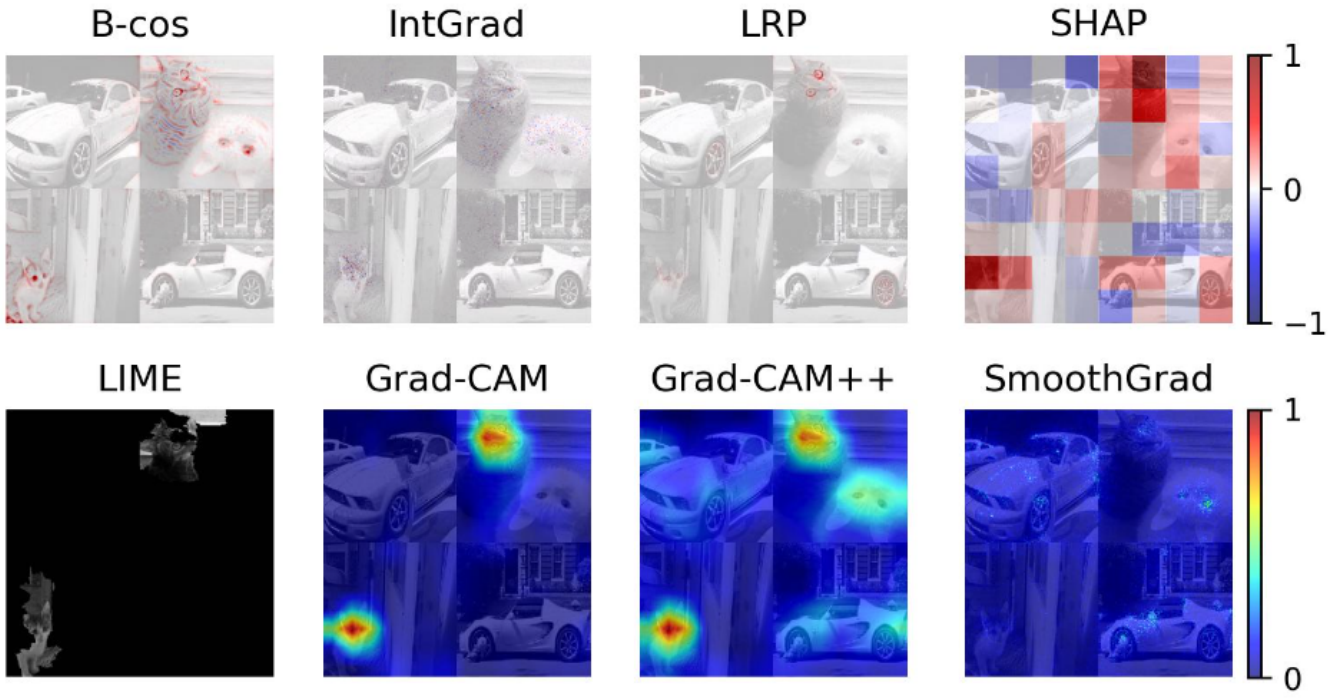


Figure 5: One sample heatmap by each saliency method for the first mosaic shown in Figure 1 of the Cars/Cats dataset, here for VGG11. The upper row shows heatmaps for methods with positive and negative FI, the lower one for methods with only positive FI. LIME uses a binary mask to highlight relevant image pieces, thus a binary masking of the original image is shown here. Note the differences to the explanations for the same image for ResNet50 in Figure 2.

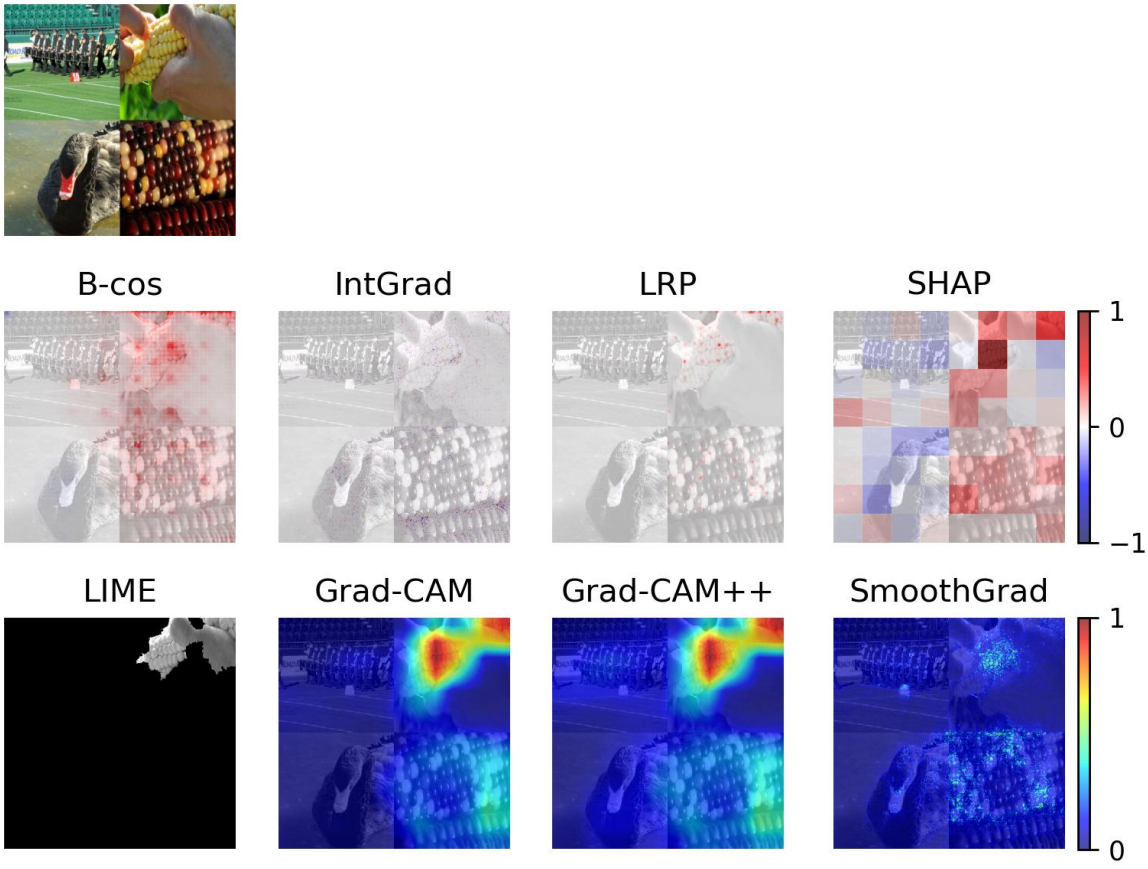
Table 5: Krippendorff's α for B-cos, IntGrad, LRP and SHAP for all metrics for ResNet50.

	Precision	Sensitivity	False-Negative-Rate	False-Positive-Rate	Specificity	Accuracy	F1-Score
Cars/Cats dataset	0.81	0.89	0.89	0.82	0.82	0.83	0.85
Mountain Dogs dataset	0.24	0.98	0.98	0.96	0.96	0.35	0.67
ImageNet	0.61	0.85	0.85	0.49	0.49	0.69	0.74

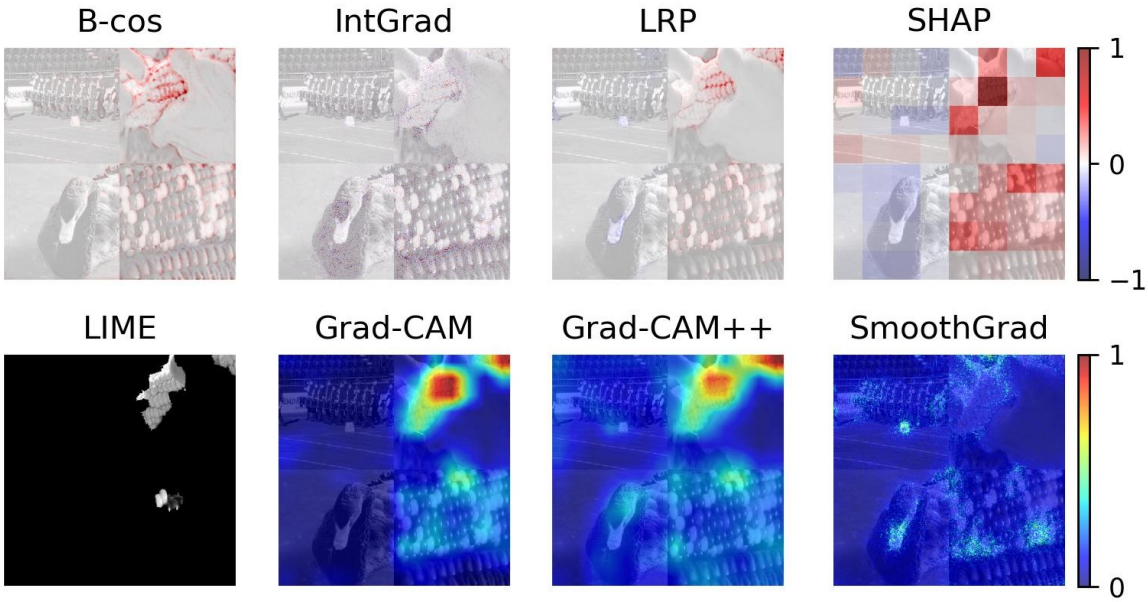
Table 6: Krippendorff's α for B-cos, IntGrad, LRP and SHAP for all metrics for VGG11.

	Precision	Sensitivity	False-Negative-Rate	False-Positive-Rate	Specificity	Accuracy	F1-Score
Cars/Cats dataset	0.56	0.59	0.59	0.79	0.79	0.64	0.56
Mountain Dogs dataset	0.15	0.72	0.72	0.67	0.67	0.14	0.52
ImageNet	0.52	0.58	0.58	0.67	0.67	0.62	0.57

(Cars/Cats, Figure 11b), no clear correlation tendencies exist, with the highest value below 0.6 and most being close to 0, with some even below 0. This shows that the saliency methods do not consistently agree on which of the mosaics in this dataset is easier or more difficult to attribute correctly. Since this dataset approximates real-world use-cases better than the Mountain Dogs dataset, it can be concluded that for real-world applications, saliency methods will likely struggle with different images and the difficulty of explaining a decision is not inherent to images but related to the used saliency method. In Section 5, this was summarized as "*The saliency methods tend to work individually but some of them fail jointly*".

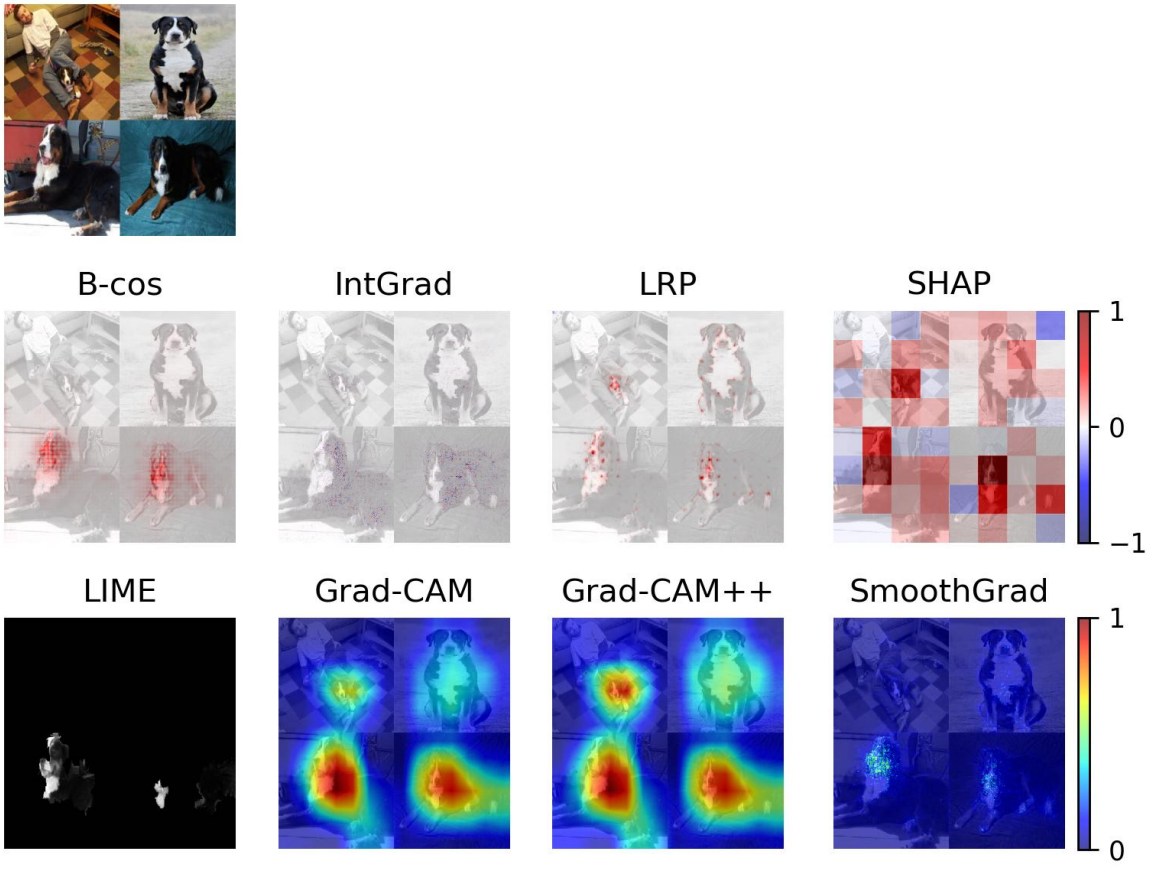


(a) Saliency maps for ResNet50 on an ImageNet mosaic.

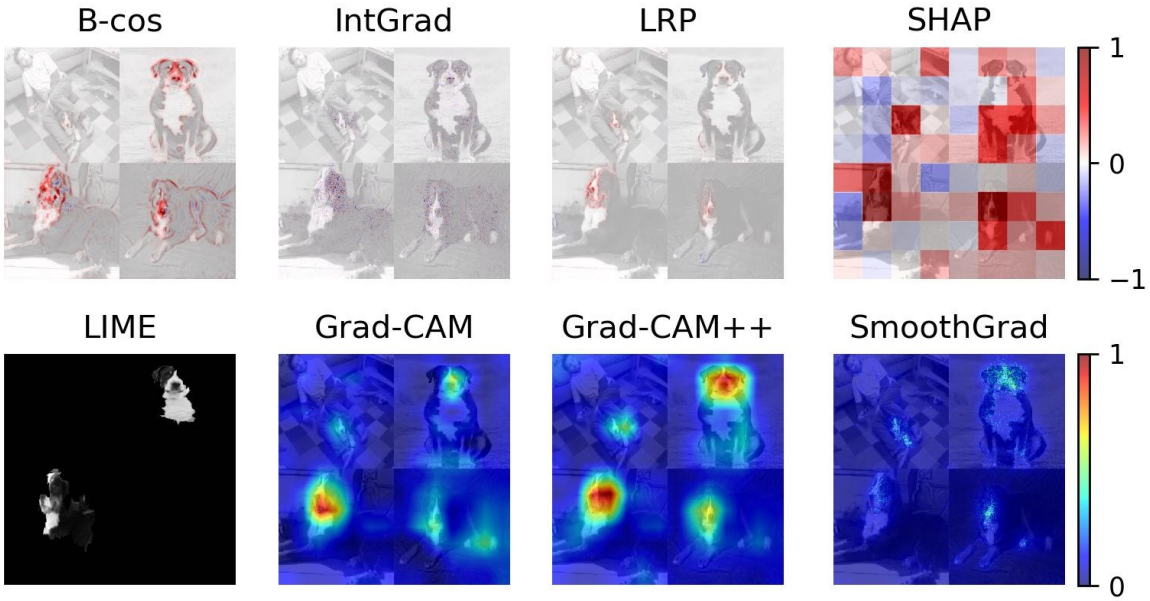


(b) Saliency maps for VGG11 on an ImageNet mosaic.

Figure 6: Saliency maps for the ImageNet mosaic shown as the first image in Figure 6a. The saliency was calculated regarding the target class "ear, spike, capitulum", to which the right two images in the mosaic belong.

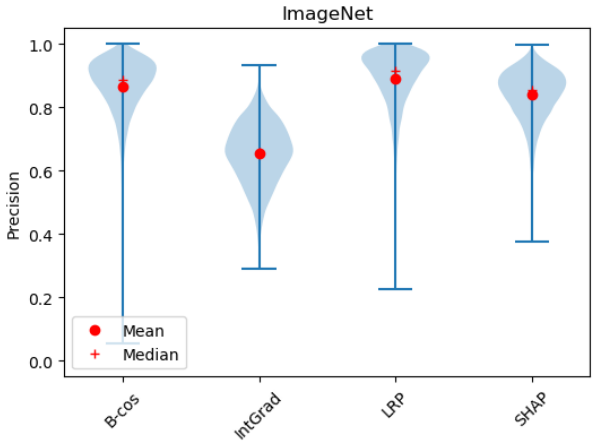


(a) Saliency maps for ResNet50 on a mosaic of the Mountain Dogs dataset.

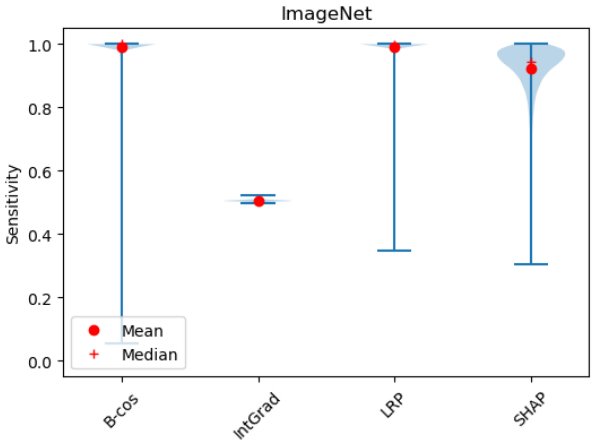


(b) Saliency maps for VGG11 on a mosaic of the Mountain Dogs dataset.

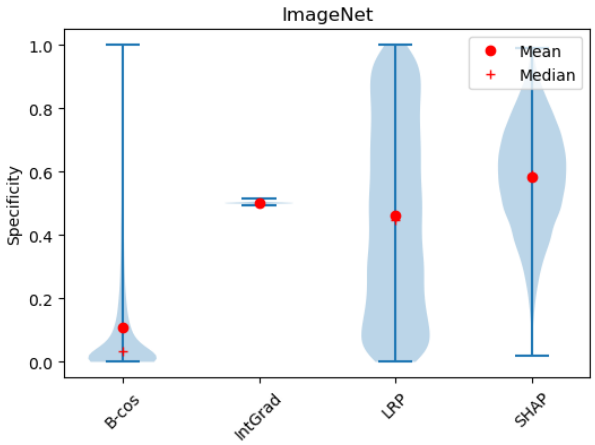
Figure 7: Saliency maps for the mosaic shown as the first image in Figure 7a. The saliency was calculated regarding the target class "Bernese Mountain Dog", to which the lower two images in the mosaic belong.



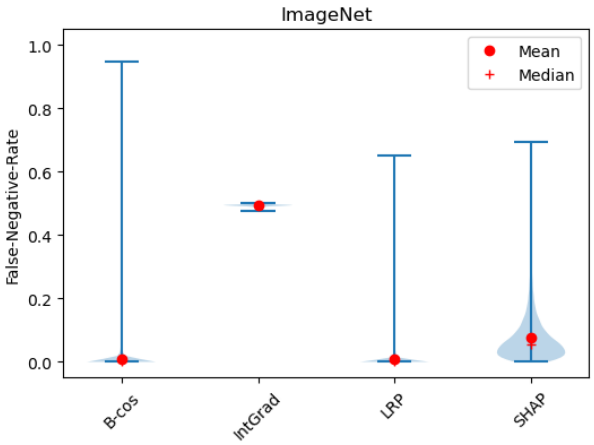
(a) Precision



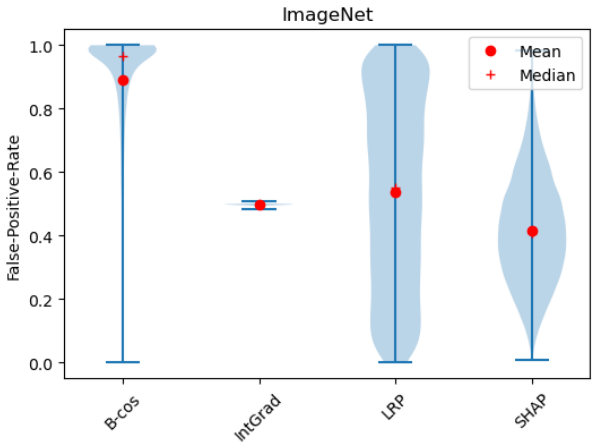
(b) Sensitivity



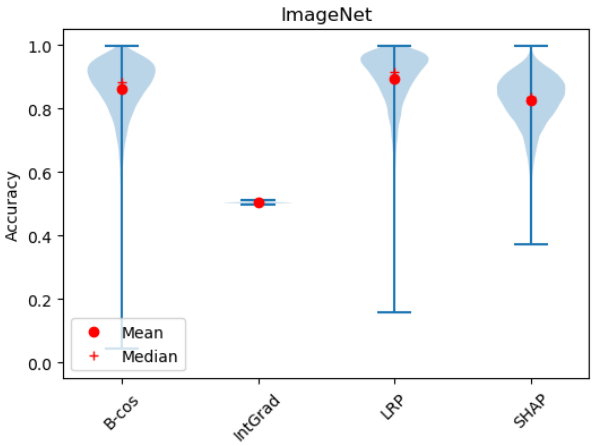
(c) Specificity



(d) False-Negative-Rate

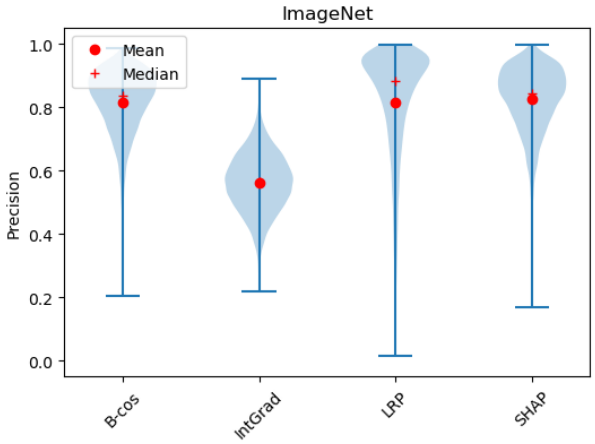


(e) False-Positive-Rate

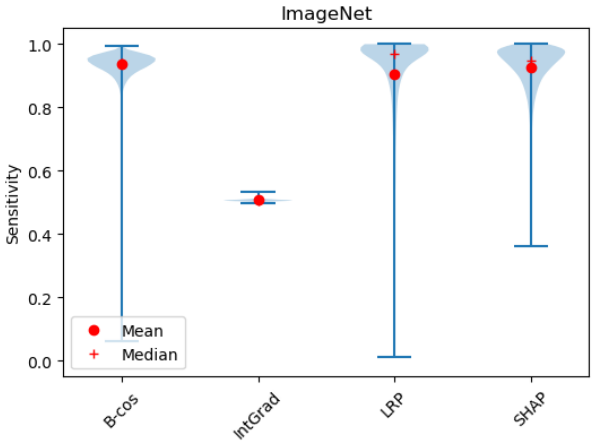


(f) Accuracy

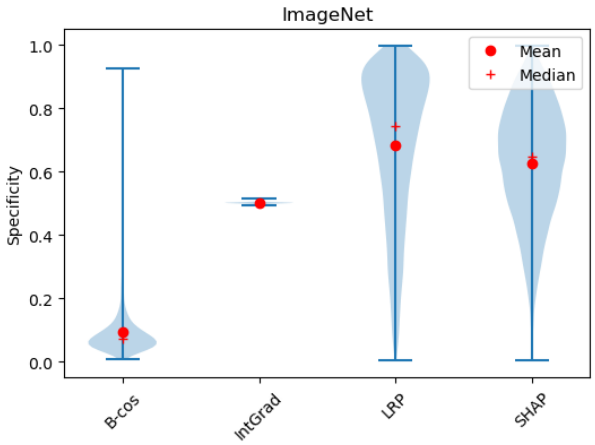
Figure 8: Results of the saliency metrics on the ImageNet mosaics for the ResNet50 model.



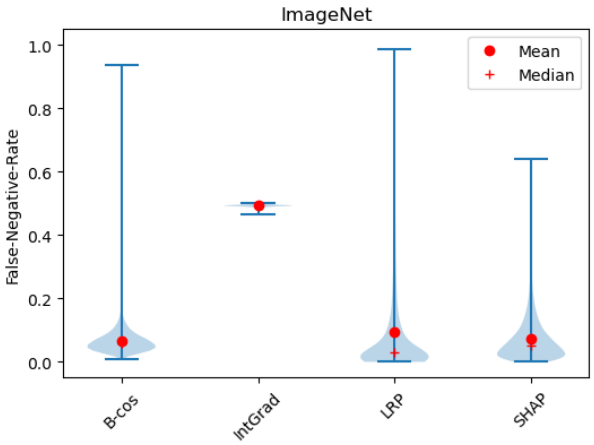
(a) Precision



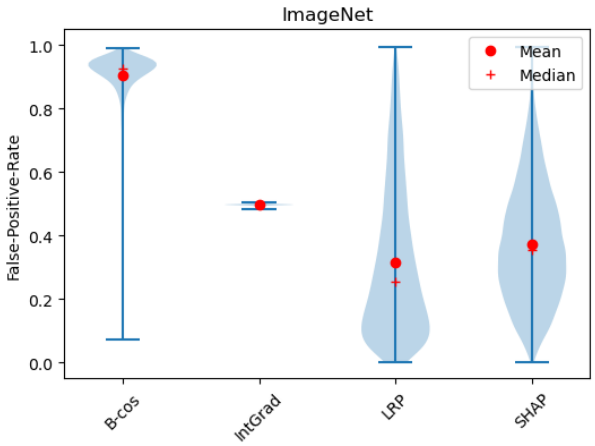
(b) Sensitivity



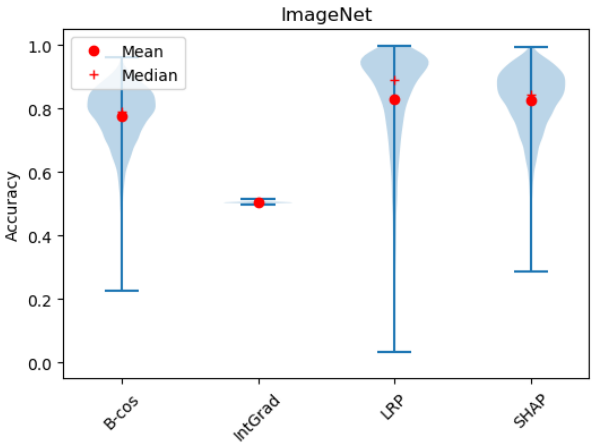
(c) Specificity



(d) False-Negative-Rate

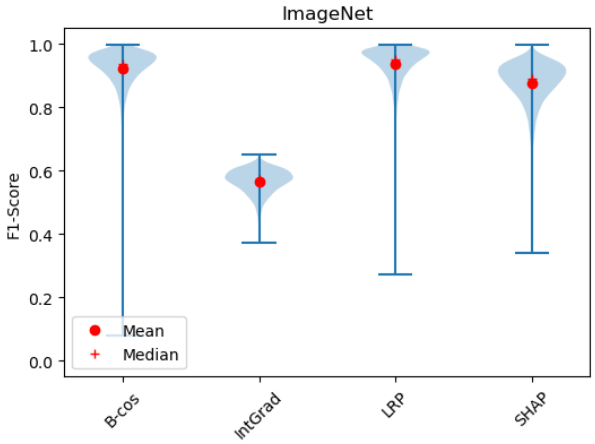


(e) False-Positive-Rate

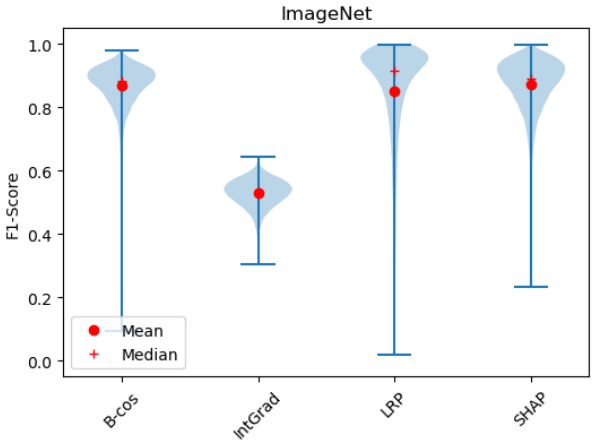


(f) Accuracy

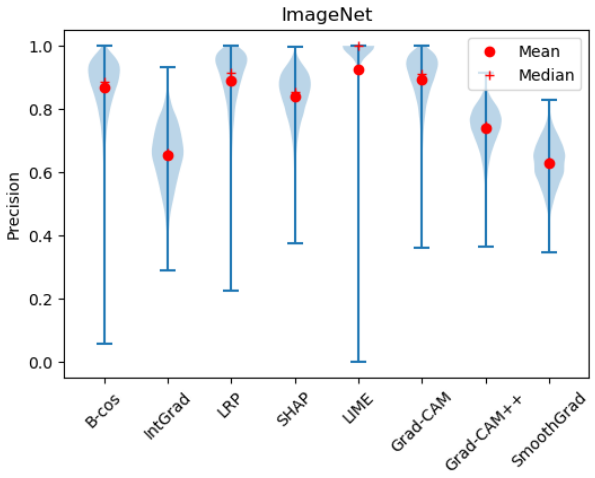
Figure 9: Results of the saliency metrics on the ImageNet mosaics for the VGG11 model.



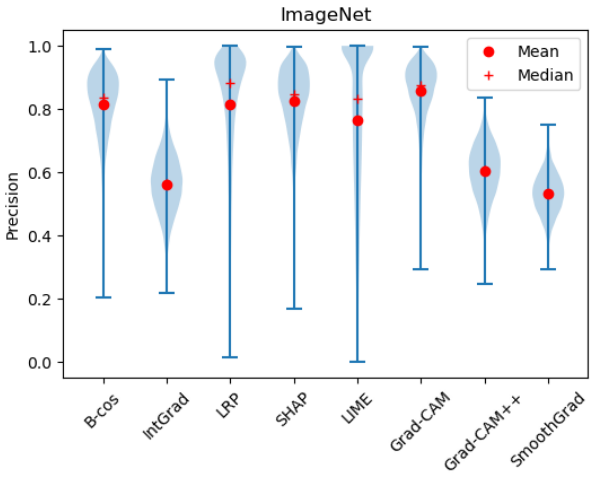
(a) F1-score for ResNet50.



(b) F1-Score for VGG11.



(c) Precision for all methods for ResNet50.



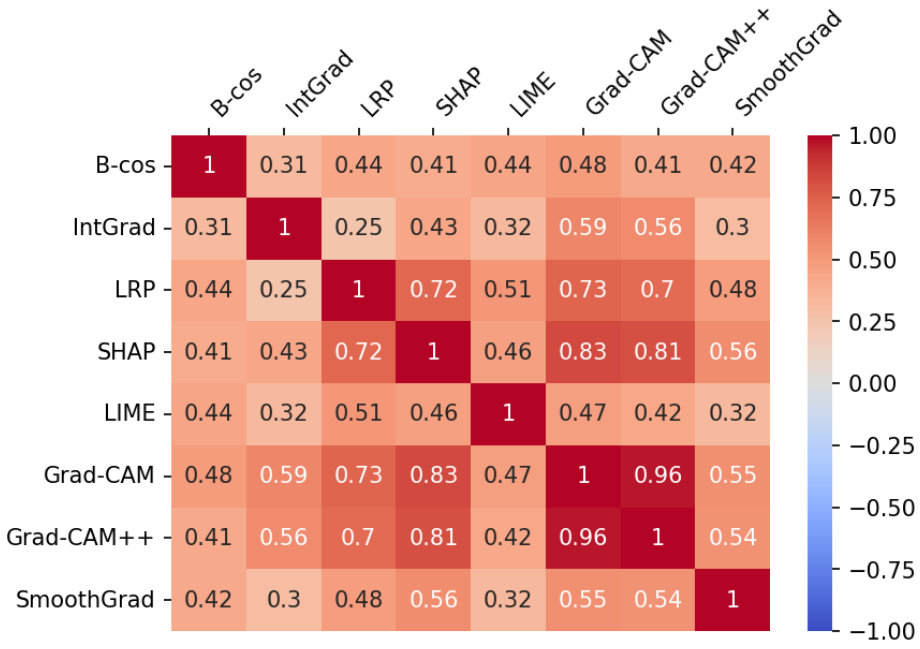
(d) Precision for all methods for VGG11.

Figure 10: F1-Score and precision for both models on the ImageNet mosaics.

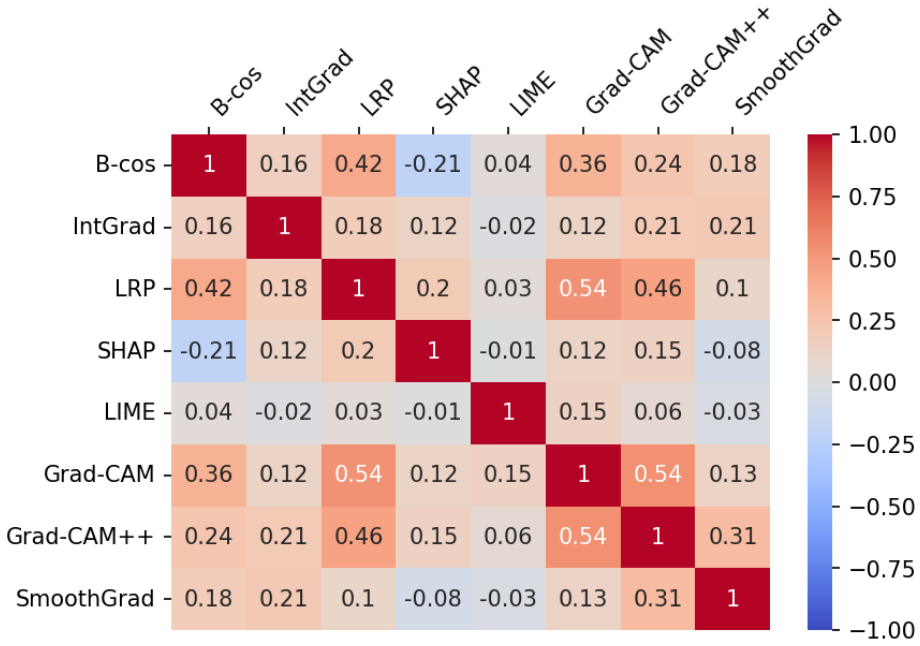
Table 7: Krippendorff's α for B-cos, IntGrad, LRP,SHAP, LIME, Grad-CAM, Grad-CAM++ and SmoothGrad for precision for ResNet50 and VGG11.

	ResNet50	VGG11
Cars/Cats dataset	0.88	0.56
Mountain Dogs dataset	0.25	0.14
ImageNet	0.71	0.56

Received 22 January 2024



(a) Spearman's ρ on the Mountain Dogs dataset for the precision metric for ResNet50.



(b) Spearman's ρ on the Cars/Cats dataset for the precision metric for ResNet50.

Figure 11: Spearman's ρ for the precision metric on different datasets. While the dataset with difficult to distinguish classes produces high correlation values (nearing 1 for Grad-CAM/Grad-CAM++), the dataset with easy to distinguish classes produces mostly random correlations, with the highest one between Grad-CAM/Grad-CAM++ and Grad-CAM/LRP with 0.54.