This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# A Confusion Matrix for Evaluating Feature Attribution Methods

Anna Arias-Duart[1]
anna.ariasduart@bsc.es

Ettore Mariotti[2]
ettore.mariotti@usc.es

Dario Garcia-Gasulla[1]
dario.garcia@bsc.es

Jose Maria Alonso-Moral[2]
josemaria.alonso.moral@usc.es

[1]Barcelona Supercomputing Center (BSC)
Universitat Politècnica de Catalunya (UPC)

[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela, Spain

## Abstract

*The increasing use of deep learning models in critical areas of computer vision and the consequent need for insights into model behaviour have led to the development of numerous feature attribution methods. However, these attributions must be both meaningful and plausible to end-users, which is not always the case. Recent research has emphasized the importance of faithfulness in attributions, as plausibility without faithfulness can result in misleading explanations and incorrect decisions. In this work, we propose a novel approach to evaluate the faithfulness of feature attribution methods by constructing an 'Attribution Confusion Matrix', which allows us to leverage a wide range of existing metrics from the traditional confusion matrix. This approach effectively introduces multiple evaluation measures for faithfulness in feature attribution methods in a unified and consistent framework. We demonstrate the effectiveness of our approach on various datasets, attribution methods, and models, emphasizing the importance of faithfulness in generating plausible and reliable explanations while also illustrating the distinct behaviour of different feature attribution methods.*

## 1. Introduction

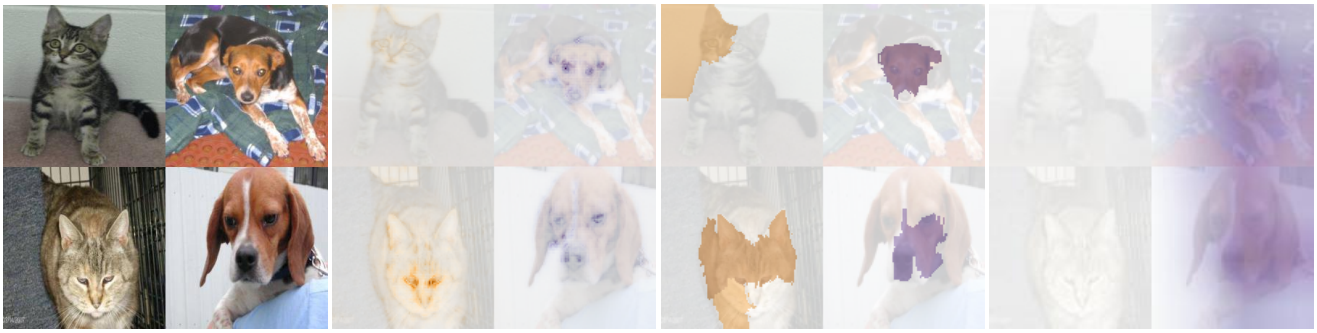
Research on eXplainable Artificial Intelligence (XAI) has gained considerable attention in recent years due to its importance in high-stakes scenarios where the consequences of a model's decision can have significant impacts on individuals or society as a whole [5]. By putting humans in the loop, XAI allows for better insights into the decision-making process of AI models, allowing for better understanding and trust in their outcomes. Additionally, regulations such as the GDPR's right to explanation have made XAI a legal requirement in certain contexts [6].

Among XAI methods, feature attribution methods aim at identifying the most relevant input features that contribute to a model's decision. These attributions must be both faithful to the model behaviour and plausible for the end user [8]. However, it's important to notice that a plausible explanation is useless if it is not faithful to the model. In this work, we aim at evaluating the faithfulness of these techniques in a quantitative manner. Assessing how faithful a deep model's explanation is can be quite challenging, given that there is no perfect ground truth to compare it against. In fact, some might argue that it is not even possible to define such a truth in the first place. This makes it tricky to evaluate different XAI methods and to determine whether a specific XAI implementation is accurate or not.

In this paper, we introduce a novel approach to evaluate the faithfulness of feature attribution methods by constructing an "Attribution Confusion Matrix" which provides familiar metrics, such as precision, accuracy, recall or the F1 score. By comparing methods in various scenarios, we aim to guide researchers and practitioners in selecting suitable attribution methods, contributing to more trustworthy and interpretable AI systems.

## 2. Related Work

Various techniques have been proposed in the Computer Vision (CV) field to evaluate the faithfulness of feature attribution methods. The absence of a ground truth explanation has led to different approaches. Many methods perturb the



(**a**) (**b**) (**c**) (**d**)

Figure 1. (**a**) Example of a mosaic made up of images from the Dogs vs. Cats[1] dataset. On the right, the explanations for the target class *dog* obtained with: (**b**) LRP (**c**) LIME (**d**) GradCAM. Purple areas correspond to positive attributions and orange to negative ones. Notice that GradCAM only provides positive attributions. The model used was a ResNet-18 architecture pre-trained on ImageNet and fine-tuned on the Dogs vs. Cats dataset.

input images according to the attribution maps to evaluate the effect on the model output [1, 3, 4, 19]. This assumes that the relevance of a feature is directly related to its effect on the model output. However, since the perturbed images fall outside the original data distribution, it is uncertain whether the change in model output is due to the absence of relevant features or simply because of the implicit distribution shift and the consequent model unpredictability. Moreover, we argue that this class of evaluation techniques tends to favour attribution maps that behave like adversarial attacks and may not necessarily be helpful in understanding the model's behaviour in real-world scenarios.

An alternative approach to evaluating feature attribution methods in deep neural networks is to use a pseudo-ground truth, such as a mask or bounding box of the target object and measure how much relevance falls within that mask [9, 23]. However, this assumption that relevance should always fall on the object is not necessarily faithful, as models may learn to distinguish classes based on the background rather than the object itself [16]. More generally, we do not have a precise understanding of what leads a model to make a particular prediction, and evaluating feature attribution scores against a bounding box that a human observer thinks is the source of evidence may introduce bias. To circumvent this problem, recent works like *Focus* [2] and [15] introduce a weaker assumption. Instead of assuming that relevance should fall on a given mask of the target object, relevance should fall on the image of the target class. In the context of image classification, the *Focus* is obtained through a set of mosaic images; a grid of images of different classes (including the target class). By doing so, a pseudo-ground truth is generated, which allows quantifying the amount of relevance that falls on those target class quadrants (see the example in Figure 1).

A limitation of the aforementioned works is that they only considers positive relevance, as they calculates the proportion of relevance that falls on the images of the target class in relation to the sum of positive relevance for the entire mosaic. However, other feature attribution methods also provide negative relevance (see Figure 1). Negative relevance is attributed to areas that provide evidence against the target class, favouring other classes instead. This issue has been addressed in the past by setting the negative impact to zero, which results in a loss of information and numerical instabilities of division by zero when all attributions are negative. Although progress has been made in evaluating feature attribution, limitations like not accounting for negative attributions or biased human-generated masks remain.

This work extends the *Focus* framework to include negative attributions as pseudo-scores, introducing the Attribution Confusion Matrix for a more comprehensive evaluation approach.

## 3. Proposed Approach

The proposed approach involves rethinking feature attribution scores as a set of classification scores that categorizes input into "relevant" and "non-relevant" classes. This is achieved by using the relaxed assumption of the *Focus*, which states that if a new image is composed of images from classes A and B in the same dataset, the attributions for class A should be mostly on images of A, while the ones *against* class A should be mostly on images of B.

### 3.1. Building the Attribution Confusion Matrix

We redefine key quantities in classification evaluation for feature attribution. To formalize this, let us first define T as the set of images belonging to the target class within the mosaic, N as the set of images not belonging to the target class, and $\alpha_i$ as the feature attributions. Therefore, for each mosaic, we define:

- True Positive evidence (TP) $= \sum_{i \in T} |\max(0, \alpha_i)|$
- False Positive evidence (FP) $= \sum_{i \in N} |\max(0, \alpha_i)|$
- True Negative evidence (TN) $= \sum_{i \in N} |\min(0, \alpha_i)|$
- False Negative evidence (FN) $= \sum_{i \in T} |\min(0, \alpha_i)|$

### 3.2. Adapting Metrics from Classification

The previous indicators are combined into an Attribution Confusion Matrix, enabling the evaluation of feature attribution performance, akin to the classic confusion matrix in classification tasks. Thanks to this link to the confusion matrix, we can borrow and extend metrics developed for the classification case to evaluate feature attribution. We propose redefining each metric X as Attribute-X. So, for example, we can define:

- Attribute-Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Attribute-Precision $= \frac{TP}{TP+FP}$
- Attribute-Recall $= \frac{TP}{TP+FN}$
- Attribute-F1 $= \frac{2 \times TP}{2 \times TP+FP+FN}$

The use of metrics adapted from classification can facilitate the evaluation of feature attribution models in a standardized and interpretable manner. We acknowledge that certain metrics may hold greater relevance in specific scenarios. For instance, in certain medical applications, we could prioritize feature attribution methods which minimize the number of false positives (*i.e.*, high Attribute-Precision) to avoid unnecessary treatments or prioritize attribution which minimizes false negatives (*i.e.*, high Attribute-Recall) to avoid non-diagnosed pathological cases. This allows practitioners to identify the most suitable option for their particular requirements. It is worth noting that the original *Focus* metric is equivalent to Attribute-Precision, thus inheriting its strengths and weaknesses.

## 4. Experiments

To be consistent with the previous *Focus* experiments [2] and for the sake of reproducibility, we use the same models. Specifically, two different architectures: VGG16 [21] and ResNet-18 [7]. The models were fine-tuned on the following datasets: the Dogs vs. Cats[1] dataset (binary problem), the MAMe [13] dataset (29 categories of art mediums and techniques) and the MIT67 [14] dataset (67 indoor scenes). The first two datasets were combined with pre-training on ImageNet [18] and the third one with the Places365-Standard dataset [24]. We evaluate the proposed metrics on each model using 4 feature attribution methods:

1. LIME [17]: based on the Tulio *et al*. implementation[2]. For each explanation, 1000 samples are used, and only the 6 superpixels with the largest attribution in absolute value are considered.
2. LRP [3]: we use the implementation of Nam *et al*. [12]. The different rules used per layer are: the $z^B$-rule [11] on the first layer, the $LRP-\epsilon$ [3] rule on fully connected layers, and the $LRP-\alpha\beta$ [3] with $\alpha = 1$ and $\beta = 0$ on convolutional layers.
3. GradCAM [20]: based on the Gildenblat *et al*. implementation[3].
4. Integrated Gradients (IG) [22]: the implementation used is from Kokhlikyan et al. [10]. We use 30 steps to approximate the integral and the black image is used as a baseline.

Notice that while LIME, LRP and IG provide both positive and negative relevance, GradCAM only generates positive attributions. These methods were chosen due to their popularity in the research community, and their diverse approaches to explainability. The source code to reproduce all experiments is available online[4].

## 5. Discussion and Future Work

The evaluation results are shown in Table 1. For each target class: 100 mosaics were built for the Dogs vs. Cats dataset (a total of 200), 100 mosaics for the MAMe (a total of 2,900) and 10 mosaics for the MIT67 (670 in total). Each mosaic is composed of two images of the target class and two other randomly chosen from the rest of the classes. Note that to keep the image content within the original data distribution (*i.e.* the one that the model has been trained on), images are not resized and therefore mosaics have a size of $448 \times 448$. It is important to acknowledge that the reliability of attribution methods is directly affected by the accuracy of the underlying model. A high-performing model (*e.g.* Dogs vs. Cats models) reinforces the plausibility of the assumption that attributions are genuinely associated with labels. In this context, our first finding is that all metrics considered appear to be equally affected by this factor, as the differences between metrics for the same attribution method remain consistent across models (*e.g.* Attribute-Precision, Attribute-Recall), irrespective of their downstream task accuracy.

Among the methods obtaining positive and negative relevance (*i.e.* LIME, LRP and IG), for the Dogs vs. Cats task (high-performing models) LIME consistently obtains

[1] https://www.kaggle.com/c/dogs-vs-cats/

[2] https://github.com/marcotcr/lime

[3] https://github.com/jacobgil/pytorch-grad-cam

[4] https://github.com/HPAI-BSC/Attribution-Confusion-Matrix

Table 1. Mean and standard deviation of the four metrics computed: Attribute-Precision, Attribute-Accuracy, Attribute-Recall and Attribute-F1. Each metric is shown grouped by column and each row shows the results for a combination of a feature attribution method, a specific architecture and a target task. For each model, the metric obtaining the highest mean is highlighted in bold. Note that since GradCAM does not provide negative relevances both $TN = 0$ and $FN = 0$, thus Attribute-Precision and Attribute-Accuracy coincide.

| | | | Attribute-Precision | Attribute-Accuracy | Attribute-Recall | Attribute-F1 |
|---|---|---|---|---|---|---|
| Dogs vs. Cats | VGG16 acc: 0.9893 | LIME | **0.9935 (± 0.0724)** | **0.9913 (± 0.0435)** | **0.9863 (± 0.0746)** | **0.9855 (± 0.0859)** |
| | | LRP | 0.9526 (± 0.0877) | 0.9343 (± 0.0835) | 0.9011 (± 0.1707) | 0.9141 (± 0.1290) |
| | | IG | 0.4973 (± 0.0912) | 0.5038 (± 0.0011) | 0.5039 (± 0.0015) | 0.4963 (± 0.0471) |
| | | GradCAM | 0.9446 (± 0.0577) | 0.9446 (± 0.0577) | - | - |
| | ResNet-18 acc: 0.9878 | LIME | **0.9913 (± 0.0739)** | **0.9853 (± 0.0786)** | **0.9796 (± 0.1131)** | **0.9776 (± 0.1154)** |
| | | LRP | 0.9741 (± 0.1018) | 0.9729 (± 0.1012) | 0.9690 (± 0.1142) | 0.9707 (± 0.1066) |
| | | IG | 0.4937 (± 0.0802) | 0.5018 (± 0.0006) | 0.5019 (± 0.0008) | 0.4944 (± 0.0419) |
| | | GradCAM | 0.9725 (± 0.0320) | 0.9725 (± 0.0320) | - | - |
| MAMe | VGG16 acc: 0.8069 | LIME | 0.7987 (± 0.2603) | 0.8048 (± 0.2373) | **0.9490 (± 0.1757)** | **0.8359 (± 0.2333)** |
| | | LRP | 0.7827 (± 0.2015) | 0.7913 (± 0.1967) | 0.9103 (± 0.2200) | 0.8311 (± 0.2001) |
| | | IG | 0.5354 (± 0.1050) | 0.5043 (± 0.0023) | 0.5065 (± 0.0035) | 0.5152 (± 0.0512) |
| | | GradCAM | **0.8665 (± 0.1123)** | **0.8665 (± 0.1123)** | - | - |
| | ResNet-19 acc: 0.8220 | LIME | 0.8020 (± 0.2520) | 0.7987 (± 0.2422) | 0.9632 (± 0.1508) | 0.8443 (± 0.2205) |
| | | LRP | 0.8864 (± 0.1268) | 0.8913 (± 0.1237) | **0.9866 (± 0.0786)** | **0.9292 (± 0.0989)** |
| | | IG | 0.6076 (± 0.1213) | 0.5027 (± 0.0015) | 0.5041 (± 0.0024) | 0.5452 (± 0.0526) |
| | | GradCAM | **0.8941 (± 0.0938)** | **0.8941 (± 0.0938)** | - | - |
| MIT67 | VGG16 acc: 0.6948 | LIME | 0.7800 (± 0.2585) | 0.7823 (± 0.2319) | **0.9390 (± 0.1823)** | **0.8218 (± 0.2280)** |
| | | LRP | 0.6012 (± 0.1918) | 0.6132 (± 0.1898) | 0.6886 (± 0.2231) | 0.6367 (± 0.2022) |
| | | IG | 0.5262 (± 0.0809) | 0.5076 (± 0.0043) | 0.5118 (± 0.0057) | 0.5157 (± 0.0401) |
| | | GradCAM | **0.8248 (± 0.1076)** | **0.8248 (± 0.1076)** | - | - |
| | ResNet-18 acc: 0.7619 | LIME | **0.9543 (± 0.1102)** | **0.9302** (± 0.1347) | 0.9611 (± 0.1282) | **0.9492 (± 0.1220)** |
| | | LRP | 0.9136 (± 0.1434) | 0.9169 (± 0.1417) | **0.9736 (± 0.1240)** | 0.9397 (± 0.1307) |
| | | IG | 0.6980 (± 0.0910) | 0.5034 (± 0.0017) | 0.5042 (± 0.0020) | 0.5829 (± 0.0334) |
| | | GradCAM | 0.9302 (± 0.0749) | **0.9302 (± 0.0749)** | - | - |

the best scores on all measures. LRP gets the second position obtaining close Attribute-Precision scores, but lower Attribute-Recalls (thus making more false negative predictions). Lastly, IG gets similarly random results on all metrics. Note that IG results may vary depending on the number of steps and the baseline image used. For the MAMe results (*i.e*. models with lower accuracy) LIME shows lower performance in all the metrics with respect to the simpler Dogs vs. Cats task, probably due to model performance drop. This also affects the variance, which increases in all metrics, particularly in Attribute-Precision (unreliable amount of false positives). Regarding the Attribute-Recall scores, both LIME and LRP maintain high mean values. Finally, for the MIT67 task (*i.e*. models with even lower performance) LIME performs better than LRP for all metrics, particularly in VGG16, with only two exceptions (Attribute-Recall and Attribute-F1 for ResNet-18).

A consistent relevant finding across the experiments is the high Attribute-Recall score of LIME and LRP (obtaining a mean greater than 0.9 in all experiments except one). That being said, underperforming models often yield lower precision scores than recall scores, indicating higher reliability of negative relevances with respect to positive relevances. However, in the case of LIME, this feature might be a consequence of the superpixels selection of LIME, since the explanation will only provide negative results, as long as these superpixels have high relevance in absolute value (being among the top 6 most attributed superpixels). This could be important for some case studies, which could motivate their use in contrast with methods which only provide positive relevance.

GradCAM generates only positive relevances, so Table 1 displays Attribute-Accuracy and Attribute-Precision, because other metrics would be misleading (*e.g*. the Attribute-Recall score would be always 1 since the false negatives always are 0 by definition). GradCAM performs well in all tasks, ranking as the top method in half of the experiments, also obtaining a small variance in general. However,

in cases where negative relevance is important, GradCAM applicability is limited.

As stated in §2, the Attribute-Precision score sometimes encounters numerical problems. This issue arises when all the attributions are negative, leading to a denominator of zero in Attribute-Precision. Consequently, this error propagates to Attribute-F1. Conversely, Attribute-Accuracy only suffers from this issue when all the attributions are zero. This is reasonable, as the accuracy of an all-zero explanation remains ambiguous.

Our proposed approach enables a comprehensive comparison of existing methods and can serve as a tool for developing and testing new methods in the CV field. This is directly applicable to other domains like natural language processing (*e.g.* sentiment analysis) and assessing transformer-based models on vision tasks. Our framework has broader implications for the validation of tools in specific use cases where the relevance of false negatives and false positives is distinct (*e.g.* systems that provide support in analyzing images in medical domains). Finally, it is worthy to remark the importance of evaluating the faithfulness of feature attribution methods, and our proposed approach provides a valuable tool for doing so. We hope that our work will inspire further research and contribute to the development of trustworthy and explainable AI systems.

## Acknowledgement

This work is conducted within the NL4XAI project which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This work is also supported by the Spanish Ministry of Science, Innovation and Universities (grants PID2021-123152OB-C21, TED2021-130295B-C33 and RED2022-134315-T) and the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19). These grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). This work is also supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu) and by the Departament de Recerca i Universitats of the Generalitat de Catalunya under the Industrial Doctorate Grant DI 2018-100.

## References

[1] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations (ICLR)*. OpenReview. net, 2018. 2

[2] 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. IEEE, 2022. 2, 3

[3] 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, 3

[4] 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

[5] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021. 1

[6] Philipp Hacker and Jan-Hendrik Passoth. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI*, volume 13200, pages 343–373. Springer International Publishing, Cham, 2022. Series Title: Lecture Notes in Computer Science. 1

[7] 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. 3

[8] 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, Online, 2020. Association for Computational Linguistics. 1

[9] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with LRP. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 2

[10] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for PyTorch. 3

[11] 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. 3

[12] Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2501–2508, 2020. 3

[13] Ferran Parés, Anna Arias-Duart, Dario Garcia-Gasulla, Gema Campo-Francés, Nina Viladrich, Eduard Ayguadé, and Jesús Labarta. The MAMe dataset: on the relevance of high resolution and variable shape image properties. *Applied Intelligence*, pages 1–22, 2022. 3

[14] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009. 3

[15] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10223–10232, 2022. 2

[16] 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, pages 1135–1144, New York, NY, USA, Aug. 2016. Association for Computing Machinery. 2

[17] 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*, pages 1135–1144, 2016. 3

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[19] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 2

[20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. 3

[21] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR)*, 2015. 3

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 3

[23] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2

[24] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3