



UNIVERSITY
OF AMSTERDAM

“Studying How to Efficiently and Effectively Guide Models with Explanations” – A Reproducibility Study

Adrian Sauter, Milan Miletic, Ryan Ott, Pemmasani Prabakaran Rohith Saai

NeurIPS 2024





“Studying How to Efficiently and Effectively Guide Models with Explanations” – A Reproducibility Study

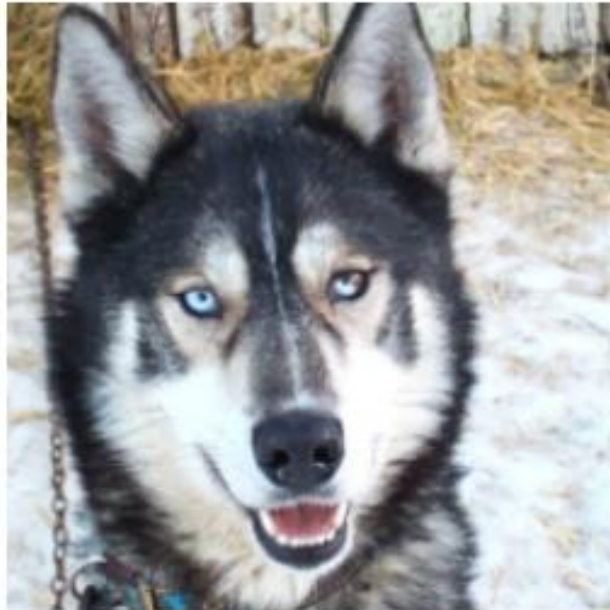
Rao et al. (2023)

Adrian Sauter, Milan Miletic, Ryan Ott, Pemmasani Prabakaran Rohith Saai

NeurIPS 2024

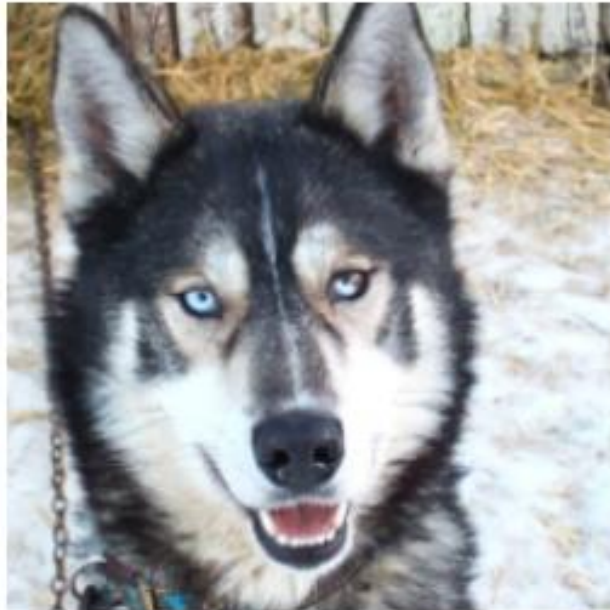


Motivation

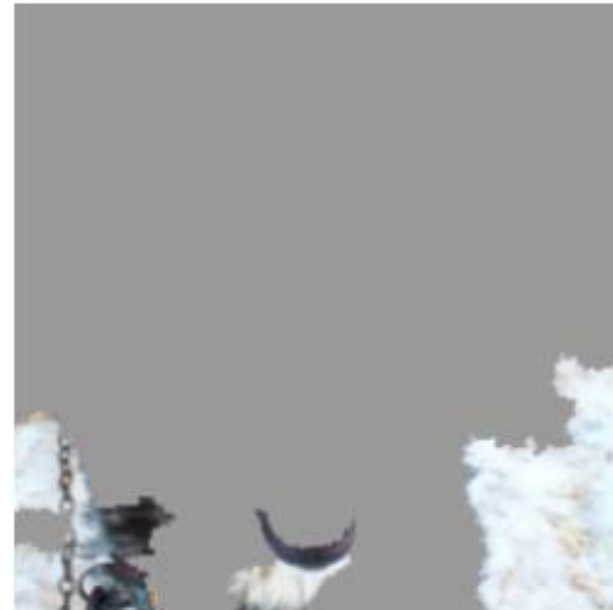


Husky classified
as a wolf

Motivation



Husky classified
as a wolf

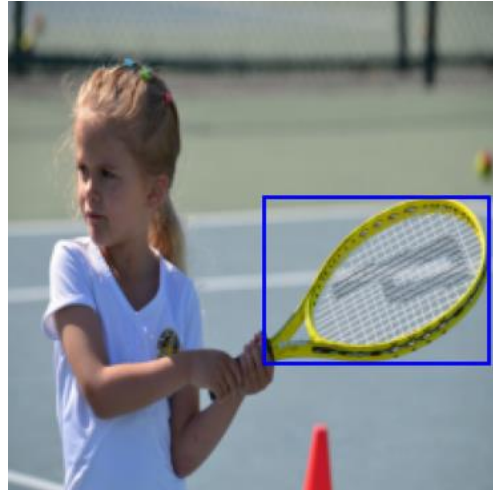


Explanation

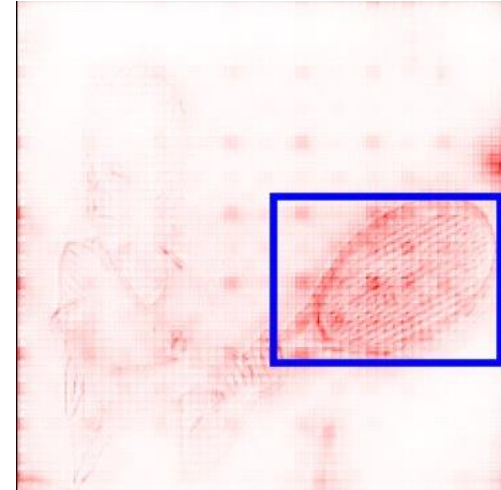
Model Guidance

- Ensuring models are “right for the right reasons”

Tennis racket

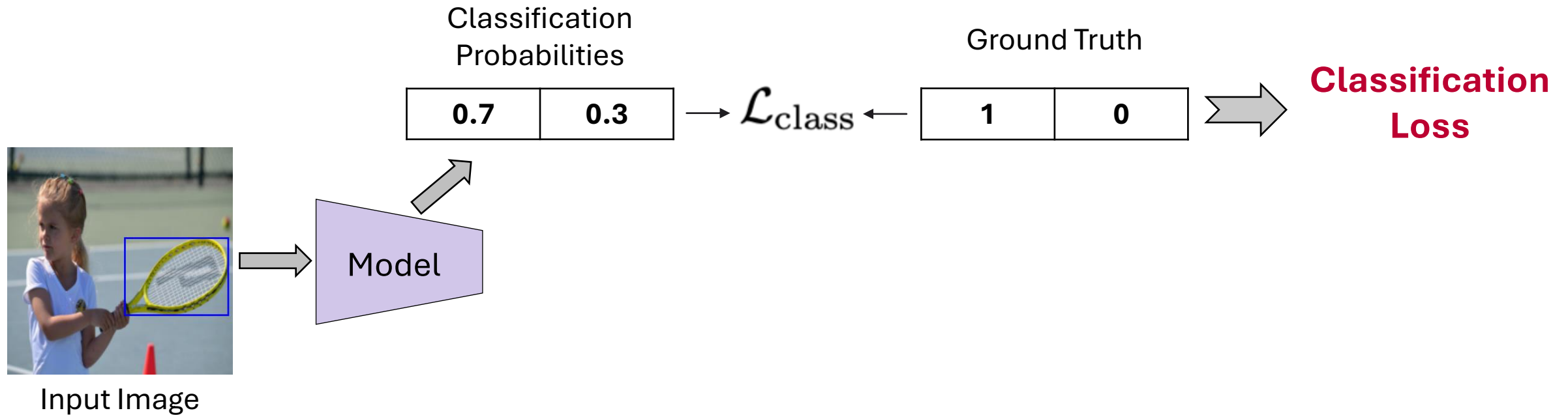


Input Image

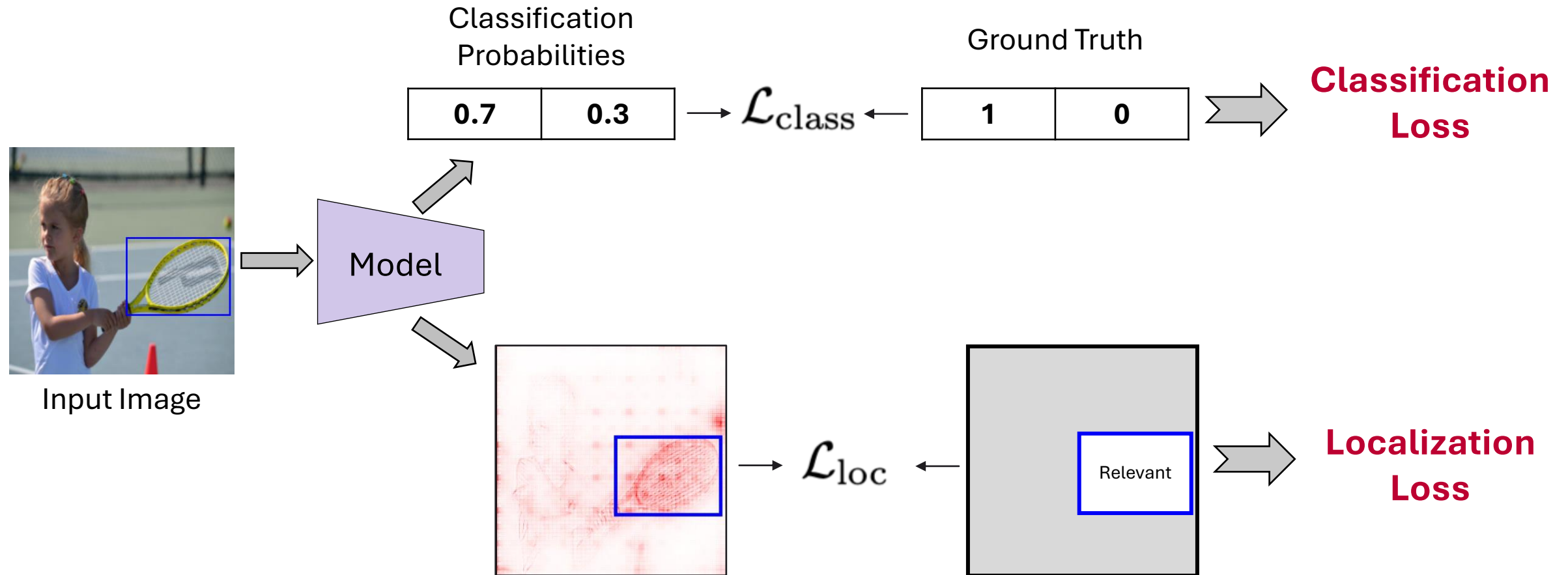


Attribution Map

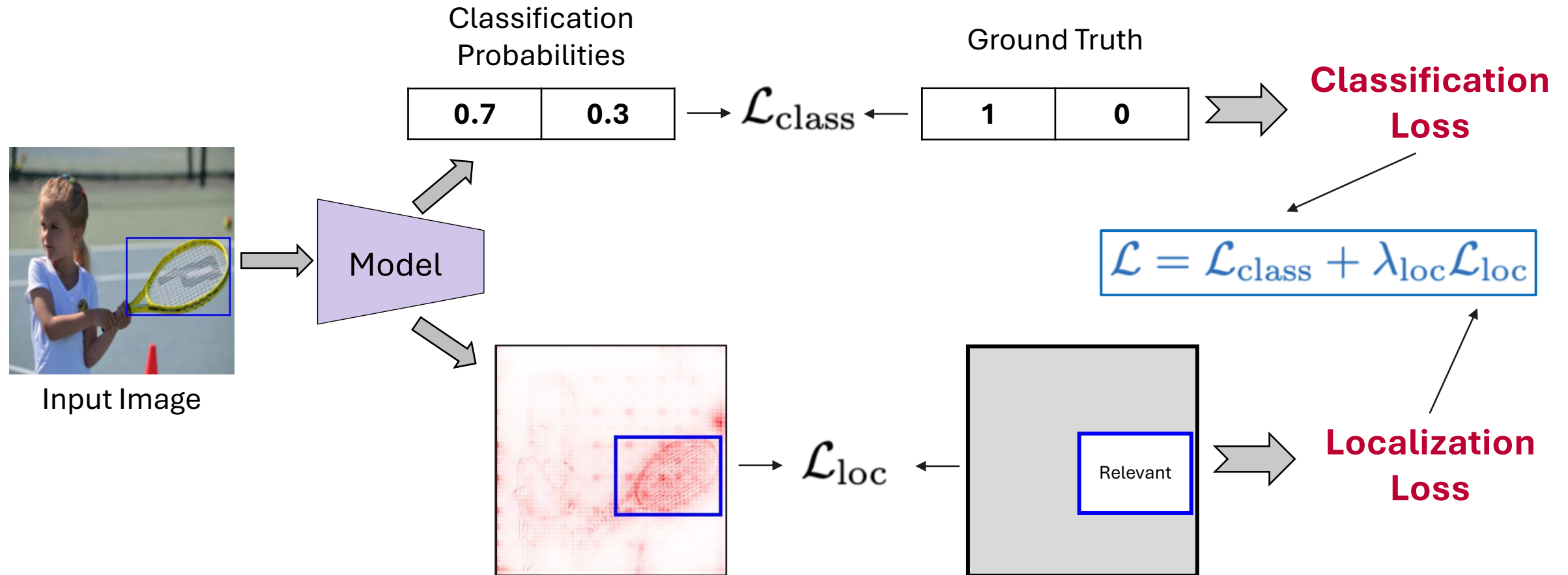
Methodology (Rao et al., 2023)



Methodology (Rao et al., 2023)

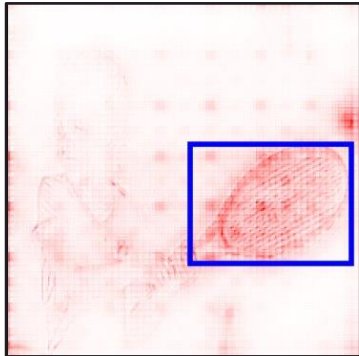


Methodology (Rao et al., 2023)



Considerations (Rao et al., 2023)

Attribution Methods



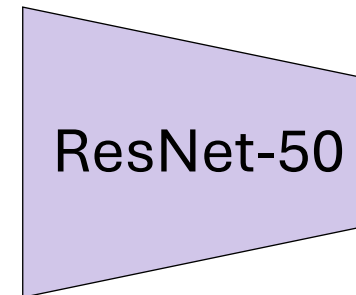
- IxG
- GradCam
- IntGrad
- B-cos

\mathcal{L}_{loc}



- L1
- Energy
- RRR*
- PPCE

Model



- Vanilla
- \mathcal{X} -DNN
- B-cos

Evaluation Metrics



- IoU
- EPG

Objectives

- 1. Reproducibility study (MLRC 2023 Challenge)**
 - Verify the main claims of Rao et al. (2023)
- 2. Extended work**
 - Address some limitations and propose solutions



Reproducibility Summary

Experiment

Claim

Comparing loss functions
for model guidance

R1

Issues with IoU

R2

R3

Comparing models and
attribution methods

R4

R5

Does not hold
for RRR* loss

Improving model accuracy
with model guidance

R6

Holds only for
B-Cos models

Efficiency and robustness
considerations

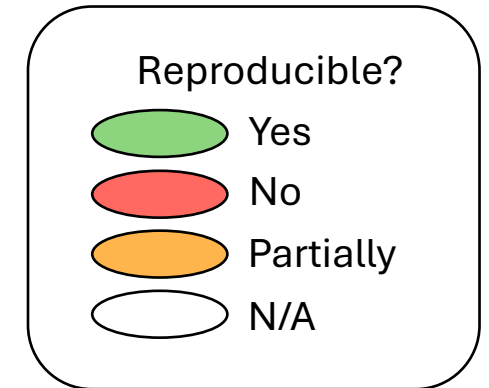
R7

R8

R9

Effectiveness against
spurious correlations

R10



Extensions

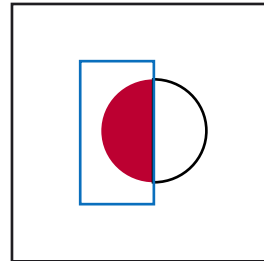
1. **X-SegEPG**
2. **EPG vs. Bounding box size**
3. **The impact of context on image classification**
4. **Segmentation masks and sparse annotations**



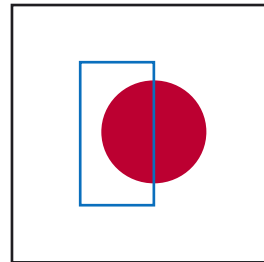
X-SegEPG

Energy-based Pointing Game (EPG)

Positive
attributions within
the bounding box



All positive
attributions



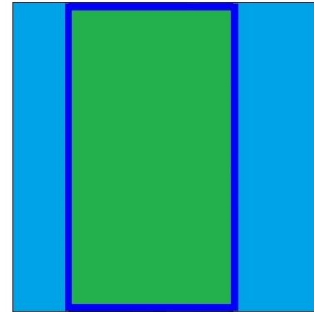
X-SegEPG

$$\text{EPG} = \frac{\text{[Diagram: Square with red circle and blue rectangle, partially overlapping]}}{\text{[Diagram: Square with red circle and blue rectangle, fully overlapping]}}$$

Input

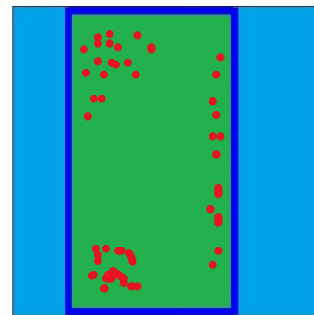


EPG



 Correct

 Incorrect



EPG = 1.0

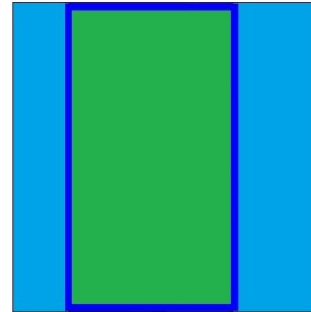
X-SegEPG

$$\text{EPG} = \frac{\text{[Diagram: Square with red circle and blue rectangle, circle partially outside rectangle]}}{\text{[Diagram: Square with red circle and blue rectangle, circle fully inside rectangle]}}$$

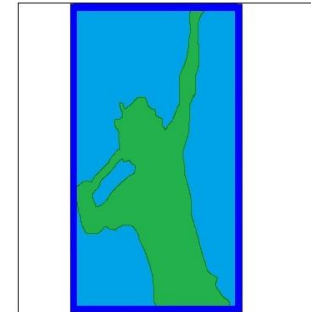
Input



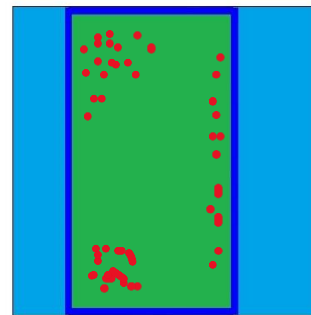
EPG



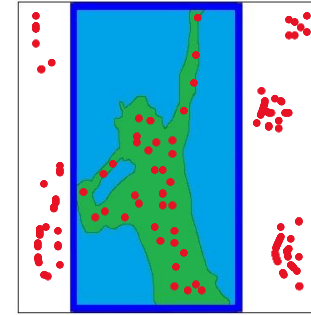
SegEPG



 Correct
 Incorrect



EPG = 1.0



SegEPG = 1.0

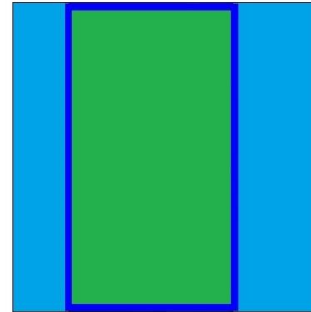
X-SegEPG

$$\text{EPG} = \frac{\text{[Diagram: Square with red circle and blue rectangle, circle partially outside]}{\text{[Diagram: Square with red circle and blue rectangle, circle fully inside]}}$$

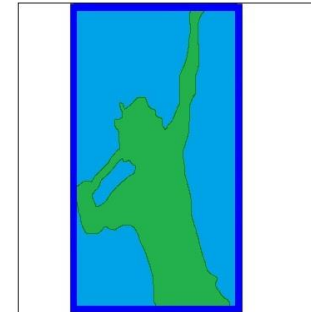
Input



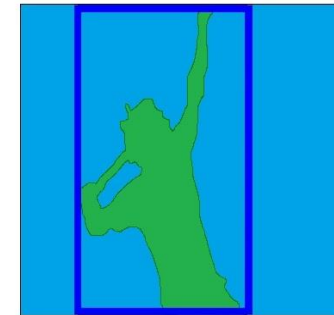
EPG



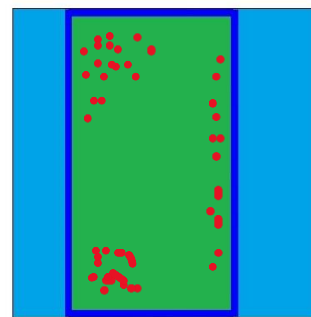
SegEPG



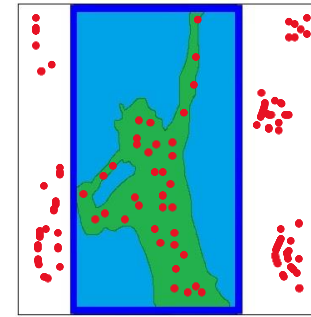
X-SegEPG



Correct
 Incorrect



EPG = 1.0



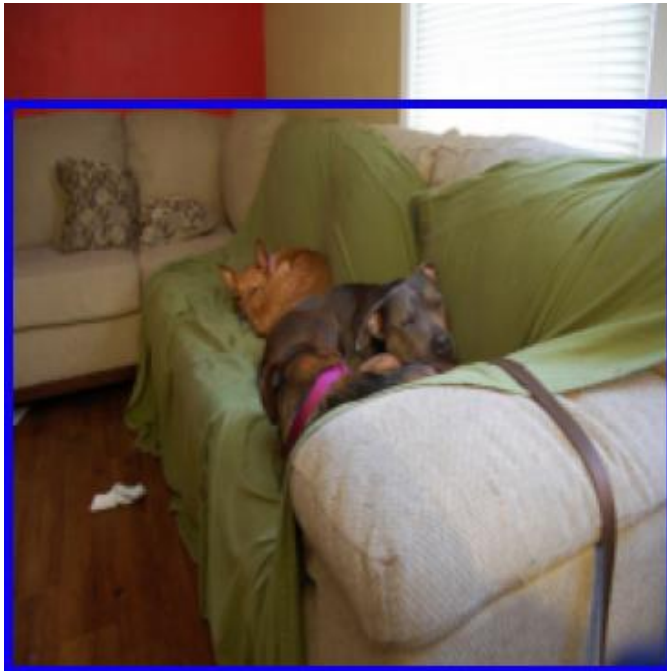
SegEPG = 1.0

X-SegEPG

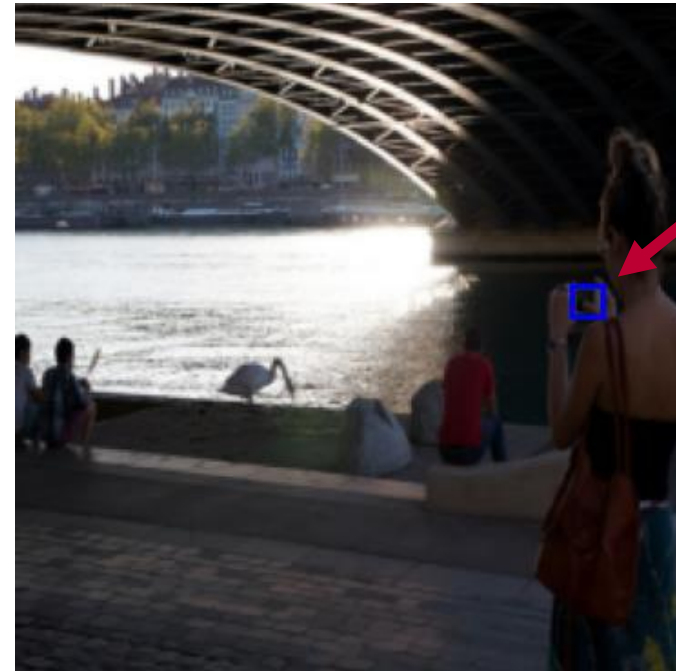
| EPG | | SegEPG | | X-SegEPG | |
|------------|------------|---------------|------------|-----------------|-----------|
| Val. | Std. | Val. | Std. | Val. | Std. |
| 0.35 | ± 0.23 | 0.33 | ± 0.25 | 0.16 | ± 0.2 |

EPG vs. Bounding Box Size

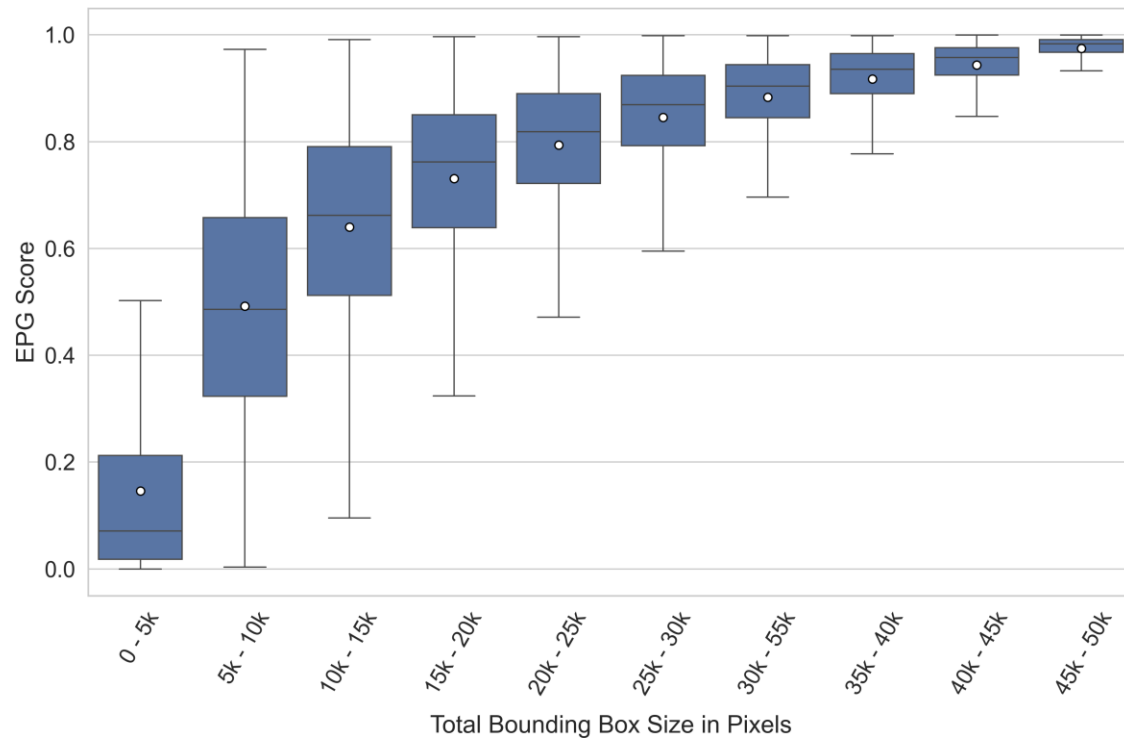
Class: Couch



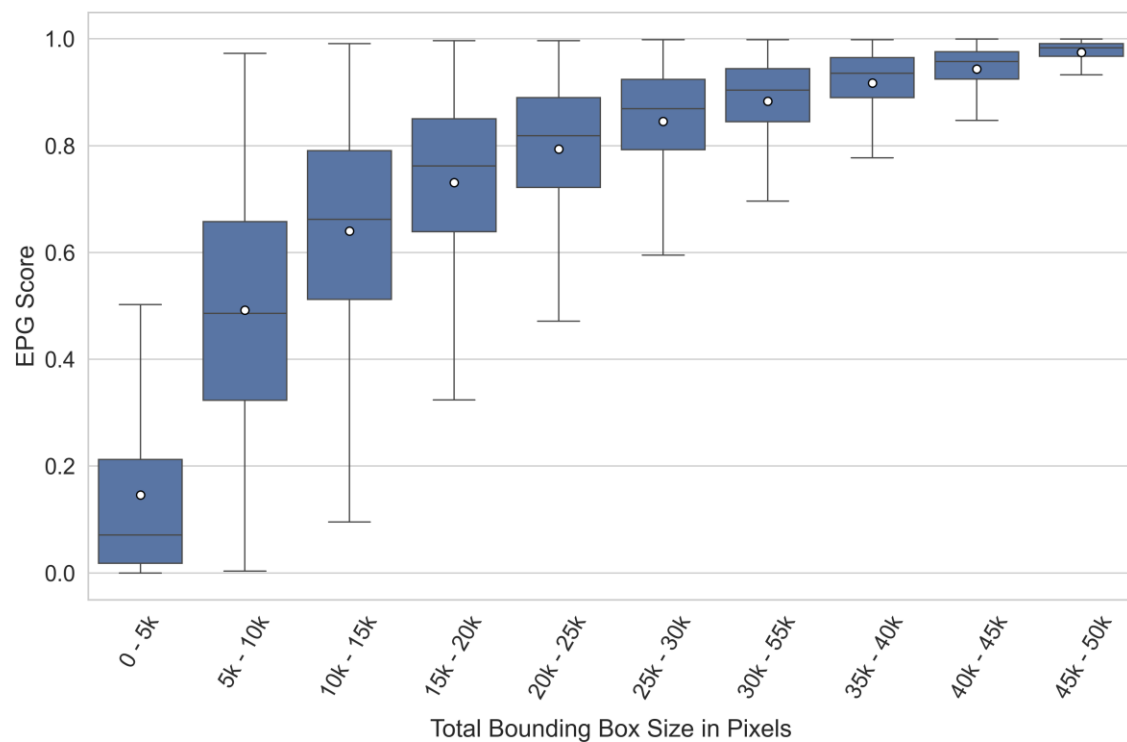
Class: Phone



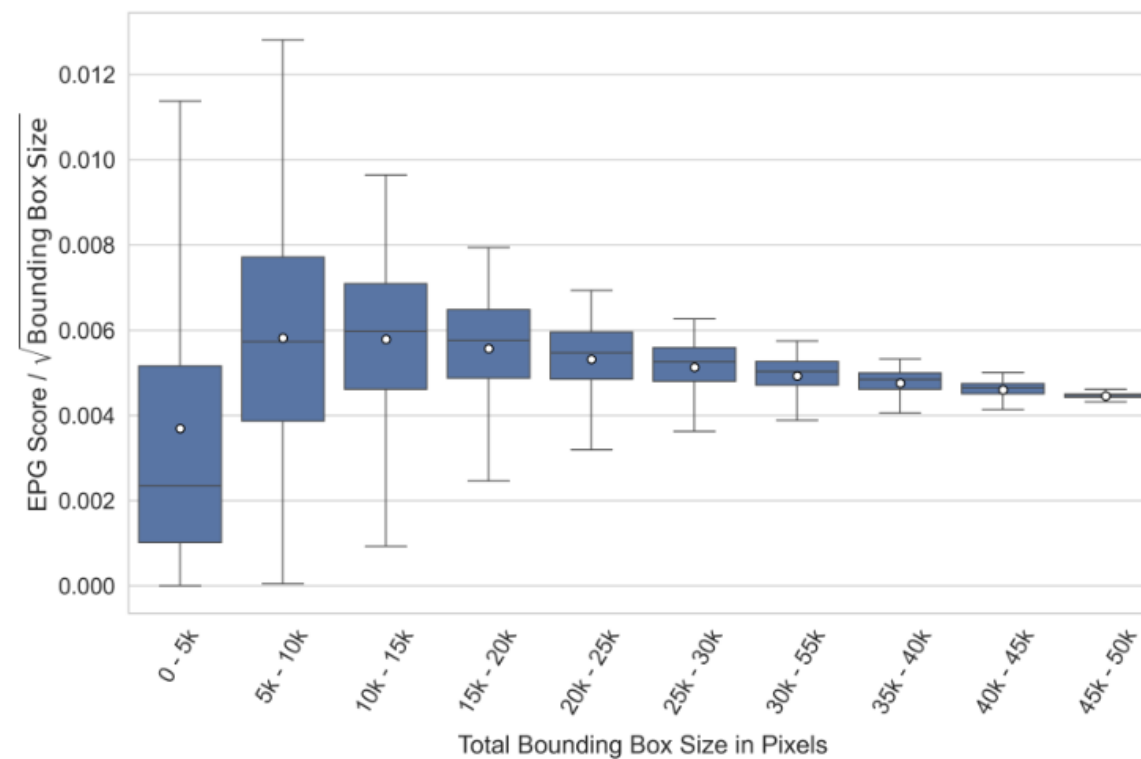
EPG vs. Bounding Box Size



EPG vs. Bounding Box Size



$$\mathcal{L}_{\text{Energy}^*} = 1 - \frac{1}{(\text{BB Area})^\alpha} \times \text{EPG}$$



Impact of Context

Is this a **real** or a **toy** car?

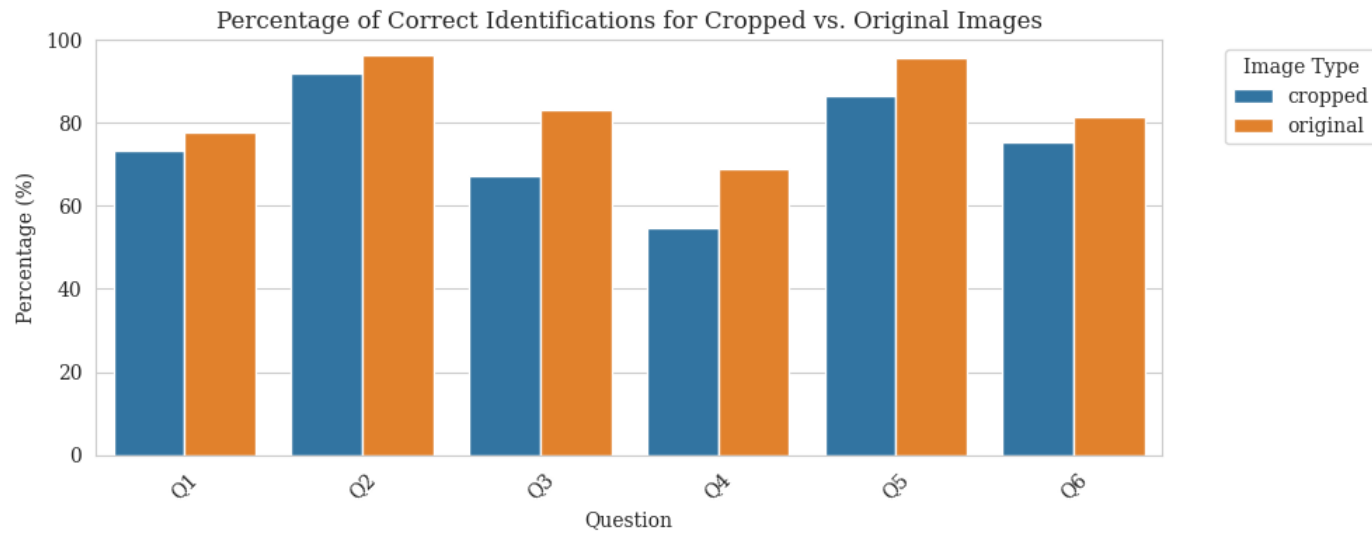


Impact of Context

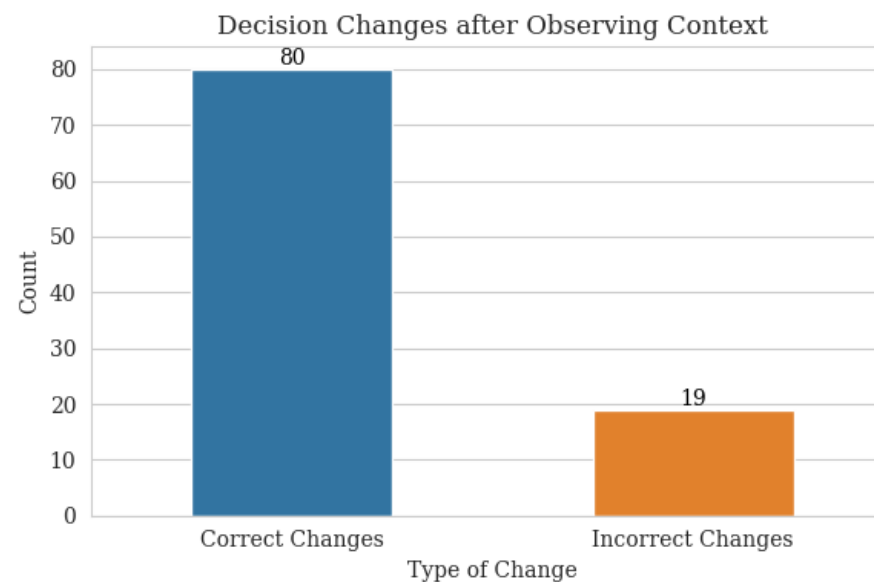
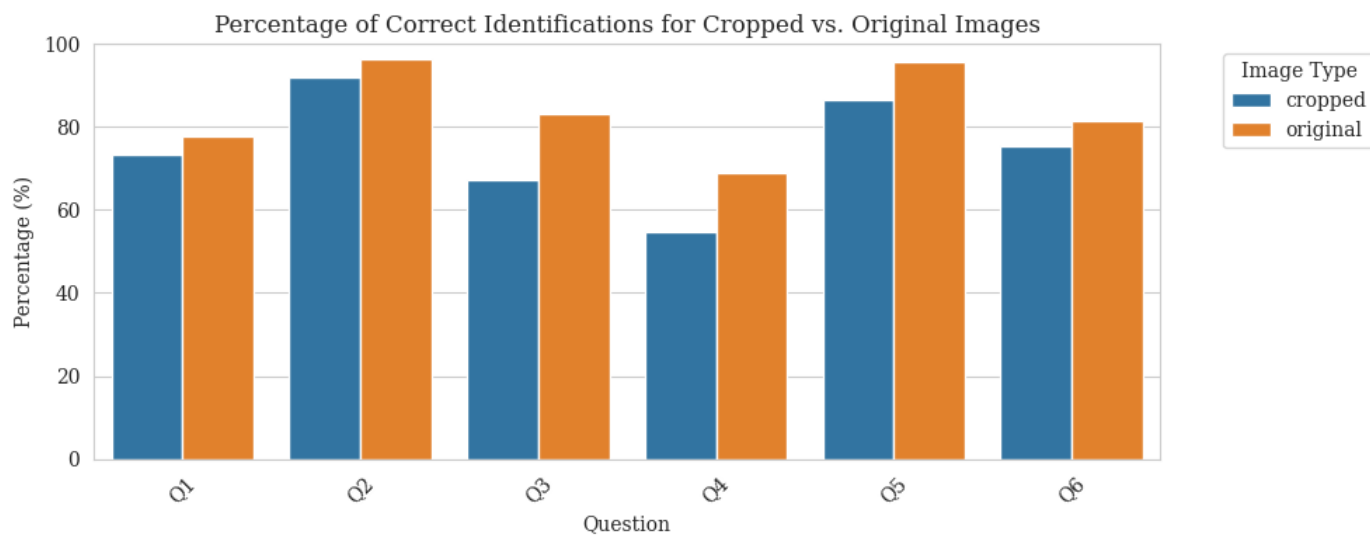
What about now?



Impact of Context

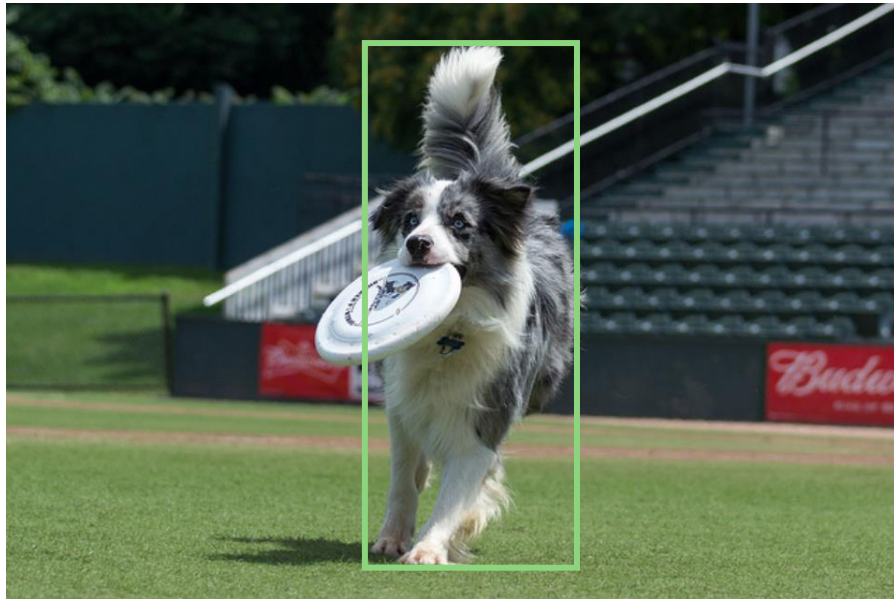


Impact of Context



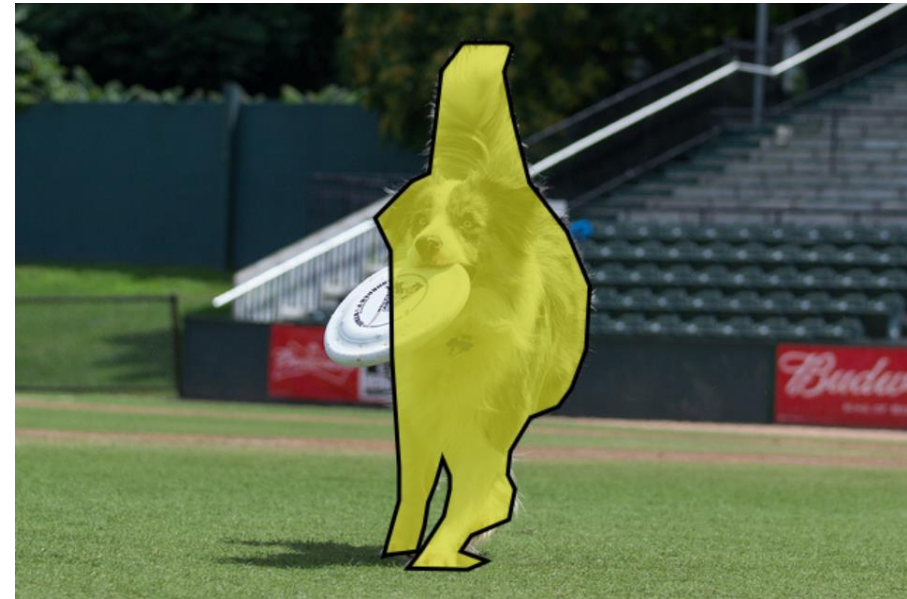
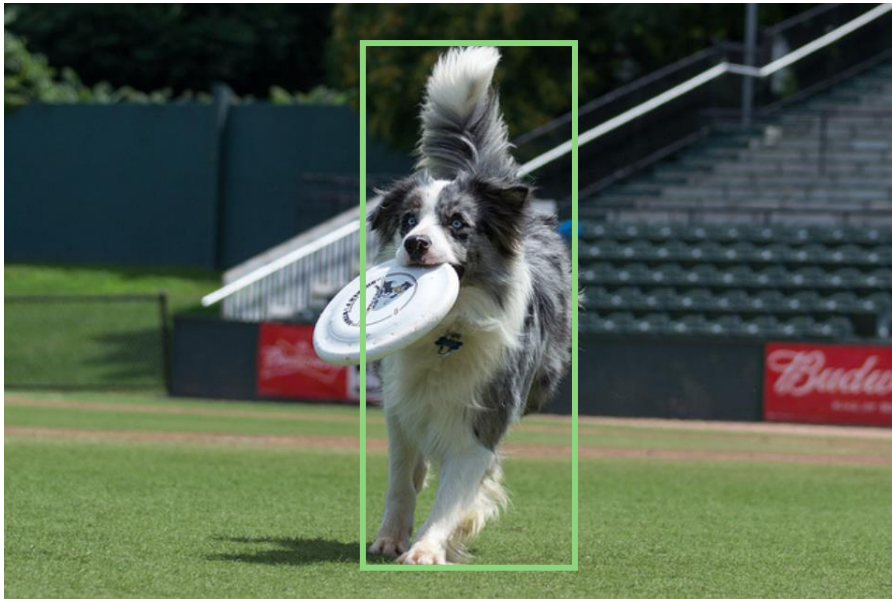
Segmentation Masks and Sparse Annotations

R10: Using as little as **1% of Bounding Boxes** improves model generalization

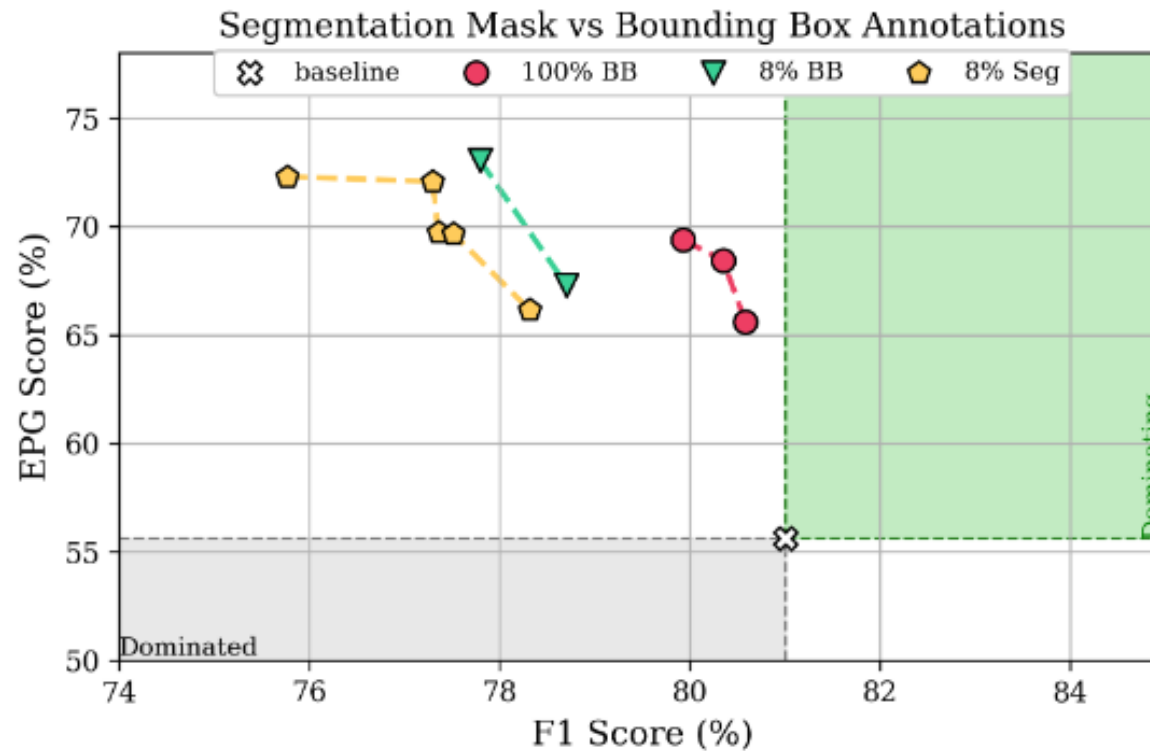


Segmentation Masks and Sparse Annotations

Does that mean that **Segmentation Masks** become affordable?



Segmentation Masks and Sparse Annotations





UNIVERSITY
OF AMSTERDAM

Thank you!

[1] 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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.

[2] Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1922–1933, 2023.

Icons: Flaticon.com

Survey images: https://drive.google.com/drive/folders/1cnnEceaKTaG456gvoUi14tzaL4IN_UU-?usp=drive_link