# ODExAI: A Comprehensive Object Detection Explainable AI Evaluation

Loc Phuc Truong Nguyen[1] , Hung Truong Thanh Nguyen[2] , and Hung Cao[2]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany
[2] University of New Brunswick, Fredericton E3B 5A3, Canada
loc.pt.nguyen@fau.de, {hung.ntt,hcao3}@unb.ca

**Abstract.** Explainable Artificial Intelligence (XAI) techniques for interpreting object detection models remain in an early stage, with no established standards for systematic evaluation. This absence of consensus hinders both the comparative analysis of methods and the informed selection of suitable approaches. To address this gap, we introduce the *Object Detection Explainable AI Evaluation* (ODExAI), a comprehensive framework designed to assess XAI methods in object detection based on three core dimensions: localization accuracy, faithfulness to model behavior, and computational complexity. We benchmark a set of XAI methods across two widely used object detectors (YOLOX and Faster R-CNN) and standard datasets (MS-COCO and PASCAL VOC). Empirical results demonstrate that region-based methods (e.g., D-CLOSE) achieve strong localization (PG = 88.49%) and high model faithfulness (OA = 0.863), though with substantial computational overhead (Time = 71.42s). On the other hand, CAM-based methods (e.g., G-CAME) achieve superior localization (PG = 96.13%) and significantly lower runtime (Time = 0.54s), but at the expense of reduced faithfulness (OA = 0.549). These findings demonstrate critical trade-offs among existing XAI approaches and reinforce the need for task-specific evaluation when deploying them in object detection pipelines. Our implementation and evaluation benchmarks are publicly available at: https://github.com/Analytics-Everywhere-Lab/odexai.

**Keywords:** Explainable AI · Object detection · Evaluation framework

## 1 Introduction

Deep neural networks (DNNs) have greatly advanced object detection, but their opaque decision-making limits interpretability [36,38,4]. This reduces trust and restricts deployment in sensitive domains such as healthcare [41,59,11,37] and security [40,6,33], where models must offer not only high accuracy but also transparent, fair, and domain-aligned reasoning [55,10,62]. Explainable AI (XAI) addresses this need by making detection decisions more interpretable.

XAI has emerged as a distinct research area, driven by advances in deep learning, DARPA-funded initiatives [17], and the recognition of a right to explanation in the GDPR [20], EU AI Act [35]. In recent years, a wide range of XAI methods for object detectors has been proposed. While this reflects growing interest in the

field, it also presents several challenges [24]. The number of available methods is increasing rapidly, and many offer only minor variations, such as Spatial Sensitive Grad-CAM [61] and Spatial Sensitive Grad-CAM++[60], or produce similar outputs despite differing implementations, as seen with D-RISE [44] and ODAM [65]. As discussed by several works [34,36,38,4,40], explanation quality must be assessed across several dimensions, such as faithfulness to the model's reasoning, conciseness, fairness, and alignment with domain knowledge. Yet two critical questions remain open: *how to select an appropriate XAI method*, and *how to benchmark it in a standardized manner* that ensures objective evaluation and supports generalizable findings.

To address these challenges, we introduce the *Object Detection Explainable AI Evaluation* (ODExAI), a structured approach for evaluating and comparing XAI methods for object detection. Recent work has focused on developing metrics to benchmark interpretability techniques across settings. These metrics enable quantitative assessment of localization, faithfulness, and complexity, offering insight into each method's strengths, limitations, and suitability. The main contributions of this work are:

1. We propose ODExAI Framework grounded in three key interpretability dimensions: faithfulness, localization, and complexity. Each dimension is supported by a set of quantitative metrics, enabling systematic and task-specific assessment.
2. We conduct a toy experiment to illustrate the use of ODExAI in evaluating three novel XAI methods: D-CLOSE [57], G-CAME [39], and D-RISE [44], across two widely used detectors, YOLOX [14] and Faster R-CNN [47], using the MS COCO [25] and PASCAL VOC [12] datasets.
3. We conduct a comprehensive analysis of the trade-offs between XAI methods, highlighting their performance across different interpretability dimensions. Based on the evaluation outcomes, we provide practical guidelines for selecting suitable XAI methods according to task-specific requirements.

## 2   Background

This section reviews object detection models and the XAI methods applied to interpret their outputs. It also outlines recent progress and key challenges, particularly the lack of a unified and standardized evaluation framework for explanation techniques.

### 2.1   Object Detection Models

DNNs have significantly advanced object detection through various techniques, each offering distinct advantages. Single-stage detectors, such as the YOLO [45] family (e.g., YOLOv3 [46], YOLOv4 [3], and YOLOX [14]) and SSD[26], are optimized for real-time detection, offering a balance between speed and accuracy. Two-stage detectors, known for their high precision, include Faster R-CNN [48],

Mask R-CNN [18] (which extends detection to instance segmentation), and Cascade R-CNN [7], which refines results through multi-stage processes. Anchor-free detectors simplify detection by directly predicting object centers or corners, with notable examples such as CornerNet [23], CenterNet [9], and FCOS [56] (Fully Convolutional One-Stage Object Detection). Transformer-based detectors, including DETR [56] (DEtection TRansformer), Deformable DETR [8] (which accelerates convergence), and Swin Transformer [27] (known for hierarchical feature extraction), leverage attention mechanisms to improve performance.

## 2.2   Explainable AI for Object Detection

While XAI techniques are widely used in classification and segmentation, their application to object detection remains limited due to constraints in flexibility, efficiency, and adaptability [16]. As noted in [39], these methods fall into two main categories: region-based and CAM-based saliency approaches. Region-based methods assess the importance of image regions by masking inputs and observing changes in output. Examples include LIME [49], RISE [43], and object detection variants such as SODEx [50], D-RISE [44], and D-CLOSE [57], which enhance precision through task-specific adaptations. These approaches are model-agnostic and intuitive. In contrast, CAM-based methods like Grad-CAM [51] use feature maps and partial derivatives to generate saliency maps, offering efficiency but often yielding less informative results for detection tasks. Designed primarily for classification, they lack dedicated adaptations for detection. The recently proposed G-CAME [39] addresses these gaps by providing stable, efficient explanations across both one-stage and two-stage detectors.

## 2.3   Development and Gaps in XAI for Object Detection

XAI is essential for object detection for understanding how models assign class labels and bounding boxes from visual features. It helps verify whether predictions rely on relevant regions, supports error analysis, and enhances transparency. In healthcare, XAI has been applied to thyroid nodule detection [41], tumor localization [11], and abnormality identification [59]. In autonomous driving, it aids in diagnosing detection failures [53] and assessing robustness under varying weather conditions [54]. In environmental monitoring, it explains outputs in deforestation tracking [5] and ecological change analysis [52]. In security, it promotes fairness in detecting individuals with disabilities [40], while in agriculture, it supports interpretation in crop health monitoring [33] and pest detection [6].

Although XAI methods are broadly applicable, their evaluation remains inconsistent and lacks standardization [24,2]. This stems from subjective assessment criteria and methodological diversity. In many cases, formal procedures are absent, leading to ad hoc strategies tailored to specific models, datasets, or use cases [28]. As a result, findings are often difficult to compare or reproduce, limiting progress toward reliable and generalizable explanations. Several initiatives have attempted to address this gap. Quantus [19], an open-source toolkit, offers computational metrics for evaluating explanation properties. Other studies seek to

clarify evaluation processes [24,34] or provide practical guidelines for designing evaluation protocols [29,2]. However, current approaches lack consensus, and no unified evaluation framework exists for XAI for object detection [41,31,21].

## 3   The proposed ODExAI Framework

The absence of a standardized evaluation framework in XAI has led to the use of ad hoc metrics, often tailored to specific methods. Although these metrics cover diverse evaluation criteria, their proliferation introduces uncertainty about task relevance, hindering reproducibility and generalization. To address this, ODExAI defines three core evaluation properties for object detection, each illustrated by schematic diagrams (Figure 2–Figure 4). Metrics are grouped by property based on shared objectives, enabling a more structured, consistent, and interpretable evaluation process.

### 3.1   Framework Overview

Our framework consists of two main components: Saliency Map Extraction with XAI and Explanation Evaluation. The overall procedure is illustrated in Figure 1. A detailed description of each component is provided below:

1. **Saliency Map Extraction with XAI**: This part focuses on generating saliency maps using XAI methods. Users begin by uploading an image and selecting the task, with object detection models assigned, e.g., Faster R-CNN [47] and YOLOX [14]. After prediction, the user selects the target object and an explanation method, e.g., D-CLOSE [57], G-CAME [39], and D-RISE [44], to generate the saliency map.
2. **Explanation Evaluation**: This part provides an overview of each explanation method across three dimensions: localization, faithfulness, and complexity. Once the explanation is generated, users can initiate the evaluation, which is visualized as a spider plot. Each dimension includes multiple metrics, with corresponding numerical values that quantify the method's performance.

### 3.2   Localization Metrics

Localization quality is measured by how well an explanation aligns with a user-defined region of interest (ROI) [4], as shown in Figure 2. This involves comparing the most relevant pixels in the explanation map to labeled regions such as bounding boxes or segmentation masks. Since the ROI is assumed to be critical to the model's decision [4,1,63], localization is considered high when relevance is concentrated within it. We quantify this using the Pointing Game (PG) [63] and Energy-based Pointing Game (EBPG) [58] metrics.
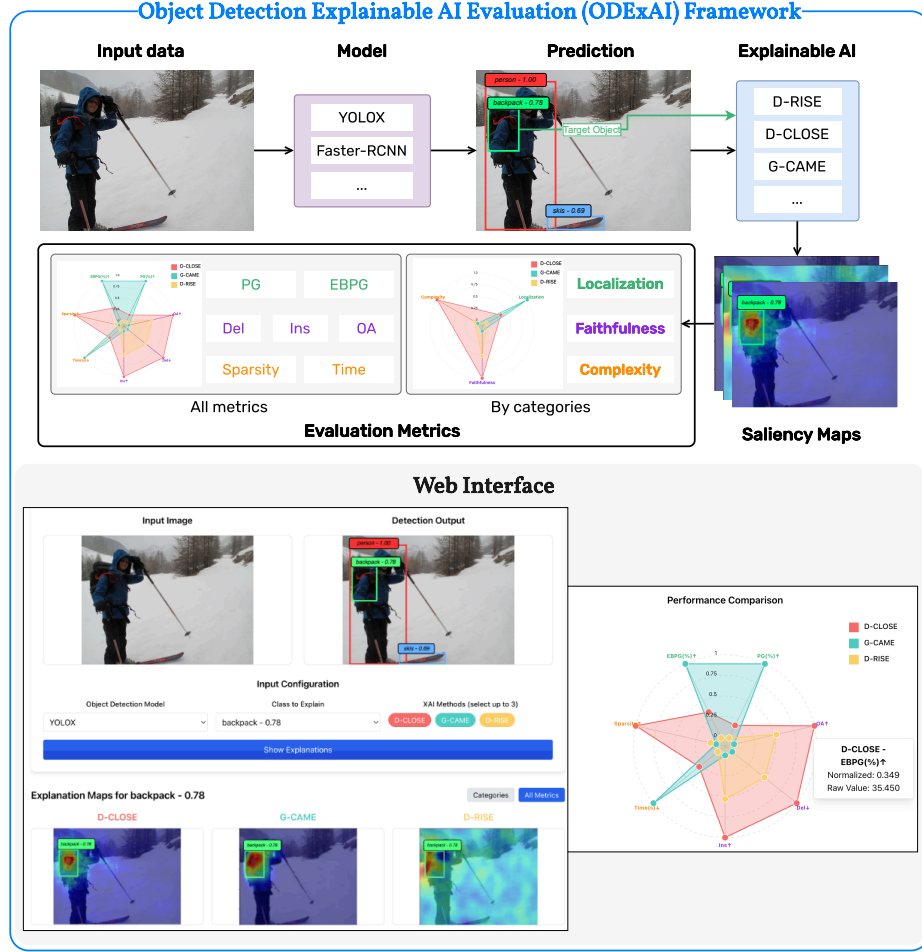
Fig. 1: Overview of the architecture and web interface of our proposed Object Detection Explainable AI Evaluation (ODExAI).

**Pointing Game (PG)** [63] evaluates the spatial precision of saliency maps by measuring their ability to localize target object categories. For each saliency map, the metric identifies the point of maximum saliency and checks if it falls within the ground truth annotation of the target object. A hit is recorded if the point lies within the annotated region; otherwise, it is counted as a miss. The localization accuracy is defined as:

$$PG = \frac{\# \text{ Hits}}{\# \text{ Hits} + \# \text{ Misses}} \tag{1}$$

where # Hits and # Misses denote the total number of correct and incorrect localizations, respectively. The overall performance is calculated by averaging the
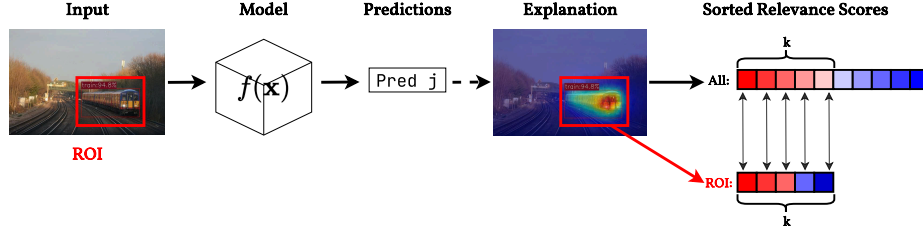
Fig. 2: Demonstration of the localization property using region-specific relevance comparisons. Localization measures how well an explanation aligns with the spatial region that is most responsible for the model's prediction. To evaluate this, a region of interest (ROI) is predefined in the input image (red box), marking the area where the most relevant evidence is expected to reside. In this example, the target is the object "train". After model inference, an explanation map is generated, and pixel-wise relevance scores are computed. The $k$ relevance scores within the ROI are then compared with the top-$k$ scores from the full explanation map. A method with strong localization is expected to assign the highest relevance to pixels inside the ROI.

accuracy across all categories. A high PG score for an XAI method indicates its ability to generate accurate and interpretable saliency maps.

**Energy-Based Pointing Game (EBPG)** [58] evaluates the spatial accuracy of saliency maps by analyzing how much of the map's energy (activation) falls within the annotated bounding box of the target object. Unlike the traditional PG, which focuses on a single maximum point, this metric considers the overall distribution of energy within the relevant regions. The localization accuracy is defined as:

$$\text{EBPG} = \frac{\sum L^c_{(i,j)\in\text{bbox}}}{\sum L^c_{(i,j)\in\text{bbox}} + \sum L^c_{(i,j)\notin\text{bbox}}} \tag{2}$$

where $L^c_{(i,j)}$ represents the saliency map intensity at pixel $(i,j)$, and bbox refers to the bounding box region of the target object. The overall performance is calculated by averaging the accuracy across all categories. A high EBPG value indicates that the XAI-generated saliency maps can effectively localize the target region and suppress noise outside the bounding box.

### 3.3    Faithfulness Metrics

Faithfulness measures how changes to highly relevant input features affect the model's prediction [57], as shown in Figure 3. It is evaluated by progressively inserting or deleting pixels with high relevance scores and comparing the resulting predictions to the original. Since relevance reflects each feature's contribution,
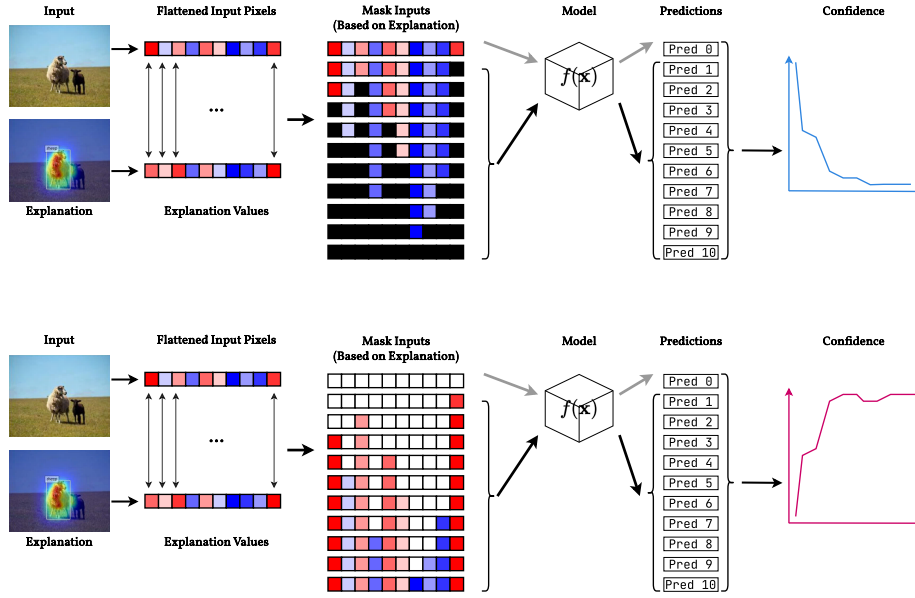
Fig. 3: Illustration of the faithfulness property using progressive perturbation. Faithfulness evaluates how the most relevant pixels affect the model's prediction. Explanation scores are first ranked to identify the most relevant features, highlighted in red. Their corresponding positions are then mapped onto the flattened input, as indicated by the arrows. In the top panel, pixels are progressively removed (masked in black), while in the bottom panel, they are progressively added to a blank input. At each step, the modified input is passed through the model, and the prediction confidence is recorded. A faithful explanation produces a steep confidence drop when highly relevant pixels are removed and a sharp increase when they are added, whereas perturbing less relevant regions yields smaller changes.

significant prediction shifts are expected when important features are modified. We quantify faithfulness using the Deletion and Insertion [43] and Over-All (OA) [64] metrics.

**Deletion (Del) and Insertion (Ins)** The Deletion [43] metric evaluates the effect of removing key features by tracking the decline in predicted class probability as important pixels, identified by the saliency map, are progressively removed. A steep probability drop, quantified by a low area under the curve (AUC), denotes a strong explanation. In contrast, the Insertion [43] metric measures the probability increase when pixels are gradually introduced, with a higher AUC indicating better explanations.
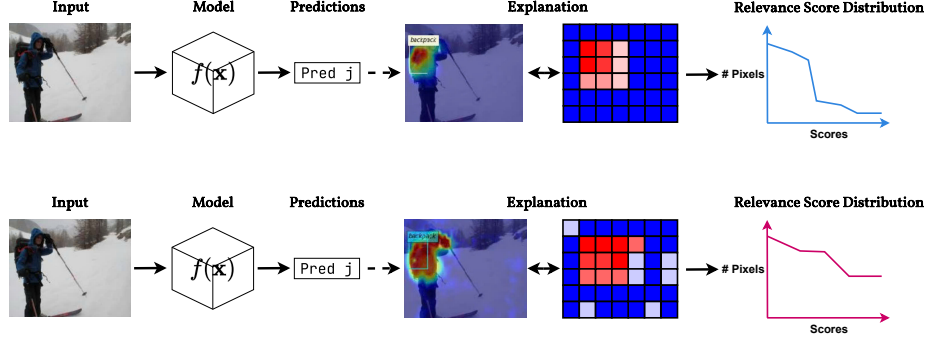
Fig. 4: Visualization of the complexity property via relevance score distributions. Complexity measures how concentrated or dispersed an explanation is over the input space. After inference, relevance scores are assigned to all input pixels and sorted to generate a distribution curve. In the top panel, low complexity is illustrated by a steep curve where few pixels carry most of the relevance. In contrast, the bottom panel shows high complexity, with scores distributed more uniformly, resulting in a flatter curve. Low-complexity explanations are desirable because they highlight key regions, reduce the effective input space, and focus the analysis on essential model components, leading to more efficient and interpretable outputs.

**Over-All (OA)** [64] metric is defined as $\text{AUC}(\text{Insertion}) - \text{AUC}(\text{Deletion})$, provides a comprehensive evaluation of saliency map quality by considering both its positive and negative effects on the model's predictions. A higher score indicates the saliency map effectively highlights critical regions (high Insertion) while minimizing the influence of irrelevant ones (low Deletion).

### 3.4   Complexity Metrics

Complexity reflects the conciseness of an explanation, ideally highlighting a small set of key features [4]. Concise explanations are easier to interpret and tend to convey more useful information with less noise [4,15]. We also incorporate computation time into the complexity measure, as high runtime can limit the practicality of XAI methods in real-time or large-scale systems [22]. Complexity is quantified using the Sparsity metric [15] and by recording the runtime of each method. The overall concept is illustrated in Figure 4.

**Sparsity** [15] metric measures how concentrated or dispersed the scores in a saliency map are, reflecting the focus of attention on specific regions. It is defined as the ratio of the maximum score ($S_{\max}$) to the mean score ($S_{\text{mean}}$):

$$\text{Sparsity} = \frac{S_{\max}}{S_{\text{mean}}} \tag{3}$$

It should be noted that all saliency map scores are min–max normalized beforehand to ensure consistency across methods. A high sparsity value corresponds to a high $S_{\max}/S_{\mean}$ ratio, i.e., a low mean score $S_{\mean}$, indicating that the map's focused areas are narrow and well-defined.

**Computation Time** records the time taken to compute explanations for a given model input. It is typically reported as the average runtime per instance and depends on factors such as model complexity, input size, and explanation method. Lower values indicate higher efficiency, while higher values reflect greater computational cost.

## 4 Demonstrative Example

This section presents a toy experiment to demonstrate the utility of the ODExAI framework. We first introduce the evaluated XAI methods, benchmark datasets, and object detection architectures, followed by the experiment and a detailed report of the results.

### 4.1 Benchmark Datasets and Object Detection Models

We evaluate XAI methods on two standard object detection benchmarks: **MS COCO 2017** [25], containing 123,287 images across 80 object categories, and **PASCAL VOC 2012** [12], with 11,530 images spanning 20 classes. These datasets are selected for their diversity in scene complexity and annotation quality. To ensure architectural variety, we employ **Faster R-CNN** [47], a two-stage detector known for high accuracy, and **YOLOX** [14], a single-stage, anchor-free model optimized for real-time performance.

### 4.2 XAI Methods for Object Detection

In this work, we focus on XAI methods from the region-based and CAM-based categories, as defined in [39] for object detection models. A brief description of the selected methods is given as follows:

1. **D-CLOSE** [57] is a region-based XAI method that generates saliency maps by segmenting the input image at multiple levels and combining their features to enhance explanation quality. Random masks are used to perturb the input, and the similarity scores between the masked outputs and the target prediction are calculated to weigh the importance of each region. The saliency maps are normalized using a density map to reduce noise, and multi-scale feature fusion aggregates fine-grained to coarse-grained features.
2. **D-RISE** [44] is a region-based XAI method designed to generate saliency maps for object detectors by jointly addressing localization and classification. It applies randomized binary masks to the input and evaluates their impact on detection outcomes. The importance of each masked region is estimated using

a similarity metric that combines Intersection over Union (IoU) for spatial alignment, cosine similarity for classification consistency, and objectness scores. The final saliency map is obtained by aggregating the masks weighted by their contributions.

3. **G-CAME** [39] improves CAM-based explanation methods for object detection by integrating a Gaussian kernel for saliency map refinement. Specifically, it combines gradient-based localization with Gaussian weighting, effectively suppressing background noise and enhancing the detection of small objects. The method consists of four steps: selection of target layers, gradient-based object localization, feature map weighting, and Gaussian masking.

### 4.3   Results

We evaluate the selected XAI methods for object detection using YOLOX and Faster R-CNN on the MS COCO and PASCAL VOC datasets across three core criteria: faithfulness, localization, and complexity, each measured using dedicated quantitative metrics. All experiments are conducted under identical settings, with 2000 perturbed masks for both D-CLOSE and D-RISE. The evaluation is run on a system with an Nvidia Tesla T4 GPU and 24GB of RAM. Results in Table 1 show that ODExAI consistently measures explanation quality and informs method selection for specific detection tasks.

**Faithfulness Results** Across both datasets and models, D-CLOSE consistently outperformed in faithfulness metrics. For example, on MS-COCO with YOLOX, it achieved the highest Insertion (0.908) and OA (0.881), with the lowest Deletion (0.027). This trend held across all settings, with D-CLOSE achieving the highest OA on three out of four model–dataset combinations. These results indicate that region-based explanations generated by D-CLOSE are tightly aligned with the model's true decision process, highlighting its high faithfulness.

**Localization Results** In contrast, G-CAME excelled in the localization dimension, achieving the highest EBPG and PG scores across all configurations.

Table 1: Evaluation results from ODExAI.

| Dataset | MS-COCO | | | | | | PASCAL VOC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | YOLOX | | | Faster R-CNN | | | YOLOX | | | Faster R-CNN | | |
| Method | D-CLOSE | G-CAME | D-RISE | D-CLOSE | G-CAME | D-RISE | D-CLOSE | G-CAME | D-RISE | D-CLOSE | G-CAME | D-RISE |
| Ins↑ | **0.908** | 0.703 | 0.812 | **0.912** | 0.718 | 0.867 | **0.804** | 0.512 | 0.775 | **0.826** | 0.534 | 0.783 |
| Del↓ | **0.027** | 0.059 | 0.043 | **0.049** | 0.169 | 0.152 | 0.128 | 0.183 | **0.103** | **0.171** | 0.284 | 0.206 |
| OA↑ | **0.881** | 0.644 | 0.769 | **0.863** | 0.549 | 0.715 | **0.676** | 0.329 | 0.672 | **0.655** | 0.250 | 0.577 |
| PG(%)↑ | 87.86 | **94.31** | 86.55 | 88.49 | **96.13** | 84.12 | 81.48 | **93.96** | 79.01 | 82.94 | **95.13** | 80.11 |
| EBPG(%)↑ | 35.45 | **67.16** | 18.47 | 36.13 | **70.11** | 20.74 | 29.23 | **58.39** | 28.11 | 31.84 | **57.84** | 29.94 |
| Sparsity↑ | **25.02** | 2.84 | 4.43 | **26.13** | 4.95 | 5.91 | **26.84** | 3.81 | 12.45 | **27.12** | 4.91 | 6.21 |
| Time(s)↓ | 70.12 | **0.43** | 98.67 | 71.42 | **0.54** | 99.11 | 70.94 | **0.48** | 99.19 | 72.94 | **0.63** | 99.98 |

For instance, on YOLOX with MS COCO, it reached PG = 94.31% and EBPG = 67.16%, significantly outperforming D-CLOSE and D-RISE. This consistent strength in localization suggests that CAM-based methods are highly effective at identifying relevant object regions, even when their explanations do not fully capture the model's internal reasoning.

**Complexity Results** G-CAME was also the most computationally efficient, with runtimes under 1s across all settings. For instance, on MS-COCO with YOLOX, it required only 0.43s, compared to 70.12s for D-CLOSE and 98.67s for D-RISE. However, its Sparsity was the lowest, indicating more diffuse explanations. In contrast, D-CLOSE produced the most concise maps but at a significantly higher cost.

## 5    Discussion

This section discusses the trade-offs between XAI methods revealed by ODExAI, offers guidance on selecting suitable approaches, acknowledges current limitations, and outlines directions for future development.

### 5.1    Trade-off Analysis and Suggestions for Method Selection

The evaluation reveals clear trade-offs among XAI methods for object detection, showing that no single approach satisfies all interpretability requirements [42,30]. Region-based methods like D-CLOSE demonstrate strong faithfulness and produce sparse saliency maps due to their perturbation-based design: they systematically mask input regions and measure changes in model predictions, directly attributing importance to causally influential areas [57,39]. This yields explanations closely aligned with the model's decision process, as confirmed by faithfulness metrics, and results in concise, object-focused maps with minimal noise [57]. However, the repeated inference across multiple segmentation levels imposes significant computational overhead, limiting practicality for real-time or resource-constrained applications [57,39].

In contrast, CAM-based methods such as G-CAME achieve superior localization and low runtime by using internal feature maps and gradient signals from specific layers to generate saliency maps [39]. On the one hand, selecting appropriate layers for explanation requires specialized knowledge from users [13,38]. On the other hand, this design enables rapid computation and produces highly localized explanations, as reflected in strong PG and EBPG scores. However, because they rely on aggregated activations rather than direct input perturbations, these methods may miss the exact features responsible for decisions, leading to lower faithfulness [39,32].

D-RISE, while not leading in any single metric, maintains consistently moderate performance across all evaluation dimensions, making it a robust option for ensemble-based detection or sensitivity analysis, where explanation diversity is valuable [57,39,44]. These results demonstrate ODExAI's effectiveness in

highlighting method-specific trade-offs and guiding the choice of XAI techniques based on task-specific interpretability goals.

### 5.2   Limitations and Future Development

Despite promising results, the current framework presents several limitations. First, the number of included XAI methods and evaluation metrics remains limited, which may restrict generalizability. Future work will expand the framework to incorporate additional methods and metrics, aiming for broader coverage and the development of more comprehensive evaluation tools. Second, some dimensions mix higher-better and lower-better metrics, which may introduce inconsistency. A possible solution is to assign weights to account for these directional differences. Third, the framework lacks a proper normalization scheme, which can be misleading when metrics differ in scale or unit. A promising approach is the skill score proposed in [4], which compares an explanation method's performance relative to a reference baseline designed to violate the assumptions of each evaluation property. The skill score quantifies how much an explanation improves over this worst-case scenario.

## 6   Conclusion

In this work, we introduce the *Object Detection Explainable AI Evaluation* (ODExAI), a structured system for assessing XAI methods for object detection. ODExAI supports the selection of explanation techniques based on three core dimensions: localization accuracy, faithfulness to model predictions, and computational complexity. To validate the framework, we evaluate several XAI methods on two widely used object detectors and standard benchmark datasets. Results show that region-based methods perform well in both localization and faithfulness but incur higher computational cost. CAM-based methods, by contrast, are more efficient and provide good localization but are less faithful. These findings reveal key trade-offs and underscore the need for task-specific evaluation when selecting XAI methods. Future work will expand the framework with additional methods and metrics, and establish consistent, interpretable evaluation procedures for each dimension.

### Acknowledgment

### References

1. Arras, L., Osman, A., Samek, W.: CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81**, 14–40 (2022). https://doi.org/10.1016/j.inffus.2021.11.008

2. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. arXiv preprint arXiv:1909.03012 (2019)

3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934 (2020)

4. Bommer, P.L., Kretschmer, M., Hedström, A., Bareeva, D., Höhne, M.M.C.: Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. Artificial Intelligence for the Earth Systems **3**(3), e230074 (2024). https://doi.org/10.1175/AIES-D-23-0074.1

5. Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek, S., Stampfer, K., Holzinger, A.: Exploring artificial intelligence for applications of drones in forest ecology and management. Forest Ecology and Management **551**, 121530 (2024). https://doi.org/10.1016/j.foreco.2023.121530

6. C, B., R, S., A, A., Moorthy, R.S.: A Novel Approach for Jute Pest Detection Using Improved VGG-19 and XAI. In: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–5 (2024). https://doi.org/10.1109/ICCCNT61001.2024.10723939

7. Cai, Z., Vasconcelos, N.:  Cascade R-CNN: Delving Into High Quality Object Detection . In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6154–6162. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2018). https://doi.org/10.1109/CVPR.2018.00644

8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. p. 213–229. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58452-8_13

9. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: Keypoint Triplets for Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6568–6577 (2019). https://doi.org/10.1109/ICCV.2019.00667

10. Ennab, M., Mcheick, H.: Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. Frontiers in Robotics and AI **11**, 1444763 (2024). https://doi.org/10.3389/frobt.2024.1444763

11. Esmaeili, M., Vettukattil, R., Banitalebi, H., Krogh, N.R., Geitung, J.T.: Explainable Artificial Intelligence for Human-Machine Interaction in Brain Tumor Localization. Journal of Personalized Medicine **11**(11) (2021). https://doi.org/10.3390/jpm11111213

12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010). https://doi.org/10.1007/s11263-009-0275-4

13. Famiglini, L., Campagner, A., Barandas, M., La Maida, G.A., Gallazzi, E., Cabitza, F.: Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. Computers in Biology and Medicine **170**, 108042 (2024). https://doi.org/10.1016/j.compbiomed.2024.108042

14. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO Series in 2021. arXiv preprint arXiv:2107.08430 (2021)

15. Gomez, T., Fréour, T., Mouchère, H.: Metrics for Saliency Map Evaluation of Deep Learning Explanation Methods. In: El Yacoubi, M., Granger, E., Yuen, P.C., Pal, U., Vincent, N. (eds.) Pattern Recognition and Artificial Intelligence. pp. 84–95. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-09037-0_8

16. Grami, A.: The Gaussian Distribution, pp. 201–238. John Wiley & Sons (2019). https://doi.org/10.1002/9781119300847.ch7
17. Gunning, D.: DARPA's Explainable Artificial Intelligence (XAI) Program. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. p. ii. IUI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1609/aimag.v40i2.2850
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
19. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023)
20. Hoofnagle, C.J., van der Sloot, B., and, F.Z.B.: The European Union general data protection regulation: what it is and what it means*. Information & Communications Technology Law **28**(1), 65–98 (2019). https://doi.org/10.1080/13600834.2019.1573501
21. Kadir, M.A., Mosavi, A., Sonntag, D.: Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In: 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES). pp. 000111–000124 (2023). https://doi.org/10.1109/INES59282.2023.10297629
22. Kakogeorgiou, I., Karantzalos, K.: Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. International Journal of Applied Earth Observation and Geoinformation **103**, 102520 (2021). https://doi.org/10.1016/j.jag.2021.102520
23. Law, H., Deng, J.: CornerNet: Detecting Objects as Paired Keypoints. International Journal of Computer Vision **128**(3), 642–656 (Mar 2020). https://doi.org/10.1007/s11263-019-01204-1
24. Le, P.Q., Nauta, M., Nguyen, V.B., Pathak, S., Schlötterer, J., Seifert, C.: Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges. In: Elkind, E. (ed.) Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. pp. 6665–6673. International Joint Conferences on Artificial Intelligence Organization (8 2023). https://doi.org/10.24963/ijcai.2023/747, survey Track
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
26. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 21–37. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows . In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00986
28. Lopes, P., Silva, E., Braga, C., Oliveira, T., Rosado, L.: XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. Applied Sciences **12**(19) (2022). https://doi.org/10.3390/app12199423

29. Mohseni, S., Zarei, N., Ragan, E.D.: A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. **11**(3–4) (Sep 2021). https://doi.org/10.1145/3387166
30. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models, pp. 39–68. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-04083-2_4
31. Moradi, M., et al.: Model-agnostic explainable artificial intelligence for object detection in image data. Engineering Applications of Artificial Intelligence **137**, 109183 (2024). https://doi.org/10.1016/j.engappai.2024.109183
32. Mumuni, F., Mumuni, A.: Explainable artificial intelligence (XAI): from inherent explainability to large language models. arXiv preprint arXiv:2501.09967 (2025)
33. Natarajan, S., Chakrabarti, P., Margala, M.: Robust diagnosis and meta visualizations of plant diseases through deep neural architecture with explainable AI. Scientific Reports **14**(1), 13695 (Jun 2024). https://doi.org/10.1038/s41598-024-64601-8
34. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Comput. Surv. **55**(13s) (Jul 2023). https://doi.org/10.1145/3583558
35. Neuwirth, R.: The EU Artificial Intelligence Act: Regulating Subliminal AI Systems. Routledge Research in the Law of Emerging Technologies, Taylor & Francis (2022)
36. Nguyen, H., Clement, T., Nguyen, L., Kemmerzell, N., Truong, B., Nguyen, K., Abdelaal, M., Cao, H.: LangXAI: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks. In: Larson, K. (ed.) Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. pp. 8754–8758. International Joint Conferences on Artificial Intelligence Organization (8 2024). https://doi.org/10.24963/ijcai.2024/1025, demo Track
37. Nguyen, H., Rahimi, A., Whitford, V., Fournier, H., Kondratova, I., Richard, R., Cao, H.: Heart2Mind: Human-Centered Contestable Psychiatric Disorder Diagnosis System using Wearable ECG Monitors. arXiv preprint arXiv:2505.11612 (2025)
38. Nguyen, H.T.T., Nguyen, L.P.T., Cao, H.: XEdgeAI: A human-centered industrial inspection framework with data-centric Explainable Edge AI approach. Information Fusion **116**, 102782 (2025). https://doi.org/10.1016/j.inffus.2024.102782
39. Nguyen, K., Nguyen, H., Nguyen, K., Truong, B., Phan, T., Cao, H.: Efficient and Concise Explanations for Object Detection with Gaussian-Class Activation Mapping Explainer. Proceedings of the Canadian Conference on Artificial Intelligence (may 27 2024)
40. Nguyen, T.T.H., Nguyen, V.T.K., Cao, Q.H., Truong, V.B., Nguyen, Q.K., Cao, H.: Enhancing the Fairness and Performance of Edge Cameras with Explainable AI. In: 2024 IEEE International Conference on Consumer Electronics (ICCE). pp. 1–4 (2024)
41. Nguyen, T.T.H., Truong, V.B., Nguyen, V.T.K., Cao, Q.H., Nguyen, Q.K.: Towards Trust of Explainable AI in Thyroid Nodule Diagnosis. In: International Workshop on Health Intelligence. pp. 11–26. Springer (2023)
42. Padmanabhan, D.C., Plöger, P.G., Arriaga, O., Valdenegro-Toro, M.: DExT: Detector Explanation Toolkit. In: Longo, L. (ed.) Explainable Artificial Intelligence. pp. 433–456. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-44067-0_23
43. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: British Machine Vision Conference (BMVC) (2018)

44. Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V.I., Mehra, A., Ordonez, V., Saenko, K.: Black-box Explanation of Object Detectors via Saliency Maps. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11438–11447 (2021). https://doi.org/10.1109/CVPR46437.2021.01128

45. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91

46. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 (2018)

47. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)

48. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)

49. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778

50. Sejr, J.H., Schneider-Kamp, P., Ayoub, N.: Surrogate Object Detection Explainer (SODEx) with YOLOv4 and LIME. Machine Learning and Knowledge Extraction **3**(3), 662–671 (2021). https://doi.org/10.3390/make3030033

51. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74

52. Song, L., Frazier, A., Kedron, P., Araujo, D., Cui, D., Enquist, B., Maitner, B., Merow, C., Moulatlet, G., Nikolopoulos, E., Roehrdanz, P.: Explainable artificial intelligence to interpret spatially-explicit impacts of future climate change on species distribution (01 2024)

53. Tahir, H.A., Alayed, W., Hassan, W.U., Haider, A.: A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration. Sensors **24**(21) (2024). https://doi.org/10.3390/s24216776

54. Tarmissi, K., Mengash, H.A., Negm, N., Said, Y., Al-Sharafi, A.M.: Explainable artificial intelligence with fusion-based transfer learning on adverse weather conditions detection using complex data for autonomous vehicles. AIMS Mathematics **9**(12), 35678–35701 (2024). https://doi.org/10.3934/math.20241693

55. Teng, Q., Liu, Z., Song, Y., Han, K., Lu, Y.: A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems **28**(6), 2335–2355 (Dec 2022). https://doi.org/10.1007/s00530-022-00960-4

56. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully Convolutional One-Stage Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9626–9635 (2019). https://doi.org/10.1109/ICCV.2019.00972

57. Truong, V.B., Nguyen, T.T.H., Nguyen, V.T.K., Nguyen, Q.K., Cao, Q.H.: Towards Better Explanations for Object Detection. In: Yanıkoğlu, B., Buntine, W. (eds.) Proceedings of the 15th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 222, pp. 1385–1400. PMLR (11–14 Nov 2024)

58. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks . In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 111–119. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2020). https://doi.org/10.1109/CVPRW50498.2020.00020

59. Wood, D.A., Kafiabadi, S., Busaidi, A.A., Guilhem, E., Montvila, A., Lynch, J., Townend, M., Agarwal, S., Mazumder, A., Barker, G.J., Ourselin, S., Cole, J.H., Booth, T.C.: Deep learning models for triaging hospital head MRI examinations. Medical Image Analysis **78**, 102391 (2022). https://doi.org/10.1016/j.media.2022. 102391

60. Yamauchi, T.: Spatial Sensitive Grad-CAM++: Improved Visual Explanation for Object Detectors via Weighted Combination of Gradient Map. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 8164–8168 (June 2024)

61. Yamauchi, T., Ishikawa, M.: Spatial Sensitive GRAD-CAM: Visual Explanations for Object Detection by Incorporating Spatial Sensitivity. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 256–260 (2022). https://doi.org/10. 1109/ICIP46576.2022.9897350

62. Yang, G., Ye, Q., Xia, J.: Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion **77**, 29–52 (2022). https://doi.org/10.1016/j.inffus.2021.07.016

63. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-Down Neural Attention by Excitation Backprop. International Journal of Computer Vision **126**(10), 1084–1102 (Oct 2018). https://doi.org/10.1007/s11263-017-1059-x

64. Zhang, Q., Rao, L., Yang, Y.: Group-CAM: Group Score-Weighted Visual Explanations for Deep Convolutional Networks. arXiv preprint arXiv:2103.13859 (2021)

65. Zhao, C., Hsiao, J.H., Chan, A.B.: Gradient-Based Instance-Specific Visual Explanations for Object Specification and Object Discrimination. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(9), 5967–5985 (2024). https://doi.org/10.1109/TPAMI.2024.3380604