ShapBPT: Image Feature Attributions Using Data-Aware Binary Partition Trees

Muhammad Rashid[1], **Elvio G. Amparore**[1], **Enrico Ferrari**[2], **Damiano Verda**[2]

[1]University of Torino, Computer Science Department, C.so Svizzera 185, 10149 Torino, Italy
[2]Rulex Innovation Labs, Via Felice Romani 9, 16122 Genova, Italy
{muhammad.rashid, elviogilberto.amparore}@unito.it, {enrico.ferrari, damiano.verda}@rulex.ai

Abstract

Pixel-level feature attributions are an important tool in eXplainable AI for Computer Vision (XCV), providing visual insights into how image features influence model predictions. The Owen formula for hierarchical Shapley values has been widely used to interpret machine learning (ML) models and their learned representations. However, existing hierarchical Shapley approaches do not exploit the multiscale structure of image data, leading to slow convergence and weak alignment with the actual morphological features. Moreover, no prior Shapley method has leveraged data-aware hierarchies for Computer Vision tasks, leaving a gap in model interpretability of structured visual data.

To address this, this paper introduces ShapBPT, a novel data-aware XCV method based on the hierarchical Shapley formula. ShapBPT assigns Shapley coefficients to a multiscale hierarchical structure tailored for images, the Binary Partition Tree (BPT). By using this data-aware hierarchical partitioning, ShapBPT ensures that feature attributions align with intrinsic image morphology, effectively prioritizing relevant regions while reducing computational overhead. This advancement connects hierarchical Shapley methods with image data, providing a more efficient and semantically meaningful approach to visual interpretability. Experimental results confirm ShapBPT's effectiveness, demonstrating superior alignment with image structures and improved efficiency over existing XCV methods, and a 20-subject user study confirming that ShapBPT explanations are preferred by humans.

Code — https://github.com/amparore/shap_bpt
Tests — https://github.com/rashidrao-pk/shap_bpt_tests
Tech. Appendix — https://zenodo.org/records/17570695

Introduction

A fundamental challenge in Machine Learning (ML) for Computer Vision is explaining how a black-box model classifies images, providing insights into the representations the model has learned from data. A key approach to this problem involves attributing importance scores to individual pixels, identifying their contribution to the model's decision-making process. This task, commonly referred to as *explaining model predictions*, plays a crucial role in enhancing interpretability and trust in AI-driven image classification. One of the most widely used methods for this purpose is SHAP (SHapley Additive exPlanations), which applies game-theoretic principles to ML explainability. SHAP combines feature removal (masking) (Lundberg and Lee 2017) with hierarchical image partitioning (Lundberg 2020), computing feature attributions over a refinable axis-aligned (AA) grid of pixels to approximate the regions most relevant to an image classifier. Another influential method is LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016), which, despite lacking theoretical guarantees, remains popular for its ability to pre-identify relevant image regions through segmentation. However, LIME and similar approaches rely on predefined segmentation matching the relevant image regions, and they cannot adaptively refine these regions if the initial segmentation is inadequate, limiting their effectiveness for complex image data.

Since models learn to recognize structured patterns from image data, an image classifier is expected to base its decisions on a hierarchical representation that captures distinct morphological characteristics—such as shape, texture, and color continuity—of the classified objects. A key challenge lies therefore in integrating theoretically sound attribution methods, such as Shapley coefficients, with data-aware image hierarchies. Computing Shapley coefficients over adaptive, data-driven hierarchical partitions can enhance interpretability by aligning attributions more closely with the model's learned representations. However, for this approach to be effective, the partitions must remain flexible and refinable, rather than being imposed a priori (as done by LIME or similar approaches).

This paper provides the following contributions:

1. A novel hierarchical model-agnostic XCV method for images, named *ShapBPT*, that integrates an adaptive multi-scale partitioning algorithm with the Owen approximation of the Shapley coefficients. We repurpose the BPT (Binary Partition Tree) algorithm (Salembier and Garrido 2000) to effectively construct hierarchical structures for explainability. This approach overcomes the limitations of the inflexible hierarchies of state-of-the-art methods such as SHAP.
2. An empirical assessment of the proposed method on natural color images showcasing its efficacy across various

scoring targets, in comparison to established state-of-the-art XCV methods, and a controlled human-subject study comparing explanation interpretability across methods.

We show that the proposed approach surpasses existing Shapley-based model-agnostic XCV methods that do not leverage on data-awareness, and at the same time it achieves a significantly faster convergence rate. This efficiency stems from the fact that, on average, fewer recursive applications of the Owen formula (i.e. expansions of the partition hierarchy) are needed to accurately localize objects when using a *data-aware* partition hierarchy, such as the proposed BPT hierarchy, compared to other hierarchies. As far as we know, this is the first XCV method that combines the Owen formula with a data-aware partition hierarchy for image data, and with this paper we prove the effectiveness of this combined strategy for interpreting ML classifiers.

Methodology

A fundamental ML objective is to discover a function $f : \mathcal{X} \to \mathcal{Y}$ that effectively approximates a response $y \in \mathcal{Y}$ corresponding to a given input $x \in \mathcal{X}$. For the sake of simplicity, we assume $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^n$. In many practical cases, only some components of x significantly influence the response $y = f(x)$. Understanding the relative importance, or *contribution*, of each component x_i of x in determining the value of y by f is a central problem in XCV. An important approach (Covert, Lundberg, and Lee 2021) for assessing these contributions is through *feature removal* (also called *masking*), where certain values of x are replaced with values from a specified context-dependent background set. Let $\nu_{f,x} : 2^{|\mathcal{X}|} \to \mathcal{Y}$ be a *masking function* for $f(x)$, where $\nu_{f,x}(S)$ represents the evaluation of the resulting model when only the elements in the subset S of x are retained, while the others are masked. In the following, we will denote $\nu_{f,x}$ as ν.

Shapley values. We consider the setup of a n-coalition game $(\mathcal{N}, \nu)$, which is analogous to an importance scores attribution task in XCV (Rozemberczki et al. 2022). The finite set $\mathcal{N} = \{1, \ldots, n\}$ is the set of players (*features*). Each nonempty subset $S \subseteq \mathcal{N}$ is a *coalition*, and $\mathcal{N}$ is itself the *grand coalition*. A *characteristic function* $\nu : 2^n \to \mathbb{R}$ assigns to each coalition S a (real) *worth value* $\nu(S)$, and it is assumed that $\nu(\varnothing) = 0$ (it is always possible to ensure $\nu(\varnothing) = 0$ by translation of the equation system). A *marginal contribution* of a player i to a coalition S (assuming $i \notin S$) is given by

$$\Delta_i(S) = \nu(S \cup \{i\}) - \nu(S) \tag{1}$$

Semivalues (Dubey, Neyman, and Weber 1981), weighted sums of marginal contributions (1), were introduced as a method for fairly distributing the total value $\nu(\mathcal{N})$ of the grand coalition $\mathcal{N}$ among its members. The Shapley value (Shapley 1953), a well-known semivalue, demonstrates favorable axiomatic properties and has been used effectively to explain ML models (Rozemberczki et al. 2022).

Hierarchical coalition structures (HCS). A fixed a-priori *coalition structure* (López and Saboya 2009; Owen 2013, 1977) for the $\mathcal{N}$ players is a finite set $\{T_1, \ldots, T_m\}$ of m partitions of $\mathcal{N}$ (i.e. $\cup_{k=1}^{m} T_k = \mathcal{N}$, and $T_i \cap T_j \neq \varnothing \Leftrightarrow i = j$). Elements T_i are usually called *partitions*, *coalitions*, *teams* or *unions*.

We consider a recursive definition of a hierarchical coalition structure, where each partition T can be either an *indivisible partition* or a *sub-coalition structure* itself $T = T_1 \cup \ldots \cup T_m$. Let $T{\downarrow}$ be the (downward) recursive partitioning of T, defined as

$$T{\downarrow} = \begin{cases} \{T_1, \ldots, T_m\} & \text{if } T \text{ admits sub-coalitions} \\ \bot & \text{if } T \text{ is indivisible} \end{cases} \tag{2}$$

We denote with $\mathcal{T}$ the HCS root, and assume w.l.o.g. that $\mathcal{T}$ contains all the elements of $\mathcal{N}$.

A special case of HCS happens when each sub-coalition structure is made by two partitions, i.e. the hierarchy forms a binary tree. We refer to these structures as *binary hierarchical coalition structures* (BHCS). In that case the recursive downward partitioning of T can be simplified as

$$T{\downarrow} = \begin{cases} \{T_1, T_2\} & \text{if } T \text{ admits a binary sub-coalition} \\ \bot & \text{if } T \text{ is indivisible} \end{cases} \tag{3}$$

The Owen approximation for Binary HCS. Computing exact Shapley values is at least #P-hard (Deng and Papadimitriou 1994), which is unfeasible for image data with hundreds or thousands of features (pixels). An approximate approach, introduced by (Owen 1977), can be used to drastically reduce the cost by grouping features into hierarchical coalitions. This concept has been pioneered for images by the SHAP Partition Explainer (Lundberg 2020; Shrikumar, Greenside, and Kundaje 2017; Lundberg and Lee 2017).

A *coalition value* $\Omega_i(\mathcal{T})$ represents the worth of the player i in a game with coalition structure $\mathcal{T}$, and is known as the Owen coalition value (Owen 1977). Computing coalition values over a binary HCS T as defined in (3) can be done by recursively composing a coalition Q using the formula

$$\Omega_i(Q, T) = \begin{cases} \frac{1}{2}\Omega_i(Q \cup T_2,\, T_1) \,+\, \frac{1}{2}\Omega_i(Q,\, T_1) & \\ \qquad\qquad \text{if } T{\downarrow} = \{T_1,\, T_2\} & \\ \frac{1}{|T|}\Delta_T(Q) \qquad \text{if } T \text{ is indivisible} & \end{cases} \tag{4}$$

with $\Omega_i(\mathcal{T}) = \Omega_i(\varnothing, \mathcal{T})$. The former case of Eq. (4) deals with coalitions T that admit a sub-coalition structure $T{\downarrow} \neq \bot$. We assume, for notational simplicity and without loss of generality, that $i \in T_1$. The latter case of Eq. (4) deals with indivisible coalitions. In that case, the formula computes a marginal contribution (uniformly divided) of all players of T w.r.t. the coalition Q formed recursively.

In the rest of the paper, we will refer to the Owen approximation of the Shapley values simply as the Shapley values. Note that Eq. (4) is not found in published literature (as far as we know), and its complete derivation is therefore provided in the Technical Appendix.

Theorem 1. ***Computational cost.*** *Consider a BHCS consisting of a balanced tree of depth d. The time complexity of Eq.* (4) *is in the order of $O(4^d)$ evaluations of the ν function.*

Proof. Derivation is in Technical Appendix. □

Theorem 1 highlights the exponential cost of Eq. (4). However, practical implementation of Eq. (4) do not rely on expanding a fully balanced BHCS tree to a fixed depth d. Instead, they employ an adaptive splitting strategy that is not limited to balanced trees. In this adaptive case, a total budget b of evaluations of the masked model ν is allocated. The adaptive algorithm then iteratively explores the tree hierarchy, at each iteration splitting the partition T that maximizes the sum of its Shapley values, $\sum_{i \in T} \Omega_i(\varnothing, \mathcal{T})$. Each partition split requires 2 model evaluations. A pseudo-code of this adaptive algorithm is provided in the Technical Appendix. Despite adaptively ignoring certain coalitions, the cost of exploring the hierarchy at depth d remains exponential, as stated in Theorem 1.

Hierarchical Coalition Structures for Images

Calculating Owen coalition values for image data necessitates a well-defined hierarchical structure that captures both spatial relationships and image semantics. Our approach is aimed at addressing limitations in existing methods, by emphasizing the importance of these factors in coalition formation. We therefore consider and compare both *data-agnostic* and *data-aware* approaches.

In a *data-agnostic* approach, partitions are created based on simple geometric divisions, like grids or quadrants. The *Axis Aligned grid hierarchy* (AA hereafter) is one such approach to building hierarchical coalition structures, adopted by the SHAP's Partition Explainer (Lundberg 2020) and by h-SHAP (Teneggi, Luster, and Sulam 2022). In an AA hierarchy, each partition T corresponds to a rectangular region within the image, and $T\downarrow$ splits the rectangular region of T in half along the longest axis. This splitting process continues until indivisible (unitary) regions (i.e. single pixels) are reached, or an evaluation budget b is consumed. The main limitation of this approach is that properly localizing the relevant regions within an image may require a large number of recursive evaluation of the Owen's formula (4), and this evaluation follows the $O(4^d)$ time cost of Theorem 1.

In a basic *data-aware* approach, morphological features within the image guide the partitioning process. This approach, pioneered by (Ribeiro, Singh, and Guestrin 2016) with LIME, utilizes a pre-defined segmentation algorithm to divide the image into regions (patches). Although effective, the main limitation is the lack of an effective feedback loop within the explanation method. If the segmentation is inaccurate, the resulting explanation is poor, and there is no opportunity for refinement.

A notable algorithm for hierarchical segmentation, that fits well with Eq. (4), is the *Binary Partition Tree* (BPT) (Randrianasoa et al. 2018), originally developed for multiscale image representation in MPEG-7 encoding (Salembier and Garrido 2000). The intuitive principle is that portions of an image with similar color and coherent shape are highly likely to have similar Shapley values, thereby maximizing the effectiveness of Eq. (4).

Theorem 1 shows that the Owen approximation cost increases rapidly if a large number of coalitions need to be evaluated recursively. Therefore, an effective BHCS must satisfy these requirements:

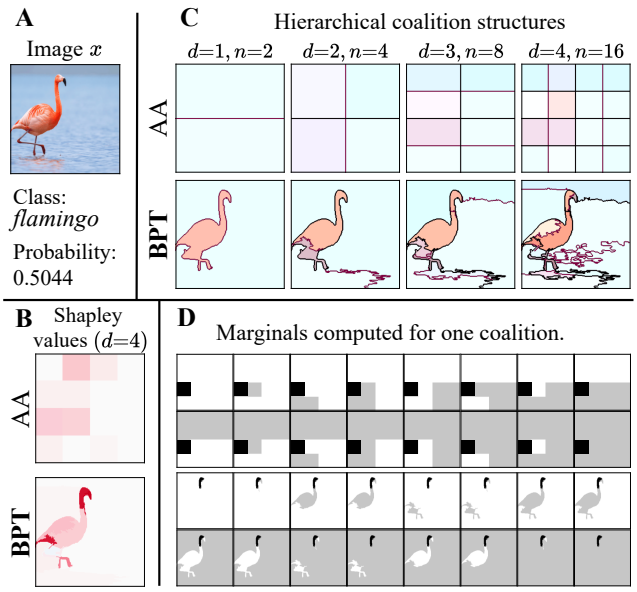

Figure 1: AA and BPT coalition structures for a sample image classification using a ResNet50 model.

R1 As few recursive cuts as possible to reach the relevant regions, as each cut increases the required evaluation budget b exponentially;

R2 Partitions should not be fixed, since the relevant regions are not known in advance.

AA hierarchies do not satisfy R1, and most a-priori segmentation algorithms do not satisfy R2. The solution that we propose, which constitutes the main contribution of this paper, is a novel hybrid method that finally satisfies the two aforementioned requirements by combining a refinable a-priori hierarchical coalition structure (the BPT) aligned with the morphological features of the image (e.g., color uniformity, pixel locality) together with an a-posteriori splitting strategy based on the distribution of Shapley values (as in the Partition Explainer). This combination results in significantly fewer recursive applications of the Owen formula needed to accurately localize objects, compared to data-agnostic coalition structures. As we shall see in the experimental section, this approach usually gets a faster convergence than other Shapley-based methods, paired with accurate shape recognition of the classified objects.

Example 1. *Figure 1 presents a sample image (A) with its Shapley explanations (B), computed using Eq.* (4) *on AA and BPT hierarchical coalition structures (C) up to depth* $d = 4$. *The first four tree hierarchy levels in (C) highlight the data-aware nature of BPT. Each coalition value is derived from a weighted sum of eight marginals* $\widehat{\varphi}_i(Q, T)$, *with the highest-value marginals shown in (D), where* Q *and* T *correspond to the grey and black regions.*

Generating BPT hierarchies. A *BPT hierarchy* captures how we can progressively merge (Randrianasoa et al. 2018)

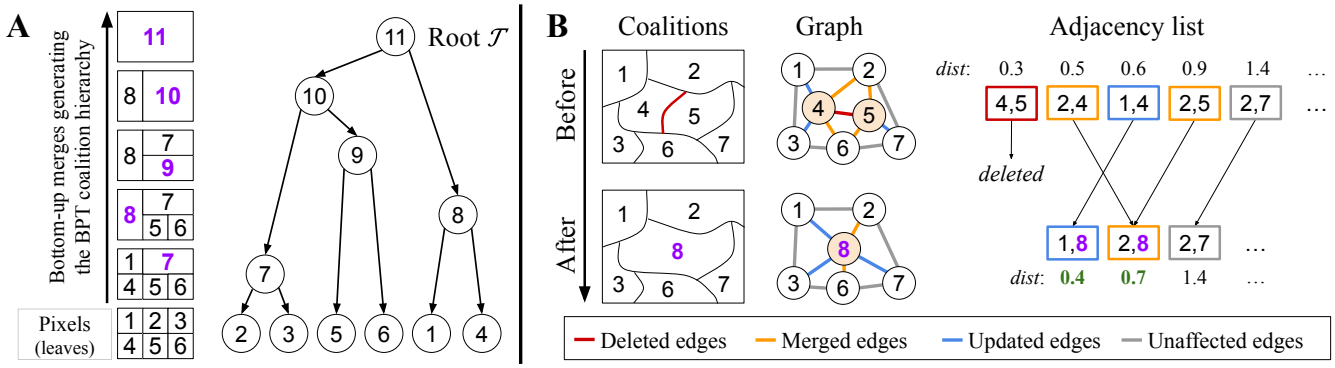

Figure 2: **(A)** BPT generating by bottom-up merging coalitions from the pixels (1–6) to the root (11). **(B)** Details of one merging step $T_8{\downarrow} = \{T_4, T_5\}$ on some arbitrary coalition structure.

the n pixels of an image x into larger regions, forming a quasi-balanced binary tree. Tree construction is bottom-up, starting from an initial coalition structure $\mathcal{T}_{[1]} = \{T_1{=}\{1\}, \ldots, T_n{=}\{n\}\}$ made by n unitary and indivisible partitions, where the features $1, \ldots, n$ represents the individual pixels of the image. Two partitions $T_i, T_j \in \mathcal{T}_{[k]}$ are *adjacent* if there is at least one pixel of T_i that is adjacent to a pixel of T_j in the image. The BPT construction involves merging adjacent partitions iteratively.

A *coalition merge* of $\mathcal{T}_{[k]}$ is a new coalition structure $\mathcal{T}_{[k+1]}$ where two adjacent partitions $T_i, T_j \in \mathcal{T}_{[k]}$ are removed and replaced by a new partition T_{n+k}, s.t. $T_{n+k} = T_i \cup T_j$ and $T_{n+k}{\downarrow} = \{T_i, T_j\}$.

The two adjacent partitions T_i, T_j of $\mathcal{T}_{[k]}$ to be merged are selected by minimizing a *data-aware* distance function. Prior work (Randrianasoa et al. 2018, 2021) on BPTs shows that color range $\times$ perimeter scores correlate with perceptual region uniformity, and area helps in keeping the tree hierarchy balanced. With this knowledge we define

$$dist(T_i, T_j) = clr^2(T_i, T_j) \cdot area(T_i, T_j) \cdot \sqrt{pr(T_i, T_j)} \quad (5)$$

as a distance criteria, where $clr^2(T_i, T_j)$ is the sum of the squared color ranges of $T_i \cup T_j$, for all color channels, and $area(T_i, T_j)$ and $pr(T_i, T_j)$ are the area and the perimeter of $T_i \cup T_j$, respectively. A sensitivity ablation analysis that supports the rationale of Eq. (5) is in the Technical Appendix.

A *merging sequence* $\mathcal{T}_{[1]} \rightarrow \mathcal{T}_{[2]} \rightarrow \ldots \rightarrow \mathcal{T}_{[n]}$ is a sequence of $n-1$ coalition merges. The sequence ends with the coalition structure $\mathcal{T}_{[n]} = \{T_{2n-1}\}$, having a single partition with all pixels. At this point, all non-unitary partitions T at any point in the merging sequence admit a binary sub-coalition structure $T{\downarrow}$. Therefore, the BPT $\mathcal{T}_{[n]}$ satisfies Eq. 3, and may become the root $\mathcal{T}$ of the BHCS. An illustration of the algorithm generating the BPT merging sequence is shown in Figure 2/A, where the unitary partitions are merged, one by one, until all pixels are merged into the root $\mathcal{T}$. The operations needed to perform a single merging step are illustrated in Figure 2/B, and a detailed pseudo-code of the BPT algorithm is provided in the Technical Appendix.

Experimental Assessment

We present a comparative analysis of the performance of the proposed Shapley method using BPT partitions, alongside other state-of-the-art image explainers.

Comparison scores. To ensure a robust and comprehensive quantitative evaluation, we consider two score categories: *response-based* and *ground-truth-based*. The *response-based* score that we consider are the *area-under-curve* (AUC) from (Petsiuk, Das, and Saenko 2018), which measure how well the ranked explanation coefficients align with the black-box model's output. These scores do not rely on any predefined notion of "correct" explanation and instead evaluate the internal consistency of the explanation with respect to the model's own behavior. Let $S^{[q]} \subseteq \mathcal{N}$ be the subset of the first q-th quantile of elements from $\mathcal{N}$ with the largest Shapley values. Define

$$AUC^+ = \int_0^1 \nu(S^{[q]})\, \mathrm{d}q, \quad AUC^- = \int_0^1 \nu(\mathcal{N} \setminus S^{[q]})\, \mathrm{d}q \quad (6)$$

With this definition AUC^+ (resp. AUC^-) evaluate the model's behavior as features are progressively included from an empty set (resp. excluded from the full set). Since we deal with regression models, we rescale (Hama, Mase, and Owen 2023) all ν values in the $[0, 1]$ range, s.t. all evaluated samples weight uniformly.

The *ground-truth-based* score we consider is the Intersection over Union (IoU) score, which compares the predicted important features with a known *ground truth* subset $G \subseteq \mathcal{N}$. Ideally G is a set for which $\nu(G) = \nu(\mathcal{N})$. This setup is relevant in the context of the *Visual Recognition Challenge* (VRC) (Russakovsky et al. 2015), where annotations provide an external reference for which image regions are expected to contribute to classification. An explanation is a *perfect match* if there is a threshold q for which $S^{[q]} = G$. Consider the standard *Intersection-over-Union* score $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ and define

$$\begin{aligned} AU\text{-}IoU &= \int_0^1 J(S^{[q]}, G)\, \mathrm{d}q \\ max\text{-}IoU &= \max_{q \in [0,1]} J(S^{[q]}, G) \end{aligned} \quad (7)$$

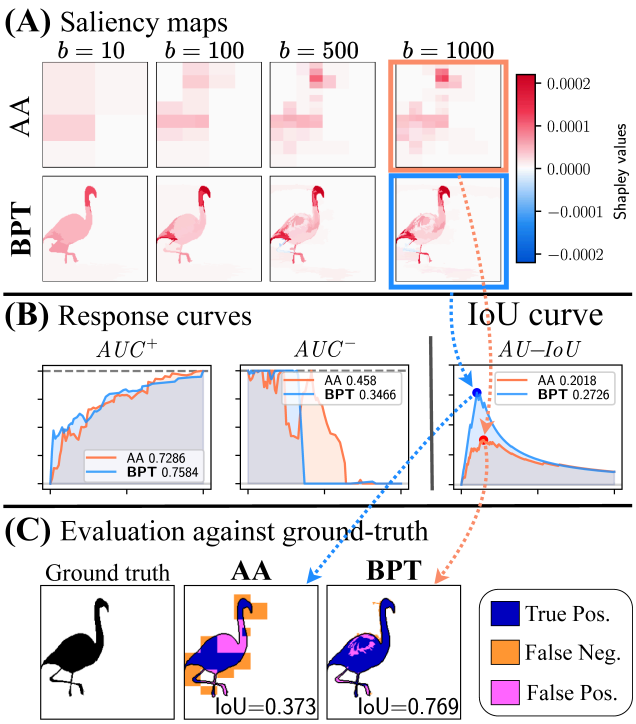

Figure 3: Shapley values for AA and BPT coalition structures, for different values of the budget b.

The score $AU\text{-}IoU$ (Gangopadhyay et al. 2023) is the area under the IoU curve, defined by the IoU values in the range $q \in [0, 1]$, and $max\text{-}IoU$ is the curve maximum. The $AU\text{-}IoU$ is maximal when the explanation perfectly matches the ground truth mask, and in such case $max\text{-}IoU = 1$.

Example 2. *Figure 3 shows the Shapley values computed using Eq.* (4) *on the AA and BPT coalition structures, by refining the most significant coalition using a budget b of model evaluations* ***(A)****, with b equal to* 10, 100, 500 *and* 1000 *samples, respectively. The area identified by the threshold q obtaining the maximal IoU is depicted in* ***(C)****. The plots* ***(B)*** *depict the response curves for the AUC scores* (6) *and* (7)*, for the case b=1000. In the example, BPT demonstrates its improved object region recognition w.r.t. AA.*

Compared methods. We run a comparative analysis using several state-of-the-art XCV methods, categorized into two groups. The first group comprises Shapley-based methods, chosen for their compatibility with our proposed approach. They include: **BPT**-b: our proposed Shapley explanation method with BPT partitions, with sample budgets b of 100, 500, and 1000 samples; **AA**-b: the SHAP Partition Explainer (Lundberg 2020), utilizing Axis-Aligned partitions with b of 100, 500, and 1000 samples; **LIME**-b: LIME[1] explanation (Ribeiro, Singh, and Guestrin 2016) with budget b and with $b/5$ segments, with b being 100, 500, and 1000.

[1]Although LIME does not generate Shapley values, it has theoretical and practical similarities to them (Lundberg and Lee 2017).

	Dataset	Size	Model	Short description
E1	ImageNet-S_{50}	574	ResNet50	Common ImageNet setup
E2	ImageNet-S_{50}	574	Ideal	Linear ideal model
E3	ImageNet-S_{50}	621	SwinViT	Vision Transformer
E4	MS-COCO	274	Yolo11s	Object detection
E5	CelebA	400	CNN	Facial attrib. localization
E6	MVTec	280	VAE-GAN	Anomaly Detection
E7	ImageNet-S_{50}	593	ViT-Base16	Vision Transformer
E8	User preference study using E1 saliency maps.			

Table 1: Summary of the experiments.

The second group consists of gradient-based methods, included in our analysis due to their widespread usage. They include: **GradExpl**: the Gradient Explainer from the SHAP package (Lundberg and Lee 2017), using the default of 20 samples; **GradCAM**: the Gradient-weighted Class Activation Mapping introduced by (Gildenblat and contributors 2021); **IDG**: the Integrated Decision Gradient method of (Walker et al. 2024); **LRP**: Layer-wise Relevance Propagation of (Bach et al. 2015; Ancona et al. 2018) from Captum; **GradShap**: gradient Shap (Sundararajan, Taly, and Yan 2017). For *GradExpl* and *IDG*, we utilize the absolute values of the produced explanations, resulting in superior scores compared to the signed values.

Experiments. ShapBPT has been tested extensively over multiple computer vision tasks, models and datasets. Table 1 reports a summary of the experiments. Figure 4 depicts examples of the generated saliency maps for the first seven experiments, which helps to get a first intuition of the characteristics of the BPT method. Each row reports the image, the ground truth G, and the saliency maps. We show all the fourteen tested method only for **E1**. The boundaries of G are drawn overlapped to the saliency maps. To illustrate the evaluation process, for the first image, we also report the optimal IoU w.r.t. G. While all the tested methods seem to somewhat agree on the recognition area, the practical behavior of BPT seems in line with its theoretical assumption that splitting the image partitions following the morphological image boundaries leads to better object recognition, and better separation from the background.

Experiments are briefly detailed in the following. Further details, examples and results are in the Technical Appendix.

Experiment **E1** uses the *1K-V2* pretrained (Vryniotis 2021) ResNet50 (He et al. 2016) model (from PyTorch, accuracy 80.858%) over the ImageNet-S_{50} dataset (Gao et al. 2022) with ground truth masks (574 images in total). Replacement value is uniform gray. In general BPT explanations (columns 3–5) show a better tendency of identifying the partition borders, cutting the recognized object from the background. In that sense, they share similarities with the explanations of LIME, but without the typical LIME noise, and without relying on a fixed, inflexible segmentation. Moreover BPT explanations look a lot more in accordance to those of GradCAM, but without the blurriness that the latter adds.

Experiment **E2** evaluates an ideal linear model which perfectly follows the ground truth. Let $\nu_{\text{lin}}(S) = \frac{|S \cap G|}{|G|}$ be an ideal linear predictor that outputs the proportion of pixels of

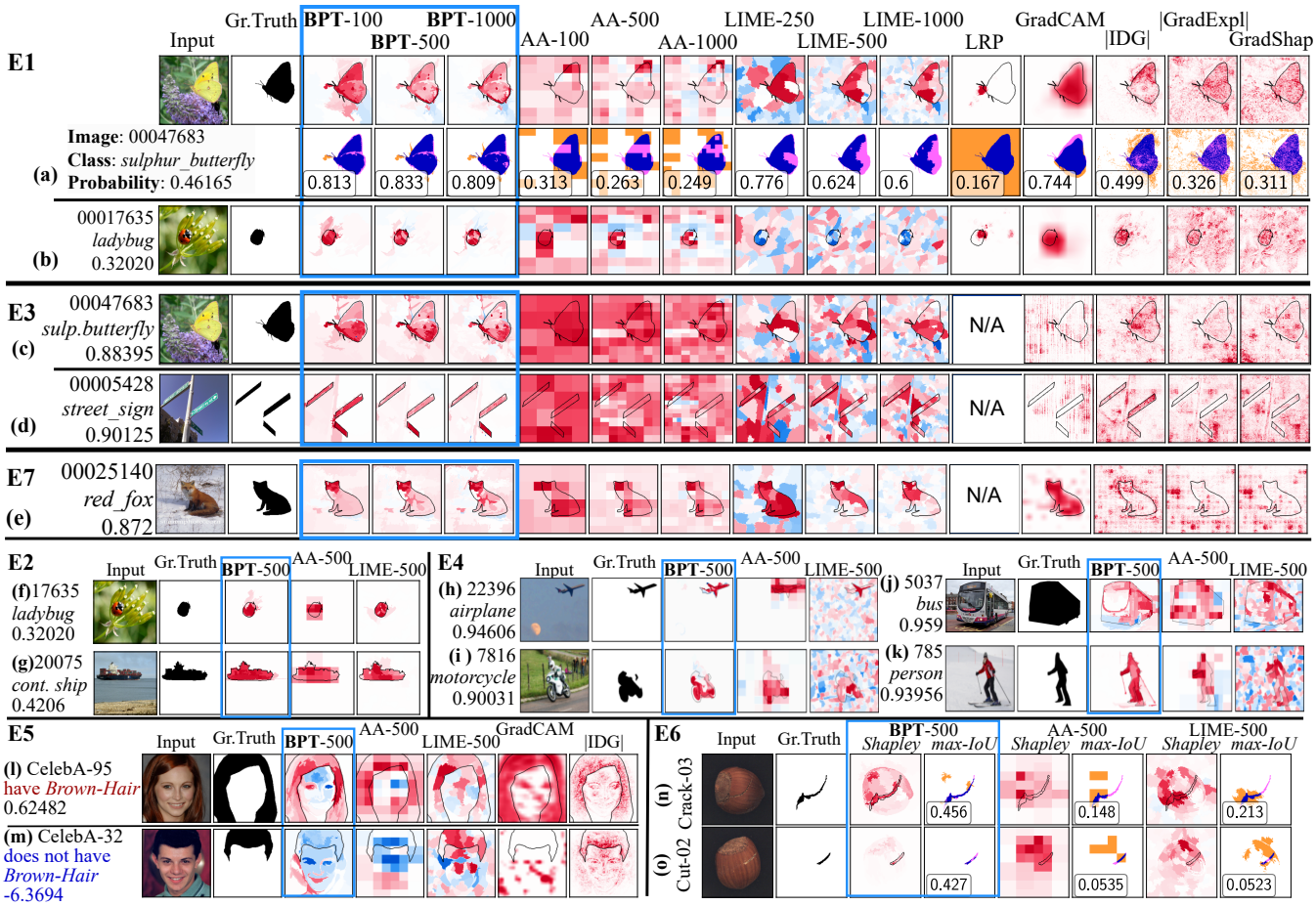

Figure 4: Selected saliency maps from experiments **E1**–**E7** (summarized in Table 1) for various computer vision ML tasks.

S that belong to the ground truth G. Since ν_{lin} is not a neural network, CAM methods cannot be used and are excluded.

Experiment **E3** uses the pretrained Vision Transformer model *SwinViT* (Liu et al. 2021) from pytorch (acc. 81.4%). It is interesting to see that all methods except BPT produce significantly more confused saliency maps, attributing a lot of importance to background features and with little focus to the actual classified objects. On the contrary, saliency maps obtained by the BPT method are clear and focused.

Experiment **E4** uses the Ultralytics Yolo11s model (Jocher and Qiu 2024) pre-trained for the MS-COCO dataset (Lin et al. 2014) which has diverse image sizes and a wider range of details than ImageNet. The XCV task involves highlighting detected objects.

Experiment **E5** uses a pre-trained CNN model (Batra 2020) to predict the presence or absence and the localization of facial features like *brown hair* and *eyeglasses* on the CelebA-HQ dataset (Karras et al. 2018). The XCV task focuses on localizing regions positively (red) or negatively (blue) influencing predictions. For this kind of tasks, Shapley values correctly distinguish positive and negative contributions, unlike CAM methods.

Experiment **E6** focuses on explaining an Anomaly Detection (AD) system. It builds on the (Ravi et al. 2021) methodology, and uses a convolutional VAE-GAN model for anomaly localization. The MVTec dataset (Bergmann et al. 2019) (*hazelnut* category) is used, with 280 high-quality images of defective (with ground truth masks) and non-defective objects. The anomaly map captures reconstruction errors, reflecting both anomalies and noise, and the XCV tasks consists in separating true anomalies from noise.

Experiment **E7** Similar to E1 using the ViT-Base 16 model (Dosovitskiy et al. 2020).

Numerical results. Figure 5 summarizes the results from all experiments, with a separate table for each of the four scores, plus one for the wall-clock evaluation time[2]. To ensure fairness across experiments, scores have been standardized accordingly. A red line indicates the overall mean across all experiments. Methods are ordered based on their mean scores from top (better) to bottom (worse). To assess statistical significance, we conducted one-way ANOVA tests for each score, testing the null hypothesis (H_0) of equal means across all sample populations, with p-value signifi-

[2]Using: Intel Core i9 CPU, Nvidia 4070 GPU, 16GB RAM.

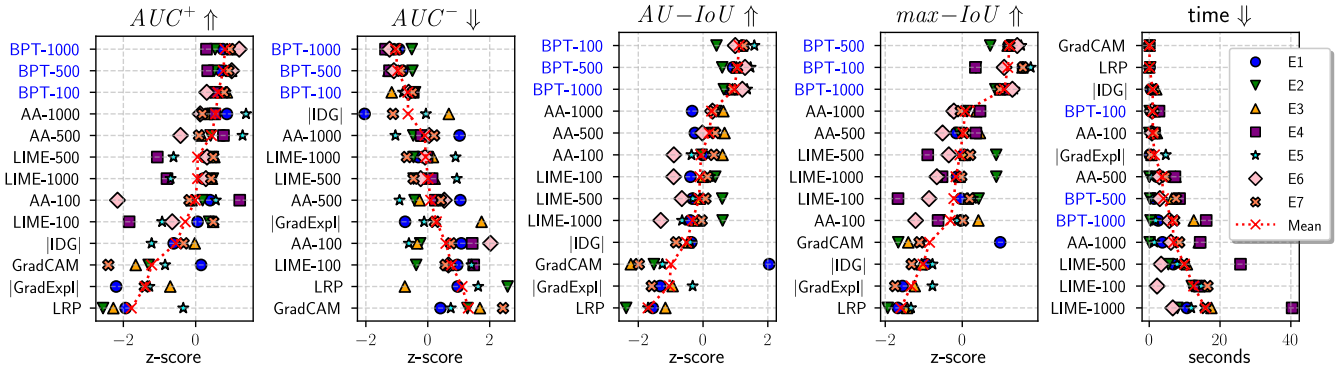

Figure 5: Results for all scores across the **E1**–**E7** experiments, with methods (on Y axis) ranked by performance (top to bottom).

cance threshold of 0.05. In all cases, the null hypothesis was rejected, indicating that the results are statistically significant. Results are reported in the Technical Appendix.

From Figure 5, we observe that BPT consistently outperforms AA and other compared methods across various models, scoring methods, and datasets. This supports the intuition that tailoring partitions to the characteristics of the data is beneficial. Moreover, BPT maintains its advantage even under resource constraints, as BPT-100 already surpasses most competing methods. This highlights its adaptability to different experimental setups and its robust generalization ability across datasets and models. In particular, ShapBPT seems to work well also with Vision Transformers (**E3** and **E7**), which are known for their robustness to partial object occlusion (Englebert et al. 2023).

E8 - User preference study. In addition to the automated metrics, we measured *perceived usefulness* with a small controlled user study with 20 participants. Each subject viewed four randomly selected images from **E1** and ranked the four explanation maps (BPT-1000, GradCAM, LIME-1000, AA-1000) from most to least helpful for understanding the model's prediction, yielding 80 rank-lists per method. A Friedman test determined the significance ($\chi^2_{(k=3)}$=19.56, p-value=0.0002, H_0 rejected). BPT emerged as the winner, ranked first in 51% of the cases with an average rank of 1.79, trailed by GradCAM (33% first-place, mean 2.41) and LIME (10%, mean 2.56), while AA was seldom preferred (6%, mean 3.24). Human preference seems therefore to partially confirm the quantitative metrics of the **E1**–**E7** experiments. Full details are in the Technical Appendix.

Discussion and Other Related Works

Our evaluation combines both *ground-truth-based* metrics (IoU) and *response-based* metrics (AUC) to provide a comprehensive and reliable assessment of the methods. IoU scores are included because they are the standard evaluation metric in object detection benchmarks (Rezatofighi et al. 2019). While deep learning models may show misalignments between the ground truth G and the model's learned representation, this should not introduce bias in the AU-IoU and max-IoU scores, as all methods are evaluated under the same conditions. Moreover, experiment **E2** is fully unbiased, since the ideal model ν_{lin} is a linear model.

A convergence analysis comparing BPT and AA across varying evaluation budgets is in the Technical Appendix.

For LIME, we generated fixed a priori partitions using the *quickshift* algorithm and also tested the more recent *SegmentAnything* (SAM) method, which improves upon *quickshift* but is significantly slower. However, neither of these methods can build the hierarchy of Shapley values adaptively. The limitation of relying on rigid, pre-defined partitions persists, an issue that is addressed by the proposed BPT approach (as outlined in requirement R2). It would be interesting to integrate SAM directly into ShapBPT and compare it against BPT. However SAM does not generate a regular HCS (Knab, Marton, and Bartelt 2025), which is a key requirement of the Owen formula. Constructing a SAM-compatible HCS therefore demands new algorithmic machinery, beyond the scope of the present study, and merits a dedicated investigation as future work. We outline the key details in the Technical Appendix. A discussion on h-Shap limits is in Technical Appendix.

We considered the *relevance mass and rank accuracy* scores (Arras, Osman, and Samek 2022) but eventually excluded them, as their reliance on non-negative values does not work well with Shapley values.

User preference results echo the objective ones: a 20-participant study confirmed that the data-aware BPT hierarchy yields explanations humans actually find most useful.

Conclusions

This paper introduces *ShapBPT*, a model-agnostic explainability method for AI classifiers in computer vision. It computes saliency maps by calculating Shapley values using the Owen formula over a data-aware *Binary Partition Tree* (BPT) of the image being explained. That captures the importance of image features in a way that is both efficient and consistent with Shapley's axiomatic properties.

Comprehensive cross-dataset benchmarks and a 20-subject preference study consistently place ShapBPT's data-aware hierarchical partitions ahead of existing XCV explainers, confirming it as a novel, robust method that delivers accurate, budget-efficient, and human-preferred explanations.

Acknowledgments

This work has received funding from the European Union's Horizon research and innovation program Chips JU under Grant Agreement No. 101139769, DistriMuSe project (HORIZON-KDT-JU-2023-2-RIA). The JU receives support from the European Union's Horizon research and innovation programme and the nations involved in the mentioned projects. The work reflects only the authors' views; the European Commission is not responsible for any use that may be made of the information it contains.

References

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations (ICLR)*.

Arras, L.; Osman, A.; and Samek, W. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81: 14–40.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140.

Batra, K. 2020. MultiLabel Classification of CelebA. https://www.kaggle.com/code/kartikbatra/multilabelclassification/output. Accessed on 2025-Nov-28.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Covert, I.; Lundberg, S.; and Lee, S.-I. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209): 1–90.

Deng, X.; and Papadimitriou, C. H. 1994. On the complexity of cooperative solution concepts. *Mathematics of operations research*, 19(2): 257–266.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Dubey, P.; Neyman, A.; and Weber, R. J. 1981. Value theory without efficiency. *Mathematics of Operations Research*, 6(1): 122–128.

Englebert, A.; Stassin, S.; Nanfack, G.; Mahmoudi, S. A.; Siebert, X.; Cornu, O.; and De Vleeschouwer, C. 2023. Explaining Through Transformer Input Sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 806–815.

Gangopadhyay, T.; Hong, S.; Roy, S.; Shah, Y.; and Cheong, L. L. 2023. Benchmarking framework for anomaly localization: Towards real-world deployment of automated visual inspection. *Journal of Manufacturing Systems*, 69: 64–75.

Gao, S.; Li, Z.-Y.; Yang, M.-H.; Cheng, M.-M.; Han, J.; and Torr, P. 2022. Large-scale Unsupervised Semantic Segmentation. *TPAMI*.

Gildenblat, J.; and contributors. 2021. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam. Accessed on 2025-Nov-28.

Hama, N.; Mase, M.; and Owen, A. B. 2023. Deletion and insertion tests in regression models. *Journal of Machine Learning Research*, 24(290): 1–38.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jocher, G.; and Qiu, J. 2024. Ultralytics YOLO11. Accessed on 2025-Nov-28.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of International Conference on Learning Representations (ICLR) 2018*.

Knab, P.; Marton, S.; and Bartelt, C. 2025. Beyond Pixels: Enhancing LIME with Hierarchical Features and Segmentation Foundation Models. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *In procs. of 13th European Conf. Computer Vision (ECCV) 2014*, 740–755. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

López, S.; and Saboya, M. 2009. On the relationship between Shapley and Owen values. *Central European Journal of Operations Research*, 17: 415–423.

Lundberg, S. 2020. The SHAP Partition Explainer. https://shap.readthedocs.io/en/latest/generated/shap.PartitionExplainer.html. Accessed on 2025-Nov-28.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774.

Owen, G. 1977. Values of games with a priori unions. In *Mathematical economics and game theory: Essays in honor of Oskar Morgenstern*, 76–88. Springer.

Owen, G. 2013. *Game theory, 4th Ed.* Emerald Group Publishing.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC) 2018*, 151. BMVA Press.

Randrianasoa, J. F.; Kurtz, C.; Desjardin, E.; and Passat, N. 2018. Binary partition tree construction from multiple features for image segmentation. *Pattern Recognition*, 84: 237–250.

Randrianasoa, J. F.; Kurtz, C.; Desjardin, E.; and Passat, N. 2021. AGAT: Building and evaluating binary partition trees for image segmentation. *SoftwareX*, 16: 100855.

Ravi, A.; Yu, X.; Santelices, I.; Karray, F.; and Fidan, B. 2021. General frameworks for anomaly detection explainability: comparative study. In *2021 IEEE International Conference on Autonomous Systems (ICAS)*, 1–5. IEEE.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. ACM SIGKDD Int. Conf., 22nd*, 1135–1144.

Rozemberczki, B.; Watson, L.; Bayer, P.; Yang, H.-T.; Kiss, O.; Nilsson, S.; and Sarkar, R. 2022. The Shapley Value in Machine Learning. In *IJCAI-22*, 5572–5579.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. of Computer Vision (IJCV)*, 115(3): 211–252.

Salembier, P.; and Garrido, L. 2000. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Trans. on Image Processing*, 9(4): 561–576.

Shapley, L. S. 1953. A value for n-person games. *The Shapley value. Essays in honor of Lloyd S. Shapley*, 31.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Teneggi, J.; Luster, A.; and Sulam, J. 2022. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4494–4503.

Vryniotis, V. 2021. How to train state-of-the-art models using torchvision's latest primitives. https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives/. Accessed on 2025-Nov-28.

Walker, C.; Jha, S.; Chen, K.; and Ewetz, R. 2024. Integrated Decision Gradients: Compute Your Attributions Where the Model Makes Its Decision. *AAAI*, 38(6): 5289–5297.