Soft Local Completeness: Rethinking Completeness in XAI

Ziv Weiss Haddad* Oren Barkan*
The Open University

Yehonatan Elisha Noam Koenigstein
Tel Aviv University

https://github.com/xaisloc/sloc

Abstract

Completeness is a widely discussed property in explainability research, requiring that the attributions sum to the model's response to the input. While completeness intuitively suggests that the model's prediction is "completely explained" by the attributions, its global formulation alone is insufficient to ensure faithful explanations. We contend that promoting completeness locally within attribution subregions, in a soft manner, can serve as a standalone guiding principle for producing faithful attributions. To this end, we introduce the concept of the completeness gap as a flexible measure of completeness and propose an optimization procedure that minimizes this gap across subregions within the attribution map. Extensive evaluations across various model architectures demonstrate that our method produces state-of-the-art results.

1. Introduction

In recent years, Explainable AI (XAI) has emerged as a critical aspect of machine learning, especially for deep learning models, which are often regarded as "black boxes" due to their lack of transparency [8, 9, 25, 41, 66]. XAI aims to provide insights into how these models make predictions, offering a way to ensure their reliability and trustworthiness [34]. A common approach to model explainability is through attribution maps, which assign importance scores to individual input features based on their contribution to the model's prediction [22, 27, 65]. These attributions serve as a tool for understanding which aspects of the input were most influential in determining the output.

A widely discussed notion in XAI is *completeness* (also known as conservation or efficiency) [4, 55, 60, 65]. Completeness is defined as the requirement that the elements in an attribution map should sum to the difference between the model's prediction for a given input and a baseline (e.g., the null representation). This difference is also known as the model's *response*. When completeness is achieved inherently by an attribution method, it is intuitively appealing, as it suggests that the attribution map fully captures the factors contributing to the model's response. However, even a low-quality attribution map can be superficially adjusted to satisfy completeness through post hoc global normalization. Yet, such an approach lacks genuine explanatory power, as the normalization is decoupled from the underlying attribution mechanism.

This paper recognizes that completeness, as defined in the literature as a *global* property that an explanation method should satisfy, is too weak to serve as a standalone criterion. Indeed, several well-regarded explanation methods [32, 39, 59] do not inherently satisfy completeness. Nevertheless, these methods have demonstrated considerable effectiveness across diverse objective evaluation metrics, providing faithful explanations and successfully passing various sanity checks [3]. This suggests that completeness is neither a necessary nor sufficient condition for generating high-quality explanations.

In this work, we differentiate between *global* completeness, as originally defined in the literature, which imposes a completeness constraint on the entire attribution map, and *local* completeness, which requires completeness within individual subregions of the attribution map. By rethinking completeness as a local and flexible guiding measure rather than a strict global constraint, we seek to overcome its limitations while harnessing the appealing motivation behind it for producing meaningful and faithful explanations. To achieve this, we require subregions of the attribution map, referred to as *sub-maps*, to promote completeness locally by accounting for the model's response to the corresponding subregions of the input image, yet in soft manner. An attribution map that adheres to this guiding principle should assign high (low) importance to image subregions where the model exhibits a strong (weak) response, thereby maintaining faithfulness to the model.

Unlike global completeness, which is a relatively weak requirement that can be easily satisfied, achieving local completeness across an arbitrary set of sub-maps may not

*Equal contribution.

always be feasible. To address this challenge, we introduce the *completeness gap* - a measure that quantifies the deviation of a sub-map from completeness. Specifically, the completeness gap measures the difference between the sum of elements in each sub-map and the model's response to the corresponding subregions of the input image. Rather than imposing a strict binary constraint, the completeness gap serves as a soft, quantifiable criterion for local completeness. We argue that minimizing this gap over a diverse set of sub-maps leads to faithful explanations.

To this end, we present **S**oft **LO**cal **C**ompleteness (SLOC) - a novel explainability method that promotes completeness locally within sub-maps of the attribution map in a soft and flexible manner. SLOC accomplishes this by seeking to minimize the completeness gap locally for each individual sub-map, simultaneously.

The motivation for minimizing the completeness gap locally is to emphasize or attenuate sub-maps based on the actual impact of their corresponding input subregions on the model's output. SLOC achieves this through a gradient-based optimization process, where each sub-map is adjusted to achieve local completeness in a soft manner. Therefore, SLOC's goal is **not** to enforce strict completeness, globally, but to use it as a **guiding principle** for refining a large set of sub-maps (subregions within the attribution), iteratively and simultaneously.

Our contribution is the introduction of the SLOC method, which facilitates a novel optimization procedure to promote completeness in a soft local manner. The effectiveness of SLOC is demonstrated through extensive experiments on various model architectures, where it is shown to produce state-of-the-art results across multiple benchmarks.

2. Related Work

The XAI literature encompasses a broad range of approaches for attributing model predictions to specific input features across different tasks [2, 4, 6, 7, 16, 24, 29, 40, 54]. Early gradient-based methods produce explanation maps by leveraging gradients directly [5, 60, 62] or through functions that combine class activation maps with or without their gradients [10, 26, 48, 58, 59, 63]. With the rise of transformer architectures, new explainability techniques for transformers have emerged [1, 4, 11, 67]. For example, Transformer Attribution [28] introduces a class-specific Deep Taylor Decomposition, applying relevance propagation for both positive and negative attributions. Building on this, Generic Attention Explainability [27] generalizes Transformer Attribution to explain bi-modal transformers.

Path integration methods form another prominent family of attribution techniques [13, 15, 18, 50, 65]. Integrated Gradients (IG) [65] computes attributions by integrating gradients along an interpolation path from a baseline to the input. A refined variant, Guided Integrated Gradients (GIG) [50], uses an adaptive path that avoids high-gradient regions, thereby reducing irrelevant attributions. Recently, Deep Integrated Explanations (DIX) [14] proposed performing integration over intermediate network representations instead of inputs, producing more faithful attribution maps than IG and GIG.

Perturbation-based (occlusion) techniques generate attribution maps by perturbing parts of the input to assess the contributions of specific elements to the prediction [17, 19, 20, 53]. For example, RISE [57], creates perturbations by masking areas in the image through the upsampling of randomly drawn low-resolution binary grids. The class score corresponding to each masked version of the image serves as an importance score for that specific mask. A linear combination of all masks, weighted by their importance, then forms the final attribution map. Another branch of perturbation methods are learning-based perturbation techniques [12, 30, 38, 39]. For example, Meaningful Perturbation (MP) [39] learns to produce perturbations by masking the smallest region possible while significantly altering the original prediction, isolating the minimum necessary content for the prediction. Alternatively, LTX [12] introduces a surrogate 'explainer' model pretrained to mask as much of the input as possible while preserving the original prediction, thereby ensuring the retained features are those most relevant to the model's prediction. Then, LTX finetunes the attribution per specific example, allowing the selection of the best-performing attribution w.r.t. to the metric at hand.

Our SLOC method falls within the family of learning-based perturbation techniques, yet it diverges from prior works in several key aspects. First, conceptually, SLOC presents a distinct methodology by promoting local completeness rather than attempting to maximize or minimize the model prediction through learned perturbations [12, 30, 39]. Consequently, SLOC does not require gradient backpropagation through the model, and incorporates the model's predictions on the masked inputs into the loss as constant nodes within the computation graph. This design yields a simple, computationally efficient gradient expression, and relies solely on the model's forward pass, hence operating in a true black-box setting. Finally, unlike LTX, SLOC avoids the overhead of learning an 'explainer' function, which would also require a dataset aligned with the training data distribution of the model being explained.

3. Soft Local Completeness

Let $f : \mathbb{R}^n \to [0,1]^C$ be a classification model that takes an input image[1] $\mathbf{x} \in \mathbb{R}^n$ and outputs a discrete probability distribution $f(\mathbf{x}) \in [0,1]^C$ over C classes. Given a target class y to explain, our goal is to produce an attribution map

[1] W.l.o.g, we represent images as vectors in $\mathbb{R}^n$.

$\mathbf{a}_{\mathbf{x}}^{y} \in \mathbb{R}^n$ that attributes the contribution of each element i in $\mathbf{a}_{\mathbf{x}}^{y}$ to the prediction score $f(\mathbf{x})[y]$ for the class y.

Completeness requires that the attributions in $\mathbf{a}_{\mathbf{x}}^{y}$ collectively account for the difference between the model's prediction on the input $\mathbf{x}$ and the *baseline* representation $\mathbf{b}$, which is set to the black image[2] in this work.

The model's *response* to $\mathbf{x}$ is defined as $r(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{b})$. Formally, the attribution map $\mathbf{a}_{\mathbf{x}}^{y}$ satisfies completeness if $\sum_{i=1}^{n} \mathbf{a}_{\mathbf{x}}^{y}[i] = r(\mathbf{x})[y]$. We note that this property can be superficially imposed on an attribution map through a post-processing step by normalizing each element using the factor $\frac{r(\mathbf{x})[y]}{\sum_{i=1}^{n} \mathbf{a}_{\mathbf{x}}^{y}[i]}$. However, artificially satisfying completeness through such normalization does not inherently yield a meaningful explanation, as the imposed completeness is an external adjustment rather than an intrinsic characteristic of the explanation method itself. For example, even a flat attribution map (where all elements are equally weighted) could be normalized to satisfy the completeness criterion, yet such a map lacks any explanatory value. Consequently, relying solely on completeness, in its global form, is deemed insufficient for producing meaningful attribution maps.

3.1. Motivation

Our SLOC approach is motivated by the **toy** example illustrated in Fig. 1, which depicts an image where the key region influencing the model's prediction for the *'great-grey-owl'* class is approximately centered. Figures 1(a)–(d) present four images: the original image (a) and three perturbed versions obtained by masking different regions (b)–(d).

The four different masks applied in Figures 1(a)–(d) are illustrated in Figure 1(e). Each mask selectively perturbs the original image by replacing certain regions with the baseline (in this work, the null pixels). For instance, the mask corresponding to Figure 1(c), shown in the top-right corner of Figure 1(e), applies a perturbation by replacing all pixels within the black rectangle (located at the bottom-right corner) with the baseline.

Notably, the model's response for the *great-grey-owl* class remains consistent across all four images (a-d), with a value of 0.8. Recall that the model response is defined as the difference between the model's class probability for a given image and that for the baseline (the null image).

As defined earlier, a **sub-map** represents a subset of elements in the attribution map that are associated with specific subregions in the input. In the context of Fig. 1, we consider four different sub-maps, each corresponding to the non-masked (visible) subregions in Figs.1(a)–(d). For instance, the sub-map corresponding to Fig. 1(c) consists of all elements in the attribution map **except** those associated with the black rectangle in the bottom-right corner of Fig.1(c). This exclusion aligns with the masked region in the input, effectively defining which attributions are considered within that sub-map. Notably, the sub-map corresponding to the original image (Fig. 1(a)) is equivalent to the **entire** attribution map, as no regions are masked. This formulation allows us to analyze how different subregions contribute to completeness by evaluating attributions over various sub-maps.

A sub-map satisfies *local completeness* if the sum of attributions within it equals the model's response for the corresponding perturbed image. For example, the sub-map induced by Fig. 1(c) satisfies local completeness if the sum of its elements equals 0.8, which is the model's response to Fig. 1(c).

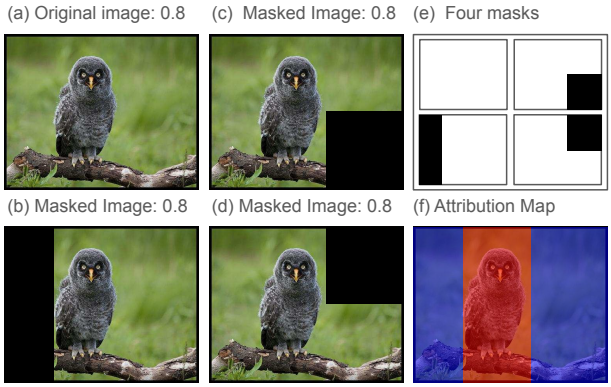

Figure 1. Subfigures (a)–(d) present the model's response (0.8) for the class *'great-grey-owl'* when applied to the original input (a) and its masked versions (b)–(d). The four masks used to generate (a)–(d) are depicted in (e). A *sub-map* refers to a subset of elements in the *attribution map* that correspond to specific subregions in the input. We consider four distinct sub-maps (not shown), each corresponding to the non-masked (visible) subregions in (a)–(d). For instance, the sub-map associated with (c) includes all elements in the attribution map *except* those corresponding to the black rectangle in the bottom-right corner of (c). This exclusion aligns with the masked region in the input, effectively determining which attributions are included in that sub-map. A sub-map satisfies *local completeness* if the sum of attributions within it equals the model's response for the corresponding perturbed image. For example, the sub-map induced by (c) satisfies local completeness if the sum of its elements equals 0.8, which is the model's response to (c). By enforcing local completeness across the four sub-maps induced by (a)–(d), the remaining central stripe, which is unmasked in all images (a)–(d), must account for the model's entire response of 0.8. This results in the attribution depicted in (f) that highlights the central stripe containing the owl in red. See 3.1 for more details.

[2]The baseline representation should resemble missing information or the neutral representation. While determining the ultimate representation for the baseline is an open question and an active research field in explainability [21, 36], in this work, we opt for the simple choice of a null baseline represented by the black image. This choice is on par with other notable works [57, 65]. Additionally, we conducted experiments using alternative baselines and observed that the resulting trends remained consistent with those obtained using the black image.

Our key insight is that enforcing local completeness across the four sub-maps induced by Figs.1(a)-(d) allows us to pinpoint the important region of the image w.r.t. the model's prediction. Specifically, the intersection of these four sub-maps forms a central stripe around the owl, illustrated in red in Fig.1(f). By evenly distributing the total attribution sum of 0.8 across the pixels within this intersection (red stripe) and assigning 0 elsewhere (highlighted in dark blue in Fig. 1(f)), we obtain an attribution map that ensures local completeness for all four sub-maps. Since each sub-map contains the intersection while the attribution values outside it are zero, the sum of attributions within each sub-map is exactly the sum of attributions within the intersection: 0.8. This resulting attribution map effectively captures the key region containing the owl.

This simple example can be slightly generalized under similar settings. We make the following assumptions: (1) The model response for a subregion is R if the subregion contains the object of interest and 0 otherwise. (2) Each sub-map either fully contains the object of interest or does not intersect with it at all. (3) One of the sub-maps corresponds to the complete attribution map. (4) All attribution values are non-negative. It can then be shown that if an attribution map satisfies local completeness for all sub-maps, the intersection defined by these sub-maps is the largest subregion that can have non-zero attribution. Simply put, this intersection highlights the important regions in the image responsible for the model's prediction. Moreover, as the number of sub-maps increases, the important regions can potentially become more precisely defined. The formalization and proof follow straightforwardly and are provided in Appendix K.

Keep in mind, however, that this setting and its assumptions are simplistic and unrealistic, serving only to illustrate the motivation. Therefore, SLOC does not actually rely on these assumptions. In more realistic scenarios, satisfying local completeness for all sub-maps is often infeasible (an example is provided in Fig. 11 in Appendix I). To address this challenge, we introduce the *completeness gap*, which quantifies the degree to which a sub-map deviates from local completeness. Our approach seeks to minimize the completeness gap across multiple sub-maps simultaneously, hence promoting completeness locally within subregions of the attribution map, in a soft manner. We formulate this as an optimization problem and employ gradient descent to find a solution.

Finally, recall that in our simple example, in the absence of additional information, we distributed the attribution uniformly across the central red stripe. When introducing the loss terms for the SLOC optimization, we will also discuss regularization terms. One such term is the Total Variation loss, which encourages spatial smoothness in the attribution. As we demonstrate in Sec. 4, by encouraging this soft form of local completeness, the resulting attribution map offers robust explanatory performance, both visually and according to a variety of evaluation metrics.

3.2. SLOC optimization

Let $\mathbf{m} \in \{0,1\}^n$ be a binary mask. We define the masked input by

$$\mathbf{x}^{\mathbf{m}} = \mathbf{x} \circ \mathbf{m} + (1 - \mathbf{m}) \circ \mathbf{b}, \tag{1}$$

where $\circ$ stands for the elementwise product. In words, $\mathbf{x}^{\mathbf{m}}$ is a perturbation (masked version) of $\mathbf{x}$, according to the mask $\mathbf{m}$, where all masked elements of $\mathbf{x}^{\mathbf{m}}$ are replaced with their respective elements from the baseline $\mathbf{b}$.

In this work, we consider binary masks constructed over patches, where each patch is randomly assigned a Bernoulli outcome with probability p. We explore two methodologies for automatically selecting p: one that determines p dynamically per input image, and another that uses a fixed value per model. The reader is referred to Appendix D for the exact details of the masks construction process, including the methodologies used for determining p.

Given an initial attribution map $\mathbf{a}_{\mathbf{x}}^{y}$ and a set of masks $\mathcal{M} \subset \{0,1\}^n$, the SLOC loss is defined as follows:

$$\mathcal{L}_c(\mathbf{a}_{\mathbf{x}}^{y}; \mathcal{M}) = \frac{1}{2|\mathcal{M}|} \sum_{\mathbf{m} \in \mathcal{M}} \frac{1}{|\mathbf{m}|} (\underbrace{r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m}}_{\text{completeness gap}})^2, \tag{2}$$

where $\cdot$ is the dot-product operator, and $|\cdot|$ denotes the L1 norm. As observed in Eq. 2, $\mathcal{L}_c$ promotes the minimization of the *completeness gap*, a soft measure of completeness, locally, for all sub-maps induced by $\mathcal{M}$. The rationale behind the local minimization of the completeness gap is that attributions within a sub-map should be strengthened or reduced so that their sum closely matches $r(\mathbf{x}^{\mathbf{m}})[y]$ - the model's response to the respective regions exposed in $\mathbf{x}^{\mathbf{m}}$. This behavior is achieved by applying gradient descent on $\mathcal{L}_c$ with respect to $\mathbf{a}_{\mathbf{x}}^{y}$.

To better understand the optimization process, one can inspect the gradient:

$$\frac{\partial \mathcal{L}_c}{\partial \mathbf{a}_{\mathbf{x}}^{y}} = -\frac{1}{|\mathcal{M}|} \sum_{\mathbf{m} \in \mathcal{M}} \frac{1}{|\mathbf{m}|} (r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m})\mathbf{m}. \tag{3}$$

For simplicity, assume $\mathcal{M}$ contains a single mask $\mathbf{m}$. Then, for each element $\mathbf{a}_{\mathbf{x}}^{y}[i]$ that is exposed in $\mathbf{x}^{\mathbf{m}}$ (i.e., $\mathbf{m}[i] = 1$), the gradient is the negative of the completeness gap normalized by the number of elements in the sub-map, i.e., $\frac{\partial \mathcal{L}_c}{\partial \mathbf{a}_{\mathbf{x}}^{y}[i]} = -\frac{1}{|\mathbf{m}|}(r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m})$, and for each i where $\mathbf{m}[i] = 0$, the gradient vanishes. Accordingly, the gradient descent update ensures an equal increase (decrease) for all elements in the sub-map when the sum of the attributions falls short of (exceeds) the model's response $r(\mathbf{x}^{\mathbf{m}})[y]$. Assuming a learning rate of 1, it follows that after the gradient descent update, the completeness gap vanishes, resulting in

an updated sub-map satisfying completeness. This can be seen mathematically as follows: for a sub-map induced by the mask $\mathbf{m}$, the sum of its elements *after* the gradient update is given by

$$\begin{aligned}
&\sum_{i:\mathbf{m}[i]=1} \mathbf{a}_{\mathbf{x}}^{y}[i] - \frac{\partial \mathcal{L}_c}{\partial \mathbf{a}_{\mathbf{x}}^{y}[i]} \\
&\quad = \sum_{i:\mathbf{m}[i]=1} \mathbf{a}_{\mathbf{x}}^{y}[i] + \underbrace{\frac{1}{|\mathbf{m}|}(r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m})}_{\text{independent of } i} \\
&\quad = |\mathbf{m}|\frac{1}{|\mathbf{m}|}(r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m}) + \sum_{i:\mathbf{m}[i]=1} \mathbf{a}_{\mathbf{x}}^{y}[i] \\
&\quad = r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m} + \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m} = r(\mathbf{x}^{\mathbf{m}})[y],
\end{aligned}$$

where the second transition follows from the fact that the normalized completeness gap is independent of i and is therefore summed $|\mathbf{m}|$ times, as the summation is taken over every i where $\mathbf{m}[i] = 1$, with $\mathbf{m}$ being a binary tensor. The penultimate transition then arises from expressing the sum of elements in the sub-map as the dot product, $\mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m}$. This derivation justifies the normalization of the squared completeness gap by $|\mathbf{m}|$ in Eq. 2, as it ensures that after the gradient update, the sub-map satisfies local completeness.

To give further intuition for using the normalized completeness gap in Eq. 2, consider two different masks: one that exposes a single pixel (element), and another that exposes a large portion of the input. Given that the gradient update of elements in the sub-map is equal for all elements, it makes sense to distribute the completeness gap among all elements. In the case of a mask exposing a single pixel, we have full confidence that this specific element is responsible for the model's response $r(\mathbf{x}^{\mathbf{m}})[y]$. Therefore, it is reasonable to expect that the gradient update for this element will match the size of the completeness gap. However, in the case of a sub-map associated with a mask exposing large portions of the input, we cannot differentiate the accountability of individual elements within the exposed portions for the model's response. Thus, without further information, it is reasonable to apply an update step that evenly distributes the completeness gap across all exposed elements. Hence, the squared completeness gap is divided by $|\mathbf{m}|$.

So far, the discussion and analysis have been limited to the case of a single mask, i.e., $|\mathcal{M}| = 1$. When considering a set of masks, the update for $\mathbf{a}_{\mathbf{x}}^{y}[i]$ is determined by the accumulation of gradients from all the sub-maps it is associated with. Specifically, sub-maps that include the i-th element and whose total sum of elements is below (exceeds) the model's response $r(\mathbf{x}^{\mathbf{m}})[y]$ will contribute to the intensification (reduction) of $\mathbf{a}_{\mathbf{x}}^{y}[i]$, with the contribution from each individual mask weighted according to the normalized completeness gap it induces.

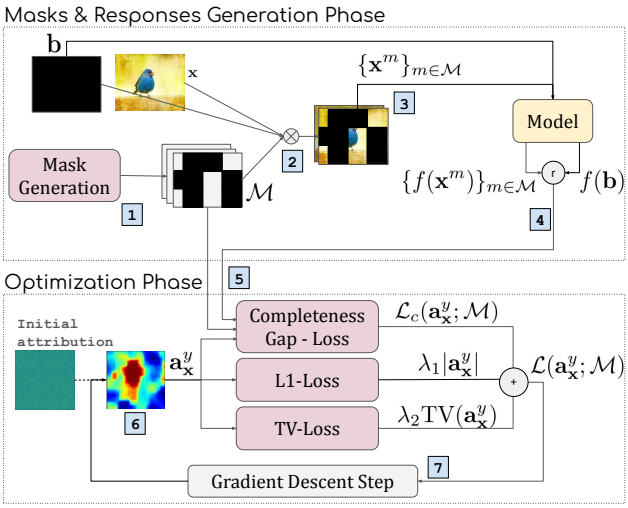

Figure 2. **SLOC overview:** The upper section illustrates the *Masks & Responses Generation Phase*. In this phase, (1) $\mathcal{M}$ masks are randomly generated. (2) Perturbations are produced by combining the original image and baseline according to the masks, as defined in Eq. (1) (3) These perturbations are passed as inputs to the model, and the corresponding outputs are obtained. (4) Subtracting the model's output for the baseline from these outputs produces the model responses. (5) These masks and their corresponding responses are then passed to the *Optimization Phase*. (6) Starting with an initial attribution. (7) Gradient descent is used to iteratively update the attribution. The loss function $\mathcal{L}$ (Eq. 4) consists of three components: the *completeness gap* (computed using the masks, responses, and attributions), as well as TV and L1 regularization. See Sec. 3.2 for details.

To promote smooth and focused attribution maps, we combine $\mathcal{L}_c$ with two types of regularization applied to the attribution map: Total Variation (TV) and L1 regularization that encourage smoothness and sparsity, respectively. The final loss function is then defined as:

$$\mathcal{L}(\mathbf{a}_{\mathbf{x}}^{y};\mathcal{M}) = \mathcal{L}_c(\mathbf{a}_{\mathbf{x}}^{y};\mathcal{M}) + \lambda_1|\mathbf{a}_{\mathbf{x}}^{y}| + \lambda_2\text{TV}(\mathbf{a}_{\mathbf{x}}^{y}), \quad (4)$$

with

$$\text{TV}(\mathbf{a}_{\mathbf{x}}^{y}) := \sum_{i,j}(\mathbf{a}_{\mathbf{x}}^{y}[i,j]-\mathbf{a}_{\mathbf{x}}^{y}[i+1,j])^2+(\mathbf{a}_{\mathbf{x}}^{y}[i,j]-\mathbf{a}_{\mathbf{x}}^{y}[i,j+1])^2 \quad (5)$$

where $\mathbf{a}_{\mathbf{x}}^{y}$ in Eq. 5 is reshaped to match the spatial dimensions of the final attribution map. Additionally, λ_1 and λ_2 are hyperparameters controlling the strength of the regularization terms. In our experiments, these hyperparameters are set based on a separate validation set.

The optimization of $\mathbf{a}_{\mathbf{x}}^{y}$ is carried out via gradient descent on $\mathcal{L}$. The optimization process allows for tracking the performance of the metric during the optimization and selecting the best-performing attribution map based on this metric from the resulting maps [12]. A diagram of the SLOC method is depicted in Fig. 2.

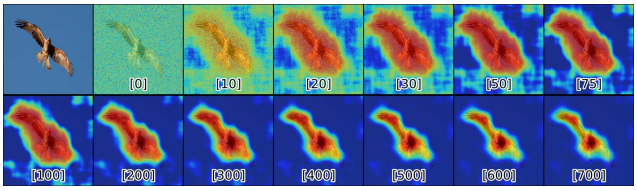

Figure 3. SLOC attribution maps across training steps. Faithful explanations emerge after a few hundred gradient updates, with an appropriate learning rate decay.

In practice, we observed that a few hundred gradient updates, combined with an appropriate learning rate decay, were sufficient to produce faithful and meaningful explanations. Therefore, in our experiments, we report SLOC results for $T = 500$. An example for SLOC's attribution across gradient update steps is shown in Fig. 3. For the exact implementation details and hyperparameter configuration, the reader is referred to Appendix D.

Due to space limitations, Appendix G provides a detailed discussion of SLOC's computational complexity. Appendix H presents a runtime analysis, comparing SLOC against various methods and demonstrating its efficiency over other state-of-the-art learning-based perturbation methods. Additionally, Appendix L examines additional axioms and explains which of them are satisfied by SLOC.

Finally, it is important to clarify that the SLOC optimization procedure does not guarantee convergence to an attribution map that fully satisfies completeness. Instead, it leverages the completeness gap as a *guiding measure* to update subregions within the attribution map, thereby promoting local completeness for these subregions in a soft manner. Nevertheless, as demonstrated in Sec. 4, SLOC has proven effective in generating attribution maps that outperform state-of-the-art explanation methods.

4. Experiments and Results

4.1. Experimental setup

The experiments were conducted on an NVIDIA DGX 8xA100 Server, utilizing the PyTorch package. As explanations are multifaceted by nature, no single evaluation metric can fully capture their quality, and no definitive metric exists [21, 23, 35, 45, 47, 51, 68]. Therefore, our evaluation encompasses several distinct protocols designed to assess explanation quality from multiple complementary perspectives, on different datasets, and across four model architectures: ResNet50 (**RN**) [43], DenseNet201 (**DN**) [46], ViT-Base (**ViT-B**) and ViT-Small (**ViT-S**) [33].

Experiment 1: Faithfulness evaluation This evaluation assesses the faithfulness (also known as correctness or fidelity) of the explanation via a set of perturbation-based metrics. These metrics are designed to reveal the actual input elements the model relies on for its predictions. Following prior works [28, 49, 50, 69], we conducted an extensive evaluation using a comprehensive set of perturbation-based explanation metrics to assess the faithfulness of the generated explanations: the Area Under the Curve (AUC) of Positive (**POS**) and Negative (**NEG**) perturbation tests [28], AUC of the Insertion (**INS**) and Deletion (**DEL**) tests [57], and AUC of the Softmax Information Curve (**SIC**) and Accuracy Information Curve (**AIC**) [49]. For POS and DEL, lower values indicate better performance, while for NEG, INS, SIC, and AIC, higher values are preferred. Since NEG and POS, as well as INS and DEL, are complementary metrics, we also report the NEG-POS Difference (**NPD**) and the INS-DEL Difference (**IDD**) as single summaries of their respective complementary performances. Faithfulness performance is reported on a sample from the ImageNet [31] ILSVRC 2012 (**IN**) validation set, which consists of 10,000 images across 1,000 classes. Since most images in the IN dataset contain a single, centrally focused object, we extend our evaluation to multi-object scenes by assessing explanation faithfulness on a sample from the Pascal VOC 2012 [37] (**VOC**) test set, which includes 1,000 images spanning 20 object classes.

Experiment 2: FunnyBirds evaluation While perturbation-based protocols are a standard approach for assessing explanation faithfulness, they are criticized for introducing domain shifts that undermine the validity of the evaluation [45]. To address this, we additionally report results on the FunnyBirds (**FB**) dataset which consists of 500 images from 50 classes, following the evaluation protocol from [44]. The FB evaluation employs controlled interventions to estimate ground-truth (GT) importance scores at the level of object *parts* rather than individual *pixels*, mitigating issues associated with pixel-wise perturbation-based evaluations [68]. Furthermore, FB avoids domain shifts by incorporating semantically meaningful interventions during training. By evaluating explanations at the part level—closer to human visual understanding—FB reduces the gap between automatic XAI evaluation (e.g., faithfulness metrics) and human-centric studies, providing a more interpretable and robust assessment of the explanations. FB evaluates explainability in three aspects: **Completeness**, **Correctness**, and **Contrastivity** - and provides a combined overall score (higher is better).

Experiment 3: Segmentation evaluation To ensure a comprehensive comparison with prior works [27, 28, 48], we conduct an extensive evaluation using segmentation tests on the ImageNet-Seg dataset (**IN-Seg**) [42], a subset of the ImageNet validation set comprising 4,276 human-annotated

ground-truth (GT) segmentation maps across 445 classes. As detailed in [28], these tests generalize the"Pointing Game" [45] and serve as a human-grounded evaluation [56] by measuring the alignment between GT segmentation maps and the attributions generated by the explanation method. While higher segmentation accuracy does not necessarily indicate greater explanatory power [57], segmentation tests assess the alignment between explanations and human-annotated GT, hence providing a complementary perspective to other evaluations. We follow the evaluation protocol from [28], assessing segmentation performance using Pixel Accuracy (**PA**), mean Intersection-over-Union (**mIoU**), and mean Average Precision (**mAP**), higher is better for all metrics.

Detailed descriptions of the evaluation protocols and metrics used in Experiments 1, 2, and 3 are provided in Appendices B.1, B.2, and B.3, respectively.

Evaluated methods We evaluate SLOC against a comprehensive suite of 14 explanation methods encompassing gradient-based, path-integration, perturbation, and CAM methods. For CNN models, we include the following methods: Grad-CAM (**GC**) [59], Grad-CAM++ (**GC++**) [26], Deep Integrated Explanations (**DIX**) [14], FullGrad (**FG**) [63], Ablation-CAM (**AC**) [32], Layer-CAM (**LC**) [48], Learning To Explain (**LTX**) [12], **RISE** [57], Meaningful Perturbation (**MP**) [39], Extremal Perturbations (**EP**) [38], Integrated Gradients (**IG**) [64], and Guided IG (**GIG**) [50]. For ViT models, we considered the following methods: Transformer Attribution (**T-Attr**) [28] and Generic Attention Explainability (**GAE**) [27], alongside DIX, RISE, LTX, EP, and MP, which are applicable to both CNN and ViT architectures. Hyperparameters for all methods were set according to the recommended settings published by the authors, unless a better configuration was found. A detailed description of all explanation methods is provided in Appendix C. Finally, we evaluated three different versions of our SLOC method on both CNN and ViT models: **SLOC**, which tunes p dynamically per input, **SLOC$_{\mathbf{xp}}$**, where p is calibrated per model (i.e., fixed across all inputs for a given model) based on a validation dataset, and **SLOC$_{\mathbf{m}}$**, which is identical to **SLOC** but additionally monitors the IDD metric on the specific input and selects the explanation with the best performance according to this metric. All versions were run with $T = 500$. The precise implementation details, including the monitoring procedure, the p tuning procedures, optimization process, and all hyperparameter configurations, are provided in Appendix D and in our GitHub repository.

4.2. Results

Tables 1 and 2 present a quantitative comparison of SLOC with other state-of-the-art explanation methods across multiple faithfulness metrics, for DN and ViT-S, respectively.

Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
SLOC$_{\mathrm{m}}$	11.61	**76.54**	9.39	**65.79**	**64.93**	**56.4**	**79.21**	**78.35**
SLOC	**11.52**	70.97	**9.12**	60.53	59.45	51.41	77.79	77.72
SLOC$_{\mathrm{xp}}$	11.78	70.20	9.34	59.75	58.43	50.41	77.33	76.80
AC	17.24	67.68	13.27	57.18	50.44	43.91	77.78	75.41
DIX	13.36	62.95	10.36	52.43	49.59	42.07	74.62	71.47
EP	16.12	65.68	12.94	55.0	49.55	42.06	77.38	74.97
FG	19.06	44.66	15.26	37.62	25.6	22.37	58.23	53.86
GC	16.88	68.54	13.04	57.95	51.66	44.91	78.38	76.01
GC++	17.36	67.29	13.34	56.75	49.93	43.41	78.04	75.62
GIG	14.81	49.96	12.28	41.93	35.16	29.65	61.12	57.6
IG	14.14	51.99	11.2	44.08	37.85	32.88	61.26	58.48
LC	17.28	67.27	13.31	56.71	49.99	43.4	77.95	75.42
LTX	16.24	71.09	12.95	59.69	54.84	46.74	78.92	76.25
MP	18.54	53.24	14.87	43.79	34.7	28.92	66.98	64.13
RISE	18.42	62.75	14.26	52.99	44.33	38.73	76.82	74.24

Table 1. Faithfulness results for all combinations of method and metric, using the DN model on the IN dataset.

Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
SLOC$_{\mathrm{m}}$	**14.83**	**81.81**	**12.27**	**70.31**	**66.98**	**58.04**	**83.78**	**83.34**
SLOC	15.25	77.97	12.47	66.51	62.72	54.04	83.19	82.88
SLOC$_{\mathrm{xp}}$	15.79	77.85	12.85	66.18	62.06	53.33	83.06	81.97
DIX	18.69	68.17	14.86	56.83	49.48	41.97	76.88	75.05
EP	27.37	72.52	21.66	60.8	45.14	39.14	79.45	77.56
GAE	19.98	66.93	15.93	55.72	46.95	39.79	75.42	73.9
LTX	20.84	68.5	16.63	56.89	47.66	40.27	74.22	71.56
MP	27.72	63.25	22.46	52.28	35.53	29.81	74.22	71.25
RISE	29.51	69.52	23.47	57.93	40.02	34.46	79.93	77.23
TATTR	19.06	67.52	14.91	56.41	48.46	41.5	78.03	75.69

Table 2. Faithfulness results for all combinations of method and metric, using the ViT-S model on the IN dataset.

Results for RN and ViT-B are available in Tabs. 7 and 8 in Appendix E, respectively. Overall, the results indicate that all SLOC versions are the best-performing methods, with LTX, DIX, RISE, and EP as runners-up, depending on the combination of architecture and evaluation metric.

Among the three versions of SLOC, SLOC$_{\mathrm{m}}$ performs best. This can be attributed to its use of the IDD metric, which measures the difference between INS and DEL performance. The fact that SLOC$_{\mathrm{m}}$ outperforms across the vast majority of metrics suggests that the IDD metric correlates well with other faithfulness metrics. SLOC and SLOC$_{\mathrm{xp}}$ follow, with SLOC performing slightly better. This suggests that tuning p per instance offers a modest advantage in metrics, at the cost of increased inference complexity. When a representative sample of data is available, p can be calibrated per model using this data. Otherwise, p should be adjusted dynamically per input during inference.

Notably, both SLOC and SLOC$_{\mathrm{xp}}$ surpass learning-based approaches such as LTX, MP, and EP, achieving state-of-the-art results without requiring gradient backpropagation through the model (see the discussion on SLOC complexity in Appendix G) or metric monitoring (as in LTX). Figures 6 and 5 present comparative examples of the top-performing methods for DN and ViT-S. Arguably, SLOC produces focused explanation maps highlighting relevant class-discriminative features in the image.

Table 3 presents faithfulness results on the VOC dataset

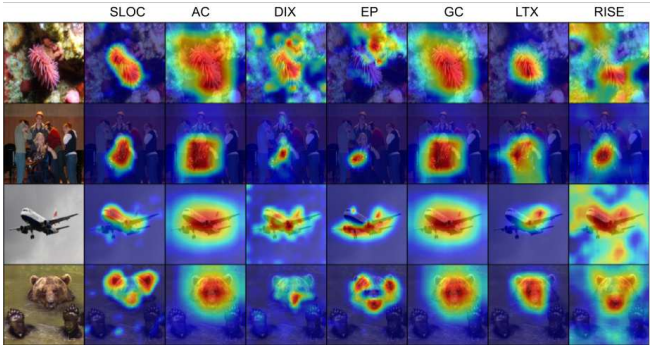

Figure 4. Qualitative comparison of attributions produced by different methods, using the DN model w.r.t. the classes (top to bottom): 'sea anemone', 'French horn', 'airliner', 'brown bear'.

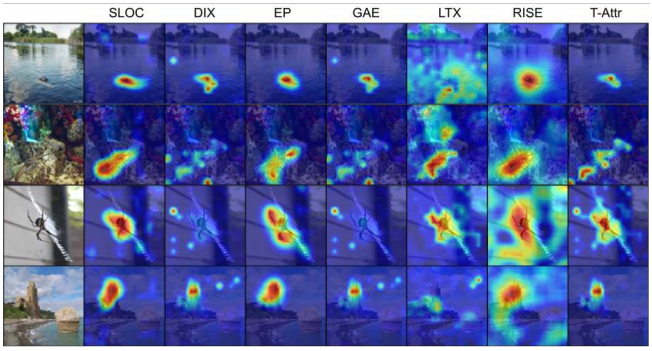

Figure 5. Qualitative comparison of attributions produced by different methods, using the ViT-S model w.r.t. the classes (top to bottom): 'American aligator', 'spiny lobster', 'black and gold garden spider', 'promontory'.

Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	**7.31**	**69.2**	**5.07**	**48.83**	**61.89**	**43.76**	**75.63**	**79.09**
SLOC	7.32	63.56	5.09	44.58	56.25	39.5	74.96	77.84
$SLOC_{xp}$	7.61	63.23	5.22	43.72	55.65	38.48	73.69	77.59
DIX	10.19	48.66	6.74	31.73	38.47	24.99	64.51	68.66
EP	14.06	55.96	9.26	36.55	41.89	27.29	68.43	72.04
GAE	11.2	47.4	7.47	30.92	36.20	23.44	64.22	67.95
LTX	12.56	49.28	8.2	32.26	36.71	24.06	56.55	63.16
RISE	15.67	53.21	10.26	35.35	37.53	25.09	66.91	70.5
T-Attr	10.37	47.56	6.81	31.35	37.19	24.54	65.62	69.53

Table 3. Faithfulness results for combinations of method and metric, using the ViT-S model on the VOC dataset.

using ViT-S. The VOC dataset contains multi-object images, allowing multiple classes to appear within the same image. These results demonstrate the superior performance of all SLOC variants not only on single-object images but also on images containing multiple objects from different classes. Results for the DN model are provided in Tab. 9 in Appendix E.

Tables 4 and 5 present the overall summarized performance scores (higher is better) on the FB benchmark for the RN and ViT-B models, respectively. Detailed fine-grained results are provided in Appendix E. Notably, SLOC

Method	SLOC	AC	DIX	EP	FG	GC	GC++	GIG	IG	LC	RISE
Score	**0.78**	0.70	0.72	0.73	0.70	0.72	0.72	0.56	0.63	0.72	0.62

Table 4. FunnyBirds results for the RN model.

Method	SLOC	DIX	EP	RISE	T-Attr
Score	**0.88**	0.87	0.79	0.77	0.87

Table 5. FunnyBirds results for the ViT-B model.

Method	SLOC	AC	DIX	EP	GC	GC++	GIG	IG	LC	LTX	RISE
mIoU↑	0.56	0.55	**0.66**	0.52	0.55	0.56	0.51	0.48	0.55	0.56	0.51
mAP↑	0.80	**0.86**	0.84	0.76	0.85	0.85	0.78	0.76	0.85	0.83	0.79
PA↑	0.77	0.72	**0.82**	0.72	0.73	0.73	0.74	0.79	0.73	0.51	0.7

Table 6. Segmentation tests results for the RN model.

achieves state-of-the-art performance on the FB benchmark as well. The distinct evaluation protocol employed in FB offers a complementary perspective to the faithfulness evaluation conducted on the IN dataset, further reinforcing our confidence in the robustness of SLOC.

Table 6 presents segmentation tests results on the IN-Seg dataset using the RN model. We observe that DIX achieves the best average performance across all metrics, while the runner-up varies depending on the specific metric. Overall, SLOC is found to be competitive, on average, with all runner-up methods. We note that higher segmentation accuracy may reflect a method's ability to facilitate strong object detection rather than to identify the most informative features that explain the model's prediction. In reality, the most explanatory features do not always cover the entire object; instead, they may correspond to a subset of features that are critical to the model's decision. Consequently, higher segmentation accuracy does not necessarily indicate superior explanatory value.

Due to space limitation, ablation studies, runtime comparison, sanity checks [3], and additional qualitative analyses, are provided in Appendices F, H, J, and I, respectively. These experiments offer further insights into the effectiveness and efficiency of SLOC.

5. Conclusion

This work introduced SLOC, a novel and efficient explainability method. By rethinking completeness as a guiding principle promoted locally rather than as a strict global constraint, SLOC exhibits state-of-the-art performance across various explainability benchmarks. These findings suggest that completeness, when promoted in a soft and local manner, provides a robust foundation for generating high-quality attribution maps that closely align with the model's predictive behavior and human comprehension. Finally, discussions on the limitations of SLOC and potential avenues for future research are provided in Appendix M.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 2

[2] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: Attention-aware layer-wise relevance propagation for transformers. In *International Conference on Machine Learning*, pages 135–168. PMLR, 2024. 2

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 1, 8, 12, 25, 26

[4] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *International conference on machine learning*, pages 435–451. PMLR, 2022. 1, 2

[5] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 2

[7] Oren Barkan, Yonatan Fuchs, Avi Caciularu, and Noam Koenigstein. Explainable recommendations via attentive multi-persona collaborative filtering. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 468–473, 2020. 2

[8] Oren Barkan, Ori Katz, and Noam Koenigstein. Neural attentive multiview machines. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2020. 1

[9] Oren Barkan, Idan Rejwan, Avi Caciularu, and Noam Koenigstein. Bayesian hierarchical words representation learning. *arXiv preprint arXiv:2004.07126*, 2020. 1

[10] Oren Barkan, Omri Armstrong, Amir Hertz, Avi Caciularu, Ori Katz, Itzik Malkiel, and Noam Koenigstein. Gam: Explainable visual similarity and classification via gradient activation maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 68–77, 2021. 2

[11] Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2882–2887, 2021. 2

[12] Oren Barkan, Yuval Asher, Amit Eshel, Yehonatan Elisha, and Noam Koenigstein. Learning to explain: A model-agnostic framework for explaining black box models. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 944–949. IEEE, 2023. 2, 5, 7, 15, 23

[13] Oren Barkan, Yehonatan Elisha, Yuval Asher, Amit Eshel, and Noam Koenigstein. Visual explanations via iterated integrated attributions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2073–2084, 2023. 2

[14] Oren Barkan, Yehonatan Elisha, Jonathan Weill, Yuval Asher, Amit Eshel, and Noam Koenigstein. Deep integrated explanations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 57–67, 2023. 2, 7, 15

[15] Oren Barkan, Yehonatan Elisha, Jonathan Weill, Yuval Asher, Amit Eshel, and Noam Koenigstein. Stochastic integrated explanations for vision models. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 938–943. IEEE, 2023. 2

[16] Oren Barkan, Tom Shaked, Yonatan Fuchs, and Noam Koenigstein. Modeling users' heterogeneous taste with diversified attentive user profiles. *User Modeling and User-Adapted Interaction*, pages 1–31, 2023. 2

[17] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. A counterfactual framework for learning and evaluating explanations for recommender systems. In *Proceedings of the ACM Web Conference 2024*, pages 3723–3733, 2024. 2

[18] Oren Barkan, Yehonatan Elisha, Yonatan Toib, Jonathan Weill, and Noam Koenigstein. Improving llm attributions with randomized path-integration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9430–9446, 2024. 2

[19] Oren Barkan, Yonatan Toib, Yehonatan Elisha, and Noam Koenigstein. A learning-based approach for explaining language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 98–108, 2024. 2

[20] Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. Llm explainability via attributive masking learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9522–9537, 2024. 2

[21] Oren Barkan, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. Bee: Metric-adapted explanations via baseline exploration-exploitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1835–1843, 2025. 3, 6

[22] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer, 2016. 1

[23] Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16143–16152, 2023. 6, 25

[24] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10329–10338, 2022. 2

[25] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[26] 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)*, pages 839–847. IEEE, 2018. 2, 7, 12, 14

[27] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 1, 2, 6, 7, 15

[28] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 2, 6, 7, 12, 15

[29] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. In *International Conference on Learning Representations*, 2019. 2

[30] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6970–6979, 2017. 2, 13

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009. 6, 25

[32] 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)*, pages 972–980, 2020. 1, 7, 15

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[34] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023. 1

[35] Yehonatan Elisha, Oren Barkan, and Noam Koenigstein. Probabilistic path integration with mixture of baseline distributions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 570–580, 2024. 6

[36] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021. 3, 30

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 6

[38] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2950–2958, 2019. 2, 7, 15

[39] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 1, 2, 7, 15, 23

[40] Keren Gaiger, Oren Barkan, Shir Tsipory-Samuel, and Noam Koenigstein. Not all memories created equal: Dynamic user representations for collaborative filtering. *IEEE Access*, 11: 34746–34763, 2023. 2

[41] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018. 1

[42] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014. 6

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[44] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods, 2023. 6, 13, 14, 16

[45] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019. 6, 7

[46] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 6

[47] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020. 6

[48] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2, 6, 7, 15

[49] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019. 6, 12, 13

[50] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021. 2, 6, 7, 14

[51] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pages 280–298. Springer, 2022. 6

[52] D Liu, R Nicolescu, and R Klette. Stereo-based bokeh effects for photography. *Machine Vision and Applications*, pages 1–13, 2016. 13

[53] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017. 2

[54] Itzik Malkiel, Dvir Ginzburg, Oren Barkan, Avi Caciularu, Jonathan Weill, and Noam Koenigstein. Interpreting bert-based text similarity via activation and saliency maps. In *Proceedings of the ACM Web Conference 2022*, pages 3259–3268, 2022. 2

[55] 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, 2017. 1

[56] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s): 1–42, 2023. 7

[57] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 2, 3, 6, 7, 12, 15, 16, 24

[58] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020. 2

[59] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2, 7, 14

[60] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017. 1, 2

[61] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 25

[62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[63] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *NeurIPS*, 2019. 2, 7, 14

[64] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 3319–3328, 2017. 7, 14

[65] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017. 1, 2, 3, 30

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[67] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019. 2

[68] Yipei Wang and Xiaoqian Wang. Benchmarking deletion metrics with the principled explanations. In *Forty-first International Conference on Machine Learning*, 2024. 6

[69] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020. 6