# BLACK-BOX EXPLANATION OF OBJECT DETECTORS VIA SALIENCY MAPS

*And Beyond.....*

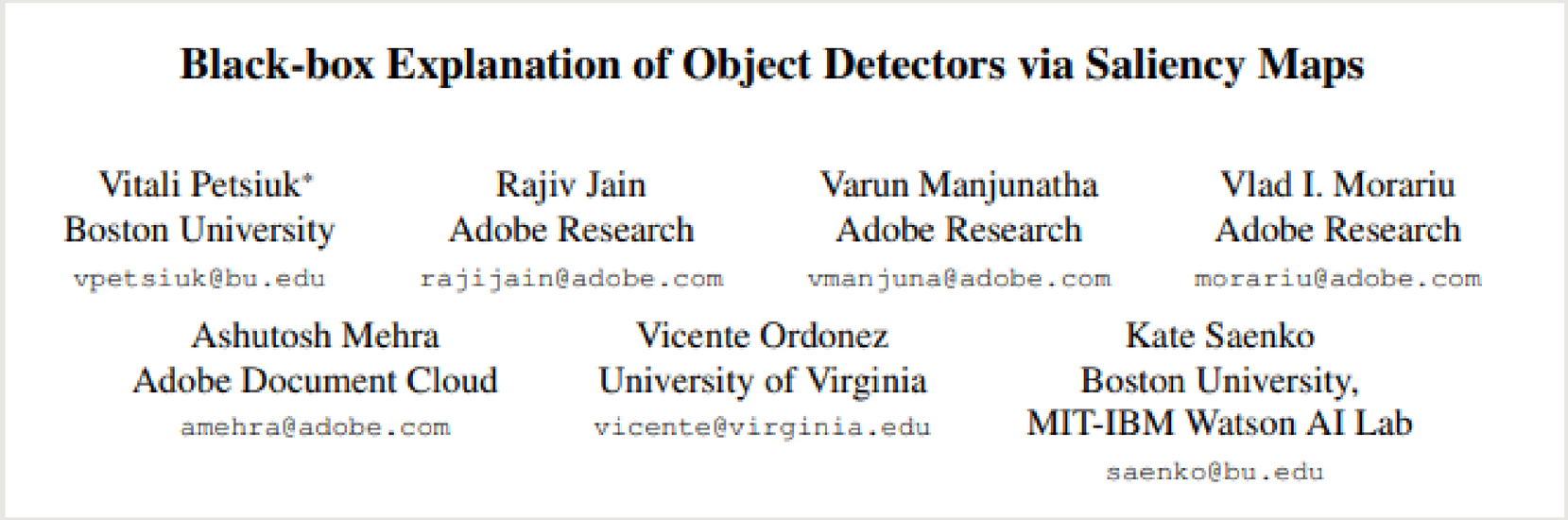
**Krishna Singh**
**2021112005**

**Sarthak Bansal**
**2021101134**

# Motivation

*Where does YOLO look?, though only once!*

Our project aims at **black-box unmasking** of popular object detector, YOLO, to gain insights about the biases the model has and how prone it is to adversarial attacks.



**Black-box Explanation of Object Detectors via Saliency Maps**


Vitali Petsiuk*
Boston University
vpetsiuk@bu.edu

Rajiv Jain
Adobe Research
rajijain@adobe.com

Varun Manjunatha
Adobe Research
vmanjuna@adobe.com

Vlad I. Morariu
Adobe Research
morariu@adobe.com

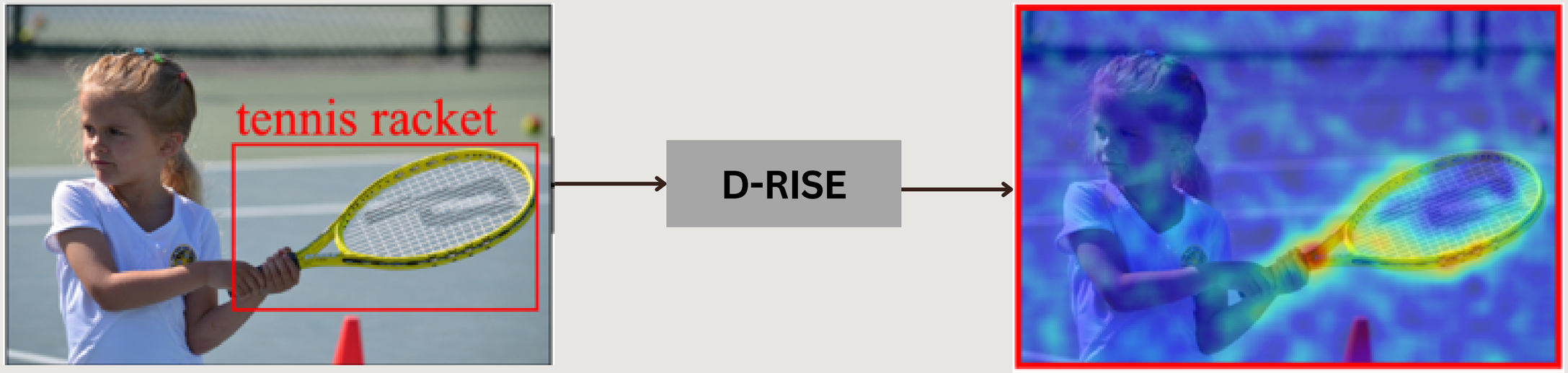
Ashutosh Mehra
Adobe Document Cloud
amehra@adobe.com

Vicente Ordonez
University of Virginia
vicente@virginia.edu

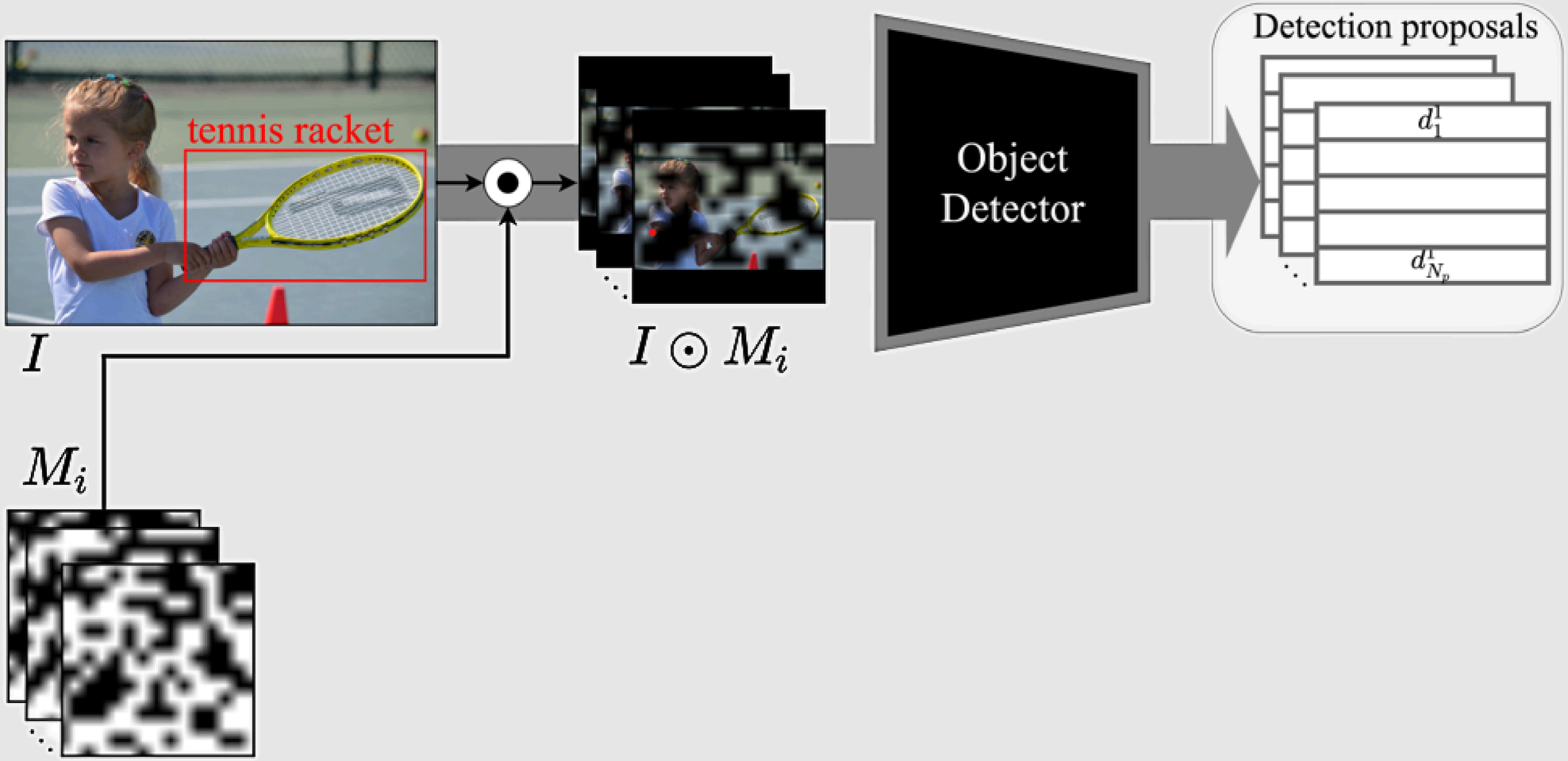
Kate Saenko
Boston University,
MIT-IBM Watson AI Lab
saenko@bu.edu

# HOW ? -> D-RISE Method

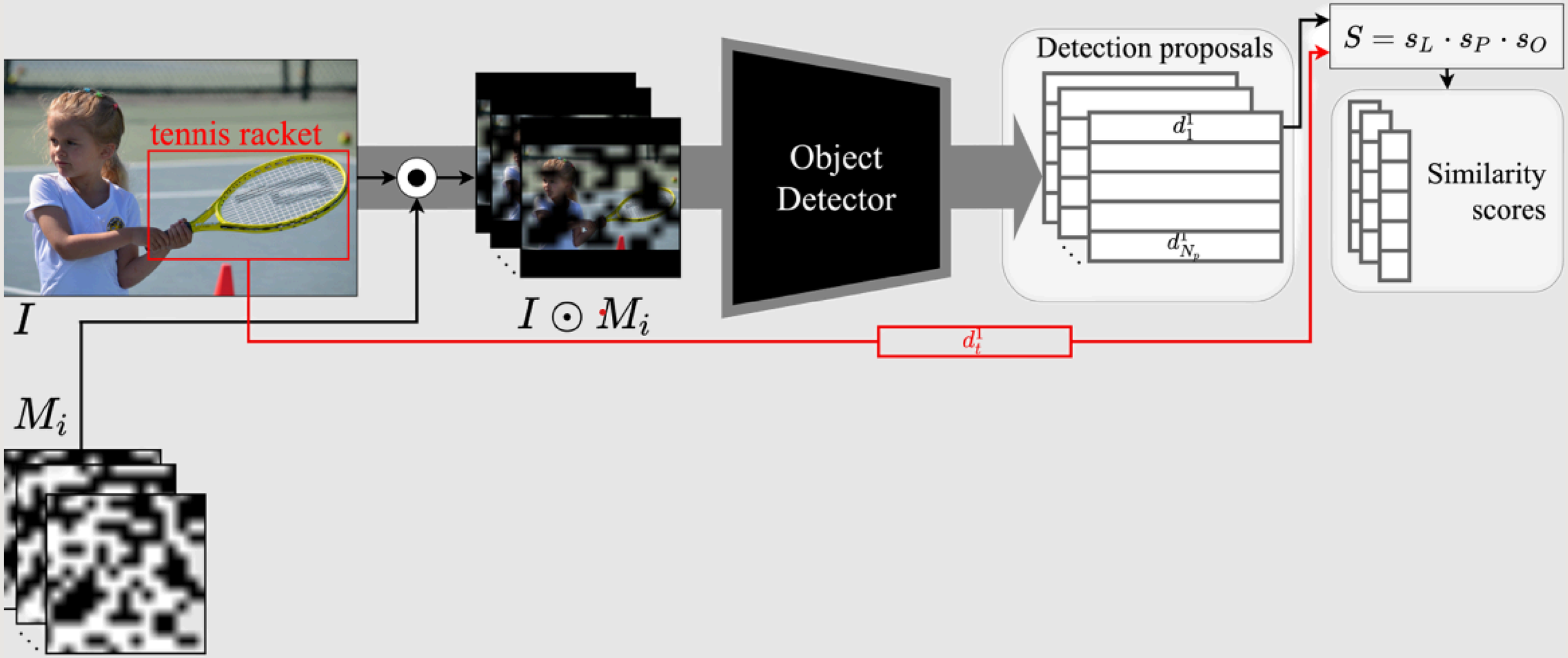
*“Generating Saliency map for any arbitrary detection without needing any prior knowledge of its working ”*



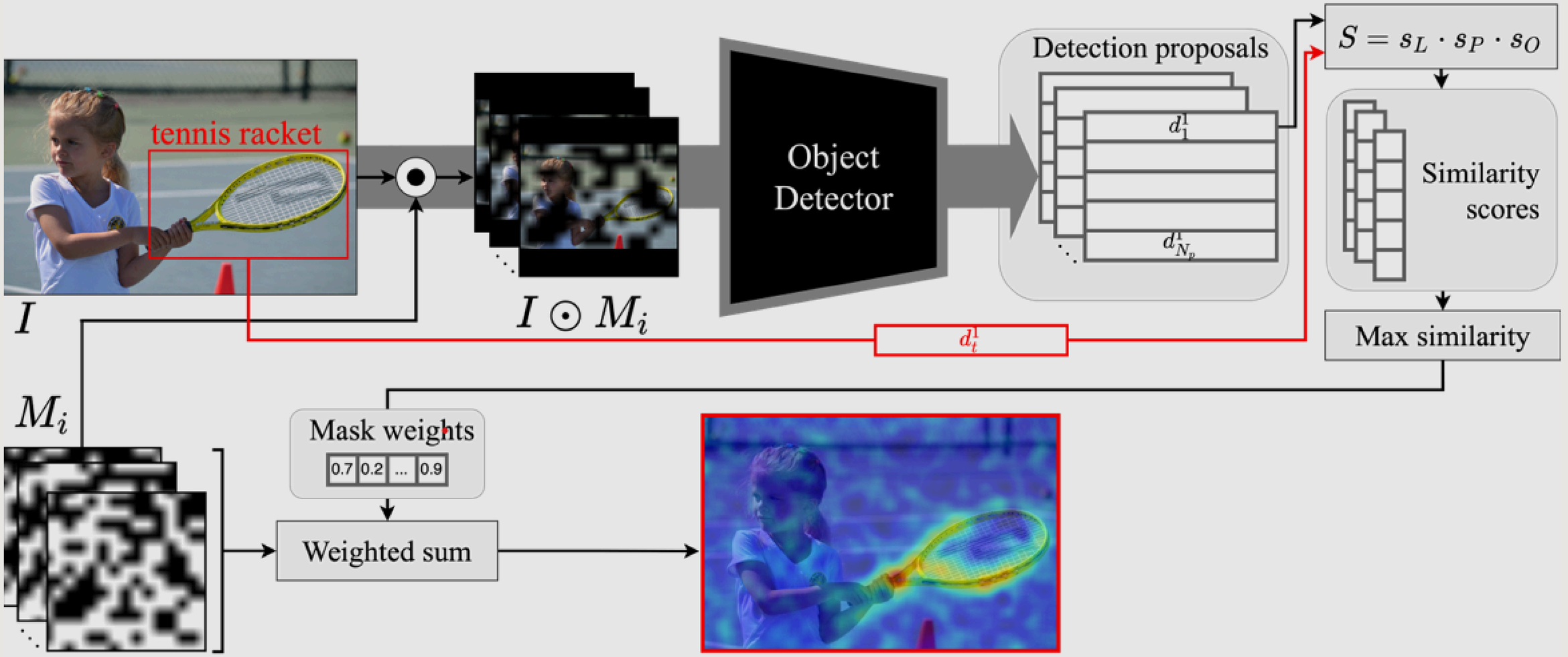
# Overview of Method



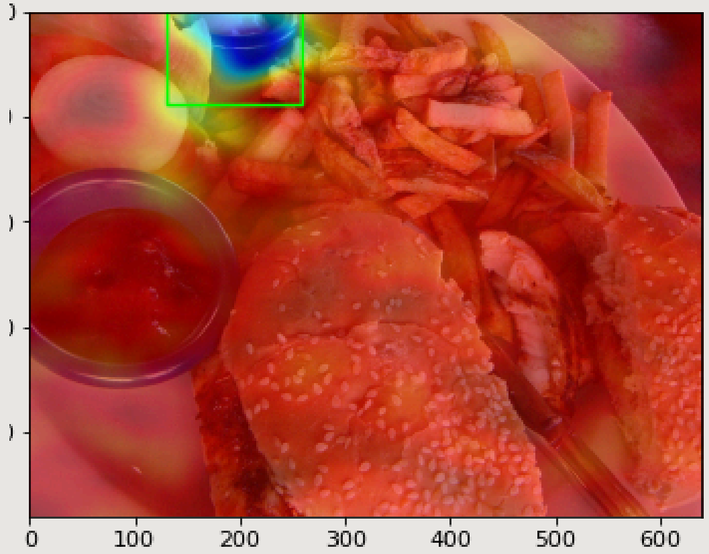
# Overview of Method



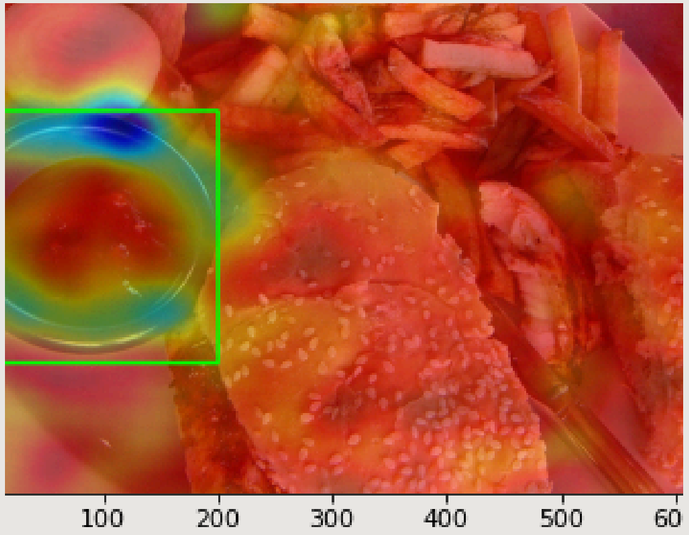
# Overview of Method



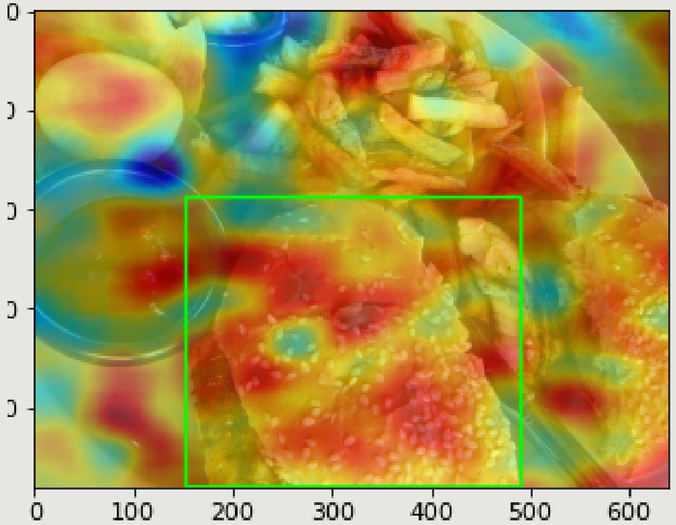
# Some Results




Model focused strongly on **almost all** of the detected regions equally

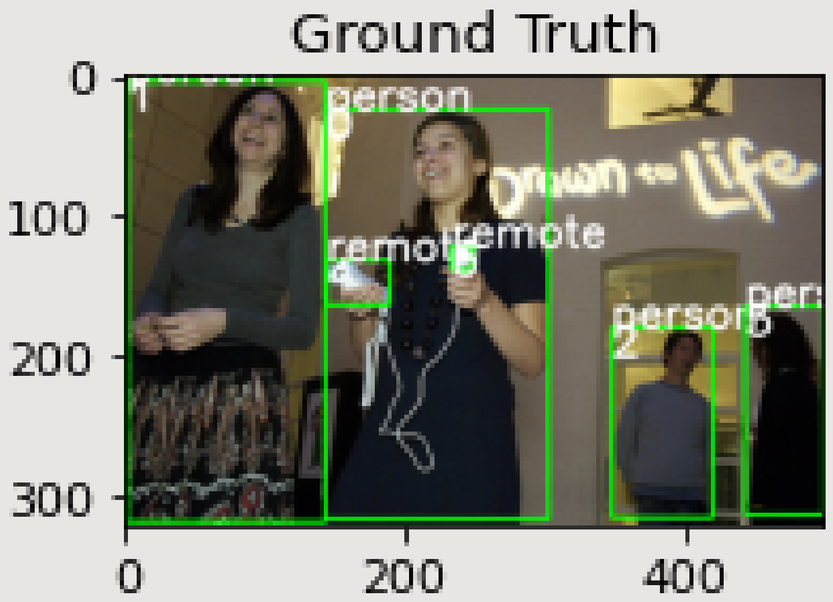


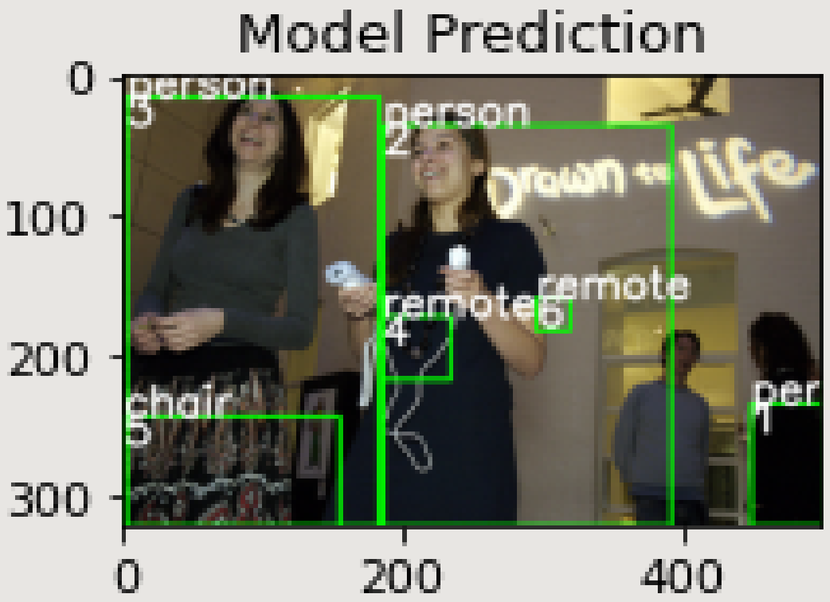

Model focused on the **brim of the bowl** strongly and didn’t focus on the inside region that much

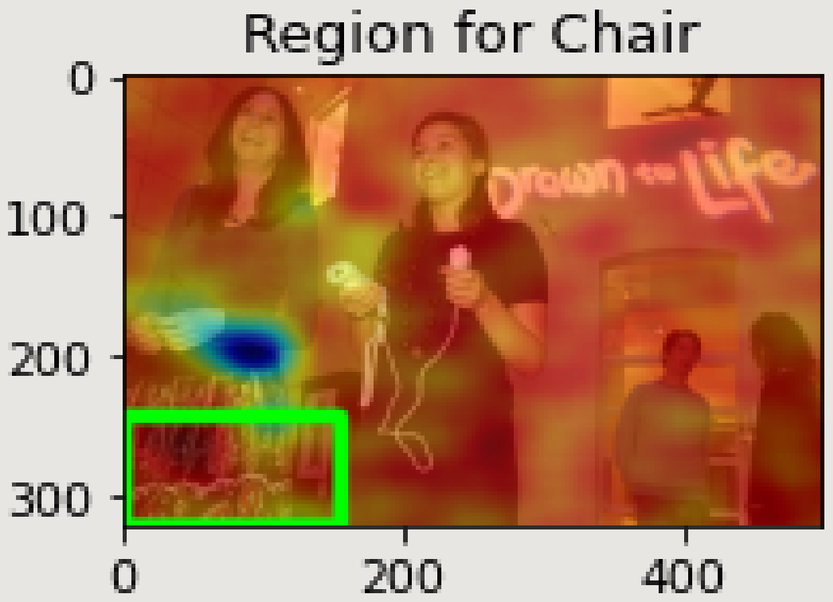



Model **didn’t** focus on the object at all but instead used **contextual** information to make prediction

# Failure Mode - 1 : False Positives



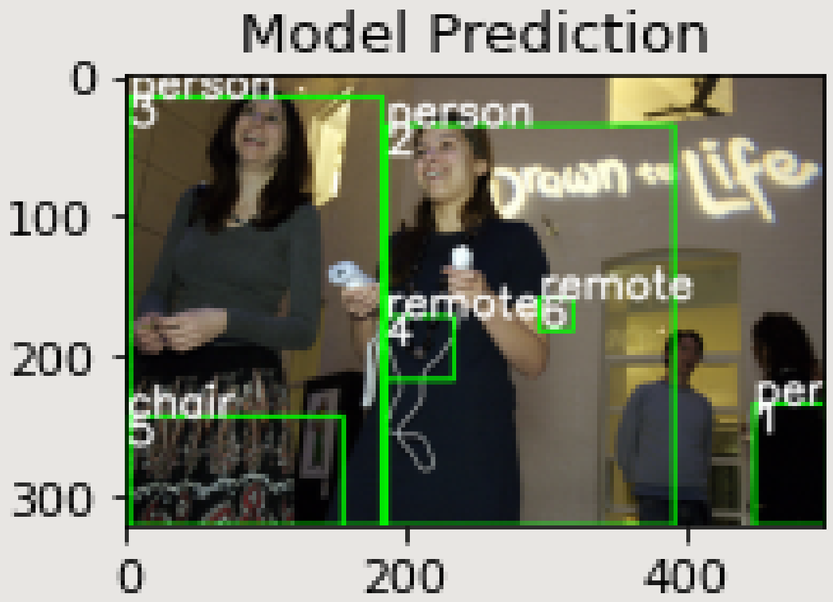



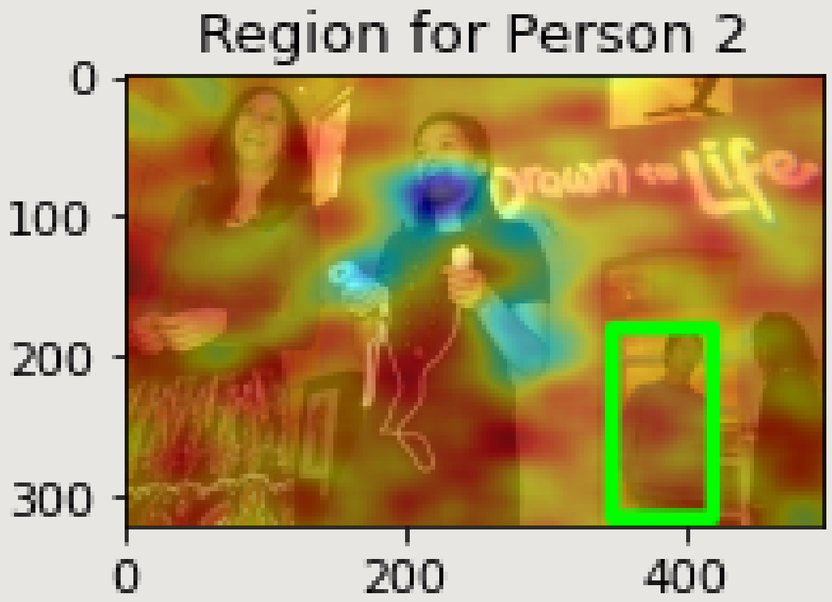




Despite the chair not being visibly present in the image, the model detects it. Analysis of the saliency map reveals that the model didn't focus on the area within the detected bounding box but relied on contextual cues for its prediction.

Such cases are common in object detection models.

# Failure Mode - 2 : Missed Object

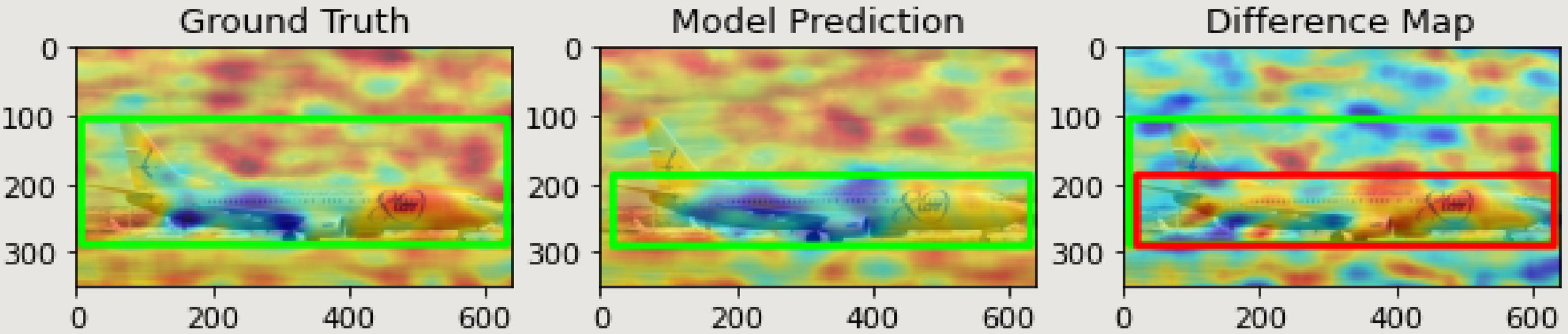




Despite the person (2) being clearly visible in the image, model fails to capture its bounding box. On analyzing the saliency map , we find that the model kept on focusing on the other person and never even looked at the contextual information around the ground truth box

Such cases are also common in object detection models.

# Failure Mode - 3 : Inconsistent bounding box

As, the D-RISE method can work on any detection, we can create difference maps to analyze the sources of error of the object detector




We can see that the model could not focus enough on some regions (for e.g. fin of the airplane) which led it to make a smaller bounding box

## What are the limitations of YOLOv5?

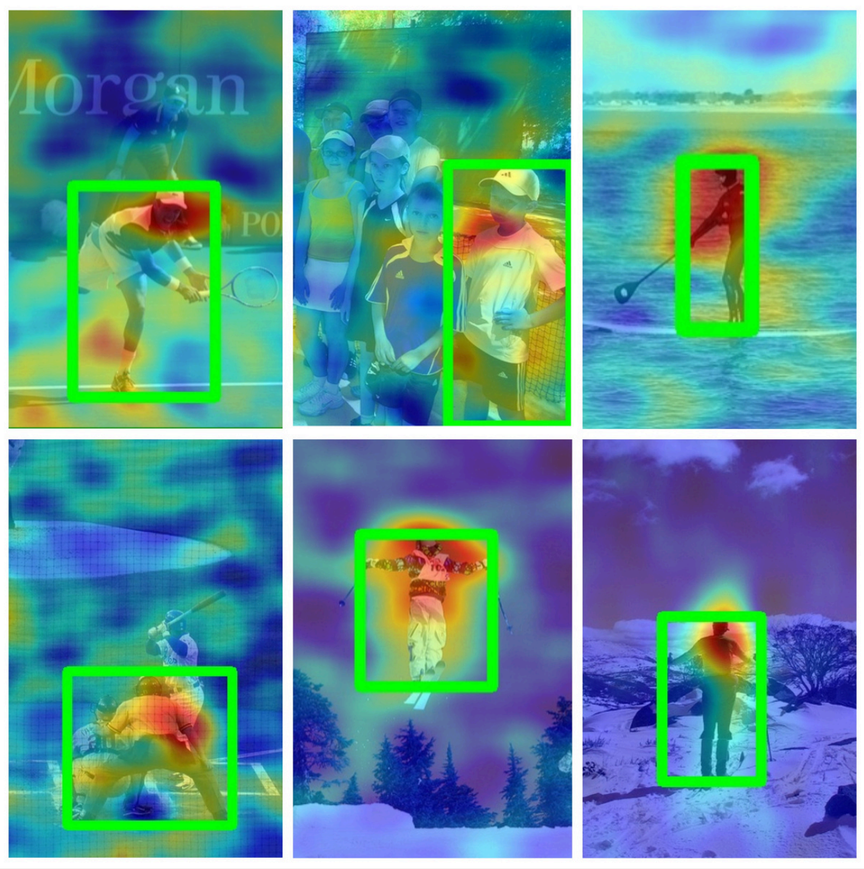
Graphical user interface | Computer science | Detector | Identification (biology) | Bottleneck

**Best insight from top research papers**

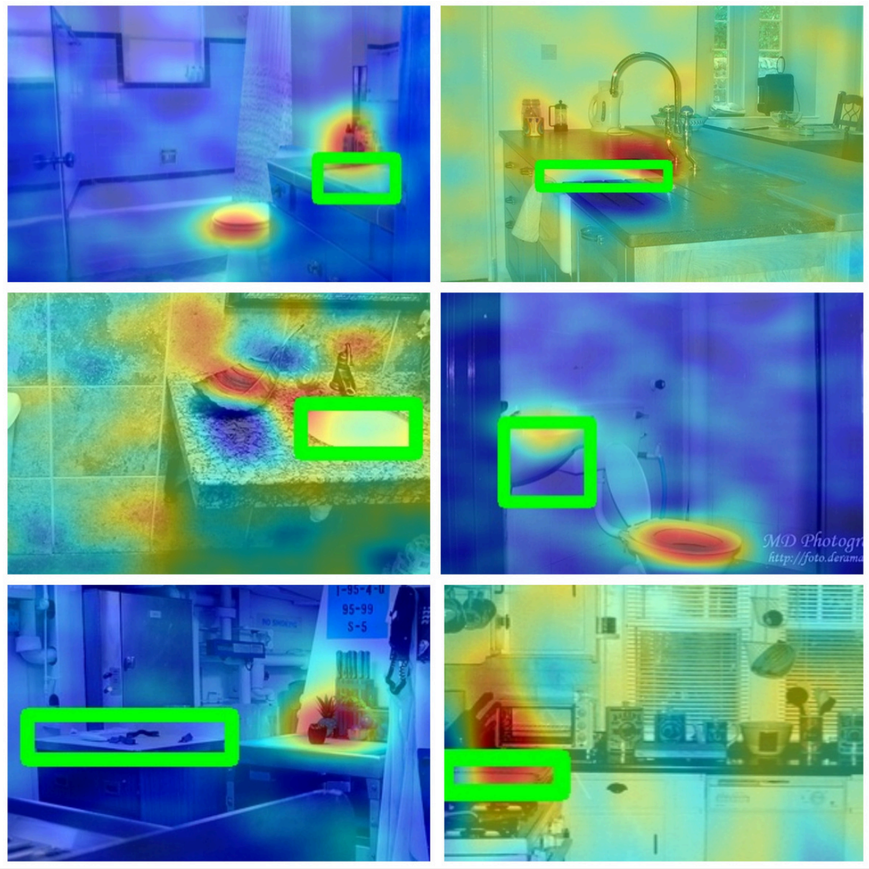
The limitations of YOLOv5 include issues such as low accuracy, a high rate of false detection, missed detection, sample imbalance, and difficulty in detecting small objects. [1] [2] [3]

YOLOv5 has limitations in detecting small objects. It suffers from false detection and missed detection of small floating objects [1]. The algorithm lacks the ability to extract sufficient semantic information from small objects and is susceptible to background interference [2]. YOLOv5 also has limitations in terms of misclassification, angle deflection, and limited scalability when objects are small in size [3]. Additionally, the low resolution of UAV viewpoint images and limited valid information affect the recognition rate of small targets in YOLOv5 [4]. The algorithm fails to effectively detect small targets due to factors such as complex backgrounds and loss of fine-grained information [5]. These limitations hinder the detection accuracy of YOLOv5 for small objects in various applications, including traffic sign detection and road crack detection .

# Per Class Analysis

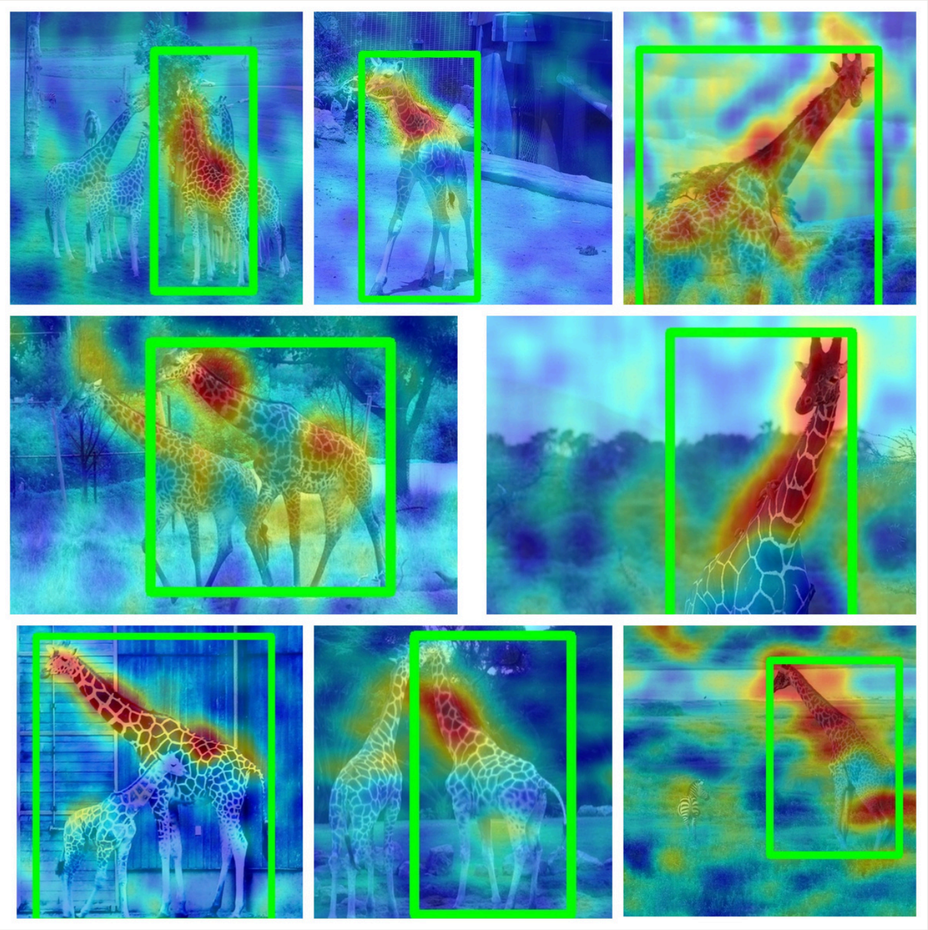


For the person class, detector usually focuses on the upper body.



For the sink class, detector usually focuses on contextual information outside the bounding box.

# Per Class Analysis



For the giraffe class, YOLO focuses on the neck region in the cases where the complete giraffe is visible. Whereas it focuses on other facial features when the complete body is not in the image.

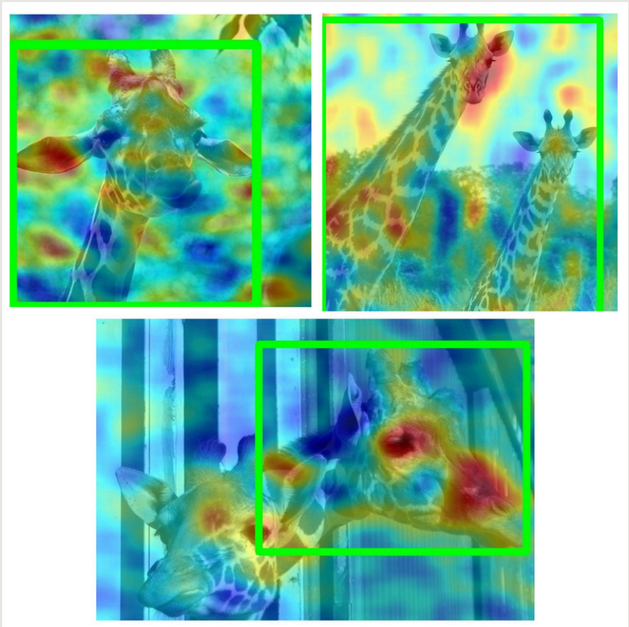
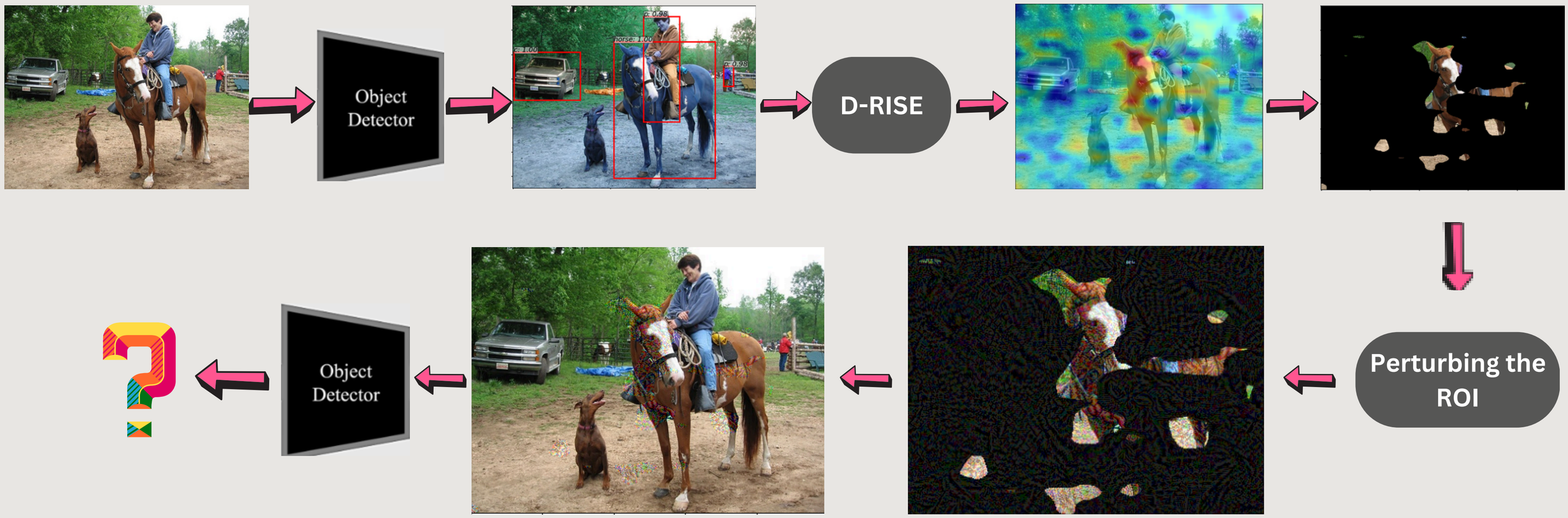


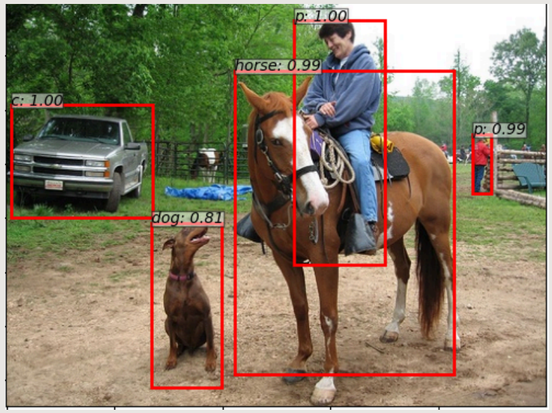

**Image Courtesy:**
**CIE Wall!**
**Next to samosa vendor.**

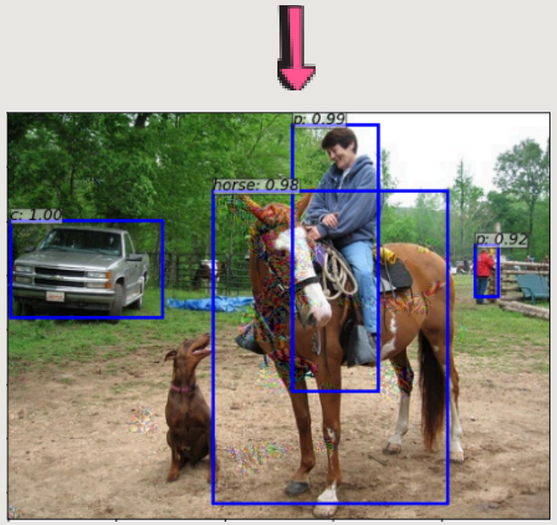
# Breaking YOLO!

*We now test YOLO's robustness by perturbing the Region of Interest (ROI) detected by D-RISE. This helps us understand what type of changes occur in the YOLO detections after the ROI is altered.*



# In some cases ...



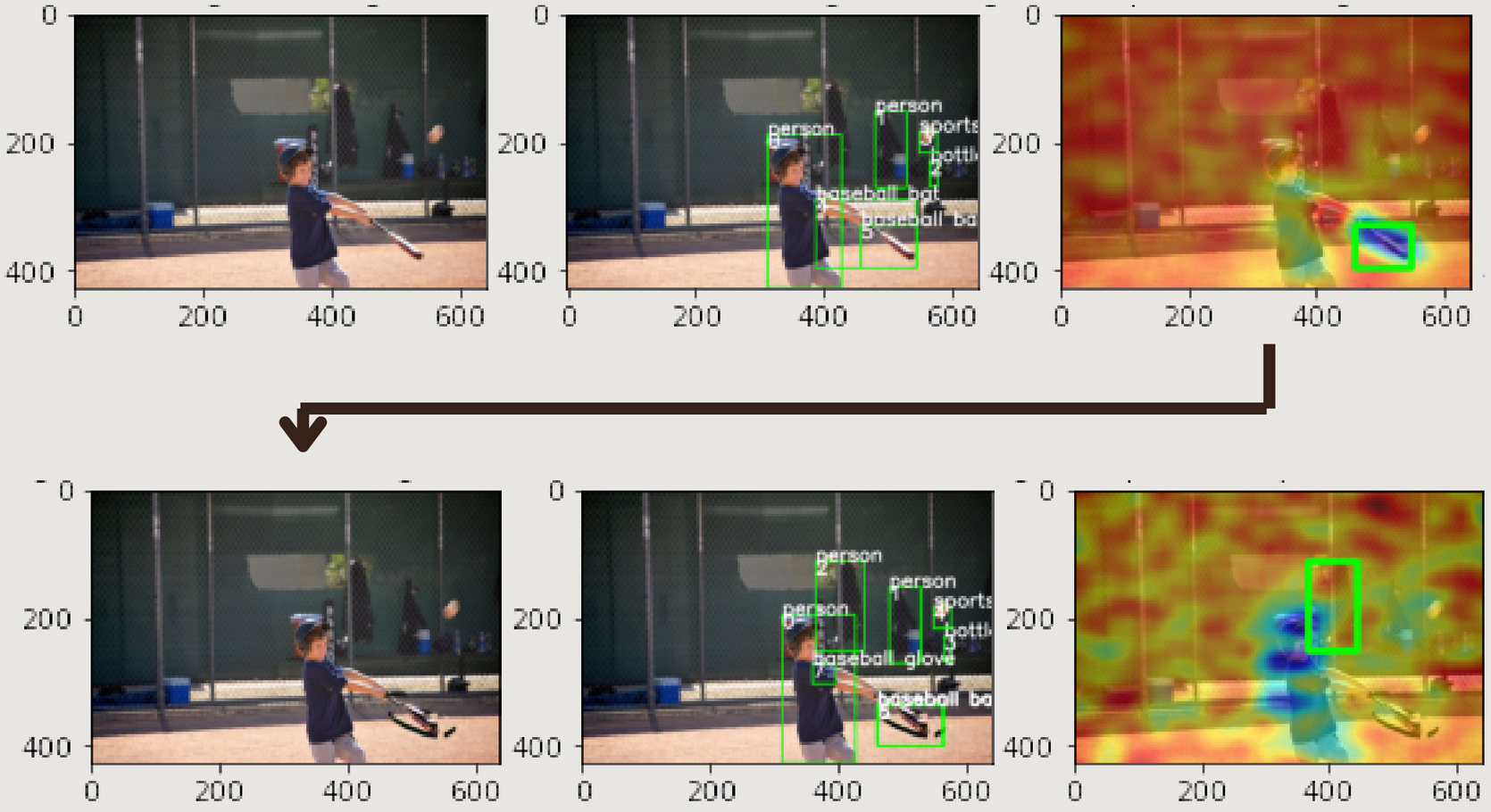




- Horse detection is taken.
- After perturbation, confidence score decreases, not only of horse but of person also.
- Bounding box shifts slightly.
- Saliency map covers person region as well as horse's. (for horse detection)

**But ...**

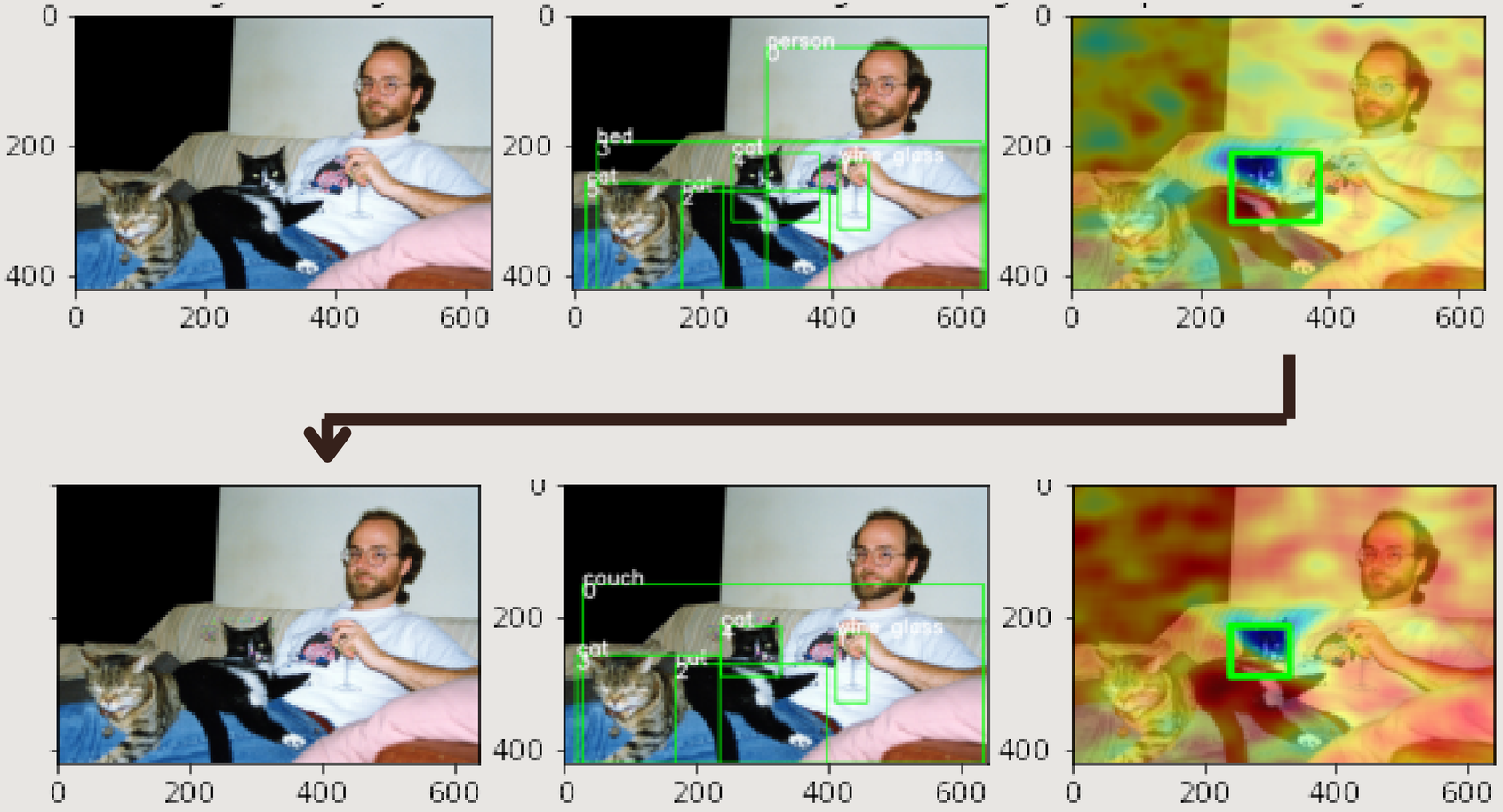
**No significant change in detections!**

# Outcome - 1 : False Positives



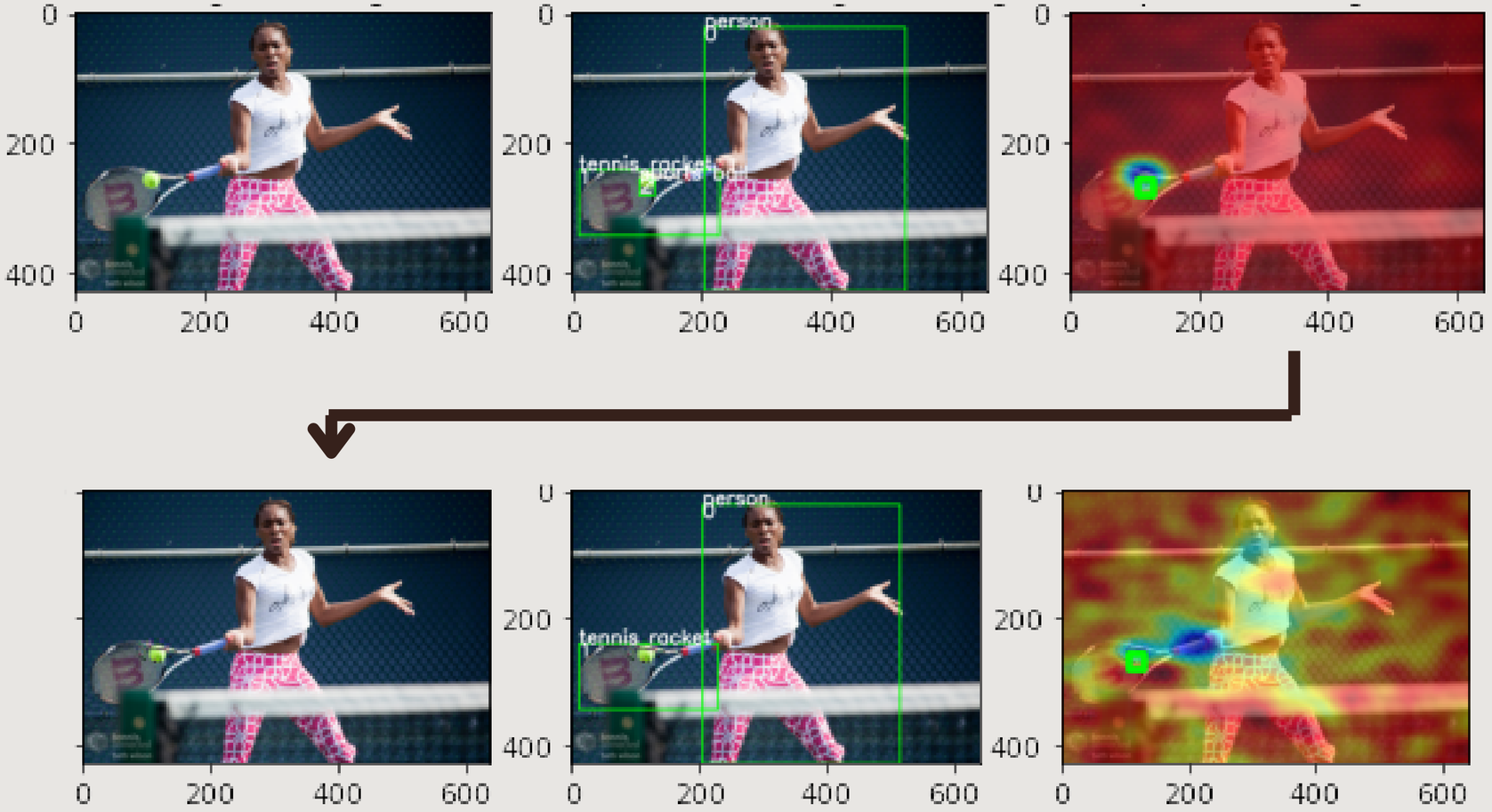

- Person (2) appears as a new detection even though the perturbation does not overlap with the detection bounding box. (Baseball bat detection is perturbed)
- The heatmap shows that YOLO focuses on foreground person for this false positive.
- YOLO can sometimes produce false positives where a background object is mistaken for a foreground object.
- Attack exploits this inherent flaw of YOLO.

# Outcome - 2 : Changed bounding box



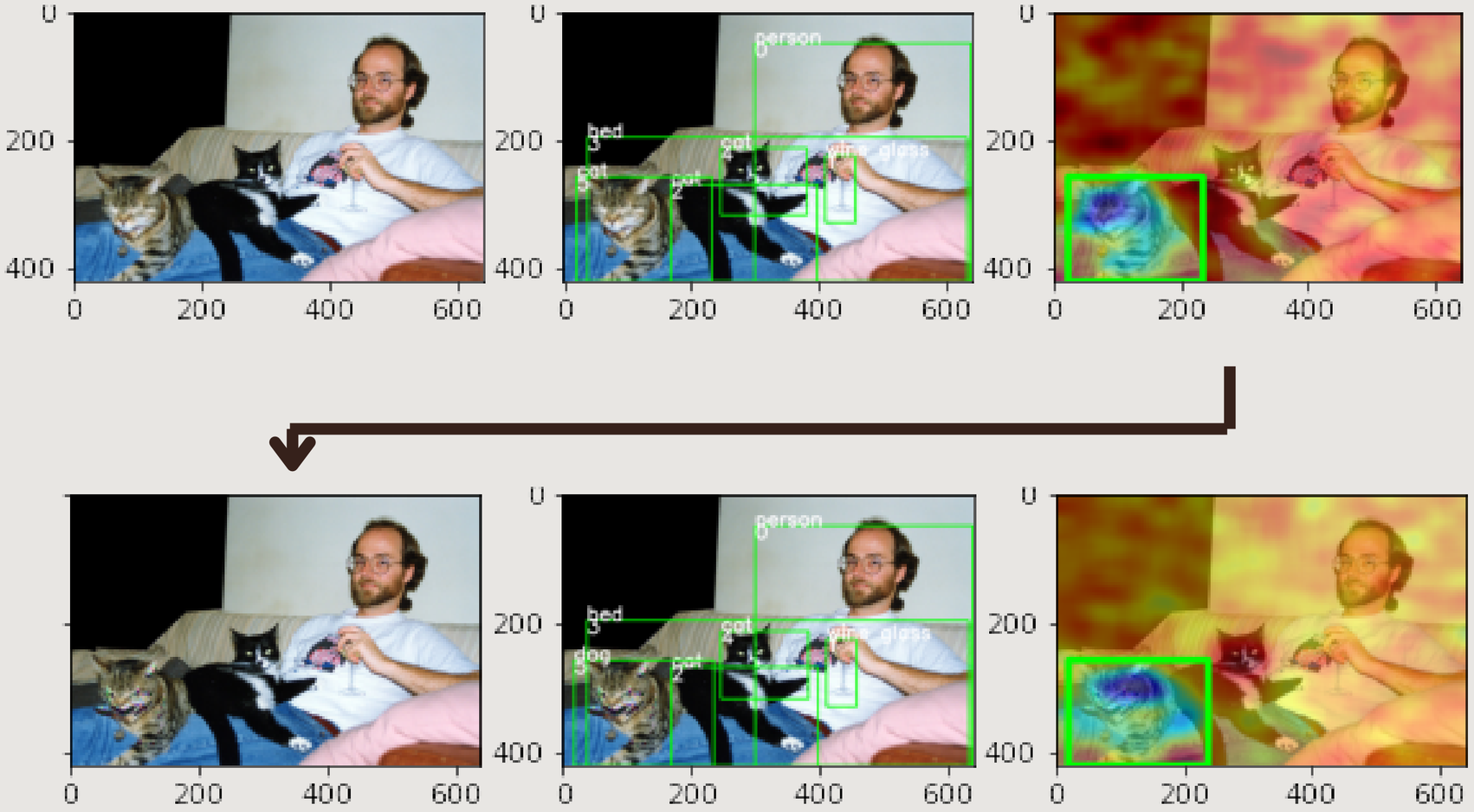

- The bounding box of cat has changed.
- The focus concentrates from the other regions of the image to face of the cat.

# Outcome - 3 : Missed Classification



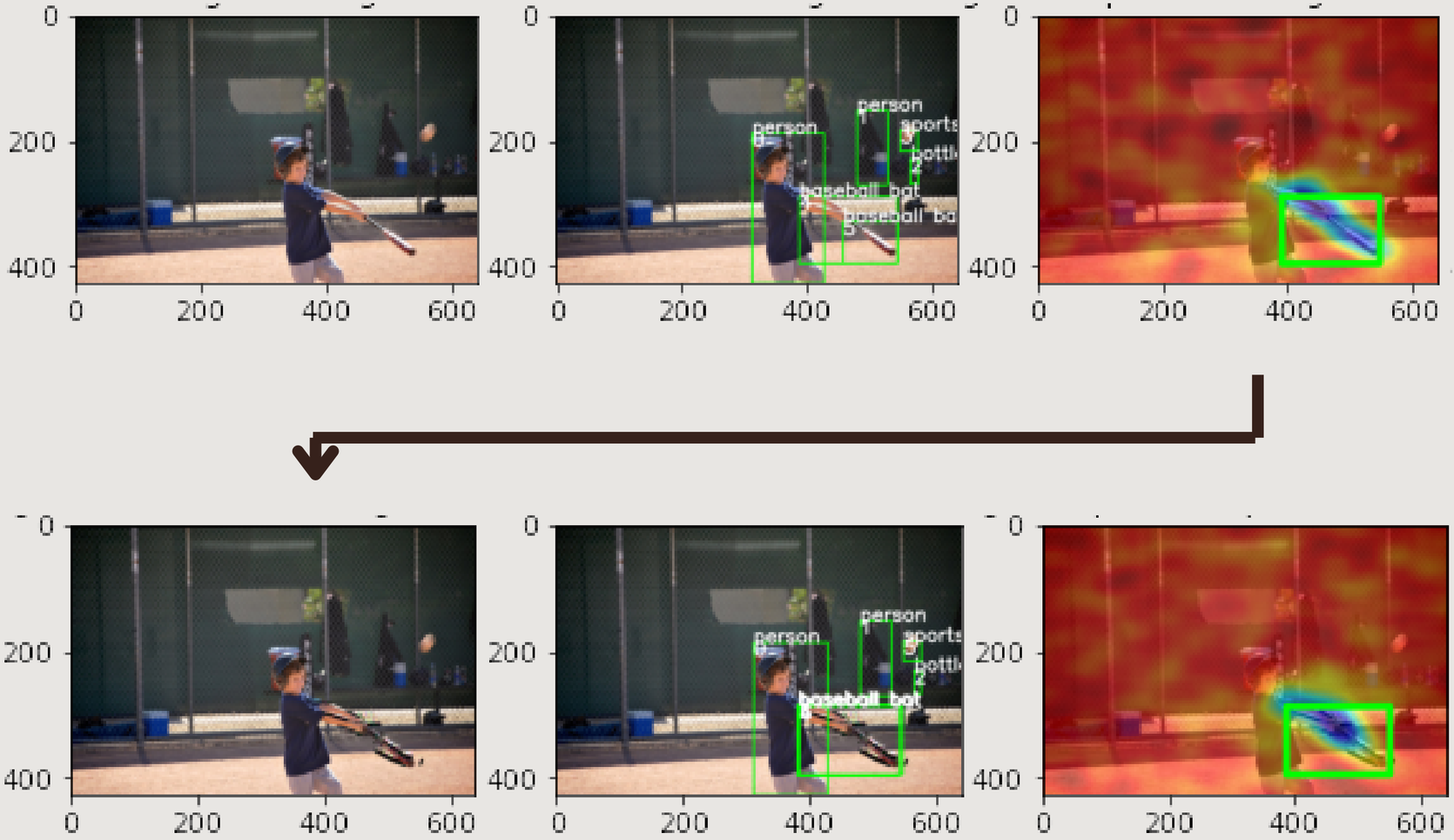

- The ball is missed from classification.
- The heatmap shows that YOLO shifts focus from ball to handle of the racquet.
- Smaller objects like the ball can be missed even when perturbation is small.

# Outcome - 4 : Misclassification




- The cat has been misclassified as a dog.
- This attack required a higher attack budget (epsilon value).
- Heatmap shows that YOLO focuses on almost the same region for the misclassified sample.

# Outcome - 5 : Not always evil !!!




- Initially the model gave a false positive baseball bat.
- Heatmap shows that the ROI is scattered in this case.
- On perturbing the image in ROI region, the false positive disappears.
- Heatmap shows that the focus of YOLO shifts to the grip/shaft of the bat.

# Future Works

- To perform simple adversarial training on YOLO and test for robustness.
- Explore other more interpretable adversarial attacks in order to specify the failure modes and perform more informed perturbations. For e.g. TOG attacks

# References


- Black-box Explanation of Object Detectors via Saliency Maps (https://arxiv.org/pdf/2006.03204)
- You Only Look Once: Unified, Real-Time Object Detection (https://arxiv.org/pdf/1506.02640)
- simple-faster-rcnn-pytorch (https://github.com/chenyuntc/simple-faster-rcnn-pytorch)
- Adversarial Detection: Attacking Object Detection in Real Time (https://arxiv.org/abs/2209.01962)

# THANK YOU!