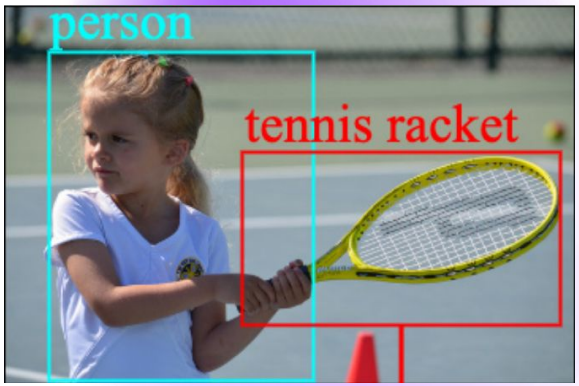
CVPR 2021

BLACK-BOX EXPLANATION OF OBJECT DETECTORS VIA SALIENCY MAPS

presented by:
Julia Chylak, Aleksandra Mysiak

D-RISE: Detector Randomized Input Sampling for Explanation

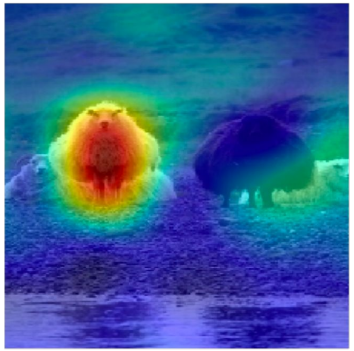


D-RISE

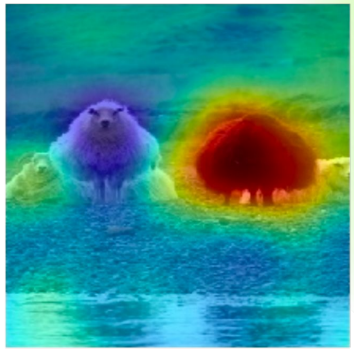
* method of creating saliency maps
* designed for object detection
* measures output disturbances with masked inputs
* based on RISE

RISE – RESULTS

(a) Sheep - 26%, Cow - 17%



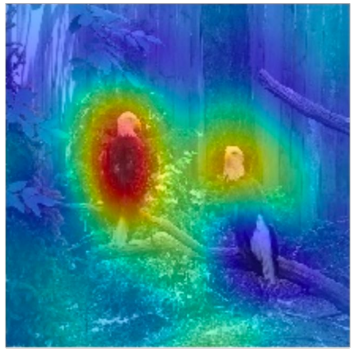
(b) Importance map of '*sheep*'



(c) Importance map of '*cow*'

(d) Bird - 100%, Person - 39%

(e) Importance map of '*bird*'



(f) Importance map of '*person*'

WHY DO WE NEED D-RISE?

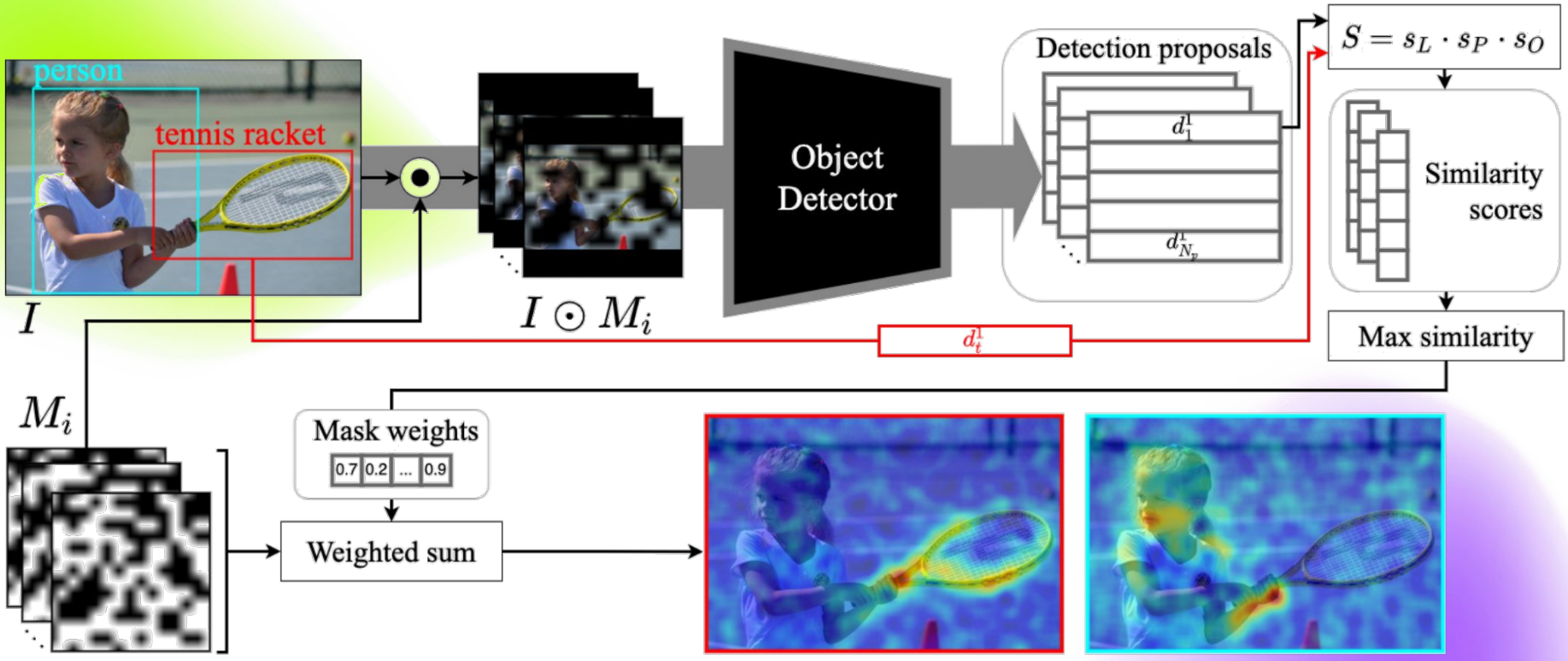
Detection is not classification:

* location matters

* multiple proposals per object

* not a single vector of class probabilities

➔ existing methods for classification are not applicable


person
tennis racket
I
$I \odot M_i$
Object Detector
Detection proposals
d_1^1
$d_{N_p}^1$
$S = s_L \cdot s_P \cdot s_O$
Similarity scores
Max similarity
d_i^1
M_i
Mask weights
0.7
0.2
...
0.9
Weighted sum

DETECTION VECTOR

$$d_i = \left[L_i, O_i, P_i\right]$$

$$= \left[(x_1^i, y_1^i, x_2^i, y_2^i), O_i, (p_1^i, \ldots, p_C^i)\right]$$

[bounding box, objectness score, class probabilities]

SIMILARITY SCORES AND SELECTION

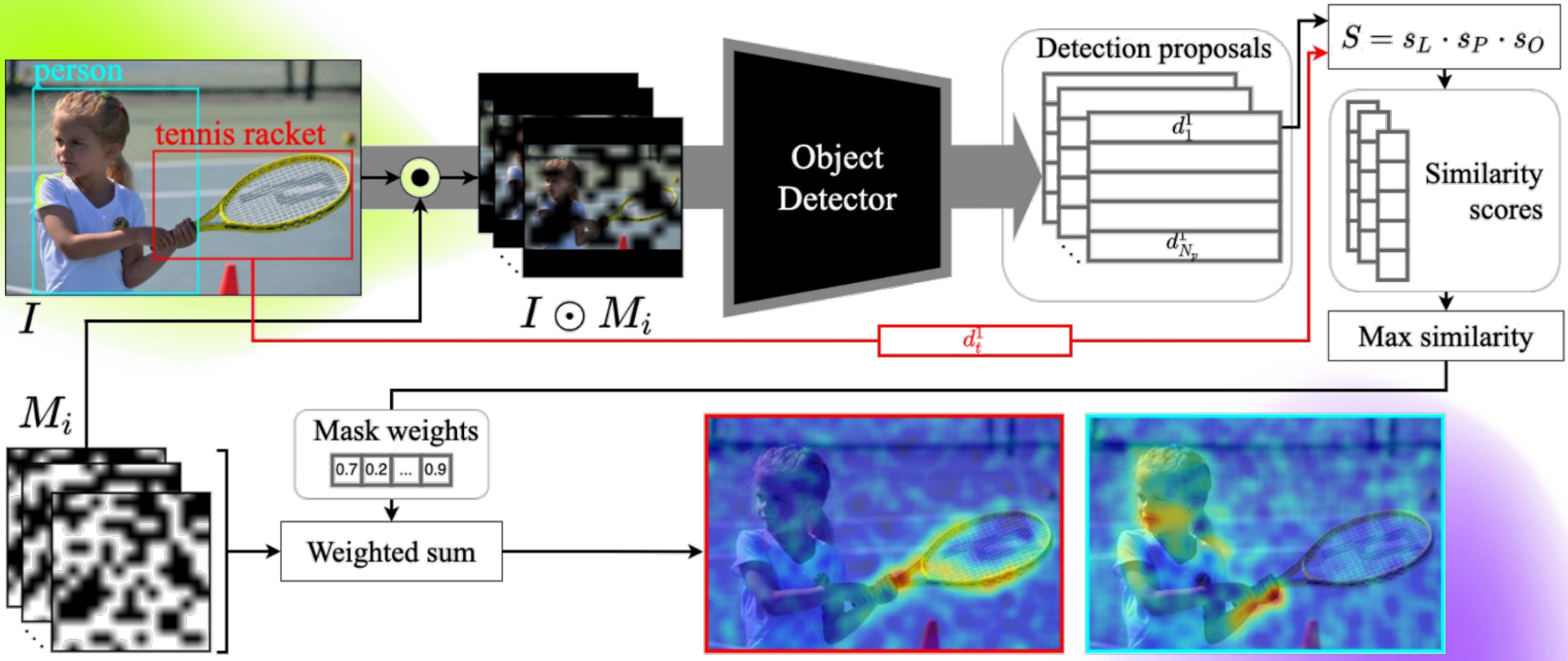
how:

$$s(d_t, d_j) = s_L(d_t, d_j) \cdot s_P(d_t, d_j) \cdot s_O(d_t, d_j)$$

$$s_L(d_t, d_j) = \mathrm{IoU}(L_t, L_j) \qquad s_P(d_t, d_j) = \frac{P_t \cdot P_j}{\|P_t\| \|P_j\|} \qquad s_O(d_t, d_j) = O_j$$

where:

$$S(d_t, f(M_i \odot I))) \triangleq \max_{d_j \in f(M_i \odot I)} s(d_t, d_j)$$


person
tennis racket
I
$I \odot M_i$
Object Detector
Detection proposals
d_1^1
$d_{N_p}^1$
$S = s_L \cdot s_P \cdot s_O$
Similarity scores
Max similarity
d_i^1
M_i
Mask weights
0.7
0.2
...
0.9
Weighted sum

ADVANTAGES OF D-RISE

first black-box method for explaining object detectors:

* takes object location into account

* allows multiple proposals per object

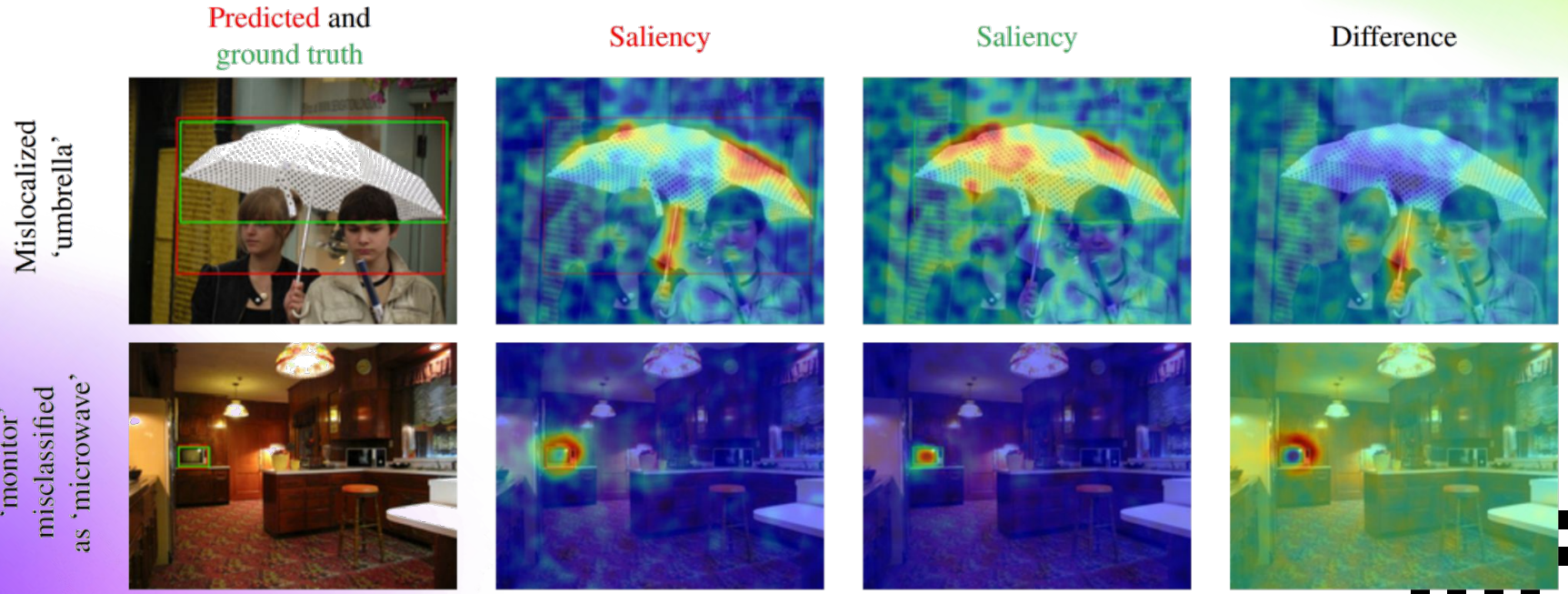
* agnostic to type of model

* can produce explanations for arbitrary detections

RESULTS

MISSED DETECTION

POOR LOCALIZATION AND MISCLASSIFICATION



CORRECT PREDICTIONS

USER TRUST

* YOLOv3 (55.3% mAP) vs YOLOv3-Tiny (33.1% mAP)

* 242 correct predictions’ explanations rated by humans

* more users found explanations from the more accurate model to be better (50.2% vs 27.4%)

SUMMARY

* D-RISE is a black-box method of explaining object detectors

* it can be used to analyse multiple types of errors

* explanations of more accurate networks seem more trustworthy

* we really liked the paper :)

THANK YOU FOR YOUR ATTENTION!