SALVAGE: SHAPLEY-DISTRIBUTION APPROXIMATION LEARNING VIA ATTRIBUTION GUIDED EXPLORATION FOR EXPLAINABLE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The integration of deep learning into critical vision application areas has given rise to a necessity for techniques that can explain the rationale behind predictions. In this paper, we address this need by introducing *Salvage*, a novel removal-based explainability method for image classification. Our approach involves training an explainer model that learns the prediction distribution of the classifier on masked images. We first introduce the concept of *Shapley-distributions*, which offers a more accurate approximation of classification probability distributions than existing methods. Furthermore, we address the issue of unbalanced important and unimportant features. In such settings, naive uniform sampling of feature subsets often results in a highly unbalanced ratio of samples with high and low prediction likelihoods, which can hinder effective learning. To mitigate this, we propose an informed sampling strategy that leverages approximated feature importance scores, thereby reducing imbalance and facilitating the estimation of underrepresented features. After incorporating these two principles into our method, we conducted an extensive analysis on the ImageNette, MURA, and Pet datasets. The results show that Salvage outperforms various baseline explainability methods, including attention-, gradient-, and removal-based approaches, both qualitatively and quantitatively. Furthermore, we demonstrate that our explainer model can serve as a fully explainable classifier without a major decrease in classification performance, paving the way for fully explainable image classification.

1 INTRODUCTION

In recent years, the expansion of artificial intelligence (AI) techniques, particularly in the field of computer vision, has revolutionized numerous industries and societal domains, ranging from healthcare to autonomous vehicles. Notably, the emergence of Vision Transformers (ViT), (Dosovitskiy et al., 2021) has lately significantly impacted the field of computer vision, establishing a new standard for image classification. Their ability to leverage self-attention mechanisms and effectively model long-range dependencies has positioned them at the forefront of research and applications across diverse domains. However, despite their remarkable performance, a critical challenge persists: the incompatibility of existing explainability techniques with vision transformers. The majority of explainability methods developed and refined for Convolutional Neural Networks (CNNs) often prove inadequate when applied to vision transformers. This difference highlights the need for architecture-independent explainability methods.

Among such methods, removal-based techniques work by iteratively masking portions of an input image to observe the resulting changes in the model's predictions. If the removal of a particular region significantly affects the model's prediction, it suggests that the region is important for the decision. Conversely, if removing a region has little to no effect, it is deemed less relevant to the model's decision. Recently, ViT Shapley (Covert et al., 2023) has been introduced, merging the principles of removal-based methods with a game-theoretic foundation. The method involves training an explainer model to estimate the Shapley values (Shapley, 1952) of the image patches, quantifying their contribution to the classifier's prediction. The Shapley value of a patch is estimated as the average change in the model's prediction when the patch is added to the image, calculated across all possible masked variations. Given that the number of possible mask combinations grows exponentially with the number of patches, ViT-Shapley leverages the FastShap (Jethani et al., 2022) algorithm, approximating the Shapley values via a least squares objective and stochastic gradient descent on randomly sampled masks, efficiently handling the computational complexity.

We propose a similar removal-based approach but with a key difference. Instead of approximating the Shapley value during training, we train an explainer model to learn a representation of the classifier's prediction distribution on masked images. At test time, the Shapley value for each patch is then derived from the learned representation. Moreover, we address two weaknesses of ViT-Shapley:

- The ViT-Shapley method often falls short by treating differences in prediction probabilities as linear scores and optimizing them using least squares, which does not adequately capture the probabilistic nature of the classifier's outputs. We address this limitation using a more conventional approach for probability distribution approximations, minimizing the divergence between the explainer's estimated distributions and the classifier's actual prediction distributions.
- The random mask sampling employed by ViT-Shapley is sample-inefficient, especially when dealing with heavy unbalanced ratios of important and unimportant patches, which can hinder effective learning. To mitigate this issue, we adopt an informative sampling strategy to enhance sample efficiency throughout training. By integrating the estimated attribution scores into the sampling process, we are able to achieve a more balanced distribution of masks with low and high prediction likelihoods, thereby facilitating the estimation of underrepresented features.

After incorporating the two proposed optimizations into a method, we refer to as Salvage (Shapley-distribution Approximation Learning Via Attribution Guided Exploration), we performed an evaluation on three datasets (ImageNette, Pet, MURA) and observed that Salvage outperforms various baseline explainability methods, including attention-, gradient-, and removal-based approaches, both qualitatively and quantitatively. Moreover, we introduce a novel concept, *classifying by explaining* which shifts the focus from explaining a classifier's behavior to aggregating the explainer's estimated feature importance scores into a classification prediction. By doing so we can guarantee the consistency between the predictions and explanations of the model. Our results demonstrate that our explainer can serve as a fully explainable classifier without a major decline in classification performance, advancing the development of more trustworthy image classifiers.

2 RELATED WORK

With the increasing demand for explainable AI, a variety of different attribution methods have been explored. These fall into five main categories.

Class Activation Maps: Convolutional Neural Networks (CNNs) have inspired the development of Class Activation Mapping (CAM) techniques to highlight important features in visual tasks. The original CAM (Zhou et al., 2015) method works for CNNs with Global Average Pooling (GAP) layers by generating attribution maps based on weighted feature maps. However, this method is limited to architectures with GAP layers. Grad-CAM (Selvaraju et al., 2019) improves upon this by using backpropagated gradients to compute feature map weights, making it more flexible. Variants like Grad-CAM++ (Chattopadhay et al., 2018), Eigen-CAM (Muhammad & Yeasin, 2020), and Ablation-CAM (Desai & Ramaswamy, 2020) explore different ways of refining these weights. Despite these advancements, CAM-based techniques were primarily designed for CNNs and often underperform when applied to transformer architectures (Covert et al., 2023).

Attention-based Methods: The attention mechanism of transformer models naturally allows insights into the information flow within the network. A straightforward method to assess importance is by analyzing the attention scores between the class token and input tokens at a specific layer (Clark et al., 2019). However, this approach gives limited insight into the overall information flow since different layers may focus on different regions, and the final output is shaped by the interaction across all layers. (Abnar & Zuidema, 2020) tackles this by modeling the information flow as a directed acyclic graph, using attention scores as edge weights. They propose two methods to extract input token relevance: attention rollout, which traces attention weights from the class token back to the input tokens, and attention flow, which estimates information flow using maximum flow computations in the graph. However, attention mechanisms often exhibit issues like high attention scores focusing on low-informative background regions (Covert et al., 2023; Darcet et al., 2023). (Darcet et al., 2023) suggest this problem stems from the use of random tokens as intermediaries for internal computations and address it by adding supplementary tokens. While attention scores can be useful in some cases, recent studies question their reliability as explanations, arguing they may not always reflect a model's true reliance on each token (Jain & Wallace, 2019; Serrano & Smith, 2019). Moreover, attention-based methods are class-agnostic, providing a single explanation per prediction and lacking class-specific insights.

Gradient-based Methods: Saliency maps (Simonyan et al., 2014) are an early method that extracts the gradient of the class score with respect to the input image. However, these gradients can be highly sensitive to small input perturbations, resulting in significant fluctuations (Smilkov et al., 2017). To mitigate this, SmoothGrad (Smilkov et al., 2017) averages the gradients over multiple noisy versions of the input image, effectively smoothing the gradients and reducing volatility. Integrated Gradients (Sundararajan et al., 2017) further improves on this by integrating gradients along the path between a baseline image (typically a black image) and the target image.

Removal-based Methods: Treating neural networks as a complete black box function, removal-based Methods measure fluctuations in the predicted class probabilities under partial information. The estimation of the prediction under partial information is achieved by inferring the classifier on masked images. RISE (Petsiuk et al., 2018) suggests measuring the contribution of each part of the image by sampling a large number of masks, computing the prediction of the network on each masked image, and finally summing up the averaged product of the masks with their corresponding predictions. However, as the number of possible masks grows exponentially in the number of image patches, a large amount of masks is required to obtain a decent estimation for each single region. To address this issue, FastSHAP (Jethani et al., 2022) proposes a game theory approach, training an explainer model to estimate the Shapley value of the image patches, which consists of the average change in the model's prediction when the patch is added to the image. Building upon this concept, ViT-Shapley (Covert et al., 2023) further extends this method by adopting a vision transformer-based architecture for the explainer model. While this method aims to train a model to approximate the Shapley value directly, our approach learns a representation of the classifier's prediction distribution from which the Shapley value can be extracted.

LRP-based Methods: Layer-Wise Relevance Propagation (LRP) (Lapuschkin et al., 2015), based on Deep Taylor Decomposition (DTD) (Montavon et al., 2017), explains model predictions by propagating the output back to the input using specific decomposition rules. While LRP has shown good results on CNNs, applying it to transformer architectures has led to unstable explanations. (Chefer et al., 2020) attributes these instabilities to skip connections and attention layers, and the authors propose alternative propagation rules for these operations, particularly combining LRP relevance with gradient-based attention rollout for attention layers. Similarly, (Ali et al., 2022) attempts to address this issue by proposing more stable rules for the self-attention and LayerNorm operations.

3 BACKGROUND

In this section, we introduce Shapley values (Shapley, 1952), which serve as the foundation of our method. Originating from cooperative game theory, Shapley values are used to fairly distribute payoffs among players based on their individual contributions to the total value. In the context of machine learning, they quantify the impact of each feature on a model's prediction by measuring the average change in prediction when the feature is included in an input subset. We begin by presenting the formal definition of Shapley values, followed by a rearrangement that enables their approximation without requiring their marginal contributions.

3.1 SHAPLEY VALUES

Let N be a set of features and $v(S)$ the prediction outcome given a feature subset $S \subset N$. The Shapley value ϕ_i of a feature i is obtained as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!\,(|N| - |S| - 1)!}{|N|!}}_{w_S} \left(v(S \cup \{i\}) - v(S)\right) \tag{1}$$

The computation of Shapley values for all features requires evaluating the predictions for every possible subset of features. However, as the number of features increases, the number of subsets grows exponentially, making the computation of the exact Shapley values infeasible for very large numbers of features. To address this challenge, FastShap (Jethani et al., 2022) proposes an approximation of the Shapley values using a least-squares objective over the feature subset distribution $p_w(S) \propto w_S$, sampled proportionally to w_S :

$$\mathbb{E}_{p_w(S)}[(v(S) - \sum_{i \in S} \phi_i)^2] \tag{2}$$

However, using the Mean Squared Error (MSE) loss to approximate probabilistic model outputs introduces significant limitations. MSE is primarily designed for comparing scalar values and is not well-suited for capturing the complexities of probability distributions, such as their inherent uncertainty, variance, and multimodal characteristics. Additionally, MSE is sensitive to scale and does not enforce the necessary constraints of probability measures, like non-negativity and normalization. As a result, MSE often yields invalid or suboptimal approximations when applied to distributions.

3.2 APPROXIMATING SHAPLEY VALUES WITHOUT MARGINAL CONTRIBUTIONS

Kolpaczki et al. (2024) suggests a rearrangement of the Shapley value formula. Instead of expressing it as the weighted average of marginal contributions, it can be viewed as the difference between the weighted average of prediction outcome when feature i is included and the weighted average of prediction outcome when feature i is excluded:

$$\phi_i = \underbrace{\sum_{S \subseteq N \setminus \{i\}} w_S \cdot v(S \cup \{i\})}_{\phi_i^+} - \underbrace{\sum_{S \subseteq N \setminus \{i\}} w_S \cdot v(S)}_{\phi_i^-} \tag{3}$$

The positive and the negative Shapley values can be seen as the expected values $\phi_i^+ = \mathbb{E}[v(S \cup i)]$ and $\phi_i^- = \mathbb{E}[v(S)]$, over the set distribution $p_w(S) \propto w$ for $S \subseteq N \setminus \{i\}$.

4 APPROACH

4.1 SHAPLEY DISTRIBUTION ESTIMATION

We build upon the concept of optimizing Shapley values without relying on marginal contributions by training an explainer model to learn both positive and negative Shapley values. During training, we sample masked images from the distribution $p_w(S) \propto w_S$. For each sampled masked image, we update the estimated positive Shapley values ϕ_i^+ for all visible image patches $i \in S$, and the negative Shapley values ϕ_j^- for all masked patches $j \notin S$. This is achieved by minimizing the difference between the sum $\sum_{i \in S} \phi_i^+ + \sum_{i \notin S} \phi_i^-$ and the actual prediction outcome $v(s)$. As mentioned in section 3.1, using the mean squared error (MSE) to approximate the target distribution $v(s)$ would yield suboptimal results because of the probabilistic nature of the classifier's output. Therefore, we propose mapping the summed term into a probability distribution $u(S)$, which we refer to as *Shapley probability distribution*:

$$u(S) = \sigma(\sum_{i \in S} \phi_i^+ + \sum_{i \notin S} \phi_i^-) \tag{4}$$

where σ denotes the softmax function in a multiclass classification setting or the sigmoid function for binary classification. The Shapley distribution of the masked image is then optimized by minimizing the Jensen–Shannon (JS) (Lin, 1991) divergence between the classifier's prediction $v(S)$ and its corresponding estimated probability distribution $u(S)$:

$$\arg\min_{\phi^+, \phi^-} \mathbb{E}_{p_w(S)}[D_{JS}(u(S)||v(S))] \tag{5}$$

At test time, the feature importance scores of each feature (image patch) are given by their estimated Shapley value $\phi_i = \phi_i^+ - \phi_i^-$.

4.2 ATTRIBUTION GUIDED SAMPLING

In our experiments, we observed that sampling from the random mask distribution p_w often results in a disproportionate number of masked images having either high or low likelihoods of the predicted class. This imbalance severely affects the minority class estimation, since a large number of samples are required to estimate the values of its members. This finding motivated us to address this imbalance through an alternative mask distribution, thereby enhancing the sampling efficiency during training. Thus, we propose exploiting the current estimates of feature importance scores ϕ to rebalance the ratio of masks, targeting the most and least informative regions of the image. Our proposed informative sampling distribution, $p_\phi(S) \propto \phi$, operates in two stages:

1. Sampling the number of masked patches: First, the number of masked patches within an image is sampled from a uniform distribution $U(1, n)$, where n represents the total number of patches in the image.
2. Selecting patches: Next, patches are selected without replacement, using the estimated feature importance scores ϕ of the target class, as sampling weights. For instance, if three patches have importance scores of 0, 1, and 3, their corresponding probabilities of being sampled would be 0, 0.25, and 0.75, respectively.

We then generate two equal mask subsets: the first, prioritizing the most informative regions from the image, and the second by masking them to target the least informative regions. A detailed description of the sampling process is provided as pseudo-code in Algorithm 3. Compared to random sampling, $p_\phi(S)$ yields a mask distribution with more balanced prediction likelihoods (see Figure 5), thereby enhancing sample efficiency during training.

4.3 FROM AN EXPLAINER TO A FULLY EXPLAINABLE CLASSIFIER

Since the classifier and the explainer are two decoupled models, the explainer merely approximates the behavior of the classifier. Thus, there is no guarantee of consistency between the classifier's predictions and the explainer's explanations, especially under domain shift. We suggest addressing this issue by using the explainer as a unified model for both classification and explanation.

Recall that Salvage is trained to minimize the divergence between its Shapley distribution $u(S)$ and the corresponding classifier prediction $v(S)$. By setting S to the full (unmasked) image N in eq. (4), we obtain the explainer's approximation for the classifier's prediction for the complete image:

$$u(N) = \sigma(\sum_{i \in N} \phi_i^+) \approx v(N) \tag{6}$$

In addition to ensuring consistency between the classification prediction and its explanation, using the explainer as a classifier offers a unique advantage. By aggregating the importance scores of each image region, we obtain a precise understanding of how each region contributes to the overall prediction. This approach results in a classifier that is fully transparent and explainable.

5 EXPERIMENTS

In this section, we first describe our experimental setup. We then conduct both a qualitative and a quantitative analysis of our method and several baselines. Next, we conduct an ablation study, showing the advantage of informative sampling. Finally, we evaluate the classification and explanation of Salvage as an explainable classification method.

5.1 EXPERIMENTAL SETUP

We adopt the experimental setting from ViT-Shapley (Covert et al., 2023), evaluating the explanation performance of our method on vision transformer classifiers across the ImageNette and Pet datasets for multi-class classification and the MURA dataset for binary classification. To do so we train a ViT classifier for 25 epochs and fine-tune it on masked images for 50 epochs. We initialize the weights of the classifier with the pre-trained parameters from Dinov2 (Oquab et al., 2024) with registers (Oquab et al., 2024). We adopted the Segmenter (Strudel et al., 2021) segmentation architecture for our explainer models and trained them using 32 masks per image for $\sim 18k$ iterations (corresponding to 50 epochs for ImageNette and MURA, and 100 epochs for Pet), all models were trained with a batch size of 64, an AdamW (Loshchilov & Hutter, 2019) optimizer with a learning rate of 1e-5 and a weight decay of 1e-5, except the explainer models of ViT-Shapley which were trained with lr 1e-4. We compare our method to 10 different baselines; GradCam (Selvaraju et al., 2019), EigenCam (Muhammad & Yeasin, 2020), Attention scores from the last layer with registers (Clark et al., 2019; Darcet et al., 2023) (Attn. last), Attention Rollout with registers (Abnar & Zuidema, 2020; Darcet et al., 2023), ViT-CX (Xie et al., 2023), Saliency maps (Simonyan et al., 2014), Integrated gradients (Sundararajan et al., 2017), LRP beyond attention (Chefer et al., 2020), RISE (Petsiuk et al., 2018), and ViT Shapley (Covert et al., 2023). As a reference, we additionally evaluate the metric scores on randomly generated maps (Random). The implementation of the CAM-based methods is adapted from (Gildenblat & contributors, 2021), while the other baselines use their original implementations.

5.2 QUALITATIVE ANALYSIS

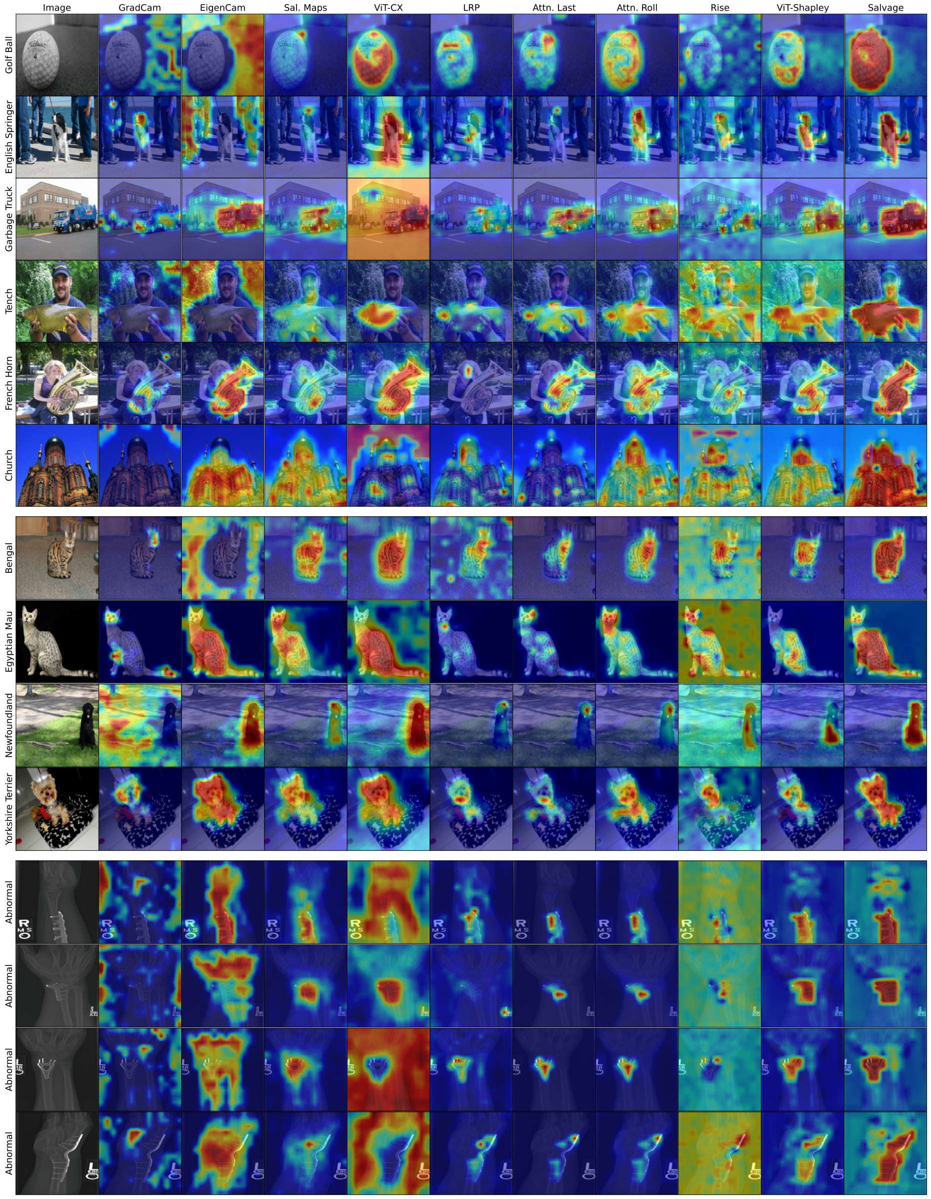


Figure 1: Qualitative examples computed on ImageNette, Pet and MURA.

We conduct a qualitative analysis comparing our method with the 6 (quantitatively) best baselines. We present 6 examples from ImageNette, 4 from Pet, and 4 from MURA in Figure 1. Saliency maps, Vit-CX, and LRP seem unreliable as their attribution maps often focus on random parts of the background. Despite its solid quantitative results, the attribution maps generated by RISE are highly noisy. Attention rollout (with registers) and ViT-Shapley yield decent results on most images with relatively low scores outside the informative regions. Solely Salvage excels at highlighting the entire relevant region of interest. Further qualitative examples can be found in the appendix in Figure 8.

5.3 QUANTITATIVE ANALYSIS

Table 1: Quantitative results computed on the Pet, ImageNette, and MURA datasets. The performance of the 10 compared methods is measured in terms of SRG, R-SRG, RMA, and RRA.

	Pet				ImageNette		MURA	
Method	SRG	R-SRG	RMA	RRA	SRG	R-SRG	SRG	R-SRG
GradCam	10.61	3.51	48.15	42.66	-1.95	-3.31	16.16	10.09
EigenCam	27.42	3.16	48.85	62.93	13.25	-3.07	0.11	-4.54
Attn. last	47.89	9.62	61.15	70.15	27.02	3.03	22.41	7.00
Attn. Roll.	51.97	11.22	51.52	74.65	32.02	3.45	17.58	6.26
ViT-CX	50.23	17.64	30.64	67.46	29.92	7.49	19.84	9.09
Sal. Maps	51.12	10.78	52.75	**76.27**	27.75	2.84	25.26	8.50
IntGrad	27.44	7.88	51.52	58.77	10.96	2.17	13.87	6.09
LRP	49.55	9.24	63.93	71.77	27.88	3.04	19.29	6.84
RISE	63.70	18.50	30.10	47.84	22.92	5.37	56.48	22.07
ViT-Shap	61.07	14.69	52.74	69.05	40.33	6.22	65.34	20.58
Salvage	**68.46**	**26.29**	**64.87**	73.52	**51.35**	**14.87**	**68.56**	**25.32**
Random	0.00	0.00	30.02	29.38	0.00	0.00	0.00	0.00

Based on our baselines, we employed various metrics to assess Salvage's performance. Since the true importance of features is unknown beforehand, evaluating explanation accuracy poses a challenge. The metrics used to evaluate our method include Most and Least Influential First (MIF, LIF) (Petsiuk et al., 2018), Symmetric Relevance Gain (SRG) (Blücher et al., 2024), Relevance Rank Accuracy (RRA) and Relevance Mass Accuracy (RMA) (Arras et al., 2022). MIF and LIF (also known as Insertion and Deletion) measure performance by progressively adding or removing image patches based on their importance scores, with the goal of maximizing MIF (adding most important features first) and minimizing LIF (adding least important features first). SRG improves upon these metrics by addressing their sensitivity to masking strategies and calculating the difference between the MIF and LIF scores to provide more consistent performance rankings. Additionally, RRA and RMA assess the alignment of feature importance scores with human-annotated regions of interest. RRA evaluates how well the top-k important patches match the annotated region, while RMA measures the proportion of attributions within the annotated area, reflecting the focus on relevant regions.

It is important to note that MIF, LIF, and SRG offer only limited insights into the quality of explanations, as they focus solely on the ranking of features while disregarding their relative differences. To address this limitation, we extend our metric selection by further including three new metrics to assess the relative score differences within an attribution map. Specifically, we extend the MIF and LIF metrics into R-MIF (Relatively Most Influential First) and R-LIF (Relatively Less Influential First). Instead of selecting patches purely based on their rank, we use estimated importance scores as weights in a sampling process to determine which patches to add first. To reduce sampling variance, we generate 128 masks for each size of feature subset. Analogously to SRG, we define R-SRG as the difference between R-MIF and R-LIF. These proposed metrics provide deeper insight into the relative importance of feature attribution scores and their influence on the model predictions. A pseudo-code and a more detailed description of the metrics is provided in appendix A.1.

Our results, as shown in Table 1 and Figure 2, clearly demonstrate that our method outperforms all baselines across both rank-based and relative-based metrics on all three datasets. We have observed in Figure 2 that all top-3 methods (Rise, ViT-Shapley, and Salvage) reach similar MIF and R-MIF scores – hinting, that all three methods highlight a small subset of important features which is sufficient for the model to recognize the classified object. However, we observed that Salvage reaches significantly better LIF and R-LIF scores, which suggests that our method is more effective at identifying a larger portion of the important features compared to the other methods. Disregarding the top-3 methods, our results are in line with Covert et al. (2023) showing poor performance of the CAM-based, attention-based, gradient-based, and LRP-based methods. The LIF, MIF, R-LIF, and R-MIF scores of all methods have been included in appendix A.3 for the sake of completeness.

Moreover, we evaluate the RMA and RRA metrics on the Pet dataset, for which human-annotated regions of interest were provided. The results (cf. Table 1) showed that Salvage achieves the best scores in terms of RMA. We additionally observe that Sal. Maps and Attn. Roll. reach slightly higher RRA scores showing a strong alignment in ranking, meaning their top attribution scores lie within the object of interest. However, these methods also show a lower attribution mass within the object (RMA) and exhibit weaker SRG and R-SRG scores, suggesting they may not focus on the most relevant parts of the object for the prediction, potentially limiting their effectiveness.

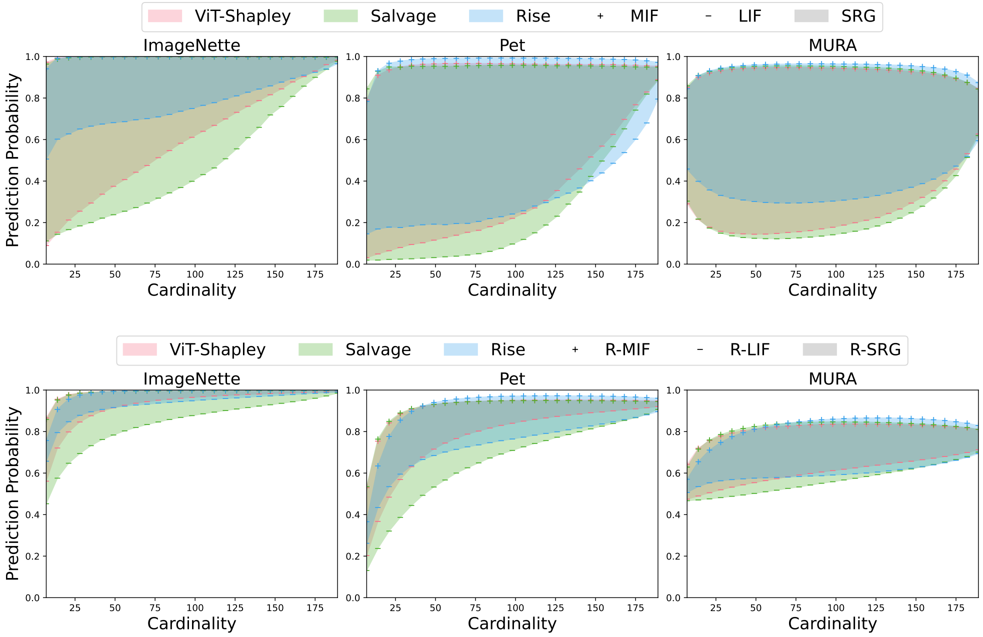
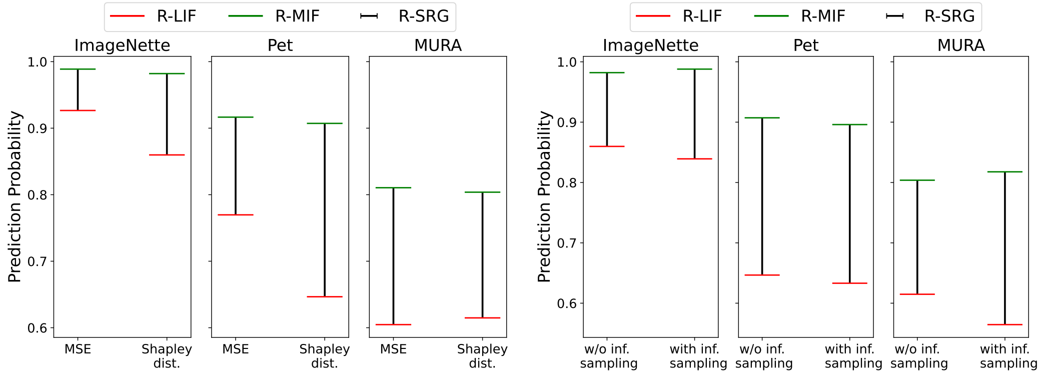


Figure 2: An Illustration of the MIF, LIF, SRG, R-MIF, R-LIF, and R-SRG across the different cardinality values (number of masked patches) for the top-3 performing methods (RISE, ViT-Shapley, and Salvage) computed on the Pet, ImageNette, and MURA datasets.

5.4 ABLATION STUDIES

In this section, we investigate the individual contribution of each core principle in our method, Shapley distribution estimation and informative sampling, through two ablation studies. The first study compares the performance of Shapley distribution estimation to the MSE-based approximation (ViT-Shapley) without informative sampling. The second study investigates the effect of informative sampling on the performance of Salvage by comparing its performance with and without the use of informative sampling.

The results of our studies, illustrated in Figure 3, reveal the following: (a) Shapley distribution estimation outperforms MSE-based approximation (ViT-Shapley) on ImageNette and Pet datasets, but shows a slightly lower performance on MURA. However, we believe the latter could benefit from adjusting the temperature parameter in the sigmoid function used during training. (b) Informative sampling clearly improves performance on MURA and ImageNette, with a modest gain of 0.26 in R-SRG observed for Pet.



(a) (b)

Figure 3: (a) Ablation study quantifying the performance gain of using Shapley distribution estimation versus MSE-based approximation (ViT-Shapley). (b) Ablation study comparing the performance of Salvage with and without informative sampling. The performance is reported in terms of R-SRG, which is given by the difference between R-MIF and R-LIF (the larger the better).

5.5 FROM EXPLAINER TO EXPLAINABLE CLASSIFIER

In this subsection, we assess the benefits of using Salvage as an explainable classifier. By directly deriving predictions from the feature importance scores, the exact contribution of each image region to the classification outcome is made explicit. This makes Salvage particularly well-suited for applications where explainability is as critical as classification accuracy.

We start by comparing the classification performance of Salvage to a baseline classifier and the ViT-Shapley explainer model. As shown in Table 2, Salvage demonstrates no major drop in performance relative to the original classifier. In contrast, the ViT-Shapley explainer model performs poorly in classification. This issue can be attributed to the explainer's attribution scores for different classes being decoupled during training due to the use of additive normalization (Covert et al., 2023).

Table 2: An overview of the classification performance of the original classifier, ViT-Shapley, and Salvage computed on Pet, ImageNette, and MURA.

Model	Pet	ImageNette	MURA			
	Accuracy	Accuracy	Precision	Recall	F1-score	MCC
Classifier	95.91%	99.64%	84.64%	78.88%	81.66%	0.66
ViT-Shapley	0.00%	0.05%	59.03%	92.74%	72.14%	0.39
Salvage	93.61%	98.88%	80.31%	80.52%	80.41%	0.62

Additionally, we analyze the predictions consistency between the classifier and explainer models in Appendix A.4, demonstrating an average agreement of 95.90% for correctly classified samples and 80.10% for misclassified ones.

Next, we analyze the explanations of Salvage in cases where its classification prediction is inaccurate. In Figure 4, we present examples where Salvage failed to produce the correct classification prediction and provide its attribution maps for both the predicted class (second row) and ground truth classes (third row). In the second row of the figure, we can see that Salvage provides clear, understandable explanations for its misclassifications. For example, it can explain errors such as mistaking a monument for a church, a construction truck for a garbage truck, or a mouse for an English Springer Spaniel. Notably, in the last row, even in these failure cases, the attribution map corresponding to the ground truth class of the image still accurately highlights the ground truth object (parachute, french horn,...). The ability to provide reliable explanations, even in cases of classification failure, underscores Salvage's strong potential to serve as a fully explainable classifier.

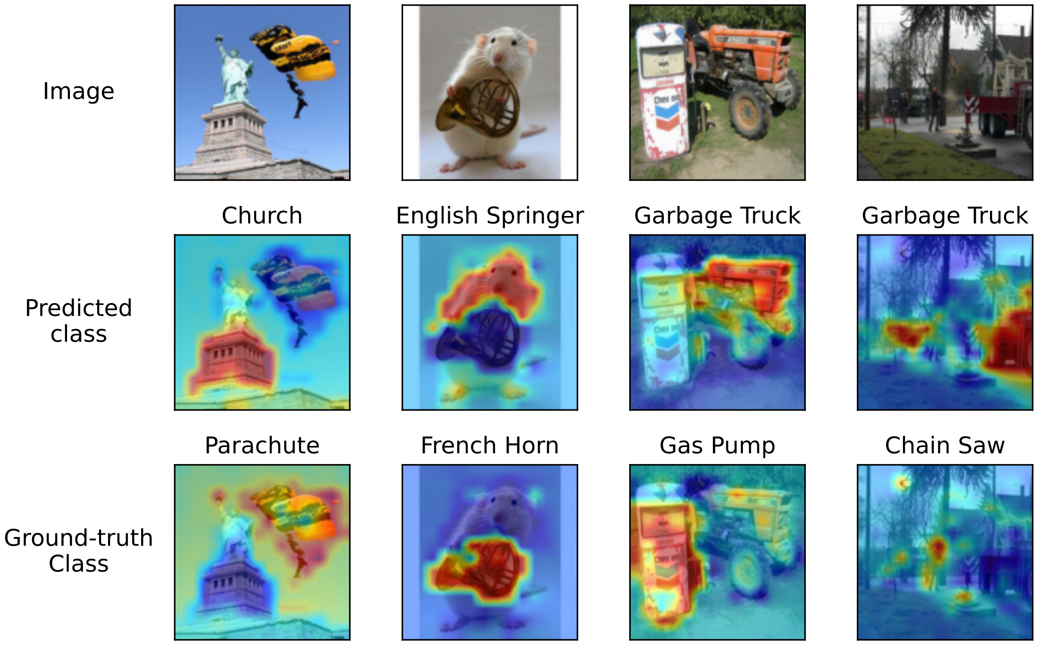


Figure 4: Examples of test images from ImageNette where Salvage fails to make a correct classification prediction. In the second row, we present the Salvage's attribution map for the (wrongly) predicted class, in the last row its attribution map for the ground truth class of the image.

6 CONCLUSION

By incorporating a novel methodology for attribution score estimation and informative sampling, we have developed a removal-based explanation method for image classification called Salvage. Our experiments demonstrate that Salvage outperforms all 10 evaluated baselines, both qualitatively and quantitatively, by delivering higher-quality explanations and clearly distinguishing relevant from irrelevant image regions. Beyond its strong explanation performance, we have also established Salvage's potential as a fully explainable classifier. While its classification accuracy is comparable to that of a classifier model, Salvage consistently provides detailed and interpretable explanations, even for images that are misclassified. This capability not only highlights the regions contributing to the predictions but also helps users understand the underlying factors leading to errors. Overall, these features underscore Salvage's strong potential to serve as a fully explainable classifier in applications where explainability is as critical as classification accuracy.

7 LIMITATIONS AND FUTURE WORK

Our work offers several promising avenues for further advancement. Future optimizations of our method could involve refining the neural architecture of the explainer model, as improved segmentation architectures may enhance its ability to capture spatial relationships and accurately estimate attribution scores for each superpixel-class pair. Moreover, introducing a temperature parameter in the softmax or sigmoid functions during the approximation of the classifier's distribution may be valuable and could offer a better alignment of the magnitude of the approximated values with the output logits of the classifier. Additionally, since Salvage adopts the principles of Shapley's additive explanation, it relies on the assumption that all features are linearly independent—an assumption that may be overly restrictive in practice. A promising direction for future work could involve extending Salvage to account for feature interactions. Motivated by the quality of the explanation maps produced by Salvage, we plan to explore its potential on different tasks and data modalities, as well as an unsupervised segmentation model in future research.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation, 2022.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.11.008. URL `https://www.sciencedirect.com/science/article/pii/S1566253521002335`.

Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Decoupling pixel flipping and occlusion strategy for consistent xai benchmarks, 2024. URL `https://arxiv.org/abs/2401.06654`.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018. doi: 10.1109/wacv.2018.00097. URL `http://dx.doi.org/10.1109/WACV.2018.00097`.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2020. URL `https://api.semanticscholar.org/CorpusID:229297908`.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention, 2019.

Ian Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

Saurabh Satish Desai and H. G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 972–980, 2020. URL `https://api.semanticscholar.org/CorpusID:214604773`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation, 2022.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions, 2024. URL `https://arxiv.org/abs/2302.00736`.

Sebastian Lapuschkin, 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:e0130140, 07 2015. doi: 10.1371/journal.pone.0130140.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017. ISSN 0031-3203. doi: 10.1016/j.patcog.2016.11.008. URL `http://dx.doi.org/10.1016/j.patcog.2016.11.008`.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2020. doi: 10.1109/ijcnn48605.2020.9206626. URL `http://dx.doi.org/10.1109/IJCNN48605.2020.9206626`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL `https://arxiv.org/abs/2304.07193`.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

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. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL `http://dx.doi.org/10.1007/s11263-019-01228-7`.

Sofia Serrano and Noah A. Smith. Is attention interpretable?, 2019.

Lloyd S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. doi: 10.7249/P0295.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L. Zhang. Vit-cx: Causal explanation of vision transformers, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.

A APPENDIX

A.1 METRICS OVERVIEW

Most influential First (MIF) and Least influential First (LIF): also known as Insertion and Deletion (Petsiuk et al., 2018) repeatedly mask the images by inserting image patches based on their ascending/descending ranking from the most important to the least important attribution scores of the method being evaluated. The area under the resulting curve (predictions/number of patches) is then computed as the performance score as illustrated in Algorithm 1. The MIF score should be maximized, which is achieved by adding the most important features first, in order to get high predictions with as few patches as possible. Conversely, the LIF score should be minimized by deleting all important features before the irrelevant ones (by assigning them higher values than the irrelevant ones). It is important to note that MIF and LIF offer only limited insights into the quality of explanations, as they focus solely on the ranking of features while disregarding their relative differences.

Algorithm 1 Most Informative First (MIF) and Least Informative First (LIF)

```
Require: $N$: Set of all features, $v$: Prediction outcome function, $\phi$: Importance scores for each feature, $metric$: 'MIF' or 'LIF'
Ensure: (averaged) MIF or LIF score
 1: $S \leftarrow \emptyset$                                   ▷ Initialize empty subset
 2: $preds \leftarrow []$                                     ▷ List to store prediction outcomes
 3: if $metric$ = 'MIF' then
 4:     $sorted\_features \leftarrow$ Sort features in descending order of importance based on $\phi$
 5: else
 6:     $sorted\_features \leftarrow$ Sort features in ascending order of importance based on $\phi$
 7: end if
 8: for $i = 1$ to $|N|$ do
 9:     Add $sorted\_features[i]$ to $S$
10:     $pred \leftarrow v(S)$                                 ▷ Evaluate the prediction outcome with the current subset
11:     Append $pred$ to $preds$
12: end for
13: $final\_score \leftarrow \frac{\sum \text{preds}}{|N|}$     ▷ Compute the average over all subset sizes
14: return $final\_score$
```

Symmetric Relevance Gain (SRG): In a recent study, Blücher et al. (2024) demonstrated the inconsistency of the MIF and LIF metrics while using different masking strategies, as these can lead to different performance rankings depending on the robustness of the masking strategy. In order to address this issue, the authors presented a simple, yet effective metric named SRG, which is given by the difference between the MIF and LIF scores:

$$SRG = MIF - LIF \tag{7}$$

The authors have shown on 40 different masking strategies that this metric breaks the inherent connection to the underlying occlusion strategy and leads to consistent rankings.

Relatively Most influential First (R-SRG), Relatively least influential First (R-LIF) MIF, LIF, and SRG provide limited insights into the quality of explanations, as they focus exclusively on feature ranking without accounting for the relative difference scores between the features. To address this limitation, we propose an extension of these metrics aimed at capturing the faithfulness of the relative differences in importance scores across different features. While MIF and LIF evaluate the prediction model by selecting the features purely based on their ranking, we propose selecting n features through a weighted sampling process. This process uses feature importance scores as sampling weights, ensuring that features are sampled in proportion to their estimated importance. By employing this method, we can assess the faithfulness of the relative differences in attribution scores, which are integrated into the sampling process. Analogous to MIF and LIF, R-MIF aims to sample the most relevant features, while R-LIF aims to sample the least relevant ones. In our experiments, we repeated the sampling process for each mask size 128 times to minimize variance in performance scores resulting from the sampling process ($n_masks = 128$). For a more detailed overview, we present the pseudo-code for the computation of the R-MIF and R-LIF scores in Algorithm 2

Algorithm 2 R-MIF and R-LIF

```
Require: $N$: Set of all features, $v$: Prediction outcome function, $\phi$: Importance scores for each feature, $metric$: 'R-MIF' or 'R-LIF', $n\_masks$: Number of subsets to sample per cardinality
Ensure: (averaged) R-MIF or R-LIF score
$S \leftarrow \emptyset$    ▷ Initialize empty subset
$preds \leftarrow []$    ▷ List to store averaged prediction outcomes
$\phi\_min \leftarrow \min(\phi)$    ▷ Find minimum feature importance score
$\phi\_max \leftarrow \max(\phi)$    ▷ Find maximum feature importance score
$\phi\_norm \leftarrow \frac{\phi - \phi\_min}{\phi\_max - \phi\_min}$    ▷ Min-max normalization ensuring positive sampling weights
if $metric$ = 'R-MIF' then
    $weights \leftarrow \phi\_norm$    ▷ Use the normalized scores as weights for R-MIF
else
    $weights \leftarrow 1 - \phi\_norm$    ▷ use 1 minus the normalized scores as weights for R-LIF
end if
for $i = 1$ to $|N|$ do
    $subset\_preds \leftarrow []$    ▷ Store prediction outcomes for this step
    for $j = 1$ to $n\_masks$ do
        $S_j \leftarrow$ draw $i$ features from $N$ using $weights$ as sampling weights, without replacement
        $pred \leftarrow v(S_j)$    ▷ Evaluate the prediction outcome with the sampled subset
        Append $pred$ to $subset\_preds$
    end for
    $avg\_pred \leftarrow \frac{\sum \text{subset\_preds}}{n\_masks}$    ▷ Average the prediction outcomes over all samples
    Append $avg\_pred$ to $preds$
end for
$final\_score \leftarrow \frac{\sum \text{preds}}{|N|}$    ▷ Compute the final average score over all subset sizes
return $final\_score$
```

Relative Symmetric Relevance Gain (R-SRG) Analogously to SRG, we combine the R-MIF and R-LIF scores by defining:

$$R\text{-}SRG = R\text{-}MIF - R\text{-}LIF \tag{8}$$

Relevance Rank Accuracy (RRA) / Relevance Mass Accuracy (RMA) Arras et al. (2022) presented two metrics measuring the consistency of the feature importance scores with a target region of interests provided by human annotations. For relevance rank accuracy, image patches are ordered based on their importance scores, and the number of top-k pixels within the ground truth mask is measured, with k set to be the number of pixels in the ground truth mask. A high relevance rank score is indicative of a strong alignment between the explanation and the human annotation. For relevance mass accuracy, the ratio of positive attributions within the ground truth mask to the sum of all positive attributions is calculated. A high relevance mass score indicates that significant attention is placed on the same region as the human annotation, with little attention directed to other regions.

A.2 SAMPLING DISTRIBUTION

For the sake of completeness, we present a pseudo code of the informative sampling procedure in Algorithm 3. Moreover, we computed the average prediction of the sampled subsets once using random sampling and once using attribution-informed sampling. The results illustrated in appendix A.2 demonstrate that the average prediction likelihood of our informative sampling technique yields a more balanced distribution than random sampling.

Additionally, we evaluated the effect of addressing this unbalance on sample efficiency. To do so, we illustrate in Figure 6 the effect of informative sampling on sample efficiency by comparing the SRG metric (higher is better) during training for models trained with and without the proposed informative sampling method. The results demonstrate that informative sampling consistently improves the SRG metric across the training process, indicating enhanced sample efficiency.

Algorithm 3 Feature Subset Sampling Using Importance Scores

```
Require: $N$: Set of all features, $\phi$: Importance scores for each feature, $n\_masks$: Total number of subsets to sample
Ensure: List of sampled subsets
$\phi\_min \leftarrow \min(\phi)$    ▷ Find minimum feature importance score
$\phi\_max \leftarrow \max(\phi)$    ▷ Find maximum feature importance score
$\phi\_norm \leftarrow \frac{\phi - \phi\_min}{\phi\_max - \phi\_min}$    ▷ Min-max normalization ensuring positive sampling weights
$samples \leftarrow []$    ▷ List to store sampled subsets
for $k = 1$ to $\frac{n\_masks}{2}$ do
    $m \leftarrow$ Sample from $U(1, |N|)$    ▷ Sample subset size from uniform distribution
    $S_\phi \leftarrow$ draw $m$ features from $N$ using $\phi\_norm$ as sampling weights, without replacement
    Append $S_\phi$ to $samples$
end for
for $k = 1$ to $\frac{n\_masks}{2}$ do
    $m \leftarrow$ Sample from $U(1, |N|)$    ▷ Sample subset size from uniform distribution
    $S_{1-\phi} \leftarrow$ draw $m$ features from $N$ using $(1 - \phi\_norm)$ as sampling weights, without repl.
    Append $S_{1-\phi}$ to $samples$
end for
return $samples$
```

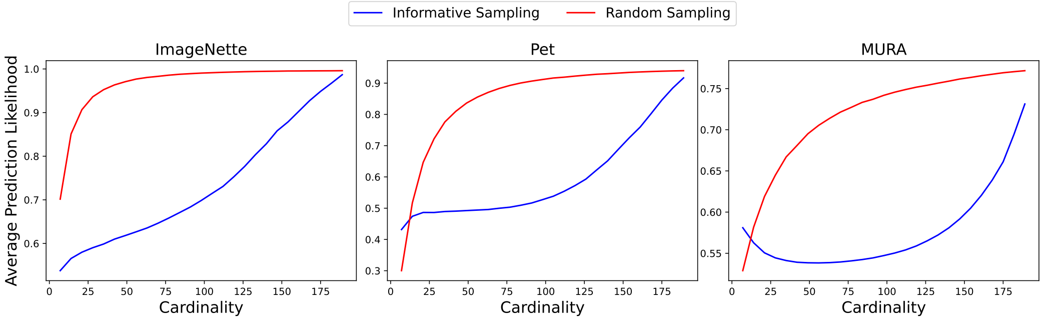


Figure 5: Average prediction likelihood of the ground truth class using informative sampling versus random sampling, computed across various mask sizes.

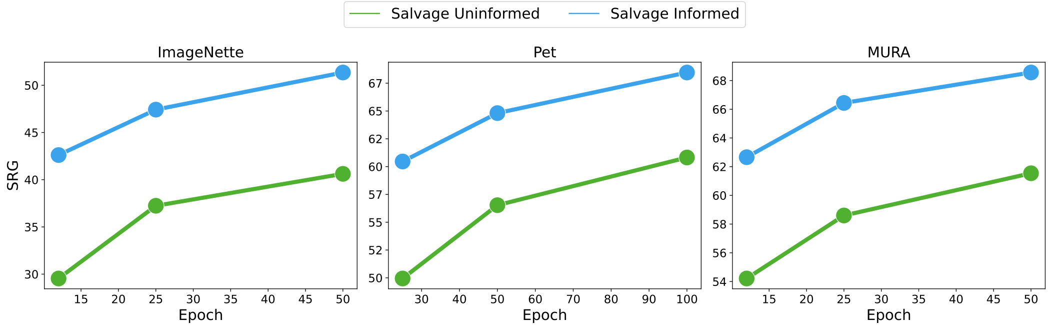


Figure 6: Comparison of the SRG metric during training of the Salvage explainer model. The SRG metric (higher is better) is evaluated for models trained with informative sampling (Salvage Informed) and without it (Salvage Uninformed) across 25%, 50%, and 100% of the overall training duration.

A.3 FURTHER QUANTITATIVE ANALYSIS

In this section, we present the averaged LIF and MIF scores for all baselines in Table 3, along with the R-LIF and R-MIF scores in Table 4. The results indicate that while the top three methods achieve relatively high MIF and R-MIF scores, our method demonstrates superior performance in the LIF and R-LIF metrics. This suggests that our model is better at identifying a larger portion of the important features, resulting in a clearer distinction between the relevant and irrelevant regions of the image. This observation is in line with the findings of our qualitative results.

Table 3: Quantitative results computed on the dataset Pet, ImageNette, and MURA. The performance of the 12 compared methods is measured in terms of MIF and LIF.

	Pet		ImageNette		MURA	
Method	MIF ↑	LIF ↓	MIF ↑	LIF ↓	MIF ↑	LIF ↓
GradCam	81.08	70.47	88.94	90.89	78.23	62.07
EigenCam	76.47	49.05	87.70	74.45	65.59	65.48
Attn. last	89.44	41.55	96.76	68.74	74.69	52.28
Attn. Roll.	89.62	37.65	96.95	64.93	73.60	56.02
ViT-CX	88.19	37.96	96.29	66.37	74.44	54.60
Sal. Maps	89.38	38.26	96.09	68.34	74.57	49.31
IntGrad	89.45	62.01	96.76	85.80	75.32	61.45
LRP	89.64	40.09	97.05	69.17	75.14	55.85
RISE	95.73	32.03	98.30	75.38	93.20	36.72
ViT-Shap	93.52	32.45	98.71	58.38	91.25	25.91
Salvage	93.21	24.75	98.44	47.09	91.94	23.38
Random	83.29	83.67	95.13	95.07	71.24	70.78

Table 4: Quantitative results computed on the dataset Pet, ImageNette, and MURA. The performance of the 12 compared methods is measured in terms of R-MIF and R-LIF.

	Pet		ImageNette		MURA	
Method	R-MIF ↑	R-LIF ↓	R-MIF ↑	R-LIF ↓	R-MIF ↑	R-LIF ↓
GradCam	85.19	81.68	92.59	95.90	78.27	68.18
EigenCam	79.15	75.99	90.03	93.10	66.30	70.84
Attn. last	91.40	81.78	98.47	95.44	75.36	68.36
Attn. Roll.	90.83	79.61	98.29	94.83	74.79	68.53
ViT-CX	89.08	71.45	97.98	90.49	74.37	65.28
Sal. Maps	90.84	80.06	98.09	95.25	75.19	66.69
IntGrad	90.62	82.74	98.20	96.03	75.38	69.29
LRP	91.32	82.09	98.52	95.47	75.66	68.82
RISE	91.62	73.12	98.26	92.89	81.91	59.84
ViT-Shap	91.66	76.97	98.89	92.67	81.05	60.47
Salvage	91.60	65.31	98.79	83.92	81.77	56.45
Random	85.20	84.91	96.66	96.68	71.53	71.79

A.4 CLASSIFICATION ANALYSIS

In Figure 7, we present the confusion matrices for the explainer and classifier outcomes across three datasets: ImageNette, Pet, and MURA, focusing on the overlap between correctly and wrongly classified samples. The high diagonal values indicate strong agreement between the explainer and classifier reaching an average value of 95.90% for correctly classified samples and 80.10% for misclassified samples (by the classifier). This overlap is particularly useful, as the explainer is designed to approximate the classifier's behavior. The consistent alignment between the two models suggests that the explainer is effectively reflecting the classifier's behavior.

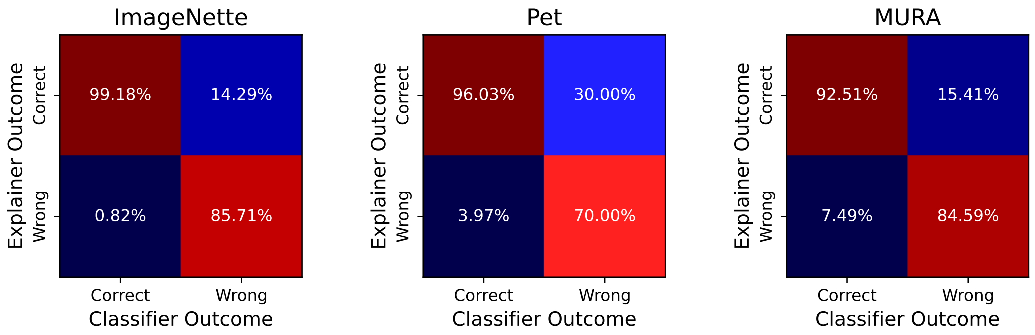


Figure 7: Confusion matrices illustrating the explainer and classifier outcomes across three datasets: ImageNette, Pet, and MURA. The matrices highlight the overlap between correctly and incorrectly classified samples. Each cell displays the percentage of samples classified correctly or incorrectly by both models, with rows representing the explainer's outcomes and columns representing the classifier's outcomes. The values are column-averaged, with each column summing the correct and incorrect outcomes of the classifier.

A.5 QUALITATIVE EXAMPLES

We present further qualitative examples comparing Salvage to the top 6 baseline methods in Figure 8. The figure includes 5 examples from ImageNette, 5 examples from Pet, and 4 examples from MURA.

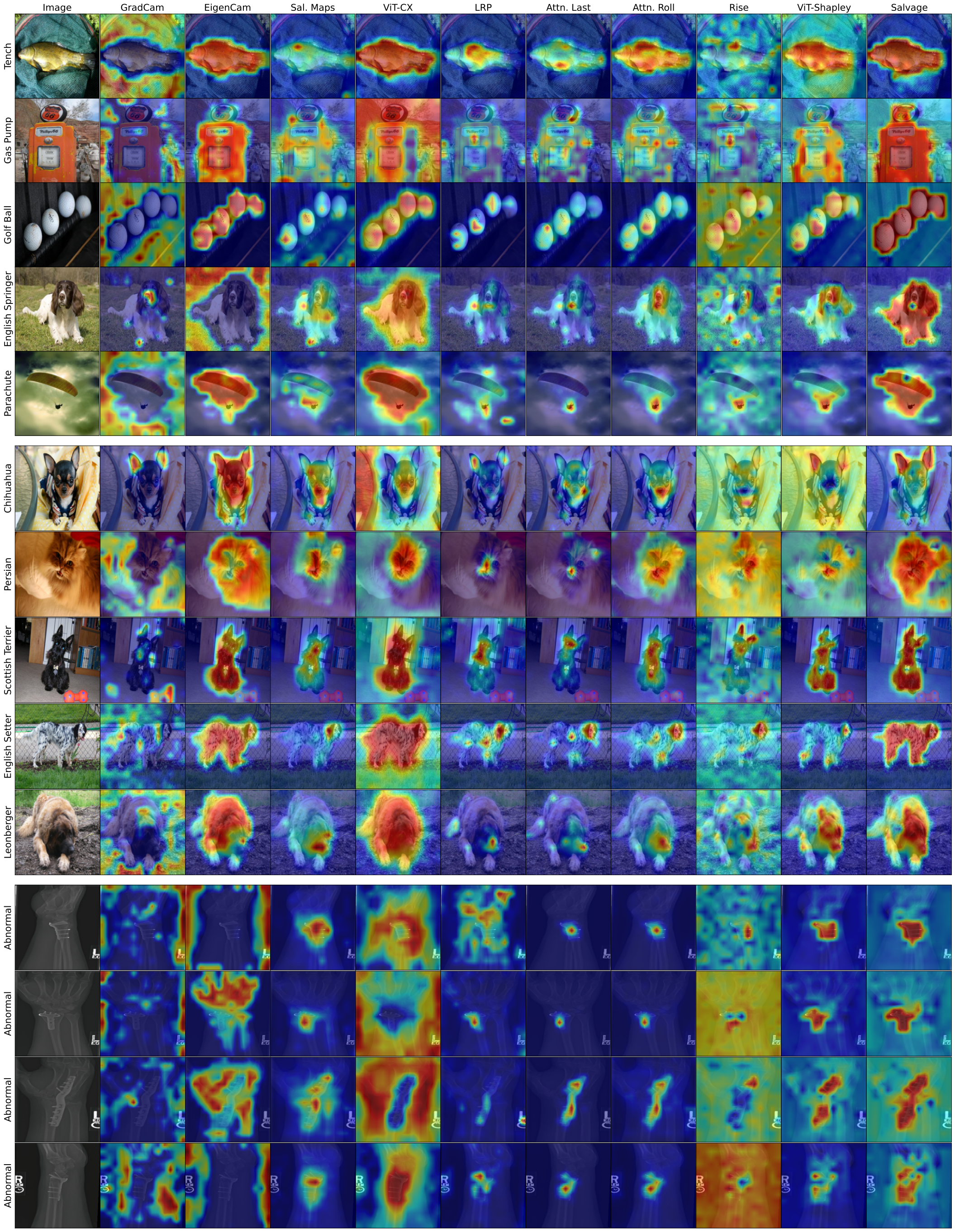


Figure 8: Qualitative examples computed on ImageNette, Pet and MURA.