Unpacking "Baby": XAI-based Intra-Class Hierarchies for Understanding Model Biases

Rodrigo Andrade Santos[1,2]
rodrigo_andrade@petrobras.com.br
Jefersson A. dos Santos[3]
j.santos@sheffield.ac.uk
Fabricio Murai[4]
fmurai@wpi.edu

[1] Universidade Federal de Minas Gerais
[2] Petróleo Brasileiro S.A.
[3] University of Sheffield
School of Computer Science
[4] Worcester Polytechnic Institute
Worcester, MA, USA

Abstract

With the widespread use of Deep Learning models in computer vision that often surpass human capabilities, many recent works have focused on eXplainable AI (XAI) to understand these black-box models and their failure modes. Over the years, this field has made large strides, proposing new feature attribution methods for explaining sample predictions, metrics for comparing these methods, and frameworks for jointly applying XAI tools. As most of these efforts try to understand *where* the network is looking at, we still don't fully understand *what* image elements consistently tend to result in stronger explanation signals. To address this gap, we propose XICH (XAI-based Intra-Class Hierarchies), a new methodology for unveiling hierarchies between clusters of same-class images based on XAI. In a nutshell, XICH computes a new metric called "dominance", which represents a statistical comparison of the explanation energy between clusters by laying out their representative samples as a mosaic. We investigate the dominance between clusters by mapping instances from the training data to clusters extracted from the evaluation data. We observe that low-dominance clusters tend to have much less corresponding instances in the training set, revealing out-of-distribution biases. The proposed metric is highly interpretable and complements existing intra-class sample hierarchization methods (e.g. based on logits or softmax probabilities), as our results shows that they are non-correlated.

1 Introduction

The recent success of deep learning models for computer vision in critical tasks such as medical image analysis, remote sensing, and autonomous driving vehicles has increased the need for transparency and accountability of model decision-making processes [14, 28, 33], and particularly through legal enforcement [1, 31, 51]. The field of Explainable AI (XAI) has emerged to shed light on models' predictions and, more recently, has also explored the use of explanations to improve models' architecture, training and debugging processes [6, 15, 25, 55].

Over the last decade, several feature attribution methods applicable to Convolutional Neural Network (CNNs) were developed for explaining sample predictions, most of which are based on specific explainable modules, output sensibility to input perturbations, gradient-activation energy measures or hybrid strategies [5, 9, 11, 18, 24, 37, 54, 58, 60, 65, 69]. With an increasing number of XAI methods, different types of explanations are generated, increasing demand for suitable XAI sanity checks [2, 13] and evaluation metrics for comparing these methods in local and general contexts [3, 10, 12, 27, 30, 36, 47, 50, 52, 70].

Most of these methods focus on generating visualizations that show *where* the network is looking at or, more precisely, what is the set of pixels that contribute the most towards the final prediction, or that aligns well with human annotations. However, no prior work has attempted to understand *which* image elements consistently tend to result in higher-energy explanations for a given class when comparing different samples. We hypothesize that this difference in explanation energy observed when comparing pairs of samples is linked to the sample representativeness in the training data. To address this gap and investigate this research question, we propose XICH (XAI-based Intra-Class Hierarchies), a new methodology for unveiling hierarchies between clusters of same-class images based on XAI. In a nutshell, XICH computes a new metric called "dominance", which represents a statistical comparison of the normalized explanation energy between clusters by laying out their representative samples as a mosaic.

The proposed method starts by deep clustering same-class non-training images (i.e., validation/test), then performing a statistical comparison of energy explanation ratio of representative images from clusters laid out as a mosaic, followed by a cluster analysis on training images. In contrast to perturbation-based XAI methods, XICH elicits coherent activation because because it uses image mosaics composed by quadrants, each sampled from the original distribution [6]. An additional advantage is that this method aids in bias detection [6]. Since the introduced noise is in-distribution, any errors in the model's explanations effectively identify and exemplify its biases. Furthermore, the proposed method defines a notion of sample hierarchy that emerges naturally when comparing the energy in explanation signals of some samples in relation to that of other samples.

Given that real applications datasets can be extremely large [21, 35, 41, 59] and hence comparing each pair of images can be unfeasible, the clustering step allows us to compare images that are semantically similar while avoiding the high computational cost.

Based on a comprehensive set of experiments, we observe that low-dominance clusters tend to have much less corresponding instances in the training set, revealing out-of-distribution biases. The proposed metric is highly interpretable and complements existing intra-class sample hierarchization methods (e.g. based on logits or softmax probabilities), as our results shows that they are non-correlated. To the best of our knowledge, this is the first effort to create a cluster hierarchy based on XAI metrics, focused on comparing samples rather than methods, which provides a seminal work for understanding model bias and sensitivity to cluster features.

2 Related Work

XAI Metrics. As surveyed in [22, 39, 46], currently there are dozens of XAI metrics. Some metrics published metrics related to dominance approach includes Focus [6], that uses image mosaics to quantify coherency of explanation between different classes further (discussed in subsection 2); Level of Strenght Explanation (LSE) [10], that compare explanations and

quantifies the extent to which the explanation produced by a post-hoc explainability method supports the class predicted by a classifier. Localization [8], a metric that assesses the ability of an explanation to highlight relevant regions or features in the input data that most influence the model's decision. Pawlicki et al. [52] also provides an extensive review on XAI metrics, and highlights that the proliferation of metrics enhances the understanding of XAI systems but simultaneously exposes challenges such as metric duplication, inefficacy, and confusion. The proposed Dominance metric, besides being interpretable, complements existing intra-class sample hierarchization methods based on logits, as our results show that they are non-correlated (see Section 4).

Mosaic in XAI applications. The use of mosaic composed of image grids in XAI applications has been increasing in the last years [6, 7, 17, 34, 55, 61]. Neural networks are known to rely on context information for their decisions [42]. Popular localization evaluation scores in classification tasks, like the Energy Pointing Game (EPG) score [65, 68], rely on object bounding boxes for localization, assume that the model only relies on information within those bounding boxes, and have the additional drawback of requiring bounding box requirement. As an alternative approach, creating a grid of images (a mosaic) and measuring localization to the entire image cell allows evaluation on datasets where bounding boxes are not available and takes into consideration the context of whole target class images. Böhle et al. [17] proposed using a grid of images composed of nine images, where every class occurred at most once on each grid, one of the different classes sampled from datasets. They propose an evaluation metric based on the ratio of the sum of positive evidence for the target class in the position of the target class and the same measure in the whole mosaic grid. Arias-Duart et al. [6] proposed "Focus", with a very similar configuration to Böhle et al. [17], a quantitative visual pseudo-precision metric for feature attribution explanation methods generated for image mosaics. The non-target images' positive relevance explanation would behave as structured noise within the dataset distribution. They also proposed the use Focus to identify and characterize biased features found in CNN models based on comparative activation of similar features present (or absent) in different classes in the mosaic. Pillai and Pirsiavash [55] proposed using a grid of images composed of four images, one of a target explanation class and three distraction images randomly sampled from other classes (as distraction images). With this configuration, the explanation of the grid for the target class would not be affected by the distraction images. Hence, the authors train a model to minimize the difference between the grid explanation output and the explanation preserved in the target class image and zeroed in all distracted images. Shah et al. [61] used a simpler grid of 2 images, one MNIST digit image and a "null block" randomly placed, the digits are progressively blurred, and the model is trained with ROAR (Remove-and-Retrain) methodology [34]. Rao et al. [57] proposed comparing mosaic explanations with different mosaic processing strategy settings (merging or disassembling mosaic images at different stages of CNN processing) to efficiently compare explanation methods.

Clustering Applications on XAI. While much of the focus in XAI has been on explaining supervised learning tasks, leveraging unsupervised learning techniques to generate explanations for supervised methods can be a powerful technique. One application is use clustering features such as embeddings, distances, prototypes, and centroids as explanations [4, 16, 48, 53, 63]. Another approach is to group samples to reduce the computational complexity of combinatorial algorithms. One of these applications is finding the minimum set of exemplars to cover a dataset is a computationally intractable problem (specifically, it is equivalent to the NP-hard set cover problem). Simultaneous Construction of Clusters-Bounded Exemplars (SCCRB), and Simultaneous Construction of Clusters and Exemplars

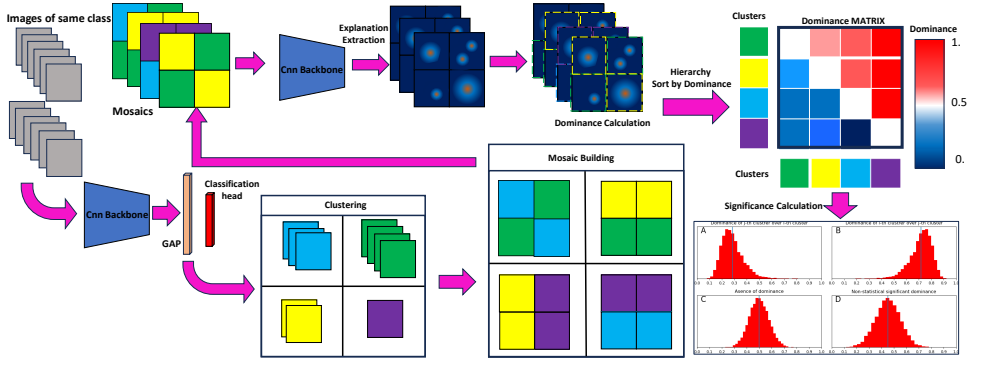

Figure 1: Proposed XICH framework for assessing explanation dominance in a pre-trained network. Images are clustered in latent space (GAP features), and synthetic 4-quadrant mosaic queries are built from exemplars of distinct clusters. Attribution maps on mosaics quantify dominance as the fraction of attribution energy within quadrants of a given cluster.

(SCCE) [20] provides efficient approximation algorithms that simultaneously partition the data into groups and find a near-optimal set of exemplars for each cluster. This theme is also discussed by Son et al. [62] and Jia et al. [38]. Clustering can also be used as a pre-processing or structuring step to improve the stability and efficiency of LIME [58] and SHAP [45] XAI methods. DLIME (Deterministic LIME) [66] proposes replacing random perturbation with a data-driven, deterministic approach based on clustering. Similarly, while SHAP (SHapley Additive exPlanations) [45] requires combinatorial analysis to evaluate input features' importance based on the interaction of features and it's outputs. C-SHAP (Clustering-Boosted SHAP) [66] addresses this efficiency bottleneck by integrating K-Means clustering as a pre-processing step grouping similar instances.

3 Methodology

The general process for computing explanation dominance values can be seen in Figure 1. Initially, the images whose predictions we want to explain are individually passed through the model. An intermediate model output (or even the logits) is used as an embedding for the images. In particular, we can use the Global Average Pooling (GAP) layer because it provides a rich semantic representation of the image with a significantly reduced spatial component. This embedding is used for clustering the images. Next, images that will compose the mosaic are sampled from two different clusters, γ_i and γ_j. The mosaic is then passed through the model, and the explanation for the mosaic is extracted. The explanation dominance for each of clusters γ_i and γ_j given the mosaic is calculated based on the sum of the energies across the explanations for the images representing the respective cluster (the formal definition is provided in Section 3.4. This process is repeated for each pairs of clusters a predefined number of times. Finally, we sort the clusters according to their relative dominance to obtain a hierarchy between clusters.

In addition, we evaluate the statistical significance of the dominance between two clusters via hypothesis testing using the dominance values obtained for each mosaic. The null hypothesis consists of the dominance mean for one cluster in relation to the other being 0.5.

If the null hypothesis is rejected, the test indicates that there are visual features in one cluster that tend to be more attended by the model.

3.1 Clustering Based on Pooling features

In this work, we used the **Global Average Pooling** (**GAP**) layer after the final convolutional block (conv-layer 4) of standard ResNet architectures as a deep feature representation of images. Let the output of the final convolutional block be a 3D tensor, or feature map, denoted as X_{feat}. This tensor has dimensions $H \times W \times C$, where H is the height and W is the width of the feature map, while C is the number of channels. For a standard ResNet-50, the shape of the feature map is typically $7 \times 7 \times 2048$ due to the adaptive average pooling, while for a standard ResNet-18, it is typically $7 \times 7 \times 500$. The GAP layer transforms this $H \times W \times C$ tensor into a $1 \times 1 \times C$ feature vector, denoted here as v_c:

$$v_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{\text{feat}}(i, j, c), \tag{1}$$

where v_c serves as a compact, semantic-rich representation of the input image. In this paper, we used K-Means with a pre-defined number of clusters on v_c features to group the images into clusters that have a clear visual coherence, in addition to reducing the computational complexity of combinations of images comparisons.

3.2 Mosaic Building

Let $I = \{img_1, img_2, .., img_N\}$ be a dataset composed by N images and $L = \{l_1, l_2, .., l_K\}$ is a set of K class labels. Every image in I is associated with a unique class label from L through a function $l(img)$. Each set of images belonging to the same class $\ell \in L$ is partitioned in a pre-defined number P of clusters $\Gamma = \{\gamma_1, \gamma_2, .., \gamma_P\}$. A mosaic is an image grid composed of several image samples[1]. For simplicity, we have used four images per mosaic arranged in a 2-by-2 grid as in previous works (see Section 2). As a slight abuse of notation, we will denote the cluster index of an image as $\gamma(img) \in \Gamma$.

In order to build a set of mosaics $M = \{m_1, m_2, .., m_J\}$, where J is the total number of mosaics in M, we create each mosaic $m \in M$ by selecting four images from two different clusters of the same class $\ell \in L$:

$$m = \{img_{i_1}, img_{i_2}, img_{i_3}, img_{i_4}\}, \tag{2}$$

where $l(img_j) = \ell$ for all $j \in \{i_1, \ldots, i_4\}$ and there exists $\gamma(img_j) \neq \gamma(img_k)$ for some $k \in \{i_1, \ldots, i_4\}$.

3.3 Explanation Extraction

GradCAM method [60] is based on use of gradients and activations of a target class score with respect to the feature maps of a specific model layer. These gradients are leveraged to compute a set of weights, α_k^ℓ, that capture the importance of each feature map k for a given class ℓ. These weights are calculated as

$$\alpha_k^\ell = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^\ell}{\partial A_{ij}^k}, \tag{3}$$

[1] Note that model training is based on individual images; mosaics are only used at inference time after rescaling.

where y^ℓ is the score for class ℓ before the softmax activation, and A^k_{ij} represents the activation at spatial location (i, j) in the k-th specific model layer feature map, denoted by A^k. The term $\frac{\partial y^\ell}{\partial A^k_{ij}}$ represents the gradient of the class score with respect to this activation, and Z is a normalization factor (the number of pixels in the feature map). These weights are then used to compute a weighted combination of the feature maps, followed by a ReLU activation, to produce the final heatmap:

$$L^\ell_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^\ell_k A^k\right) \tag{4}$$

This resulting heatmap, $L^\ell_{\text{Grad-CAM}}$, can be upsampled and overlaid on the original image to provide a clear visual indication of the regions the model focused on for its prediction. In this paper, we use a publicly available implementation of GradCAM [26].

3.4 Dominance Calculation

Arias-Duart et al. [6] defines the **positive relevance** of an explanation for an image *img* within a mosaic m with respect to a class ℓ using Grad-CAM as

$$R_\ell(img) = \sum_{(i,j)\in img} L^\ell_{\text{Grad-CAM}}. \tag{5}$$

Building on this definition, we define the **dominance** of cluster γ_i over cluster γ_j for a given mosaic m as

$$d_{\gamma_i,\gamma_j}(m) = \frac{\sum_{img_k\in\gamma_i} R_\ell(img_k)}{\sum_{img} R_\ell(img)}. \tag{6}$$

Since all images belong to the same class ℓ, the dominance of cluster γ_j over cluster γ_i is the complement of the dominance of cluster γ_i over cluster γ_j, i.e.:

$$d_{\gamma_j,\gamma_i}(m) = \frac{\sum_{img_k\in\gamma_j} R_\ell(img_k)}{\sum_{img} R_\ell(img)} = 1 - d_{\gamma_i,\gamma_j}(m). \tag{7}$$

Considering that images sampled from each class can be very different and interact in different ways, the final dominance score D_{γ_j,γ_i} is the average over a set M consisting of J mosaics produced by the images sampled from clusters γ_i and γ_j (we used $J = 25$):

$$D_{\gamma_i,\gamma_j} = \sum_{m\in M} \frac{d_{\gamma_i,\gamma_j}(m)}{J}. \tag{8}$$

Values of $D_{\gamma_i,\gamma_i} \approx 0.5$ indicate absence of dominance between clusters, $D_{\gamma_i,\gamma_j} < 0.5$ indicates dominance of cluster γ_j over cluster γ_i, and $D_{\gamma_i,\gamma_j} > 0.5$ indicates dominance in the opposite direction. However, to determine statistical significance, alongside the matrix of dominance, we compute a matrix containing the p-values for the dominance between every pair of clusters based on the t-test where the null hypothesis corresponds to a dominance mean of 0.5 (i.e., absence of dominance). After that, we sort the clusters according to their relative dominance to yield a final hierarchy between clusters.

4 Experiments and Results

In this paper, we used ResNet-18 and ResNet-50 [29]. It was selected because the Residual Network (ResNet) family has a wide range of explanation libraries and is well-known in the scientific community. Both models are fine-tuned from Imagenet-1k pre-trained weights. During the training, we apply label smoothing [49, 64] to prevent overconfidence, a learning rate schedule that combines a linear warm-up with cosine annealing [43], and the AdamW optimizer for better generalization through decoupling the weight decay from the gradient update step [44]. For explanations, we have focused on GradCAM explanations, as it fast to compute, and is proven to be best in Focus metric [8], closely related to this paper.

All tests were done on the CIFAR-10 and CIFAR-100 datasets [40] that has been used in several XAI applications [19, 23, 60, 67]. In Figure 2, we can see an example of the baby and train classes on CIFAR-100, which are categorized into 5 clusters based on GradCAM explanations. Note that in Example 1, the trained model emphasizes explanations related to babies seen from the front rather than side-view images of babies. On the other hand, in Example 2, we found that side-view baby images have stronger explanatory signals than those seen further away in bed. In Example 3, we found that clustering grouped similar train images into different clusters, and therefore, the explanatory signals have similar intensities. As expected, there is no dominance between these clusters. In Example 4, we see that colored trains seen from the side have much stronger explanatory signals (they exhibit dominance) compared to darker train images seen from the front.

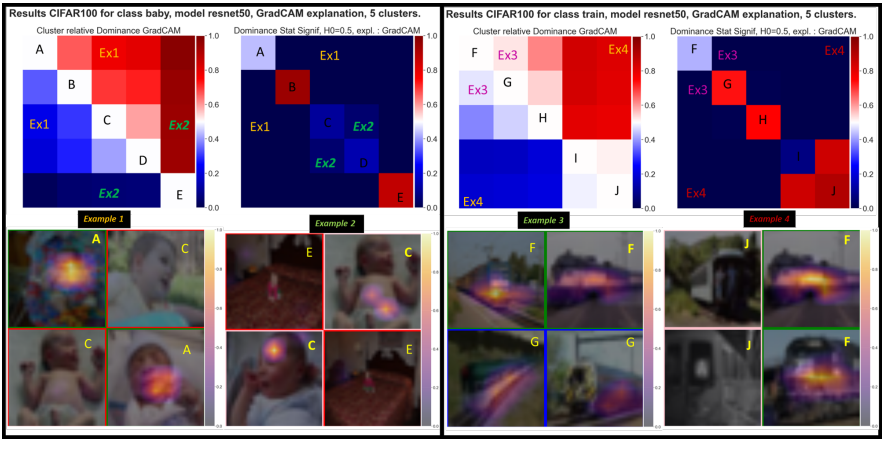

Figure 2: Cluster Dominance for CIFAR-100, classes baby and train with 5 clusters, based on GradCAM explanations. Matrix of relative dominance and matrix of statistical significance for dominance of clusters (top). Example 1: a baby class mosaic image based on samples of cluster A and C with overlaid explanation (bottom left). Example 2: a baby class mosaic image based on samples of cluster C and E with overlaid explanation (middle left). Example 3: a train class mosaic image based on samples of cluster F and G with overlaid explanation (middle right). Example 4: a train class mosaic image based on samples of cluster F and J with overlaid explanation (middle right).

Note that Grad-CAM explanations are directly proportional to the activation energy (see Equation 4). Therefore, it is natural to expect that high explanation energies in the last convolutional block would be correlated to high output logits and/or probabilities. However, our

analysis shows that this is not the case (see Figure 3). For this graph, we processed each image as an individual cluster, this approach is expensive but allows us to compare existing typically used intra-class sample hierarchization variables (logits or softmax probabilities) with dominance [32]. Dominance is non-correlated and can serve as a single or complementary parameter for sample hierarchy purposes.

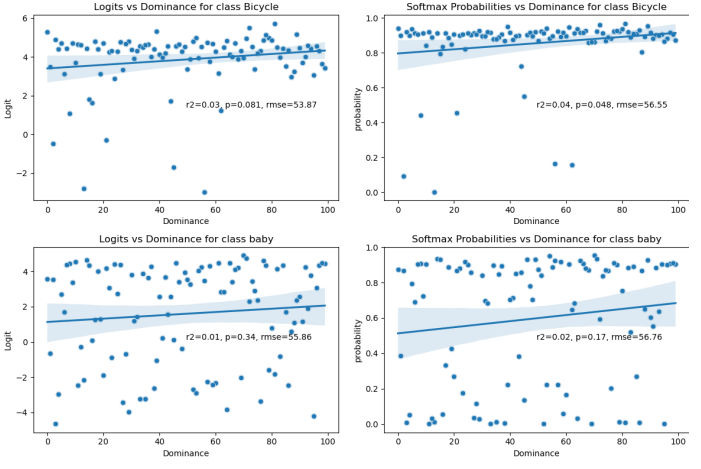

Figure 3: Cluster Dominance vs Logits and softmax probabilities for classes baby and bicycle CIFAR-100 dataset, for this graph, we have done tests considering each image as cluster, and thus it was obtained an absolute relation between images.

We found out that the test samples with lower dominance in the evaluation data are generally related to samples that have little representatives in the training data. In Figure 4, although we cannot directly relate the clusters with the highest dominance to the high number of training samples, we can see the relationship between the cluster with the lowest dominance and the lowest number of training samples (cluster with dominance ranks 3 and 4). This means that samples whose representation is relatively rarely seen during the training process present relatively weaker explanation signals (low dominance). This is achieved by fitting the training data into the evaluation clustering instance.

The same behavior can be seen on CIFAR-10 dataset (Figure 5). Note that the airplane class has a different behavior. This happens because cluster 0 has only 3 samples, and they are highly similar to the other clusters. Cluster 4 is also highly similar to the other clusters; therefore, there is almost no dominance between clusters in this class (see the first dominance matrix related to this class).

5 Conclusions and Future Work

We introduced a novel measure of hierarchy based on clustering and a post-hoc XAI method, the dominance. Our measure is not intended to replace the other existing evaluation approaches in the literature, but rather to complement the evaluation and debugging process, as it measures a different and novel aspect to be considered in the relation model-dataset. It can be used to understand model explanations, unveil out-of-distribution samples, or shed light on sample imbalance bias detection.

Calculation of dominance (and its relation to out-of-distribution interpretation) has multiple steps related to embedding, clustering, and explanation extraction; each of these steps can be unfolded for further improvements. On the other hand, new applications such as

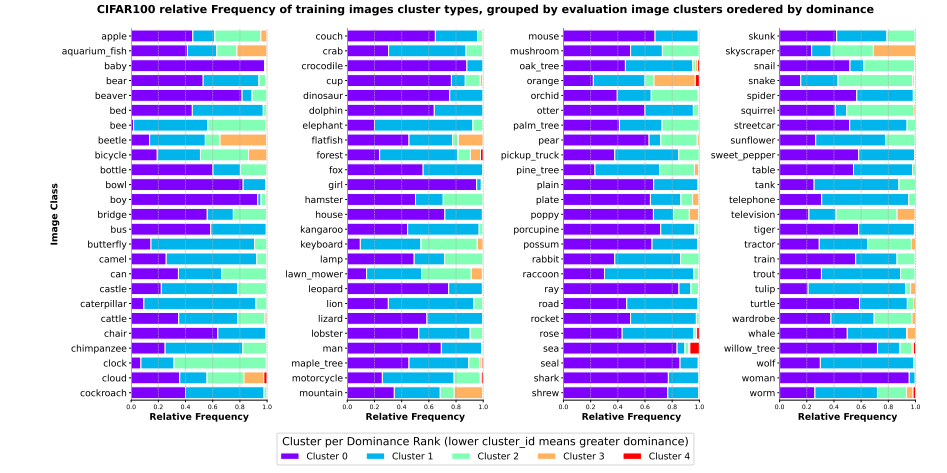

Figure 4: Relative Distribution of samples per clusters (cluster ID refers to dominance rank, 0 is the most dominant and 4 is the least) in CIFAR-100 dataset for all classes, with 5 clusters. Clusters are fitted on the evaluation dataset and predicted on the training dataset. Note that the least dominant group (clusters 3 and 4) is consistently underrepresented in the training dataset, meaning that the model gives less explanation importance to this image cluster.

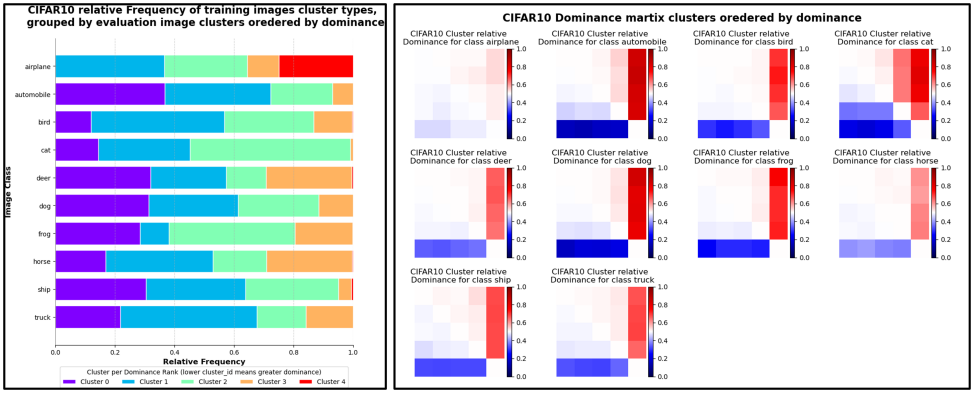

Figure 5: Relative Distribution of samples per cluster (cluster ID refers to dominance rank; 0: most dominant, 4: least dominant) in CIFAR-10 dataset for all classes. Dominance matrix for all classes on the right.

XAI dominance curriculum-learning and dominance-based adversarial explanations are unexplored opportunities for other researchers to dive deep with this new metric and findings.

Acknowledgements

The authors thank Petróleo Brasileiro S.A (Petrobrás) for their support, and Universidade Federal de Minas Gerais (UFMG), Worcester Polytechnic Institute (WPI), and the University of Sheffield for essential collaboration and resources.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

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

[3] Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations. In *ICLR 2022 Workshop on PAIR 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*.

[4] Miguel Alvarez-Garcia, Raquel Ibar-Alonso, and Mar Arenas-Parra. A comprehensive framework for explainable cluster analysis. *Information Sciences*, 663:120282, 2024. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2024.120282. URL https://www.sciencedirect.com/science/article/pii/S0020025524001956.

[5] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[6] Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Giménez-Ábalos. Focus! rating xai methods and finding biases. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2022. doi: 10.1109/FUZZ-IEEE55066.2022.9882821.

[7] Anna Arias-Duart, Ettore Mariotti, Dario Garcia-Gasulla, and Jose Maria Alonso-Moral. A confusion matrix for evaluating feature attribution methods. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3709–3714, 2023. doi: 10.1109/CVPRW59228.2023.00380.

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

[9] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Visual explanations for deep convolutional neural networks. *SN Comput. Sci.*, 2(1), January 2021. doi: 10.1007/s42979-021-00449-3. URL https://doi.org/10.1007/s42979-021-00449-3.

[10] Marilyn Bello, Rosalís Amador, María-Matilde García, Javier Del Ser, Pablo Mesejo, and Óscar Cordón. The level of strength of an explanation: A quantitative evaluation technique for post-hoc xai methods. *Pattern Recognition*, 161:

111221, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2024.111221. URL https://www.sciencedirect.com/science/article/pii/S0031320324009725.

[11] Deepshikha Bhati, Fnu Neha, Md Amiruzzaman, Angela Guercio, Deepak Kumar Shukla, and Ben Ward. Neural network interpretability with layer-wise relevance propagation: Novel techniques for neuron selection and visualization. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00441–00447, 2025. doi: 10.1109/CCWC62904.2025.10903721.

[12] Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3016–3022, 2021.

[13] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16143–16152, 2023. doi: 10.1109/CVPR52729.2023.01549.

[14] Baidyanath Biswas, Arunabha Mukhopadhyay, Ajay Kumar, and Dursun Delen. A hybrid framework using explainable ai (xai) in cyber-risk management for defence and recovery against phishing attacks. *Decision Support Systems*, 177:114102, 2024. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2023.114102. URL https://www.sciencedirect.com/science/article/pii/S016792362300177X.

[15] Jessica Y Bo, Pan Hao, and Brian Y Lim. Incremental xai: Memorable understanding of ai with incremental explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642689. URL https://doi.org/10.1145/3613904.3642689.

[16] Szymon Bobek, Michal Kuk, Maciej Szelążek, and Grzegorz J. Nalepa. Enhancing cluster analysis with explainable ai and multidimensional cluster prototypes. *IEEE Access*, 10:101556–101574, 2022. doi: 10.1109/ACCESS.2022.3208957.

[17] Moritz Böhle, Mario Fritz, and Bernt Schiele. Optimising for interpretability: Convolutional dynamic alignment networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6): 7625–7638, June 2023. ISSN 0162-8828. doi: 10.1109/TPAMI.2022.3226041. URL https://doi.org/10.1109/TPAMI.2022.3226041.

[18] 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, 2018. doi: 10.1109/WACV.2018.00097.

[19] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[20] Ian Davidson, Michael Livanos, Antoine Gourru, Peter Walker, Julien Velcin, and S. Ravi. Explainable clustering via exemplars: Complexity and efficient approximation algorithms, 09 2022.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[22] Benjamin Fresz, Lena Lörcher, and Marco Huber. Classification metrics for image explanations: Towards building reliable xai-evaluations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1–19, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658537. URL https://doi.org/10.1145/3630106.3658537.

[23] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

[24] Ruigang Fu, Qingyong Hu, Dong Xiaohu, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. 01 2020. doi: 10.5244/C.34.146.

[25] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi: 10.1145/3644073. URL https://doi.org/10.1145/3644073.

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

[27] 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.

[28] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37): eaay7120, 2019.

[29] 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 (CVPR)*, pages 770–778, 2016.

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

[31] Victoria Hendrickx. Rethinking the judicial duty to state reasons in the age of automation? *Cambridge Forum on AI: Law and Governance*, 1:e26, 2025. doi: 10.1017/cfl.2025.11.

[32] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[33] Francisco Herrera. Reflections and attentiveness on explainable artificial intelligence (xai). the journey ahead from criticisms to human–ai collaboration. *Information Fusion*, 121:103133, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103133. URL https://www.sciencedirect.com/science/article/pii/S1566253525002064.

[34] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. *A benchmark for interpretability methods in deep neural networks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[35] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henry Phillips, Ope Shpanskaya, Matthew Flegg, Robyn Ball, et al. Chexpert: A large chest xray dataset with uncertainty labels and an expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[36] Tommi Jaakkola and David Alvarez Melis. Towards robust interpretability with self-explaining neural networks. 2018.

[37] Wohuan Jia, Shaoshuai Zhang, Yue Jiang, and Li Xu. Interpreting convolutional neural networks via layer-wise relevance propagation. In *Artificial Intelligence and Security: 8th International Conference, ICAIS 2022, Qinghai, China, July 15–20, 2022, Proceedings, Part I*, page 457–467, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-06793-8. doi: 10.1007/978-3-031-06794-5_37. URL https://doi.org/10.1007/978-3-031-06794-5_37.

[38] Yangqing Jia, Jingdong Wang, Changshui Zhang, and Xian-Sheng Hua. Finding image exemplars using fast sparse affinity propagation. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 639–642, 2008.

[39] Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 000111–000124, 2023. doi: 10.1109/INES59282.2023.10297629.

[40] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[41] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[42] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[43] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[45] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30*, pages 4765–4774, 2017.

[46] Miquel Miró-Nicolau, Antoni Jaume i Capó, and Gabriel Moyà-Alcover. A comprehensive study on fidelity metrics for xai. *Information Processing & Management*, 62(1):103900, 2025. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2024.103900. URL https://www.sciencedirect.com/science/article/pii/S0306457324002590.

[47] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[48] Andrea Morichetta, Pedro Casas, and Marco Mellia. Explain-it: Towards explainable ai for unsupervised network traffic analysis. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, Big-DAMA '19, page 22–28, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369992. doi: 10.1145/3359992.3366639. URL https://doi.org/10.1145/3359992.3366639.

[49] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[50] An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *CoRR*, 2020.

[51] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. The role of explainable ai in the context of the ai act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1139–1150, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594069. URL https://doi.org/10.1145/3593013.3594069.

[52] Marek Pawlicki, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. Evaluating the necessity of the multiple metrics for assessing explainable ai: A critical examination. *Neurocomputing*, 602:128282, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.128282. URL https://www.sciencedirect.com/science/article/pii/S0925231224010531.

[53] Xi Peng, Yunfan Li, Ivor W. Tsang, Hongyuan Zhu, Jiancheng Lv, and Joey Tianyi Zhou. Xai beyond classification: interpretable neural clustering. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

[54] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *BMVC2018*, abs/1806.07421, 2018. URL http://bmvc2018.org/contents/papers/1064.pdf.

[55] Vipin Pillai and Hamed Pirsiavash. Explainable models with consistent interpretations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2431–2439, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16344.

[56] Golshid Ranjbaran, Diego Reforgiato Recupero, Chanchal K. Roy, and Kevin A. Schneider. C-shap: A hybrid method for fast and efficient interpretability. *Applied Sciences*, 15(2), 2025. ISSN 2076-3417. doi: 10.3390/app15020672. URL https://www.mdpi.com/2076-3417/15/2/672.

[57] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10213–10222, 2022. doi: 10.1109/CVPR52688.2022.00998.

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[60] 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. *Int. J. Comput. Vision*, 128(2):336–359, February 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007/s11263-019-01228-7.

[61] Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[62] Youngdoo Son, Sujee Lee, Saerom Park, and Jaewook Lee. Learning representative exemplars using one-class gaussian process regression. *Pattern Recognition*, 74:185–197, 2018. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.09.002. URL https://www.sciencedirect.com/science/article/pii/S0031320317303461.

[63] Jonathan Svirsky and Ofir Lindenbaum. Interpretable deep clustering for tabular data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[64] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[65] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, 2020. doi: 10.1109/CVPRW50498.2020.00020.

[66] Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 2021. ISSN 2504-4990. doi: 10.3390/make3030027. URL https://www.mdpi.com/2504-4990/3/3/27.

[67] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[68] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *Int. J. Comput. Vision*, 126 (10):1084–1102, October 2018. ISSN 0920-5691. doi: 10.1007/s11263-017-1059-x. URL https://doi.org/10.1007/s11263-017-1059-x.

[69] Tianyou Zheng, Qiang Wang, Yue Shen, Xiang Ma, and Xiaotian Lin. High-resolution rectified gradient-based visual explanations for weakly supervised segmentation. *Pattern Recognition*, 129:108724, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.108724. URL https://www.sciencedirect.com/science/article/pii/S0031320322002059.

[70] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.