Understanding Deep Neural Networks Via Smooth Masks and Extremal Perturbations

Ruth Fong*, Mandela Patrick*, Andrea Vedaldi

ICCV 2019

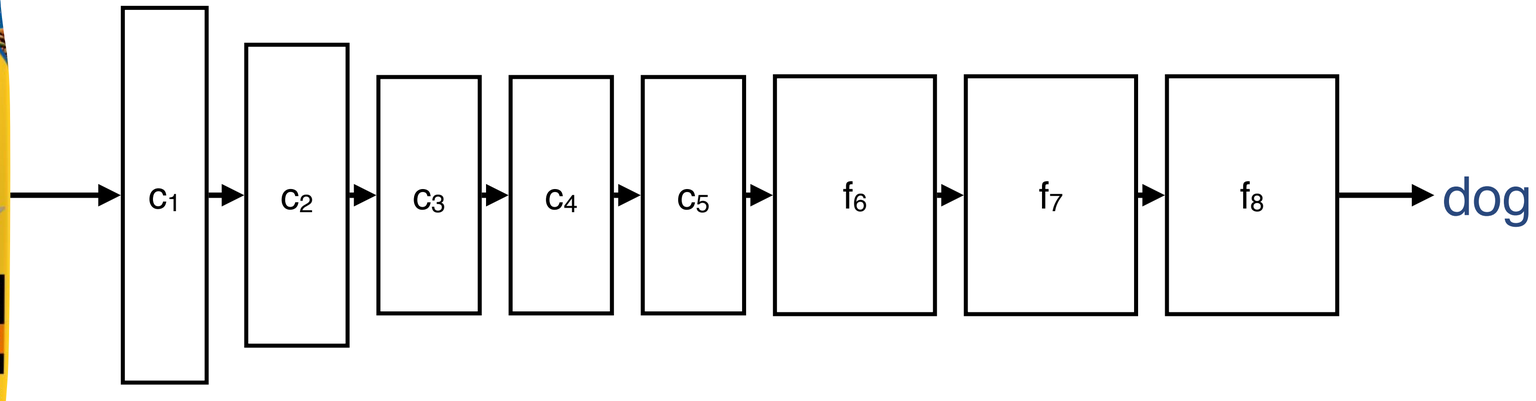
* equal contribution

UNIVERSITY OF OXFORD

FACEBOOK

Attribution

Where is the model looking?



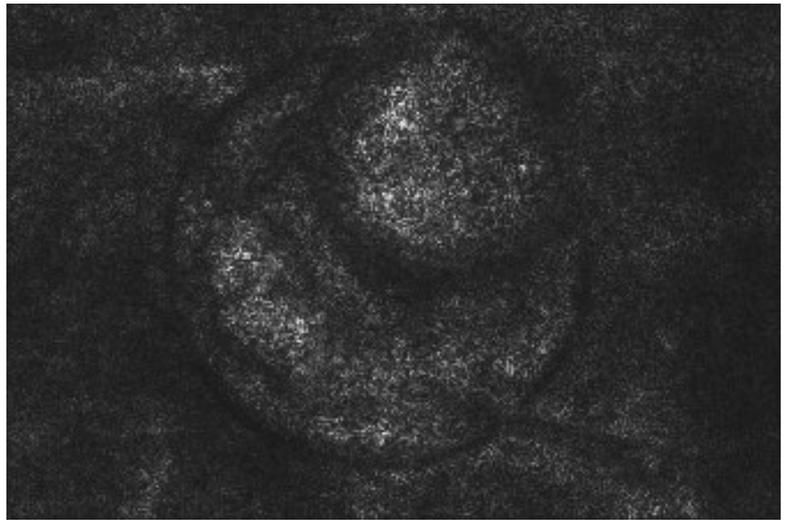
?

Backpropogation

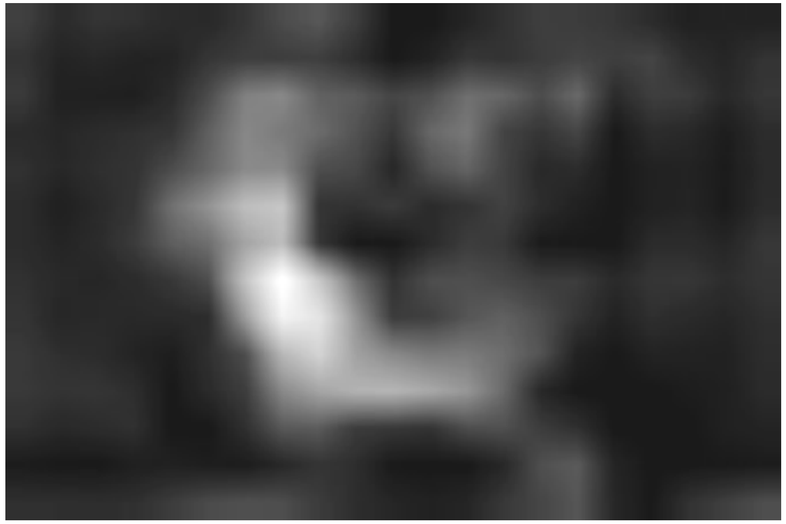
Combine network activations and gradients

Input

Gradient



GRAD-CAM



Fast, but difficult to characterize

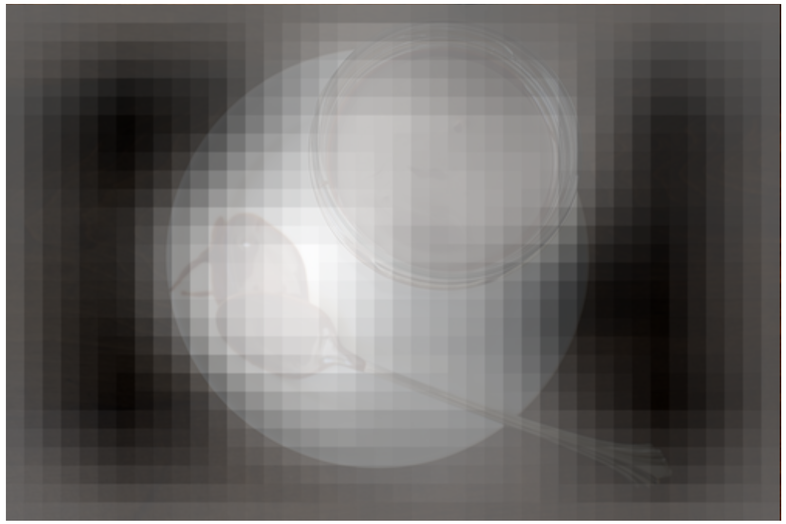
[Simonyan et al., ICLR Workshop 2014; Selvaraju et al., ICCV 2017]
[Mahendran and Vedaldi, ECCV 2016; Adebayo et al., NeurIPS 2018]

Perturbation

Change the input and observe the effect on the output

Input

Occlusion



RISE

Clear meaning, but can only test a small number of occlusion patterns

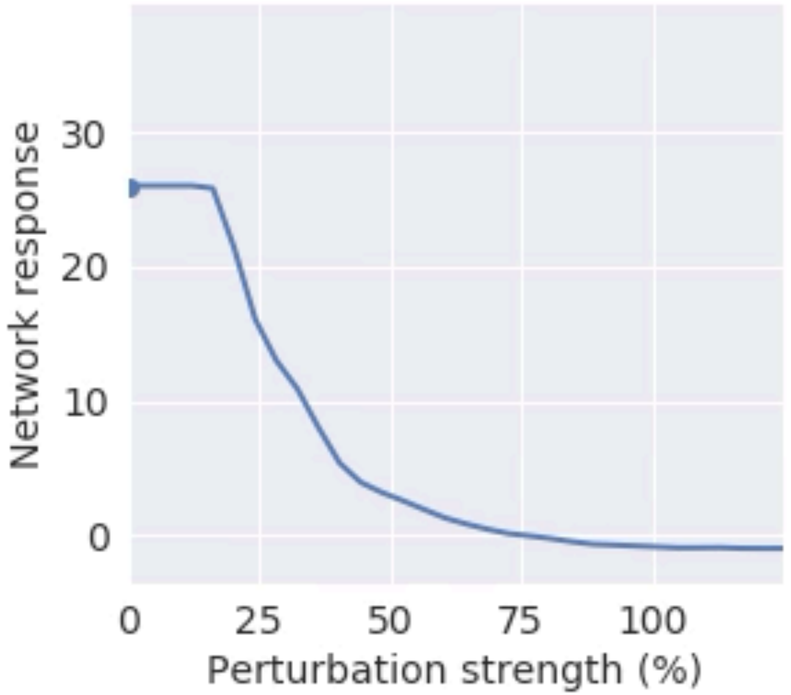
[Zeiler and Fergus, ECCV 2014; Petsiuk et al., BMVC 2018]

Blur everywhere => response suppressed

Retained region

Perturbed stimulus

Response: 26.01

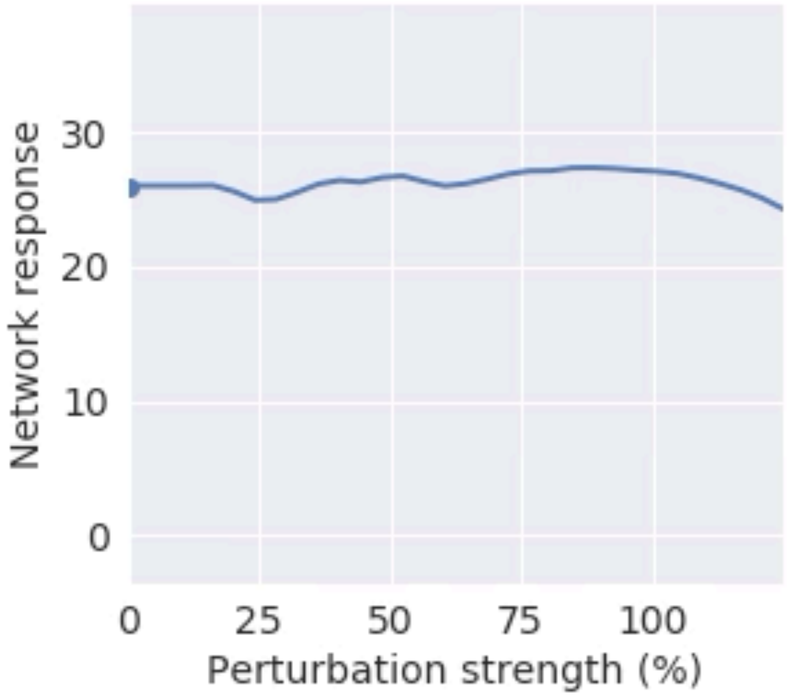


Preserve 10% => response preserved

Retained region

Perturbed stimulus

Response: 26.01

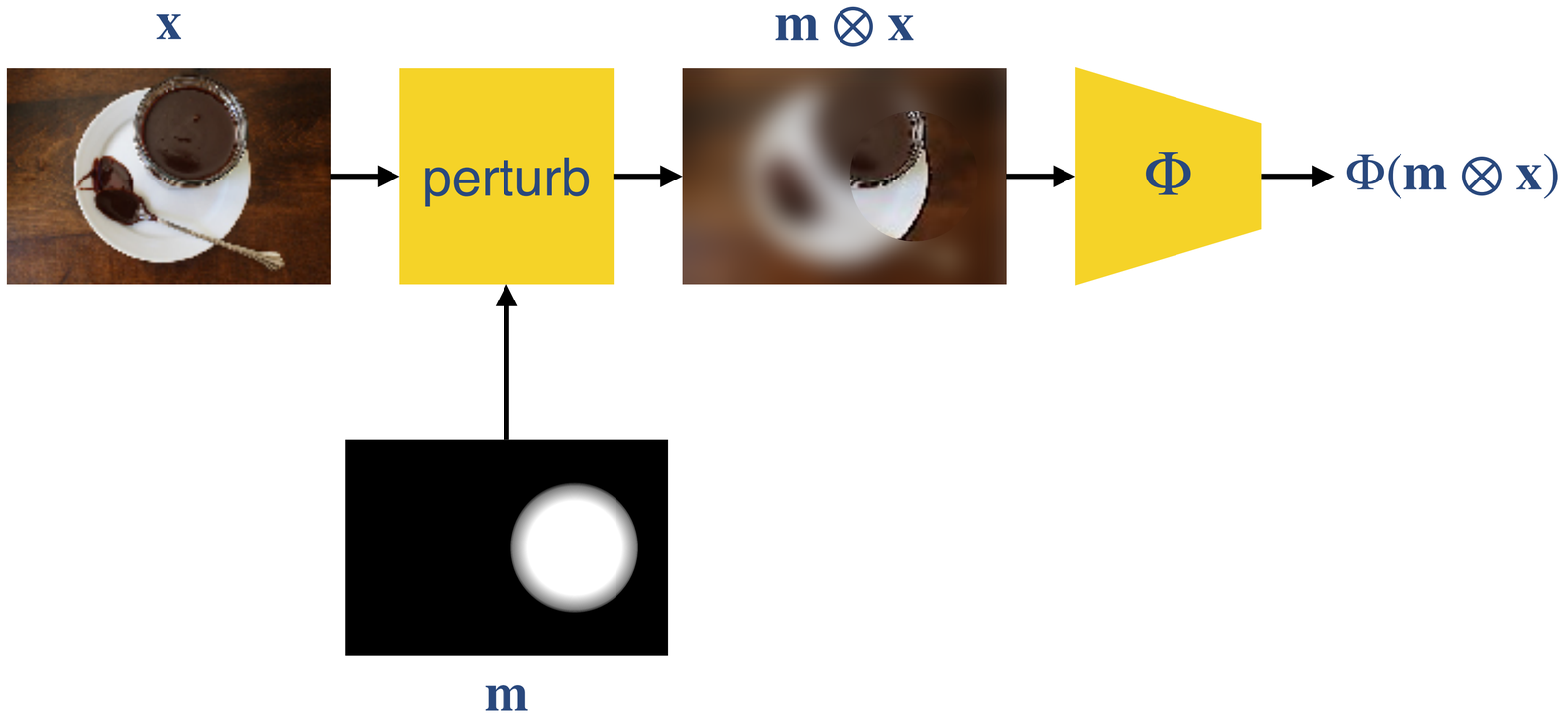


Extremal Perturbations

A mask is optimized to maximally excite the network:

$$\underset{\mathbf{m}}{\operatorname{argmax}}\ \Phi(\mathbf{m} \otimes \mathbf{x})$$

subject to $\operatorname{area}(\mathbf{m}) = a$

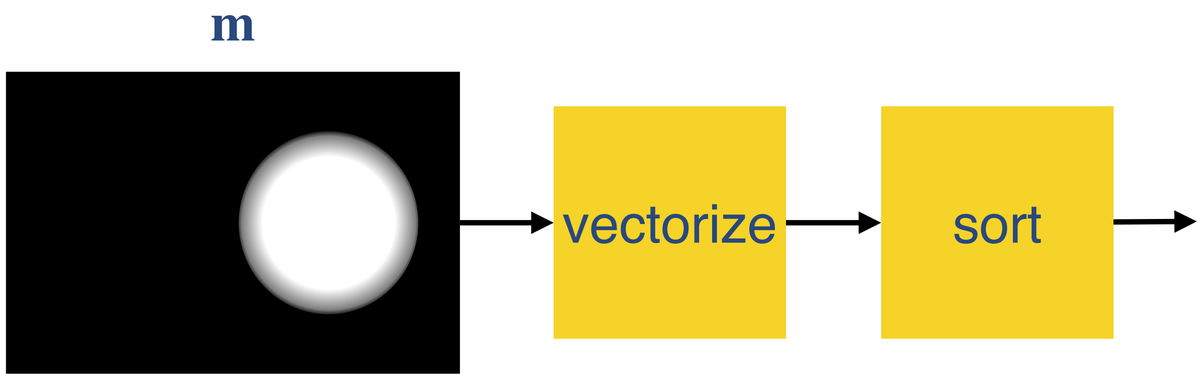


Area Constraint

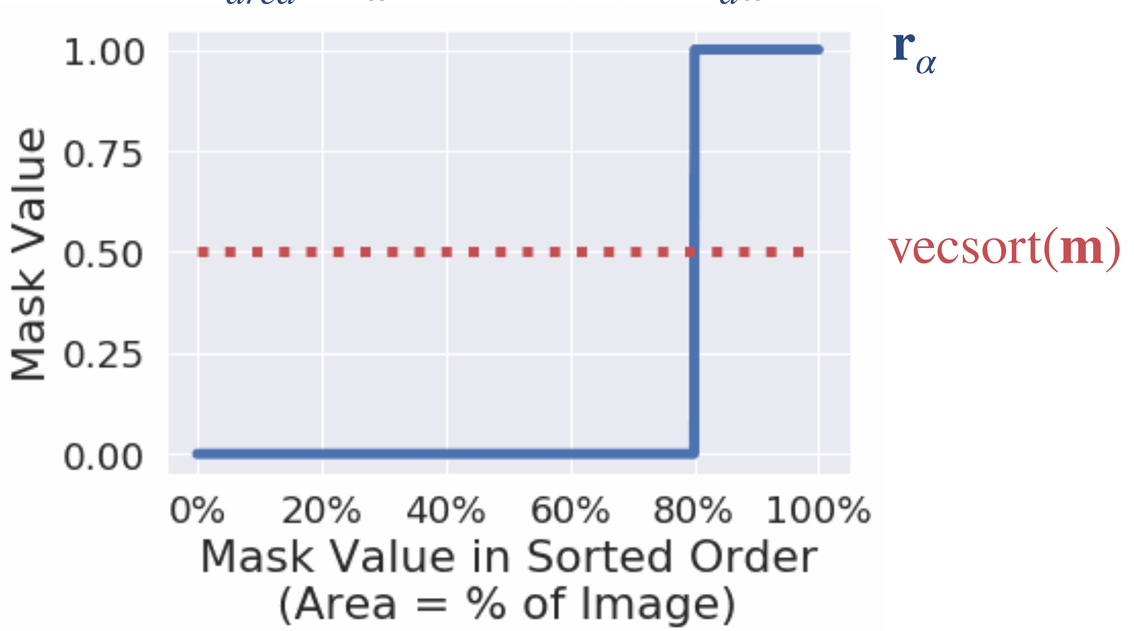
Optimizing w.r.t. to an area constraint is challenging

Here we re-formulate it as matching a **rank statistics**

$$L_{area} = \| \text{vecsort}(\mathbf{m}) - \mathbf{r}_a \|^2$$

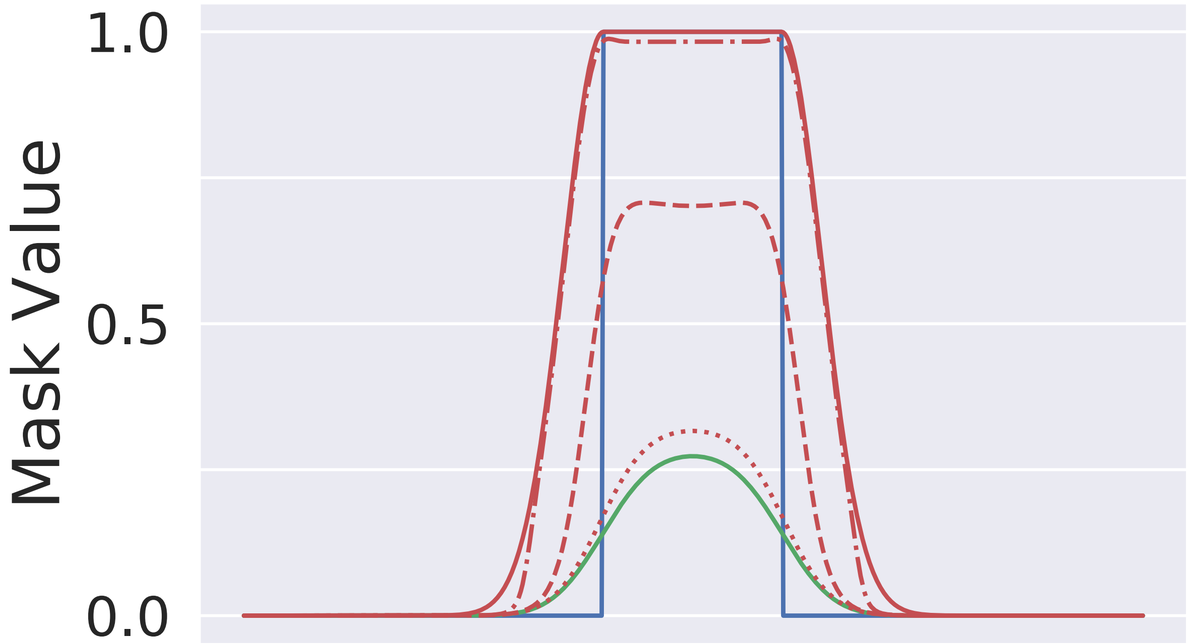


subject to $\text{area}(\mathbf{m}) = a$



Concurrent work: [Kapishnikov et al., arXiv 2019]

Smooth Masks



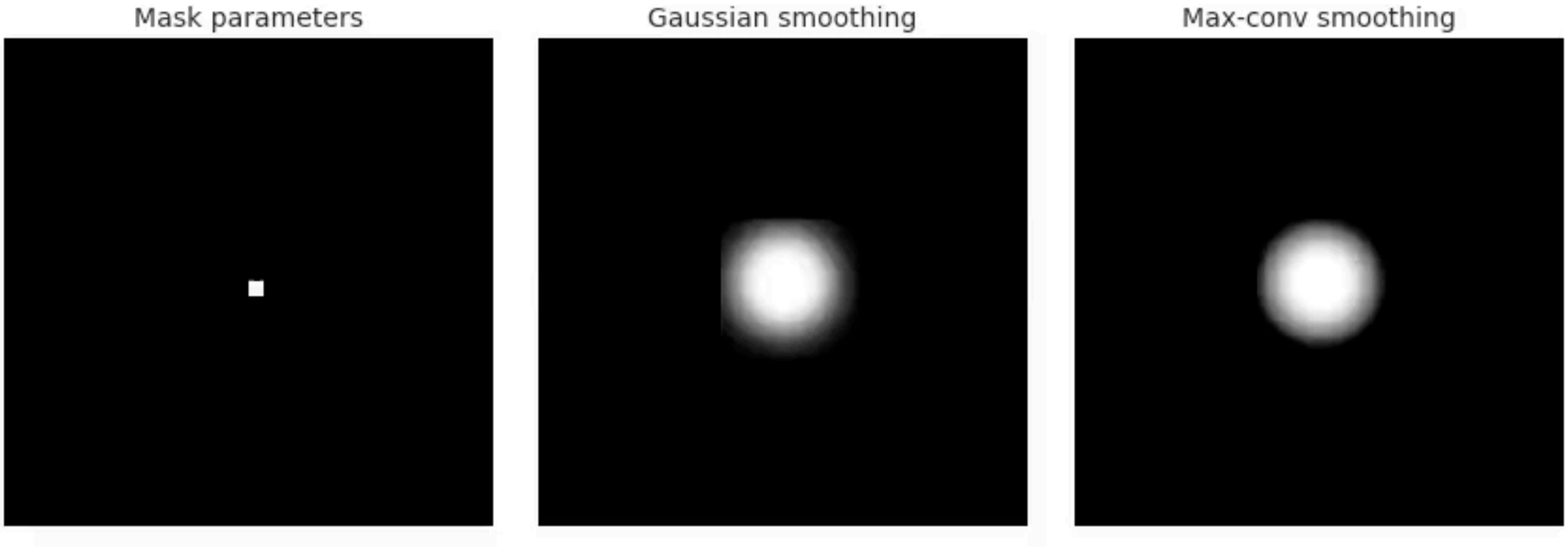
$m(v)$: mask

$$\text{conv}(u; m; k) = \frac{1}{Z} \sum_{v \in \Omega} k(u - v) m(v)$$

$$\text{maxconv}(u; m; k) = \max_{v \in \Omega} k(u - v) m(v)$$

$$\text{smoothconv}(u; m; k; T) = \text{smax}_{v \in \Omega; T}\, k(u - v) m(v)$$

Smooth Masks



Algorithm

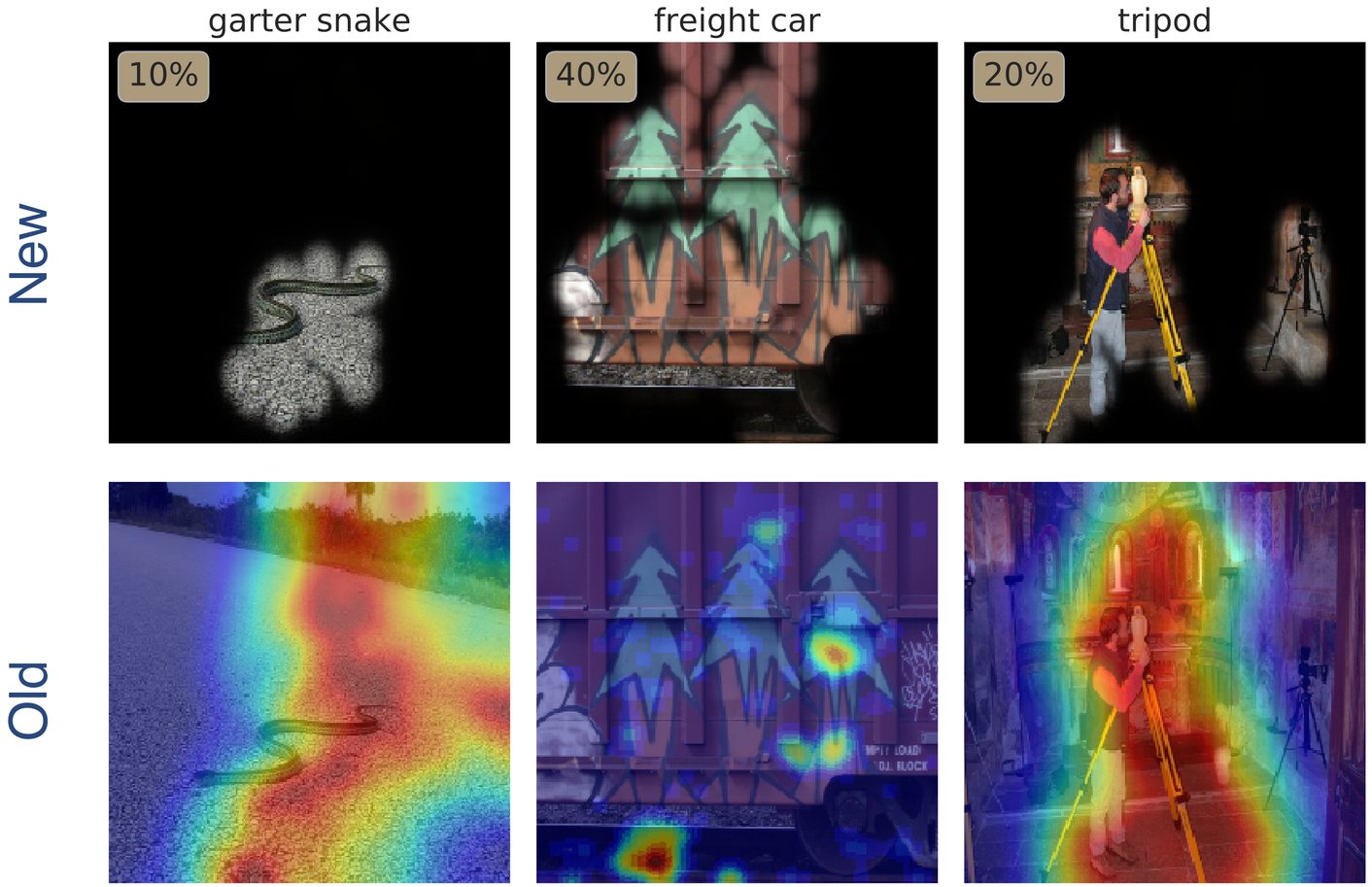
1. Pick an area a

2. Use SGD to solve the optimization problem for a large λ:

$$\underset{\mathbf{m}}{\operatorname{argmax}}\ \Phi(\text{smooth}(\mathbf{m}) \otimes \mathbf{x}) - \lambda \| \text{vecsort}(\text{smooth}(\mathbf{m})) - \mathbf{r}_a \|^2$$

3. If needed, sweep a and repeat

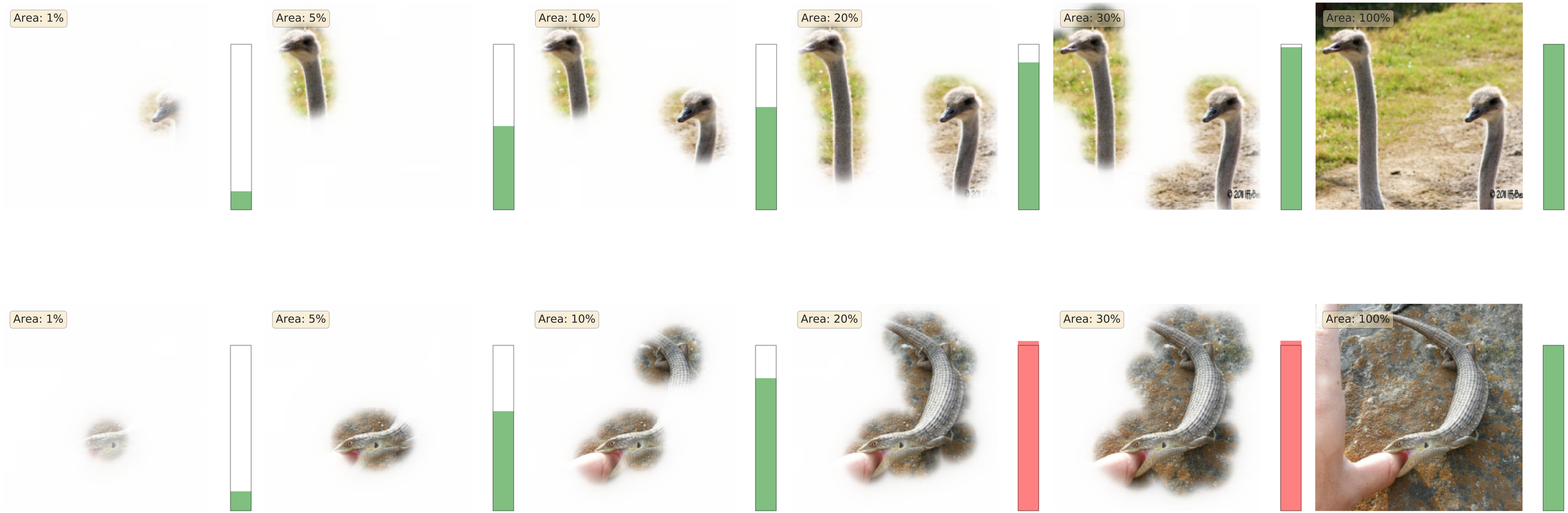
Comparison to prior work on “meaningful perturbations”



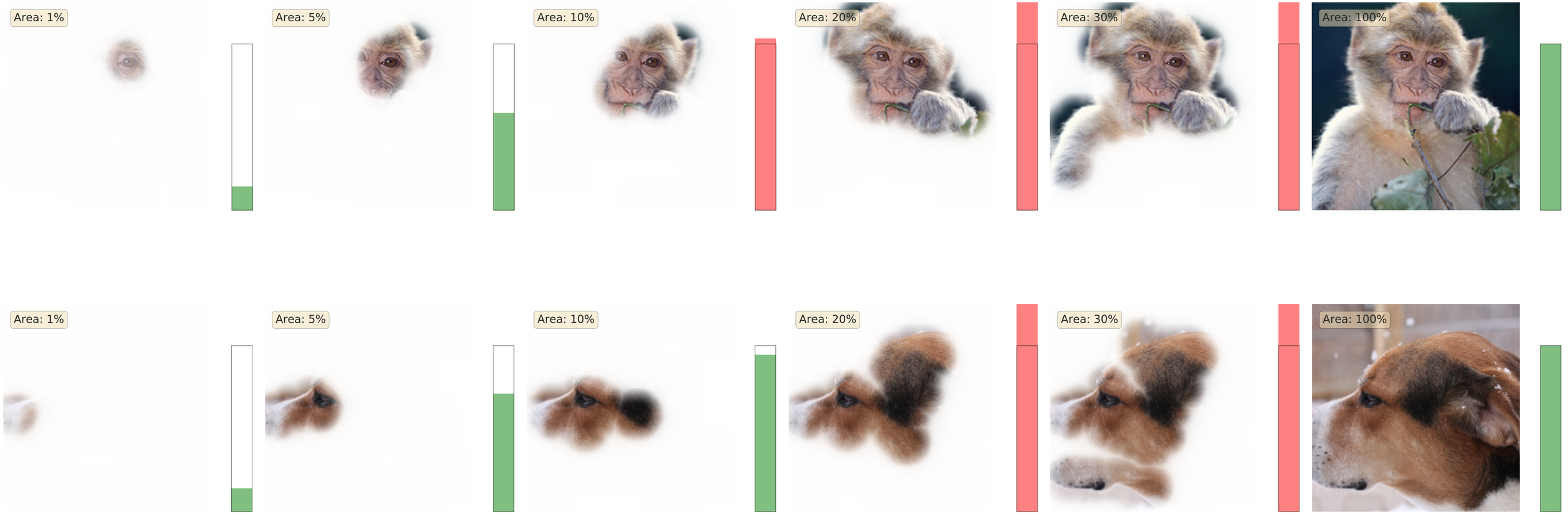
[Fong & Vedaldi, ICCV 2017]

Results

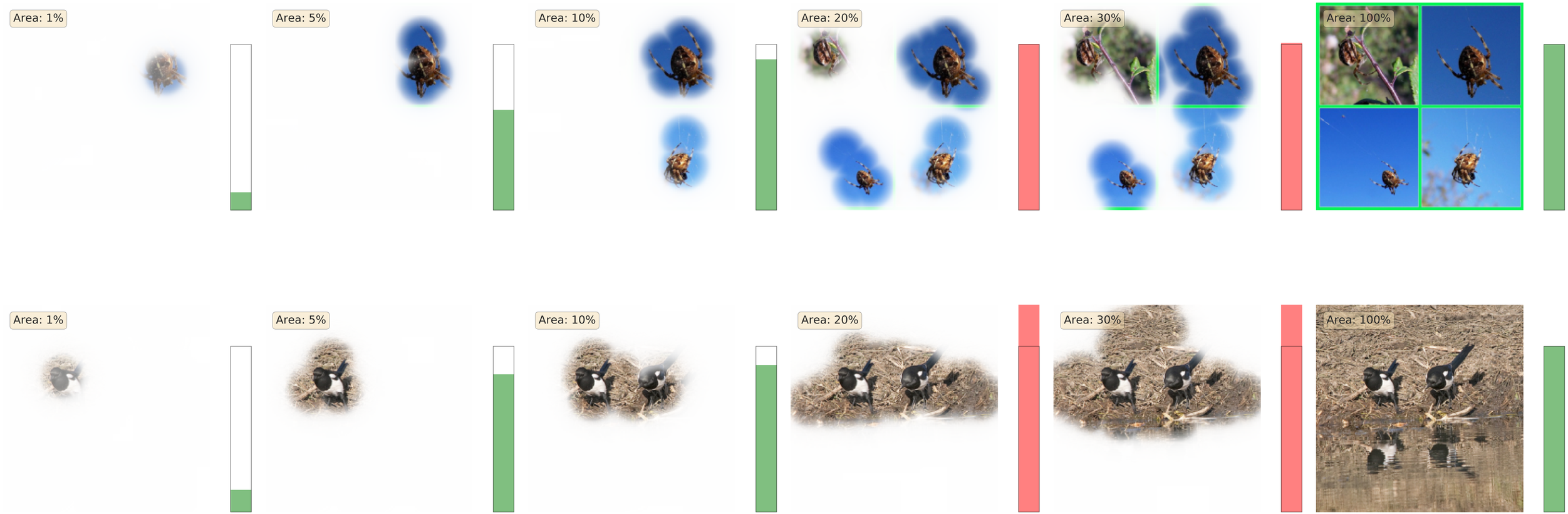
Foreground evidence is usually sufficient



Large objects are recognised by their details

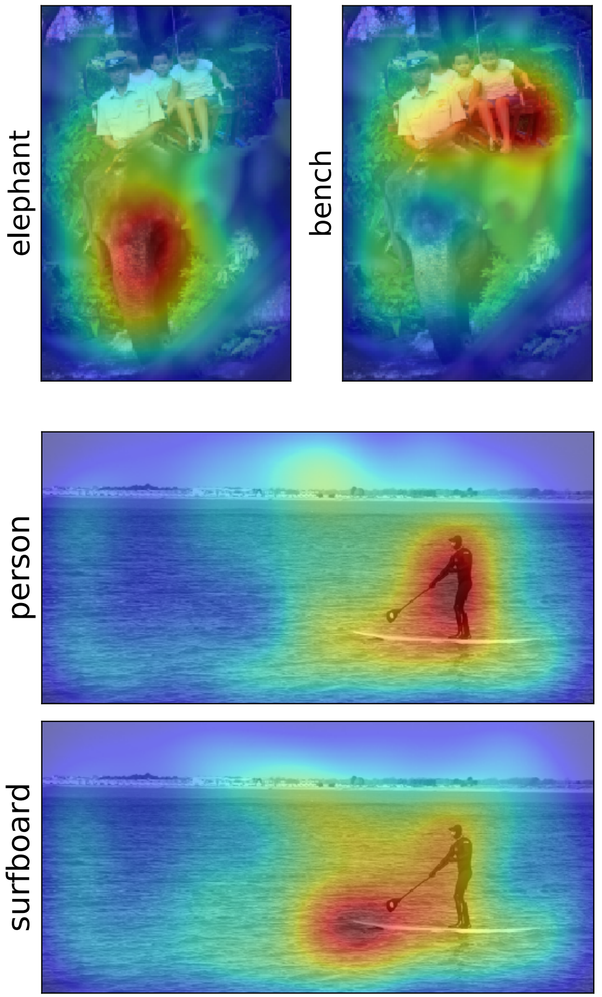


Multiple objects contribute cumulatively



Pointing game: weak localization

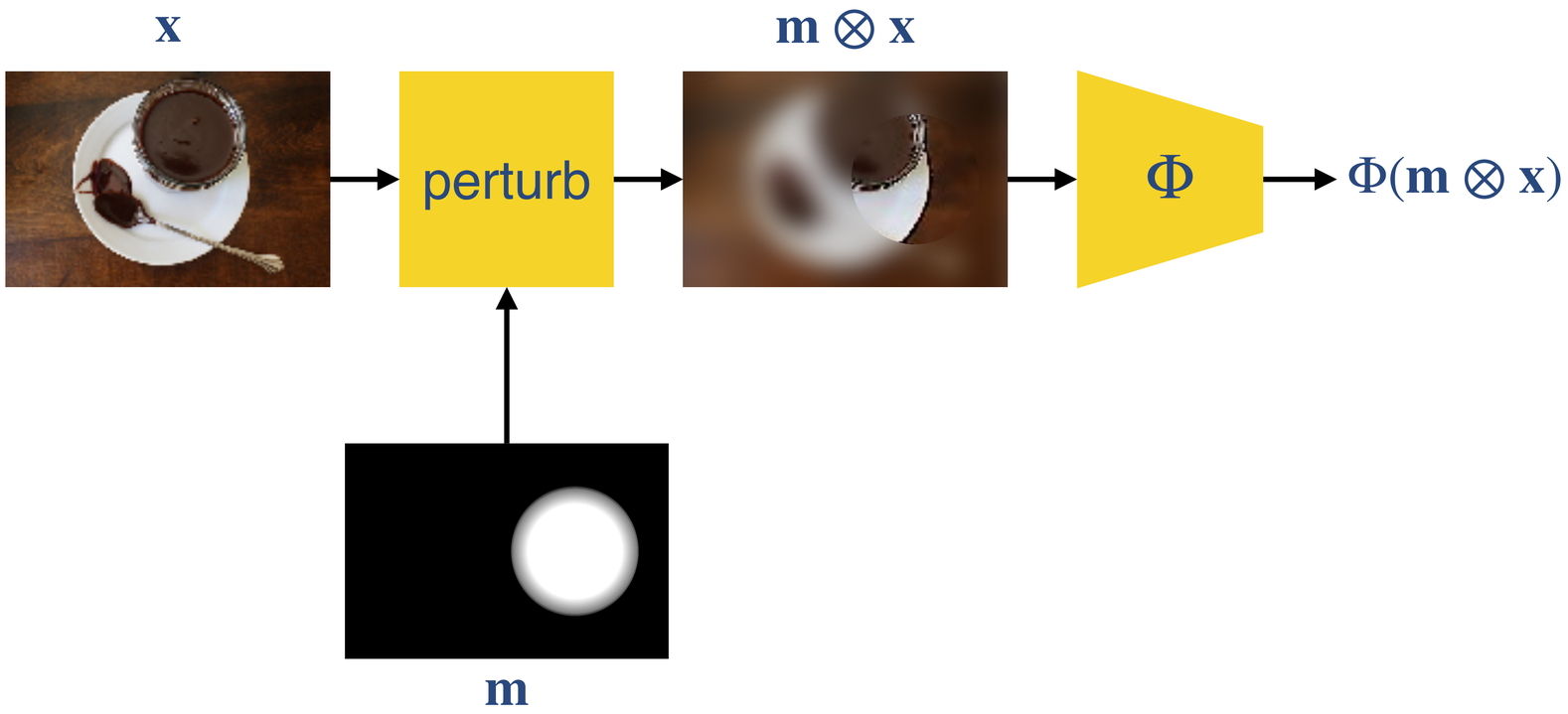
	VOC07 Test (All/Diff)		*COCO14 Val (All/Diff)*	
Method	*VGG16*	*ResNet50*	*VGG16*	*ResNet50*
Grad	76.3/56.9	72.3/56.8	37.7/31.4	35.0/29.4
DConv	67.5/44.1	68.6/44.7	30.7/23.0	30.0/21.9
Guid.	75.9/53.0	77.2/59.5	39.1/31.4	42.1/35.3
MWP	77.1/56.6	84.4/70.8	39.8/32.8	49.6/43.9
cMWP	79.9/66.5	**90.6**/82.2	49.7/44.3	**58.5/53.6**
RISE*	87.3/——	88.9/——	50.7/——	55.6/——
GCAM	86.6/74.0	90.4/**82.3**	**54.2/49.0**	57.3/52.3
Ours	**88.7/75.5**	86.3/73.4	53.4/47.7	55.7/46.9



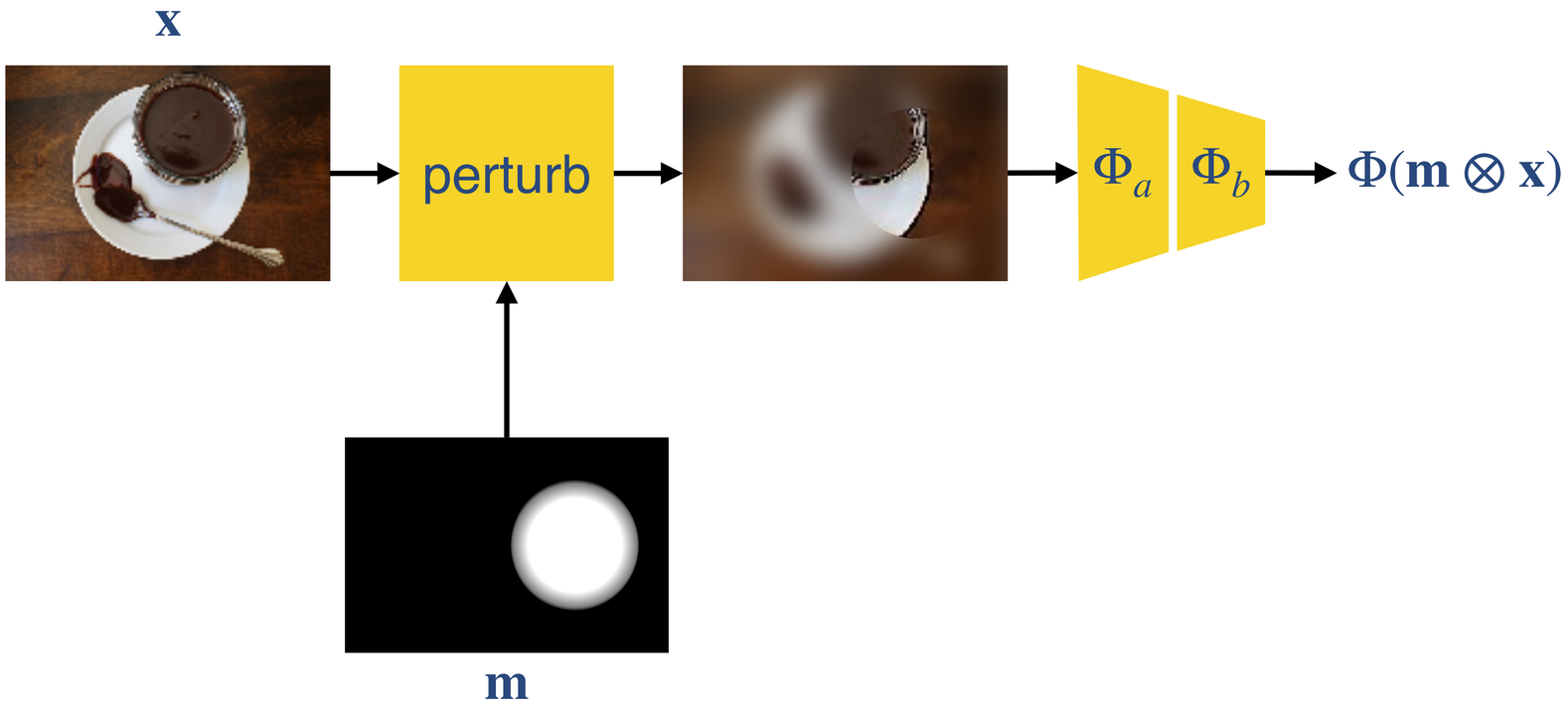
[Zhang et al., ECCV 2016]

Attributing channels at intermediate layers

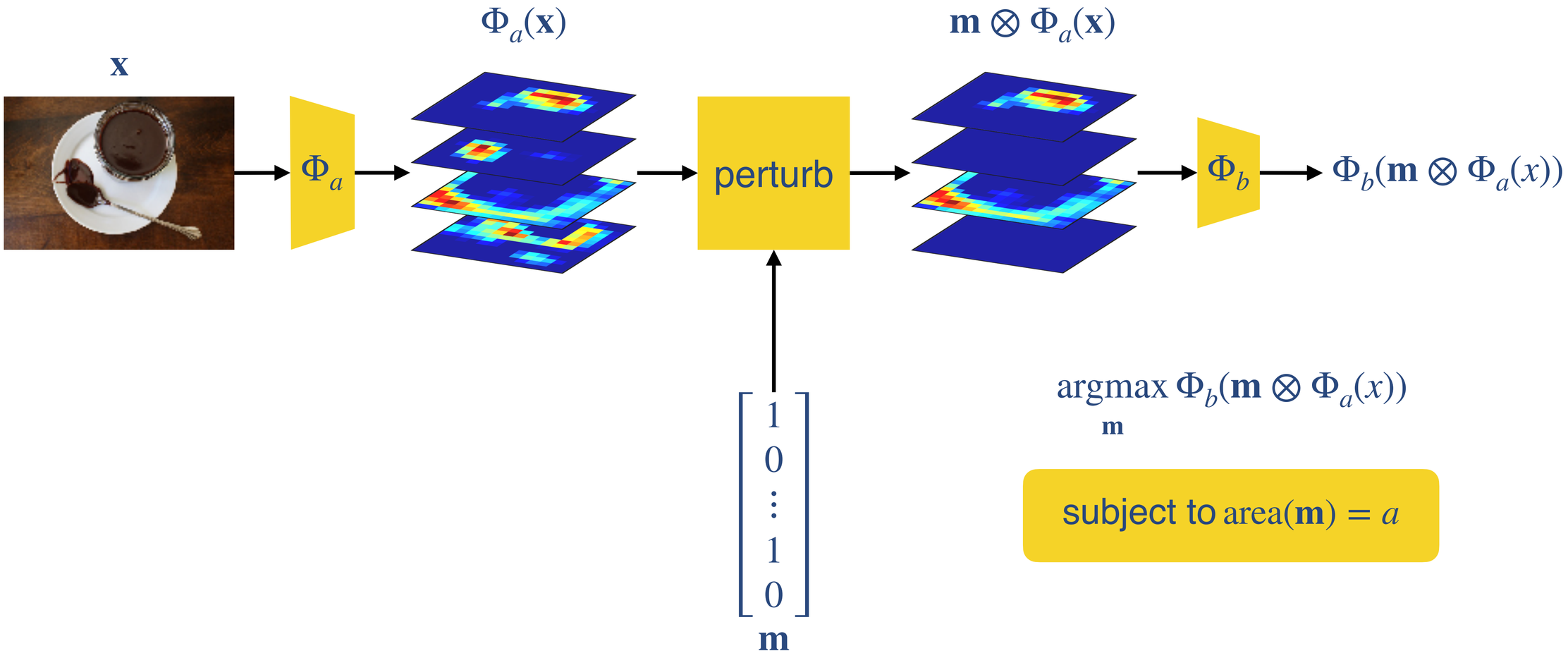
Spatial Attribution



Channel Attribution



Channel Attribution



Activation “diffing”

[Olah et al., Distill 2017]