BOSTON UNIVERSITY

Adobe

Black-Box Explanation of Object Detectors via Saliency Maps

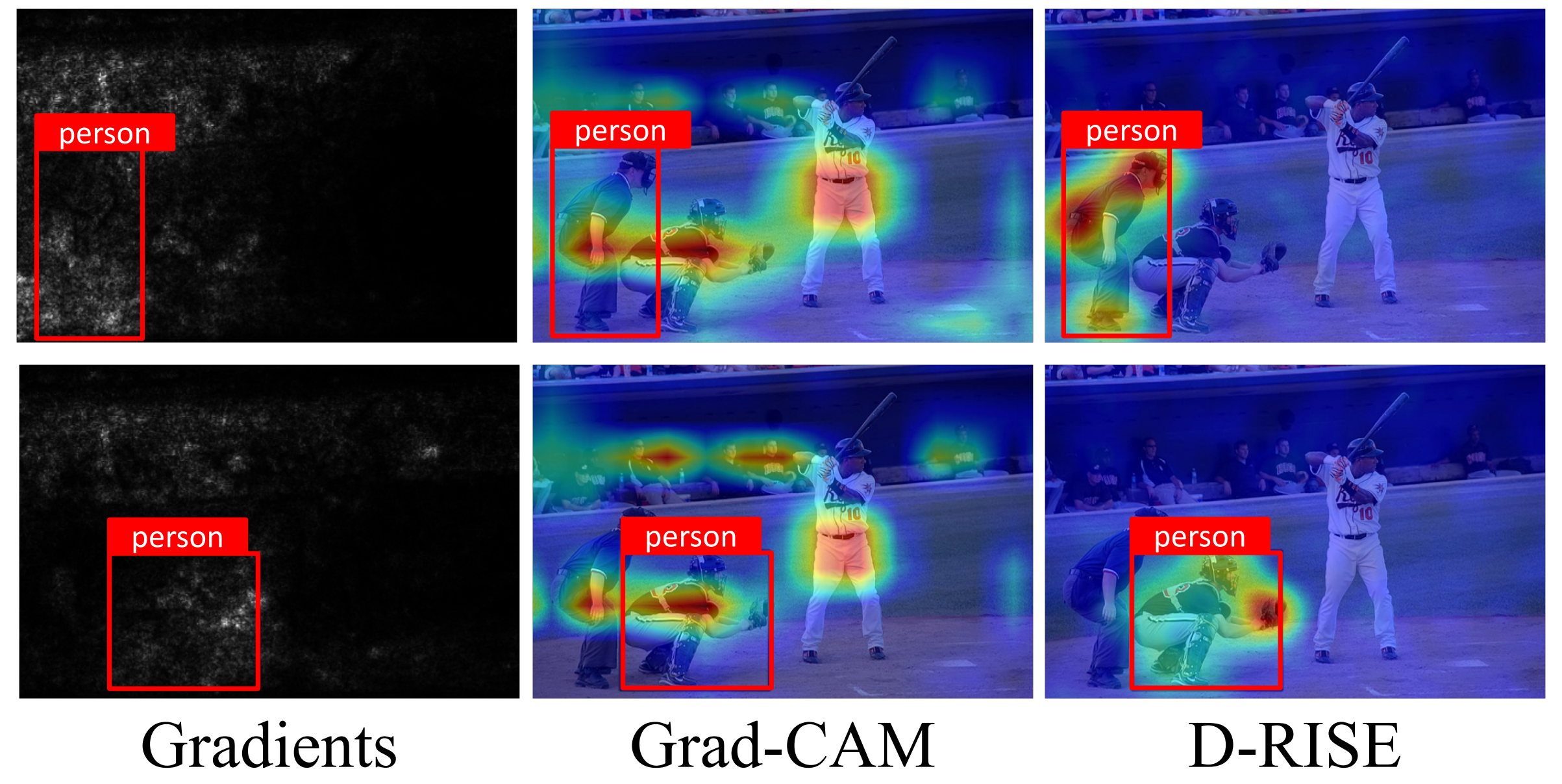
Vitali Petsiuk[1], Rajiv Jain[2], Varun Manjunatha[2], Vlad I. Morariu[2], Ashutosh Mehra[2], Vicente Ordonez[4], Kate Saenko[1,3]

Boston University, Adobe Inc., MIT-IBM Watson AI Lab, University of Virginia

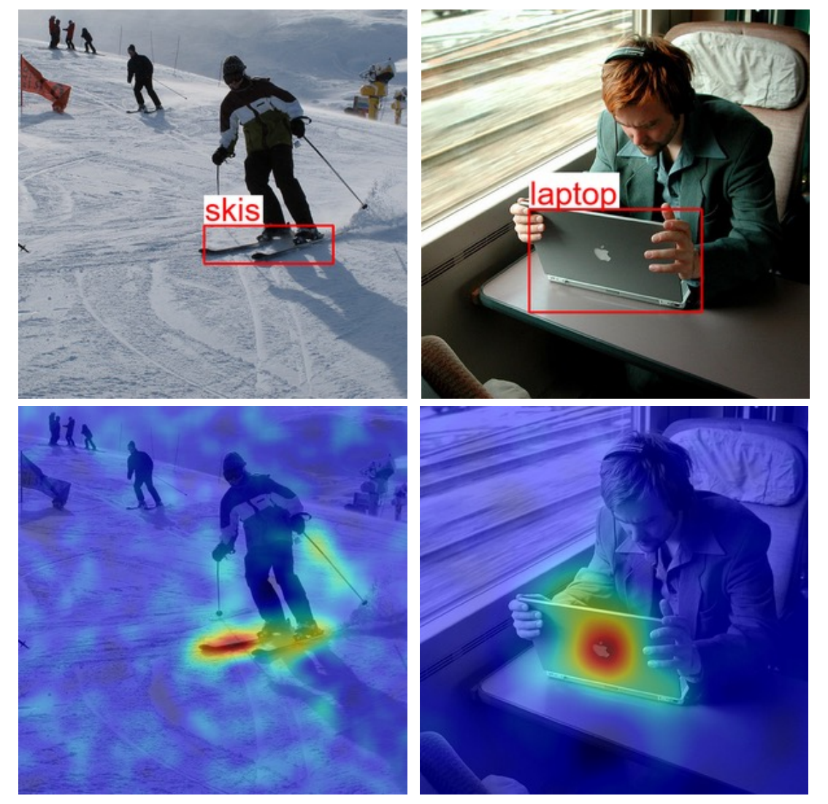
CVPR VIRTUAL JUNE 19-25

Overview.

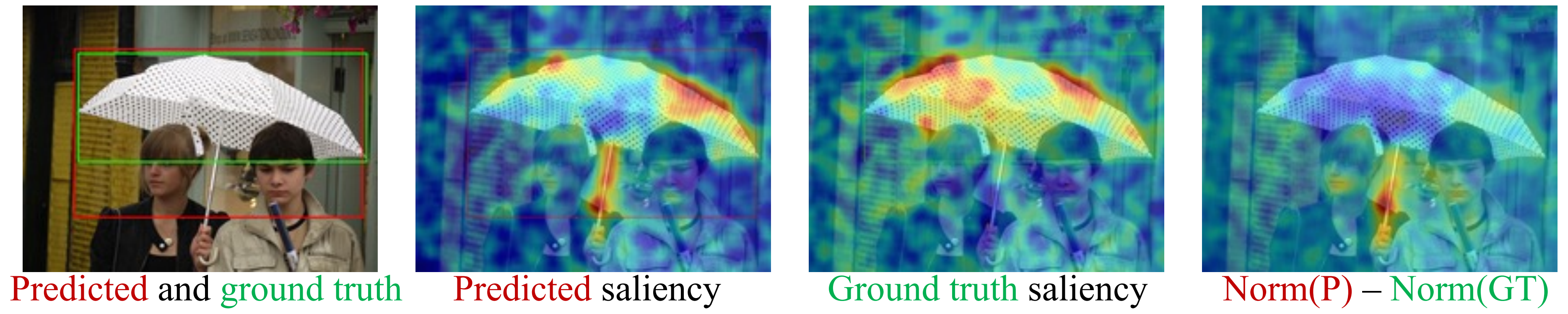
We propose D-RISE, a method for generating visual explanations for the predictions of object detectors. Utilizing the proposed similarity metric that accounts for both localization and categorization aspects of object detection allows our method to produce saliency maps that show image areas that most affect the prediction. D-RISE can be considered "black-box" in the software testing sense, as it only needs access to the inputs and outputs of an object detector.

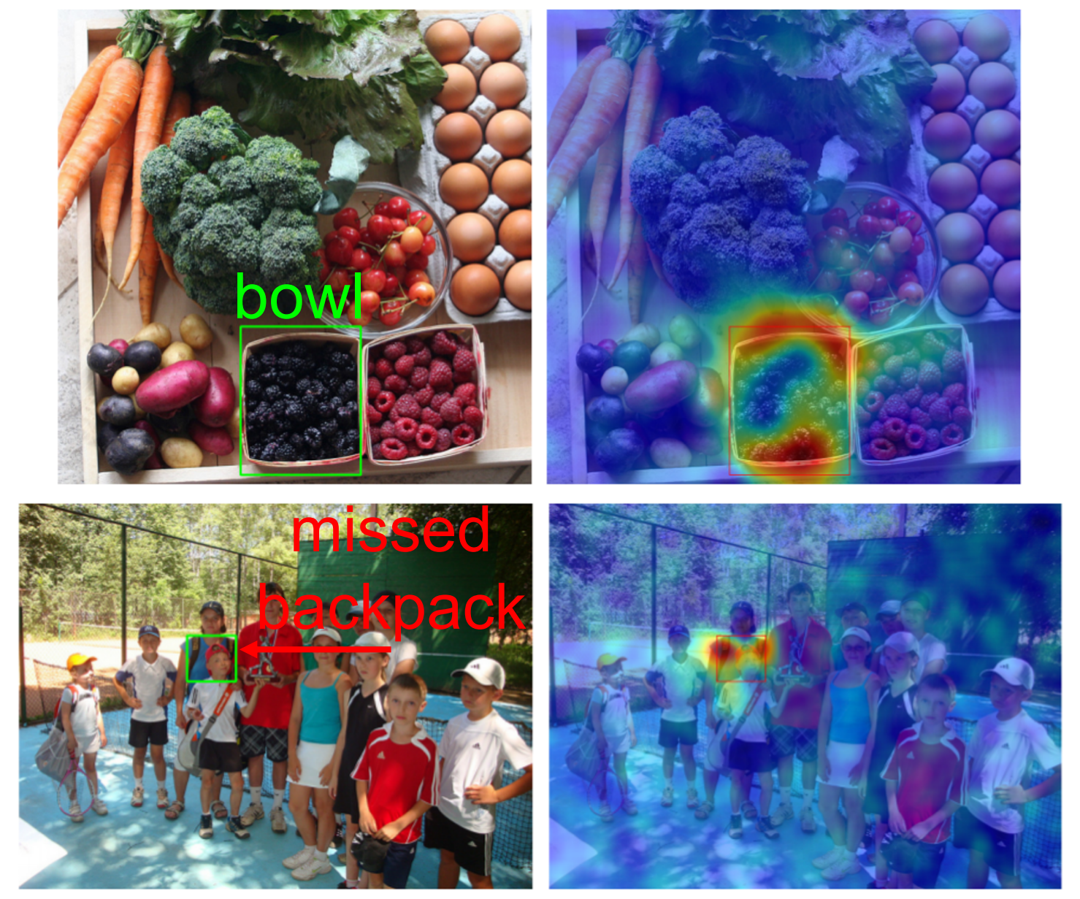


Classification saliency methods applied to detection model vs. D-RISE



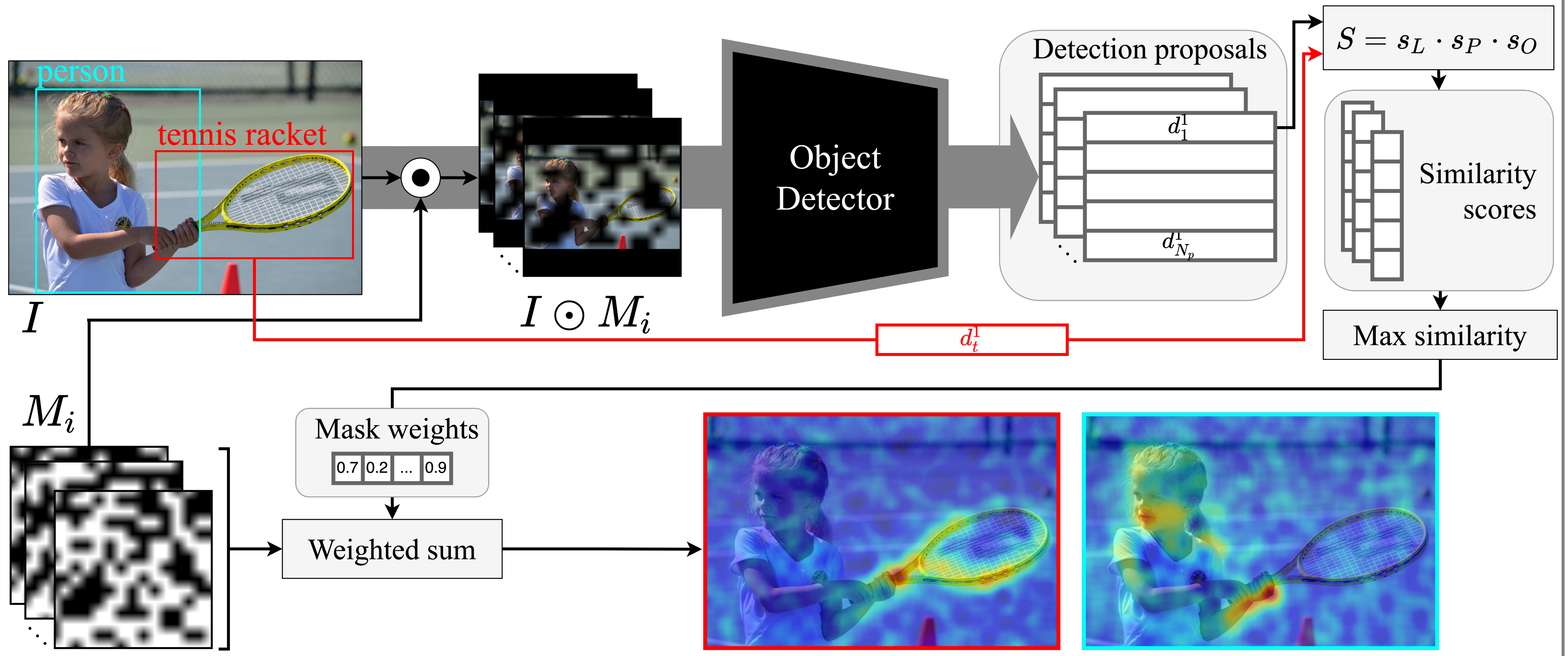
1. Model could be using contextual information.
2. Object parts are not equally important.





Analyzing failure modes

An object detector's errors may be categorized into the following modes of failure: 1) missing an object entirely, 2) detecting an object with poor bounding box localization and 3) correct localization but misclassification of an object (which includes confusion with similar classes, with dissimilar classes or with background). We show that our method can be used to analyze each of these specific types of errors.

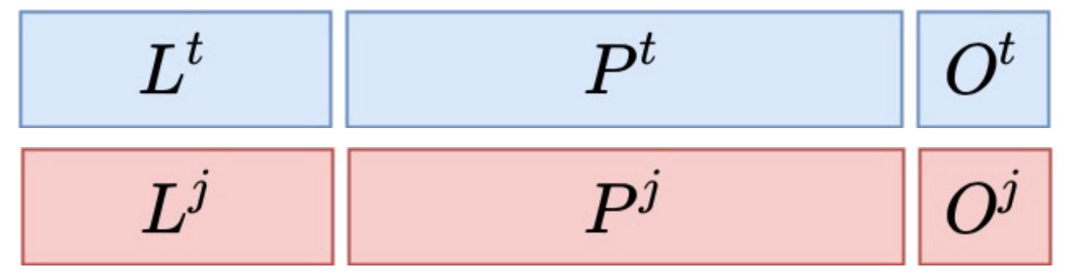


D-RISE

Our method D-RISE attempts to explain the detections (bounding-box+category) produced for this image by an object detector. We convert target detections that need to be explained into detection vectors dt. We sample binary masks and run the detector on the masked images to obtain proposals. We compute pairwise similarities between targets and proposals to obtain weights for each mask. Finally, the weighted sum of masks is computed to produce saliency maps.

Similarity metric

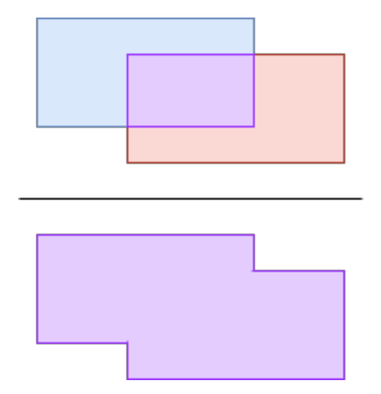
To compute the similarity score between the target vector and the proposal vector, three components should be considered: localization (L), classification (P) and objectness (O).



$$s(d_t, d_j) = s_L(d_t, d_j) \cdot s_P(d_t, d_j) \cdot s_O(d_t, d_j)$$

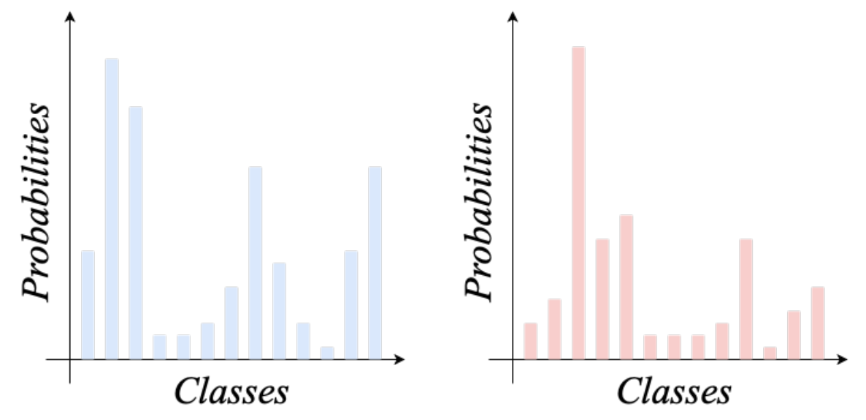
$$S(d_t, f(M_i \odot I))) \triangleq \max_{d_j \in f(M_i \odot I)} s(d_t, d_j)$$

Intersection over Union



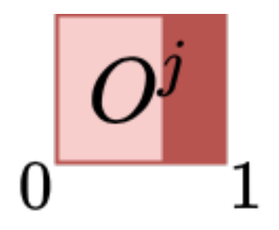
$$s_L(d_t, d_j) = \mathrm{IoU}(L^t, L^j)$$

Cosine similarity



$$s_P(d_t, d_j) = \frac{P^t \cdot P^j}{\|P^t\|\|P^j\|}$$

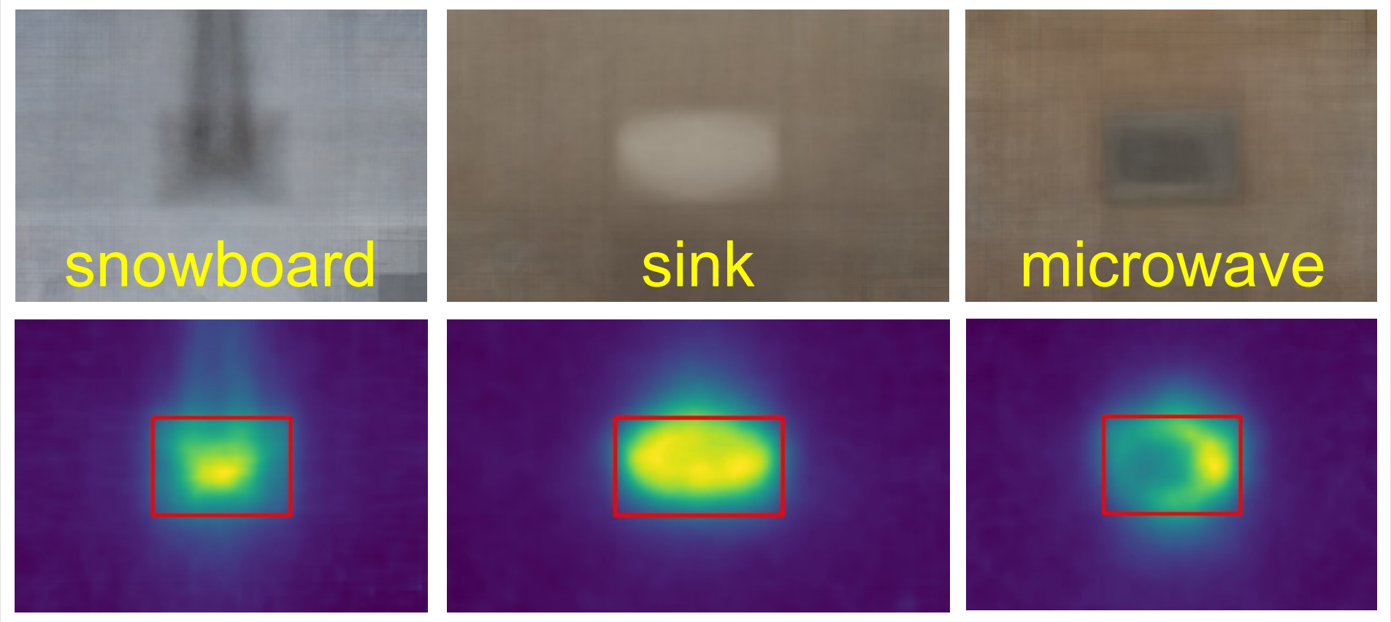
Objectness score (Optional)



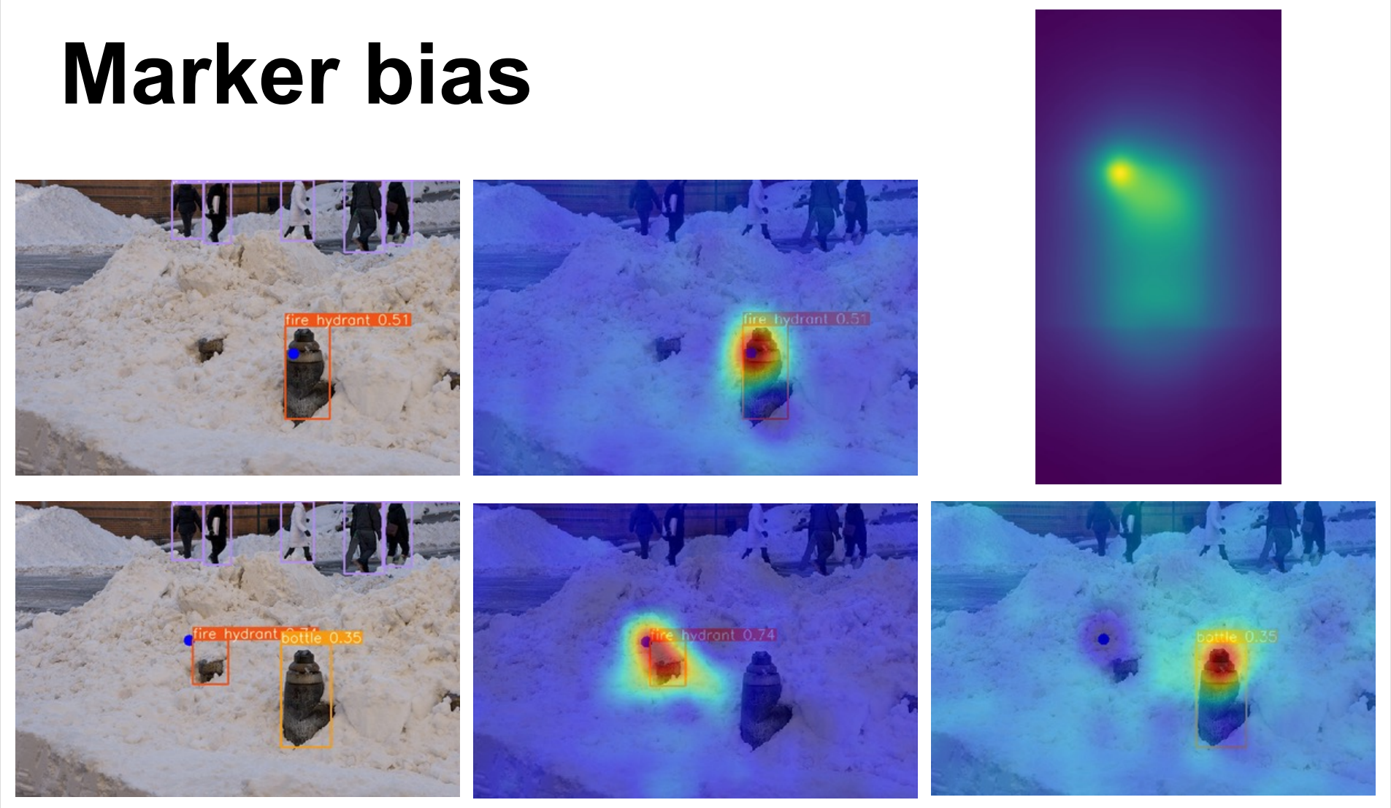
$$s_O(d_t, d_j) = O^j$$

Average saliency

To explain the model from a more holistic perspective and find common patterns we compute average saliency maps for each category of MS-COCO dataset.



Marker bias



D-RISE correctly points to the blue dot as a reason for the detections, and the average saliency map shows a significant artifact in the top-left corner.

Evaluation

We evaluate the adapted classification saliency metrics, Pointing Game and Deletion/Insertion, to compare it against the classification-based methods.

We also evaluate the ability of saliency maps help user identify which of the two models is better.

Method	PG (mask) ↑	PG (bbox) ↑	Del. ↓	Ins.↑
Gradient	0.7304	0.5195	0.0464	0.4561
Grad-CAM	0.5232	0.4209	0.0762	0.4050
D-RISE	**0.9656**	**0.8458**	**0.0440**	**0.5622**

Conclusions

We propose a novel approach for providing saliency-based explanations for black-box object detectors. Our method is general enough to be applied to many different object detection architectures.

We demonstrate the usefulness of our method in aiding error analysis and in providing insights to model developers by means of per-class average saliency maps.

We have shown that our method is able to detect pathological biases in model behavior.

https://cs-people.bu.edu/vpetsiuk/drise