ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA250809

D-CRISP: Explaining Object Detectors by Combining Randomized and Segment-Based Perturbations

Alain Andres[a,b,*] **and Javier Del Ser**[a,c]

[a]TECNALIA, Basque Research and Technology Alliance (BRTA), 20009 Donostia-San Sebastian, Spain
[b]University of Deusto, 20012 Donostia-San Sebastian, Spain
[c]Department of Mathematics, University of the Basque Country (UPV/EHU), 48940 Leioa, Spain

Abstract. Explaining the decisions issued by Machine Learning models for object detection tasks is essential in high-stakes decision making scenarios, such as medical image processing and vehicular perception for autonomous driving. Despite the proliferation of post-hoc perturbation-based methods for generating visual explanations, most eXplainable AI (XAI) approaches rely exclusively on either random image masking or selective segmentation-based occlusion, missing the opportunity to synergistically leverage both strategies in a complementary fashion. In this paper we address this gap by proposing D-CRISP (Detector-Combining Randomized Input and Segment Perturbations), a novel post-hoc explanation method for object detection models. D-CRISP unifies both random and region-based occlusions derived from image segmentation, producing multiscale saliency maps that capture both granular (pixel-level) and semantic (region-level) cues about the objects detected by the model. Experiments on the MS-COCO dataset show that D-CRISP significantly outperforms random-masking approaches in terms of explanation faithfulness and localization, while requiring slightly more computation effort than these methods. At the same time, it achieves comparable or better performance than segmentation-based methods, yet with substantially lower mask generation latencies. These results position D-CRISP as a highly effective and efficient XAI alternative for object detection models, particularly suited for time-constrained applications requiring timely, accurate, and interpretable decisions.

1 Introduction

Deep learning-based object detection models such as Faster R-CNN [17], DETR [6], and the YOLO series [23]– has driven significant progress in critical domains like autonomous driving, security surveillance, and industrial automation. However, the black-box nature of these models presents a major challenge in high-risk environments, where understanding the rationale behind object predictions is essential for establishing trust and ensuring their reliability in real-world settings [2, 14].

To address this challenge, the field of eXplainable Artificial Intelligence (XAI) has emerged, aiming to improve the interpretability of Artificial Intelligence (AI) systems and, in turn, foster the trust of their audience in their outputs [5]. XAI methods are broadly categorized into white-box and black-box (also known as *model-agnostic*) approaches. White-box methods require access to internal model components, such as weights, activations, or gradients. While they can provide deep insights, their reliance on specific architectures limits their generalizability. In contrast, black-box methods treat the model as an opaque system, generating explanations based solely on its input-output behavior, making them more versatile and applicable across a wider range of models and scenarios.

Within black-box XAI, perturbation-based techniques have become especially prominent due to their model-agnostic nature. Among them, RISE [15] and its adaptation for object detection, D-RISE [16], generate explanations by first applying binary masks sampled uniformly at random over the input image, and secondly by measuring the impact of such masks on model predictions. These methods are efficient and image-agnostic, making them well suited for large-scale or time-constrained scenarios. However, the resulting saliency maps may lack semantic structure, which can hinder their interpretability. More recently, some approaches such as D-MFPP [3], D-CLOSE [13] or BODEM [11] have proposed the use of segmentation-based masks to provide more human-interpretable and better spatially localized explanations. These methods leverage image segmentation algorithms to occlude semantically meaningful regions (e.g., superpixels or object parts), which can lead to better interpretable saliency maps. Nevertheless, this gain in interpretability comes at a penalty: segmentation is inherently image-dependent, introduces computational overhead and limits scalability in real-time or resource-constrained environments. Moreover, segmentation masks may inadvertently group both relevant and irrelevant pixels into a single region, which can degrade the spatial precision of the explanations.

Despite the complementary advantages of these two masking strategies (the efficiency and generality of uniform random mask, and the semantic coherence of segmentation-based masks), existing perturbation-based approaches that hinge on masking use exclusively one or the other. To the best of our knowledge, no previous work has explored unified strategies that combine both types of perturbations in a principled way. This paper covers this research gap by introducing **D-CRISP** (***D**etector-**C**ombining **R**andomized **I**nput and **S**egment **P**erturbations*), a novel post-hoc XAI method for object detection models. D-CRISP integrates random occlusions with segmentation-based occlusions to generate multiscale saliency maps that reflect both fine-grained (pixel-level) and semantic (region-level) relevance. The use of random masks helps mitigate the tendency of segmentation-based approaches to group both relevant and irrelevant pixels into a single region, thereby improving the spatial accuracy of

* Corresponding Author:
alain.andres@tecnalia.com (https://aklein1995.github.io/)

the explanations. At the same time, D-CRISP maintains low computational cost by reusing image-agnostic random masks and limiting the number of segmentation masks per image, making it especially suitable for large-scale or time-sensitive applications. The key features of D-CRISP can be summarized as follows:

- *Model-agnostic*: D-CRISP works with any object detector, requiring only the model's predicted outputs.
- *Hybrid masking*: D-CRISP simultaneously leverages the efficiency of random perturbations and the semantic richness of segmentation-based masks.
- *Multiscale saliency*: D-CRISP produces explanations that operate at multiple granularity levels, enhancing the interpretability of the identified semantic predictors.
- *Low computational overhead*: By reusing image-agnostic masks and limiting per-image segmentation masks, D-CRISP is faster than segmentation-only baselines.

We validate our method on the MS-COCO dataset to show that D-CRISP produces explanations that are more robust, better aligned with ground-truth bounding boxes, and less sensitive to minor image transformations than existing perturbation-based XAI methods. Moreover, our experiments confirm that D-CRISP requires less computation time than segmentation-based approaches, making it well suited for deployment in time-sensitive applications.

The rest of the manuscript is organized as follows: Section 2 briefly revisits related work in XAI for object detection. Next, Section 3 details the proposed D-CRISP method. Section 4 presents the experimental setup, including datasets, object detection models, and evaluation metrics. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes the paper with a summary of our findings and directions for future research.

2 Related Work

Prior to detailing the design of the proposed D-CRISP method, we review the literature on XAI for image classification and object detection, thereby positioning the contributions of this work within the context of existing research.

Explainable AI for Image Classification. Recent advancements in XAI for classification have introduced a variety of methods aimed at improving the interpretability of model predictions. Traditional gradient-based approaches, such as Grad-CAM [21] and Score-CAM [24], provide class activation maps by leveraging model gradients, but they are inherently model-dependent. This might prevent end-user their adoption due to the lack of expertise in complex neural network architectures.

To address this limitation, LIME [18] pioneered a model-agnostic approach by training local interpretable surrogate models around each prediction, offering human-understandable explanations independent of model architecture. With a similar idea, Occlusion [28] and RISE [15] introduced model-agnostic perturbation-based techniques that generate saliency maps by passing masks created in different ways (e.g. randomized) through the model and analyzing output variations. Similarly, MFPP [27] proposed the use of morphological fragments at multiple scales to occlude semantically meaningful regions, thereby refining the attribution process. Further improving on multi-scale feature extraction, SISE [19] introduced a multi-layer aggregation strategy, where feature maps from different layers are combined through a pointwise multiplicative fusion, iteratively normalized to produce a sharper and more reliable explanation. Ada-SISE [22] extends this concept by adaptively selecting the most important layers per input, making the feature aggregation process more dynamic and input-specific.

Considering that most perturbation techniques are computationally intensive, SeCAM [12] sought a balance between fidelity and efficiency, achieving interpretability levels comparable to LIME [18] while maintaining inference speeds close to CAM-based approaches. In addition, HiPe [8] proposed a hierarchical perturbation strategy by perturbing only the most influential regions with increasing resolution and discarding irrelevant areas early, which reduces explanation generation time by an order of magnitude. Nevertheless, the majority of these methods continue to face challenges in generating explanations confined to small, interpretable regions. This limitation has recently been addressed in [7] through the reformulation of the attribution task as a submodular subset selection problem.

Explainable AI in Object Detection. Explainable AI for object detection has also gained significant attention, focusing on adapting saliency and attribution methods to the more complex setting of localizing multiple objects within an image. A pioneering effort in this area was D-RISE [16], which extended the RISE framework to object detectors such as YOLOv3 by introducing a new similarity score that accounts for both classification and localization performance, thereby providing richer and more detection-specific explanations. Concurrently, SODEx [20] proposed an abstract algorithm that first converts the outputs of any object detector into binary classification tasks via a surrogate classifier, and then applies LIME to generate explanations. While offering broader model compatibility, SODEx tends to be less generalizable and robust compared to D-RISE due to its reliance on external surrogate models.

Building upon the strengths of D-RISE, D-CLOSE [13] combined the use of segmentation-based masks with the multi-layer fusion strategy originally introduced in SISE. Similarly, D-MFPP [3] extended MFPP to the object detection domain, leveraging multi-scale morphological perturbations to boost explanation quality and outperform earlier object detection XAI approaches. In a slightly different direction, GSM-HM [25], MAPSM [26] and BODEM [11] proposed hierarchical masking strategies that generate multiple levels of saliency maps, iteratively refining them to minimize the number of perturbations needed for high-quality explanations. Although these approaches leverage multi-stage refinement, a key distinction is that BODEM eliminates the dependency on the objectness scores of detections, requiring only the bounding box coordinates to generate explanations, thereby enhancing its general applicability.

Contribution. Existing explainability methods for object detection face a trade-off between efficiency and semantic quality. Random masking approaches such as D-RISE offer a fast, model-agnostic way to probe model behavior but often suffer from poor spatial and semantic coherence. In contrast, segmentation-based occlusion methods like D-CLOSE and D-MFPP operate over meaningful image regions aligned with object structures, but rely heavily on image content. Hierarchical strategies such as MAPSM and BODEM further refine explanations across multiple levels, but require computing a saliency map at each stage, leading to repeated forward passes through the object detector (with its subsequent additional inference time).

The method proposed in this work, D-CRISP, unifies the strengths of random and segmentation-based strategies by applying both perturbations jointly and aggregating their contributions into a single saliency representation. Crucially, D-CRISP generates the random masks only once and reuses them across all images, substan-

tially reducing the overall computational cost while maintaining semantic richness and precision in the explanations, leveraging the segmentation-based masks specifically generated for each image.

3 Proposed D-CRISP Framework: Explaining Object Detectors with Hybrid Masks

We now proceed by presenting the algorithmic details of our proposed local post-hoc explanation method for object detection models. Figure 1 illustrates the overall framework of D-CRISP. As shown in this plot, D-CRISP leverages both random binary masks and segmentation-based masks to generate saliency maps that highlight the most influential regions influencing the detections. In what follows we describe the procedure for generating such masks (Subsection 3.1) and computing the saliency map (Subsection 3.2).

3.1 Mask Generation

As in many perturbation-based approaches, D-CRISP requires specifying the total number of masks N to be applied per image, as well as the percentage p of pixels to be occluded in each mask. Additionally, we introduce a new hyperparameter $\alpha \in [0, 1]$, which defines the proportion of random versus segmentation-based masks used during explanation generation. Specifically, we compute:

$$N_r = \alpha \cdot N, \quad N_s = N - N_r = (1 - \alpha) \cdot N, \tag{1}$$

where N_r and N_s denote the number of random and segmentation-based masks, respectively.

Random Masks. We follow the masking strategy introduced in RISE [15] to generate a set of N_r random masks, independently of the input images. Each mask is denoted as $\mathbf{M}_r \in [0, 1]^{h \times w}$, where $r = 1, \ldots, N_r$ refers to the index within the set of precomputed random masks. These masks are generated by sampling low-resolution binary grids and upsampling them to the input resolution using bilinear interpolation. Then, given an input image $\mathbf{X}_i$, we apply the N_r precomputed random masks $\mathbf{M}_r$ to generate a set of perturbed images:

$$\mathbf{X}'_{i,r} = \mathbf{X}_i \odot \mathbf{M}_r, \tag{2}$$

where $\odot$ denotes element-wise (Hadamard) product. Since the random masks $\mathbf{M}_r$ are *image-agnostic*, they are generated once and subsequently reused across all images, resulting in significant computational savings. As the number of random masks N_r is controlled by the hyperparameter $\alpha \in [0, 1]$ introduced in Equation (1)– which determines the proportion of random masks relative to the total number of masks N– higher values of α lead to an increased reuse of precomputed masks, thereby reducing the computational cost per image.

Segmentation-based Masks. In parallel, we compute N_s segmentation-based masks $\mathbf{M}_{i,s} \in [0, 1]^{h \times w}$ for each input image $\mathbf{X}_i$ using a segmentation algorithm (e.g., SLIC [1]). Each segmented region—commonly referred to as a *superpixel*—defines a binary region within the image. For each mask, a subset of superpixels is randomly selected such that approximately p-percentage of the image's pixels are occluded. These selected regions are set to zero (masked out), while the rest of the image is preserved. As a result, we obtain a set of perturbed images from segmentation masks:

$$\mathbf{X}'_{i,s} = \mathbf{X}_i \odot \mathbf{M}_{i,s}, \tag{3}$$

where $s = 1, \ldots, N_s$ indexes the segmentation-based masks computed specifically for image $\mathbf{X}_i$.

3.2 Saliency Map Computation

To estimate the relevance of each pixel for a given detected object within the input image $\mathbf{X}_i$, we apply the object detector to each perturbed image and evaluate how it affects the prediction corresponding to the detected object. As introduced earlier, we denote as $\{\mathbf{X}'_{i,r}\}_{r=1}^{N_r}$ and $\{\mathbf{X}_{i,s}\}_{s=1}^{N_s}$ the perturbed versions of input image $\mathbf{X}_i$ generated using random and segmentation-based masks, respectively. For each perturbed image, we obtain a set of detection proposals by passing it through the object detector model $f(\cdot) : \mathbb{R}^{3 \times h \times w} \mapsto \mathcal{D}$, where $\mathcal{D} = \{\mathbf{D}_j\}$ stands for a set of detection proposals $\mathbf{D}_j$ defined as:

$$\mathbf{D}_j = (\mathbf{BB}_j, \mathbf{P}_j, O_j), \tag{4}$$

where $\mathbf{BB}_d \in \mathbb{R}[0, 1]^4$ denotes the bounding box coordinates of the detected object (normalized w.r.t. the size of the input image $h \times w$); $\mathbf{P}_j \in \mathbb{R}[0, 1]^C$ is the class probability vector associated to the object in the bounding box (with C denoting the number of known classes, $\mathbf{P}_j = \{P_j(c)\}_{c=1}^{C}$ and $\sum_{c=1}^{C} P_j(c) = 1$); and $O_j \in \mathbb{R}[0, 1]$ denotes the objectness score associated with the detection proposal. Detection proposals generated by detector $f(\cdot)$ from the masked images are then given by:

$$\mathcal{D}_{i,r} = f(\mathbf{X}'_{i,r}), \quad \mathcal{D}_{i,s} = f(\mathbf{X}'_{i,s}), \tag{5}$$

namely, multiple detection proposals are extracted for each perturbed image. Similarly to D-RISE [16], we compute a similarity score between each proposal $\mathbf{D}_j \in \mathcal{D}_{i,r} \cup \mathcal{D}_{i,s}$ and the target detected object $\mathbf{D}_T = (\mathbf{BB}_T, \mathbf{P}_T, O_T)$ for which a saliency map explanation is sought. This similarity score is computed as:

$$\text{sim}(\mathbf{D}_T, \mathbf{D}_j) = \text{IoU}(\mathbf{BB}_T, \mathbf{BB}_j) \cdot \text{cosine}\,(\mathbf{P}_T, \mathbf{P}_j) \cdot O_j, \tag{6}$$

where the components represent spatial overlap (Intersection over Union of the detected bounding boxes), class similarity (cosine similarity between class probability vectors) and detection confidence (objectness of each detection in the perturbed image). Multiple proposals may be returned for a given perturbed image $\mathbf{X}'_{i,r}$ or $\mathbf{X}'_{i,s}$ (i.e. $|\mathcal{D}_{i,r}| \geq 1$ and/or $|\mathcal{D}_{i,s}| \geq 1$ for a given r and s), for each perturbed image we retain only the one with the highest similarity to $\mathbf{D}_T$, yielding an N-sized set of similarity scores $\{\text{S}_{T,n}\}_{n=1}^{N}$ associated with perturbed image $\mathbf{X}_i$, detected object $\mathbf{D}_T$ and mask $\mathbf{M}_n \in \{\mathbf{M}_r, \mathbf{M}_s\}$ ($n = 1, \ldots, N$), computed as:

$$\text{S}_{T,n} = \max_{\mathbf{D}_j \in \mathcal{D}_n} \text{sim}(\mathbf{D}_T, \mathbf{D}_j), \tag{7}$$

where $\mathcal{D}_n$ denotes the detection proposals associated with the n-th masked image $\mathbf{X}'_{i,s}$ ($\mathcal{D}_n \equiv \mathcal{D}_{i,s}$) or $\mathbf{X}'_{i,r}$ ($\mathcal{D}_n \equiv \mathcal{D}_{i,r}$); and $S_{T,n}$ is the final relevance score assigned to mask $\mathbf{M}_n$ for detection $\mathbf{D}_T$. Finally, these scores are used to construct a $h \times w$ saliency map $\text{SM}(\mathbf{D}_T)$ corresponding to detection $\mathbf{D}_T$ as a weighted combination of masks:

$$\text{SM}(\mathbf{D}_T) = \frac{1}{N} \sum_{n=1}^{N} S_{T,n} \cdot \mathbf{M}_n. \tag{8}$$

4 Experimental Setup

In order to assess the performance of the proposed D-CRISP framework, we design an experimental setup comprising datasets and models (Subsection 4.1), baselines and hyperparameter configuration (Subsection 4.2), and explanation quality metrics (Subsection 4.3). The code with which all experiments have been carried out can be found in https://github.com/aklein1995/D-CRISP.

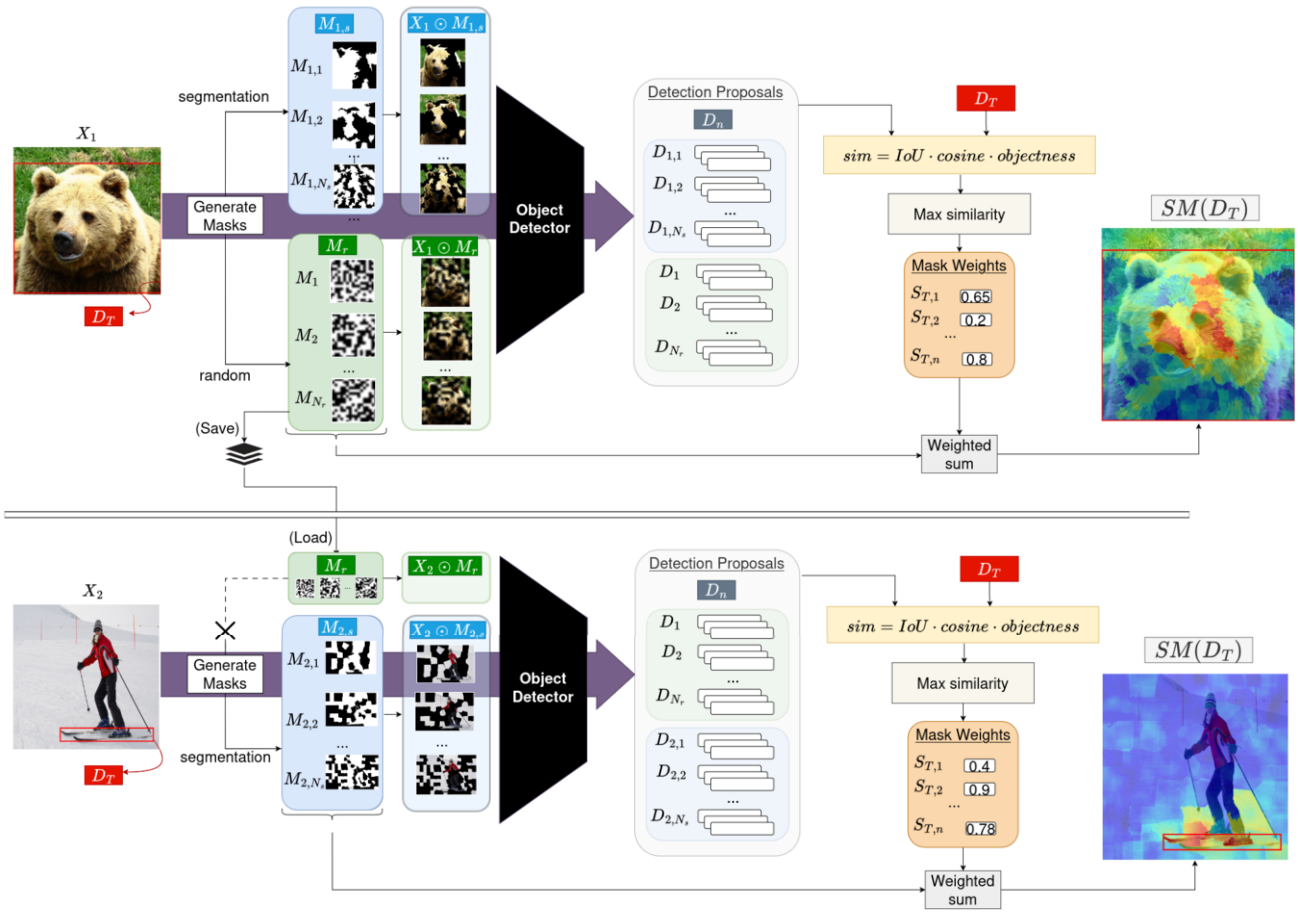


Figure 1: General diagram of the proposed D-CRISP framework. First, N_r random binary masks and N_s segmentation-based masks are generated for the input image $\mathbf{X}_1$. In subsequent images (e.g., $\mathbf{X}_2$), the previously generated random masks $\mathbf{M}_r$ (green background), are re-used and only the segmentation-based masks $\mathbf{M}_s$ have to be generated.

4.1 Dataset and Model

Following prior work, we evaluate our method on the MS-COCO validation dataset [10], a widely used benchmark in object detection explainability. The dataset consists of 5,000 annotated images containing diverse object categories and scene contexts.

Our proposed D-CRISP method is model-agnostic and can be applied to any object detector that outputs bounding box coordinates ($\mathbf{BB}$), objectness score (O), and/or class probabilities ($\mathbf{P}$). Previous studies in object detection explainability have employed different versions of the YOLO architecture [23], including YOLOv3 [16], YOLOv4 [20], YOLOv5 [11, 26], YOLOX [13], and YOLOv8 [3]. Following these preceding studies and given the computational demands of XAI evaluations, we limit our experimentation to YOLOv8, which offers a suitable trade-off between predictive accuracy and efficiency[1]. Moreover, only detections surpassing a confidence threshold of 0.7 are considered.

4.2 Baselines and Hyperparameter Configuration

In order to evaluate our proposal, we compare the performance against current state of the art model-agnostic XAI methods for object detection models, namely, D-RISE [16], D-CLOSE [13] and D-MFPP [3]. For D-RISE we use the parameters proposed in the original approach with $N = 5{,}000$ masks, probability $p = 0.5$ and a resolution of $h \times w = 16 \times 16$ pixels. For methods that rely on segmentation-based masking (D-CLOSE and D-MFPP), we use SLIC [1] to split the image into 5 levels with $[50, 100, 200, 400, 800]$ superpixels, as recommended in prior work [27].

[1] We note that D-CRISP can be directly applied to any object detection model providing the outputs collected in $\mathbf{D}$ as per Expression (4).

4.3 Explanation Quality Metrics

To quantitatively evaluate the quality of the generated explanations, we employ metrics that fall into two main categories: faithfulness and localization.

Faithfulness This criterion measures how well the explanation reflects the model's actual behavior, based on the assumption that more important features should have a stronger impact on the output [9]. To this end, the most important features given by the saliency map can be iteratively be removed with *Deletion* [4, 15], which measures the decrease in the class probability of the proposal detections given by the model as more pixels are deleted. More specifically, *D-Deletion* [3] evaluates the probability decay conditioned on the IoU with respect to the target detection. Consequently, lower *D-Deletion* scores are preferred, indicating that the removal of key features significantly reduces the model's confidence. Conversely, *Insertion* [15] assesses how quickly the confidence increases when important pixels are gradually restored. Akin to D-Deletion, D-Insertion is con-

ditioned to have an IoU>0 with the target detection $\mathbf{D}_T$. In this case, higher *D-Insertion* scores are desirable, showing that restoring important areas quickly improves the prediction strength.

Localization This family of metrics gauges whether the explanation is spatially aligned with a human-understood region of interest. *Pointing Game* (PG) [29] checks if the most important pixel features by the XAI method falls within the target bounding box. Here, higher *PG* scores mean better localization, aligning saliency with the true object region. *Energy-based pointing game* (EBPG) [24] generalizes this notion by summing the total importance mass that falls within the ground-truth box. Similarly, higher *EBPG* scores are preferred, indicating stronger focus on the object.

In the following results we make use of D-Deletion, D-Insertion, PG and EBPG to quantitatively analyze the outcomes.

5 Results and Discussion

We present and discuss the results from the experimental setup, evaluating the performance of D-CRISP and comparing it against state-of-the-art post-hoc explainability methods designed for object detection models. We organize our discussion in five parts: quality of the explanations produced by D-CRISP and the comparison baselines (Subsection 5.1), impact of bounding box size (Subsection 5.2), robustness of D-CRISP with varying number of masks (Subsection 5.3), and computational efficiency (Subsection 5.4).

5.1 Explanation Quality Comparison

We begin by analyzing the explanation quality metrics presented in Table 1. This table shows D-Deletion, D-Insertion, PG and EBPG scores averaged over all detections produced by YOLOv8 over the images within the MS-COCO validation dataset for all explainability methods considered in our benchmark. As can be observed in this table, D-RISE performs worst across all metrics, while D-MFPP consistently achieves top performance, particularly in D-Deletion and PG. D-CLOSE obtains competitive results, especially in EBPG, where it reaches an outstanding value of 48.383. This is due to its feature fusion strategy, which assigns a higher weight to the saliency maps produced at coarser segmentation levels—those with fewer and larger superpixels[2]. Our proposed D-CRISP technique performs competitively, especially for $\alpha = 0.25$ (i.e. 25% of the masks are generated at random). In this latter case, D-CRISP outperforms D-CLOSE in D-Deletion (0.119 vs. 0.139, a 15% improvement) and achieves the best D-Insertion score overall.

Table 1: Micro-average performance of each method across all detections in the MS-COCO validation set. The best score for each metric across all XAI methods is highlighted in gray.

Method	D-Deletion (↓)	D-Insertion (↑)	PG (↑)	EBPG (↑)
D-RISE	0.228	0.831	0.911	13.211
D-MFPP	0.118	0.858	0.984	14.412
D-CLOSE	0.139	0.856	0.974	48.383
D-CRISP ($\alpha = 0.25$)	0.119	0.859	0.974	13.832
D-CRISP ($\alpha = 0.50$)	0.125	0.857	0.973	13.555
D-CRISP ($\alpha = 0.75$)	0.144	0.854	0.968	13.367

To mitigate the potential bias introduced by class imbalance, Table 2 reports the explanation quality metrics using a macro-average: metrics are first computed per class (i.e., averaging over detections belonging to each object category) and then averaged across all classes. This adjustment is motivated by the strong imbalance in the MS-COCO validation set, where a few frequent categories dominate the detection count[3]. Based on this evaluation, D-CRISP ($\alpha = [0.25, 0.5]$) yields the best D-Deletion score (0.071), even outperforming D-MFPP. PG results remain comparable, with D-CRISP maintaining a small gap from the top-performing D-MFPP (0.972 vs. 0.981). This exposes the robustness of explanations issued by our D-CRISP method across different object categories.

Table 2: Macro (class-wise) average performance across object categories to account for class imbalance. The best score for each metric is highlighted in gray.

Method	D-Deletion (↓)	D-Insertion (↑)	PG (↑)	EBPG (↑)
D-RISE	0.101	0.822	0.931	13.464
D-MFPP	0.075	0.859	0.981	14.887
D-CLOSE	0.088	0.849	0.967	49.310
D-CRISP ($\alpha = 0.25$)	0.071	0.851	0.972	14.159
D-CRISP ($\alpha = 0.50$)	0.071	0.850	0.971	13.849
D-CRISP ($\alpha = 0.75$)	0.074	0.847	0.967	13.641

5.2 Impact of Object Size

Given the variability observed between the comparative results reported in Tables 1 and 2, we now investigate the behavior of the methods in the benchmark when dealing with objects of different sizes. To this end, akin to the analysis made in D-CLOSE [13], we collect all bounding box proposals generated by YOLOv8 over the entire MS-COCO validation dataset, and categorize them using k-means into three groups based on their spanned area:

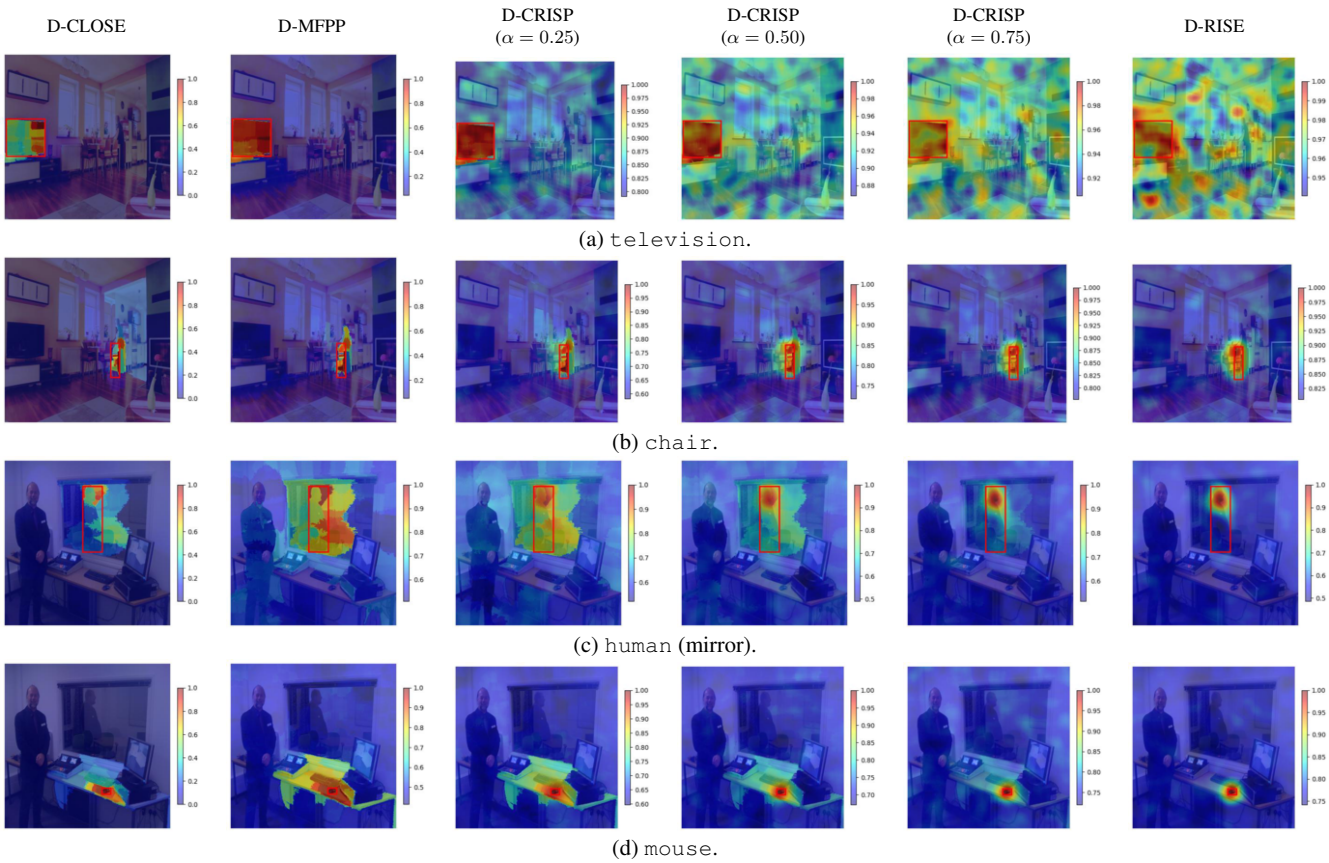
- `Small`, composed by 10,593 detections with areas ranging from 64 to 66,636 square pixels.
- `Medium`, which includes 2,594 detections with areas between 66,750 and 196,004 square pixels.
- `Large`, which comprises 867 detections with areas between 196,372 and 406,894 square pixels.

Interestingly, the distribution of detected bounding boxes is skewed towards smaller objects. Based on these categories, Table 3 presents the micro-averaged explanation quality metrics grouped by category. As can be observed in this table, D-CRISP with $\alpha = 0.25$ achieves the best D-Deletion score on `small` objects (0.077), improving upon D-MFPP while maintaining competitive results for `medium` and `large` objects.

To further illustrate these results, Figures 2.a to 2.d (next page) depict the saliency maps for representative cases of 4 of the classes present in the MS-COCO validation dataset: `television`, `chair`, `human` and `mouse`. For small objects such as the `chair` and the `mouse`, segmentation-based methods can potentially over-highlight irrelevant surrounding regions. In contrast, for larger objects like the `television`, these methods focus on informative areas more accurately. A particularly revealing example is the `human` detection, which actually corresponds to a reflection in a mirror. In this case, segmentation-based methods tend to highlight the entire mirror, even

[2] Since EBPG favors broad saliency spread within the ground-truth bounding box, this emphasis can disproportionately inflate its score, even if fine-grained localization is less precise.

[3] Out of a total of 36,781 detections across 5,000 validation images, 11,004 detections correspond to class `person`, 1,932 to `car`, 1,791 to `chair`, and 1,161 to `book`. These four categories account for approximately 43% of all detections (15,888/36,781), with `person` alone representing ∼30%.



(a) `television`.

(b) `chair`.

(c) `human` (mirror).

(d) `mouse`.

Figure 2: Saliency maps generated by D-CLOSE, D-MFPP, D-CRISP, and D-RISE for 4 object instances across two different images of MS-COCO validation dataset (IDs: '000000000139' and '000000009483'): a `television` (`large` object category), a `chair` (`small` object category), a `human` reflection in a mirror (`large` category), and a `mouse` (`small` object).

Table 3: Micro-average performance grouped by object size (`small`, `medium`, `large`). The best score for each metric within each size category is highlighted in gray.

Category	Method	D-Deletion (↓)	D-Insertion (↑)	PG (↑)	EBPG (↑)
Small	D-RISE	0.175	0.826	0.941	4.906
	D-MFPP	0.078	0.849	0.984	5.944
	D-CLOSE	0.098	0.840	0.971	36.683
	D-CRISP ($\alpha = 0.25$)	0.077	0.850	0.973	5.447
	D-CRISP ($\alpha = 0.50$)	0.081	0.849	0.976	5.207
	D-CRISP ($\alpha = 0.75$)	0.096	0.847	0.973	5.038
Medium	D-RISE	0.380	0.854	0.810	28.775
	D-MFPP	0.228	0.892	0.979	30.650
	D-CLOSE	0.253	0.892	0.982	73.549
	D-CRISP ($\alpha = 0.25$)	0.232	0.892	0.973	29.728
	D-CRISP ($\alpha = 0.50$)	0.242	0.890	0.960	29.926
	D-CRISP ($\alpha = 0.75$)	0.275	0.886	0.947	29.017
Large	D-RISE	0.408	0.815	0.863	65.629
	D-MFPP	0.267	0.867	0.995	66.662
	D-CLOSE	0.288	0.867	0.992	88.582
	D-CRISP ($\alpha = 0.25$)	0.277	0.862	0.985	66.114
	D-CRISP ($\alpha = 0.50$)	0.293	0.858	0.983	65.879
	D-CRISP ($\alpha = 0.75$)	0.325	0.849	0.973	65.739

though the person occupies less than approx. 30% of its surface. Interestingly, D-RISE focuses more on the head, suggesting a better localized and semantically more meaningful explanation.

5.3 *Robustness of D-CRISP with Fewer Masks*

We further evaluate the robustness of each XAI method under a low-mask regime by reducing the number of masks N from 5,000 to 500. Table 4 shows that although all methods experience some performance degradation, D-CRISP remains highly competitive. For instance, D-CRISP ($\alpha = 0.25$) shows a moderate drop in *PG*, from 0.974 to 0.915, yet still outperforms D-RISE using 5,000 masks in multiple metrics. This suggests that D-CRISP can maintain high-quality explanations even when computational resources are limited. In contrast, D-RISE suffers a substantial performance degradation, particularly in terms of *PG* (from 0.911 to 0.736) and *D-Insertion* (from 0.831 to 0.802), underlining its sensitivity to the number of masks. On the other hand, segmentation-based methods such as D-CLOSE and D-MFPP degrade more gracefully. Notably, D-CLOSE still achieves the highest *EBPG* score with only 500 masks (38.090),

Table 4: Performance comparison using a reduced number of masks ($N = 500$) versus the standard setting ($N = 5,000$). The table highlights the robustness of each method under limited mask budgets.

Method	N	D-Deletion (↓)	D-Insertion (↑)	PG (↑)	EBPG (↑)
D-RISE	5,000	0.228	0.831	0.911	13.211
	500	0.255	0.802	0.736	13.215
D-MFPP	5,000	0.118	0.858	0.984	14.412
	500	0.137	0.853	0.968	14.362
D-CLOSE	5,000	0.139	0.856	0.974	48.383
	500	0.146	0.853	0.961	38.090
D-CRISP ($\alpha = 0.25$)	5,000	0.119	0.859	0.974	13.832
	500	0.139	0.850	0.915	13.818
D-CRISP ($\alpha = 0.50$)	5,000	0.125	0.857	0.973	13.555
	500	0.149	0.845	0.921	13.556
D-CRISP ($\alpha = 0.75$)	5,000	0.144	0.854	0.968	13.367
	500	0.181	0.837	0.819	13.356

although this represents a noticeable decrease from its 5,000-mask score (48.383).

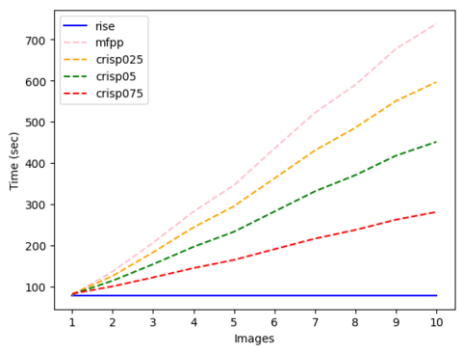
Notably, D-RISE, despite utilizing 10 times more masks, is outperformed by all segmentation-based approaches that use only 500 masks. This performance gap arises from RISE's image-agnostic design: its masks are generic and do not adapt to the specific scene content, which may include detections of varying sizes [13]. In contrast, segmentation-based masks are derived from the image's actual structure, enabling them to more precisely target relevant regions and produce more informative saliency maps under diverse conditions.

5.4 Computation Time and Efficiency

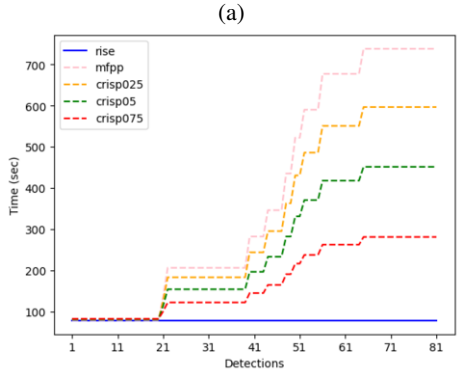
While D-MFPP consistently ranks among the top-performing methods in terms of explanation quality, its main limitation is the computational overhead involved in generating segmentation-based masks for each new image. In contrast, D-CRISP benefits from the ability to precompute image-agnostic random masks, which can be reused across samples, substantially reducing runtime.

Figure 3 shows the cumulative time required to compute the masks for the first 10 images in the MS-COCO validation set. All runtime measurements were obtained using a Tesla V100-SXM2-16GB GPU and an Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz. Although generating RISE-style random masks is time-consuming ($\sim$ 78 seconds), the cost is quickly amortized by reusing the same masks across multiple detections and images.

Figure 3.a considers the case where each image contains a single detection to be explained. The plotted curves assume that random masks are computed once, as they are image-agnostic and can be reused across detections and images. In contrast, segmentation-based masks must be recalculated for each image, since they depend on image content, resulting in a higher cumulative computational cost.



(a)



(b)

Figure 3: (a) Cumulative time required to compute $N = 5,000$ masks for the first 10 images in the validation set of MS-COCO. (b) Cumulative time required for computing these masks per detection, considering the actual number of detections for the 10 images ($[20, 1, 18, 4, 4, 2, 2, 4, 9, 17]$). Here we do not account for the time needed to calculate the saliency map.

However, this assumption rarely holds in real-world applications, where a single image typically contains multiple detections to be explained. This is reflected in Figure 3.b, which shows a more realistic case considering the actual number of detections per image (e.g., [20, 1, 18, 4, 4, 2, 2, 4, 9, 17]). While segmentation-based masks can be reused across detections within the same image – leading to some amortization of its computational cost – their cost still scales with the number of images, as new masks must be generated for each image. By contrast, random masks are generated once and reused globally, making them substantially more efficient in large-scale settings.

This analysis underscores D-CRISP's advantage: it combines both strategies to provide a balanced trade-off between computation and explanation quality, making it suitable for applications requiring efficient processing of multiple detections across many images.

6 Conclusions

In this paper we have introduced D-CRISP, a novel post-hoc local explanation method for object detection models that combines random and segmentation-based perturbations to generate multiscale saliency maps. By systematically combining random and segmentation-based masks, our approach bridges the gap between computationally efficient but coarse methods like D-RISE, and semantically richer but costly methods such as D-MFPP and D-CLOSE.

An extensive evaluation of the proposed techniques and other baselines on the MS-COCO validation set has exposed that segmentation-based methods generally provide more robust and interpretable explanations, thanks to their adaptability to different object shapes and scales. However, their requirement of per-image segmentation significantly increases the computational overhead, particularly when explanations must be generated for a large number of images or detections. On the other hand, D-RISE, while computationally cheaper due to its image-agnostic design, fails to adapt to object-specific characteristics, resulting in poorer performance—especially in settings with high variability in object size and structure. Our proposed method, D-CRISP, offers a compromise between these extremes, remaining competitive even with a low number of masks and performing particularly well on small objects.

While D-CRISP may not always outperform all methods for every object type, its tunable α parameter offers a controllable trade-off between random and segmentation-based masking. This flexibility enables D-CRISP to balance precision and semantic structure based on the context and/or size of each detection. Consequently, in the future we plan to explore strategies to dynamically adjust the α parameter based on properties of the input or target object (e.g. size, shape). Additionally, we envision that generating RISE masks at multiple resolutions can potentially enhance the quality of explanations delivered by D-CRISP to varying object scales.

Acknowledgements

The authors acknowledge funding support from FaRADAI project (ref. 101103386) funded by the European Commission under the European Defence Fund (EDF-2021-DIGIT-R). Their work is also supported by the Basque Government through the consolidated research group MATHMODE (ref. IT1456-22).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels. *EPFL Technical Report 149300*, 2010.

[2] I. Ahmed, G. Jeon, and F. Piccialli. From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8):5031–5042, 2022.

[3] A. Andres, A. Martinez-Seras, I. Laña, and J. Del Ser. On the black-box explainability of object detection models for safe and trustworthy industrial applications. *Results in Engineering*, 24:103498, 2024.

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

[5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with Transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[7] R. Chen, H. Zhang, S. Liang, J. Li, and X. Cao. Less is more: Fewer interpretable region via submodular subset selection. In *International Conference on Learning Representations (ICLR)*, 2024.

[8] J. Cooper, O. Arandjelović, and D. J. Harrison. Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping. *Pattern Recognition*, 129:108743, 2022.

[9] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. 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.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 13, pages 740–755, 2014.

[11] M. Moradi, K. Yan, D. Colwell, M. Samwald, and R. Asgari. Model-agnostic explainable artificial intelligence for object detection in image data. *Engineering Applications of Artificial Intelligence*, 137:109183, 2024.

[12] P. X. Nguyen, H. Q. Cao, K. V. Nguyen, H. Nguyen, and T. Yairi. SeCAM: Tightly accelerate the image explanation via region-based segmentation. *IEICE Transactions on Information and Systems*, 105(8): 1401–1417, 2022.

[13] T. T. H. Nguyen, V. T. K. Nguyen, Q. K. Nguyen, Q. H. Cao, et al. Towards better explanations for object detection. In *Asian Conference on Machine Learning*, pages 1385–1400, 2024.

[14] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, and E. 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*, page 1139–1150, New York, NY, USA, 2023.

[15] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized input sampling for explanation of black-box models. *arXiv:1806.07421*, 2018.

[16] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko. Black-box explanation of object detectors via saliency maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452, 2021.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you? explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[19] S. Sattarzadeh, M. Sudhakar, A. Lem, S. Mehryar, K. N. Plataniotis, J. Jang, H. Kim, Y. Jeong, S. Lee, and K. Bae. Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 11639–11647, 2021.

[20] J. H. Sejf, P. Schneider-Kamp, and N. Ayoub. Surrogate object detection explainer (SODEx) with YOLOv4 and LIME. *Machine Learning and Knowledge Extraction*, 3:662–671, 2021.

[21] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Why did you say that? *arXiv:1611.07450*, 2016.

[22] M. Sudhakar, S. Sattarzadeh, K. N. Plataniotis, J. Jang, Y. Jeong, and H. Kim. Ada-SISE: adaptive semantic input sampling for efficient explanation of convolutional neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1715–1719, 2021.

[23] J. Terven, D.-M. Cordova-Esparza, and J.-A. Romero-Gonzalez. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023.

[24] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[25] Y. Yan, X. Li, Y. Zhan, L. Sun, and J. Zhu. GSM-HM: generation of saliency maps for black-box object detection model based on hierarchical masking. *IEEE Access*, 10:98268–98277, 2022.

[26] Y. Yan, T. Jiang, X. Li, L. Sun, J. Zhu, and J. Lin. Model-agnostic progressive saliency map generation for object detector. *Image and Vision Computing*, 145:104988, 2024.

[27] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu. MFPP: Morphological fragmental perturbation pyramid for black-box model explanations. In *International Conference on Pattern Recognition (ICPR)*, pages 1376–1383, 2021.

[28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, volume 13, pages 818–833, 2014.

[29] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.