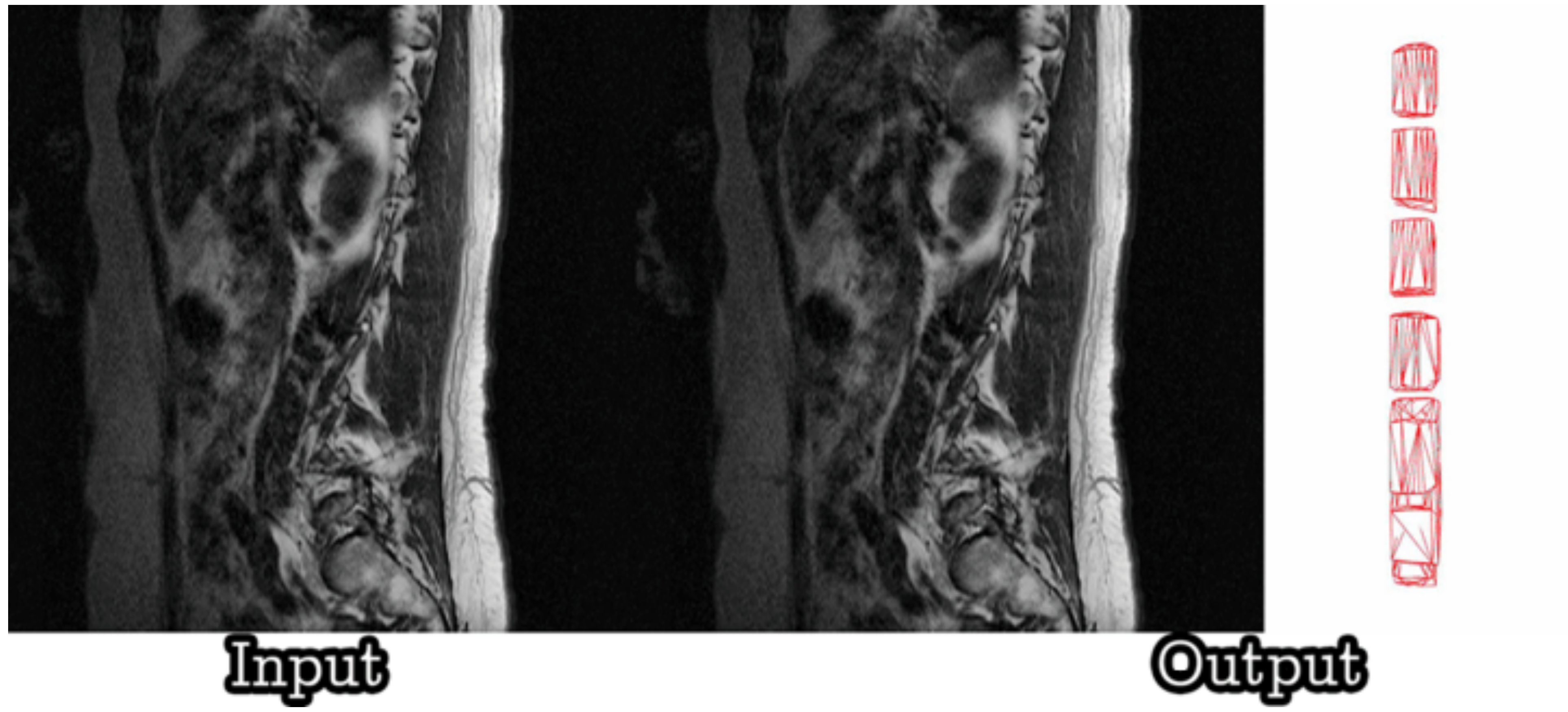
Understanding Deep Neural Networks

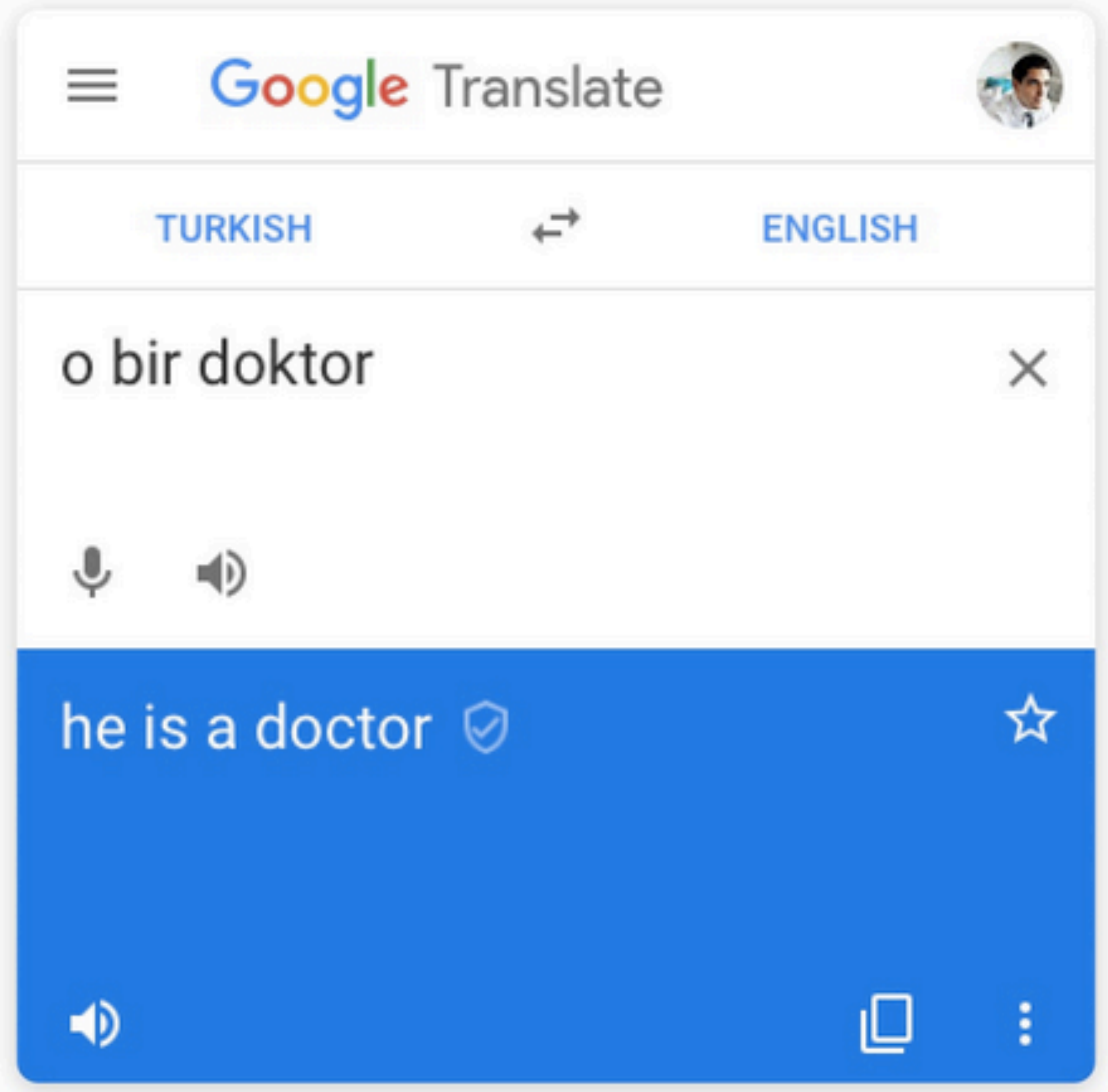
Ruth Fong

CVPR 2020 Tutorial on Interpretable Machine Learning for Computer Vision

Saturday, 13 June 2020

Applications of Deep Learning





Interpretability tools are crucial for high-impact, high-risk applications of deep learning.

[Jamaludin et al., 2017; https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html]

A Brief Primer on Deep Learning

Supervised Learning

(

, “[illegible]”)

x

y

[Russakovsky et al. , IJCV 2015]

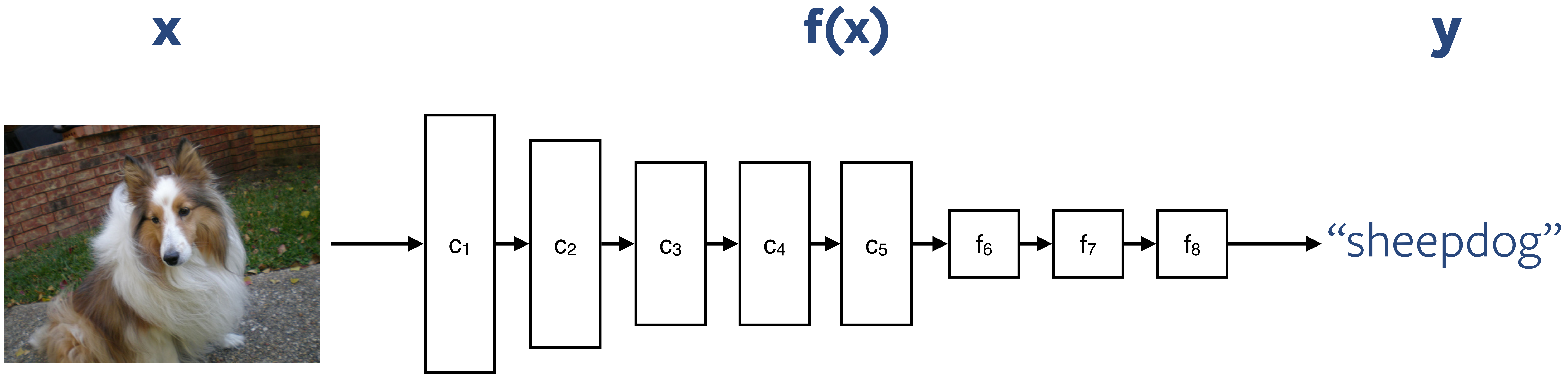
Supervised Learning

( , "sheepdog")

x **y**

[Russakovsky et al. , IJCV 2015]

Deep Learning

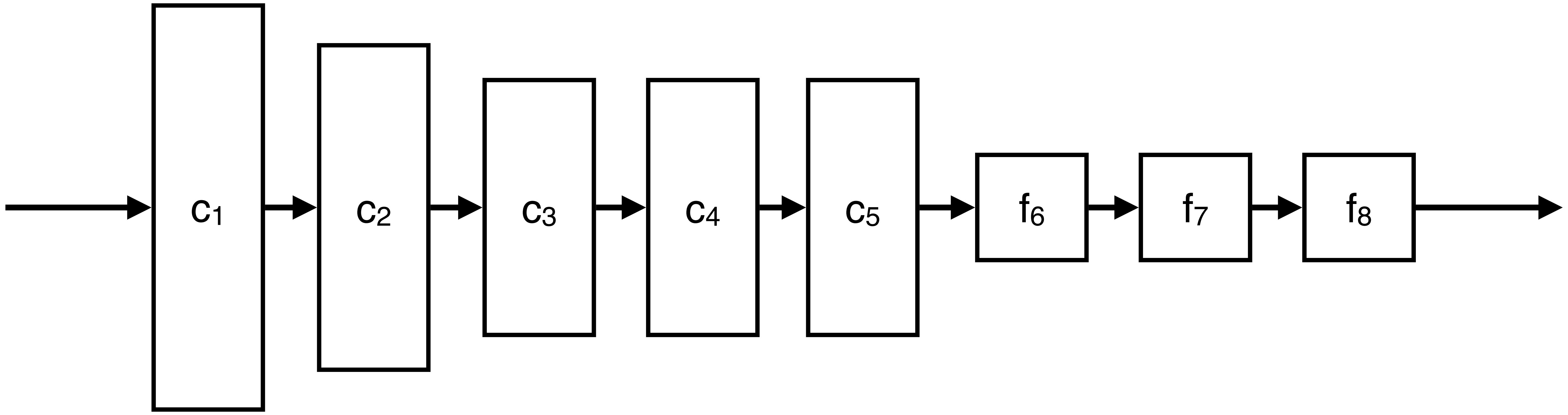


Network built up of layers, with weights θ connecting one layer to the next

Update rule: $\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$, maximizes probability of correct prediction

Deep Learning

x **f(x)** **ŷ**

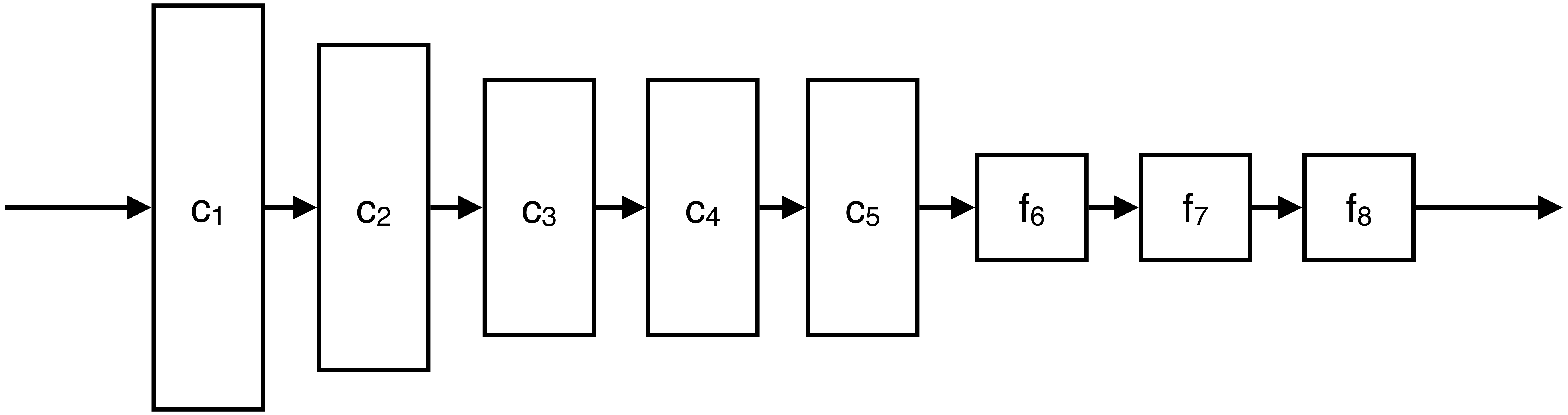


10%: “sheepdog”
10%: “sea snake”
10%: “alp”
10%: “soup bowl”
...

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Deep Learning

x **f(x)** **ŷ**



33%: “sheepdog”

5%: “sea snake”

8%: “alp”

4%: “soup bowl”

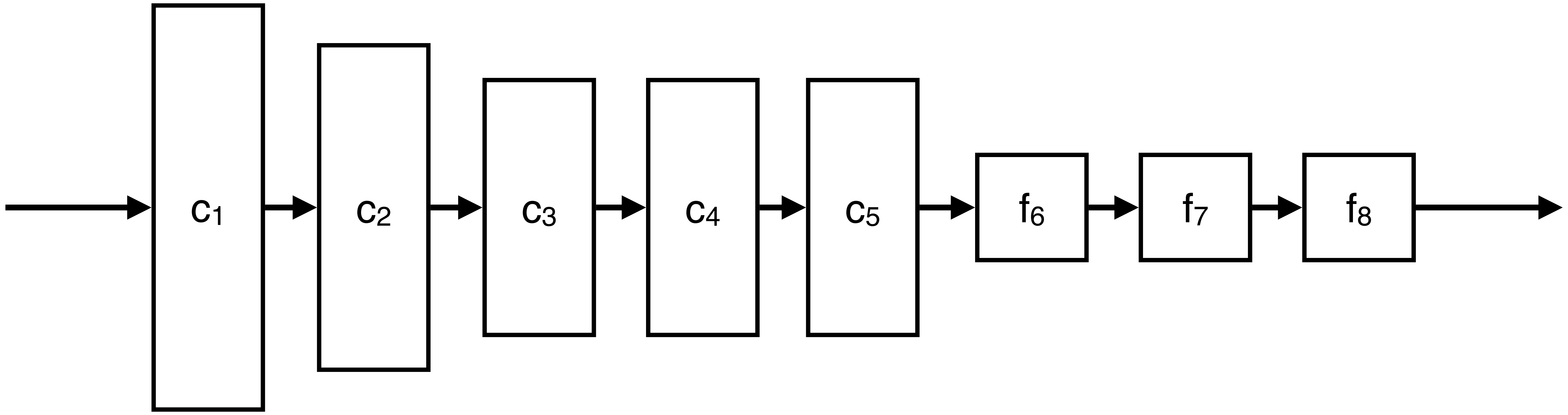
…

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Deep Learning

x

f(x)



ŷ

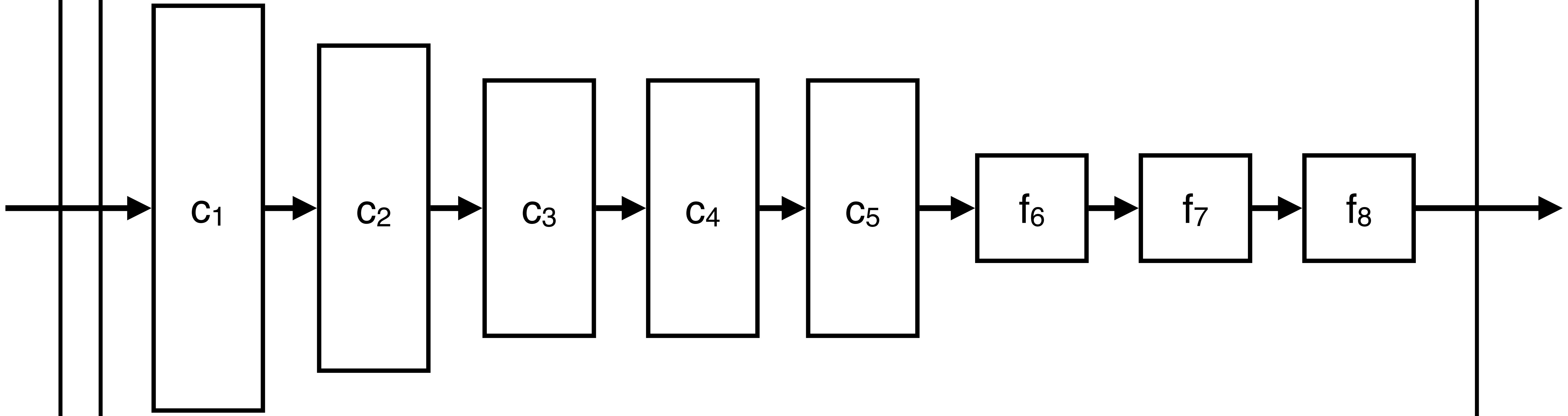
95%: “sheepdog”
1%: “sea snake”
2%: “alp”
0%: “soup bowl”
...

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Research Themes

Inputs: What is f(x) looking at?

Internal Representation: What & how does f(x) encode?



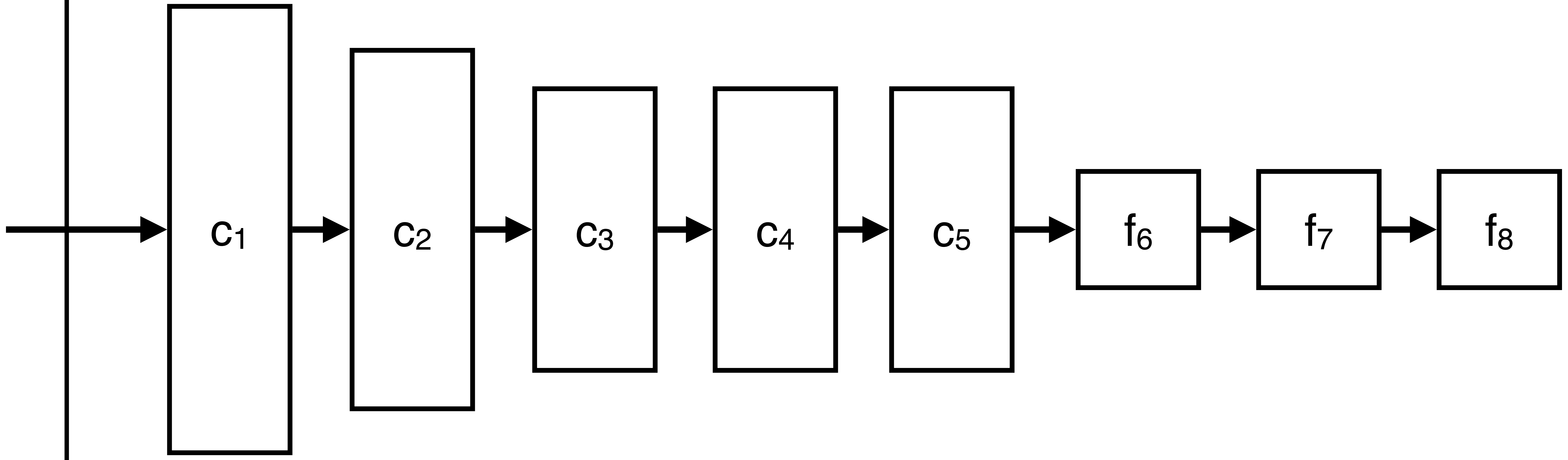
95%: “sheepdog”
1%: “sea snake”
2%: “alp”
0%: “soup bowl”
...

Training Procedure: How can we improve f(x)?

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

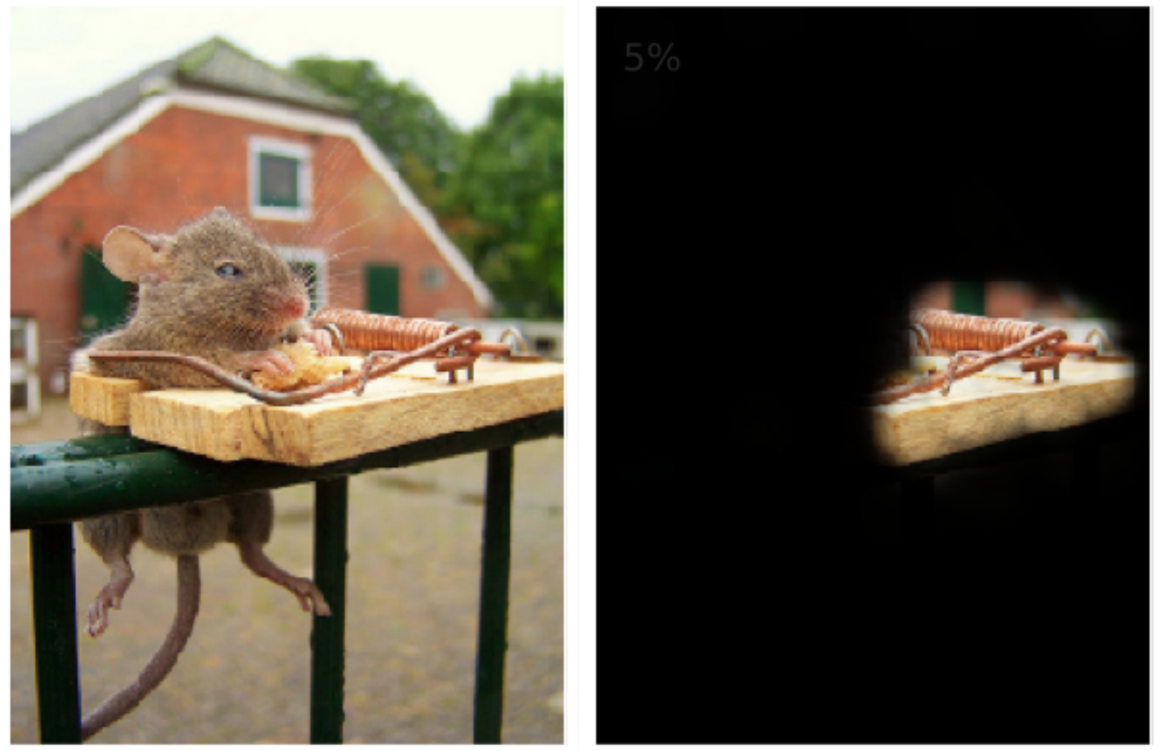
Research Themes

Inputs: What is f(x) looking at?



$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Fong & Vedaldi, ICCV 2017



Fong et al., ICCV 2019

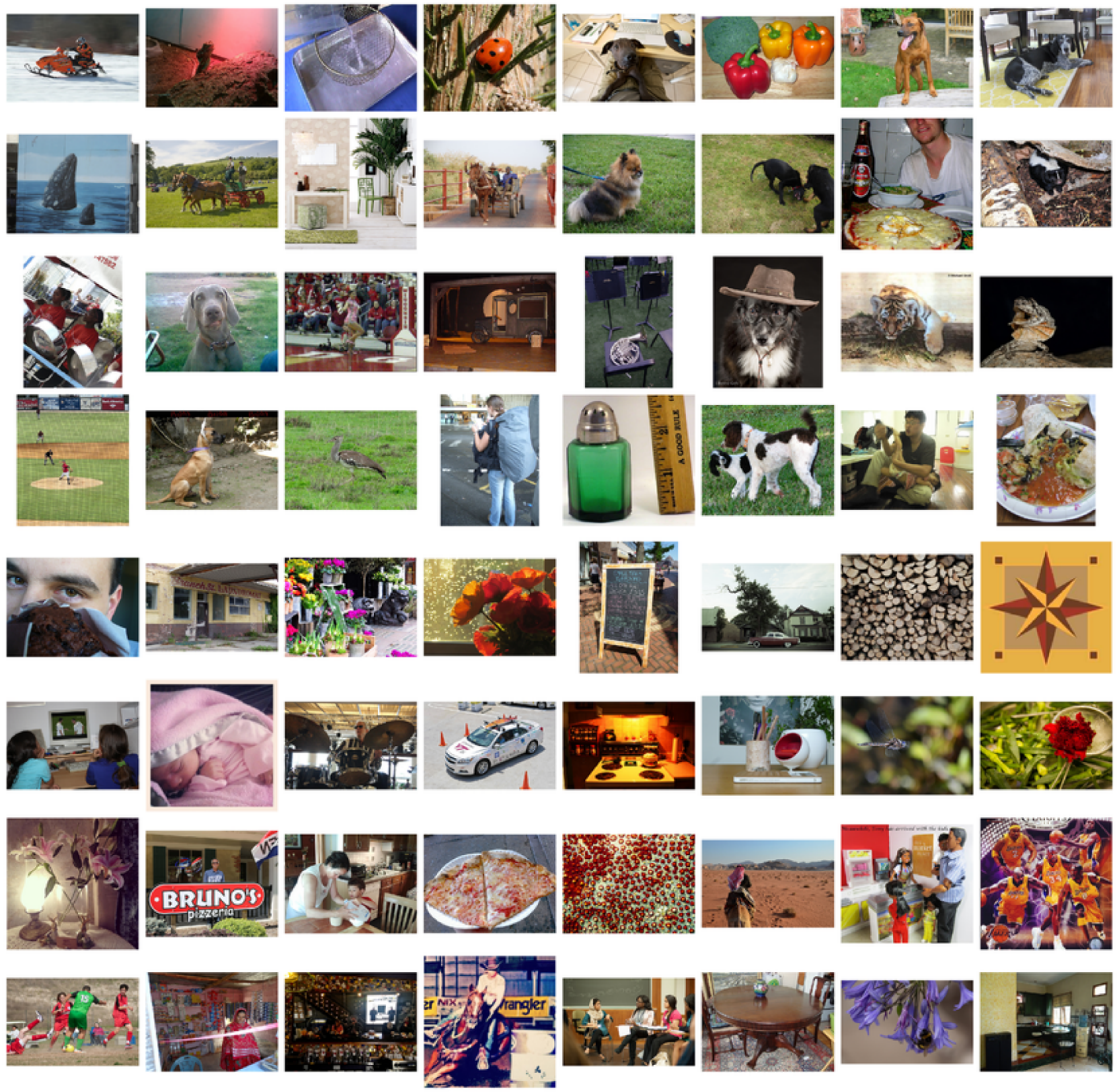
“Math whiz” Clever Hans horse

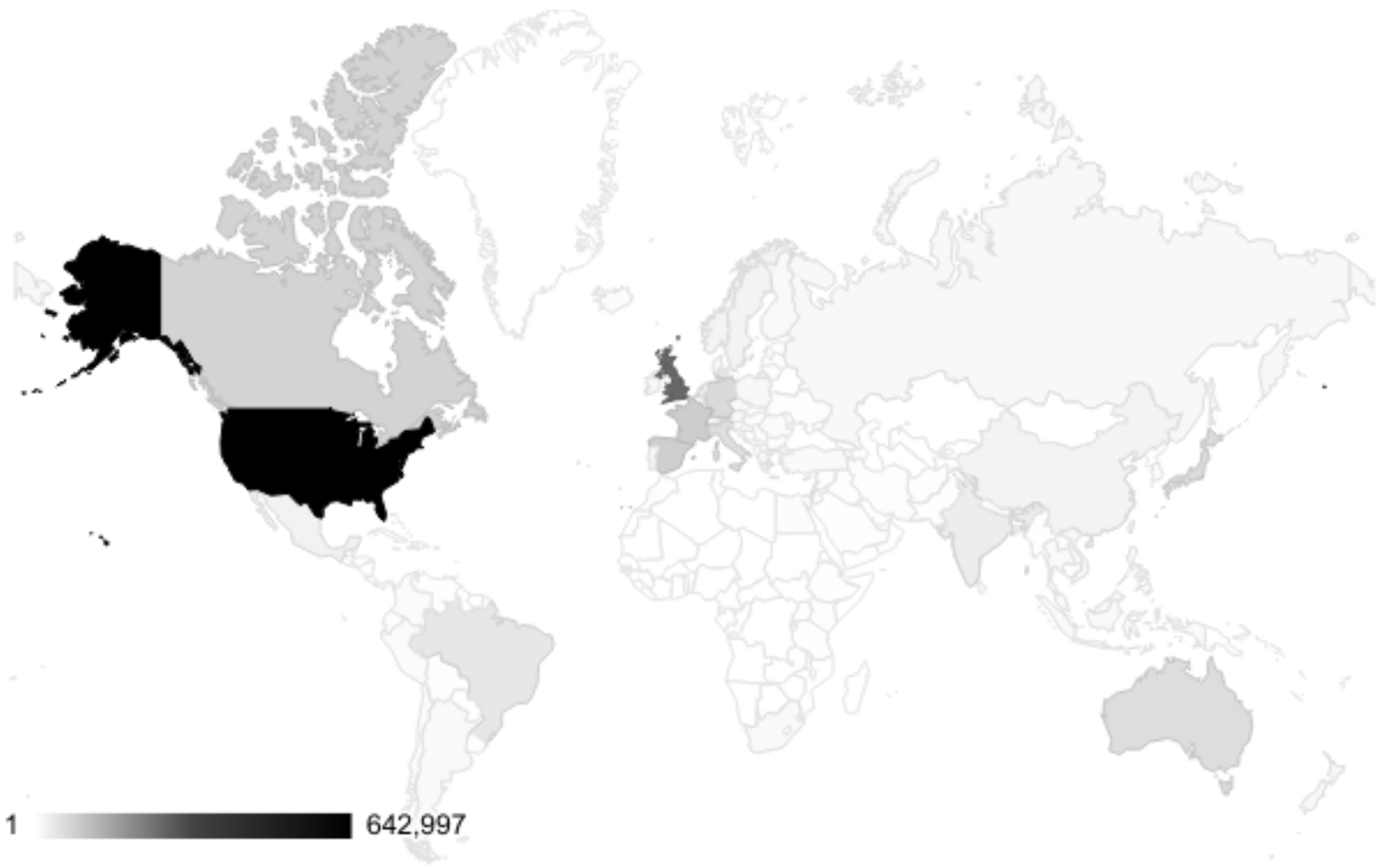
“horse”

“not horse”

PASCAL object detection dataset

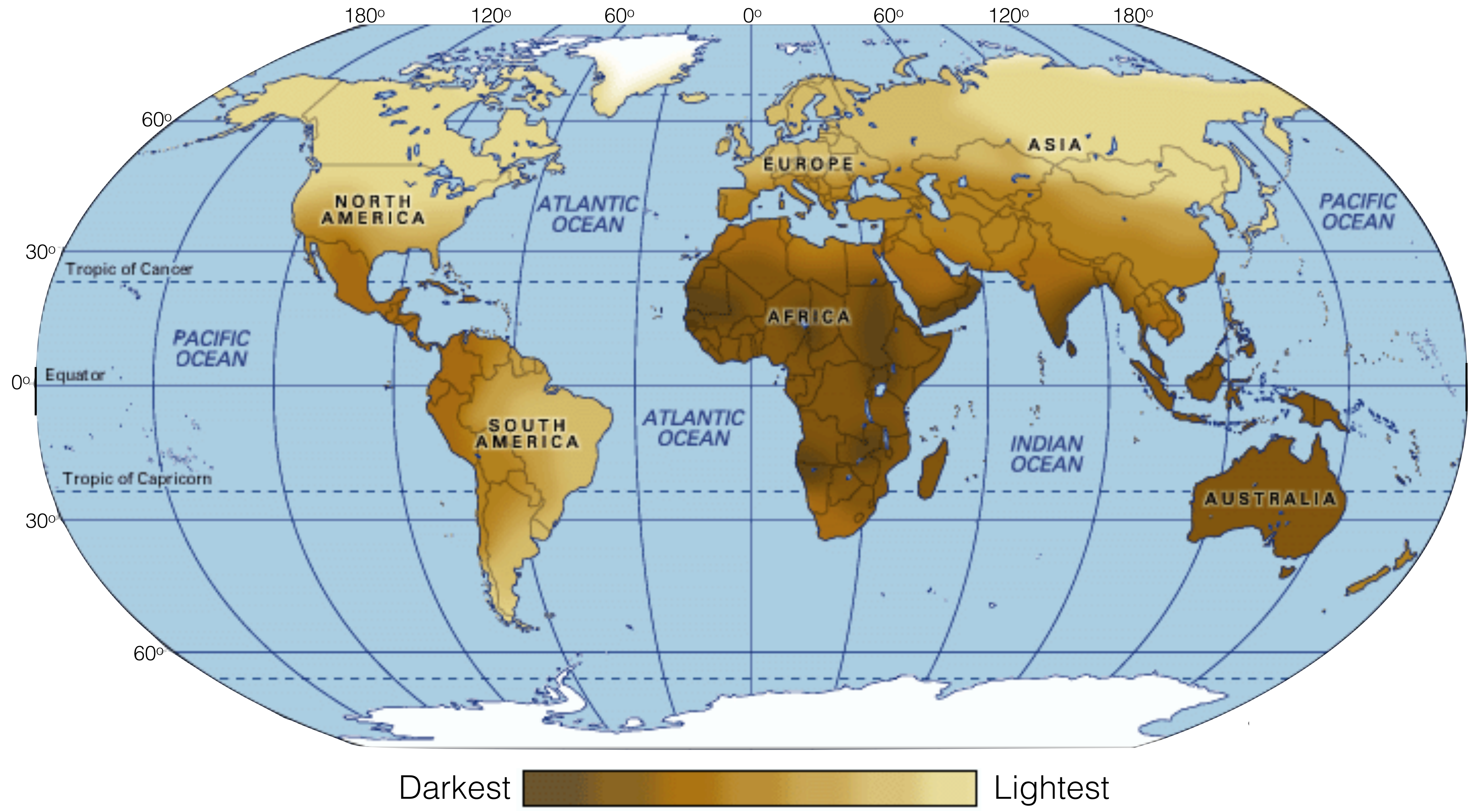
[Everingham et al., IJCV 2010; Lapuschkin et al., Nat. Commun. 2019]

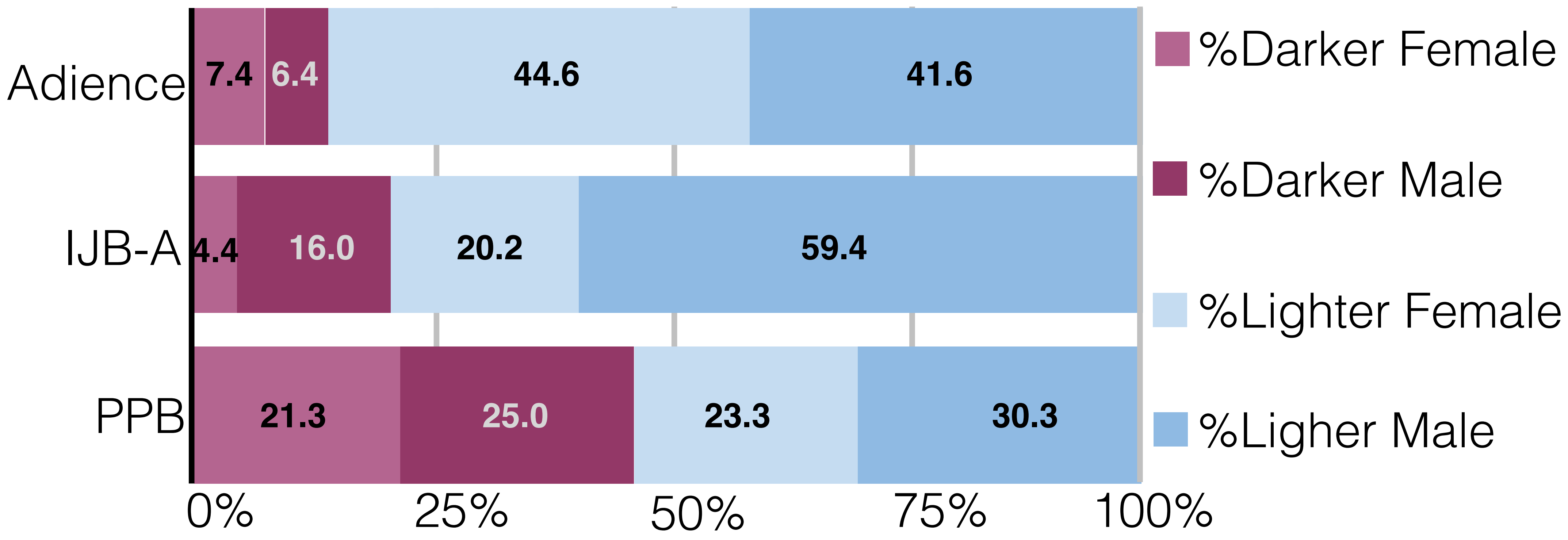




ImageNet object recognition dataset

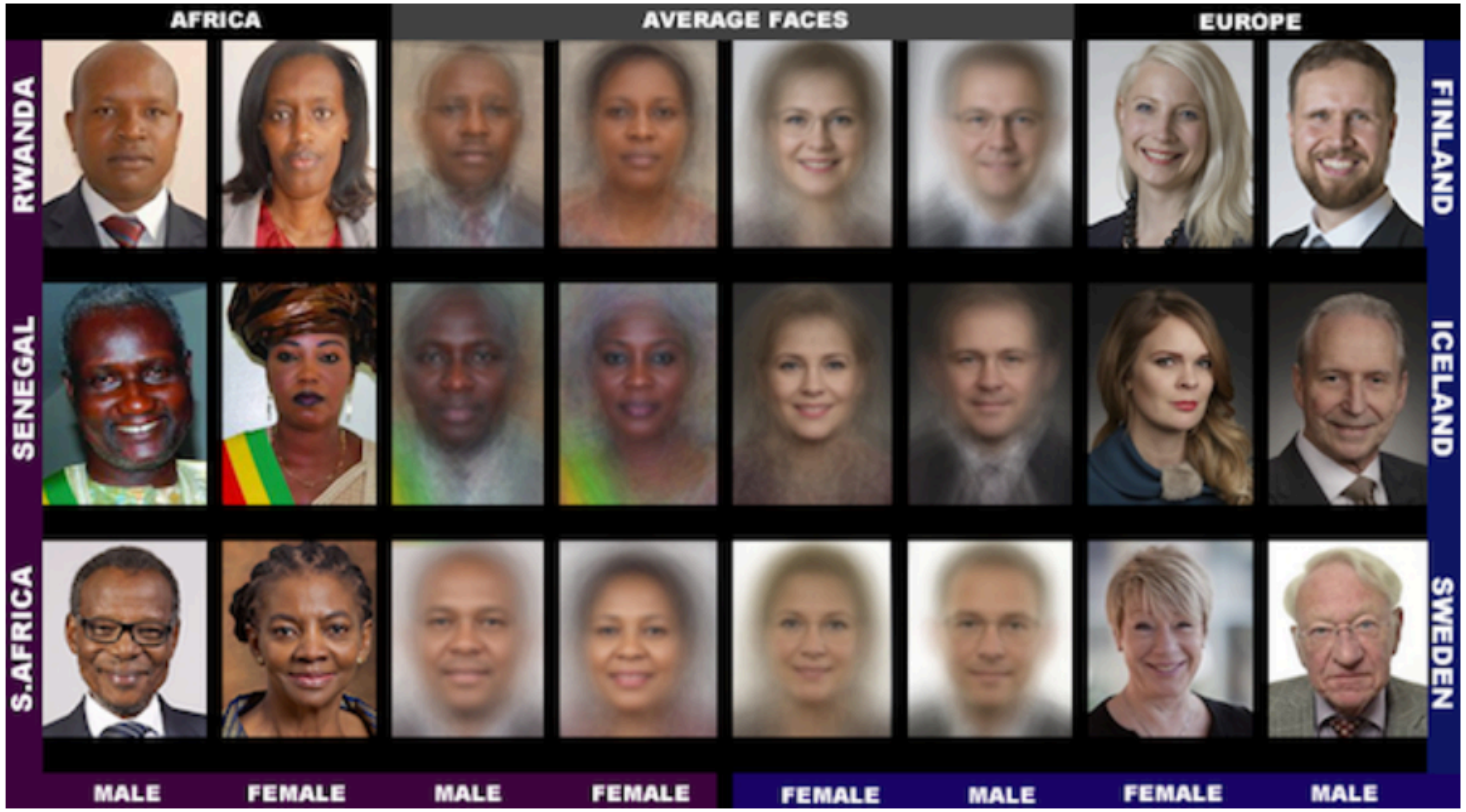
[Russakovsky et al., IJCV 2015; Shankar et al., NeurIPS Workshop 2017]





Face datasets

[Buolamwini & Gebru, JMLR 2018; globe image from Encyclopedia Britannica]

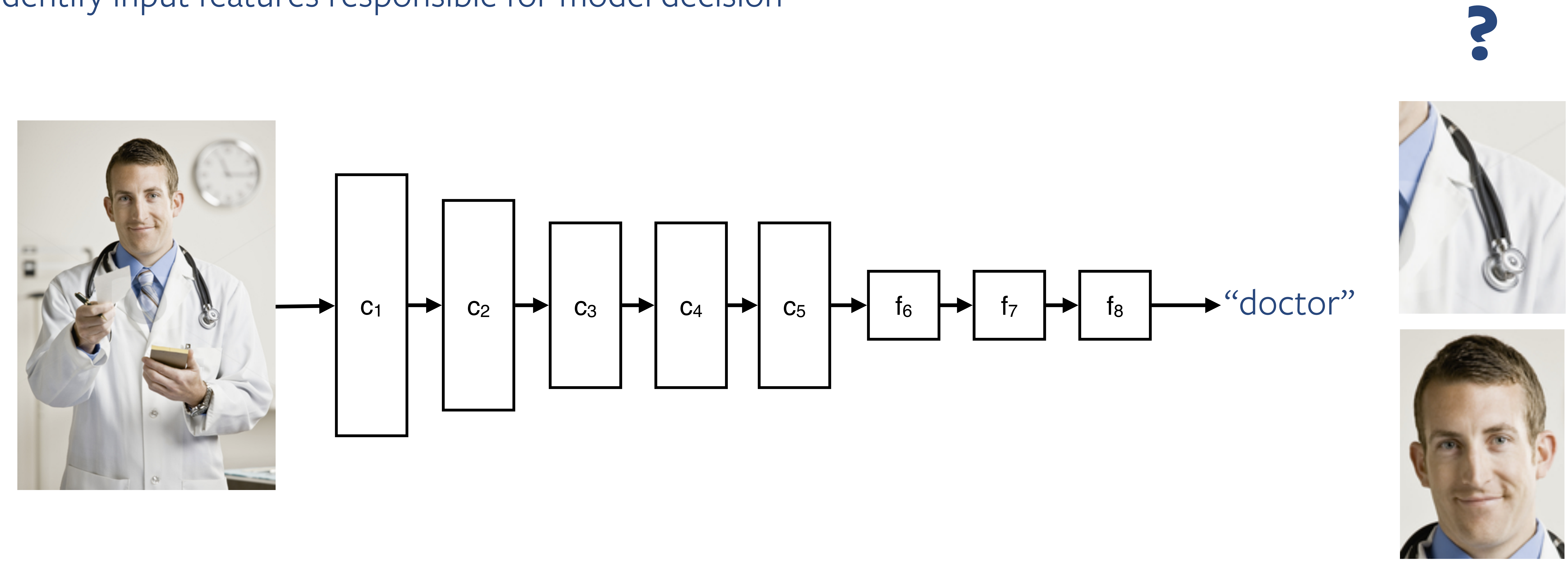


Face datasets

[Buolamwini & Gebru, JMLR 2018; globe image from Encyclopedia Britannica]

Attribution

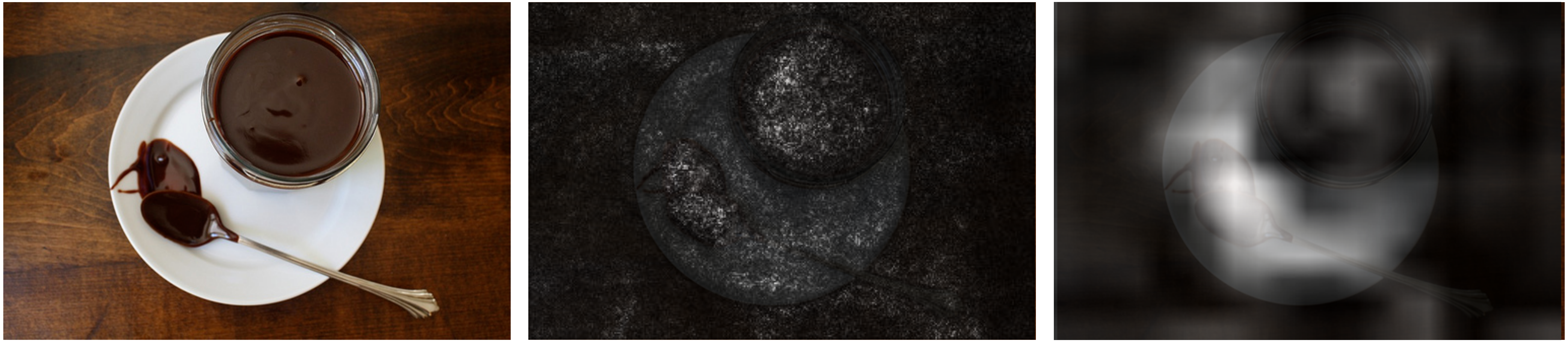
Identify input features responsible for model decision



Prior Work: Propagation-based methods

Combine network activations and gradients

Input Gradient Grad-CAM



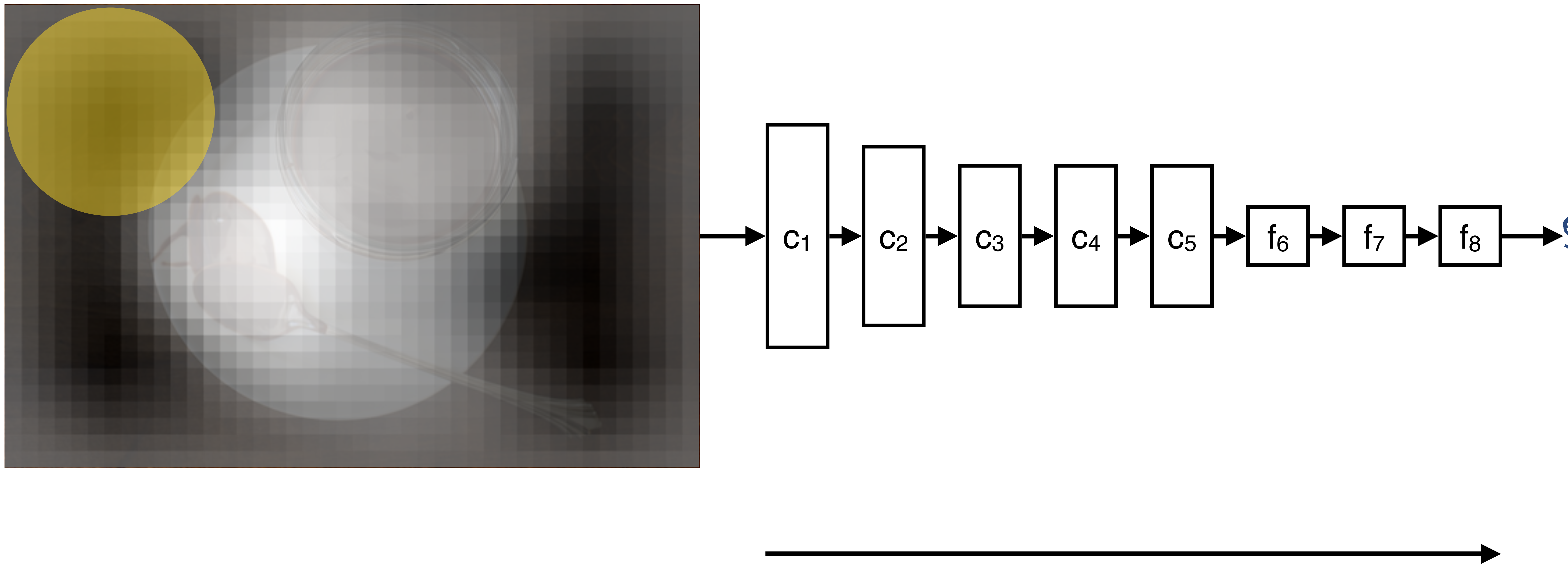
Fast, but difficult to interpret

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

Prior Work: Perturbation Approaches

Change the input and observe the effect on the output

Occlusion

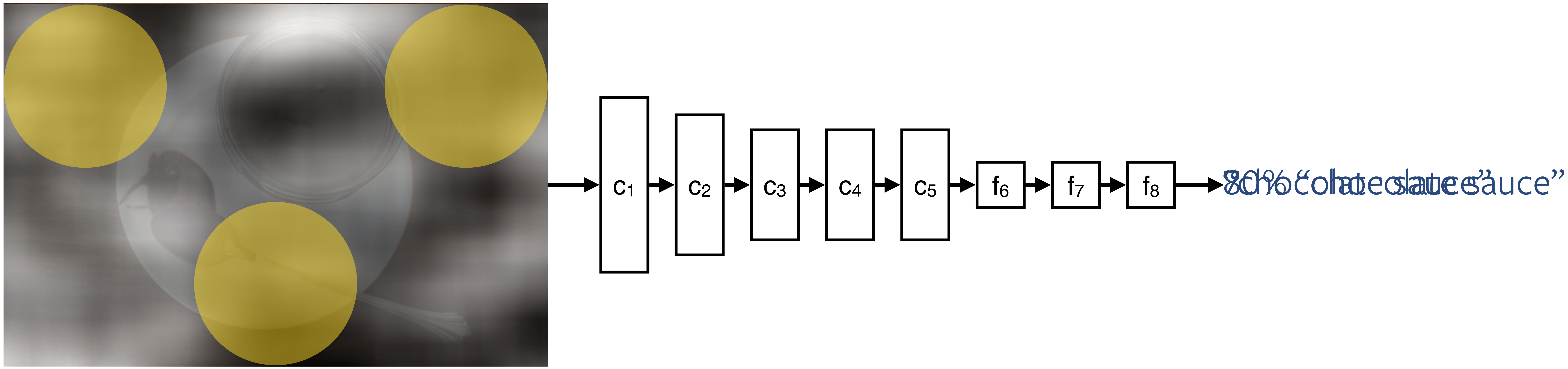


[Zeiler and Fergus, ECCV 2014]

Prior Work: Perturbation Approaches

Change the input and observe the effect on the output

RISE

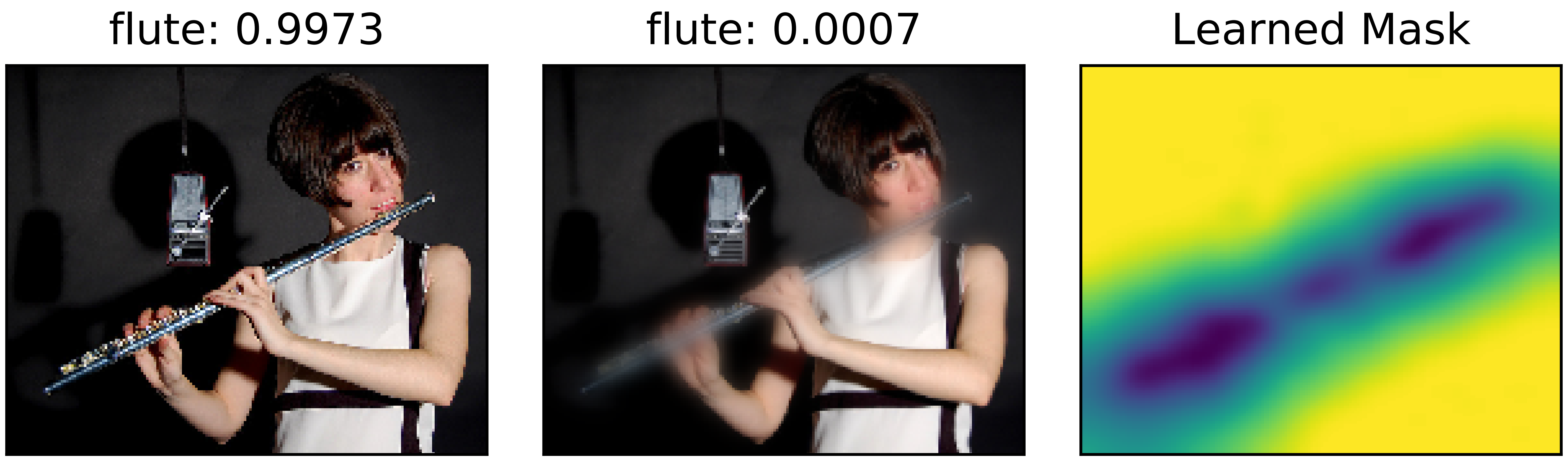


Clear meaning, but can only test a small range of occlusions

[Petsiuk et al., BMVC 2018]

Desired Approach

Our Approach: Meaningful Perturbations

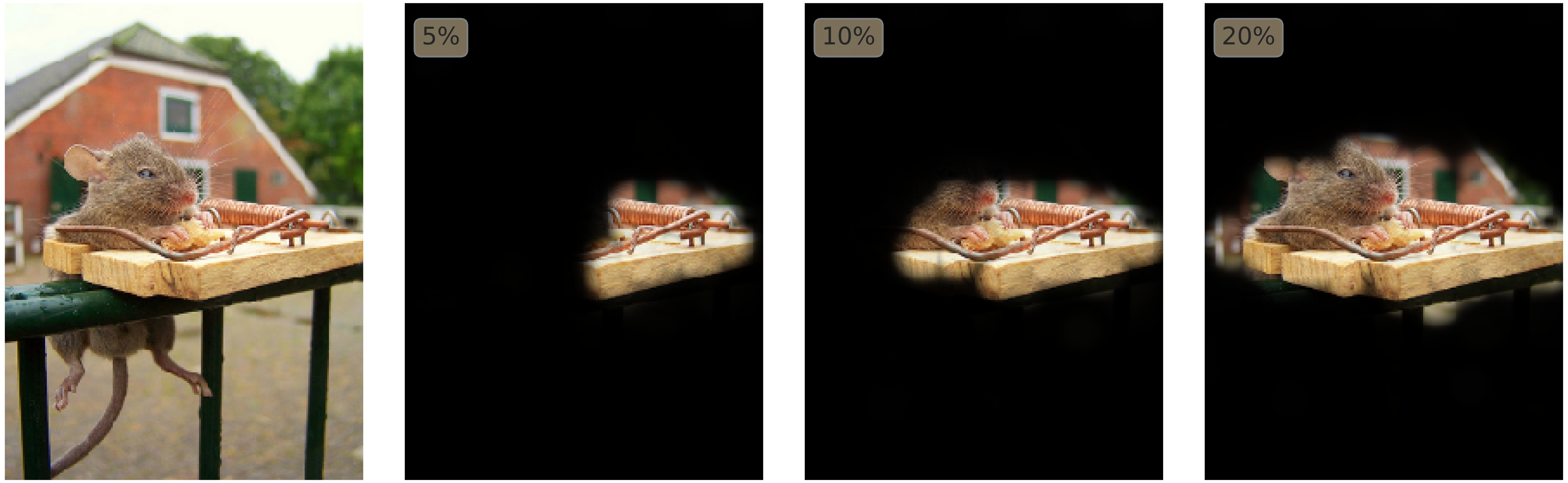


Learn a **minimal** mask **m** to perturb input **x** that maximally affects the network's output

Our method considers a wide range of occlusion sizes and shapes.

[Fong & Vedaldi, ICCV 2017]

Our Approach: Extremal Perturbations



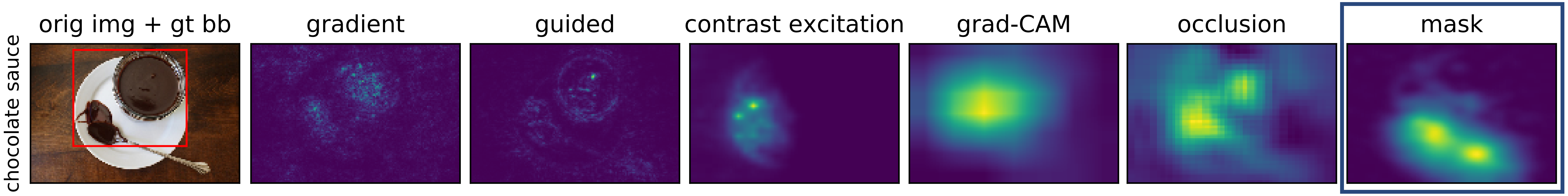
Learn a **fixed-sized** mask **m** to perturb input **x** that maximally **preserves** the network's output

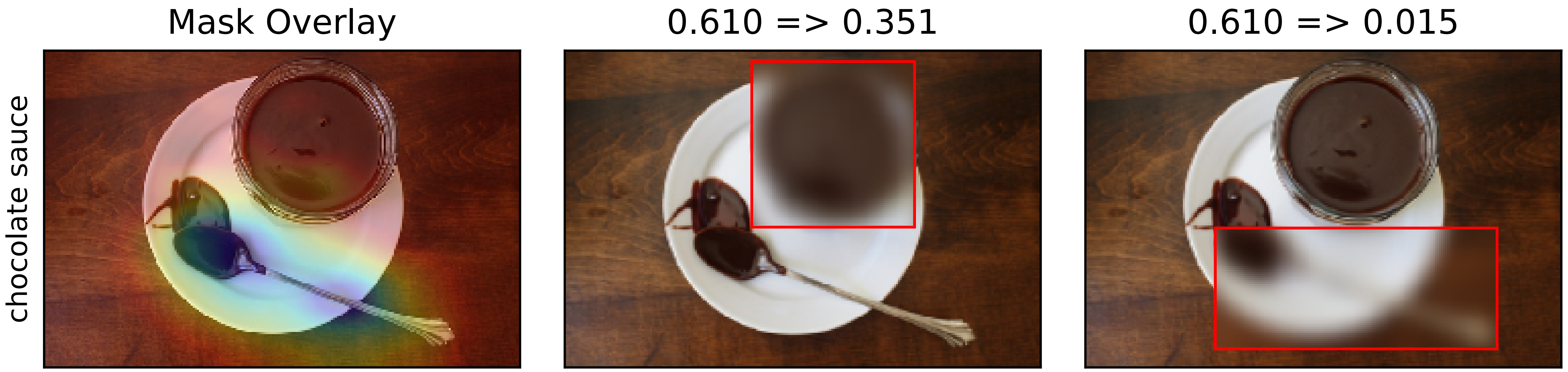
[Fong et al., ICCV 2019]

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

Results

Interpretability

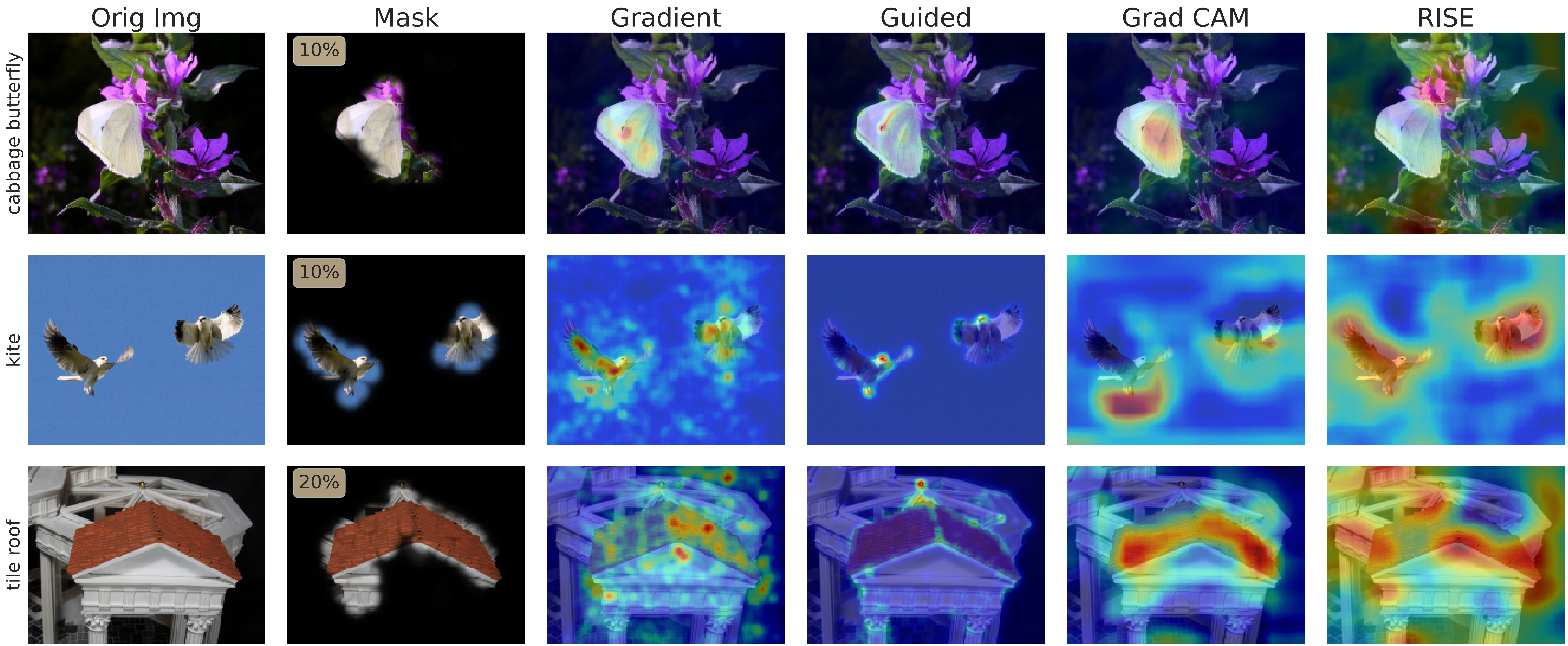




An explanation should be **falsifiable**.

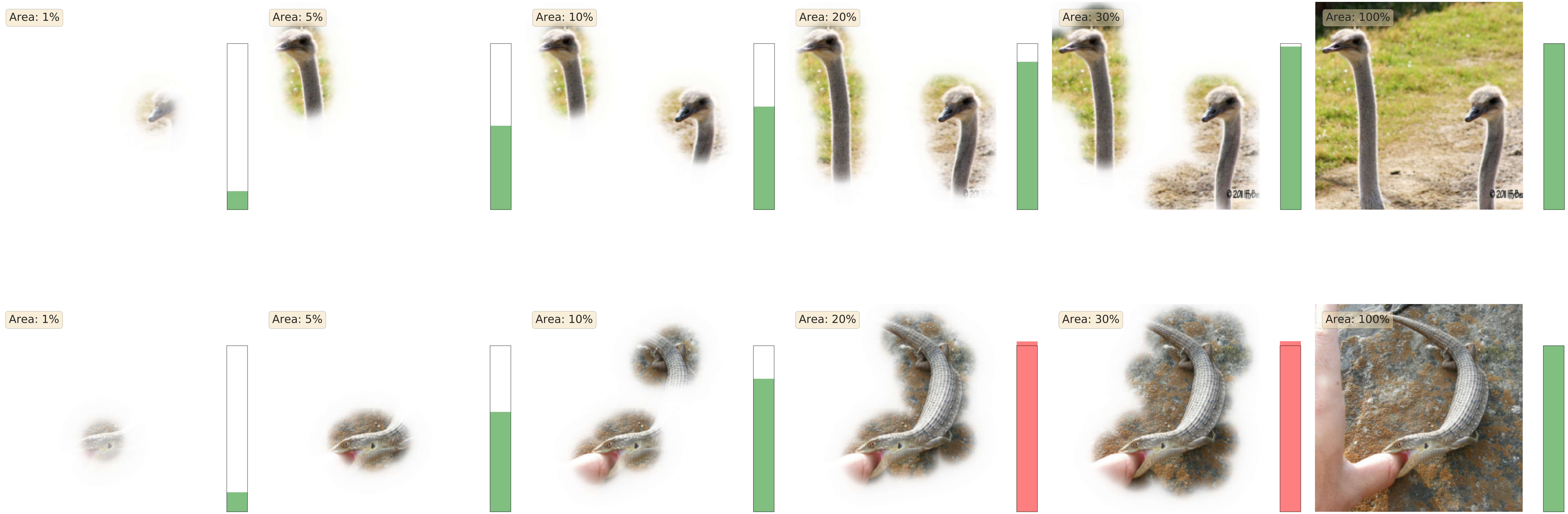
[Fong & Vedaldi, ICCV 2017]

Comparison



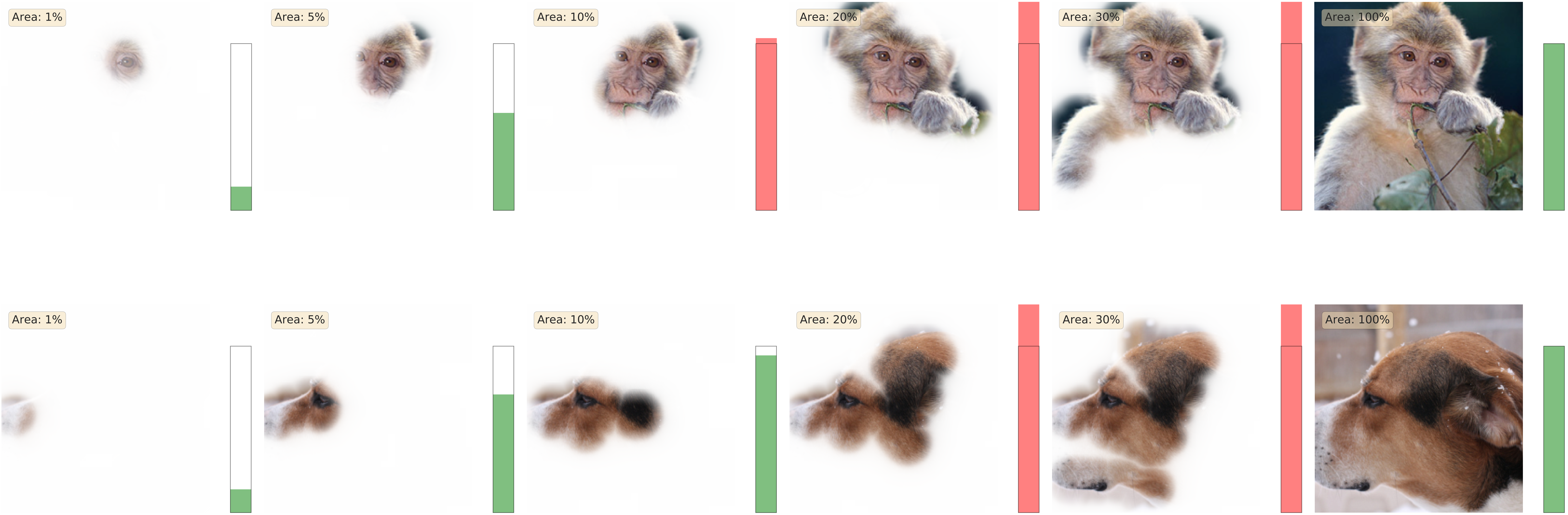
[Fong et al., ICCV 2019]

Foreground evidence is usually sufficient



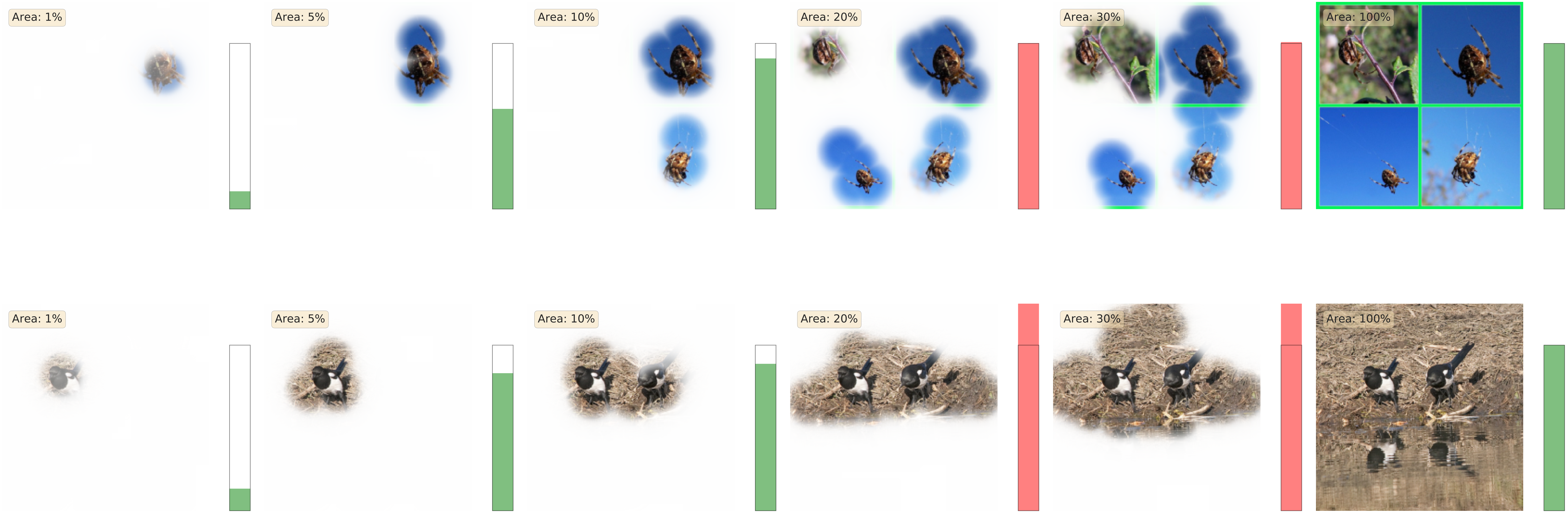
[Fong et al., ICCV 2019]

Large objects are recognized by their details



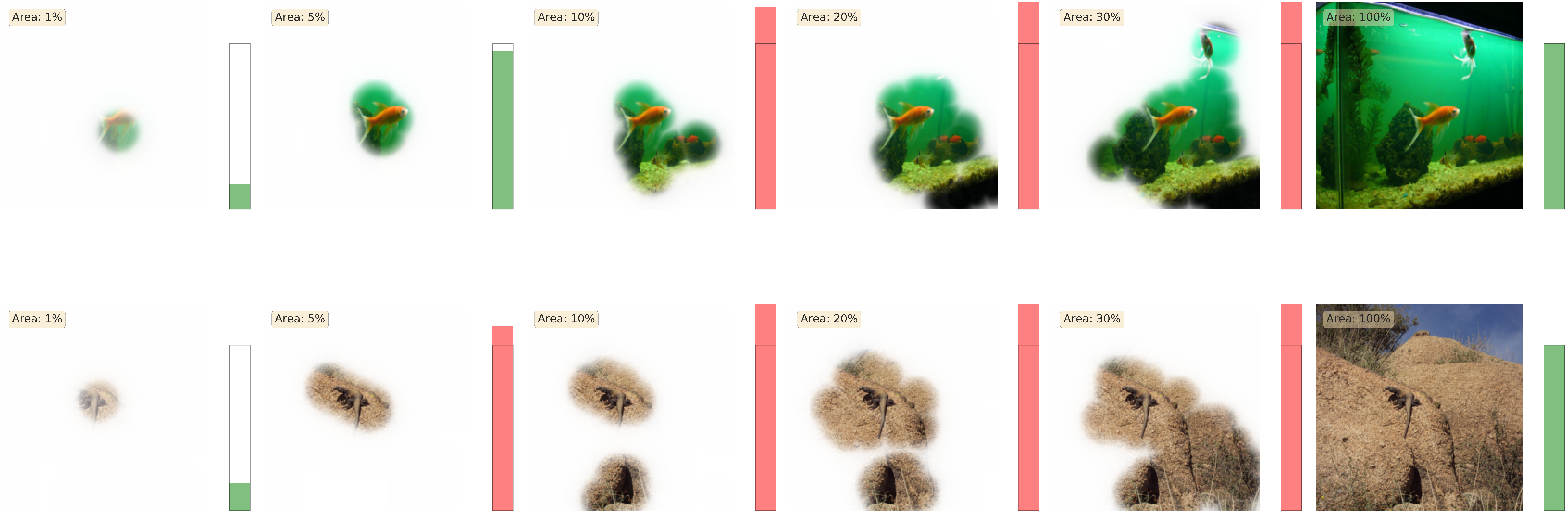
[Fong et al., ICCV 2019]

Multiple objects contribute cumulatively



[Fong et al., ICCV 2019]

Suppressing the background may overdrive the network

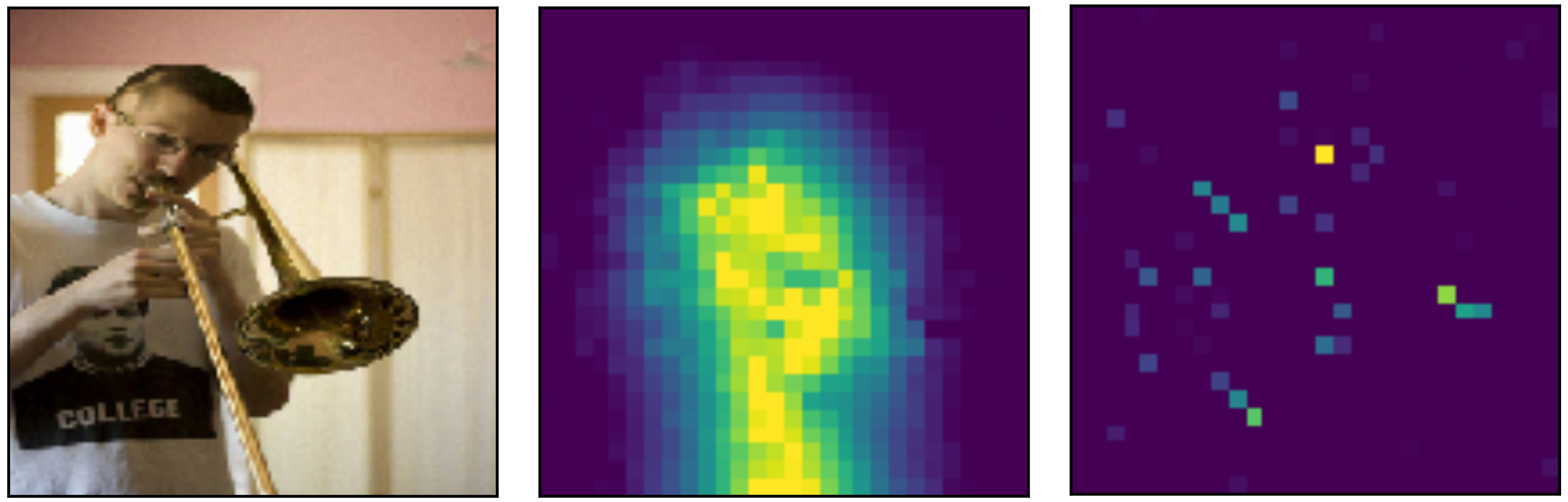


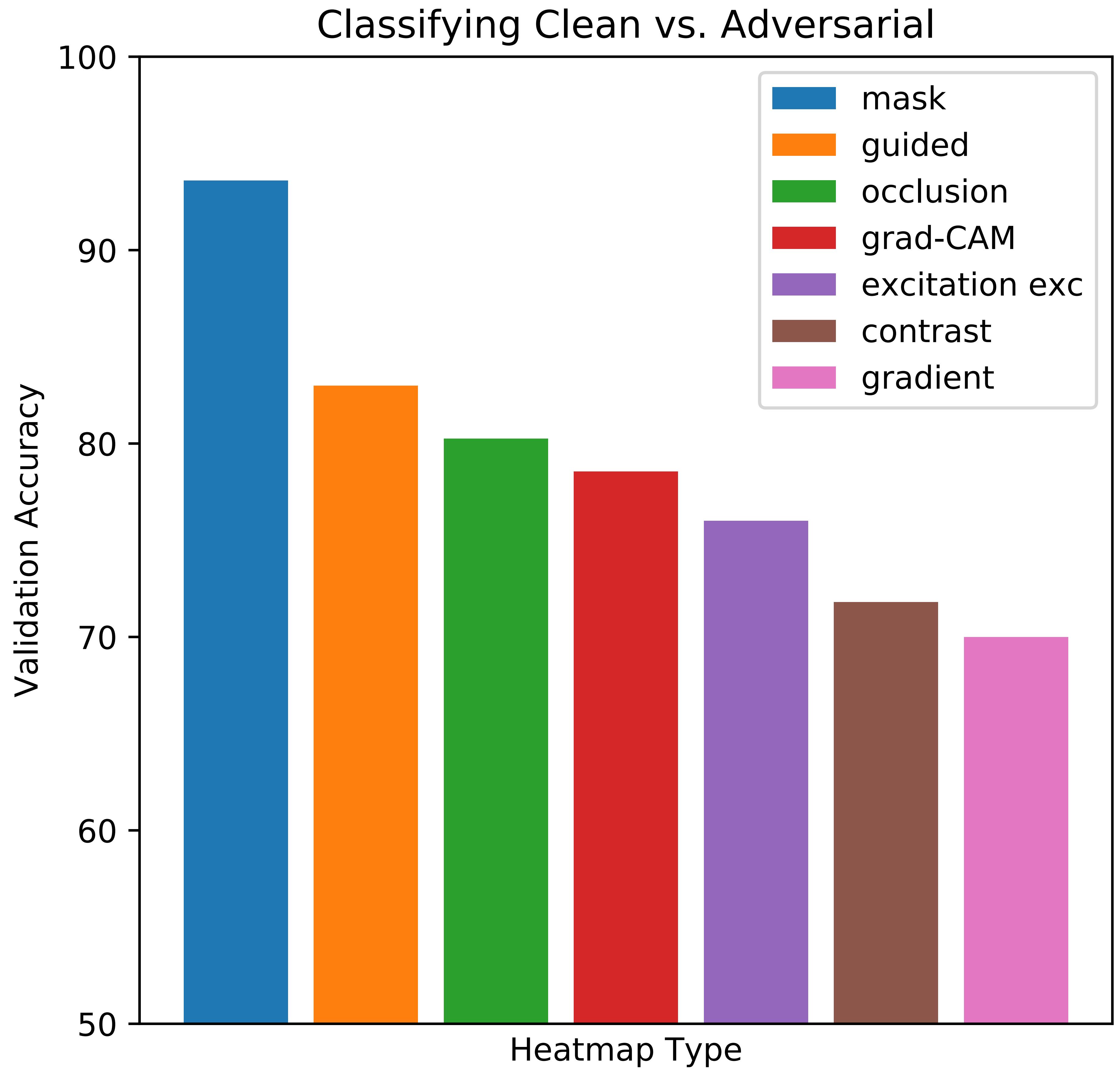
[Fong et al., ICCV 2019]

Adversarial Defense

Mask on Clean Image

Mask on Adv. Image





Our method allows us to defend **any model** against adversarial attacks.

[Fong & Vedaldi, ICCV 2017]

Details

Regularization to mitigate artifacts

network

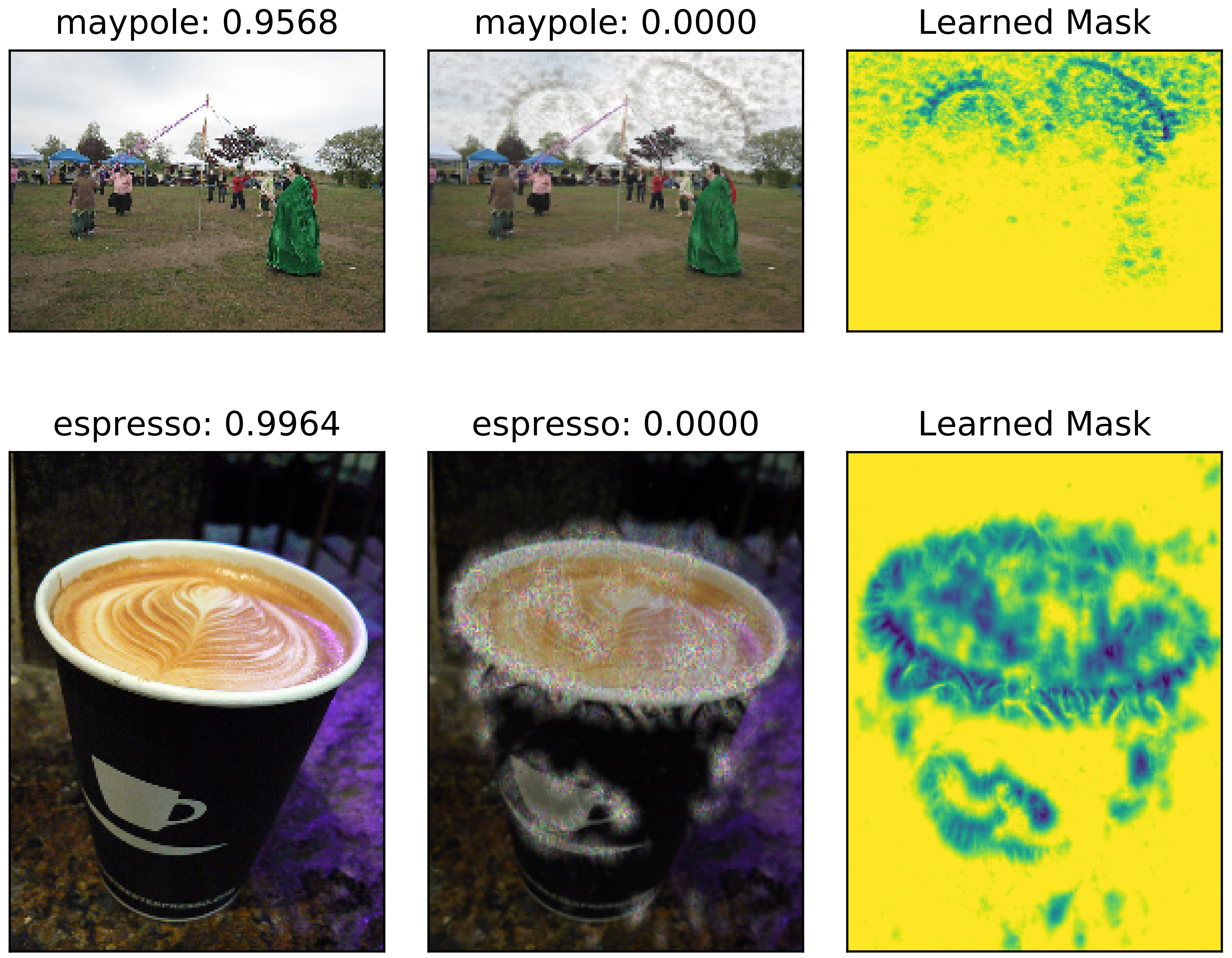
mask

image

v1: $\mathbf{m}^*(\lambda) = \operatorname{argmin}_{\mathbf{m}} \Phi(\mathbf{m} \otimes \mathbf{x}) + \lambda \, \text{area}(\mathbf{m})$

v2: $\mathbf{m}^*(\lambda_1, \lambda_2) = \operatorname{argmin}_{\mathbf{m}} \mathbb{E}_{\text{jitter}}[\Phi(M_{\text{upsample}}(\mathbf{m}) \otimes \mathbf{x})]$

$+\lambda_1 \, \text{area}(\mathbf{m}) + \lambda_2 \, \text{smooth}(\mathbf{m})$



Tradeoff between **attribution objective** and **regularization**

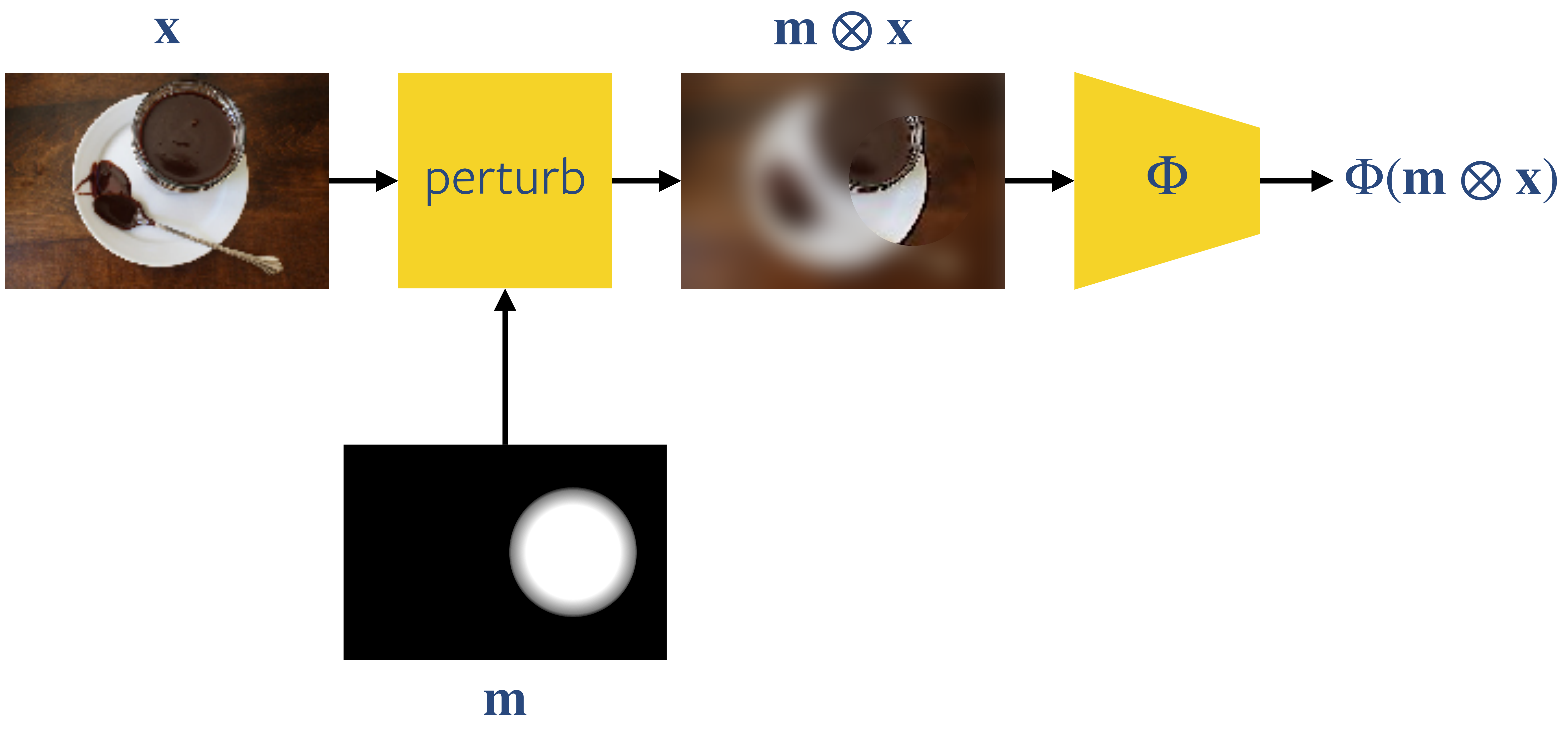
[Fong & Vedaldi, ICCV 2017]

Extremal Perturbations

A mask is optimized to maximally excite the network:

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

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



[Fong et al., ICCV 2019]

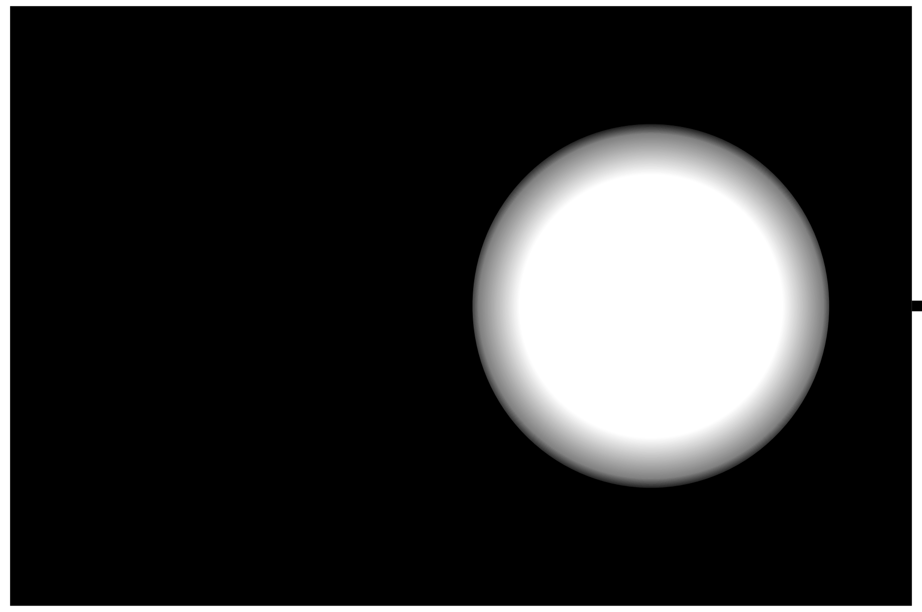
Area Constraint

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

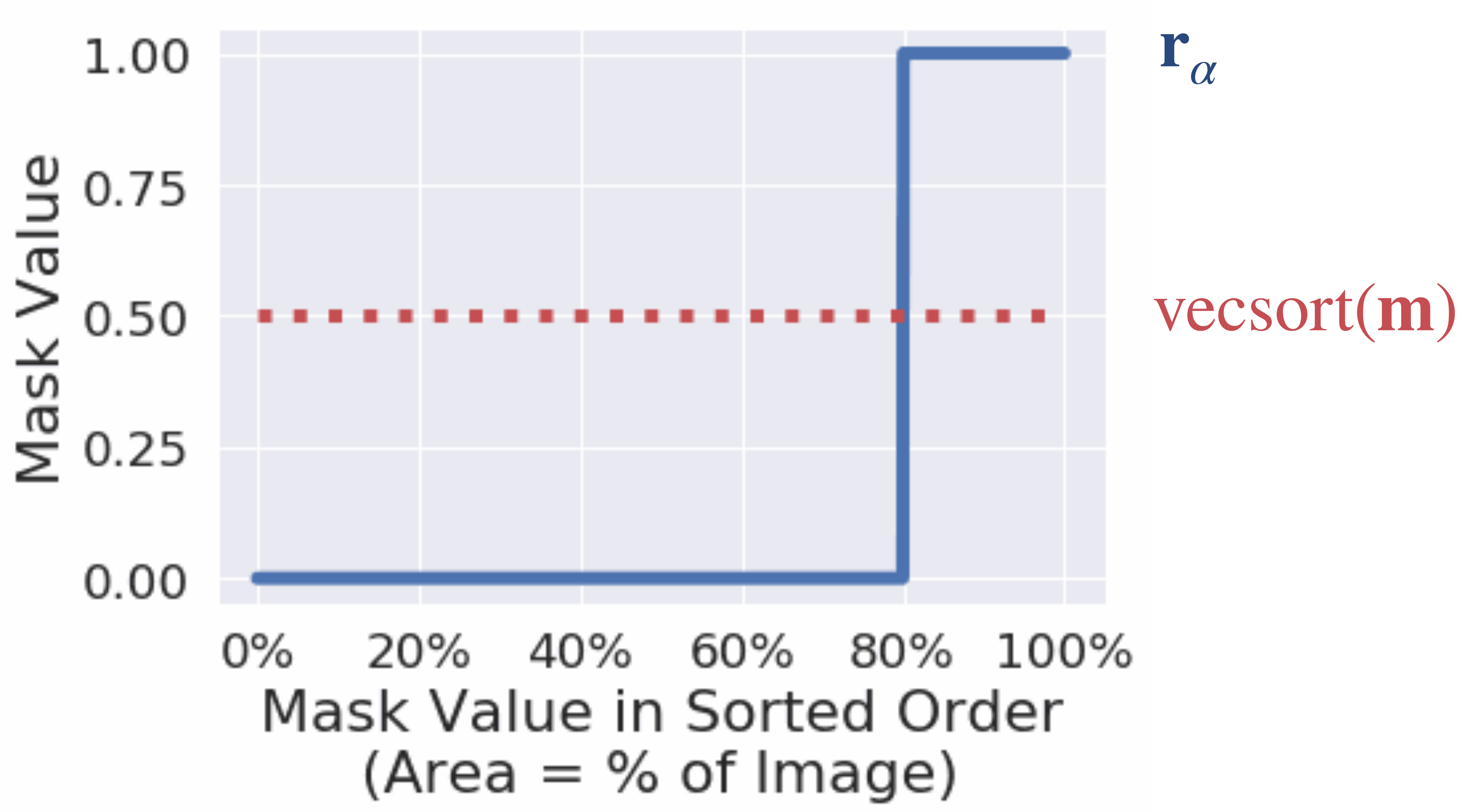
Optimizing w.r.t. to an area constraint is challenging

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

$\mathbf{m}$



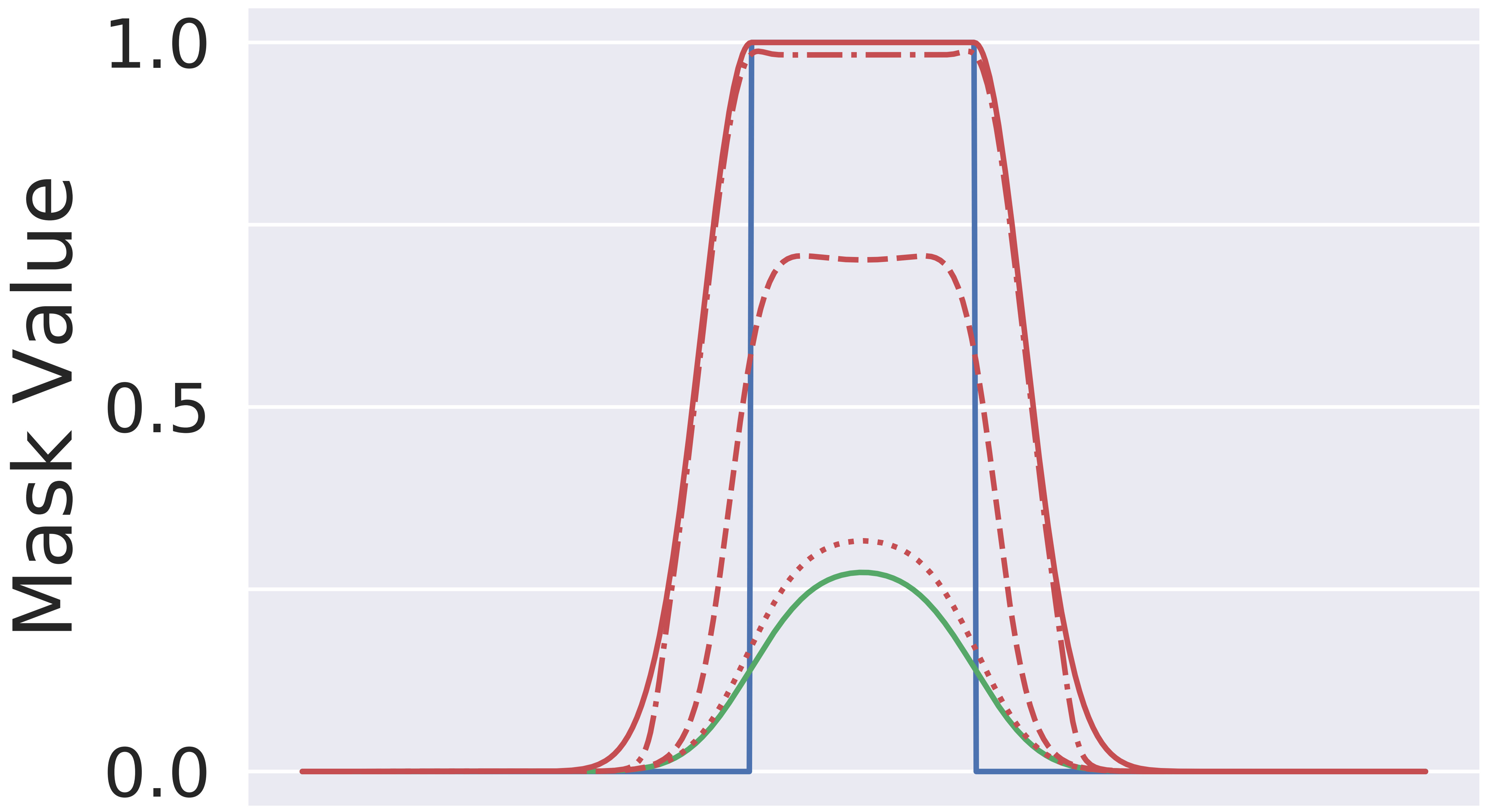
vectorize → sort →



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

[Fong et al., ICCV 2019]

Smooth Masks



$m(v)$: mask

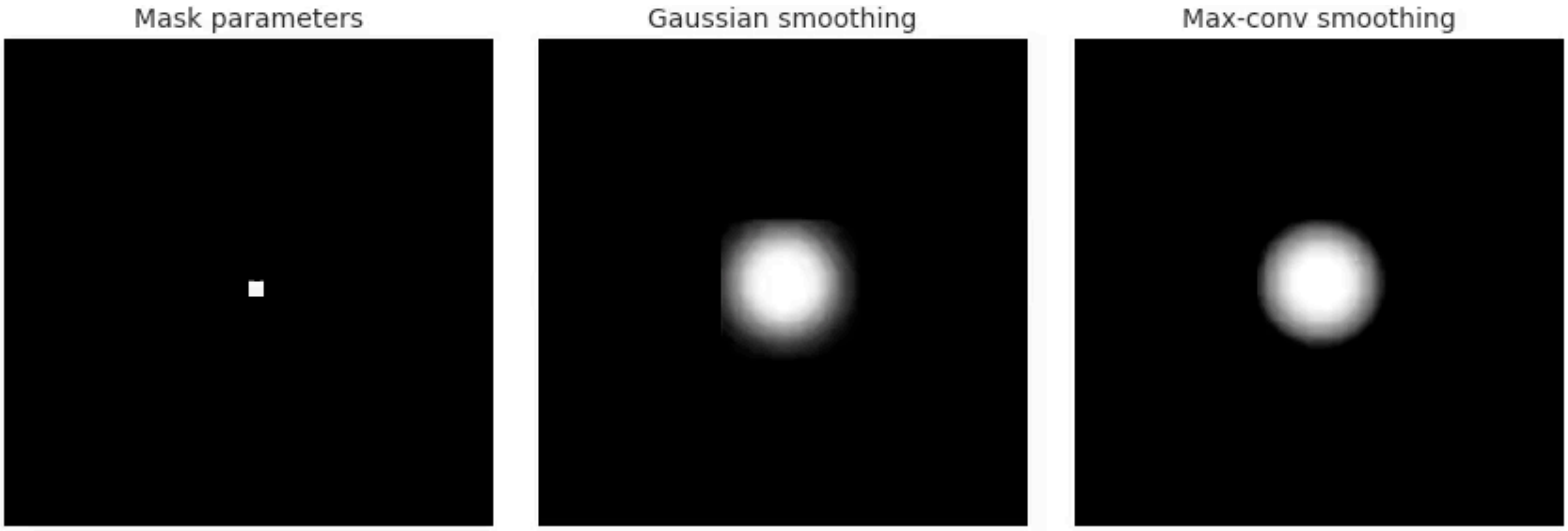
$$\mathrm{conv}(u; m; k) = \frac{1}{Z} \sum_{v \in \Omega} k(u - v) m(v)$$

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

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

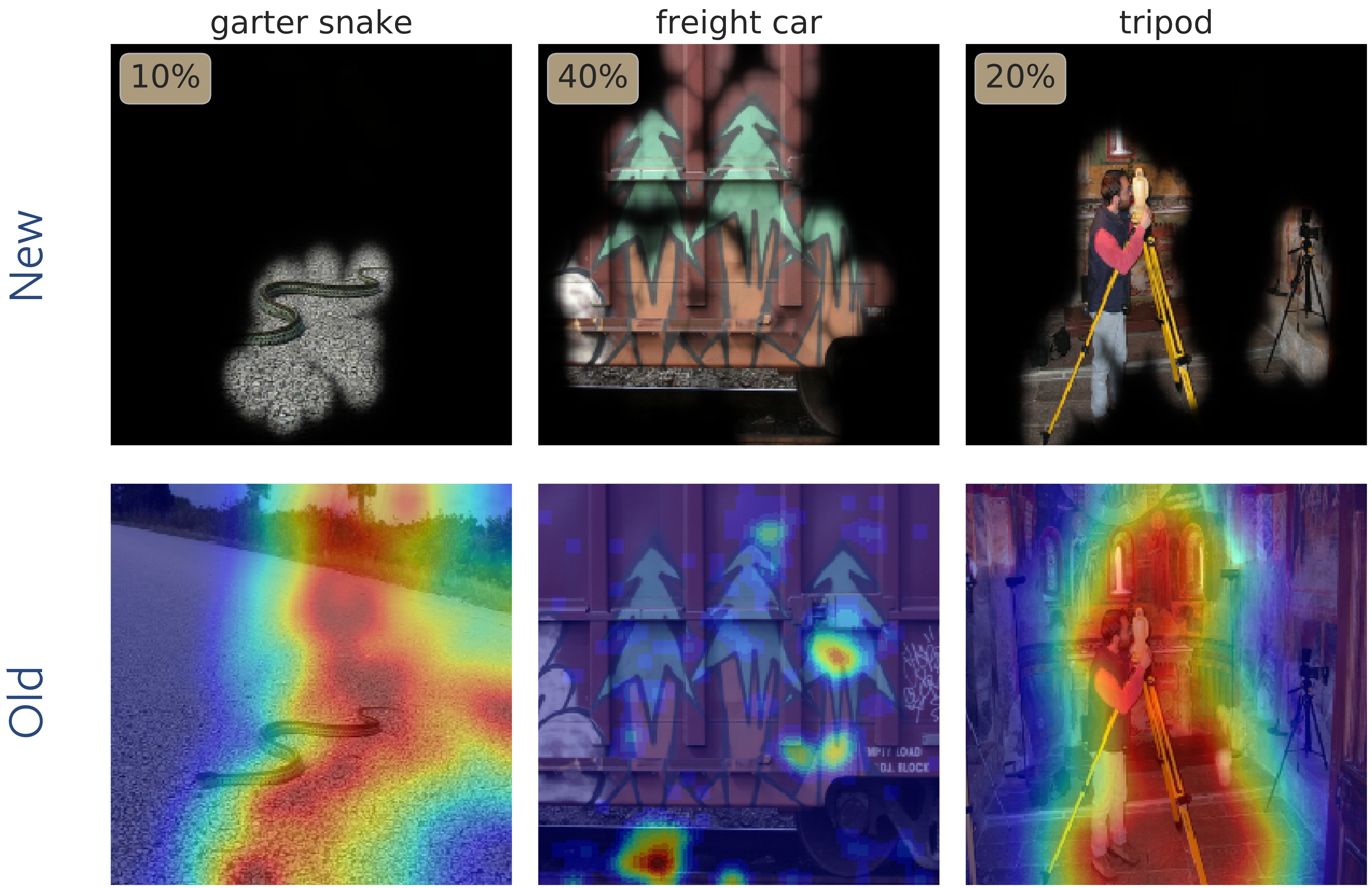
[Fong et al., ICCV 2019]

Smooth Masks



[Fong et al., ICCV 2019]

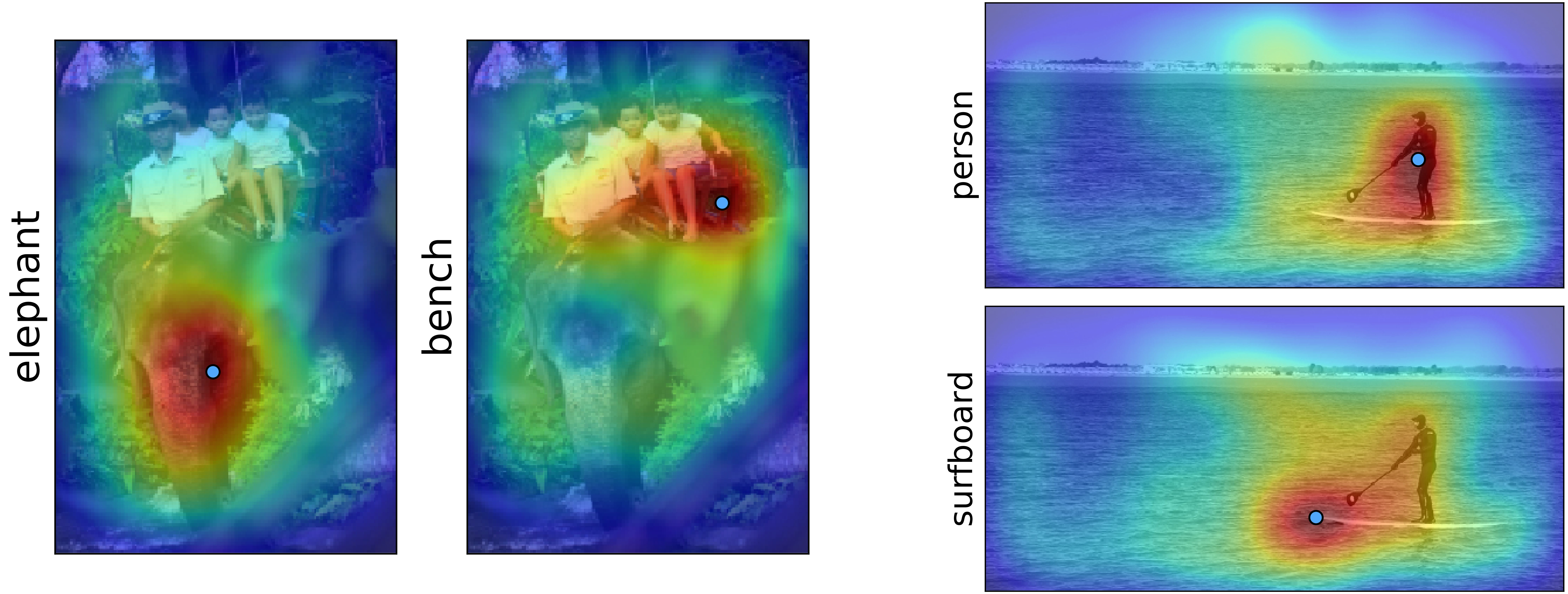
Comparison with Prior Work



[Fong & Vedaldi, 2017; Fong et al., ICCV 2019]

Evaluating and using attribution heatmaps

Measure Performance on Weak Localization

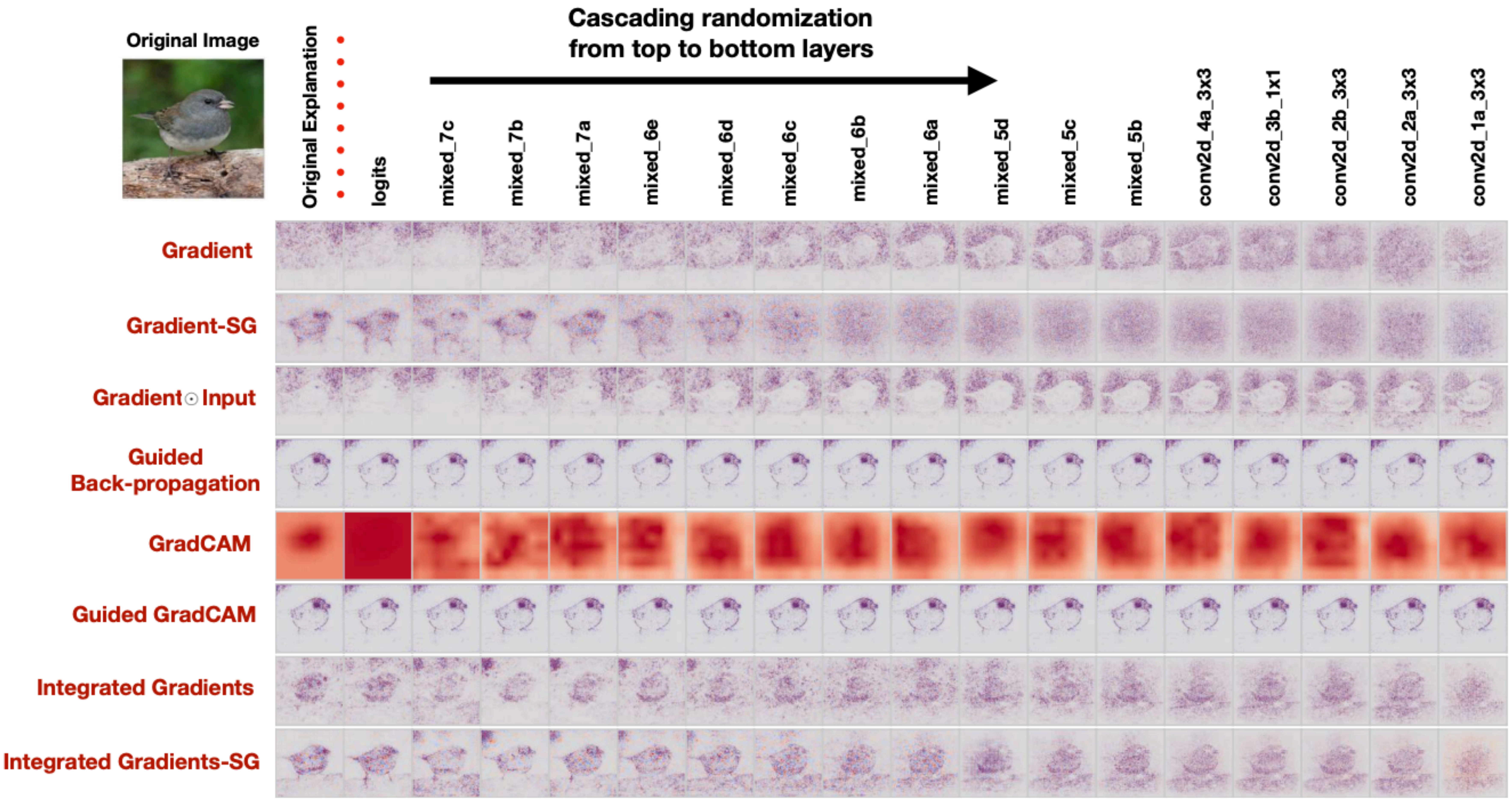


[Zhang et al., ECCV 2016]

Selectivity to Output Class

[Mahendran & Vedaldi, ECCV 2016; Rebuffi et al., CVPR 2020]

Sensitive to Model Parameters



[Adebayo et al., NeurIPS 2018]

Research Development: Critically design and evaluate attribution methods

General Usage: Assume a model has failures and use attribution methods to understand them

[Kindermans et al., arXiv 2017; Hooker et al., NeurIPS 2019; Yang & Kim, arXiv 2019]

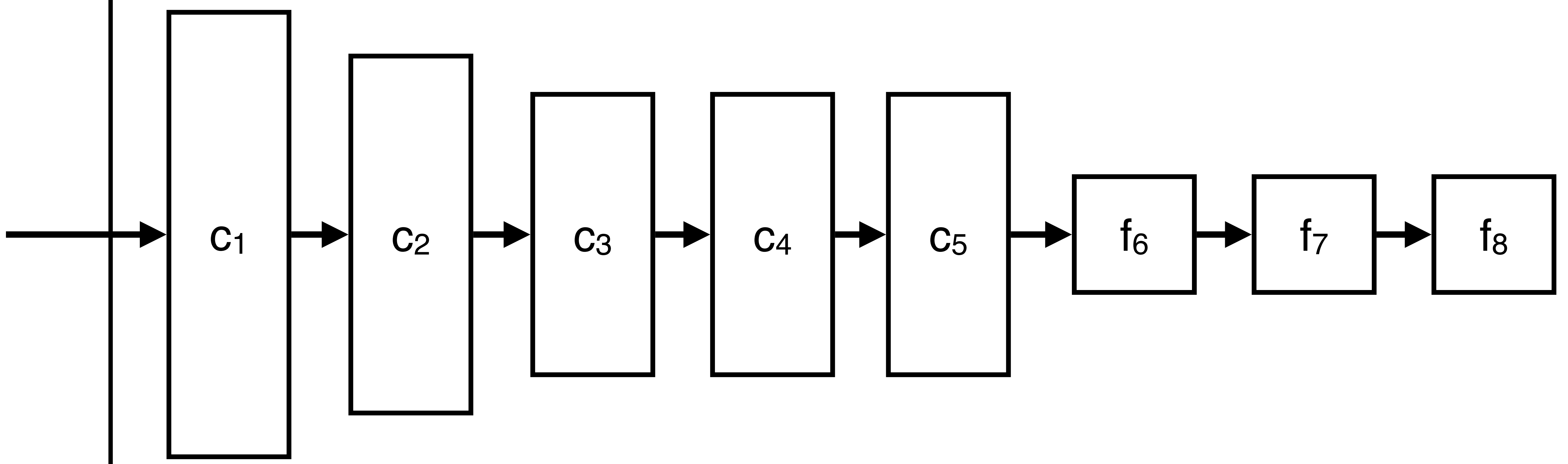
TorchRay: PyTorch interpretability library

github.com/facebookresearch/torchray

PyTorch

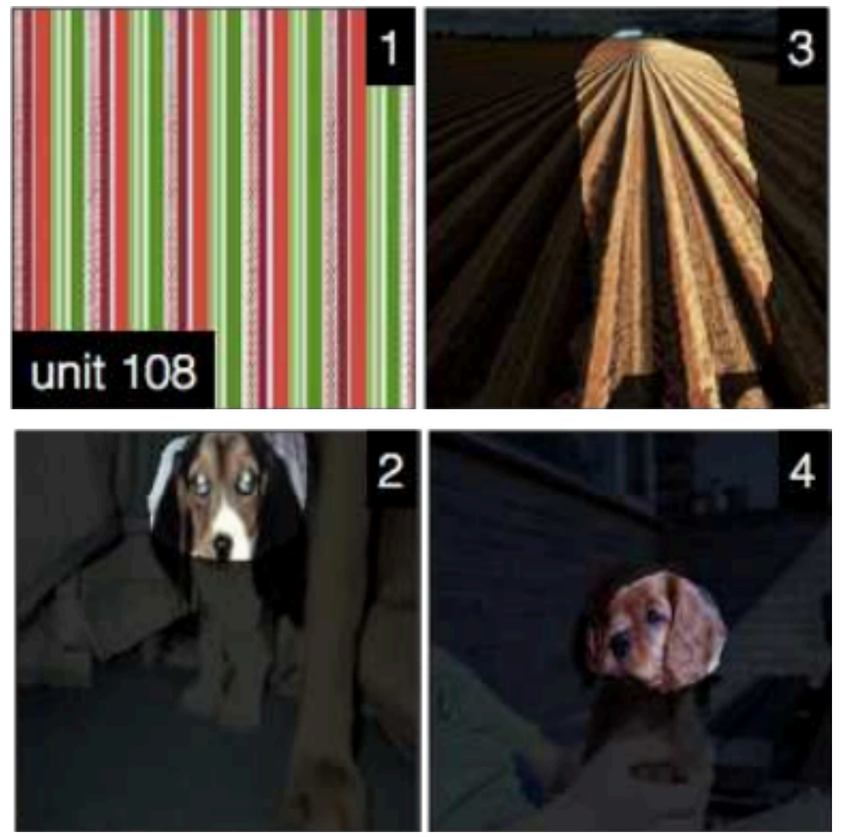
Research Themes

Internal Representation:
What & how does f(x) encode?

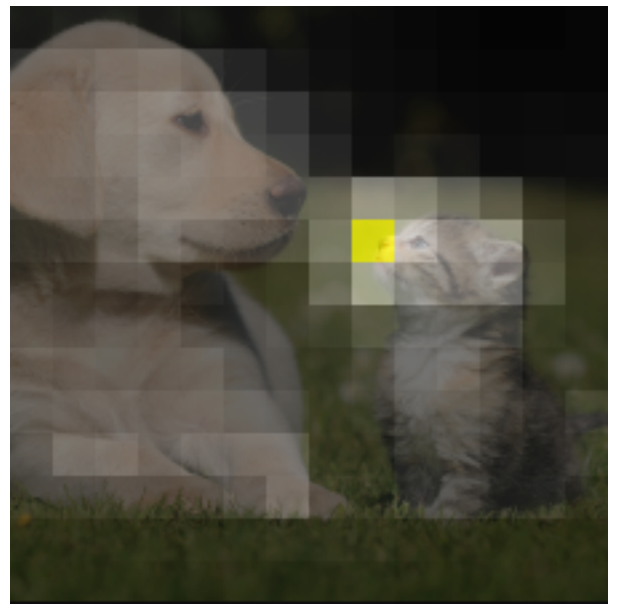


$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Fong et al., ICCV 2019

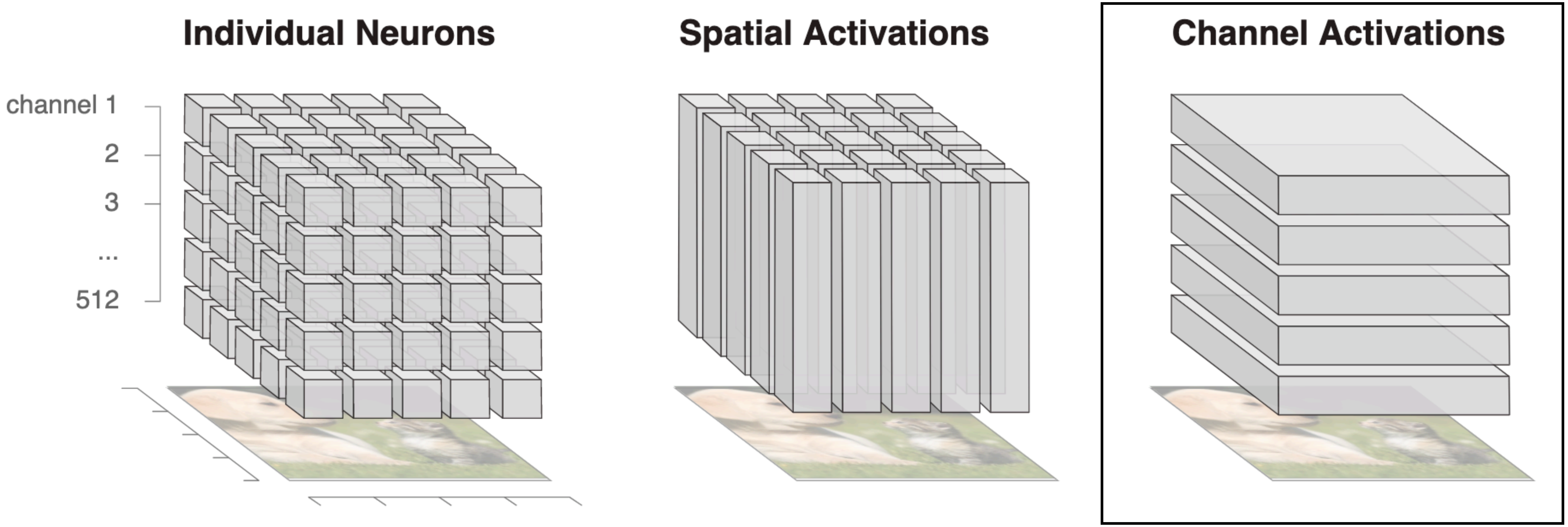


Fong & Vedaldi, CVPR 2018



Fong et al., 2020 (in prep.)

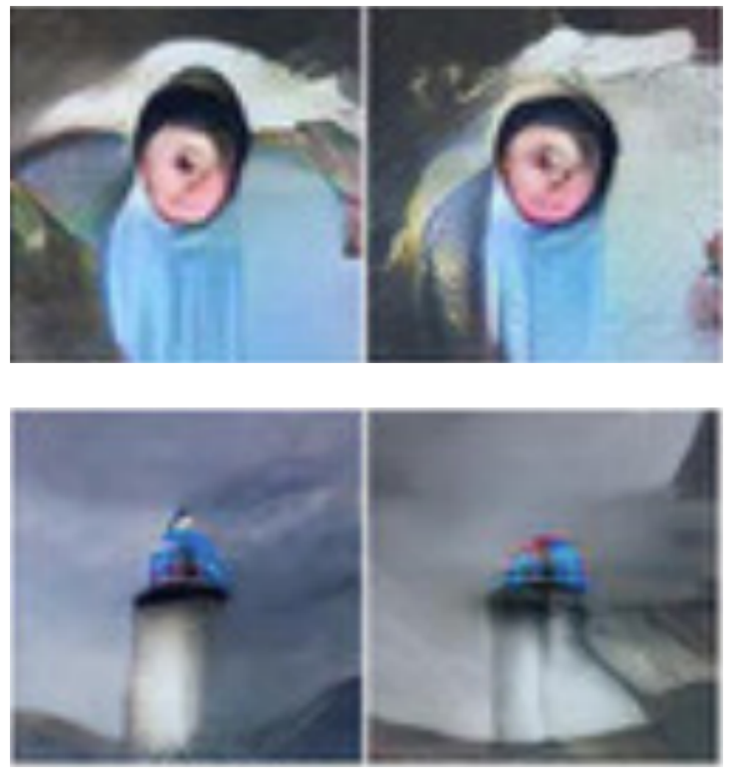
Intermediate Activations



[Olah et al., Distill 2018]

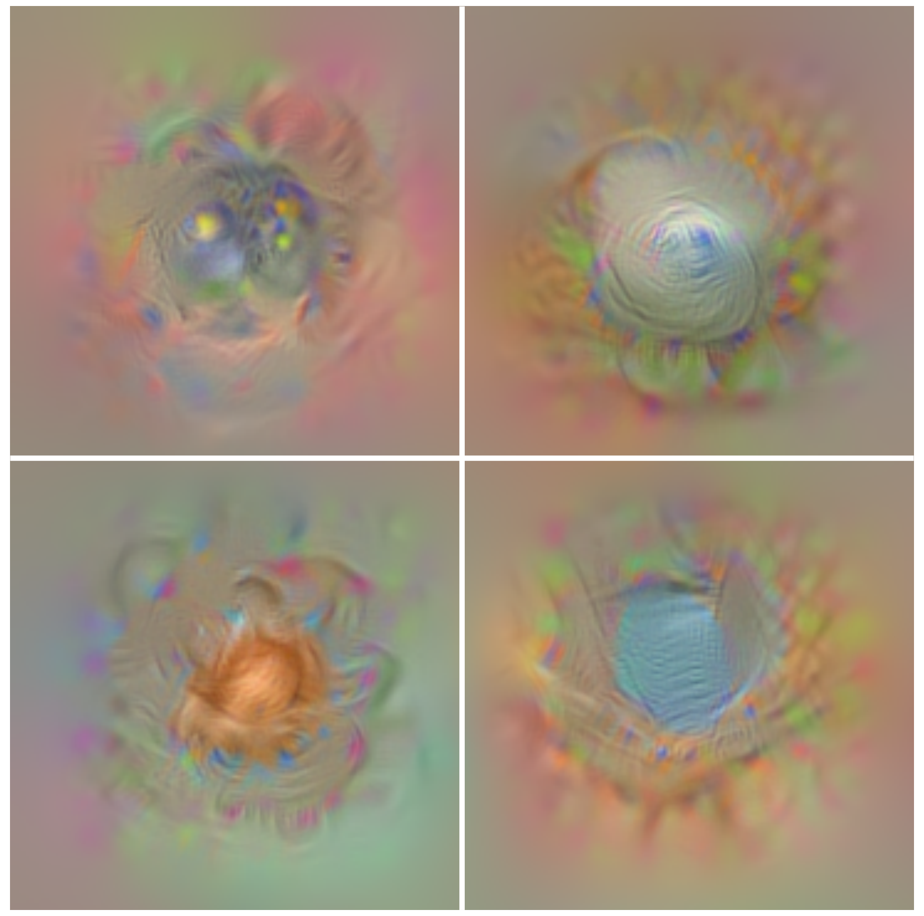
Zeiler & Fergus, ECCV 2014

Nyugen et al., NIPS 2016



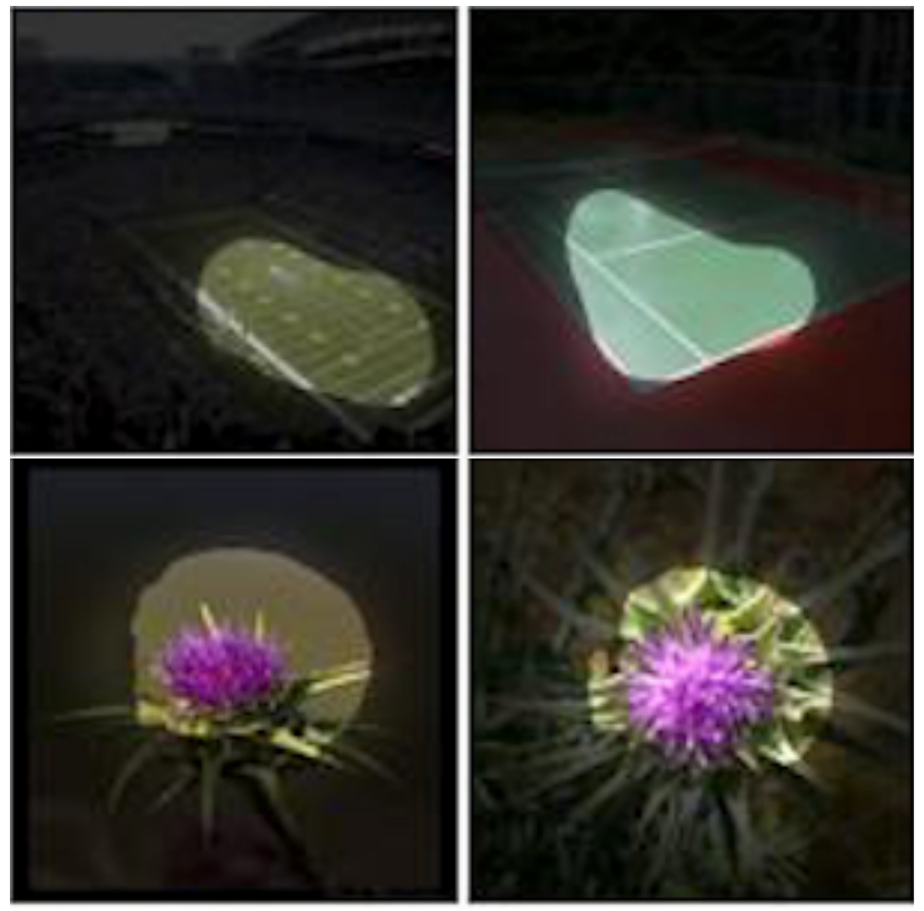
Zhou et al., ICLR 2015

Mahendran & Vedaldi, IJCV 2016



Olah et al., Distill 2017

Bau et al., CVPR 2017

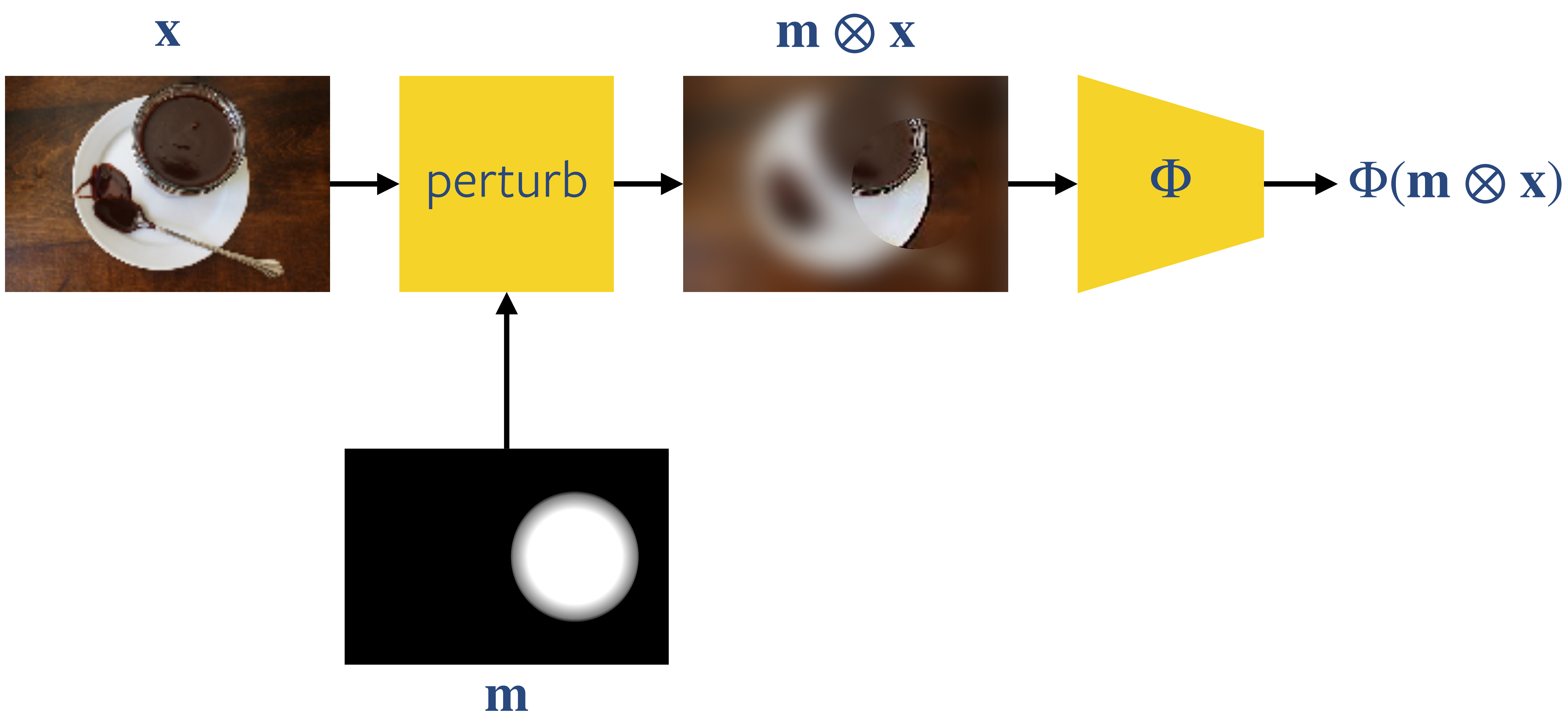


Most prior work focuses on visualizing **single channels**.

A. Attributing channels in intermediate activations

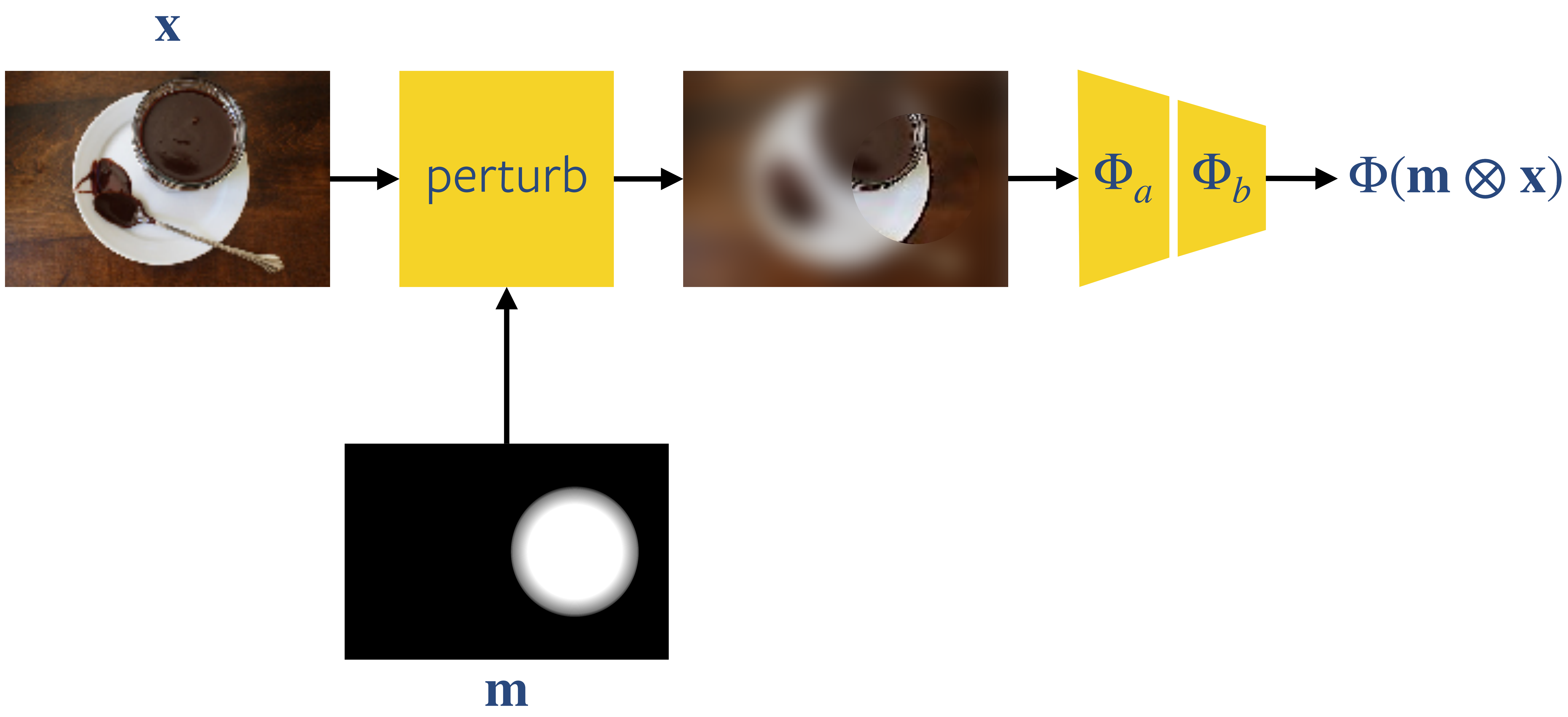
[Fong et al., ICCV 2019]

Spatial Attribution



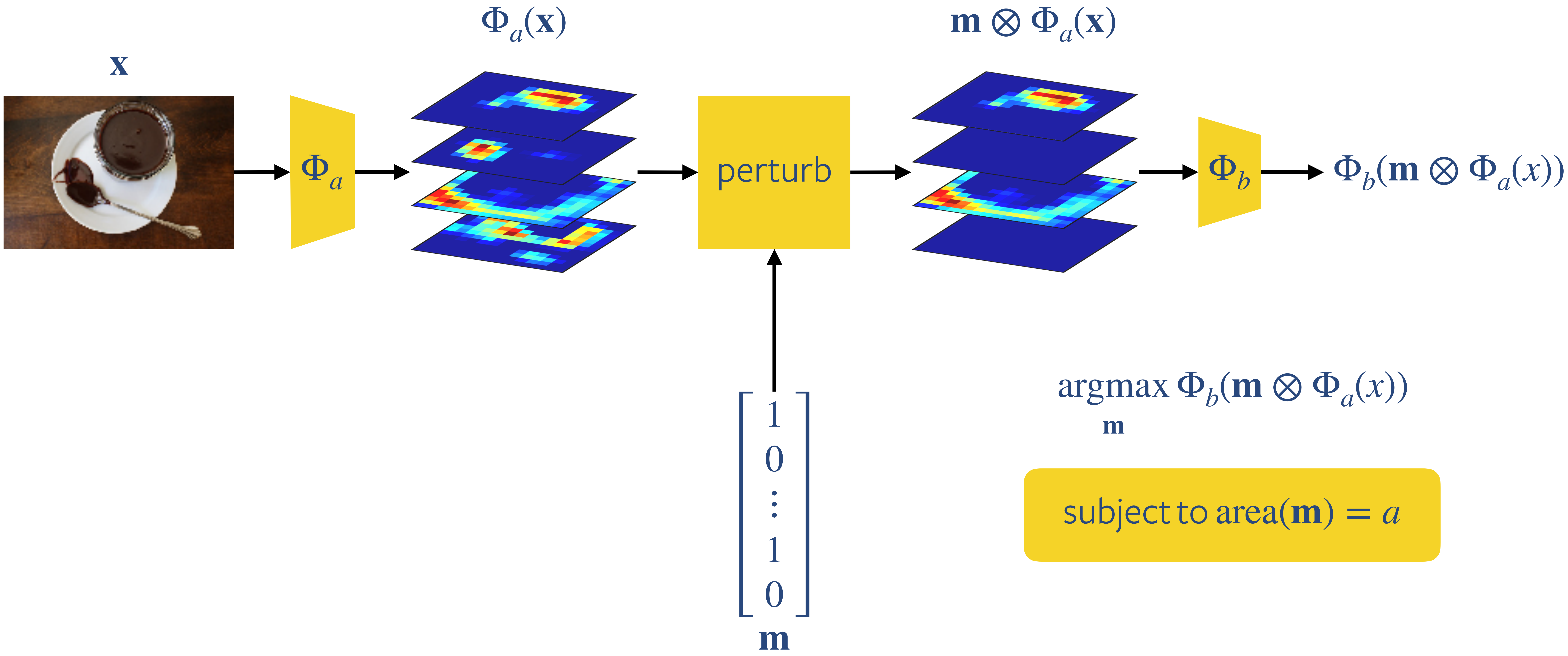
[Fong et al., ICCV 2019]

Channel Attribution



[Fong et al., ICCV 2019]

Channel Attribution



$$\underset{\mathbf{m}}{\operatorname{argmax}} \, \Phi_b(\mathbf{m} \otimes \Phi_a(x))$$

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

[Fong et al., ICCV 2019]

Activation "Diffing"

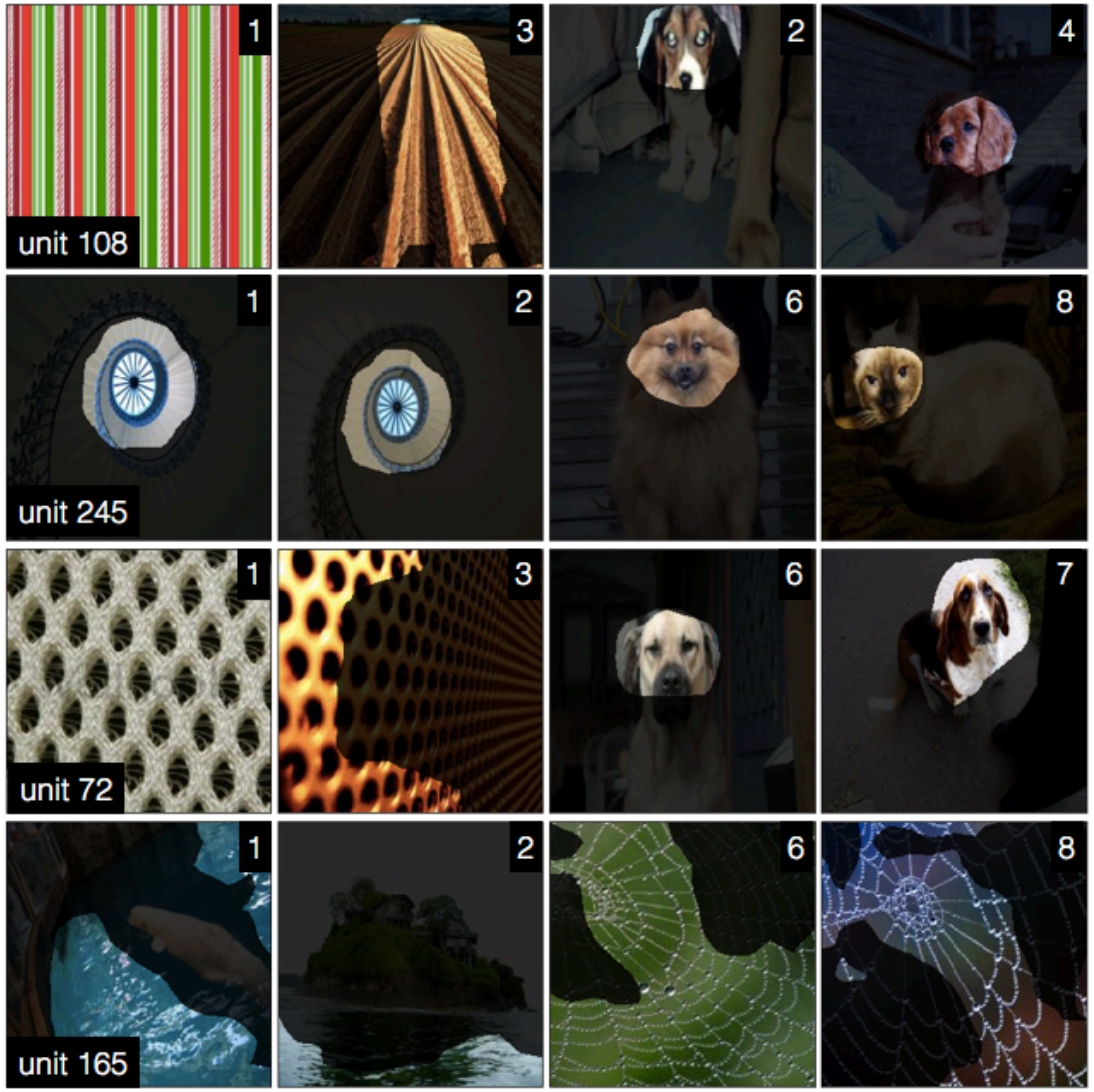
[Olah et al., Distill 2017; Fong et al., 2019]

B. Understanding how semantic concepts are encoded

[Fong & Vedaldi, CVPR 2018]

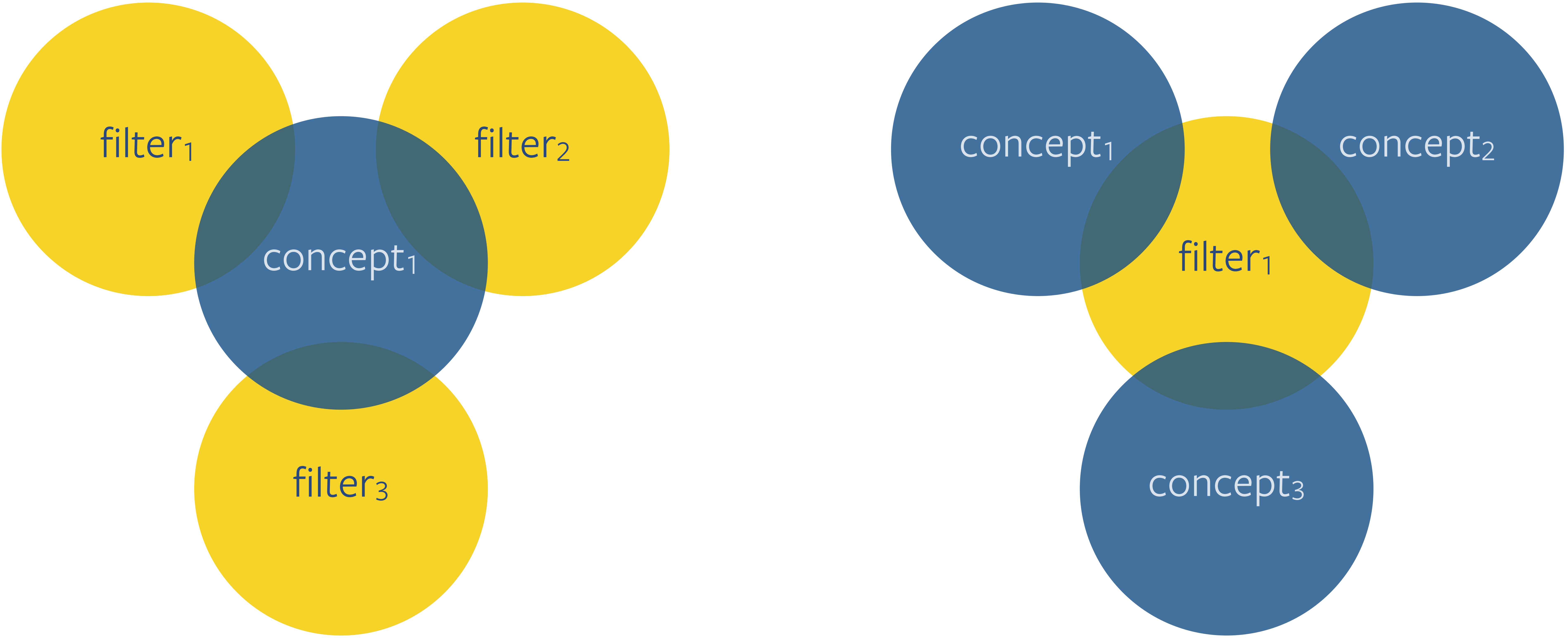
Filter-Concept Overlap

Top activated patches for specific units in AlexNet conv5 filters points to a **packing phenomenon**.



[Bau et al., CVPR 2017]

Filter-Concept Overlap



[Fong & Vedaldi, CVPR 2018; Olah, Github 2018]

Net2Vec

Learn **concept vectors** that describe how a concept is encoded **across** channels.

Probe a **network** with a **concept dataset** and learn to perform a **task** using **activations** at a given layer.

[Fong & Vedaldi, CVPR 2018; concurrent: Kim et al., ICML 2018; Zhou et al., ECCV 2018]

Results

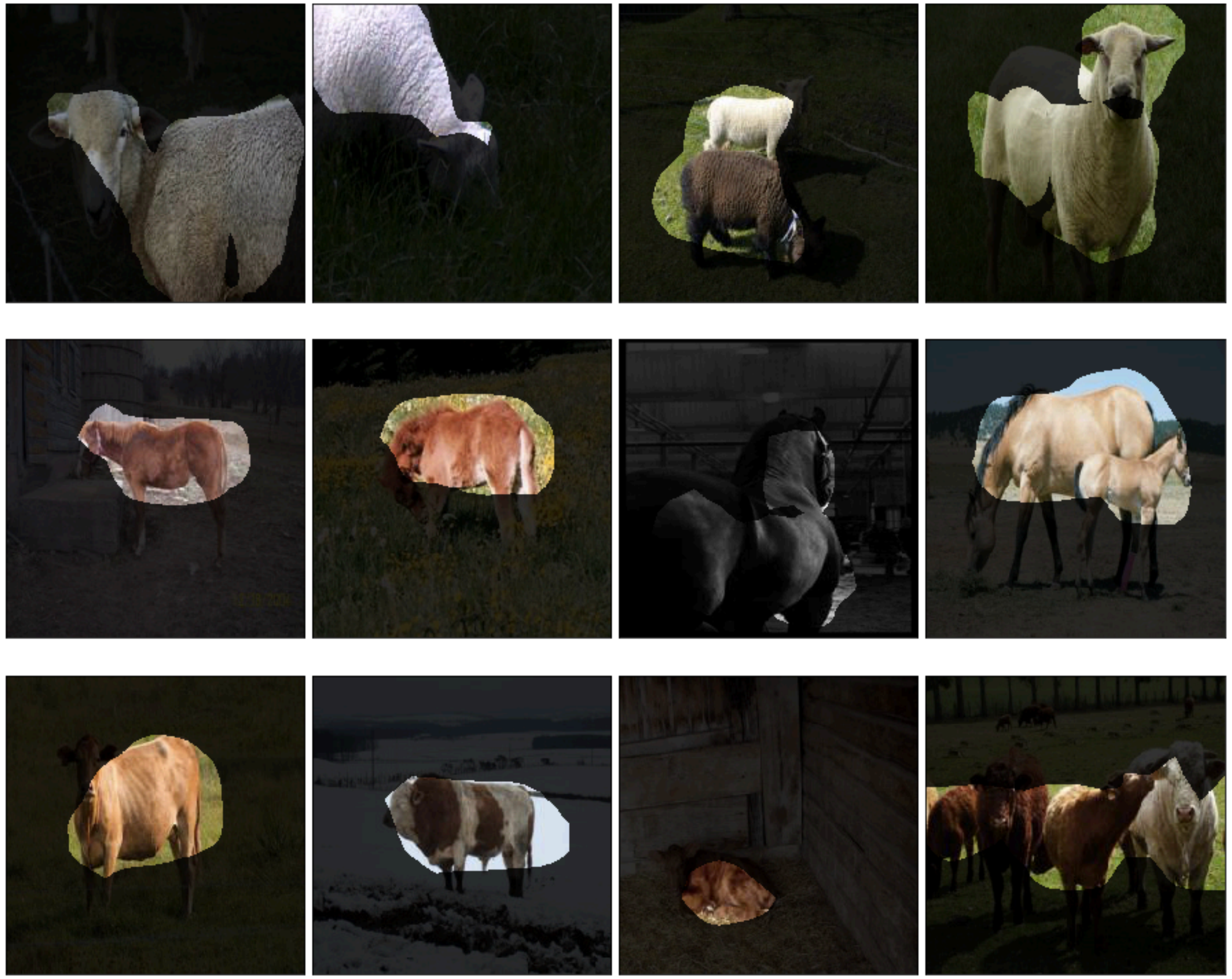
Concepts per Filter

AlexNet conv5 unit 66 is highly selective for various farm animals

Sheep
(IoU_{set} = .21)

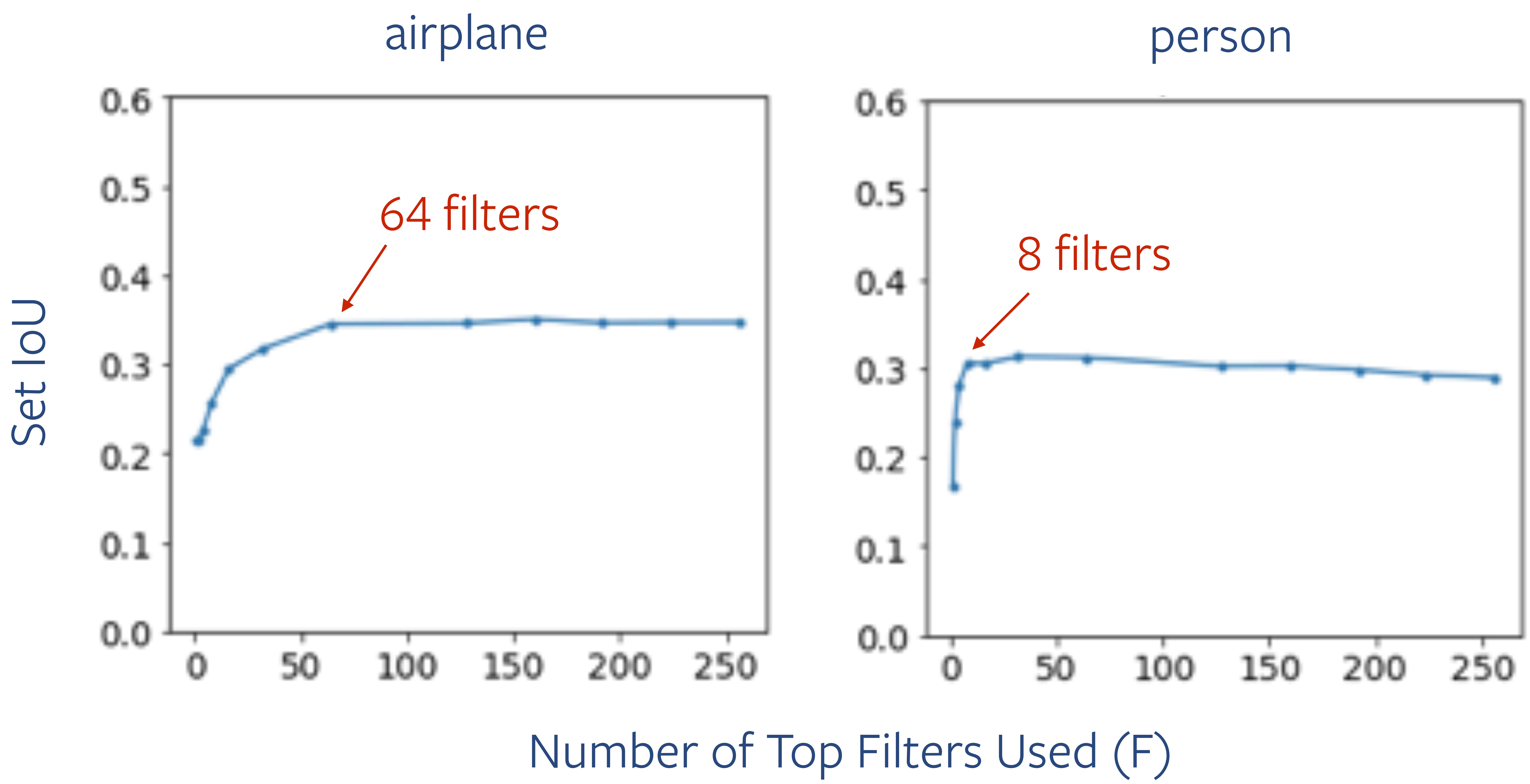
Horse
(IoU_{set} = .21)

Cow
(IoU_{set} = .20)



[Fong & Vedaldi, CVPR 2018]

Filters per Concept



Different concepts require different number of filters for encoding.

[Agrawal et al., ECCV 2014; Fong & Vedaldi, CVPR 2018]

Supervised Learning
Self-Supervised Learning

( , “sheepdog”)

x **y**

[Russakovsky et al. , IJCV 2015; Zhang et al., ECCV 2016]

Filters: Supervised vs. Self-Supervised

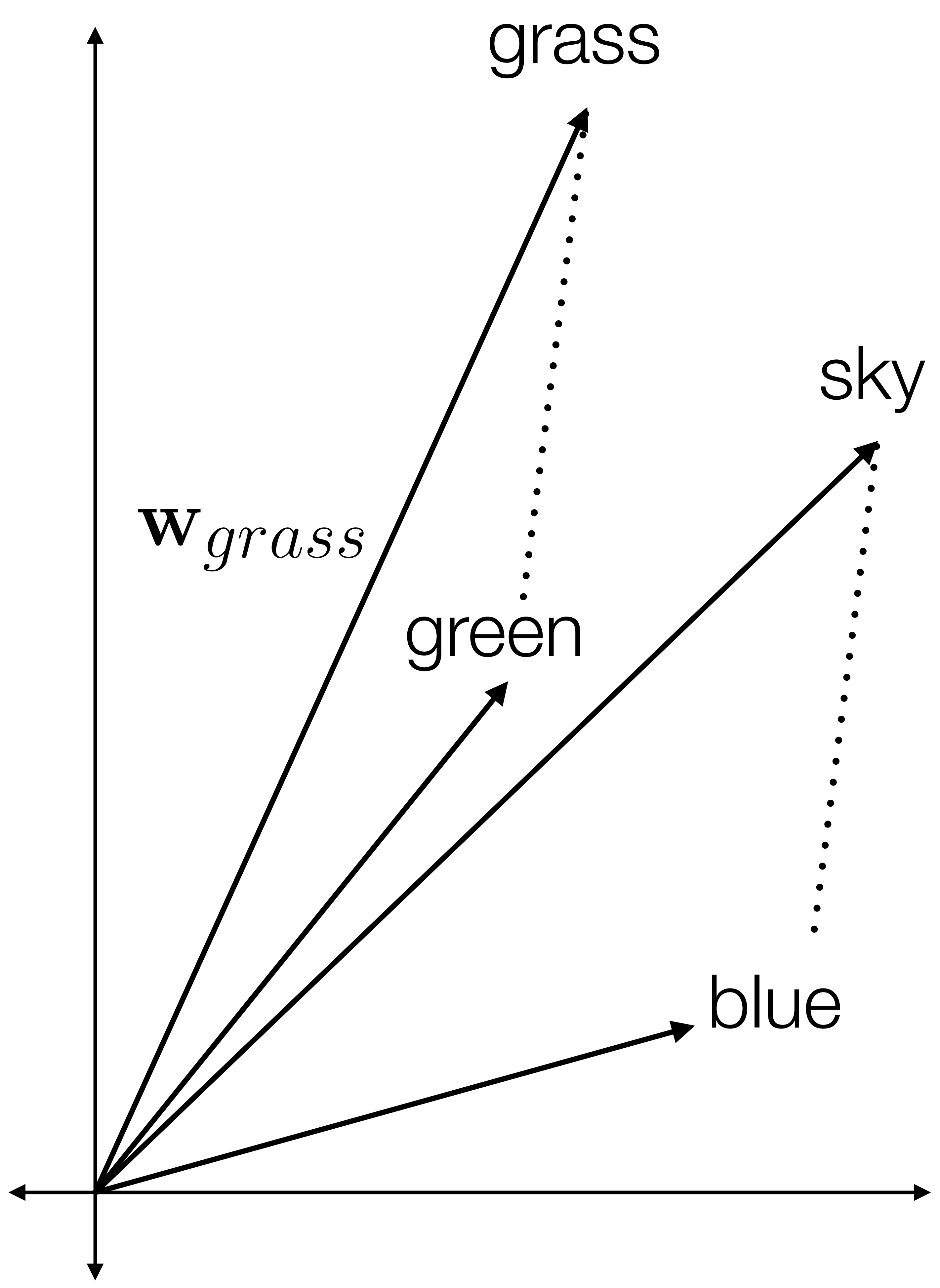
Performance Improvement (Single Filter → All Filters):

- Self-supervised networks: 5-6x
- Fully-supervised networks: 2-4x

Self-supervised networks encode connects more distributively.

[Fong & Vedaldi, CVPR 2018]

Comparing Concept Embeddings



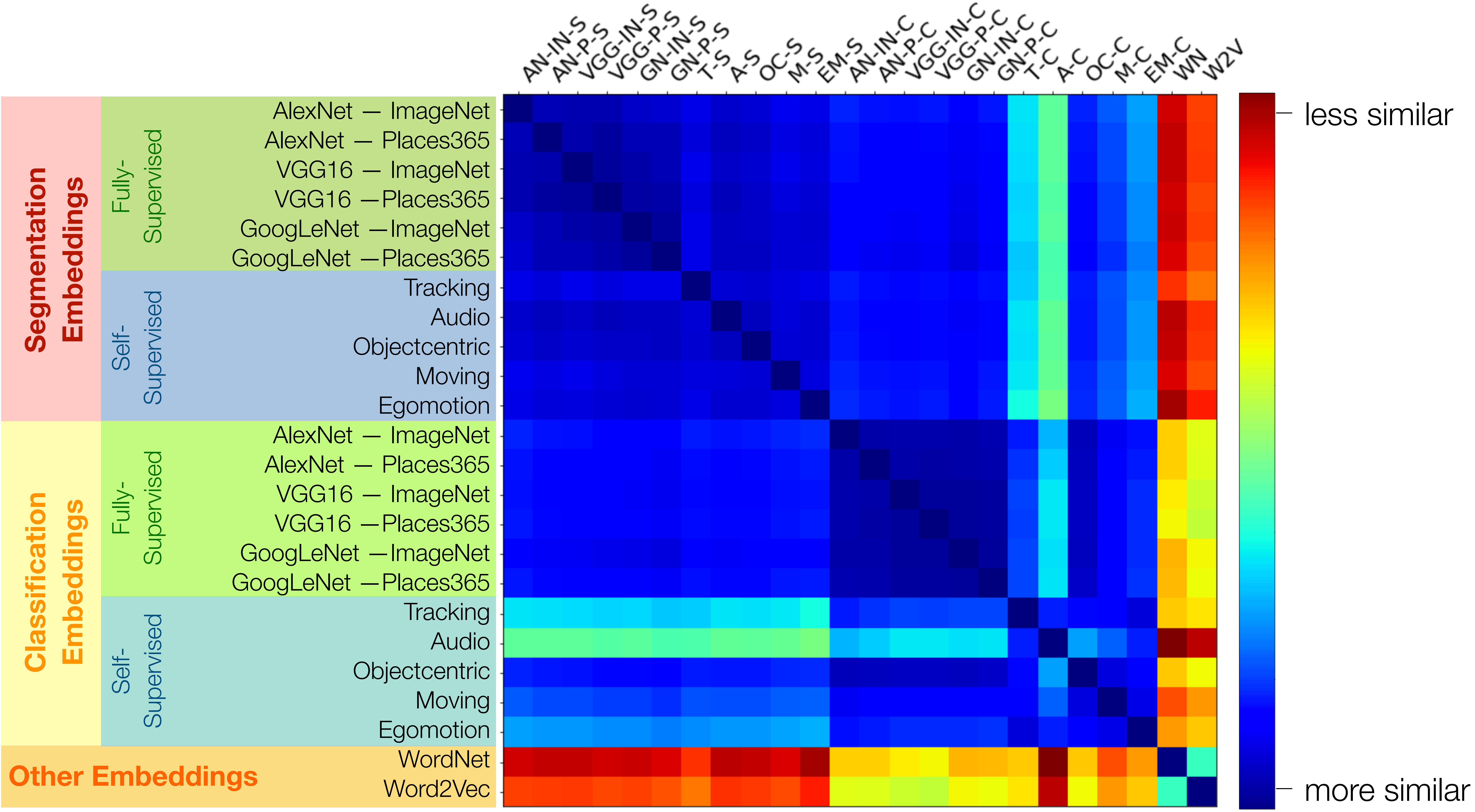
grass + blue — green = sky

tree — wood = plant

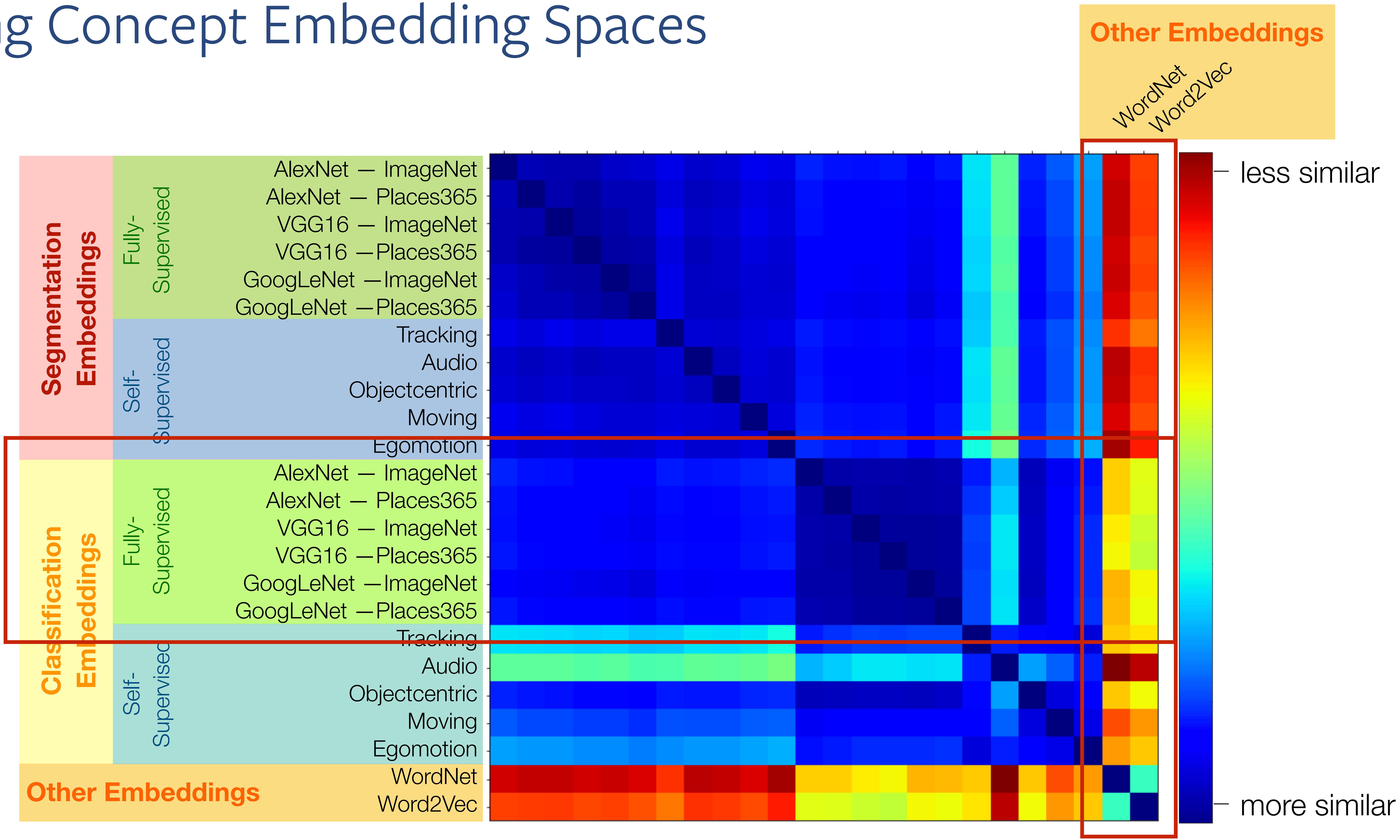
person — torso = foot

[Fong & Vedaldi, CVPR 2018]

Comparing Concept Embedding Spaces

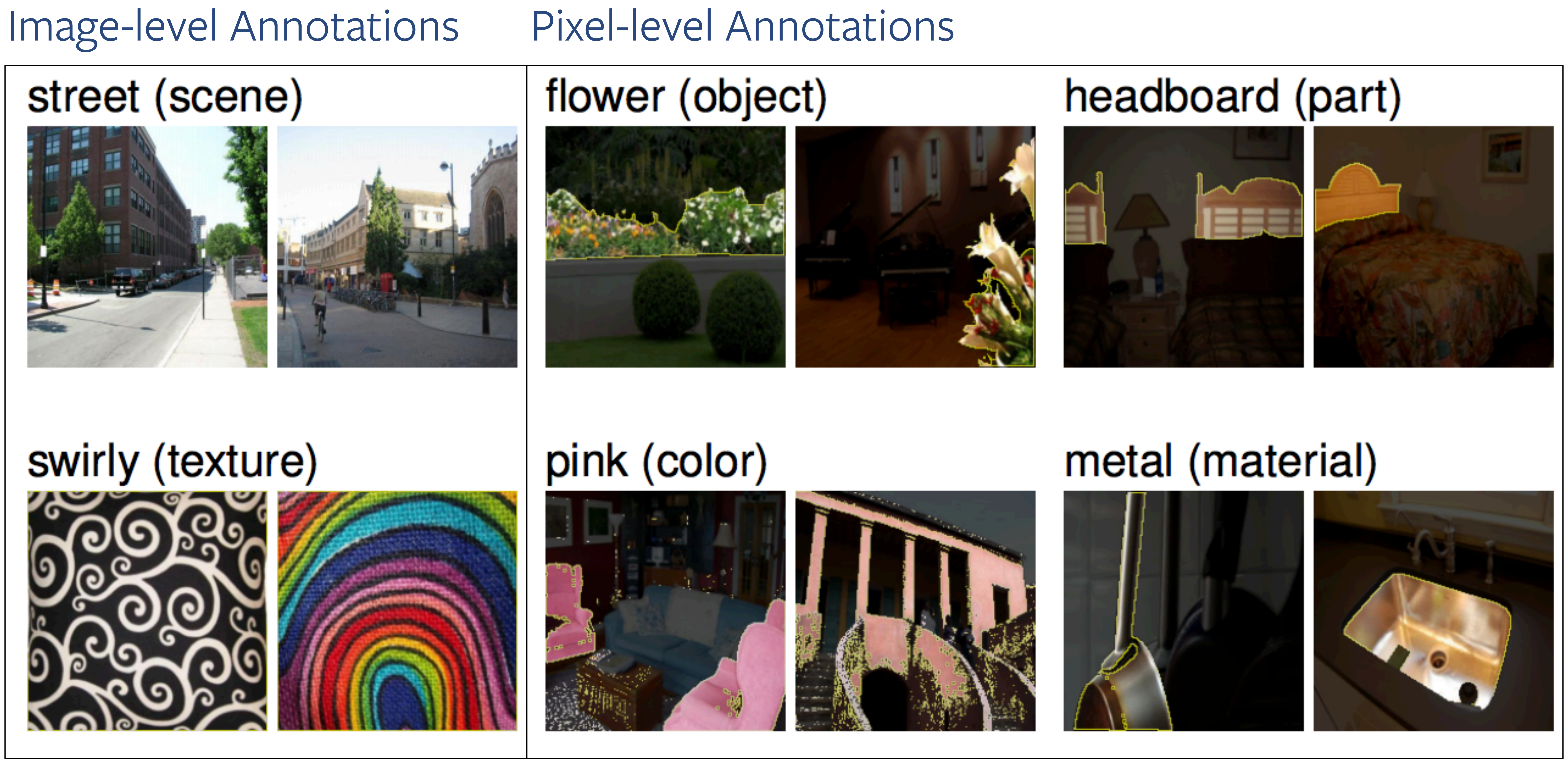


Comparing Concept Embedding Spaces



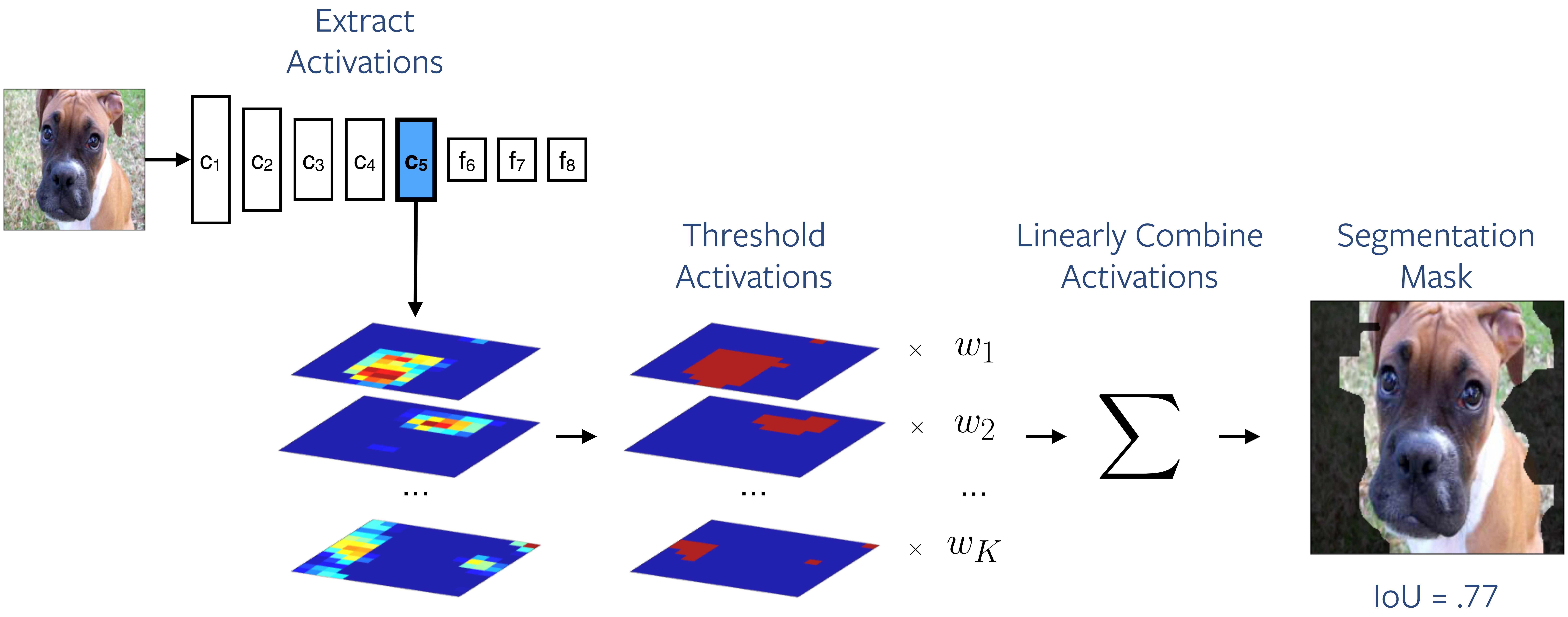
Details

BRODEN dataset



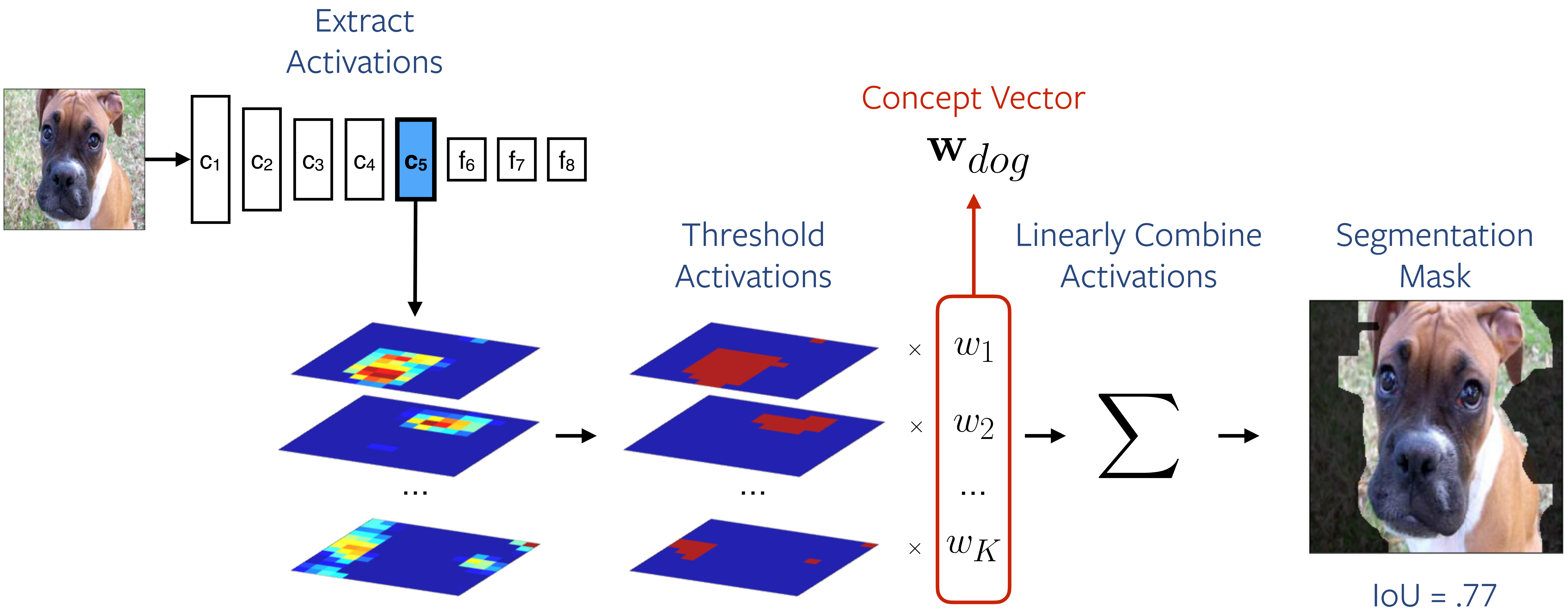
[Bau et al., CVPR 2017]

Segmentation



[Fong & Vedaldi, CVPR 2018]

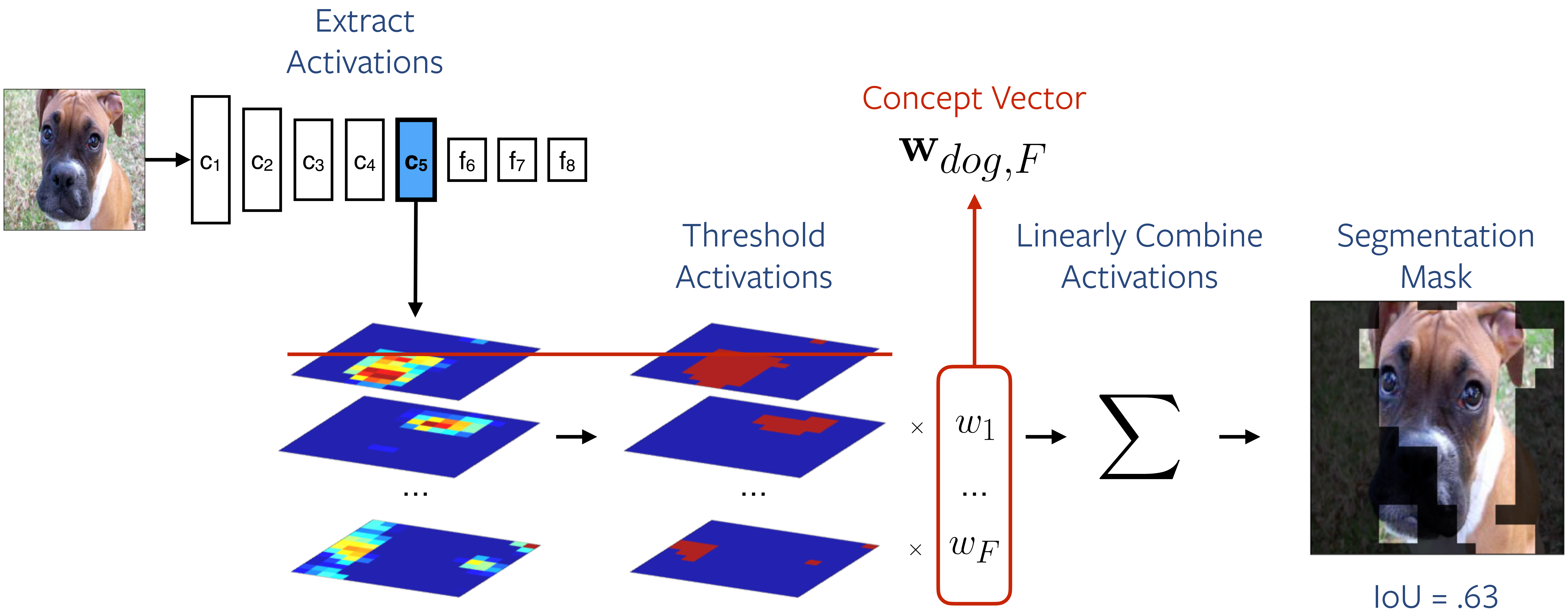
Segmentation



[Fong & Vedaldi, CVPR 2018]

Segmentation

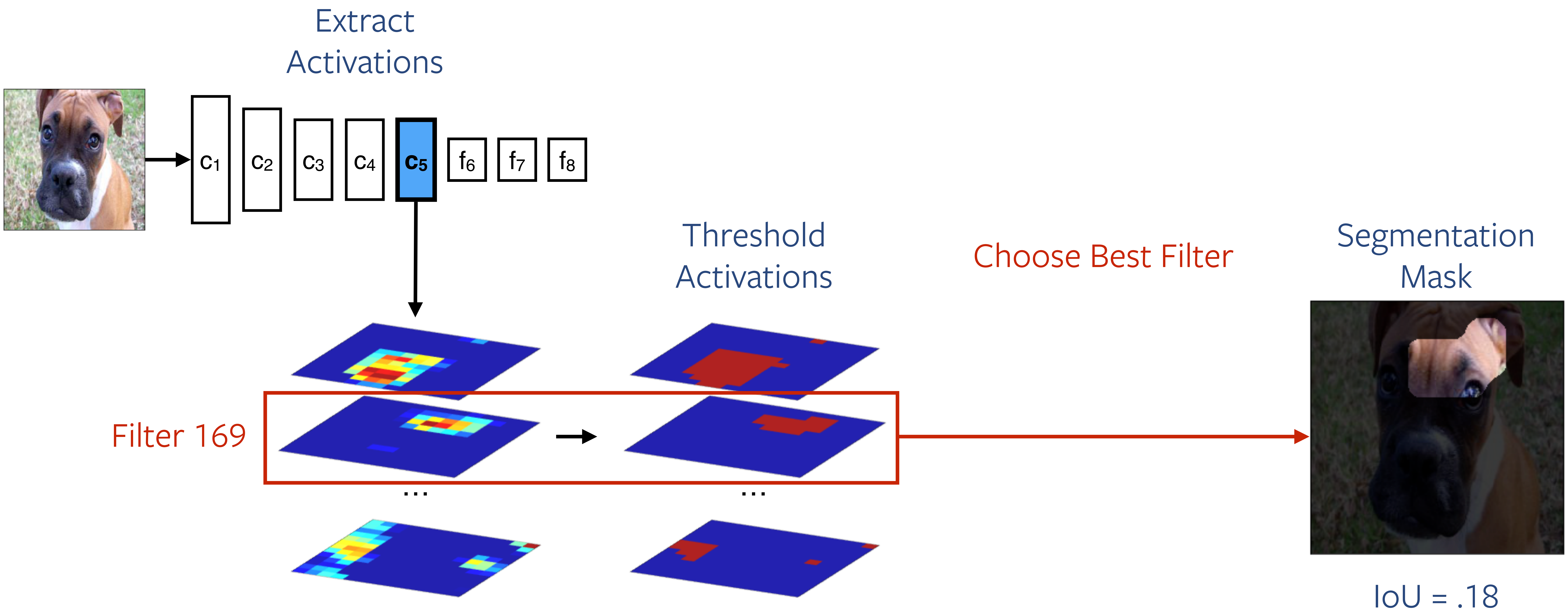
Subset:
Only use top *F* filters,
chosen by magnitude
(e.g., *F* = 4)



[Fong & Vedaldi, CVPR 2018; Agrawal et al., ECCV 2014]

Single Filter
[Bau et al., 2017]

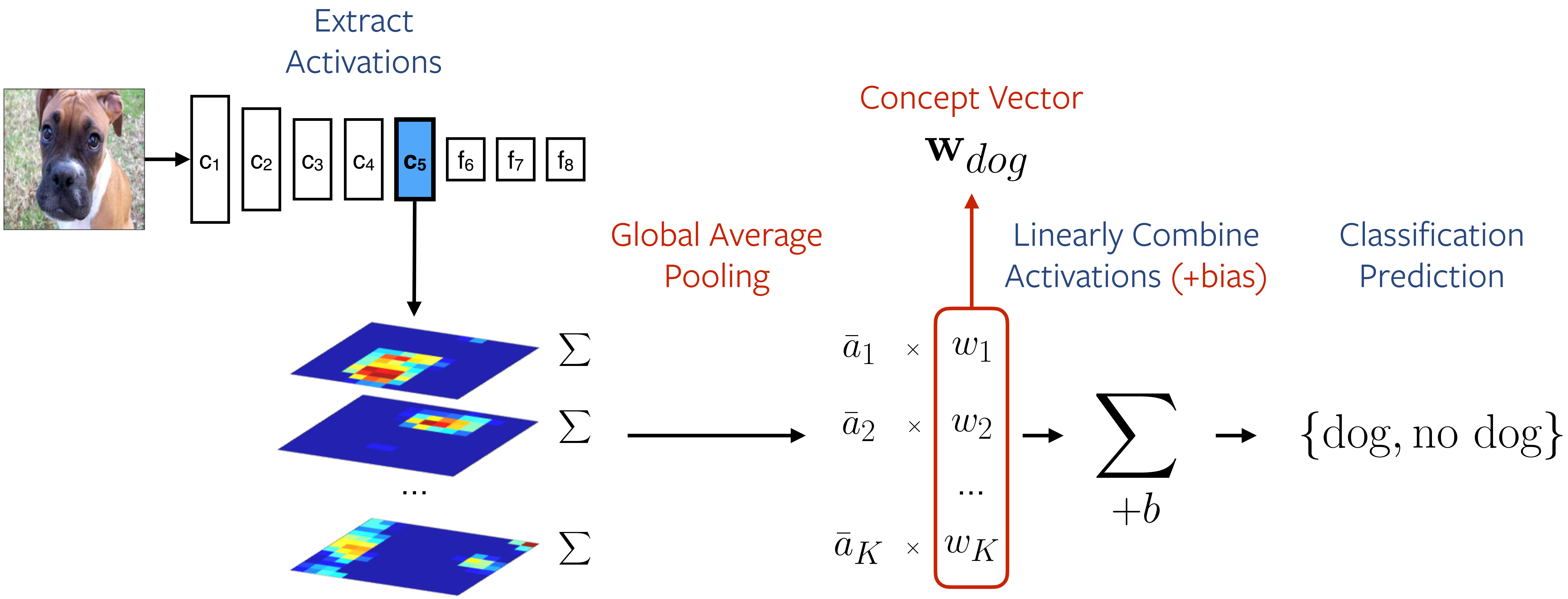
Segmentation



IoU = .18

[Fong & Vedaldi, CVPR 2018]

Classification

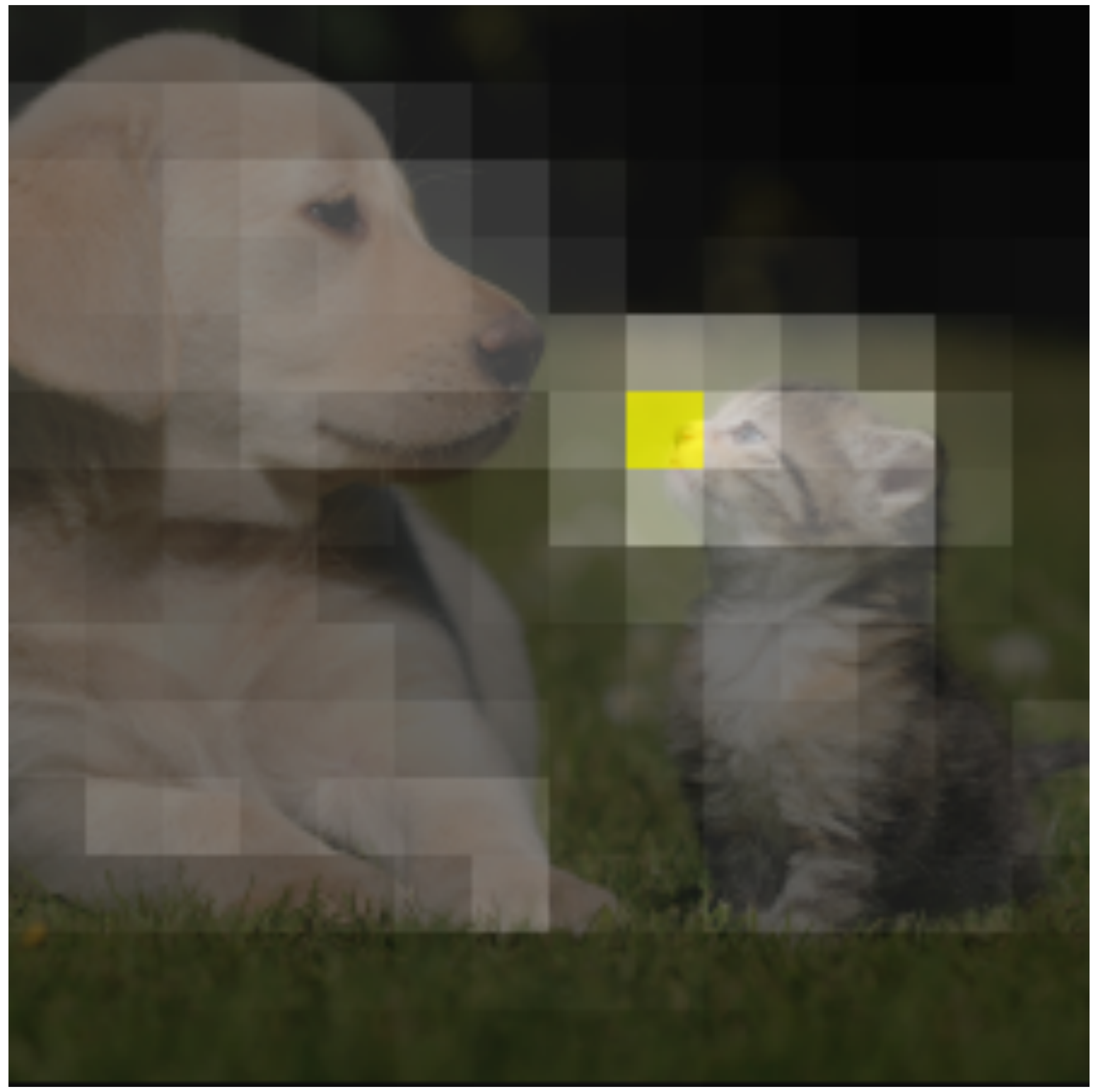


[Fong & Vedaldi, CVPR 2018]

C. Exploring activations via interactive visualizations

[Fong et al., 2020 (in prep.)]

Preview: Interactive Similarity Overlays



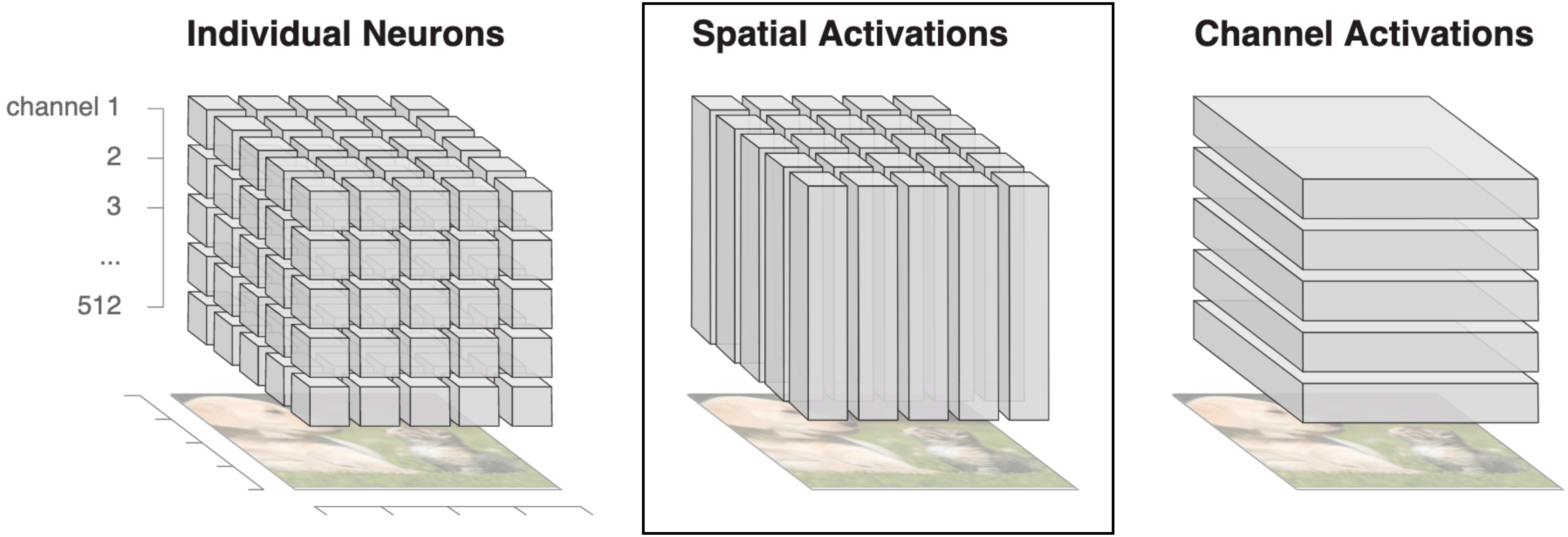
Interactive visualizations empower practitioners to easily understand model behavior.

[Fong et al., 2020 (in prep.)]

Live Preview

[Fong et al., 2020 (in prep.)]

Intermediate Activations



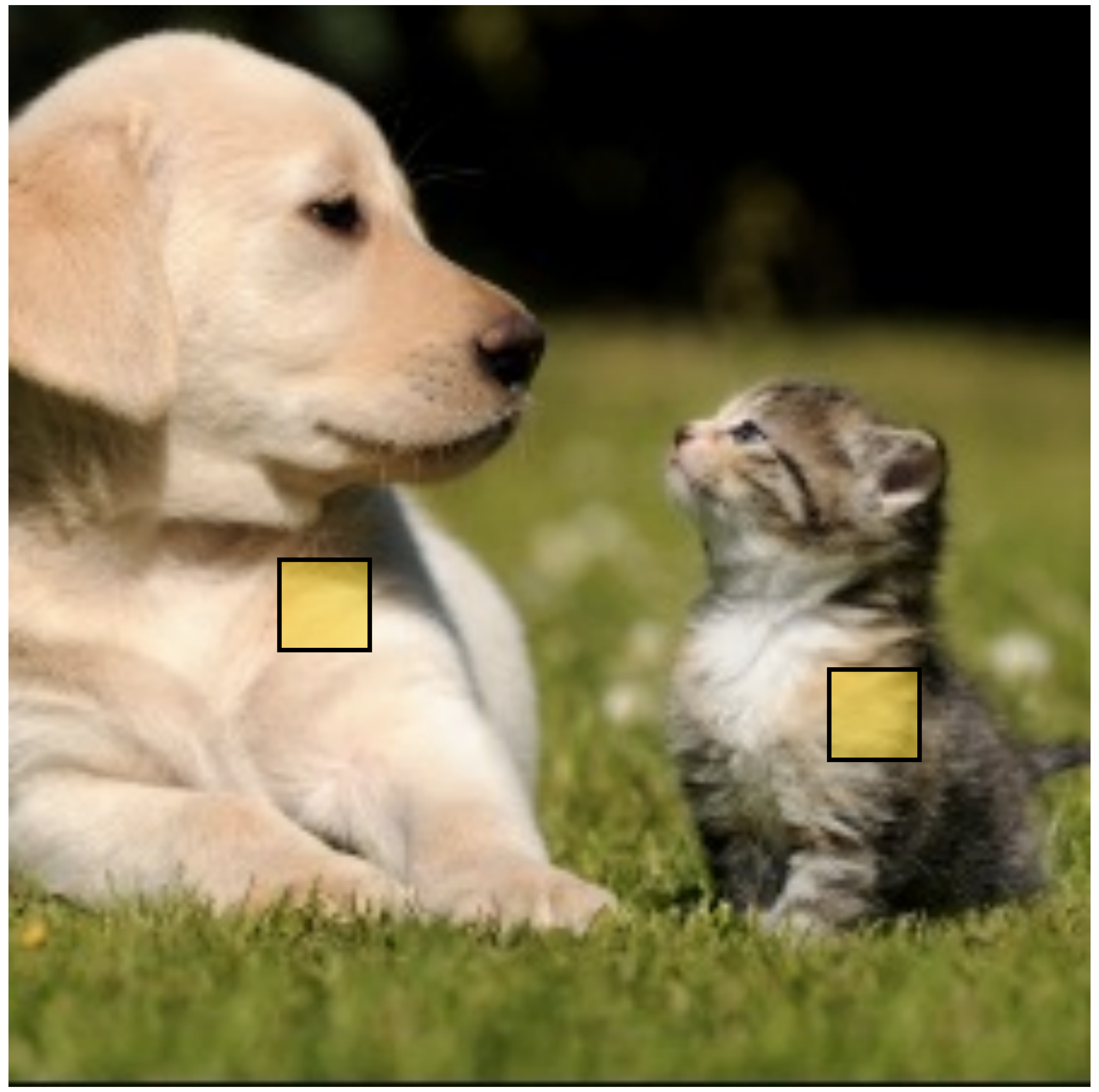
[Olah et al., Distill 2018]

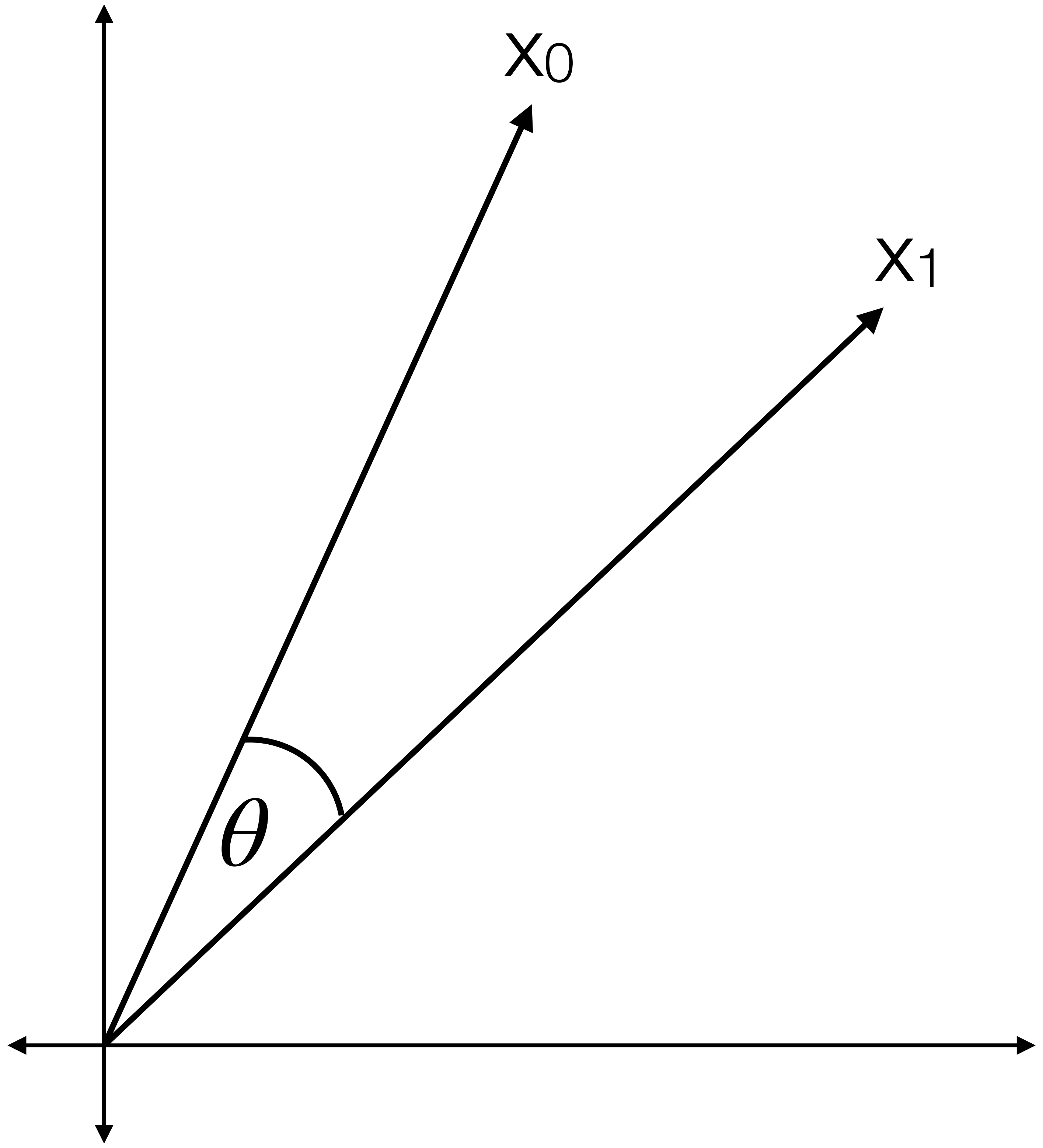
Preview: Interactive Similarity Overlays

$a_{6,5}$ = [17.7, 0, 103.4, 6.81, 0, 0, 0, 0, 32.0, 0, 0, 0, ...]

[Olah et al., Distill 2018]

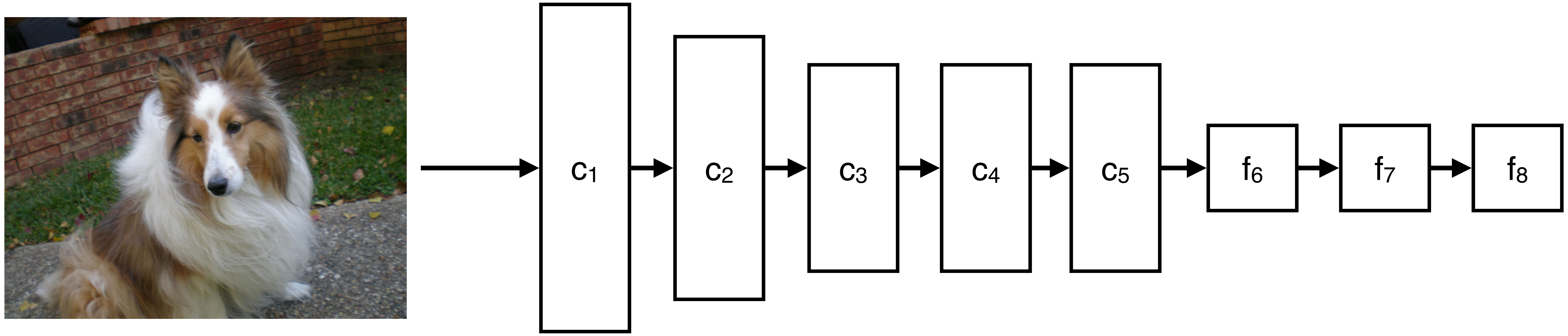
Preview: Interactive Similarity Overlays

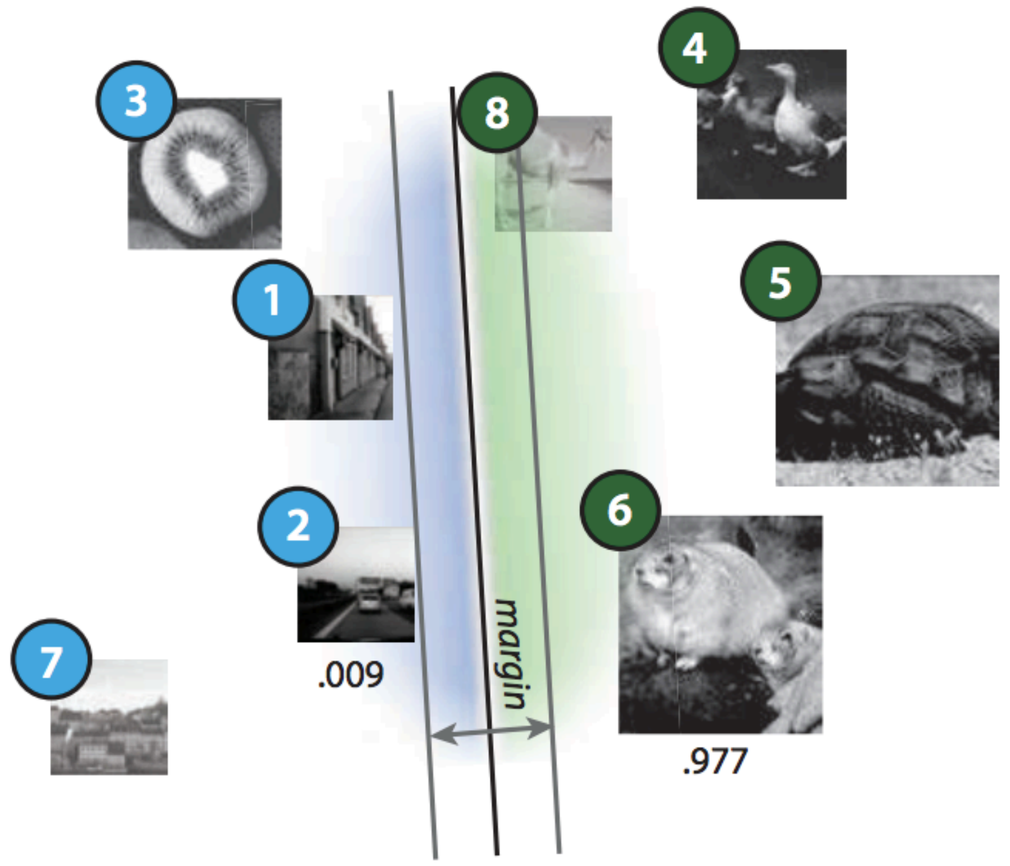




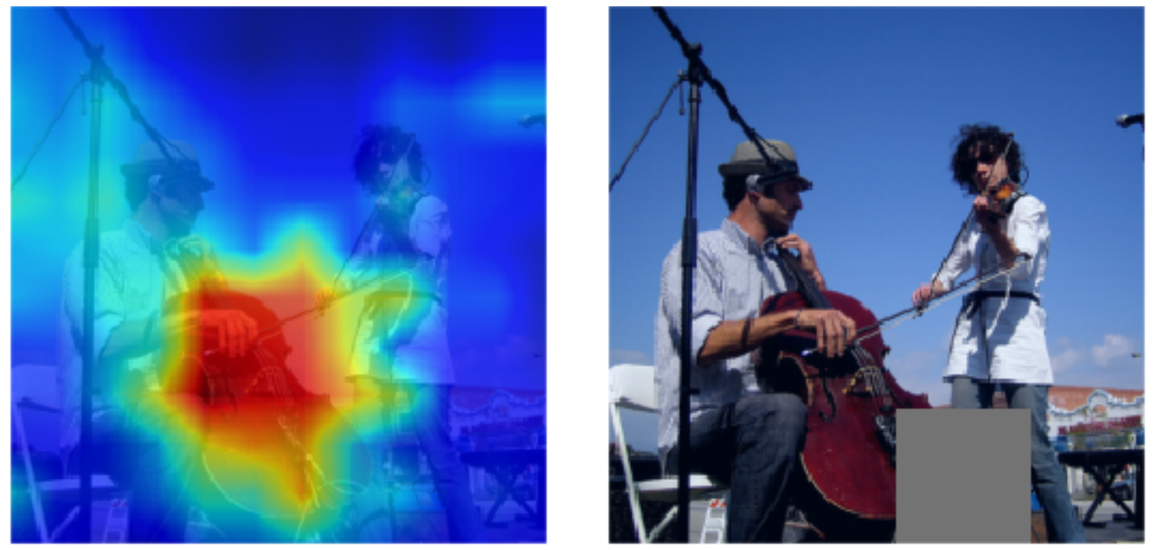
[Fong et al., 2020 (in prep.)]

Research Themes





Fong et al., Sci. Reports 2018

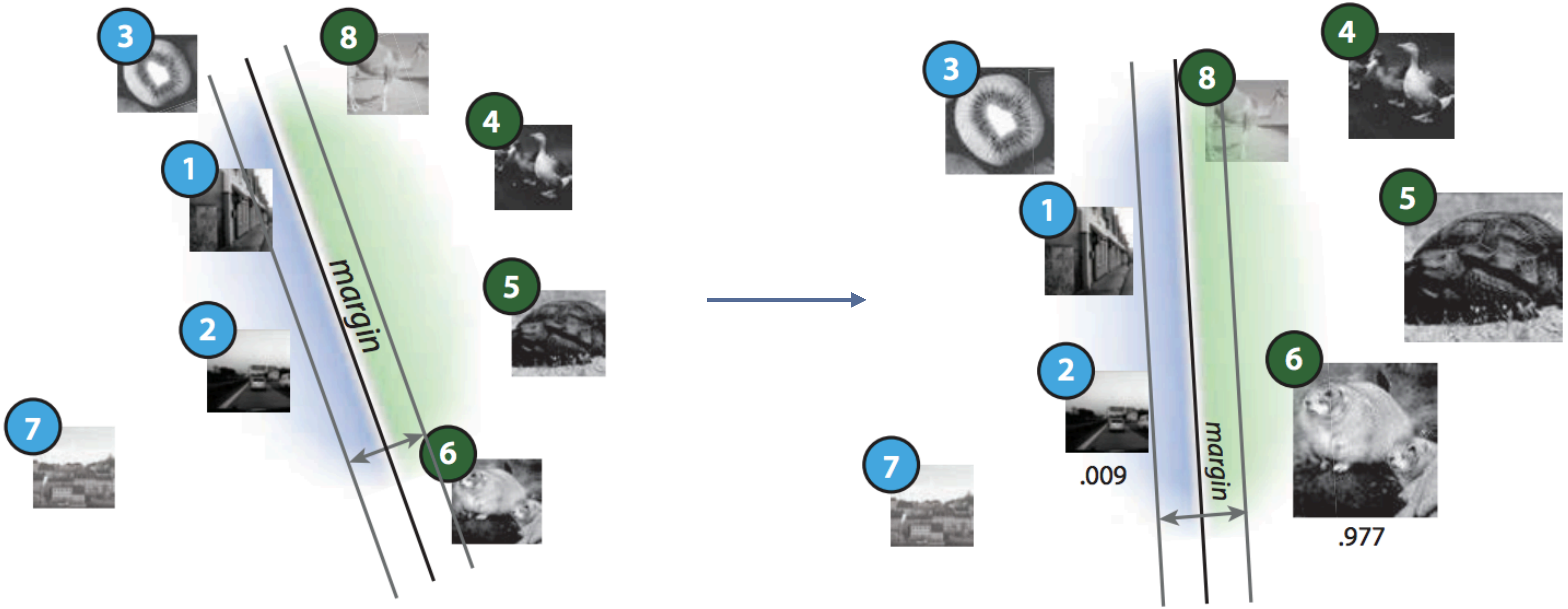


Fong & Vedaldi, ICCVW 2019

Training Procedure: How can we improve f(x)?

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Human-Guided Machine Learning

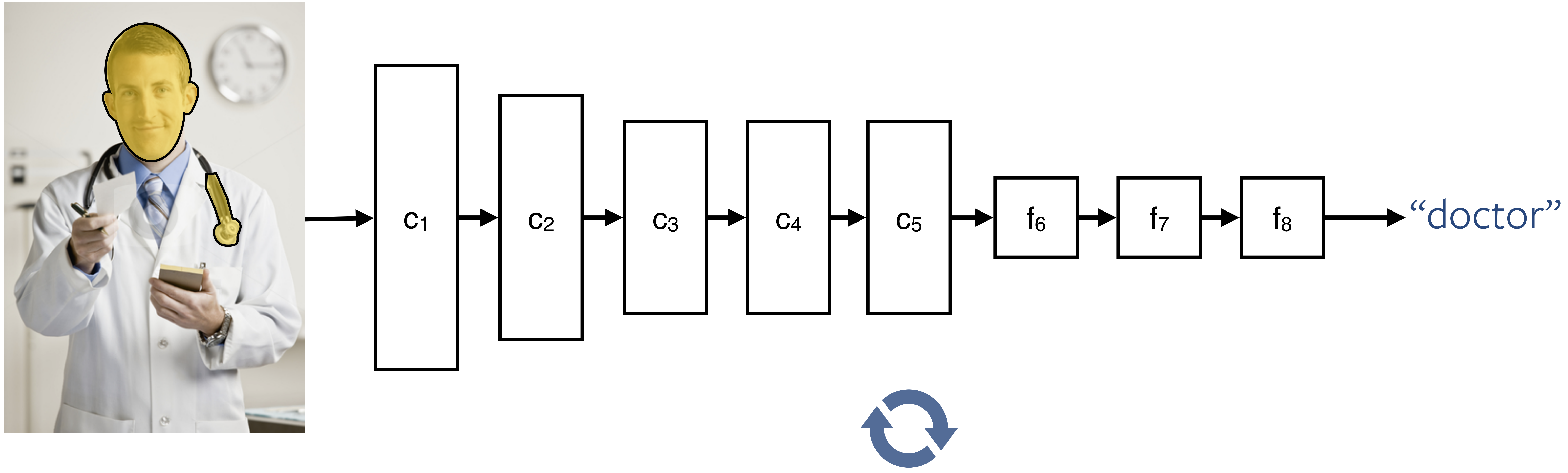


Align machine decisions with human decisions from brain activity.

[Fong et al., Sci. Reports 2018]

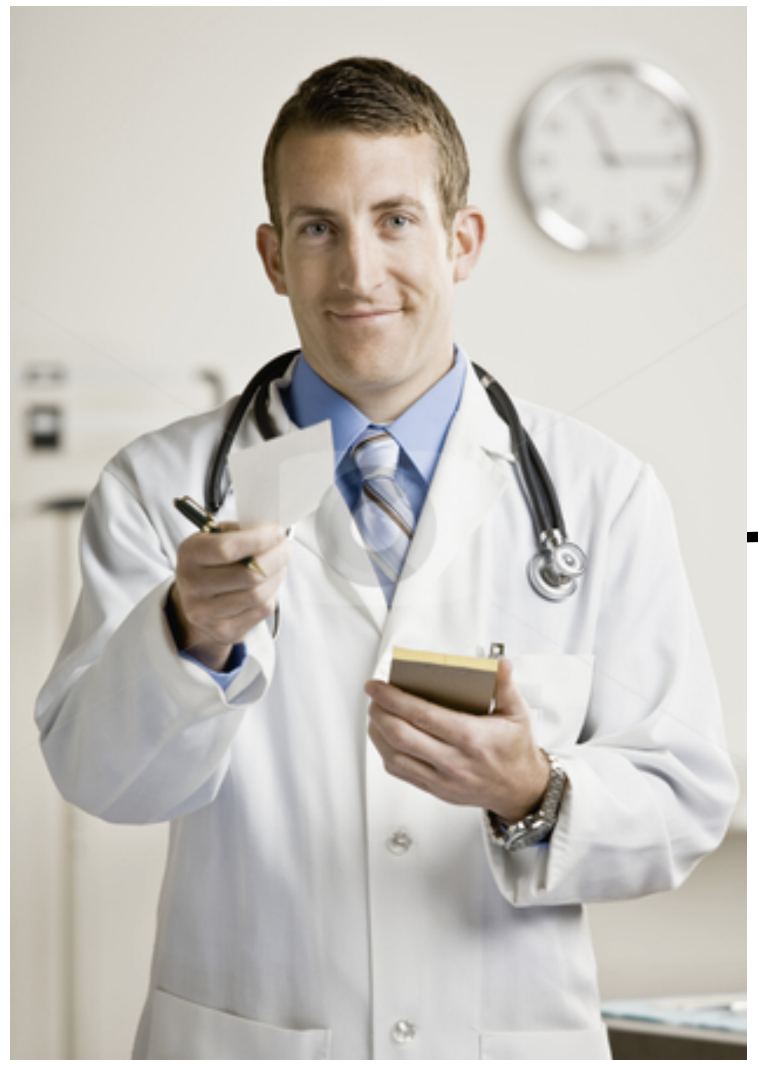
Future Work: Model Debugging

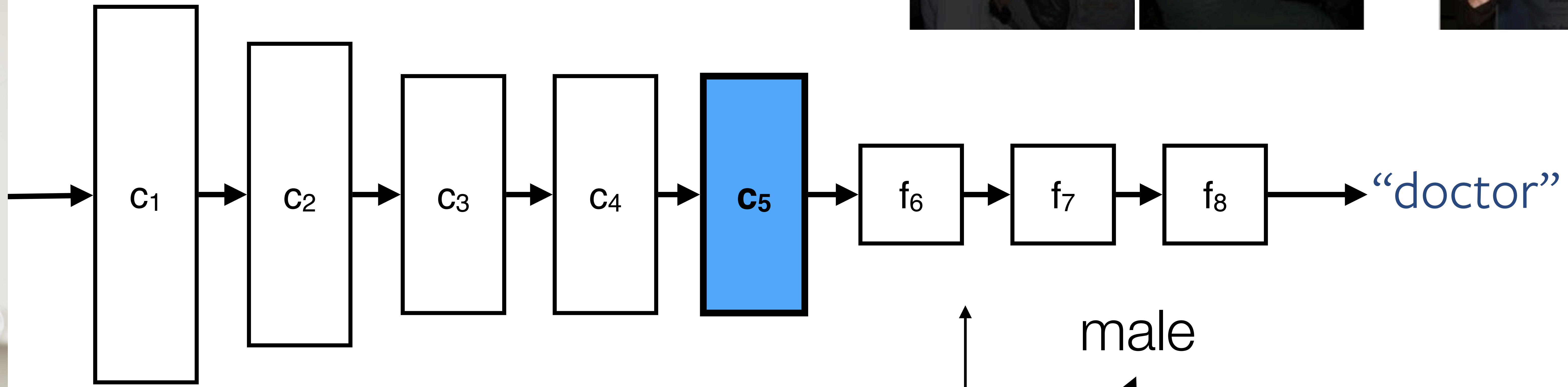
Identify and correct systematic mistakes

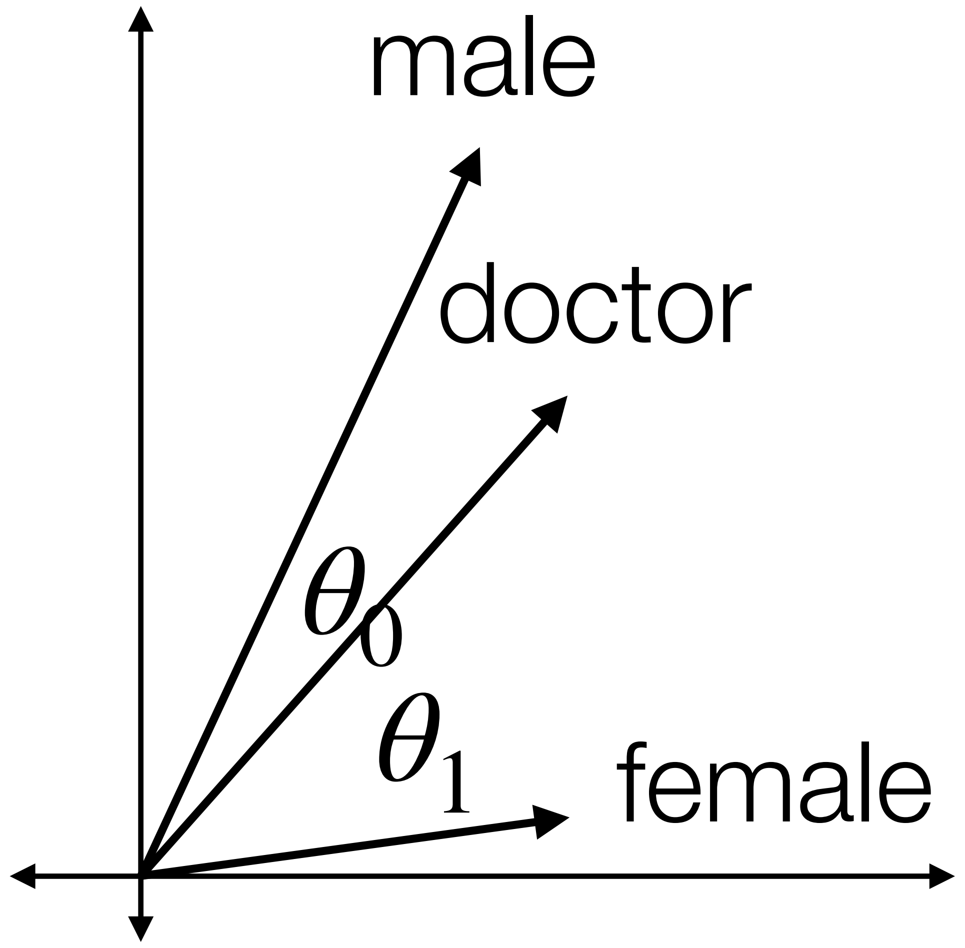


Future Work: Model Debugging

Identify and correct systematic mistakes







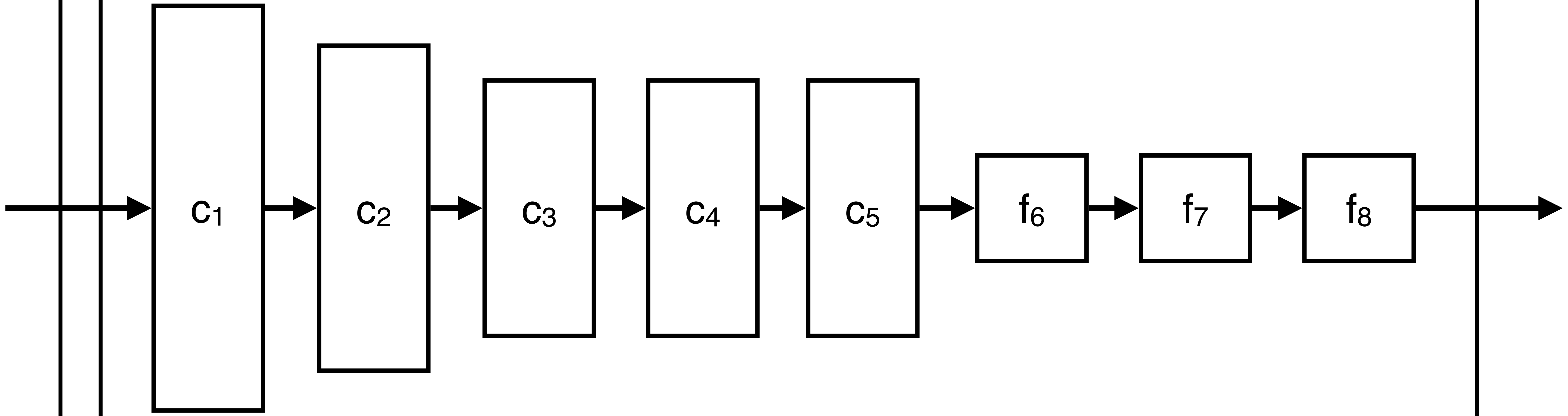
$\theta_0 = \theta_1$

[Bolukbasi et al,. NeurIPS 2016;
David et al., (in prep)]

Research Themes

Inputs: What is f(x) looking at?

Internal Representation: What & how does f(x) encode?



95%: “sheepdog”

1%: “sea snake”

2%: “alp”

0%: “soup bowl”

…

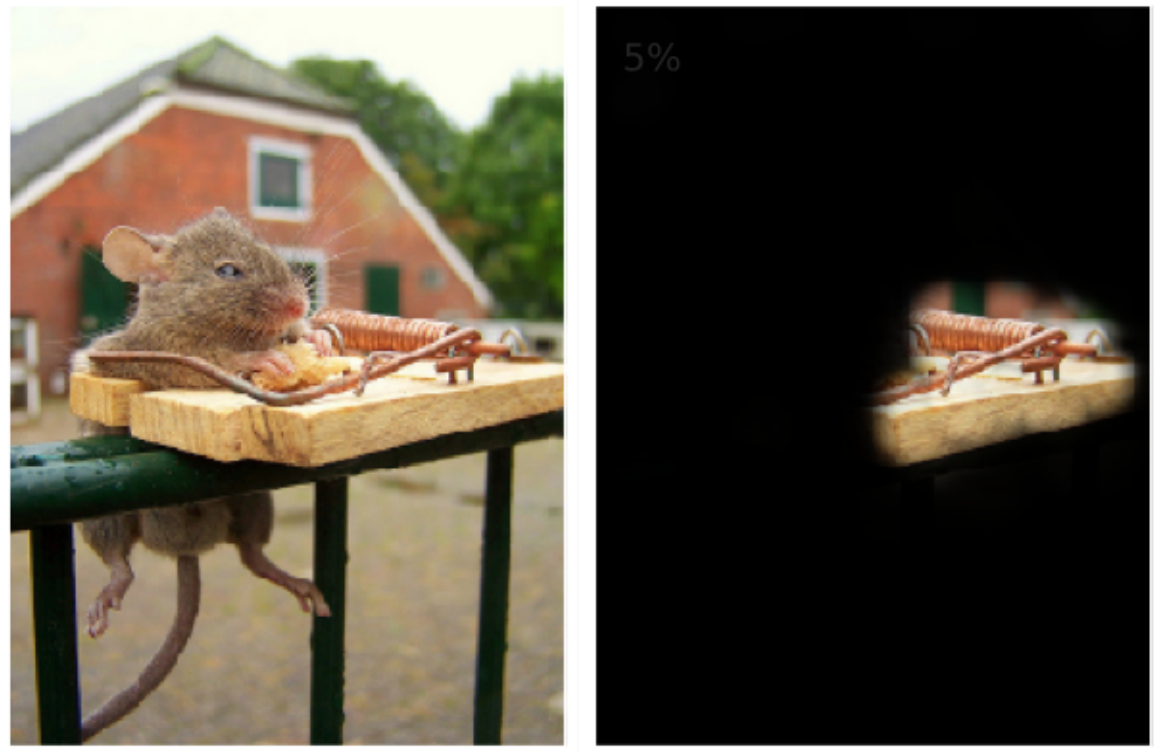
Training Procedure: How can we improve f(x)?

$$\theta \leftarrow \theta - \eta \frac{dL}{d\theta}$$

Research Themes

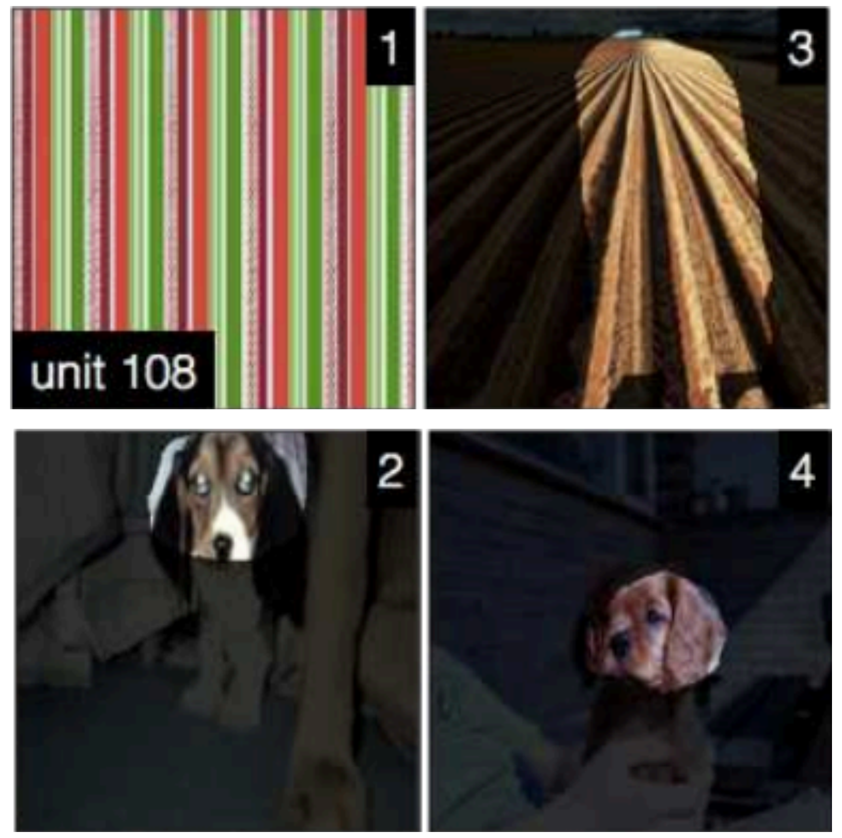
What is f(x) looking at?

Fong & Vedaldi, ICCV 2017

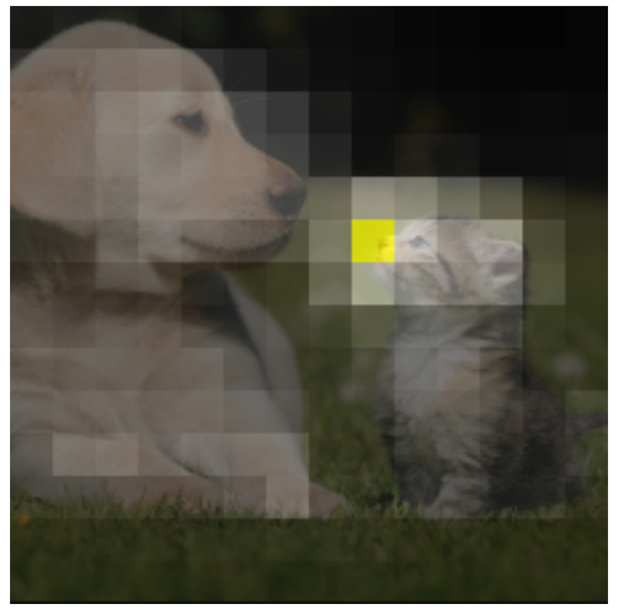


Fong et al., ICCV 2019

What & how does f(x) encode?

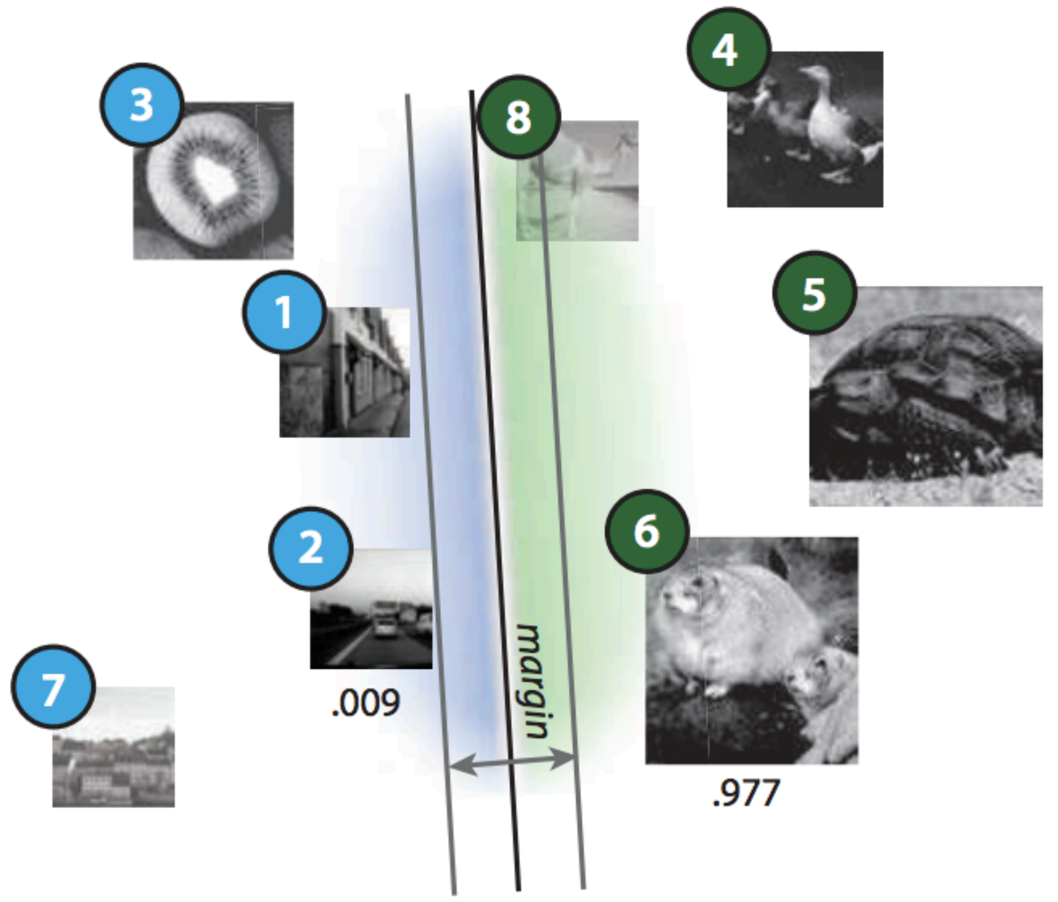


Fong & Vedaldi, CVPR 2018

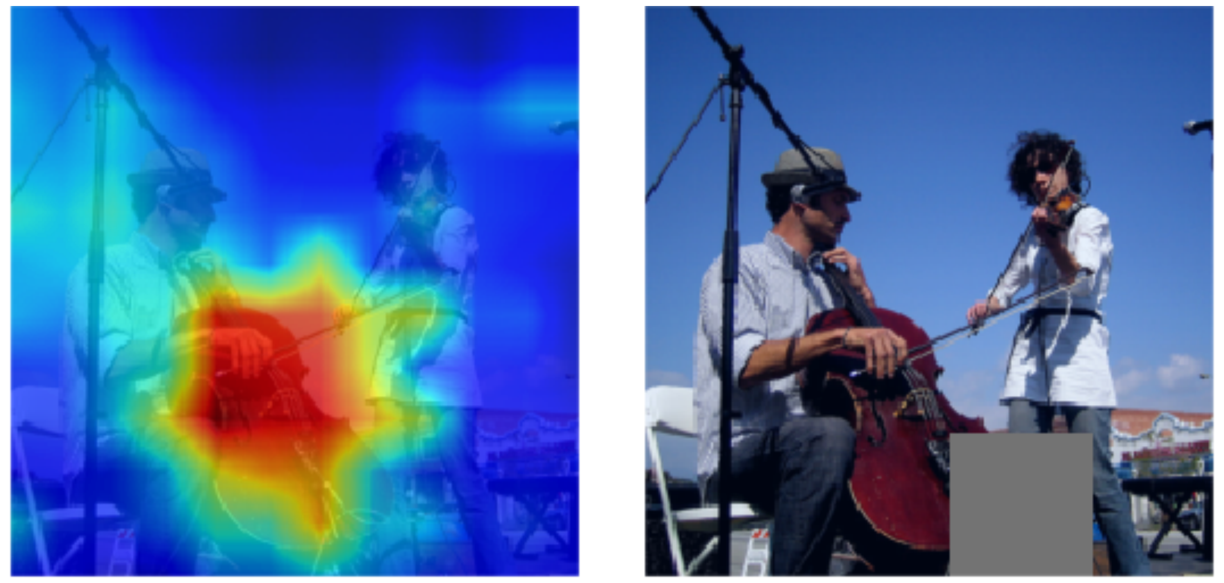


Fong et al., 2020 (in prep.)

How can we improve f(x)?



Fong et al., Sci. Reports 2018



Fong & Vedaldi, ICCVW 2019

Andrea Vedaldi

Mandela Patrick

Chris Olah

Alexander Mordvintsev

Walter Scheirer

David Cox

RHODES TRUST

Open Philanthropy Project

Thank you