Fraunhofer Heinrich Hertz Institute
Technische Universität Berlin
BIFOLD

Revealing Hidden Context Bias in Segmentation and Object Detection through Concept-specific Explanations

Maximilian Dreyer, Reduan Achtibat, Thomas Wiegand, Sebastian Lapuschkin, Wojciech Samek

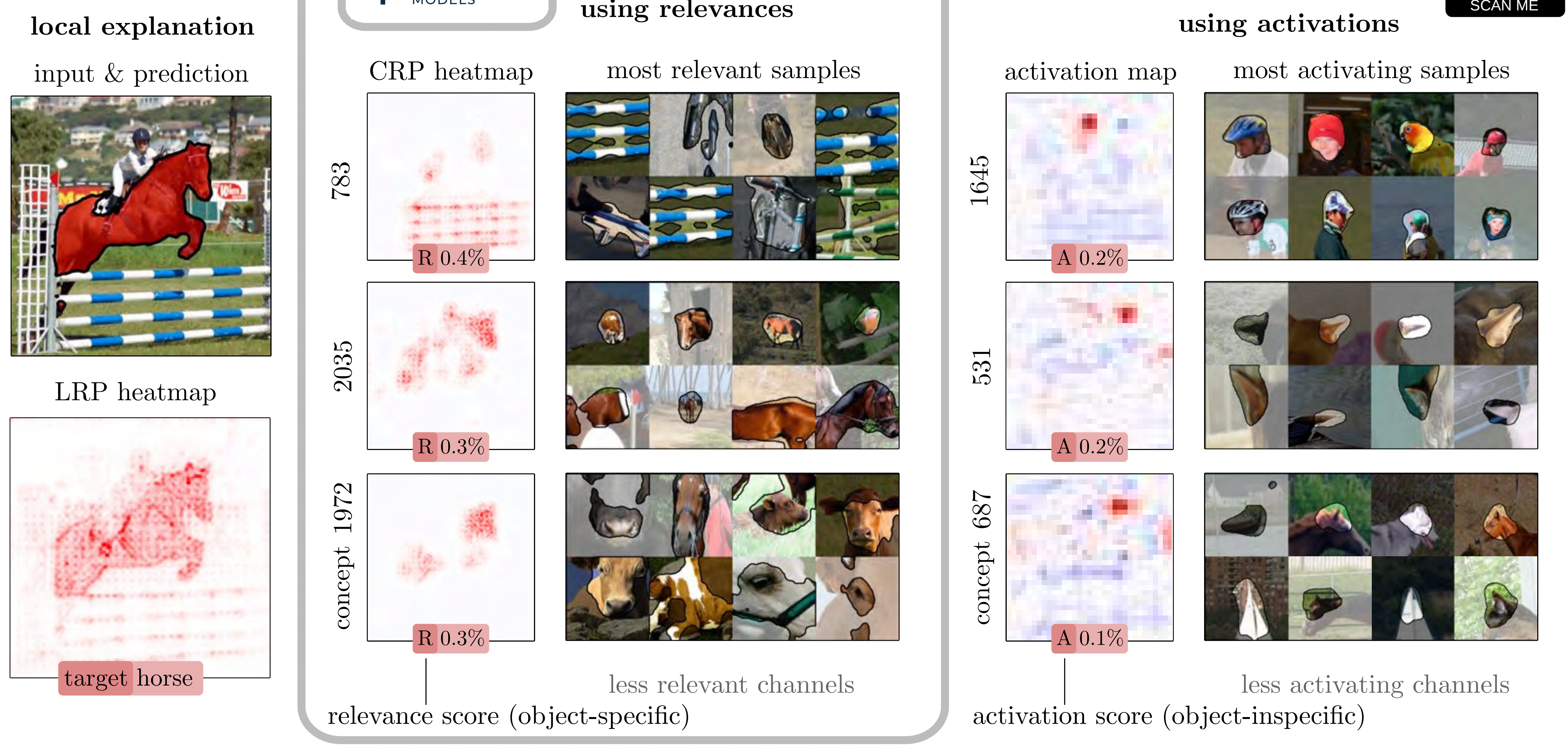
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Concept-specific Explanations for Localization Models

A Overview

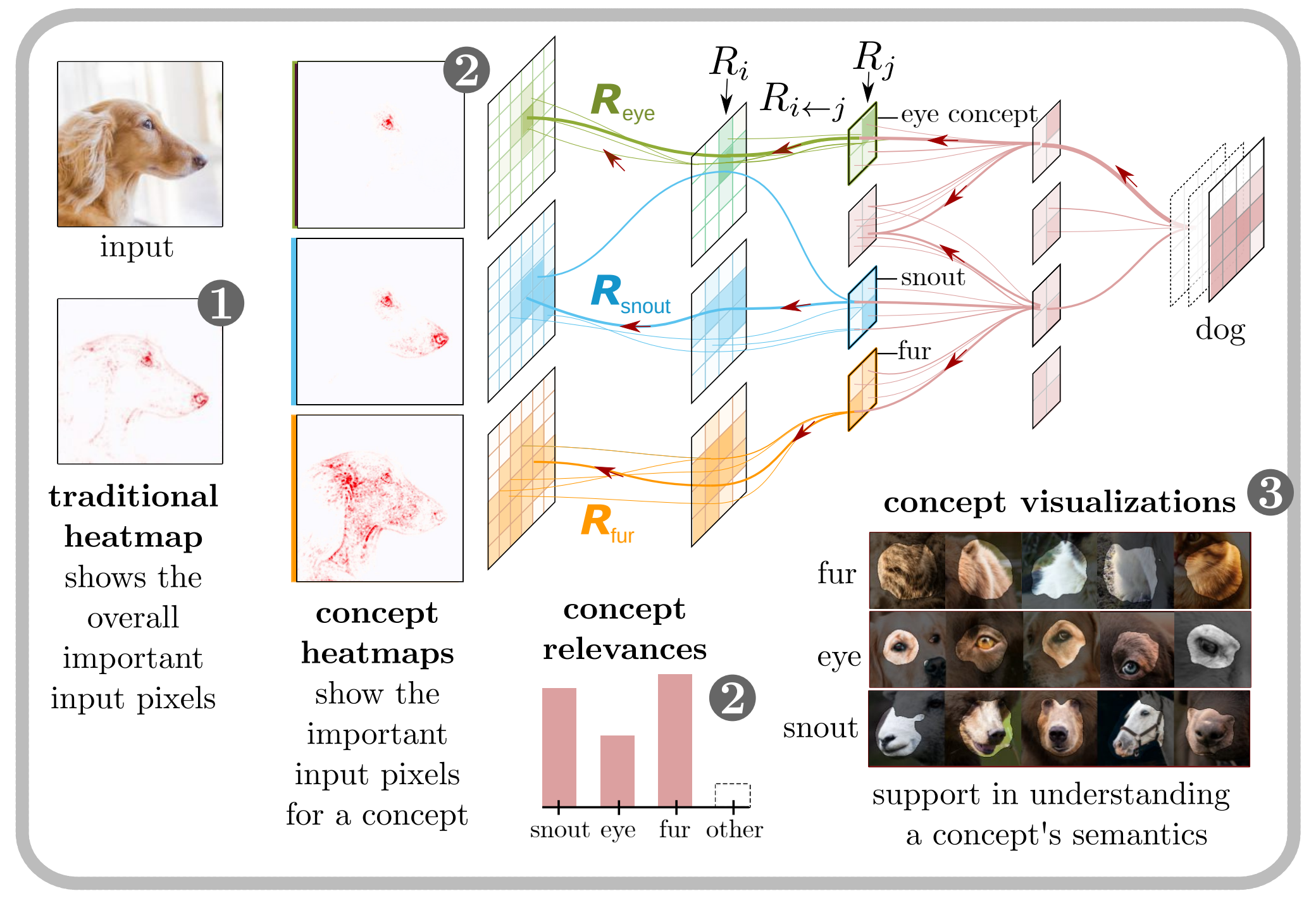
Want to generate explanations, too?
Check our repository! We currently support SSD, YoloV5+V6, DeepLabV3+ and UNet.

SCAN ME



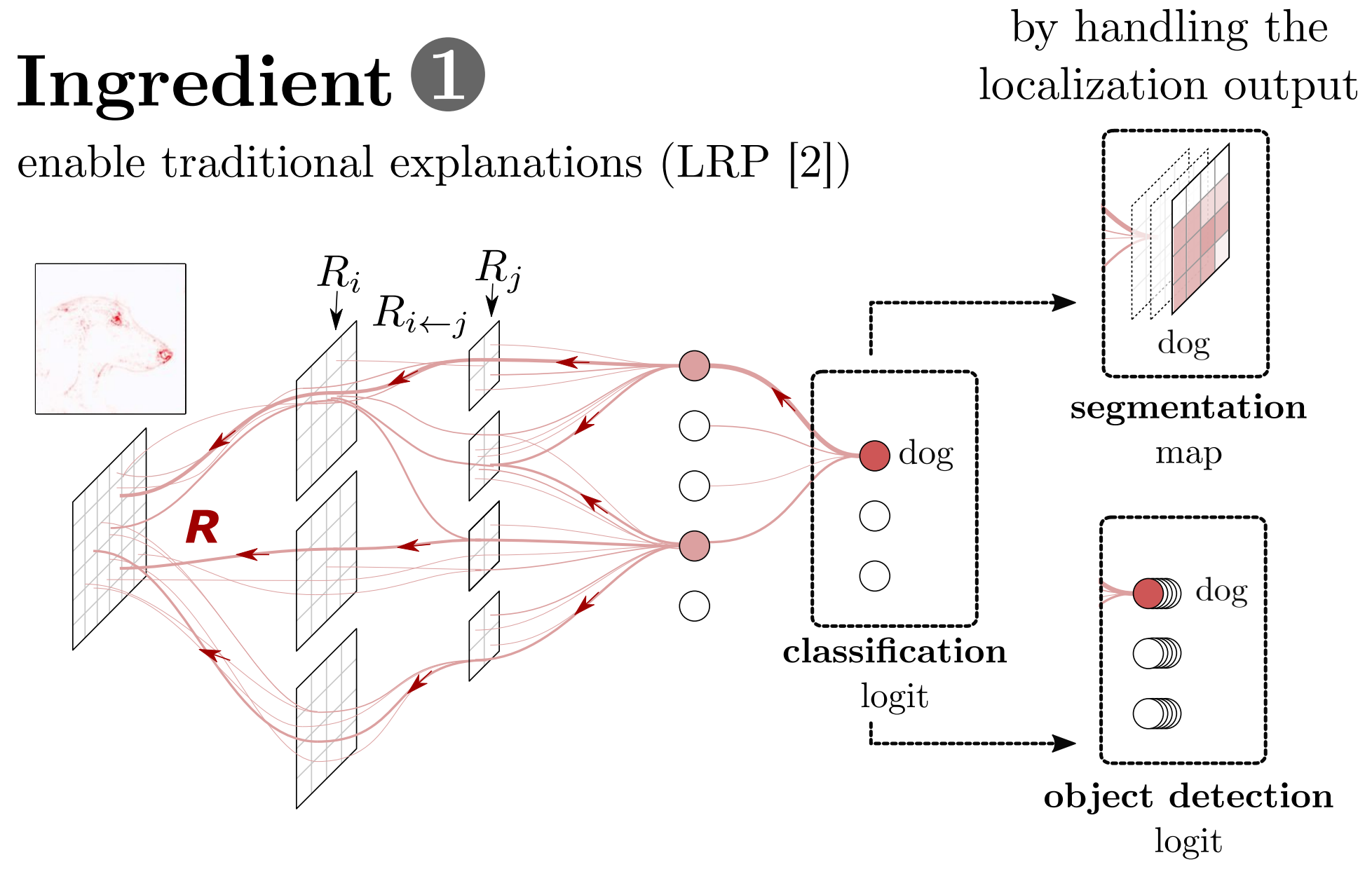
L-CRP communicates the most relevant concepts used in predictions of segmentation or object detection models.

B How does L-CRP work?



Ingredient ❶

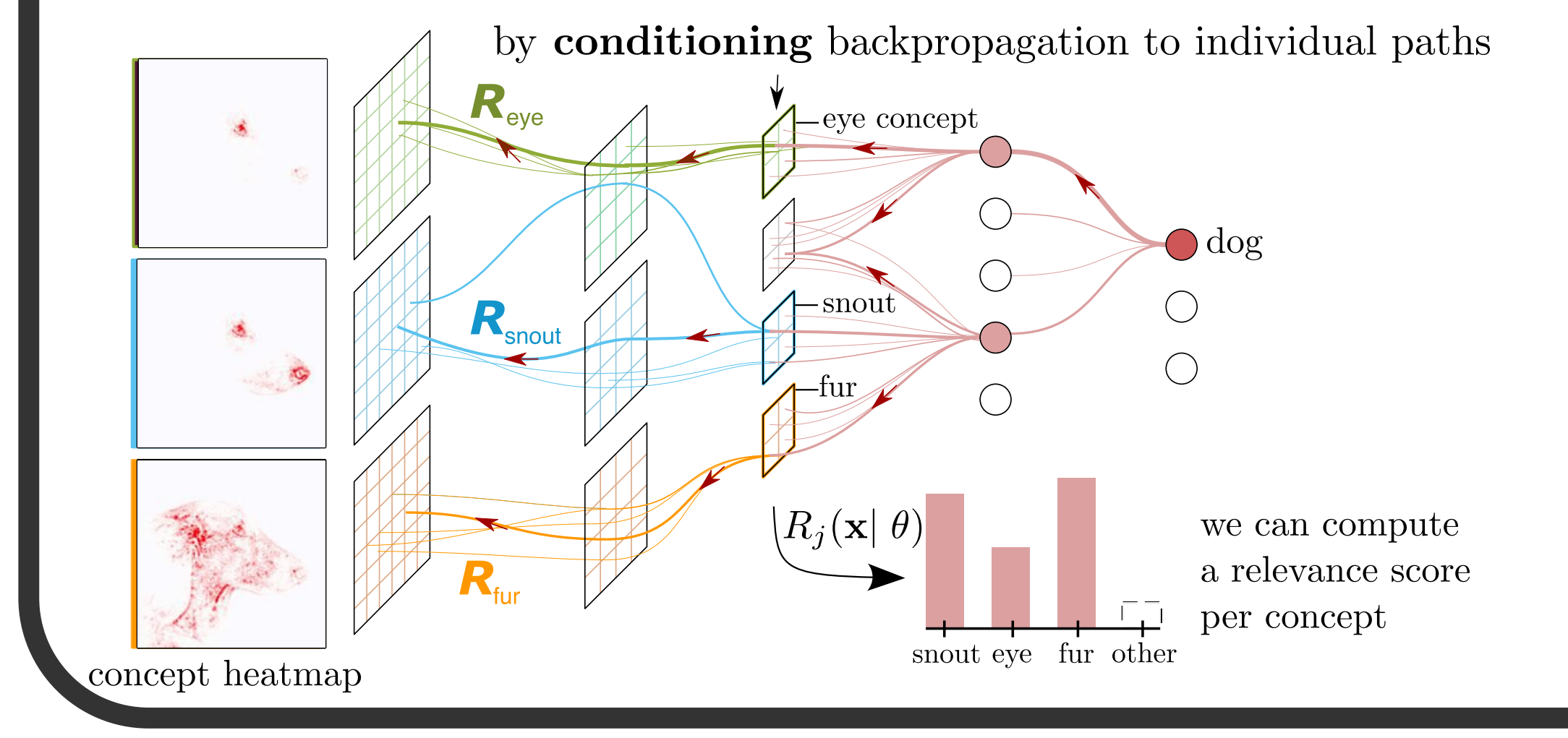
enable traditional explanations (LRP [2])



Other attribution methods can be used with L-CRP!
See paper for comparison regarding faithfulness and complexity.

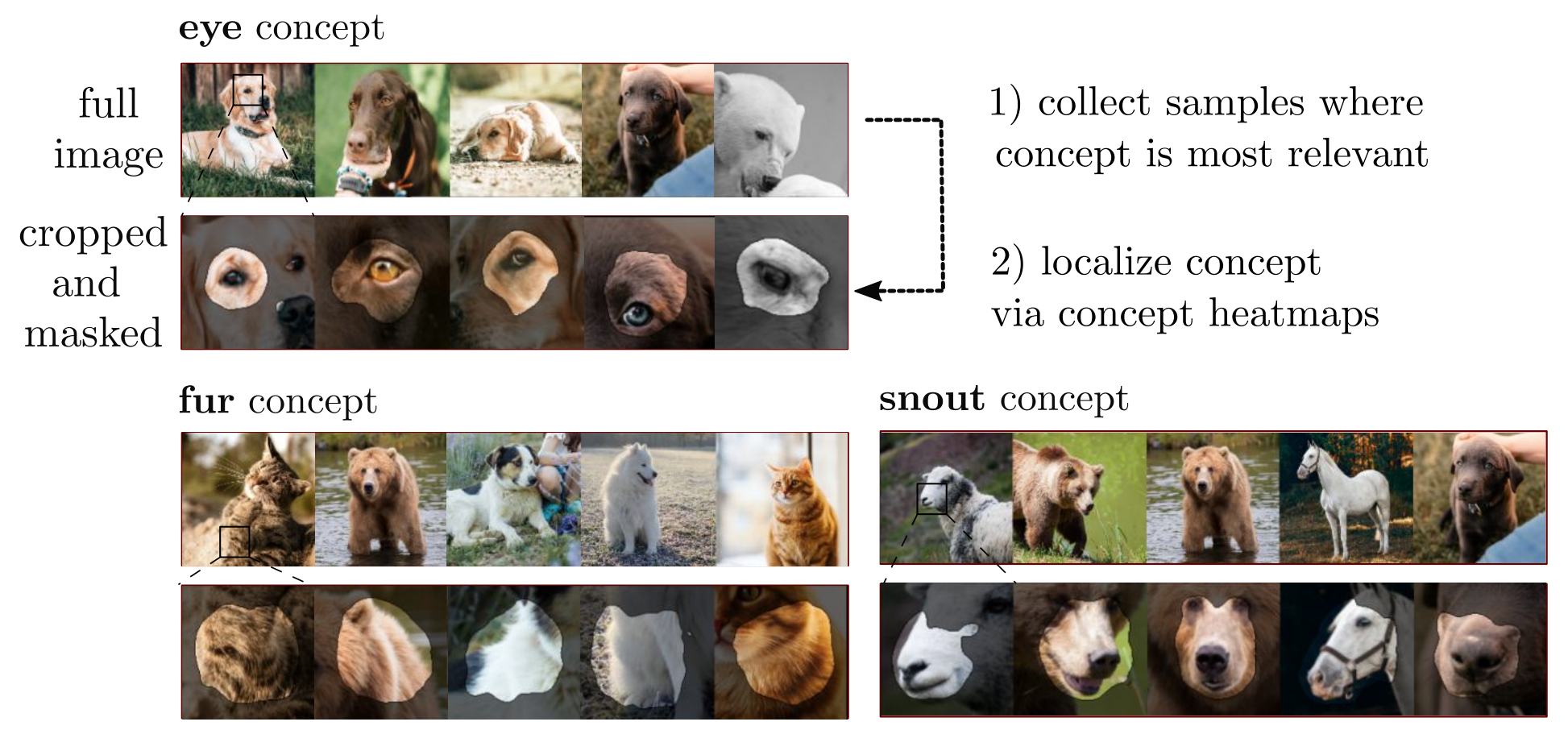
Ingredient ❷

enable concept-conditional explanations with CRP [1]



Ingredient ❸

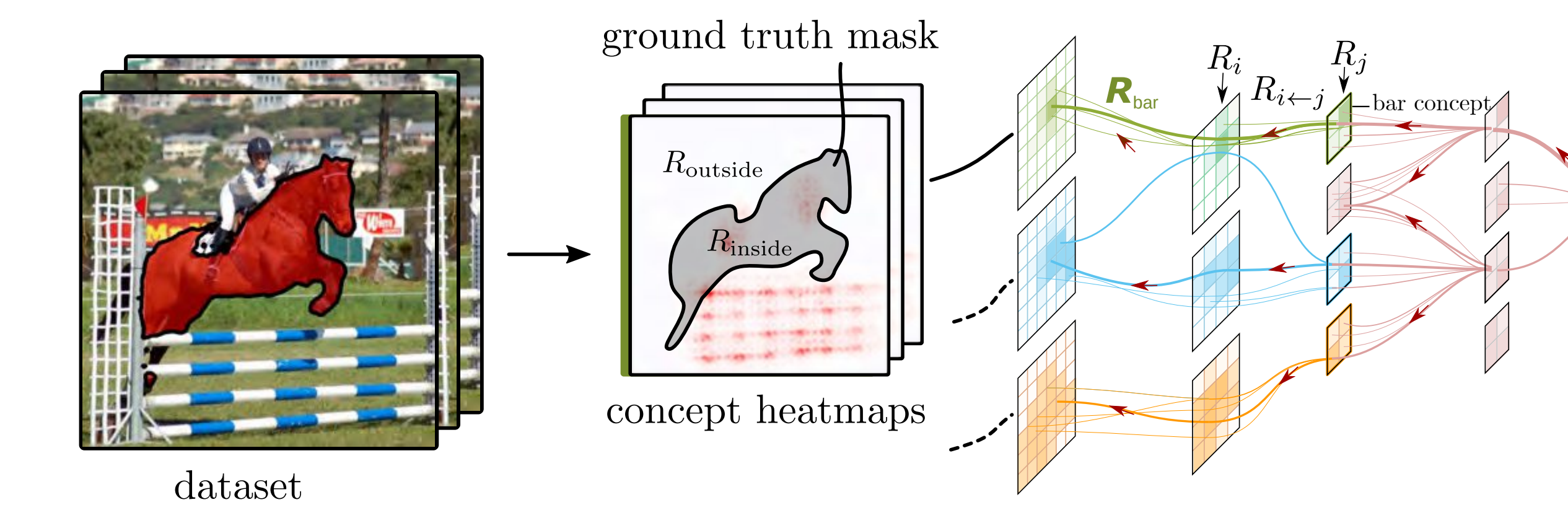
visualize concepts with RelMax [1] and concept heatmaps



Revealing Hidden Context Bias

A Computing Context Scores

We propose L-CRP context scores to find concepts that are highly used for background features.



context score is given by

$$C = \frac{R_{\text{outside}}}{R_{\text{outside}}+R_{\text{inside}}}$$

amount of relevance on background

B The Best Way to Compute Context Scores?

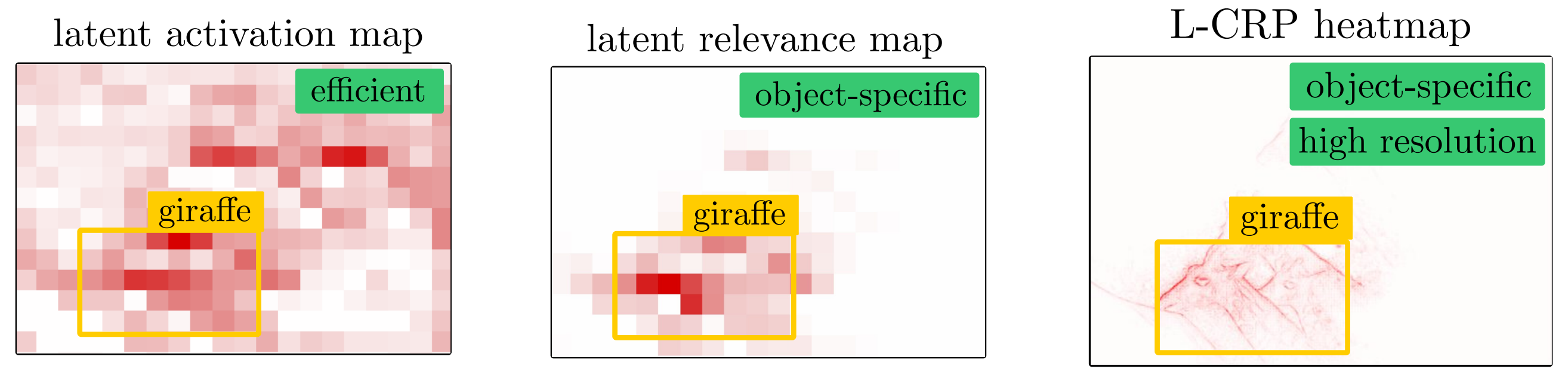
In principle, context scores can be computed by other means, however:

giraffe prediction

investigated concept

B1 Qualitative Evaluation

L-CRP provides both **object-specificity** and **high resolution**, leading to the most faithful context scores.



B2 Quantitative Evaluation

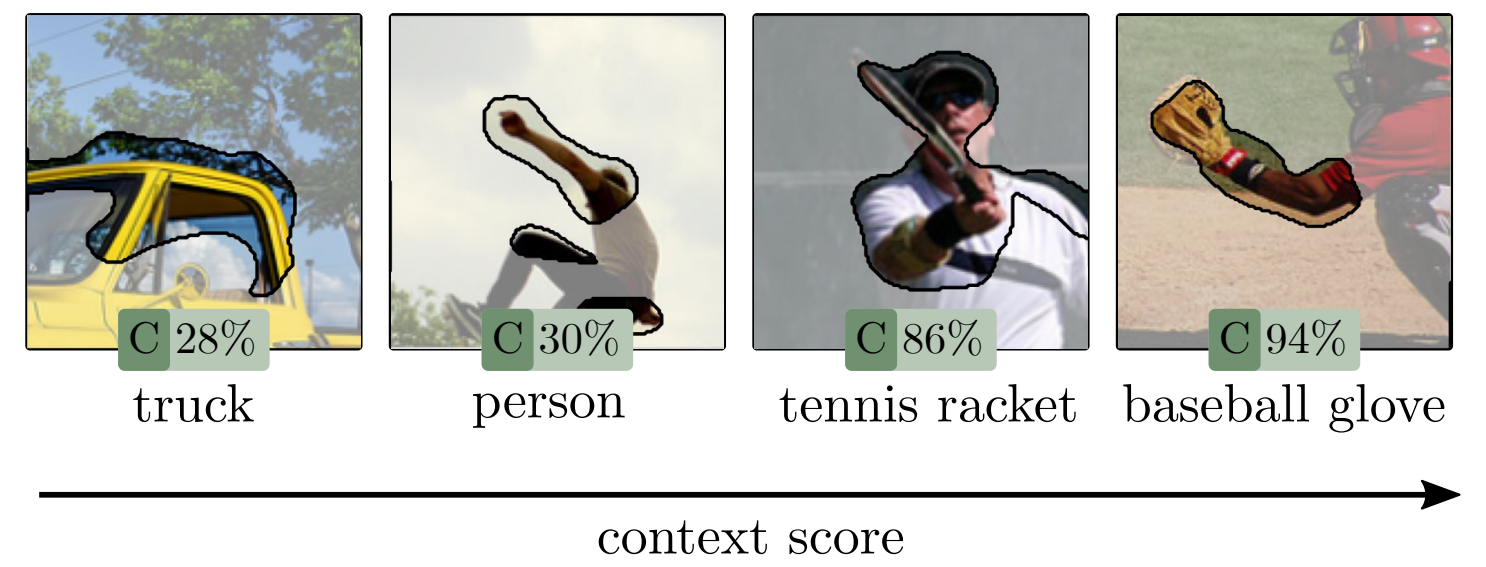
To quantitatively check the context scores' (C) faithfulness, we propose to evaluate the **background sensitivity** (S) of concepts by perturbing the object background and measuring the effect on concept relevances.
L-CRP leads to the highest correlation and smallest RMSD between C and S.

C Example of Context Bias

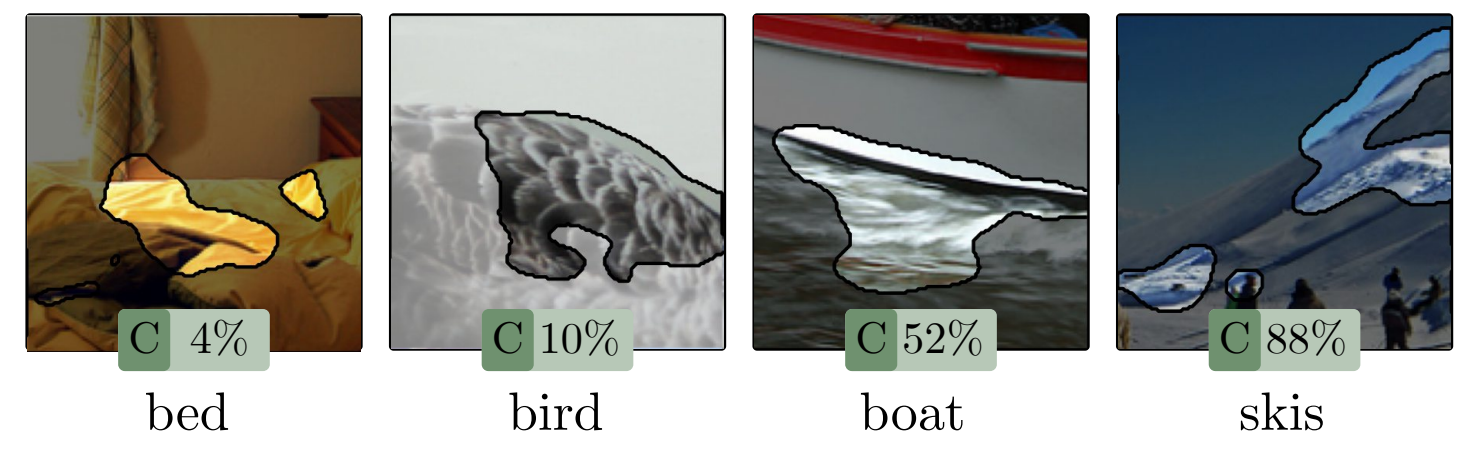
Examples of concepts with a high contex score for the person class of the YOLOv5 model trained on MS COCO 2017:

D Context Scores are Class-specific

We observe, that concepts are used differently throughout classes.
An 'arm'-like concept can be used to detect a person, but it is, in this example, also used to detect tennis rackets or baseball gloves.

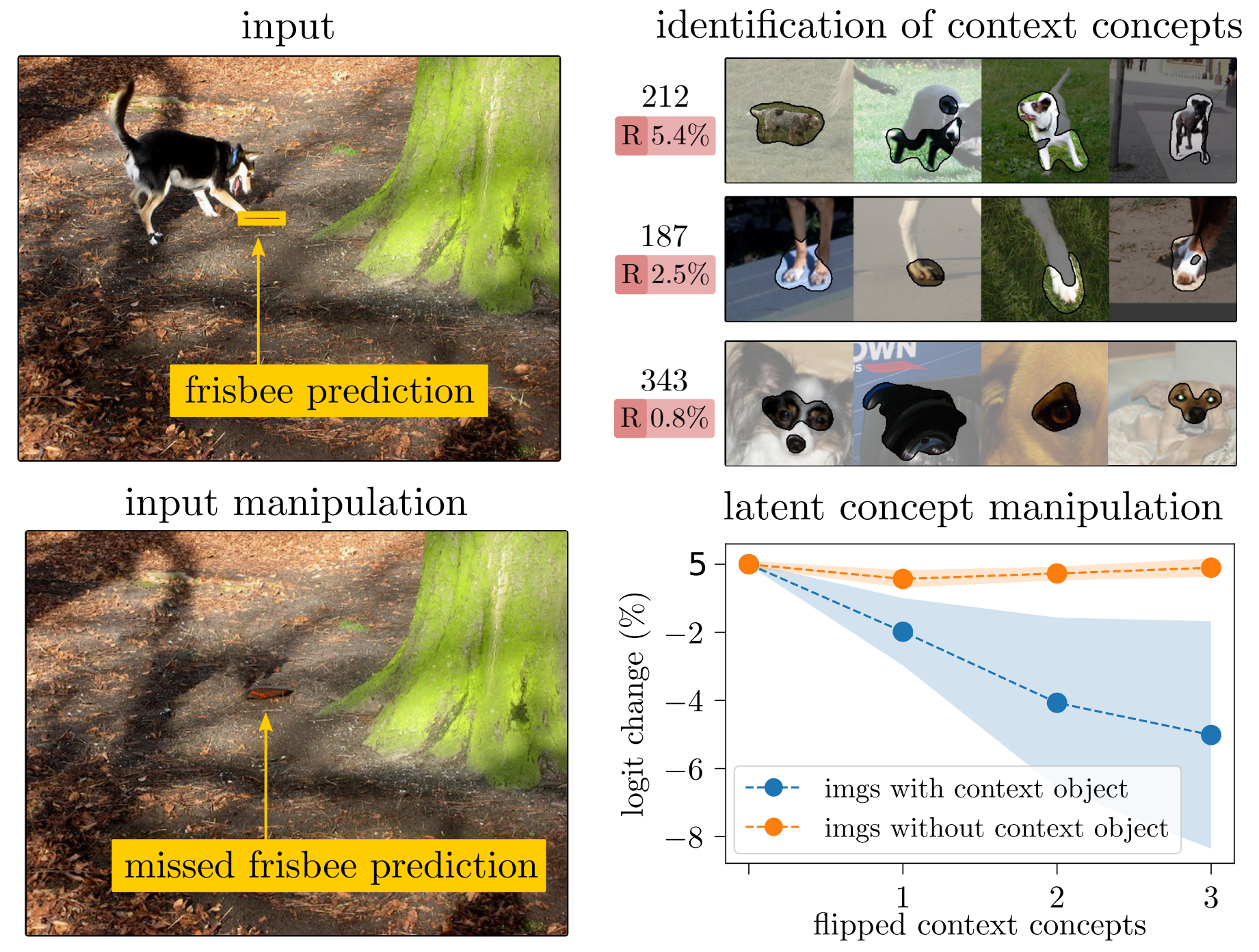


A 'wavy'-like concept can be used to detect a bed (blanket), boat (water) or skis (snow on mountain).



Interacting with the Model

With L-CRP, we can automatically find relevant concepts that focus strongly on the background.
As we know which neurons are responsible, we can start interacting with the model.

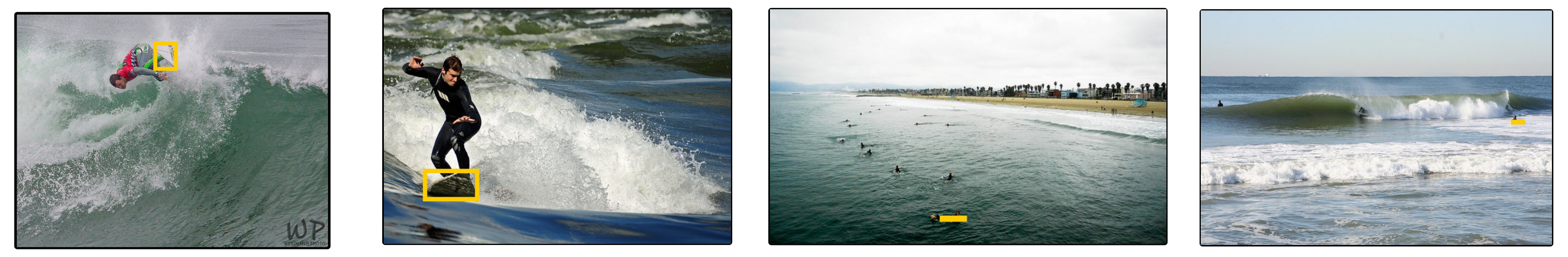


To give an example, we have found that the YOLOv5 model uses background concepts to detect frisbees. As frisbees are often depicted together with a dog, the model has learned to often rely on **dog-concepts for frisbee detections.**

To probe the model, we remove the dog from the image, leading to a missed frisbee prediction, confirming the use of dog features.
We can further flip (remove) the corresponding neurons (background concepts) and can measure a decrease in the output logit.
This decrease also correlates with the presence of a dog in an image.

In fact, objects of different classes occur often together in the MS COCO dataset, leading to background biases. Other examples include the person class and frisbees or surfing boards.

Flipping only a few background neurons can lead to missed surfing board predictions, as in the examples shown to the right.



Conclusion and Outlook

- L-CRP enables **concept-based explanations** for segmentation and object detection models.
- Having ground truth object localizations available allows us to measure to what extend **concepts focus on the background**.
  - Compared to using latent activation or relevance maps, L-CRP offers the most faithful context scores due to **object-specificity** and **high resolution** (input-level resolution).
  - We can find several background concepts, often corresponding to objects of another class. Such model behavior is not surprising, as objects are often depicted together with another object class (e.g. frisbees and dogs), allowing the model to learn short-cuts (e.g. background biases).
- With L-CRP, we can localize the background concepts in the model architecture, enabling us to **interact with the model.**
- In future work, it will be interesting to unlearn the background bias.

[1] Achtibat, Reduan, et al. "From 'Where' to 'What': Towards Human-Understandable Explanations through Concept Relevance Propagation." arXiv preprint arXiv:2206.03208 (2022).
[2] Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

BerDiBa
BERLINER DIGITALER BAHNBETRIEB

This work was partly supported by the German Ministry for Education and Research under grants [BIFOLD (01IS18025A, 01IS18037I)], the European Union's Horizon 2020 research and innovation programme as grant [iToBoS (965221)], the German Research Foundation (ref. DFG KI-FOR 5363) and the state of Berlin within the innovation support program ProFIT as grant [BerDiBa (10174498)].

Link to our Paper →

SCAN ME