Efficient and Concise Explanations for Object Detection with Gaussian-Class Activation Mapping Explainer

Khanh Nguyen[1], Hung Nguyen[1,2*], Khang Nguyen[1], Binh Truong[1], Tuong Phan[3], Hung Cao[2]

[1]Quy Nhon AI, FPT Software, Vietnam
[2]Analytics Everywhere Lab, University of New Brunswick, Canada
[3]University of Waterloo, Canada

Outline

01 Motivation

02 Methodology

03 Experiments and Results

04 Conclusion

Motivation

Why AI Trustworthy and Transparency?

In sensitive contexts…

AVOIDING BLAME | FEB 12, 1:09 PM EST *by* VICTOR TANGERMANN

Tesla Driver Says He's Not Sure If He Killed a Pedestrian Because He Was on Autopilot

This is getting ridiculous.

/ Advanced Transport / Autopilot / Full Self Driving / Tesla

Tangermann, V. (2024, February 12). *Tesla Driver Says He's Not Sure If He Killed a Pedestrian Because He Was on Autopilot*. Futurism; Futurism. https://futurism.com/tesla-driver-not-sure-full-self-driving

Opinion | When a Computer Program Keeps You in Jail (Published 2017). (2024). *The New York Times*. https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

INNOVATIONS

ChatGPT invented a sexual harassment scandal and named a real law prof as the accused

The AI chatbot can misrepresent key facts with great flourish, even citing a fake Washington Post article as evidence

By Pranshu Verma and Will Oremus
April 5, 2023 at 2:07 p.m. EDT

Verma, P., & Oremus, W. (2023, April 5). *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused*. Washington Post; The Washington Post. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/

Stanford MEDICINE | News Center

Email Tweet

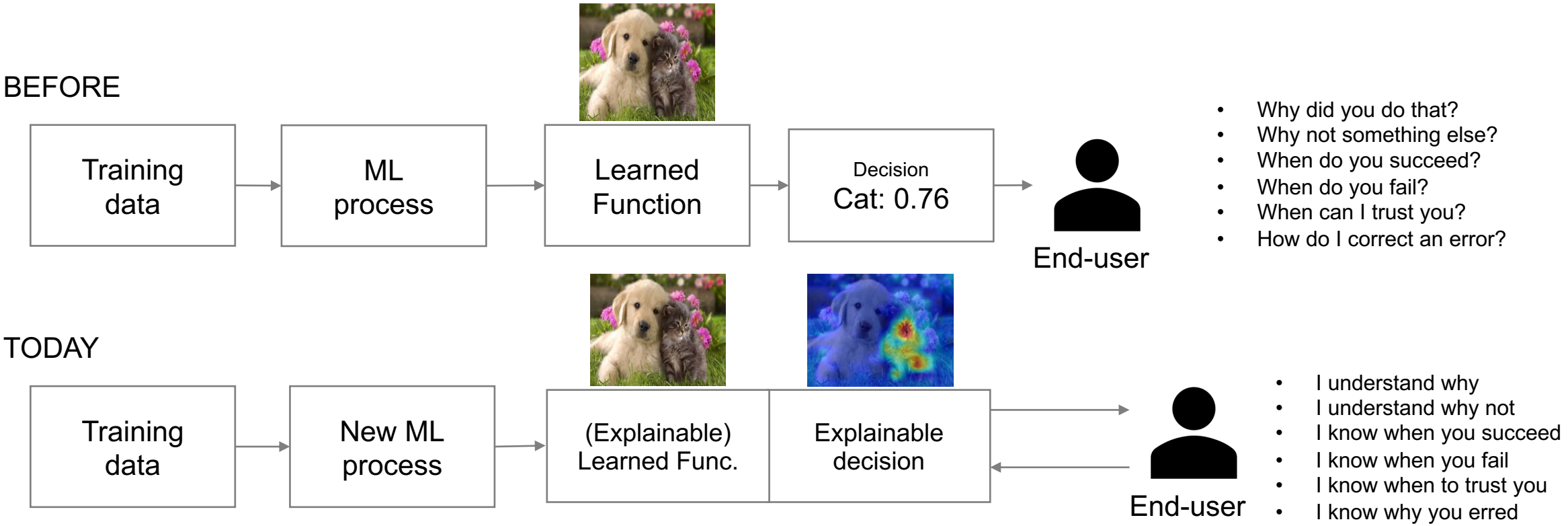
Researchers say use of artificial intelligence in medicine raises ethical questions

In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

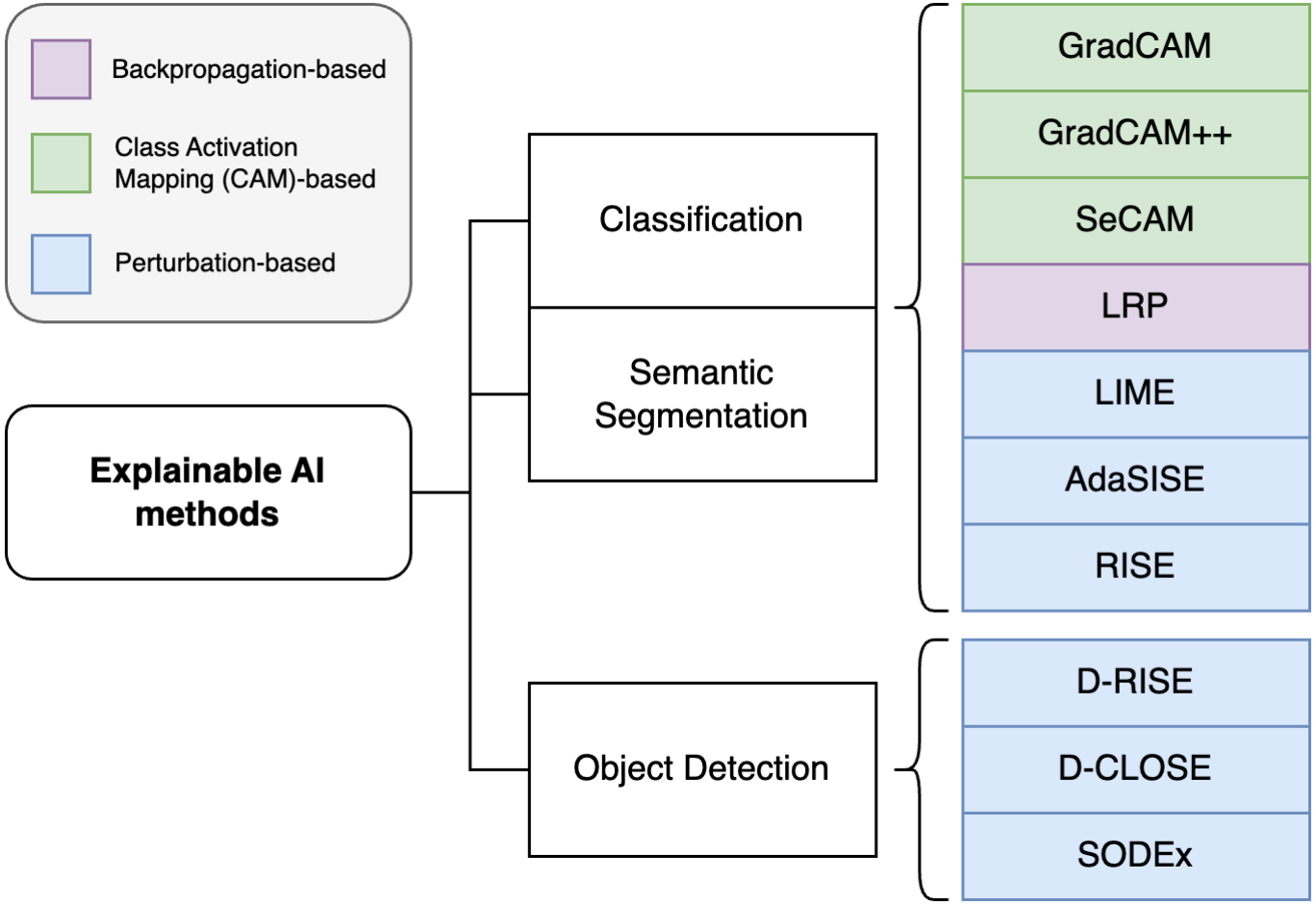
Patricia Hannon ,https://med.stanford.edu/news/all-news/2018/03/researchers-say-use-of-ai-in-medicine-raises-ethical-questions.html

Explainable AI (XAI)

To elevate the interpretability of model's decisions



Existing XAI methods for Computer Vision

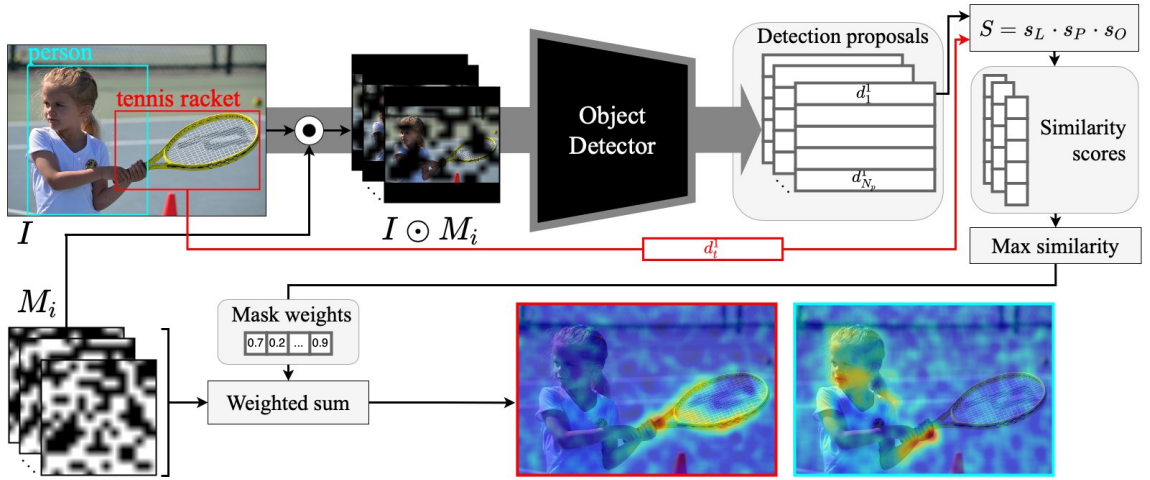


Existing XAI methods for Object Detection

D-RISE, D-CLOSE, SODEx

Only perturbation-based XAI methods for object detection:

- **Hyperparameters sensitivity:** many potential explanations for a single object.
- **Careful fine-tuning hyperparameters:** to achieve a clear and satisfactory explanation
- **Long running time:** to perturb images and generate an explanation.



D-RISE Architecture (Petsiuk et al., 2021)

G-CAME – Gaussian Class Activation Mapping Explainer

By adding the Gaussian kernel as the weight for each pixel in the feature map, G-CAME:

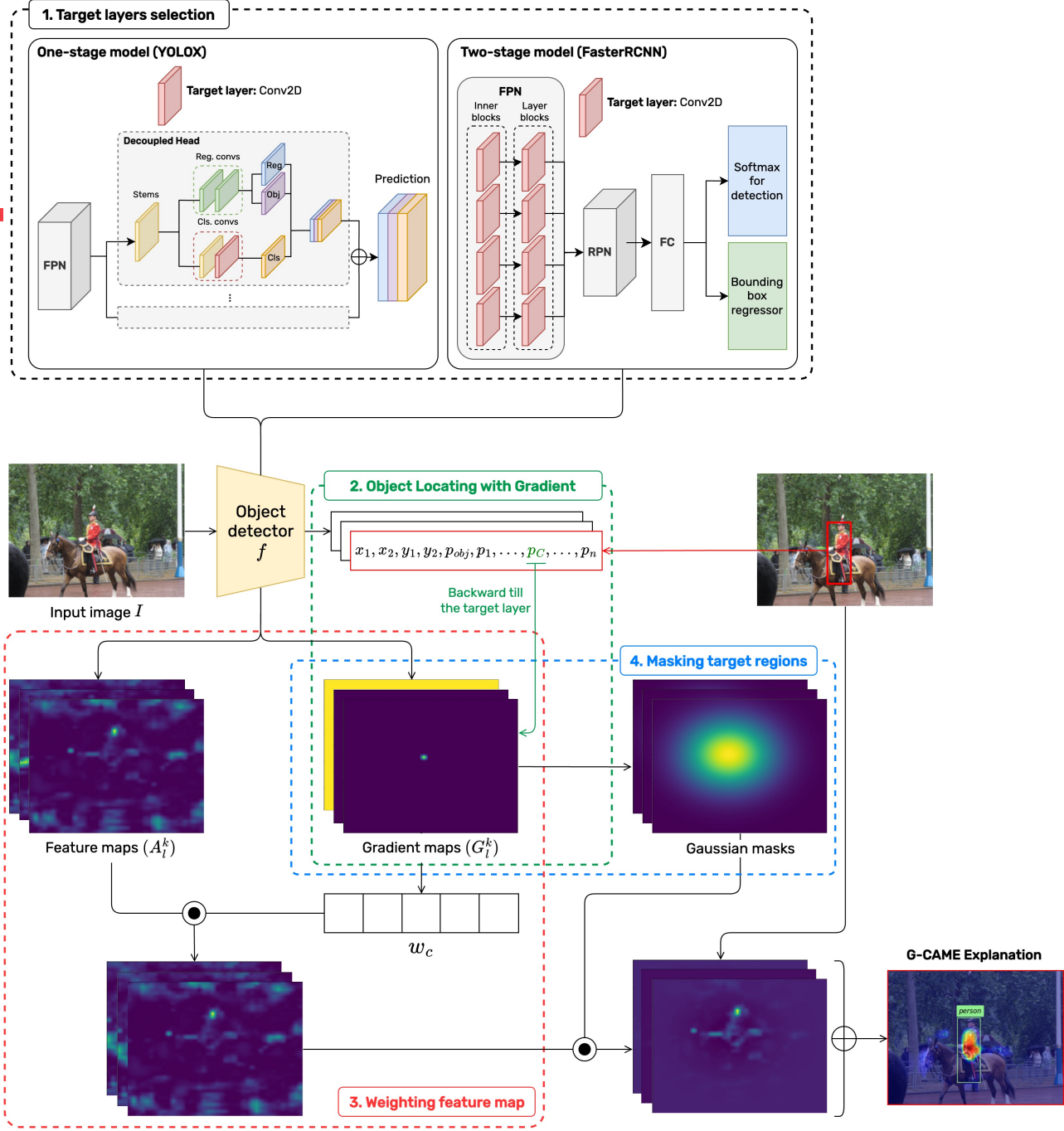
- Becomes the first CAM-based method that can explain object detectors for a specific target object.
- Runs in a short time manner compared with perturbation-based methods.
- Produces better plausible and information-faithful explanations than previous methods.

Methodology

G-CAME Architecture

4 Blocks to generate an explanation

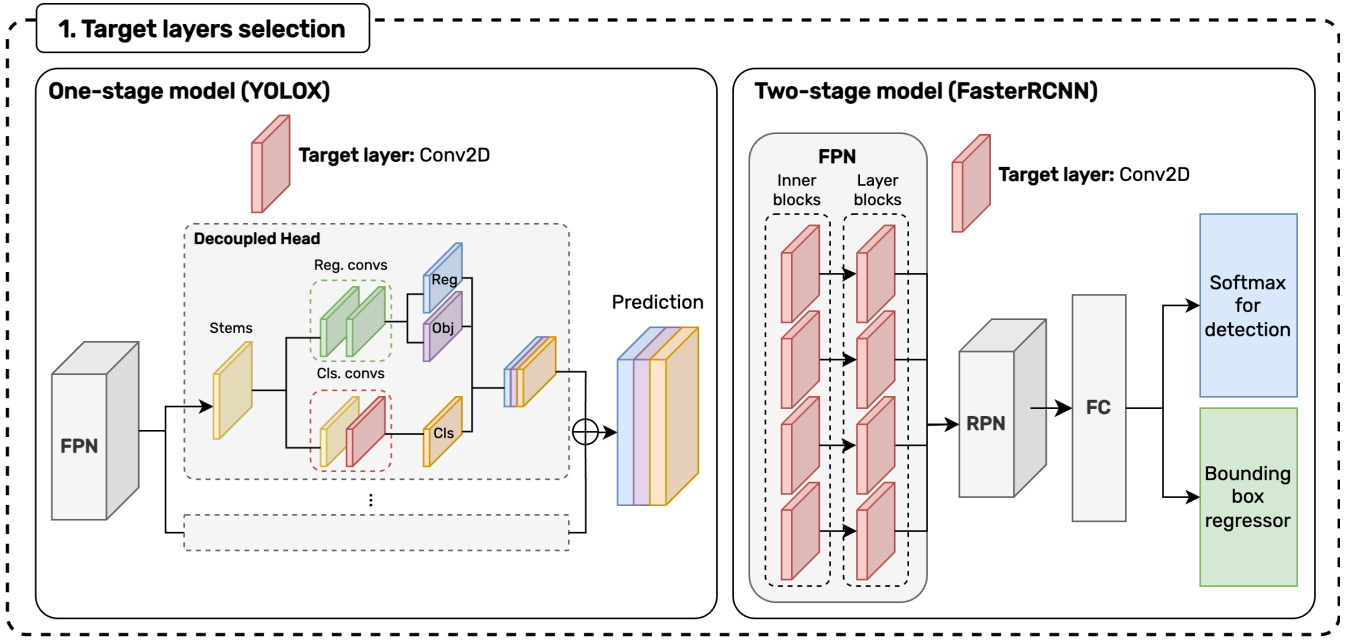
1. **Target Layers Selection:** Set the target layers from one-stage/two-stage object detection models
2. **Object Locating with Gradient:** Take the derivative of the target box with the final feature map to get the location map
3. **Weighting Feature Map:** Assign importance to each feature map based on its contribution to the target object's prediction
4. **Masking Target Regions:** Focus the saliency map on the target object and reduce noise from unrelated region



Block 1 – Target Layers Selection

To extract the spatial information from the model's convolution layers to generate the saliency map:

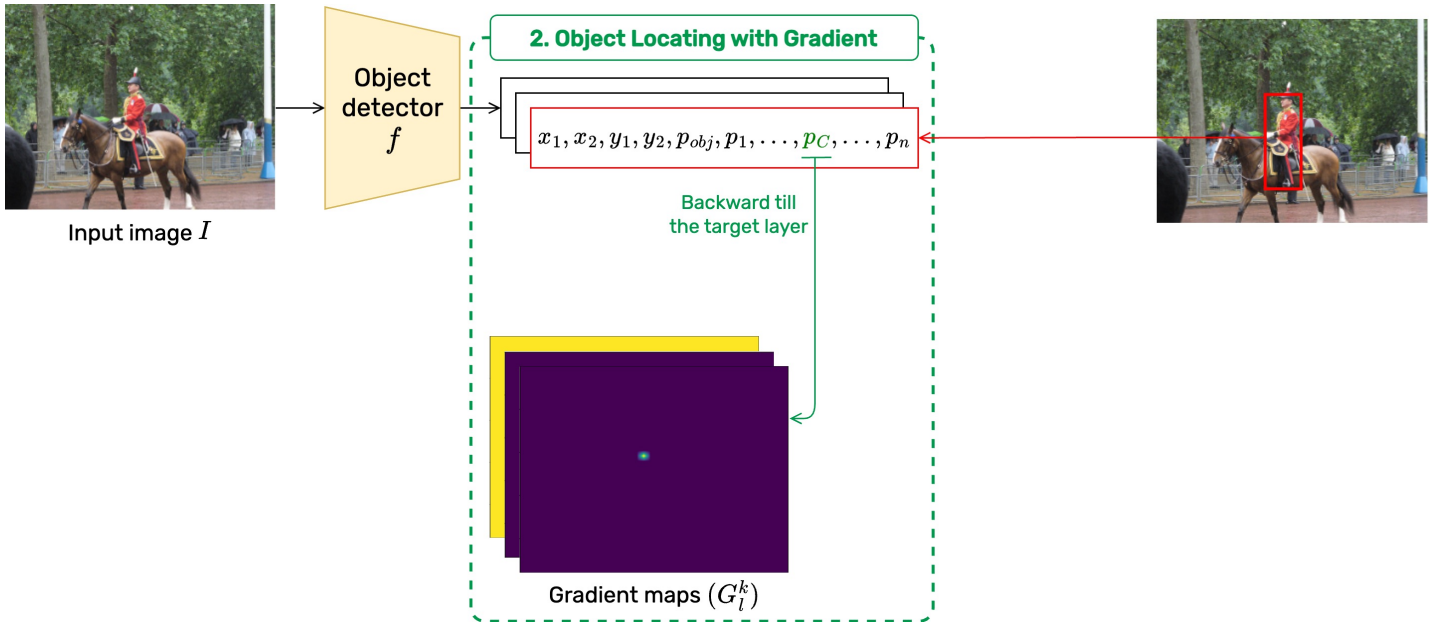
- **For one-stage detectors (e.g., YOLOX):** Choose the final convolution layer in each branch as the target layer
- **For two-stage detectors (e.g., Faster-RCNN):** Utilize the convolution layers in the Feature Pyramid Network (FPN) as the target layers



Block 2 – Object Localization with Gradient

To identify the location of the target object in the feature map:

- **For one-stage detectors:** The pixel in the gradient map represents the center of the object
- **For two-stage detectors** (regression and classification tasks are in separate branches)**:** The pixel with the highest value in the gradient map is used as an estimate of the object's center

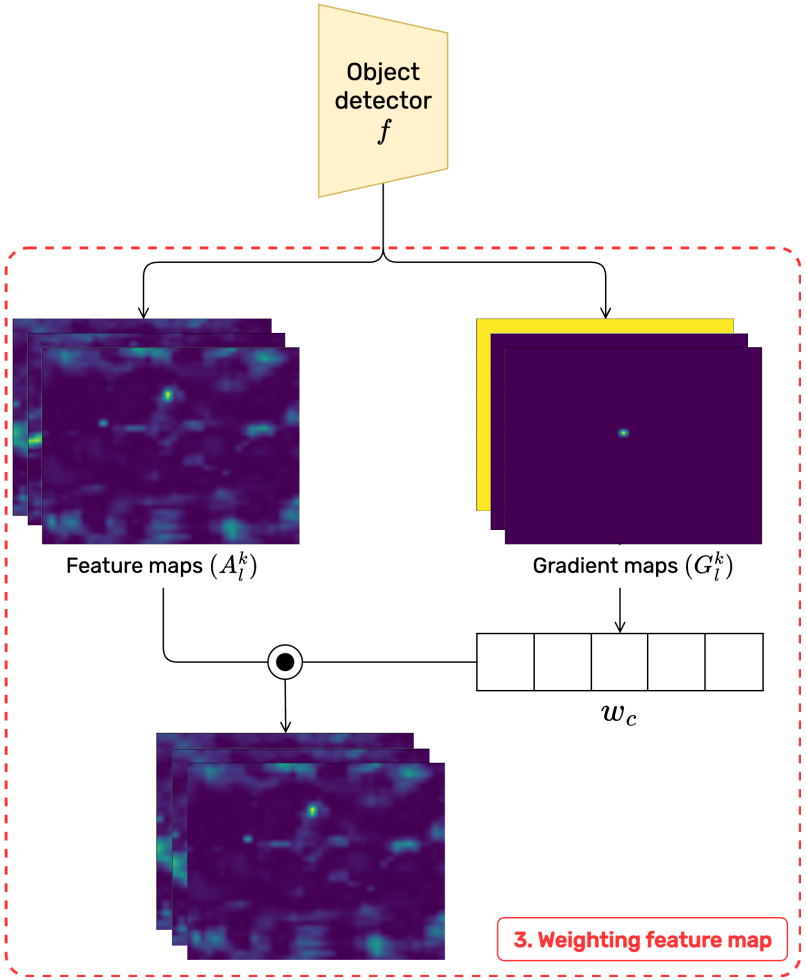


Block 3 – Weighting Feature Map

Assigning importance to each feature map based on its contribution to the target object's prediction:

- Dividing feature maps into positive and negative parts to create a smoother saliency map by considering the different effects of each part on the prediction
- Calculating the weight for each feature map using the mean value of the gradient map provides a measure of its importance

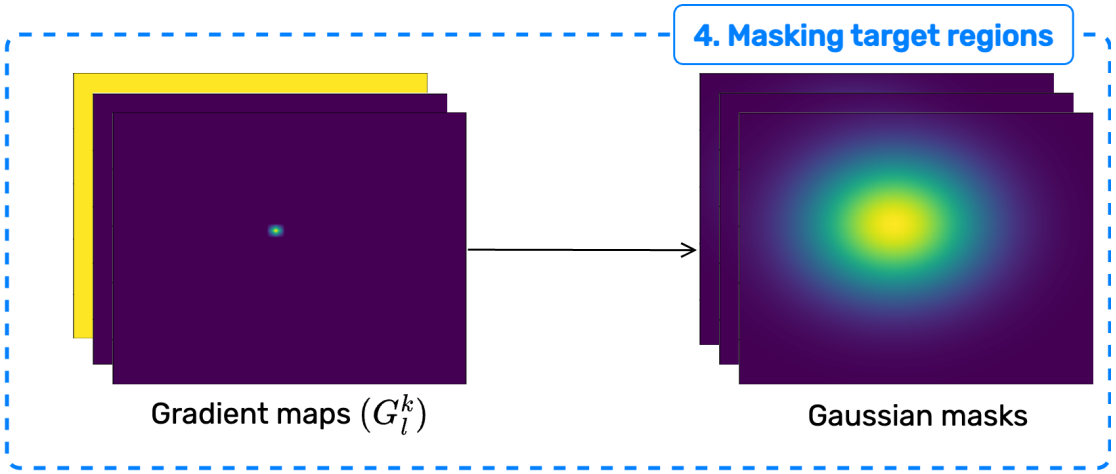
$$A_{k_2}^{c(-)} = \alpha_{k_2}^{c(-)} A_{k_2}^{c}$$
$$A_{k_1}^{c(+)} = \alpha_{k_1}^{c(+)} A_{k_1}^{c}$$
$$L_{\text{CAM}}^{c} = ReLU\left(\sum_{k_1} A_{k_1}^{c(+)} - \sum_{k_2} A_{k_2}^{c(-)} \right)$$



Block 4 – Masking Target Regions

To focus the saliency map on the target object and reduce noise from unrelated regions:

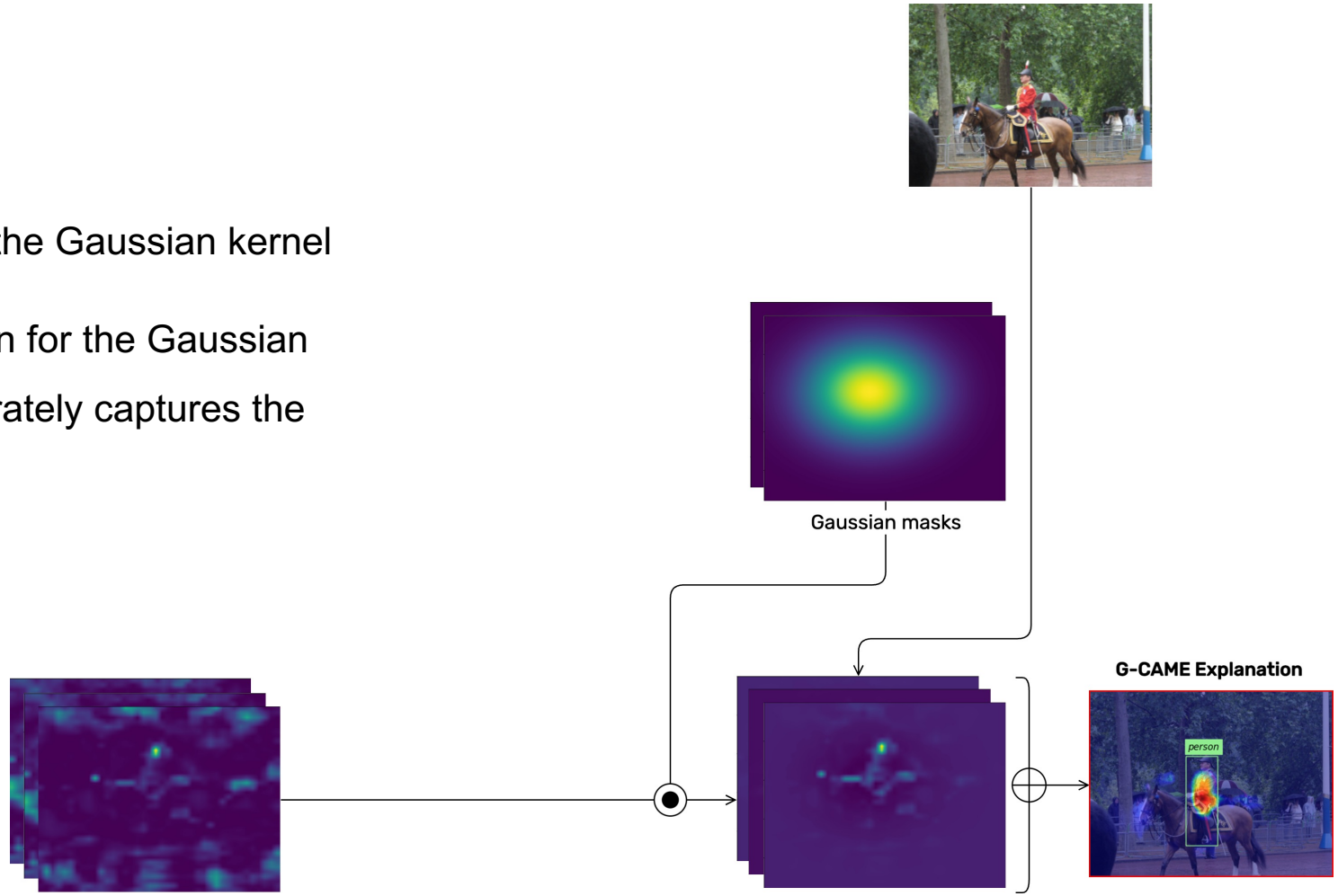
- Using a Gaussian mask as a weight for each pixel in the weighted feature map to estimate the object region based on the center of the object



Explanation Map Generation

To generate the final explanation map:

- Combine the weighted feature maps using the Gaussian kernel
- Choosing the appropriate standard deviation for the Gaussian mask to ensure that the saliency map accurately captures the object's size and location

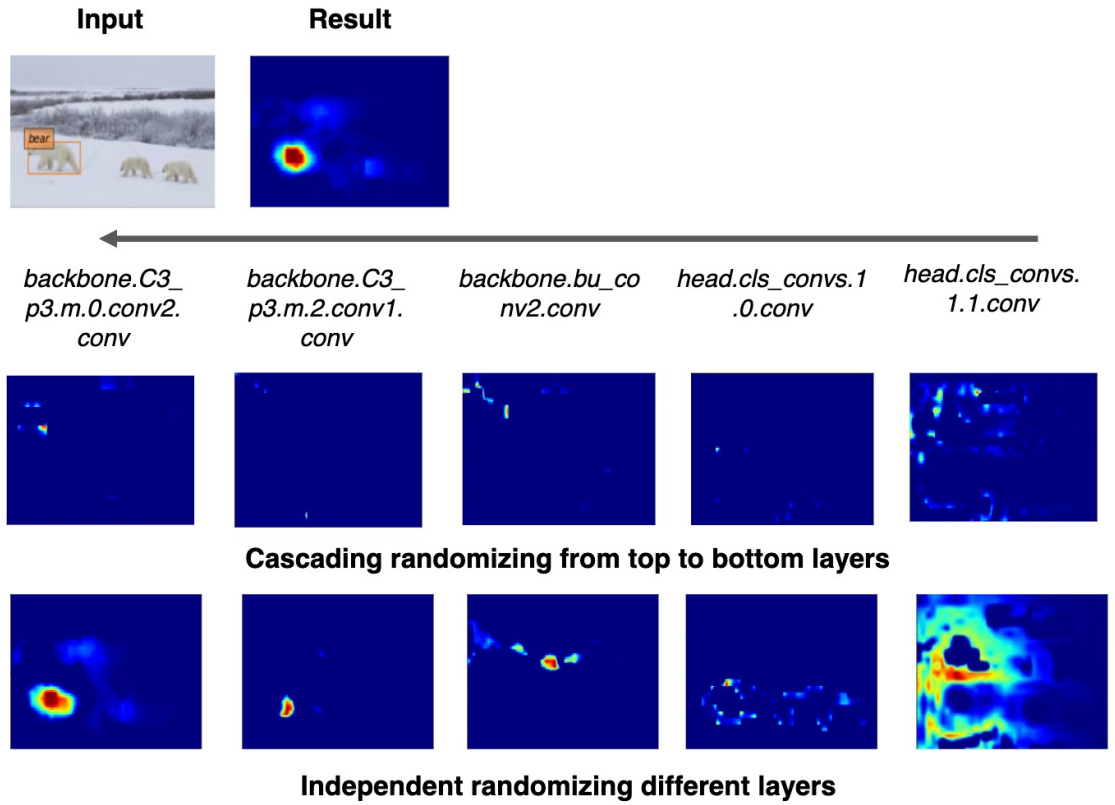


Experiments and Results

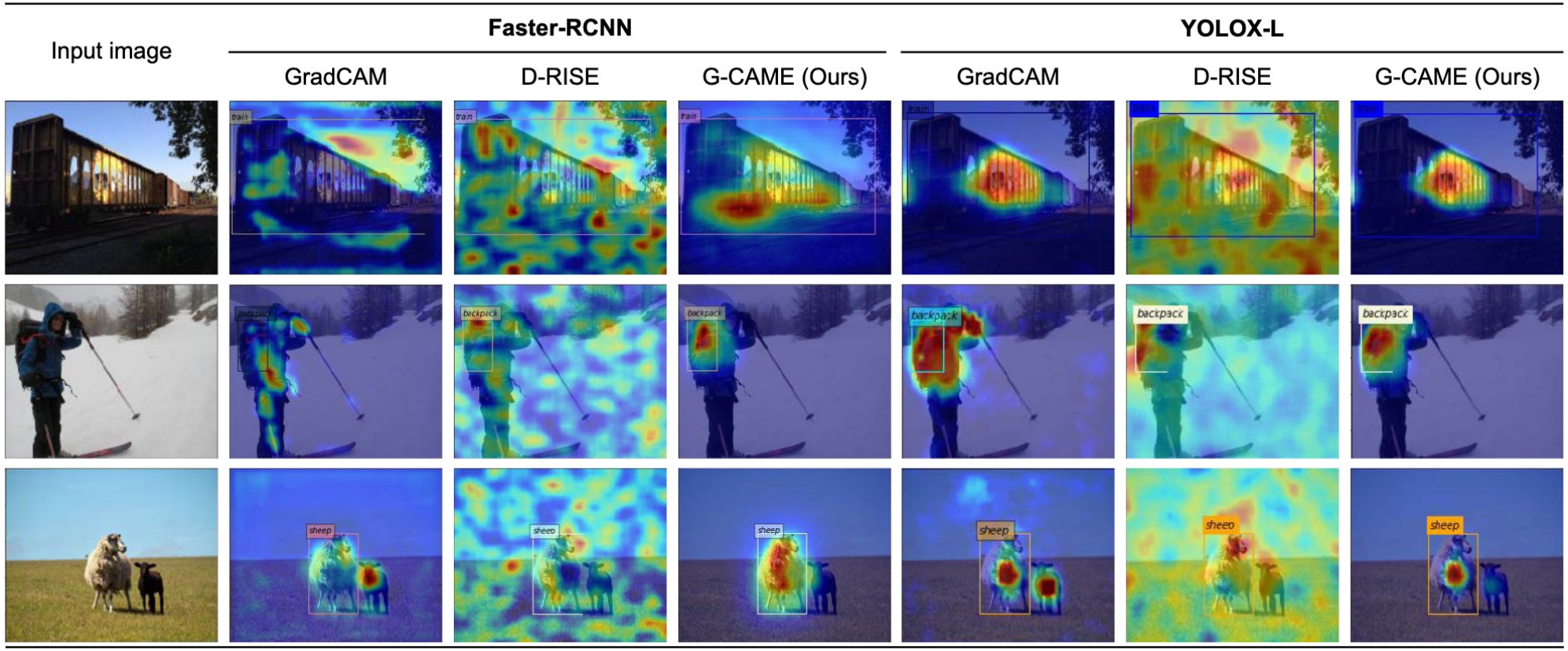
Sanity check

To validate whether the saliency map is a faithful explanation of the model's prediction

1. Perform Cascading Randomization and Independent Randomization tests
2. Results: G-CAME method produces valid and reliable explanations that are sensitive to the model's parameters



Qualitative Evaluation



Visualization results of GradCAM, D-RISE, and G-CAME on samples of MS-COCO 2017 dataset.
G-CAME can generate the least noisy saliency maps for explaining a specific object.

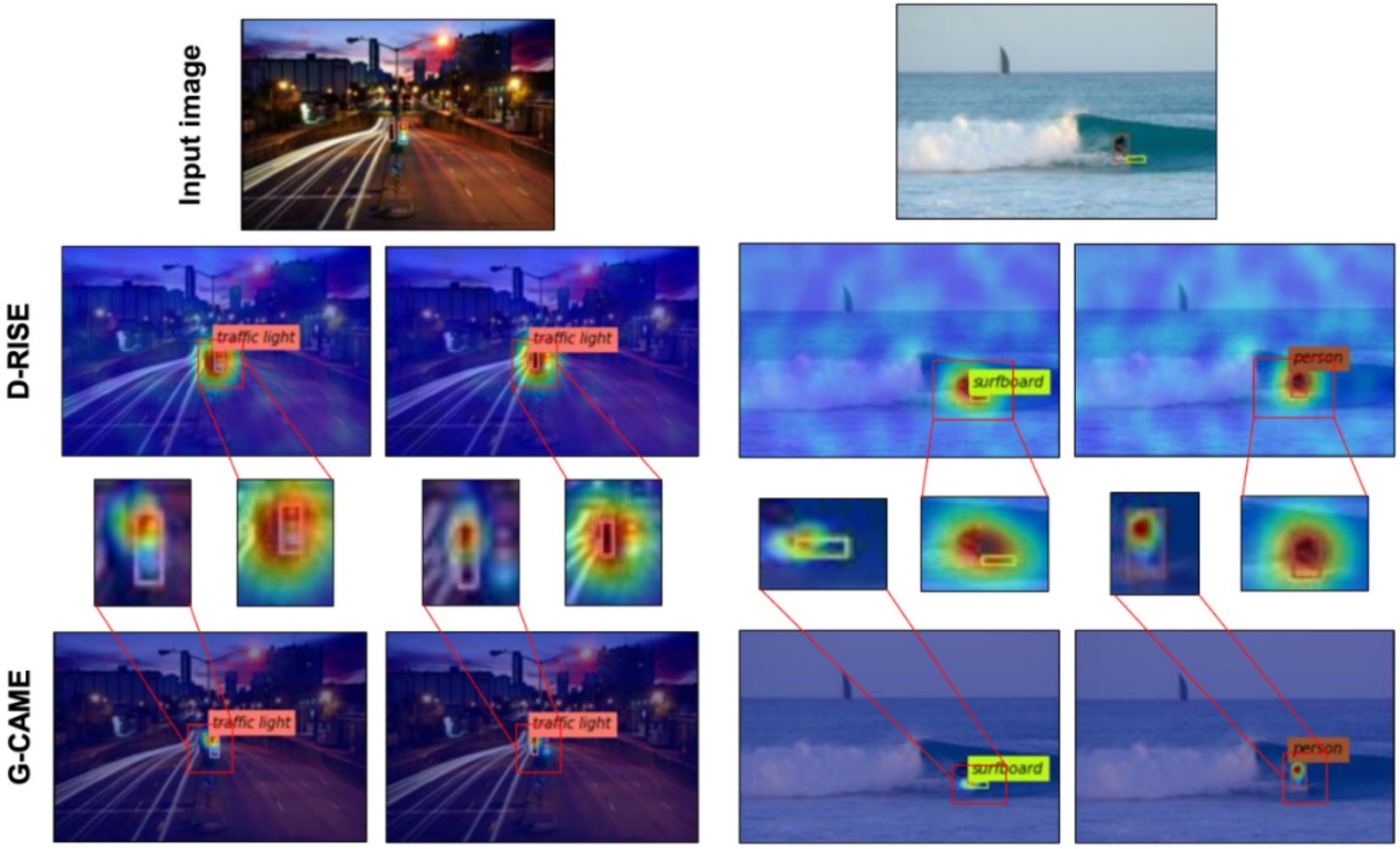
Qualitative Evaluation on Tiny Objects

The saliency map of D-RISE and G-CAME for tiny objects prediction:

(a) tiny objects of the same class lying close together

(b) multiple tiny objects of different classes lying close together.

In both cases, G-CAME can clearly distinguish each object in its explanations.



(a) Explanation for tiny objects of different class

(b) Explanation for tiny objects of the same class

Quantitative Evaluation

Plausibility evaluation: measure the correlation between the saliency map and the human-labeled ground truth

- Pointing Game (PG) and Energy-based Pointing Game (EBPG)

Faithfulness evaluation: assess the completeness and consistency of the explanations for the model's predictions

- Confidence Drop and Information Drop

Method	D-RISE	G-CAME (Our)
PG%↑ (Overall | Tiny object)	0.86 | 0.127	**0.98** | **0.158**
EBPG%↑ (Overall | Tiny object)	0.184 | 0.009	**0.671** | **0.261**

Method	D-RISE	G-CAME (Our)
Confidence Drop%↑	**42.3**	36.8
Information Drop%↓	31.58	**29.15**
Running time(s)↓	252	**0.435**

Comparison of D-RISE and G-CAME (Our) on the MS-COCO 2017 validation dataset with the YOLOX model.

Conclusion

Conclusion

Perturbation-based

D-RISE, D-CLOSE

CAM-based

G-CAME

Model-agnostic nature

Higher faithfulness explanation

Higher plausibility explanation

Less noisy explanation

Higher information retainment

Better running time

References

Truong, V. B., Nguyen, T. T. H., Nguyen, V. T. K., Nguyen, Q. K., & Cao, Q. H. (2024, February). Towards Better Explanations for Object Detection. In Asian Conference on Machine Learning (pp. 1385-1400). PMLR.

Kapishnikov, A., Bolukbasi, T., Viégas, F., & Terry, M. (2019). Xrai: Better attributions through regions. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4948-4957).

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., & Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11443-11452).

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

Dr Hung Cao
Assistant Professor, Lab Director
Analytics Everywhere Lab
University Of New Brunswick, Canada
hcao3@unb.ca

Dr Francis Palma
Assistant Professor
Faculty of Computer Science
University Of New Brunswick, Canada

Dr Monica Wachowicz
Adjunct Professor
Associate Dean Geospatial Science
RMIT University, Australia

Dr Trevor Hanson
Professor
Faculty of Civil Engineering
University Of New Brunswick, Canada

Hung Nguyen
PhD Student

Asfia Kawnine
MSc Student

Atah Nuh Mih
MSc Student

Alireza Rahimi
MSc Student

Simran Dadhich
MSc Student

...

Khanh Nguyen
AI Engineer

Khang Nguyen
AI Scientist

Binh Truong
AI Engineer

Tuong Phan
Undergrad Student