ODAM: Gradient-based Instance-specific Visual Explanation for Object Detection

Zhao Chenyang

Antoni Bert CHAN

- Motivation
- Method & Visualizations
- Results

Visual explanation

The explanation approaches produce heat maps locating the regions in the input images that the model looked at, and representing the influence of different pixels on the model's decision.

- Classification

Target: Dog

Grad-CAM

Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

Visual explanation

The explanation approaches produce heat maps locating the regions in the input images that the model looked at, and representing the influence of different pixels on the model's decision.

- On Classification

Target: Dog

Grad-CAM

- On Object detection

Target: the object in the white box

Grad-CAM

Visual explanation

The explanation approaches produce heat maps locating the regions in the input images that the model looked at, and representing the influence of different pixels on the model’s decision.

- On Classification

Target: Dog

Grad-CAM

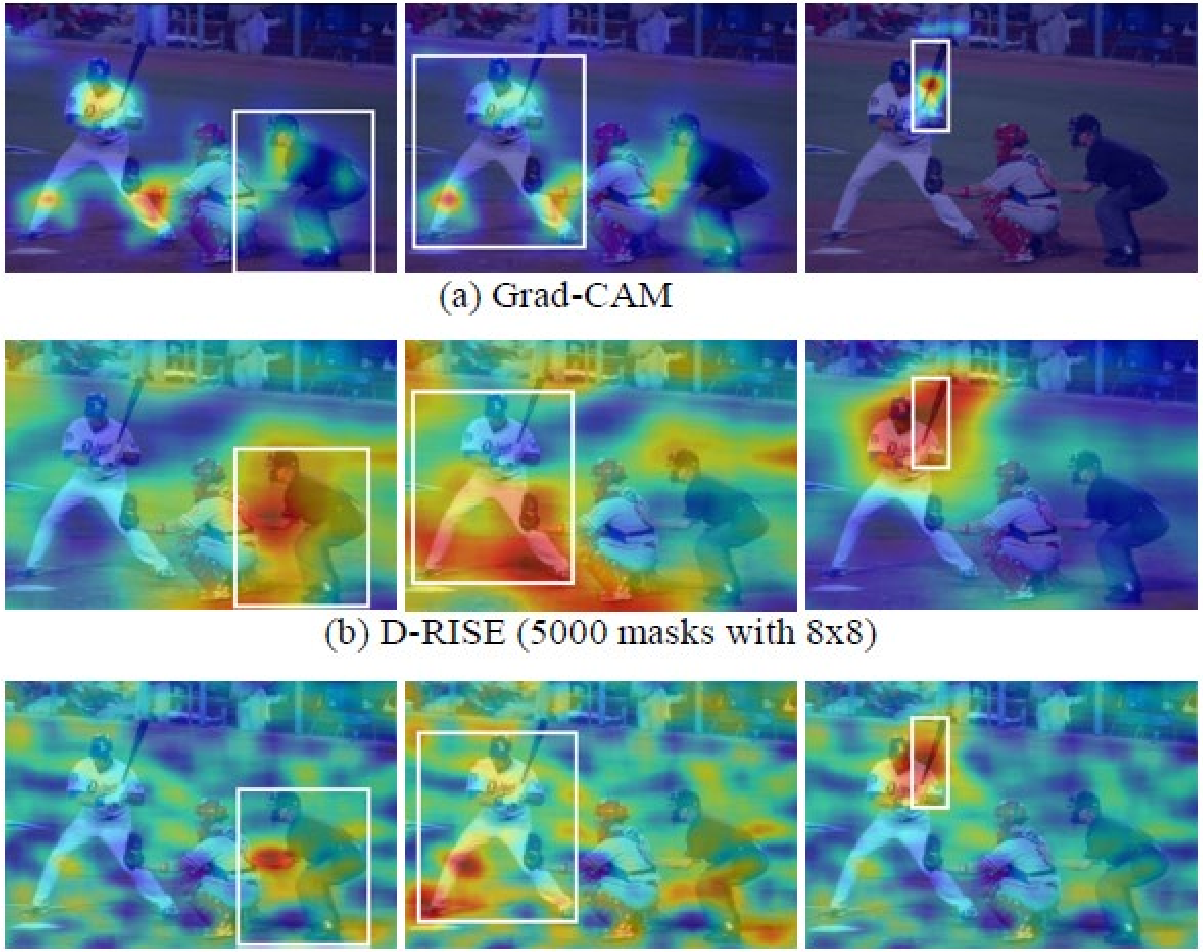
- On Object detection

Target: the object in the white box

Grad-CAM

The Classification-specific explanation highlights all objects of the same category (person) instead of the specific object instance.

Visual explanation



(a) Grad-CAM

(b) D-RISE (5000 masks with 8x8)

(c) D-RISE (5000 masks with 16x16)

D-RISE:

- Designed for object detection
- Visual explanation for the specific prediction

Drawbacks:

- Noisy background
- Influenced by the mask resolution
- Time consuming
- Disable to explain separate detection attribute (e.g., classification score and bounding box coordinates)

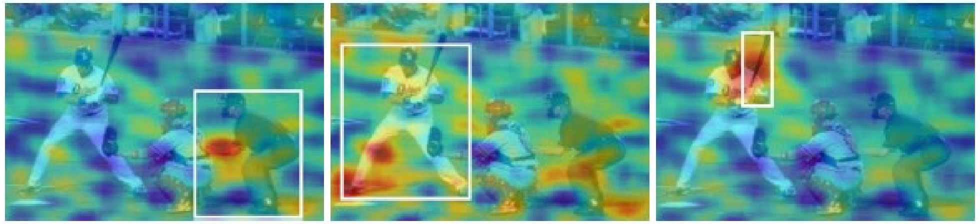
Petsiuk V, Jain R, Manjunatha V, et al. Black-box explanation of object detectors via saliency maps[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11443-11452.

Visual explanation

- Object specification: what features are important for making the predictions?
- Object discrimination: which object was detected?

(a) Grad-CAM

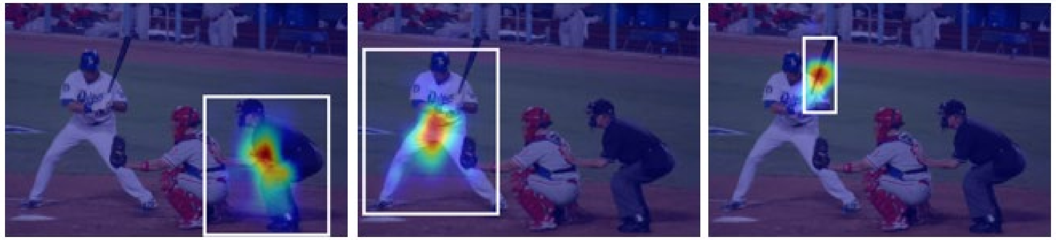
(b) D-RISE (5000 masks with 8x8)



(c) D-RISE (5000 masks with 16x16)

ODAM

Object specification

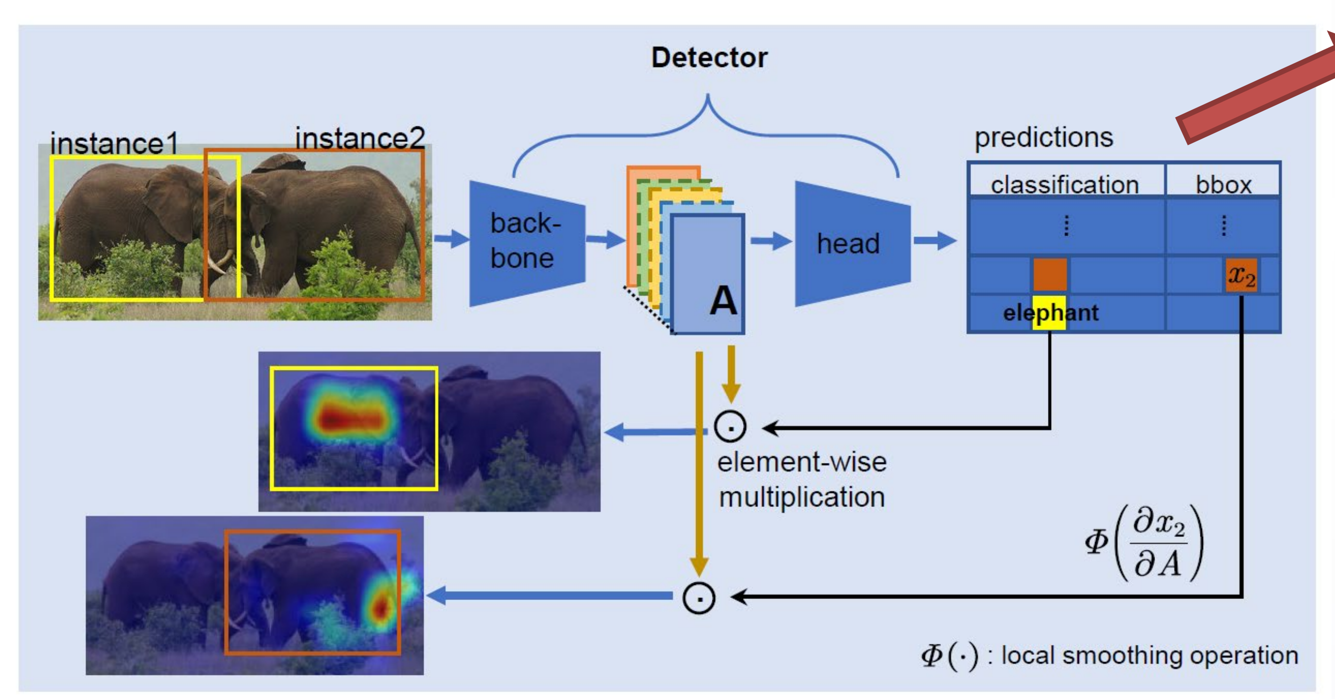


ODAM w/ Odam-Train

Object discrimination

- Motivation
- Method & Visualizations
- Results

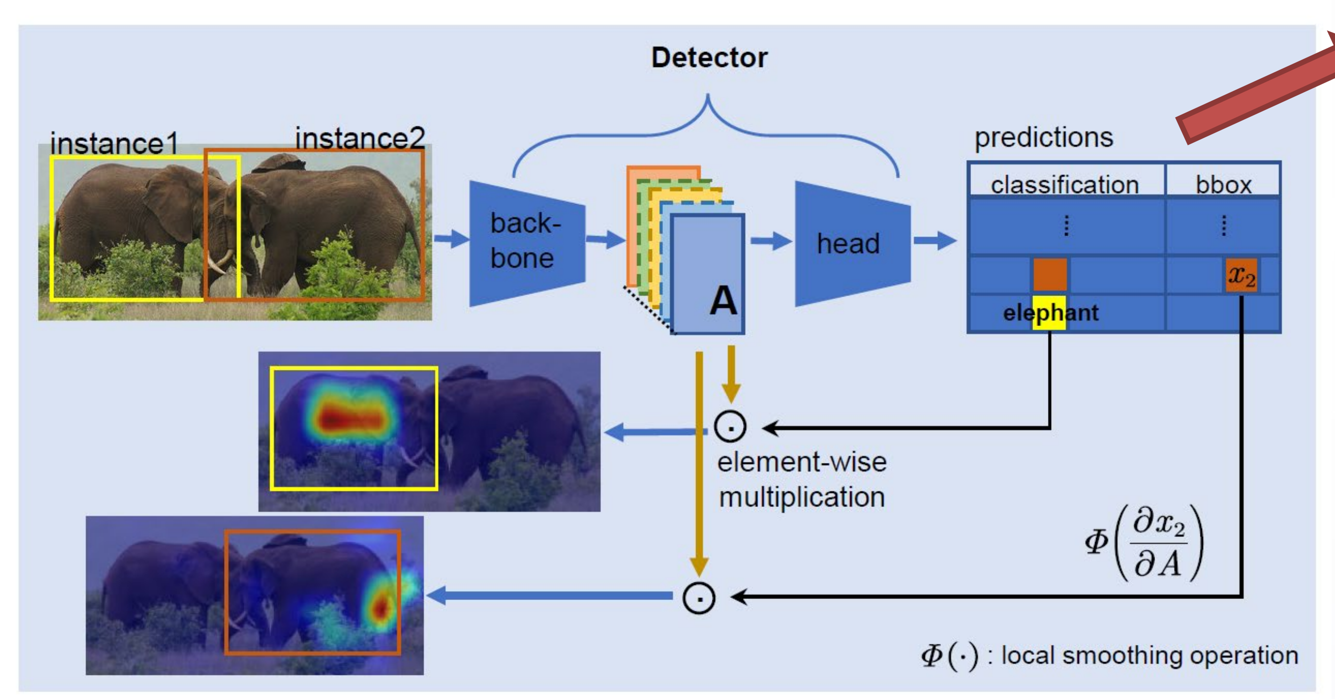
ODAM: gradient-based object detector activation maps



Prediction in Object Detection :

- Classification score: $s_c^{(p)}$
- Bounding box: $B^{(p)} = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$

ODAM: gradient-based object detector activation maps



Prediction in Object Detection :

- Classification score: $s_c^{(p)}$
- Bounding box: $B^{(p)} = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$

ODAM for predicted object attribute in object detection:

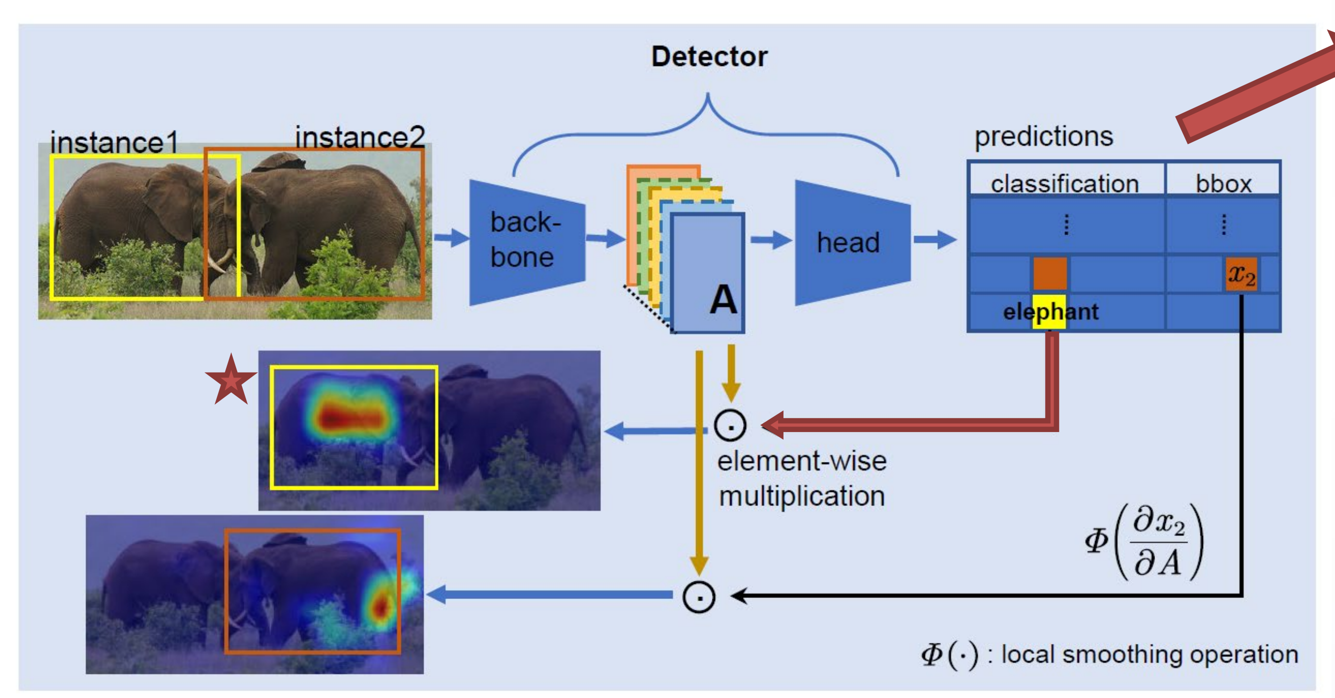
- Assume any predicted object attribute scalar $Y^{(p)}$ can be written as a linear element-wise weighted combination of feature map:

$$Y^{(p)} = \sum_k \sum_{ij} w_{ijk}^{(p)} A_{ijk} \;, \qquad H_{ij}^{(p)} = \sum_k w_{ijk}^{(p)} A_{ijk}$$

- Set the importance weight map as:

$$w_k^{(p)} = \Phi\left(\frac{\partial Y^{(p)}}{\partial A^k}\right), \qquad H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} A^k\right)$$

ODAM: gradient-based object detector activation maps



Prediction in Object Detection :

- Classification score: $s_c^{(p)}$
- Bounding box: $B^{(p)} = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$

ODAM for predicted object attribute in object detection:

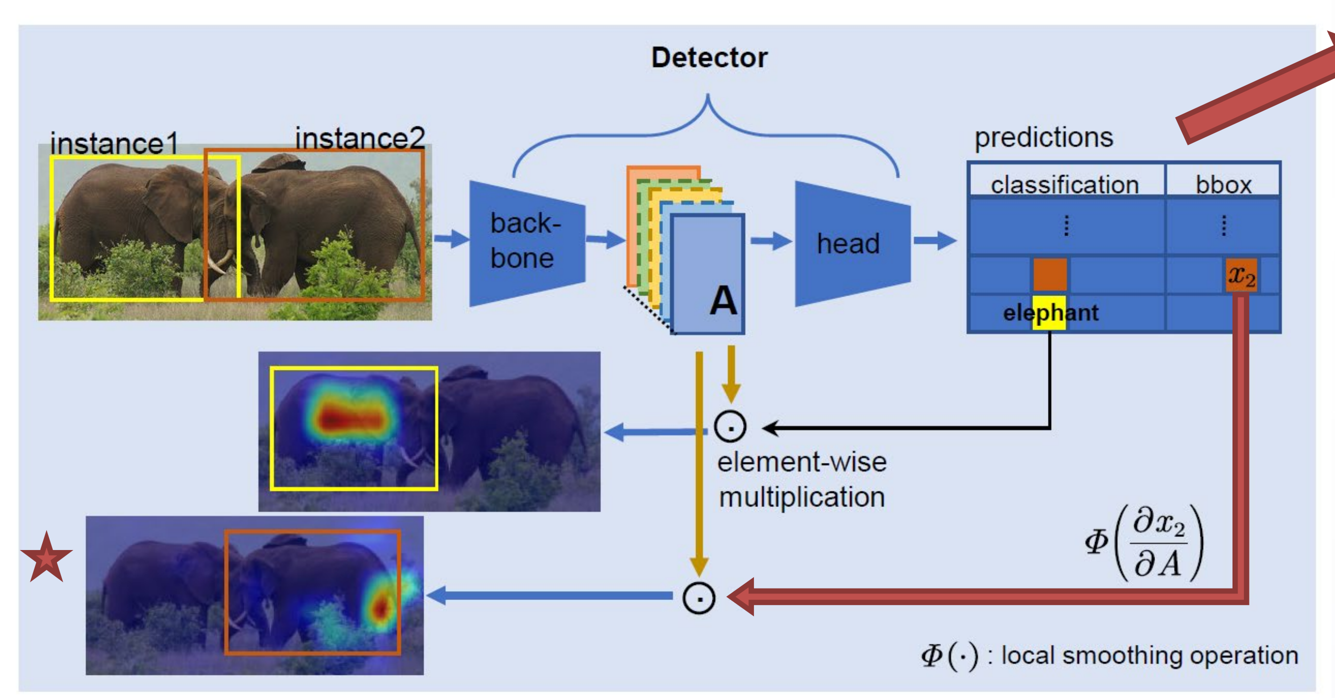
- Assume any predicted object attribute scalar $Y^{(p)}$ can be written as a linear element-wise weighted combination of feature map:

$$Y^{(p)} = \sum_k \sum_{ij} w_{ijk}^{(p)} A_{ijk} \ , \qquad H_{ij}^{(p)} = \sum_k w_{ijk}^{(p)} A_{ijk}$$

- Set the importance weight map as:

$$w_k^{(p)} = \Phi\left(\frac{\partial Y^{(p)}}{\partial A^k}\right), \qquad H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} A^k\right)$$

ODAM: gradient-based object detector activation maps



Prediction in Object Detection :

- Classification score: $s_c^{(p)}$
- Bounding box: $B^{(p)} = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$

ODAM for predicted object attribute in object detection:

- Assume any predicted object attribute scalar $Y^{(p)}$ can be written as a linear element-wise weighted combination of feature map:

$$Y^{(p)} = \sum_k \sum_{ij} w_{ijk}^{(p)} A_{ijk} \ , \qquad H_{ij}^{(p)} = \sum_k w_{ijk}^{(p)} A_{ijk}$$

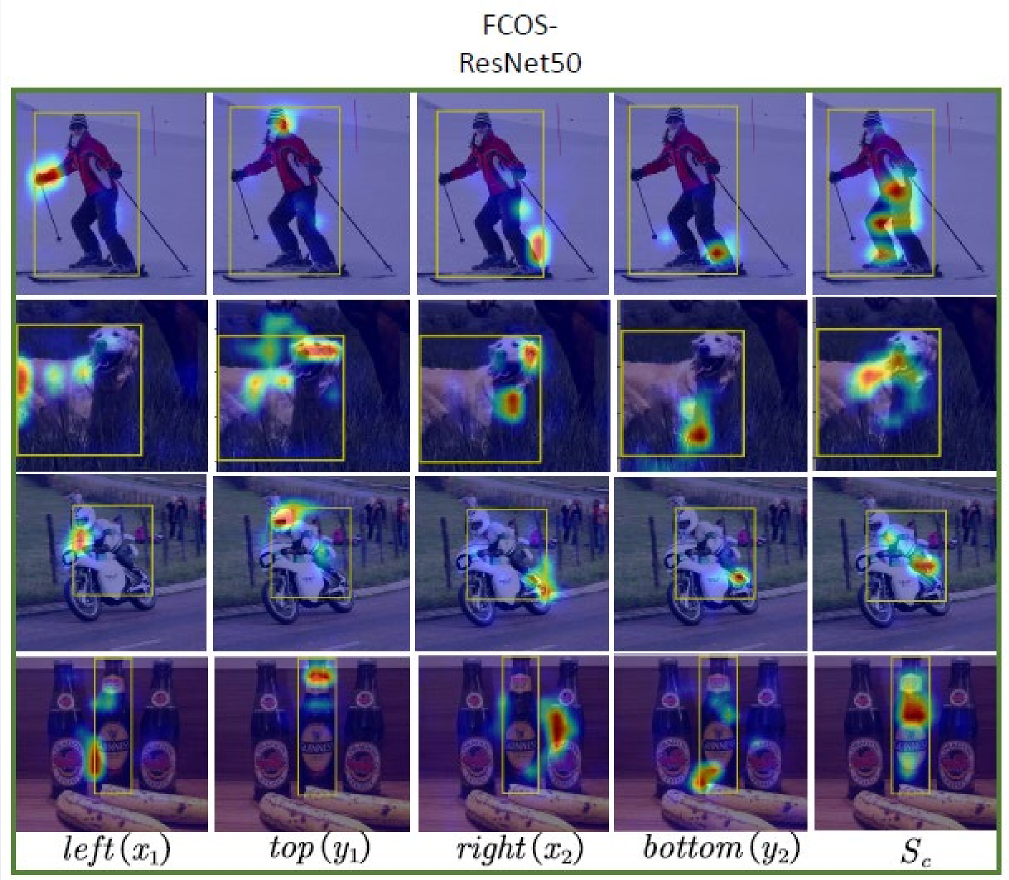
- Set the importance weight map as:

$$w_k^{(p)} = \Phi\left(\frac{\partial Y^{(p)}}{\partial A^k}\right) , \qquad H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} A^k\right)$$

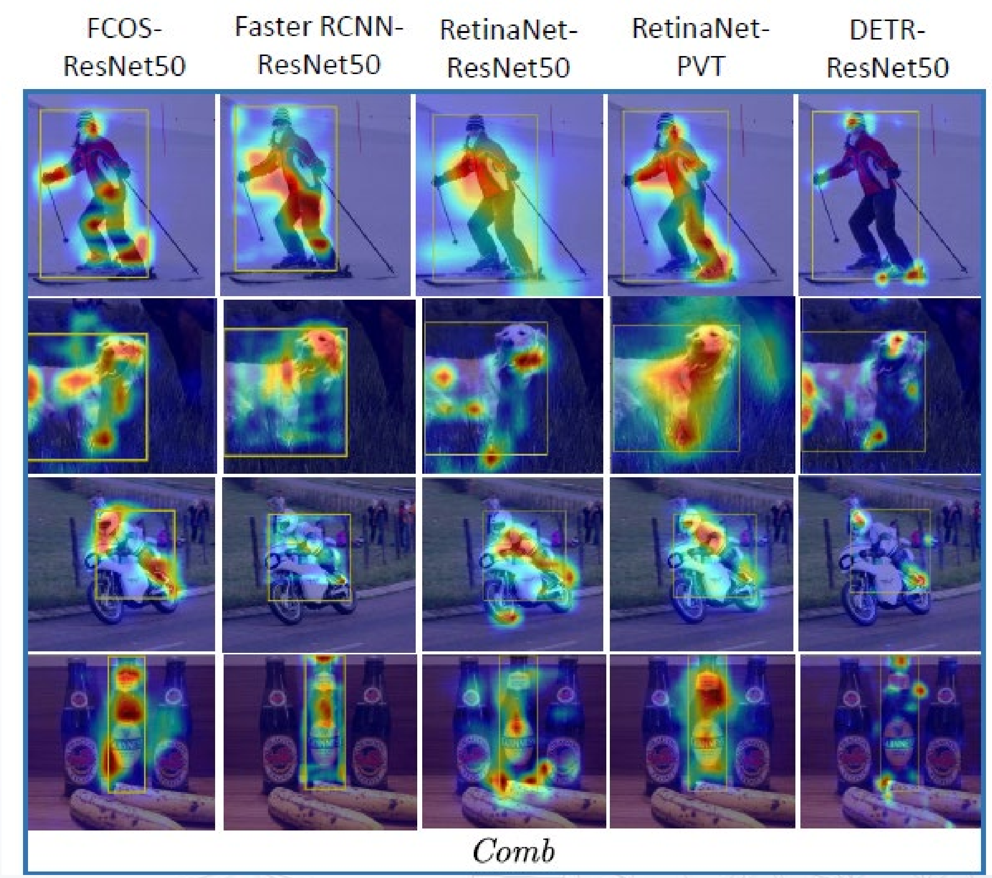
Visualizations of ODAM heat maps on object specification

Object specification: what features are important for making the predictions?

- **The heat maps explain important regions for each predicted attribute (class score and bbox coordinates) from FCOS.**



- **Heat map explanations of instances computed from different detectors.**

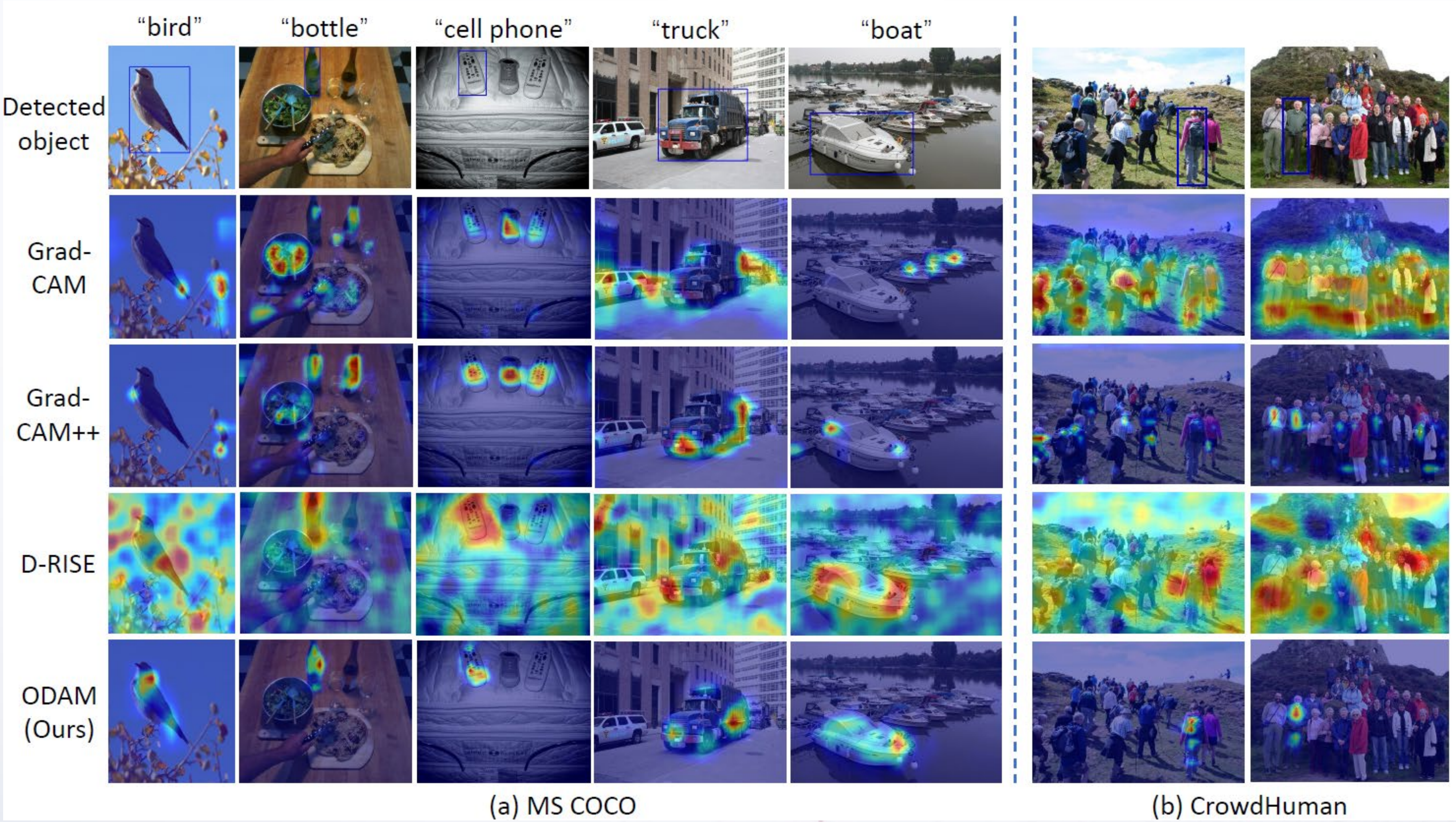


$$H_{comb} = \max\left(H_{x_1}, H_{y_1}, H_{x_2}, H_{y_2}, H_{s_c}\right)$$

Visualizations of ODAM heat maps on object specification

Object specification: what features are important for making the predictions?

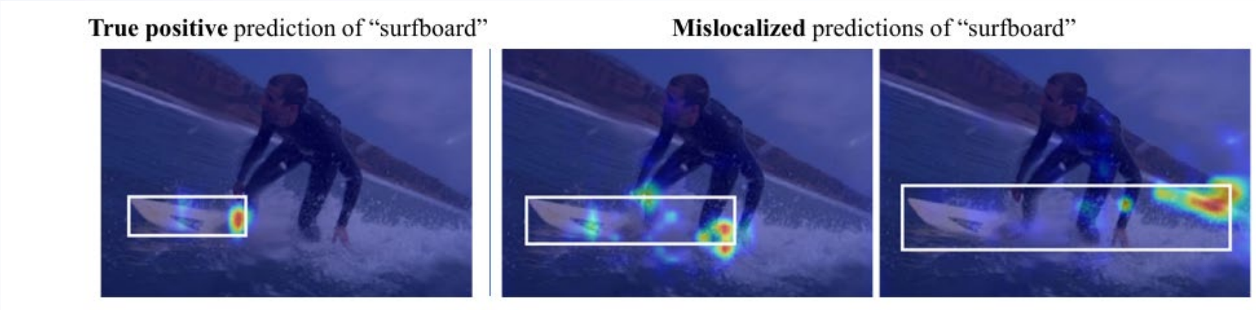
- **Comparison of heat maps from different explanation methods**



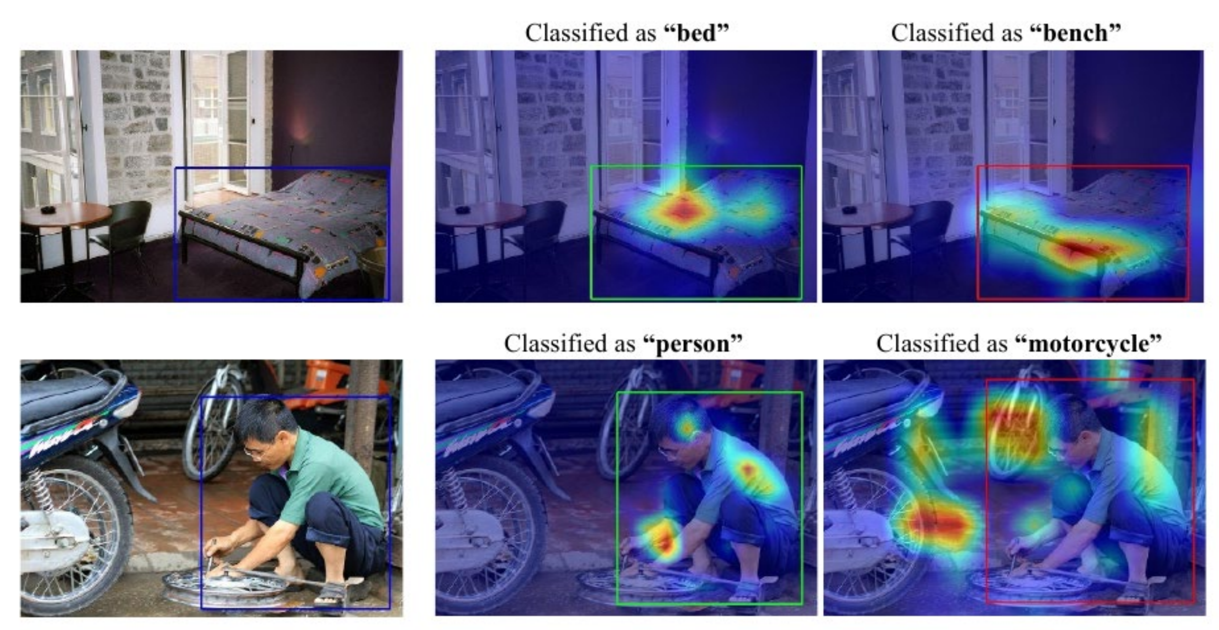
Visualizations of ODAM heat maps on object specification

Object specification: what features are important for making the predictions?

- **Error Mode Analysis**



Explanations of the predictions of the right extent for different predictions of ``surfboard''. The heat maps for the mislocalized predictions highlight the visual features that induced to the wrong extents (the leg on the right, and the sea horizon).



Explanations of the class scores of different predictions. In the first row, the model predicts ``bench'' when it puts attention on only the frame at the end of the bed. In the second row, the model is negative influenced by the context feature and misclassifies a ``motorcycle'' on a ``person''.

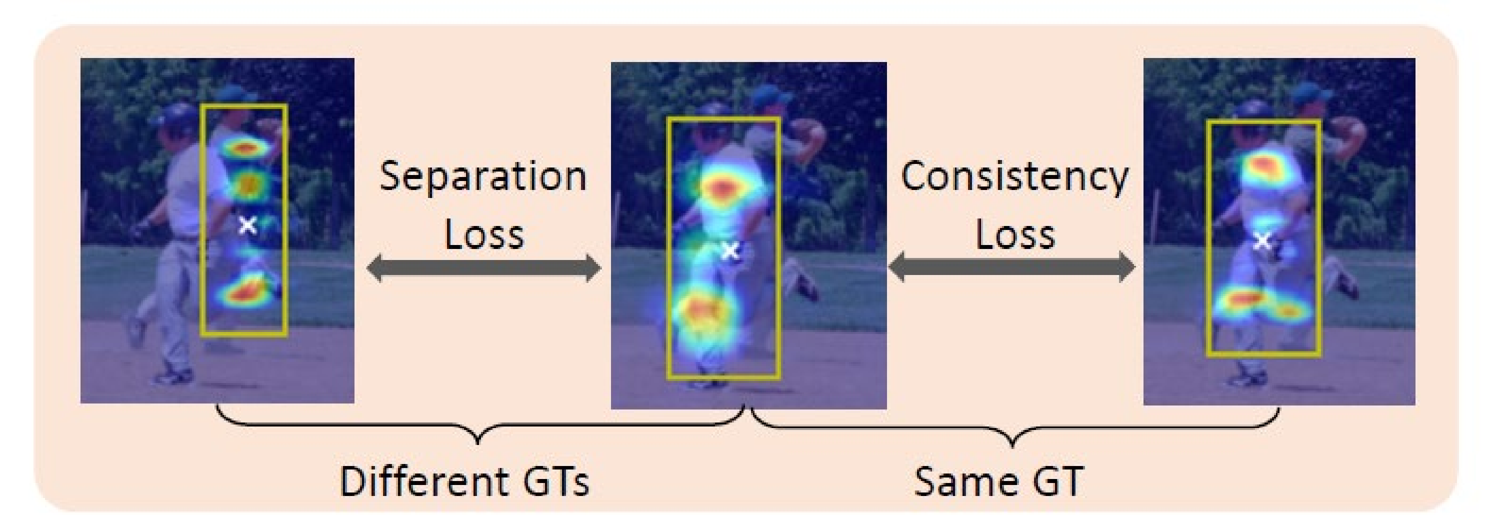
Object Discrimination

Object discrimination: Which object was being detected?

- **Odam-Train:** a training method for improving the heat maps for object discrimination, to better explain which object was being detected.

$$L_{con} = \sum_{p \in GT} \sum_{n \in \mathcal{P}^{(p)}} -\log \cos\left(H_{best}^{(p)},\ H_n^{(p)}\right),$$

$$L_{sep} = \sum_{p \in GT} \sum_{m \notin \mathcal{P}^{(p)}} -\log \left(1 - \cos\left(H_{best}^{(p)},\ H_m^{(-p)}\right)\right)$$

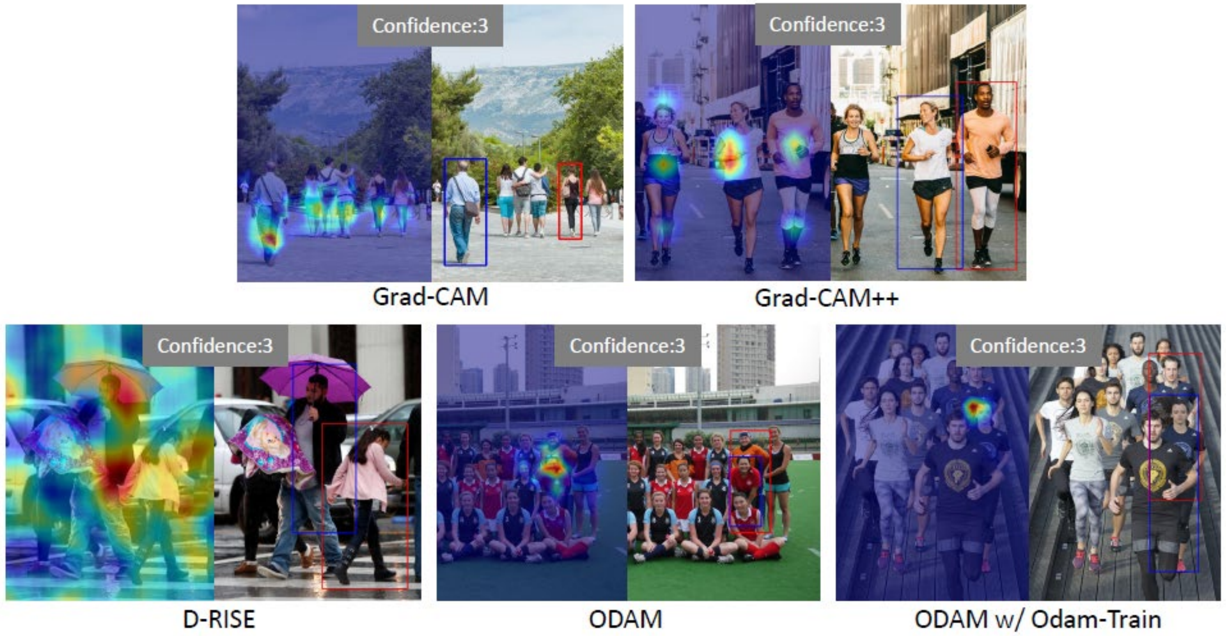


User Test on Object Discrimination

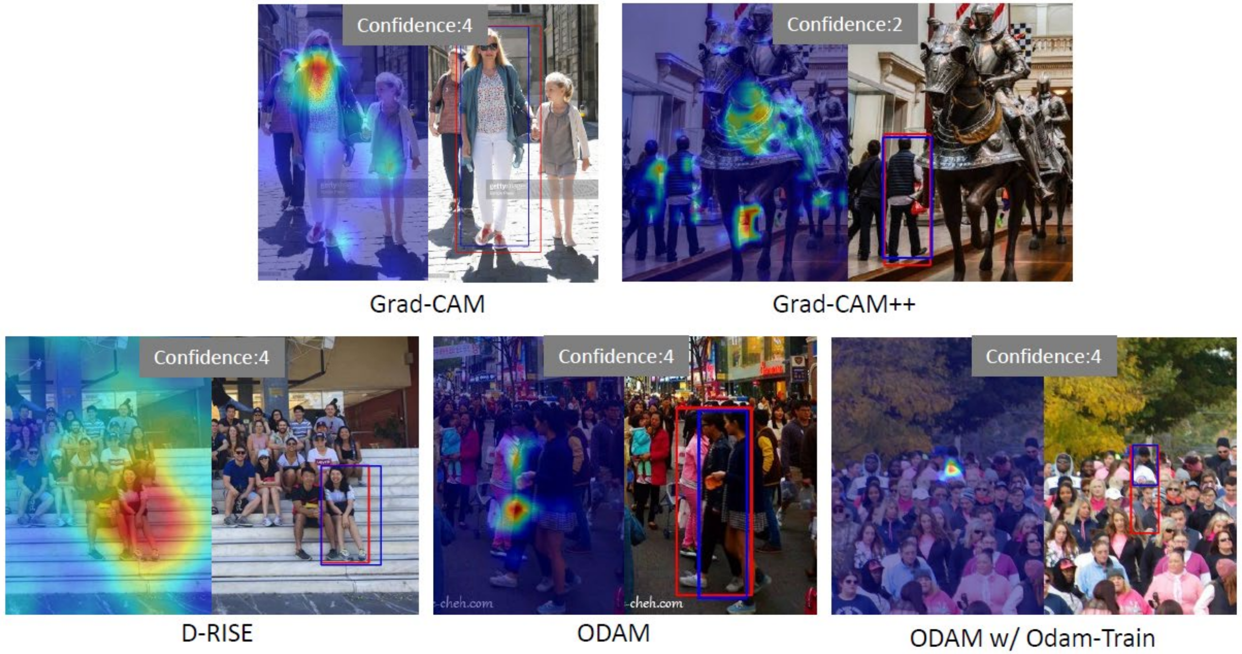
A user test about object discrimination:

- users are asked to draw the bounding box of the object which was detected based on the given heat map.
- Blue boxes are those drawn by users, while red boxes are those of the ground truth objects.

Confidence	Grad-CAM	Grad-CAM++	D-RISE	ODAM	ODAM w/ Odam-Train
1 (least)	53.38	63.76	20.67	0	0
2	30.41	24.16	36.67	0.67	1.35
3	10.14	6.71	26.01	7.33	3.33
4	5.41	4.70	11.98	21.34	11.33
5 (most)	0.68	0.67	4.68	70.68	83.99
avg. conf.	1.70	1.54	2.43	4.62	4.78
accuracy	14.19	18.79	60.67	94.00	94.67



(a) Examples of user's incorrect choice



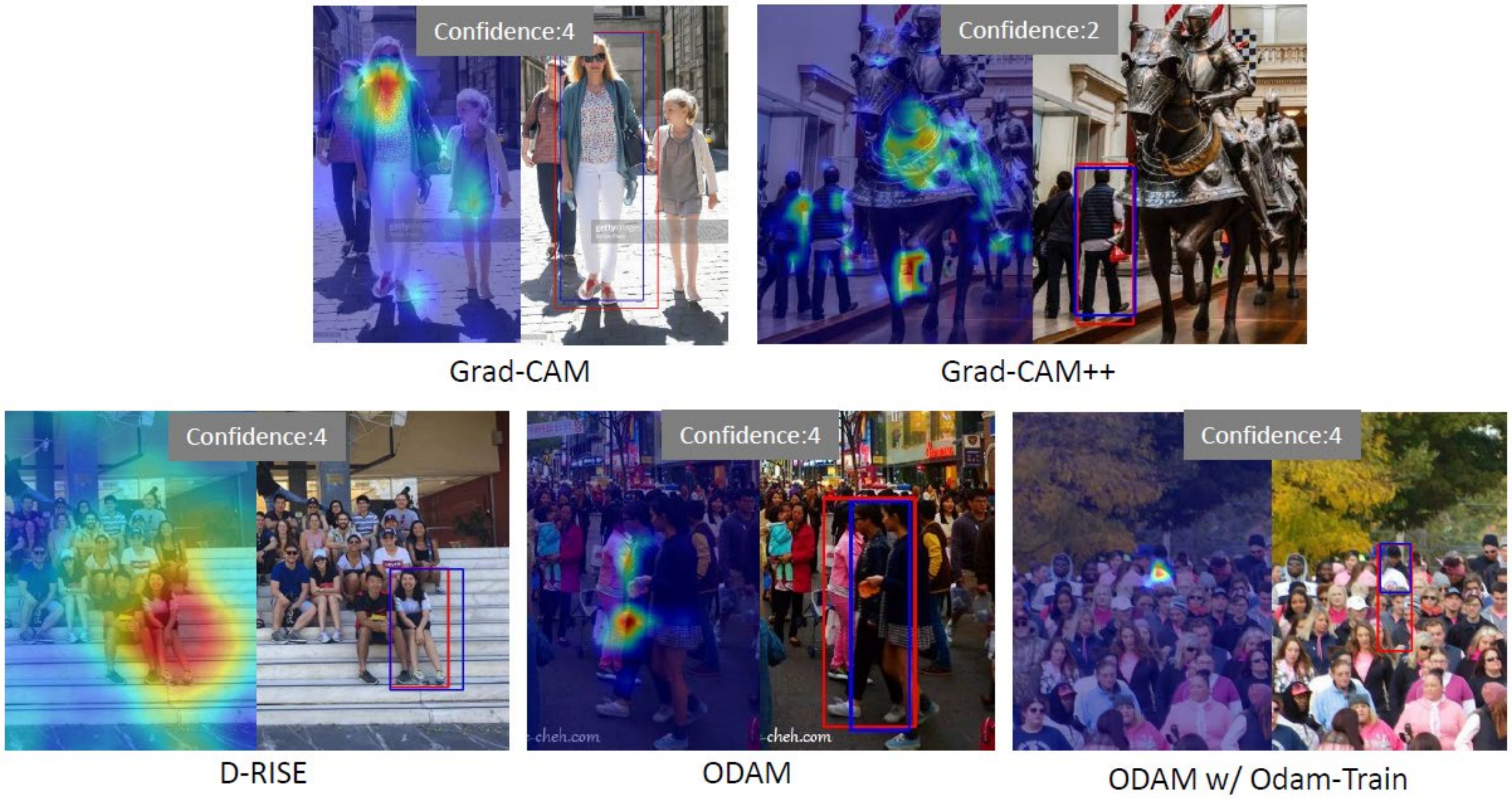
(b) Examples of user's correct choice

User Test on Object Discrimination

A user test about object discrimination:

- users are asked to draw the bounding box of the object which was detected based on the given heat map.
- Blue boxes are those drawn by users, while red boxes are those of the ground truth objects.

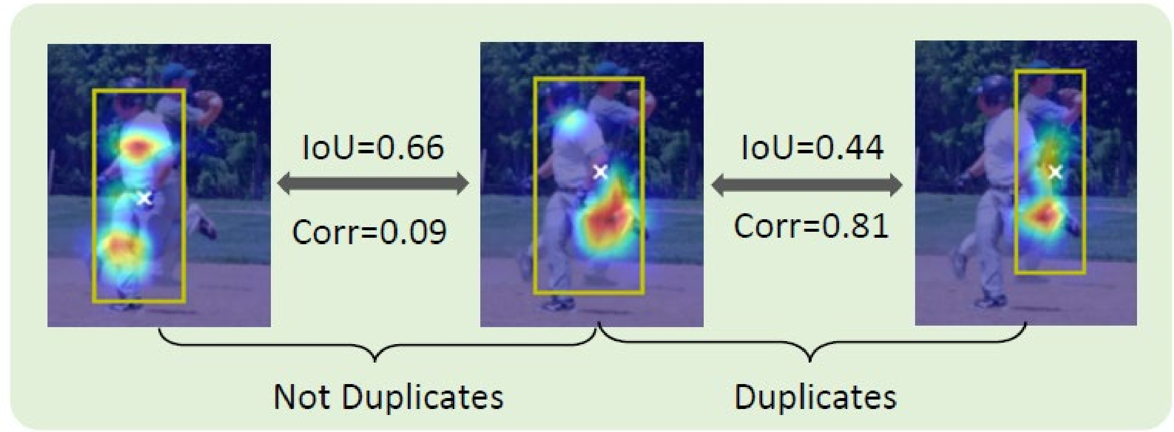
Confidence	Grad-CAM	Grad-CAM++	D-RISE	ODAM	ODAM w/ Odam-Train
1 (least)	53.38	63.76	20.67	0	0
2	30.41	24.16	36.67	0.67	1.35
3	10.14	6.71	26.01	7.33	3.33
4	5.41	4.70	11.98	21.34	11.33
5 (most)	0.68	0.67	4.68	70.68	83.99
avg. conf.	1.70	1.54	2.43	4.62	4.78
accuracy	14.19	18.79	60.67	94.00	94.67



(b) Examples of user's correct choice

Object Discrimination --- Applied to Odam-NMS

- Odam-NMS: Using ODAM w/ Odam-Train for object discrimination to help with NMS.



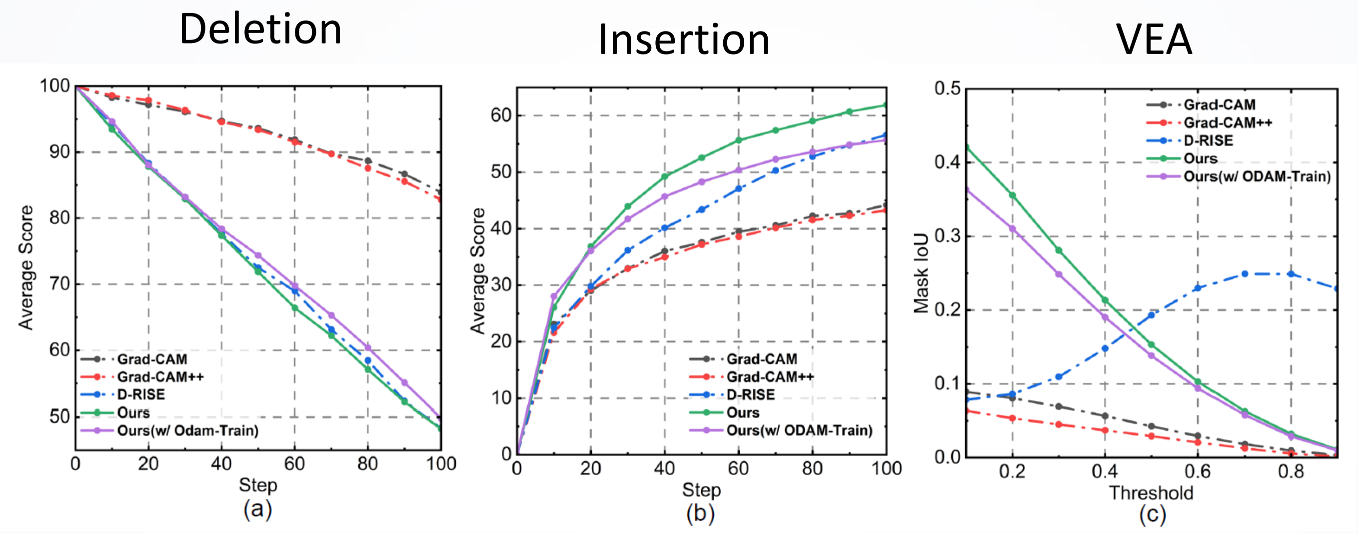
Odam-NMS:

$P \leftarrow GetPredictions(imageI)$
$P \leftarrow SORT(P)$
$D \leftarrow \varnothing$
while $P \neq \varnothing$ **do**
  $p \leftarrow POP(P)$
  $isDuplicate \leftarrow false$
  for $d \in D$ **do**
    $iou \leftarrow GetIoU(p, d)$
    $corr \leftarrow NormCorrelation(S^{(p)}, S^{(d)})$
    if $iou \geqslant T_{iou}$ **and** $corr > T^{l}$ **then**
      $isDuplicate \leftarrow true$
    else if $iou < T_{iou}$ **and** $corr > T^{h}$ **then**
      $isDuplicate \leftarrow true$
    end if
  end for
  if $\neg isDuplicate$ **then**
    $PUSH(p,\ D)$
  end if
end while

- Motivation
- Method & Visualizations
- Results

Quantitative evaluation of ODAM on object specification

- Faithfulness evaluation: Deletion, Insertion and Visual Explanation Accuracy (VEA)

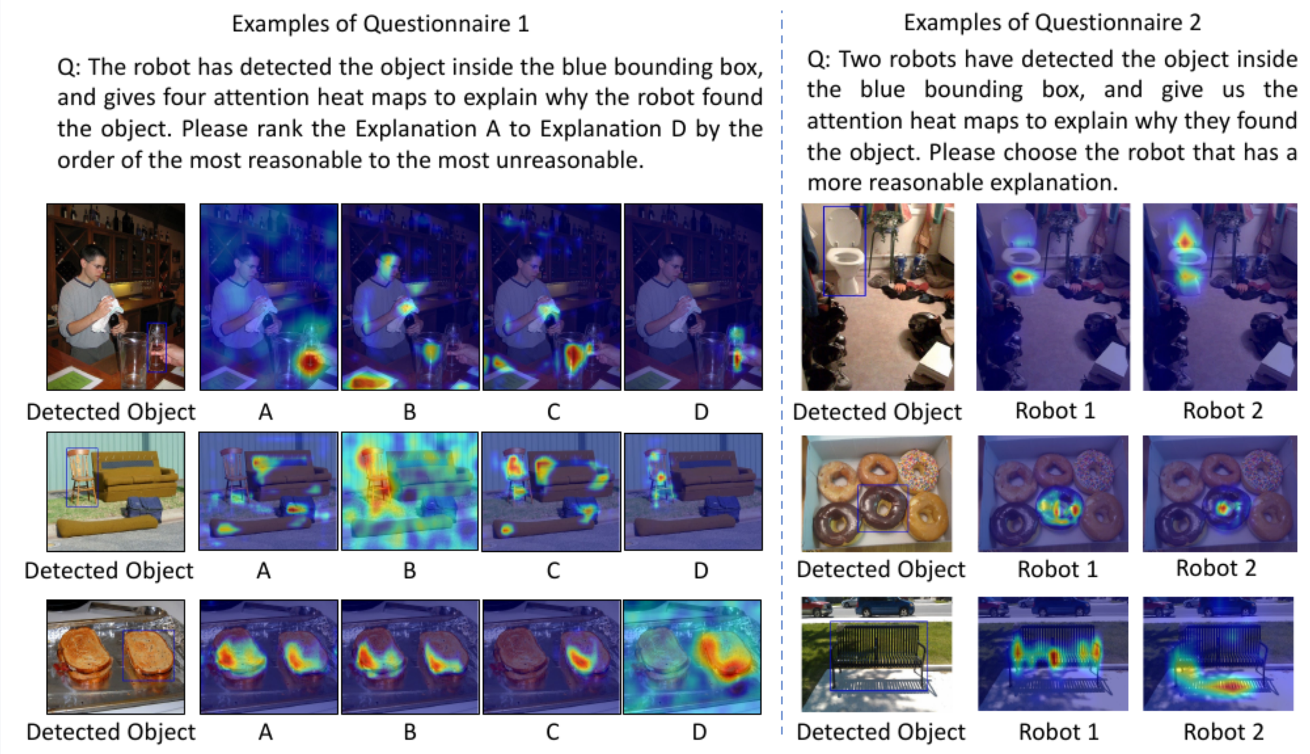


AUC for Deletion, Insertion and VEA curves.

Method	Deletion↓	Insertion↑	VEA↑
Grad-CAM	92.79	36.78	0.039
Grad-CAM++	92.52	36.18	0.027
D-RISE	73.35	43.35	0.157
ODAM	**72.68**	**50.33**	**0.163**
w/ Odam-Train	74.45	46.66	0.143

Quantitative evaluation of ODAM on object specification

- User Trust



Result of Q1: Percentage of rankings for each method and the average rank

Method	1st	2nd	3rd	4th	AR
Grad-CAM	3.9	12.9	30.5	52.7	3.3
Grad-CAM++	7.3	22.2	43.1	27.5	2.9
D-RISE	35.1	29.5	17.5	17.9	2.2
ODAM	**53.8**	35.4	8.9	1.9	**1.6**

Result of Q2: The better model received more responses that its explanations were more trustworthy (38.2% vs. 28.6%).

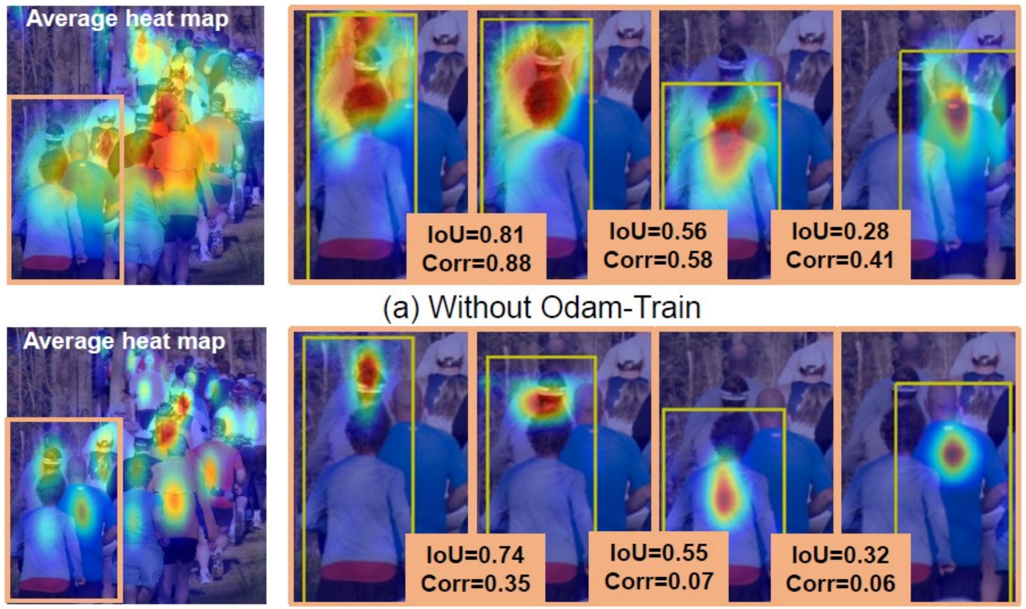
Quantitative evaluation of ODAM on object discrimination

- Localization evaluation: Point Game (PG) and Object Discrimination Index (ODI)

Comparison of Pointing Game (PG) accuracy with ground-truth bounding boxes (b) or segmentation masks (m), energy-based PG with box or mask, Heat Map Compactness (Comp.), Object Discrimination Index (ODI).

	MS COCO							CrowdHuman			
	PG(b) ↑	PG(m) ↑	enPG(b) ↑	enPG(m) ↑	Comp. ↓	ODI(b) ↓	ODI(m) ↓	PG(b) ↑	enPG(b) ↑	Comp. ↓	ODI(b) ↓
Grad-CAM	26.7	22.5	20.7	15.0	4.34	77.0	72.7	15.7	9.7	3.99	91.4
Grad-CAM++	26.6	20.2	20.0	14.8	4.91	77.3	73.2	15.4	11.4	3.84	92.0
D-RISE	82.6	68.0	17.4	12.0	5.17	71.0	66.3	1.5	1.7	3.53	95.3
ODAM	<u>91.9</u>	<u>82.6</u>	<u>73.1</u>	<u>57.1</u>	<u>1.36</u>	<u>34.8</u>	<u>19.5</u>	<u>95.5</u>	<u>79.5</u>	<u>1.04</u>	<u>56.9</u>
w/ Odam-Train	**93.3**	**83.9**	**79.6**	**63.9**	**1.32**	**34.1**	**18.7**	**97.3**	**83.9**	**0.91**	**51.3**

Evaluation of Odam-Train and Odam-NMS



(a) Without Odam-Train

(b) With Odam-Train

Comparison of **recalls** on the “crowd” and “sparse” set from CrowdHuman validation set.

		Faster RCNN			FCOS		
	Ground truth	NMS	+Odam-NMS	Δ	NMS	+Odam-NMS	Δ
Total	99,481	79,090 (79.5%)	**80,111 (80.5%)**	**+1%**	74,946 (75.3%)	**80,650 (81.1%)**	**+5.8%**
Sparse	78,273	65,480 (83.6%)	**65,639 (83.8%)**	**+0.2%**	61,890 (79.0%)	**64,726 (82.7%)**	**+3.7%**
Crowd	21,208	13,610 (64.2%)	**14,472 (68.2%)**	**+4%**	13,056 (61.6%)	**15,924 (75.1%)**	**+13.5%**

Comparisons of NMS strategies on CrowdHuman validation set.

	FCOS					Faster RCNN				
	AP	JI	MR	Recall	time(s/img)	AP	JI	MR	Recall	time(s/img)
NMS	87.8	78.4	45.5	93.2	0.114	86.9	79.5	43.2	90.3	0.092
Soft-NMS	80.8	74.9	89.0	93.0	0.47	76.5	61.9	84.8	92.3	0.284
FeatureNMS	89.3	78.1	45.6	95.4	0.145	82.0	65.7	68.8	**94.9**	0.120
Odam-NMS	**89.3**	**81.1**	**44.5**	**95.5**	0.178	**88.1**	**80.5**	**42.8**	91.5	0.140

Thanks for watching!