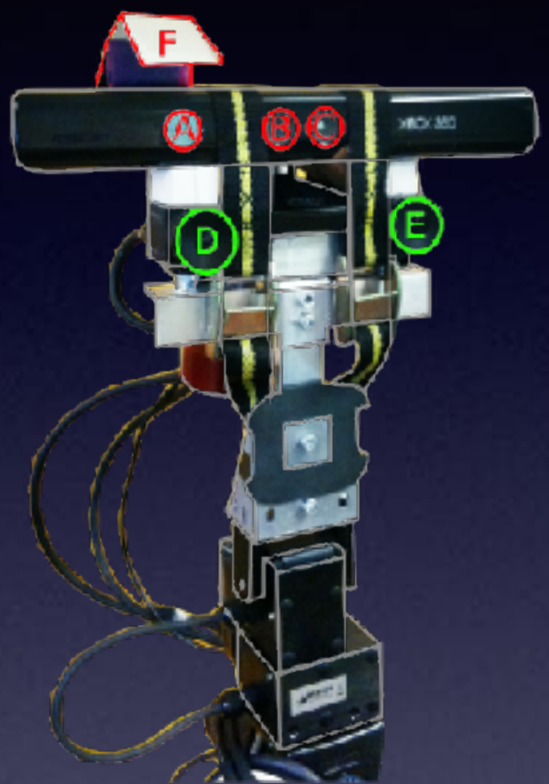
Attentional Masking for Pre-trained Deep Networks

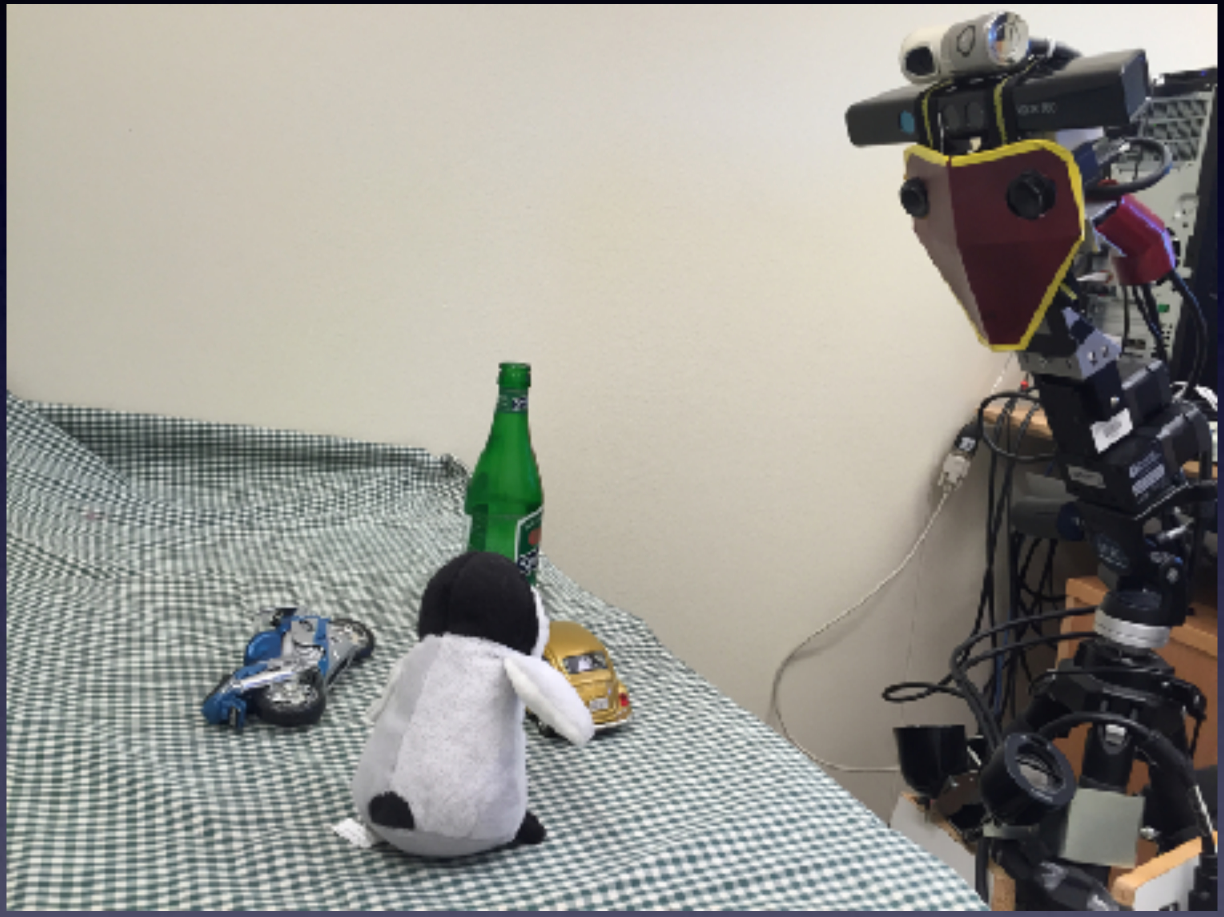
IROS 2017

Marcus Wallenberg and Per-Erik Forssén
Computer Vision Laboratory
Department of Electrical Engineering
Linköping University

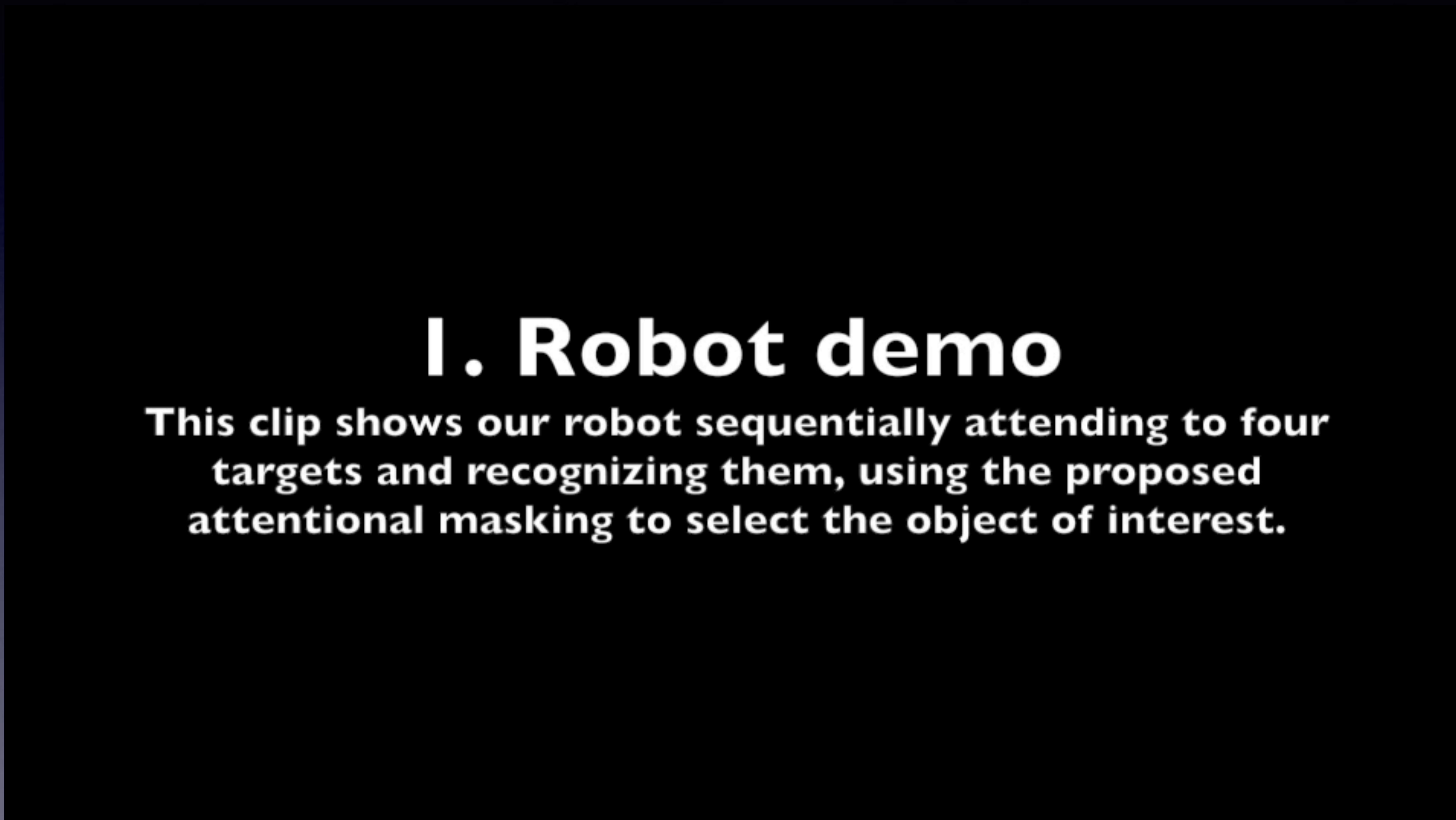
Attention for robots



A: SLP projector (unused)
B: RGB camera (unused)
C: NIR camera (unused)
D: Right wide-angle camera
E: Left wide-angle camera
F: SLP diffusor (unused)



Attention for robots



Attention for robots

Left camera

Right camera

- What is the robot looking at?

LINKÖPING UNIVERSITY

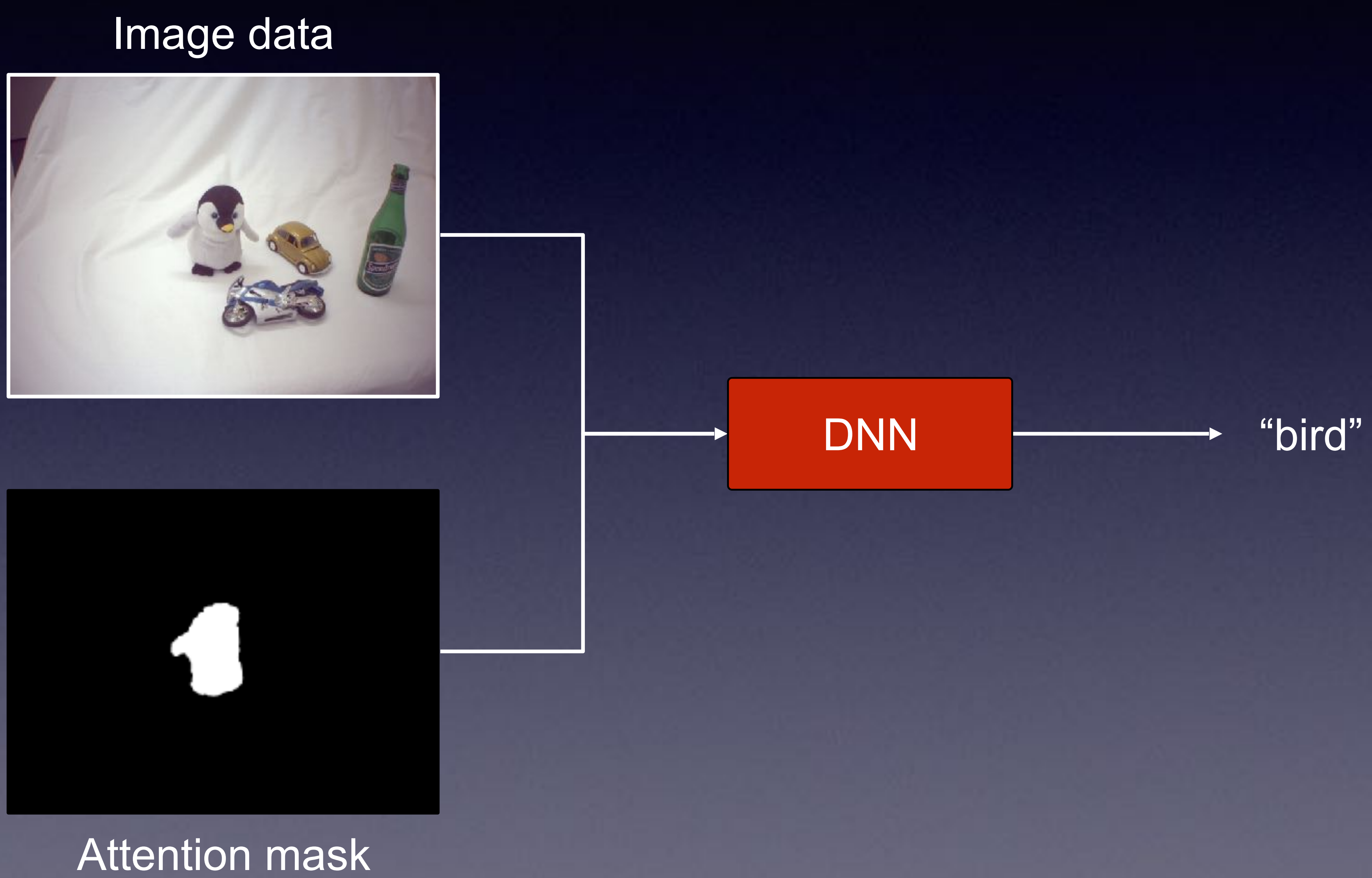
Attention for robots

Left camera

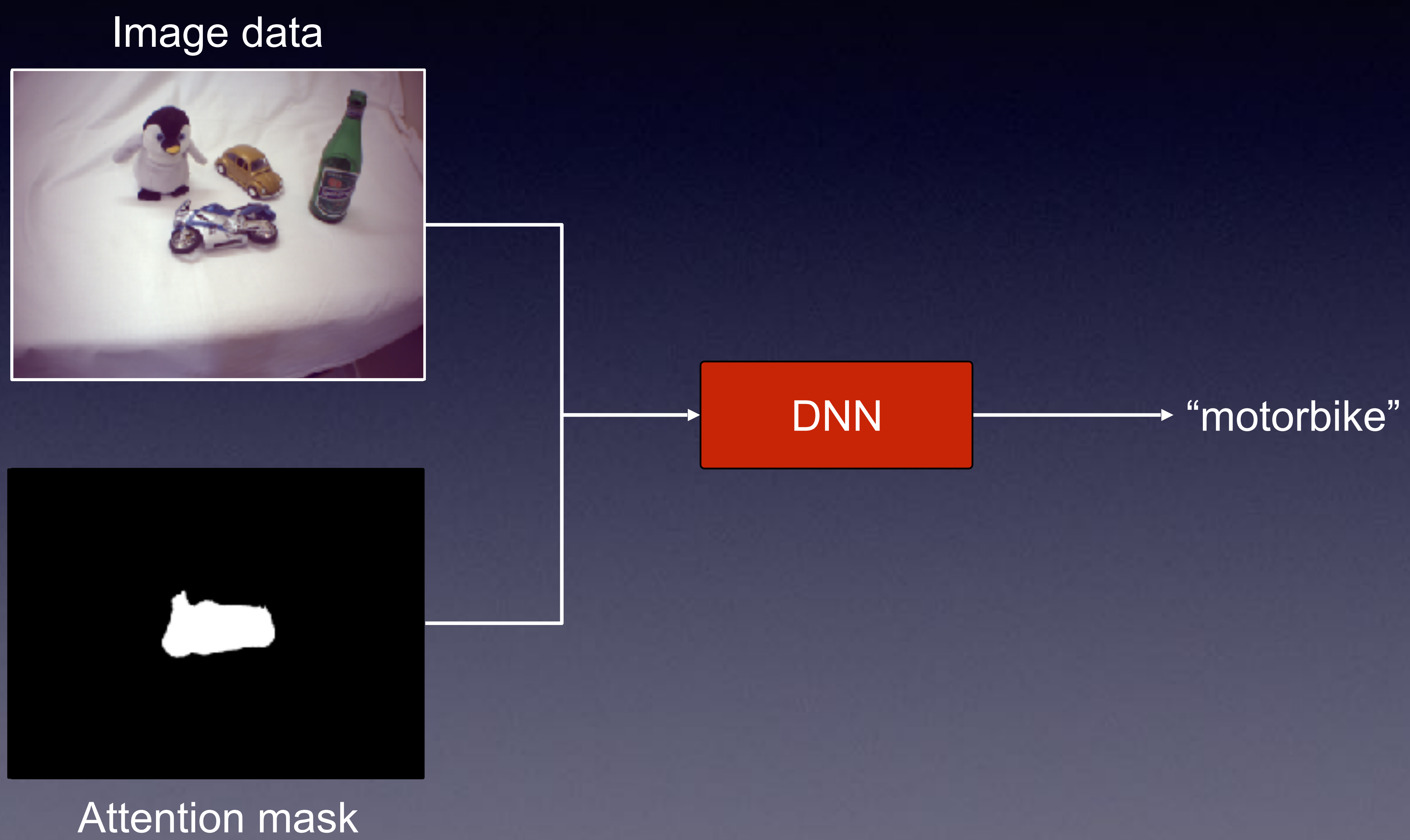
Right camera

- What is the robot looking at?
  This is an ill-posed classification problem.
  Attention is the missing constraint

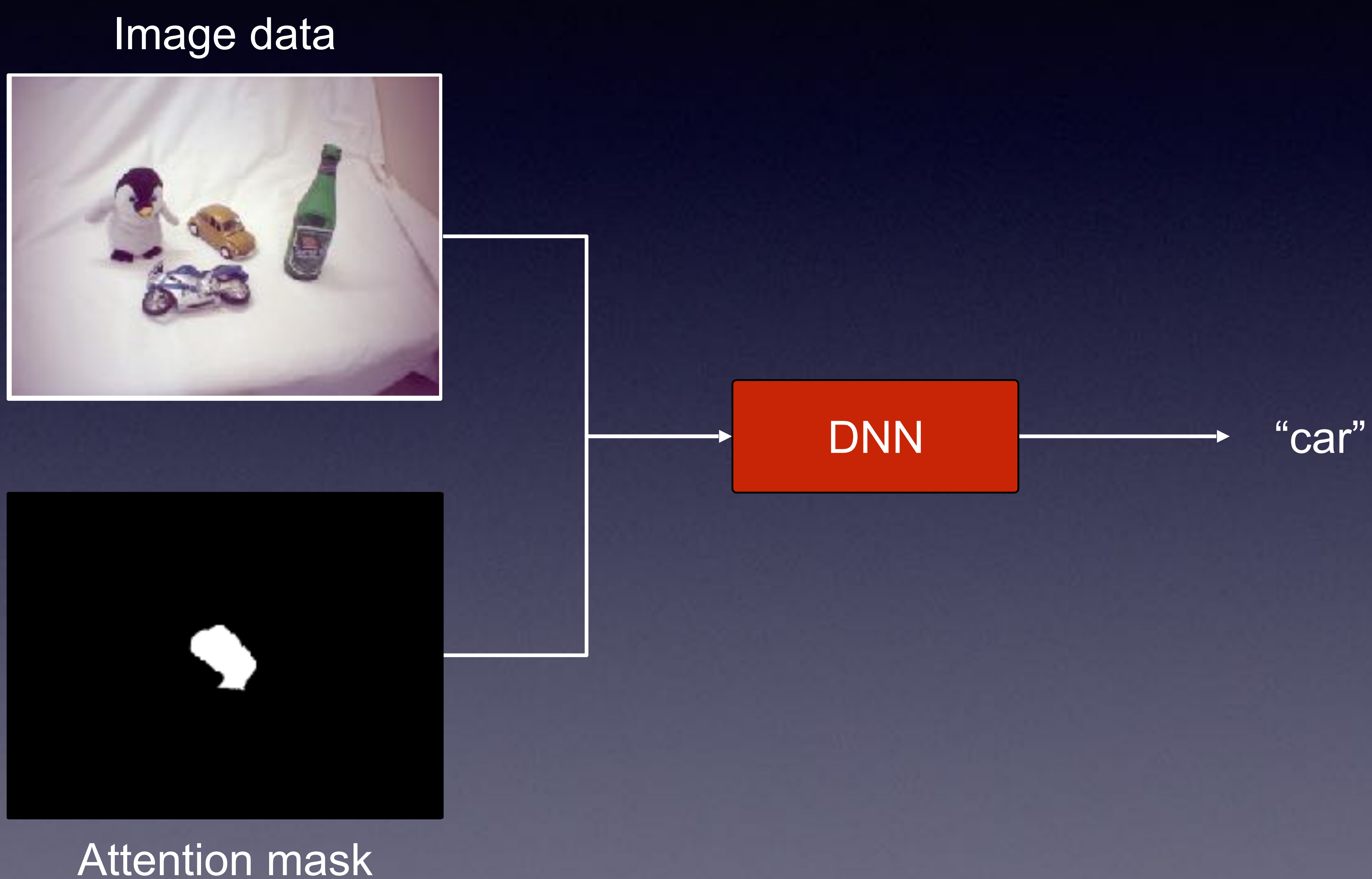
Attentional masking



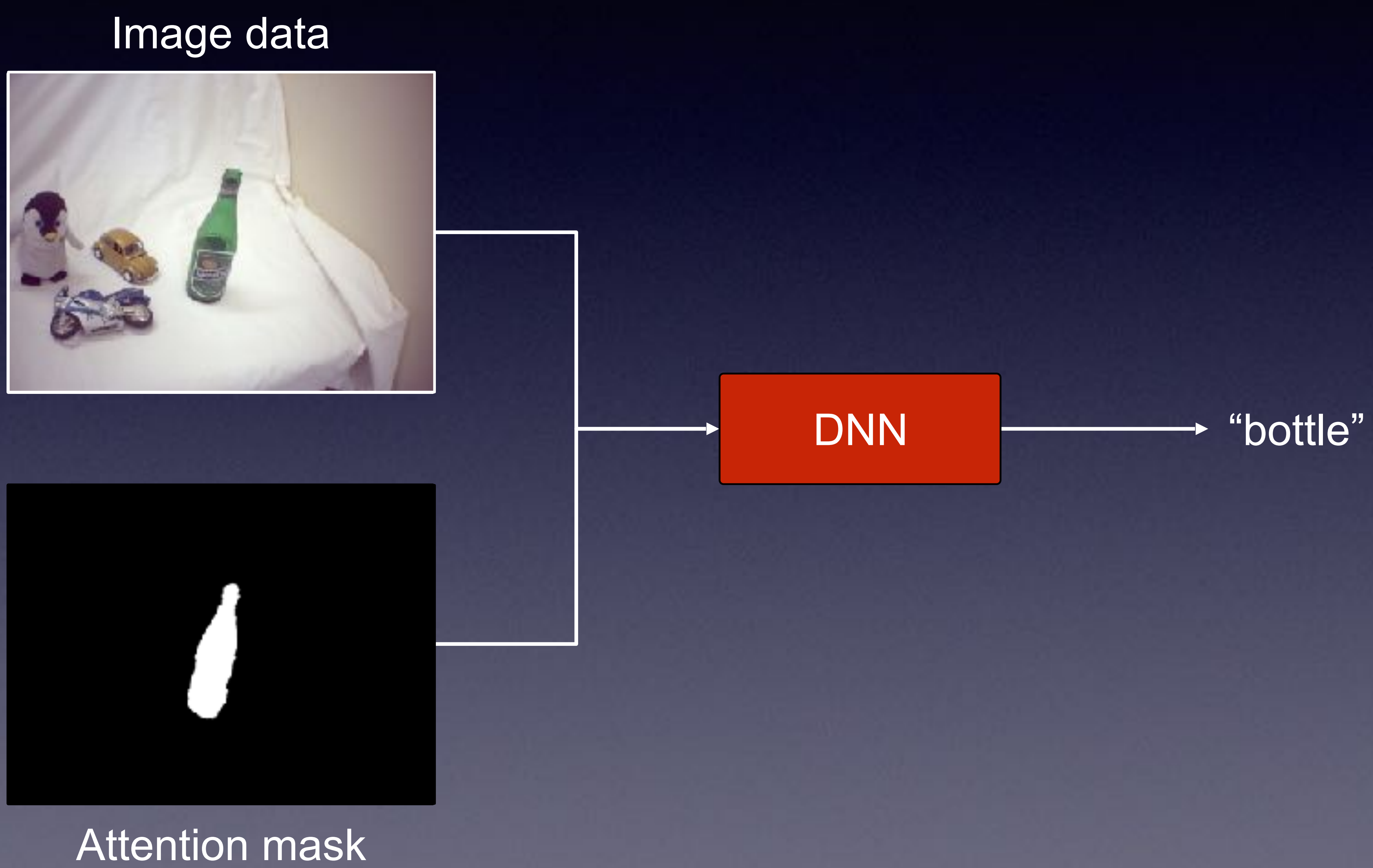
Attentional masking



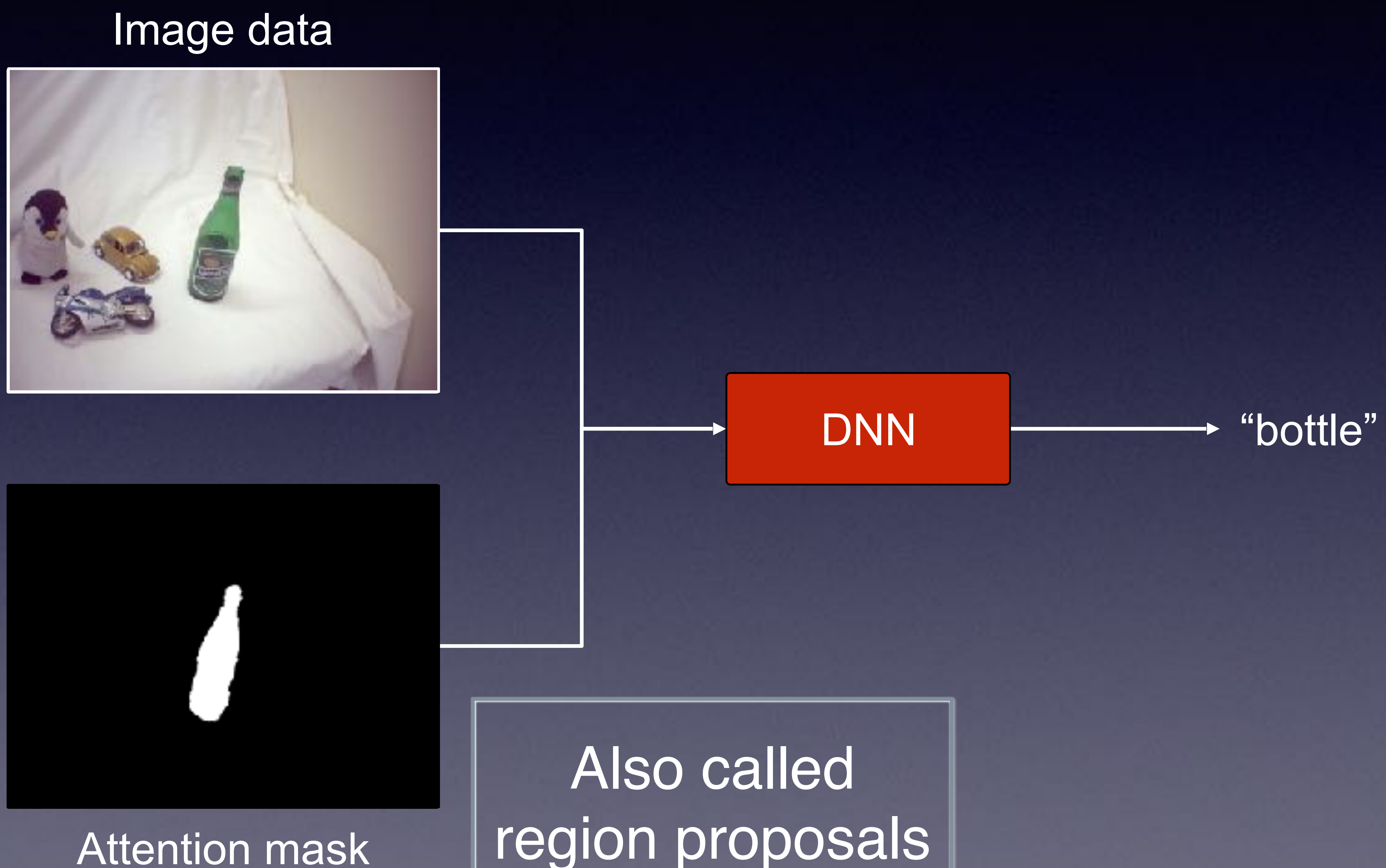
Attentional masking



Attentional masking



Attentional masking

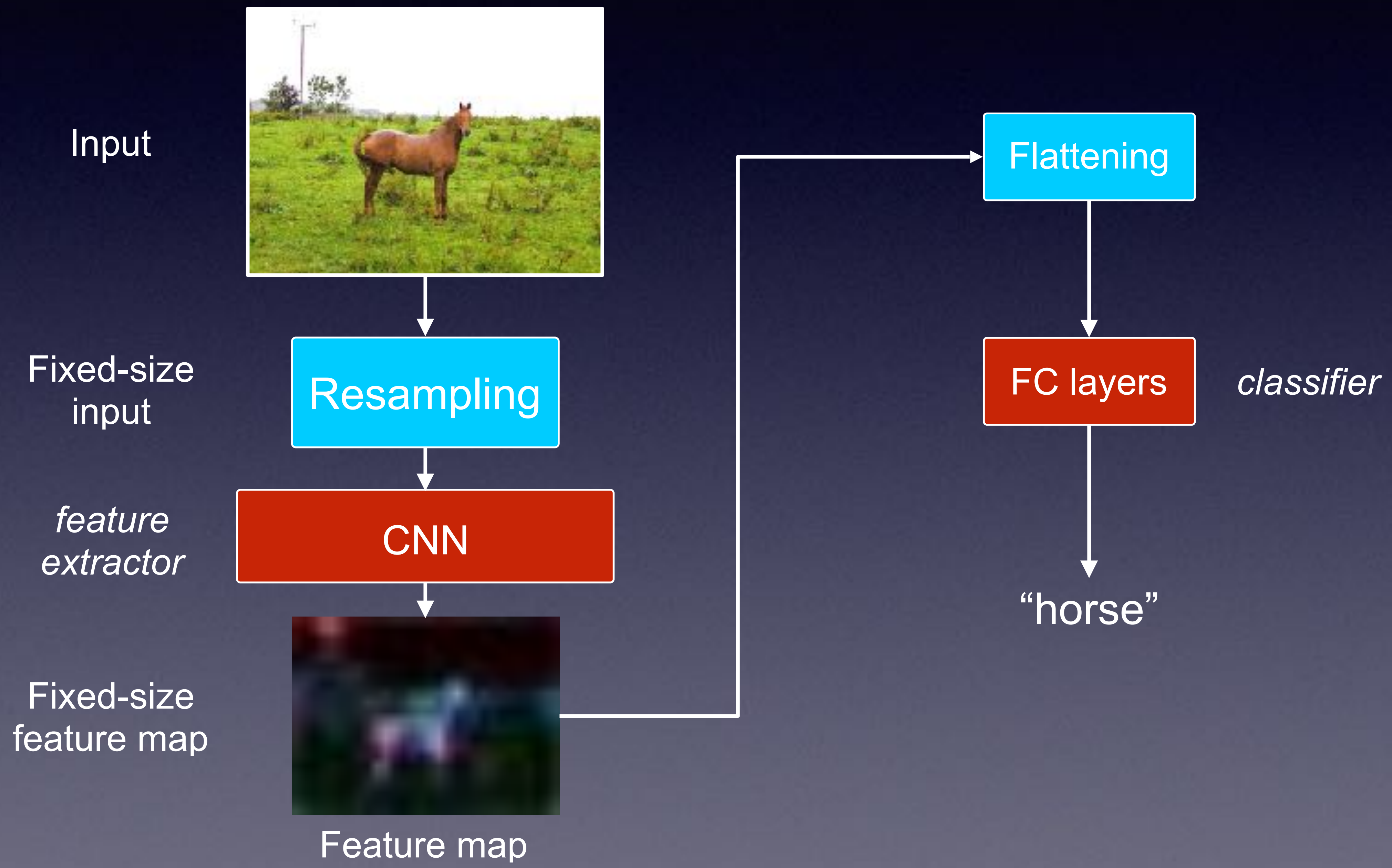


Also called region proposals

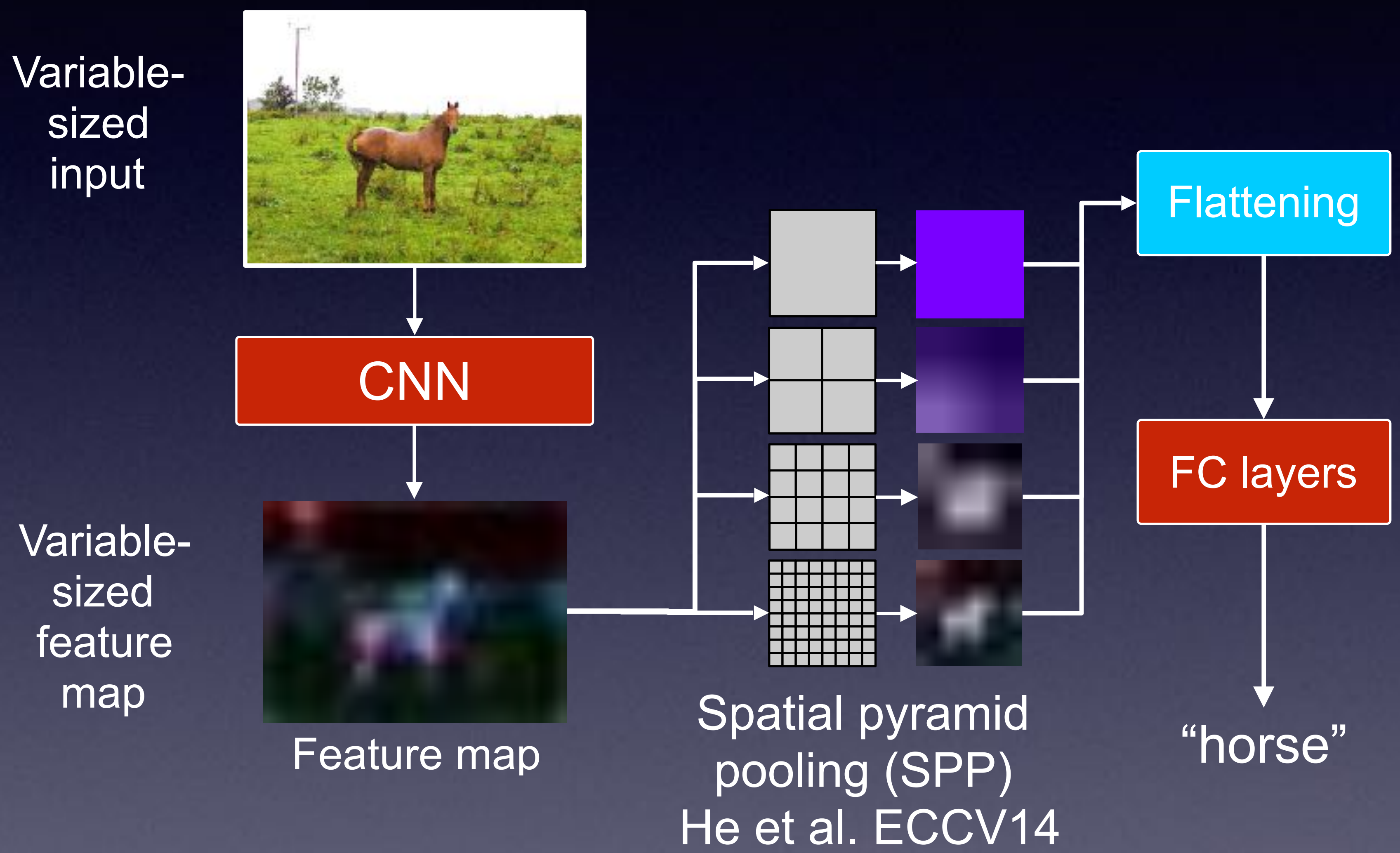
Attentional masking

- Q0: How can we apply attention in a CNN framework?
- A0: Select a region of interest bounding box (a rectangle)
- Then show the network only the bounding box contents
- Q1: Can we do better?

Classical classification network

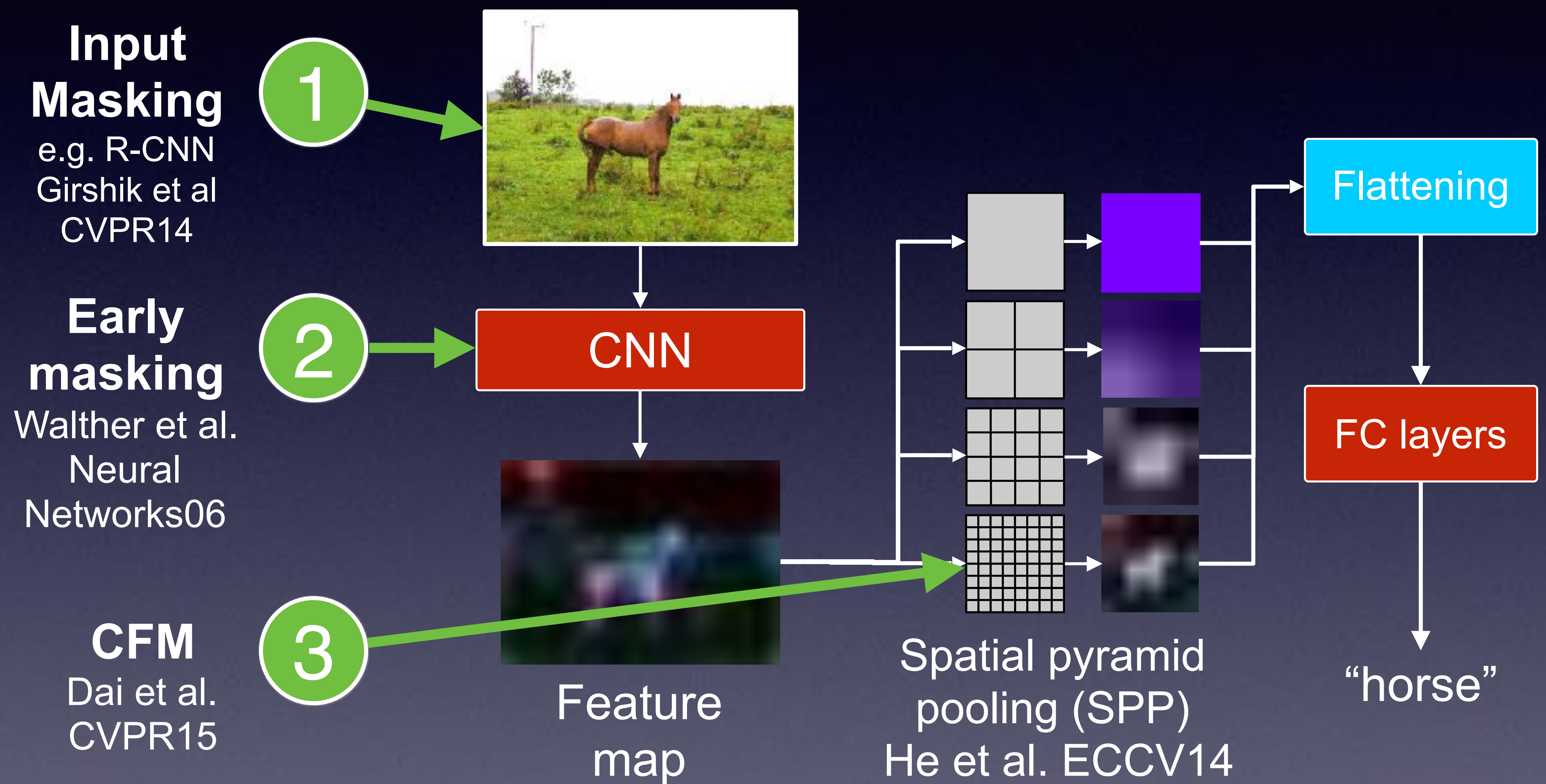


Modified classification network

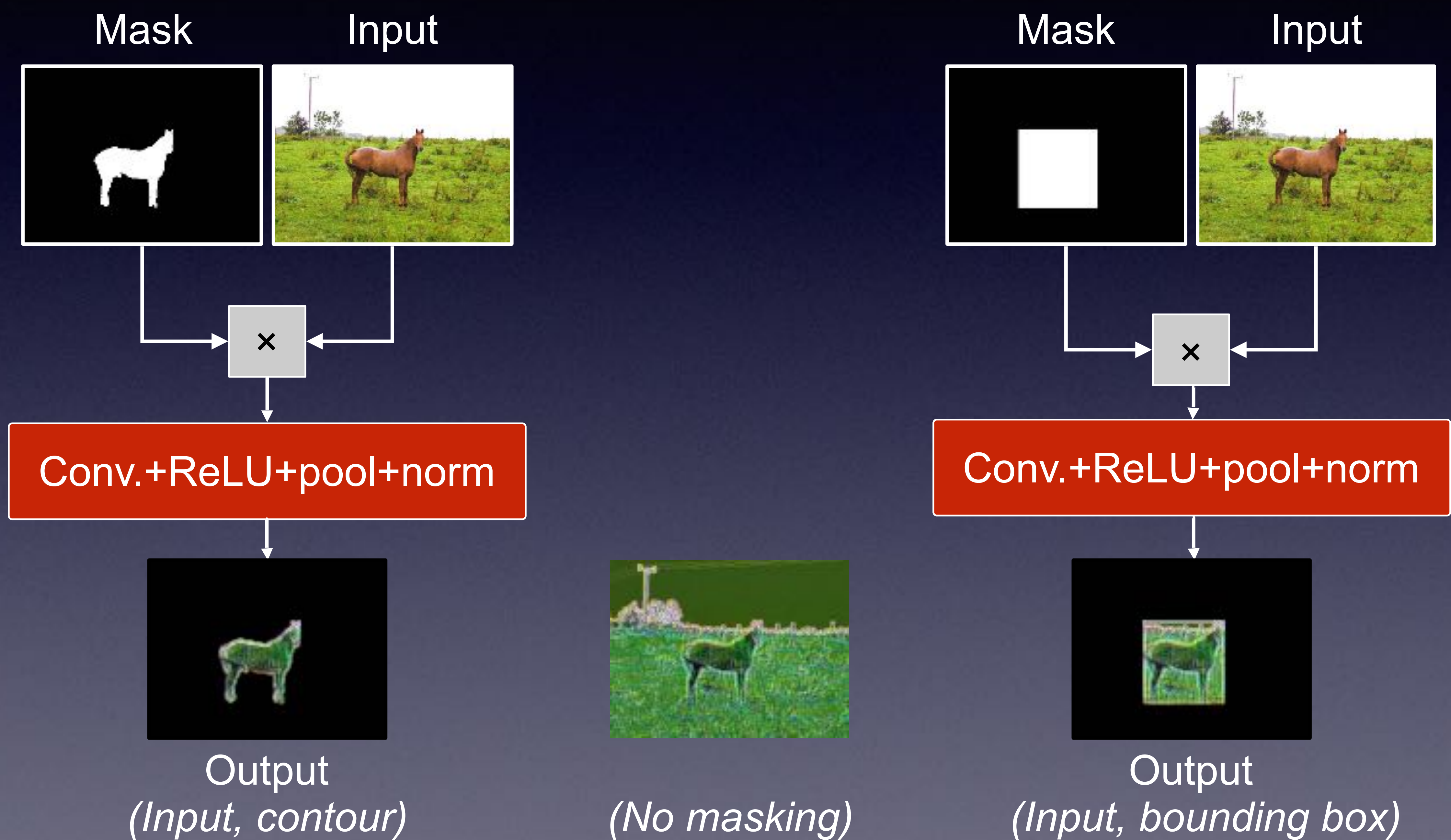


(Structure from Convolutional feature masking, Dai et al. CVPR'15)

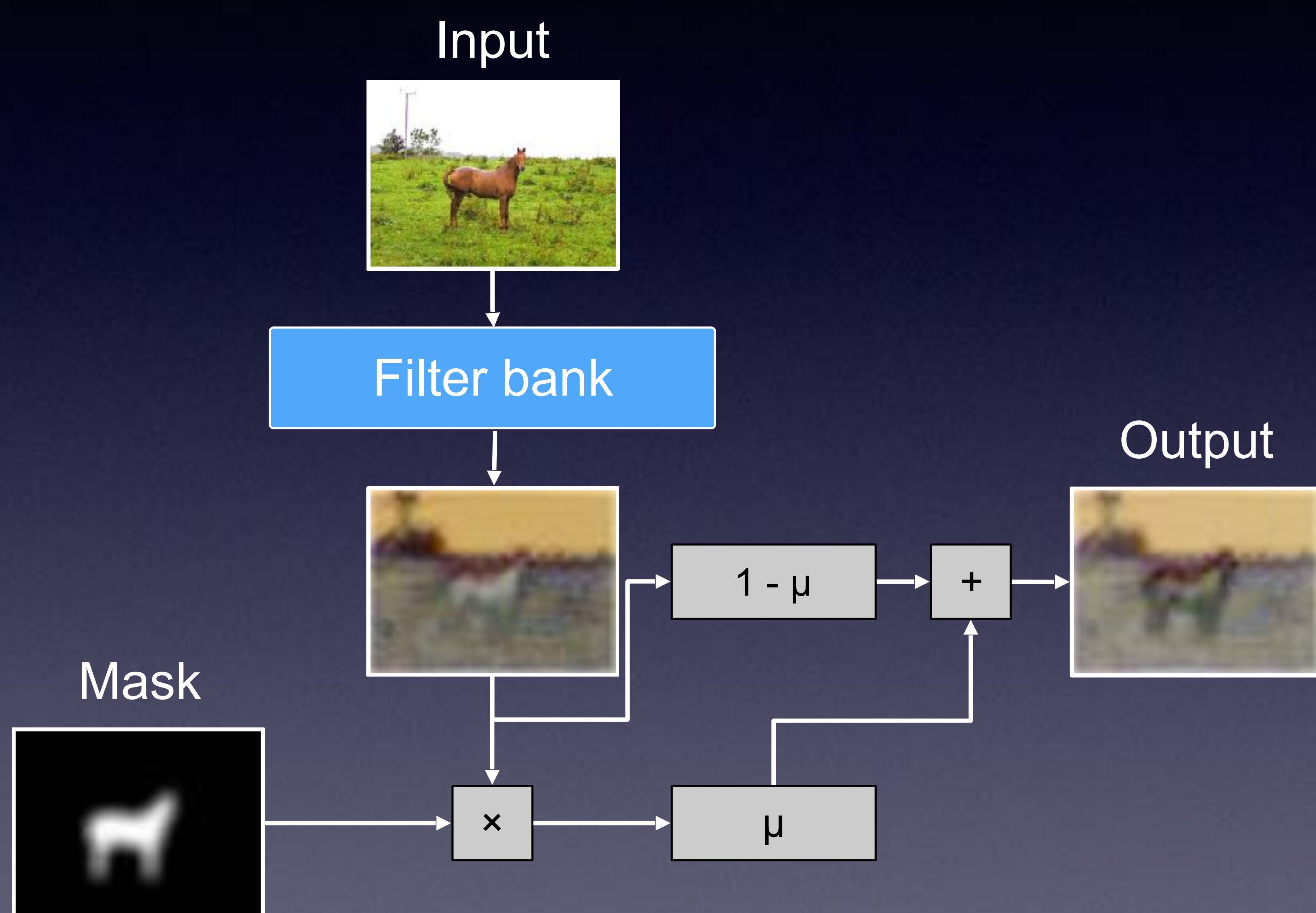
Where to apply attentional modulation?



Input masking

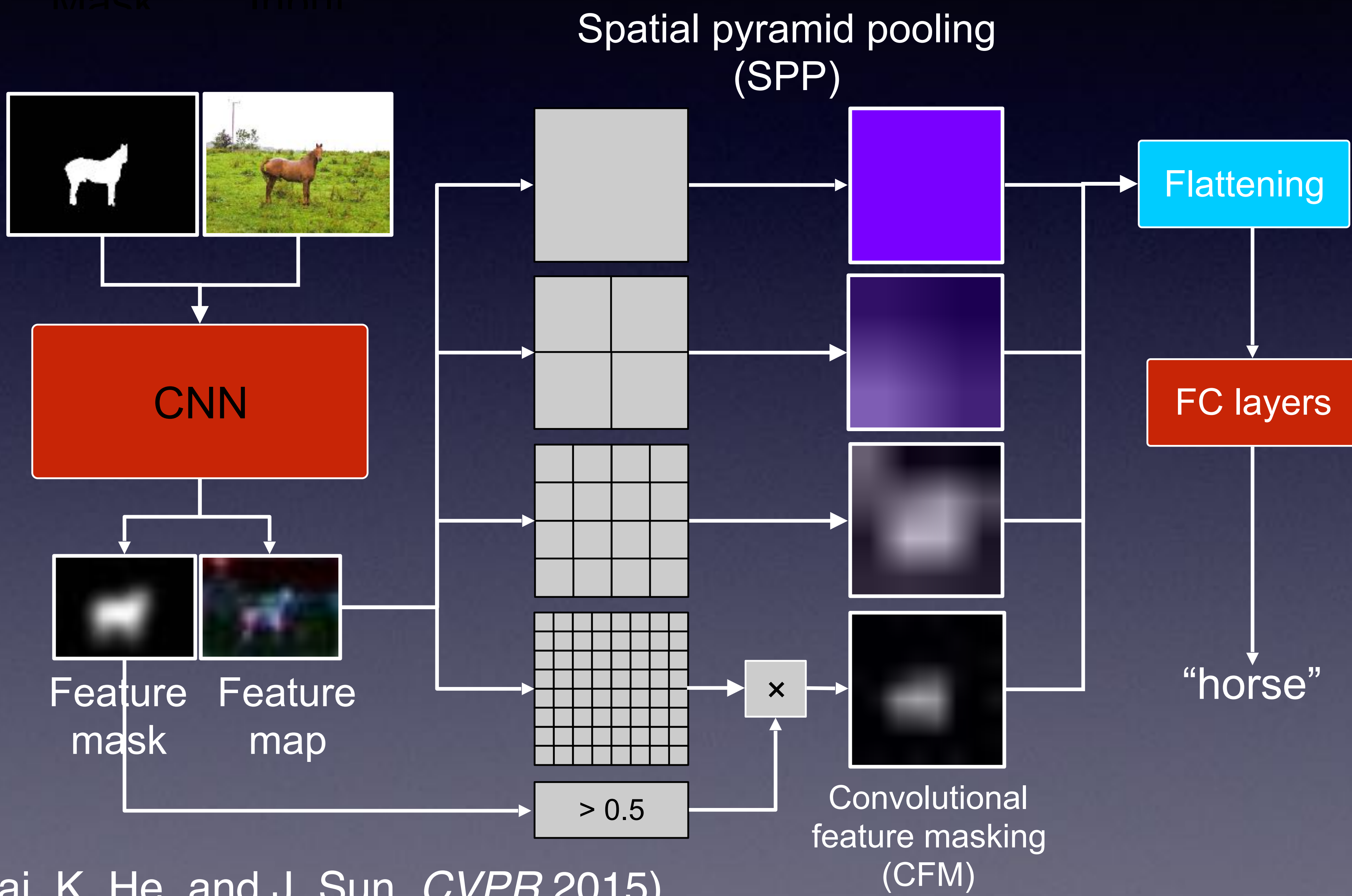


Mask and blend



(Walther & Koch, Neural Networks 2006)

Convolutional Feature Masking



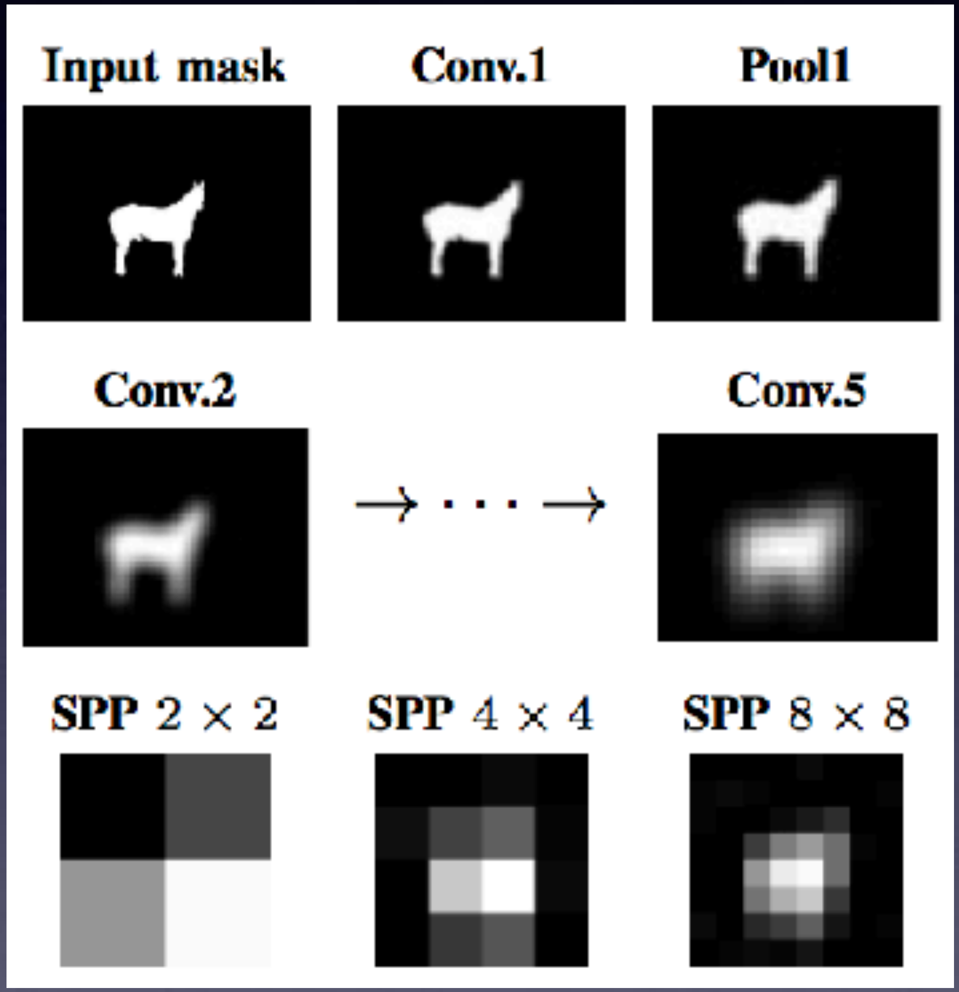
(J. Dai, K. He, and J. Sun, *CVPR* 2015)

Multi-layer, continuous valued masking (MC-CFM)

- Idea: Apply mask at all convolutional levels, and only to some degree, like Walther and Koch did:

$$R(x,y) = \mu R_M(x,y) + (1-\mu)R_0(x,y)$$

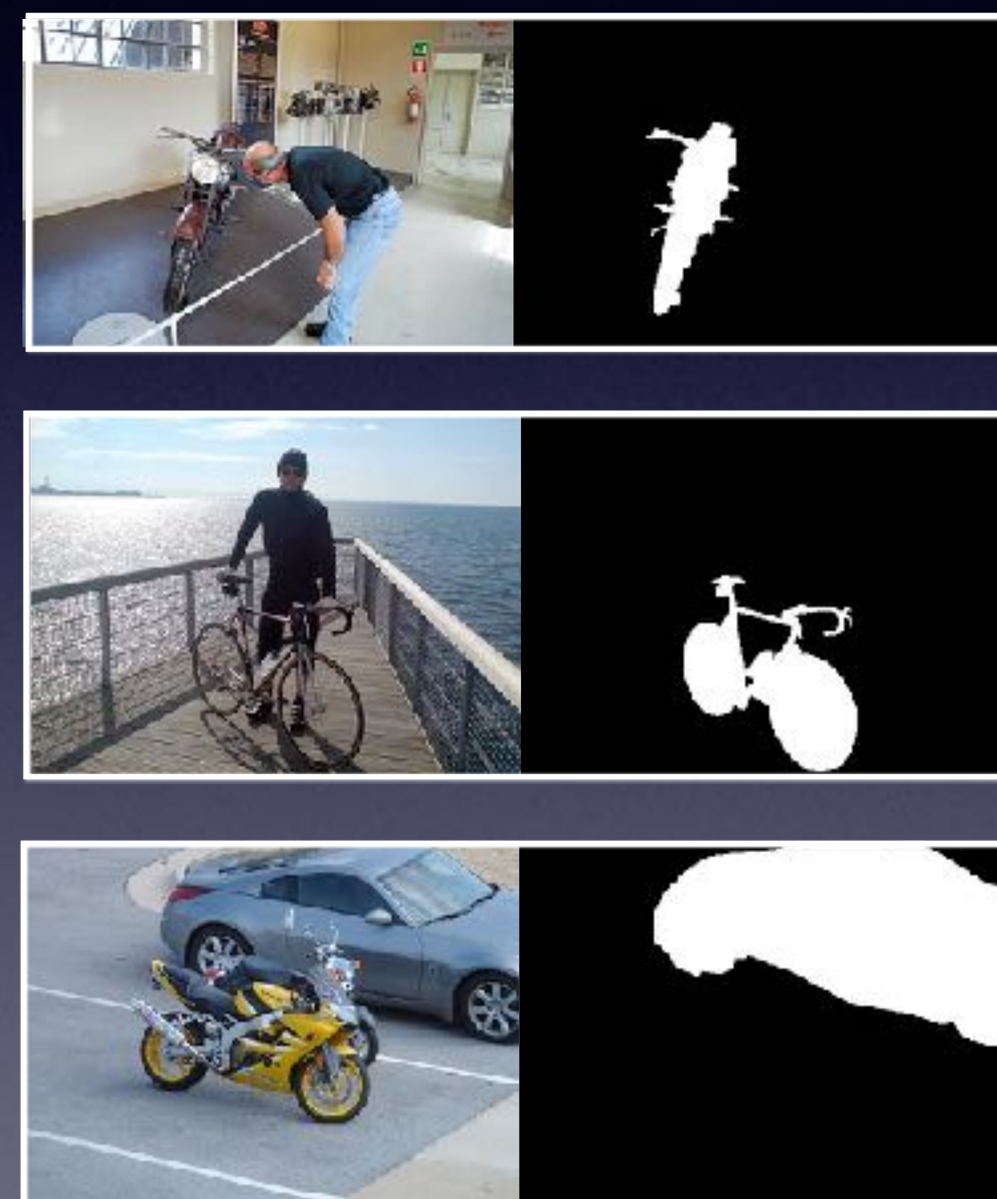
- We now get another parameter μ, one for each layer in the network.



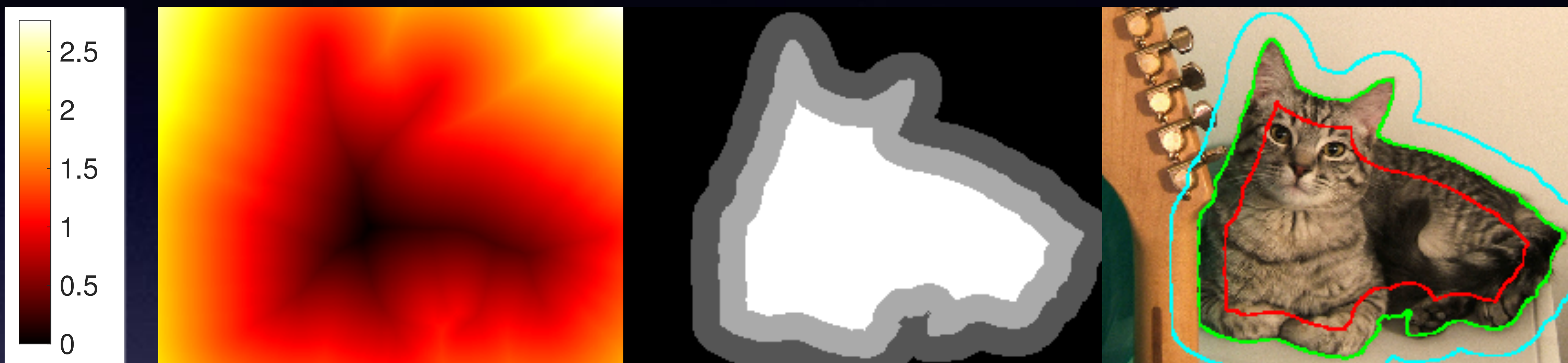
Training and Evaluation

- We want to decouple accuracy of attention masks from the performance of the masking procedures.
- For this reason we want to start with ”perfect” masks that we can distort in a controlled way.

PASCAL-S masks



Mask perturbation



t_d=0.75

t_d=1.0

t_d=1.25

- To test sensitivity to mask errors, we grow and shrink the mask using the signed distance transform

Training and Evaluation

- Main network is VGG-F
  (similar to AlexNet, 5 conv, 3 fc)
  Pre-trained on ImageNet

PASCAL-S example

PASCAL-Context example

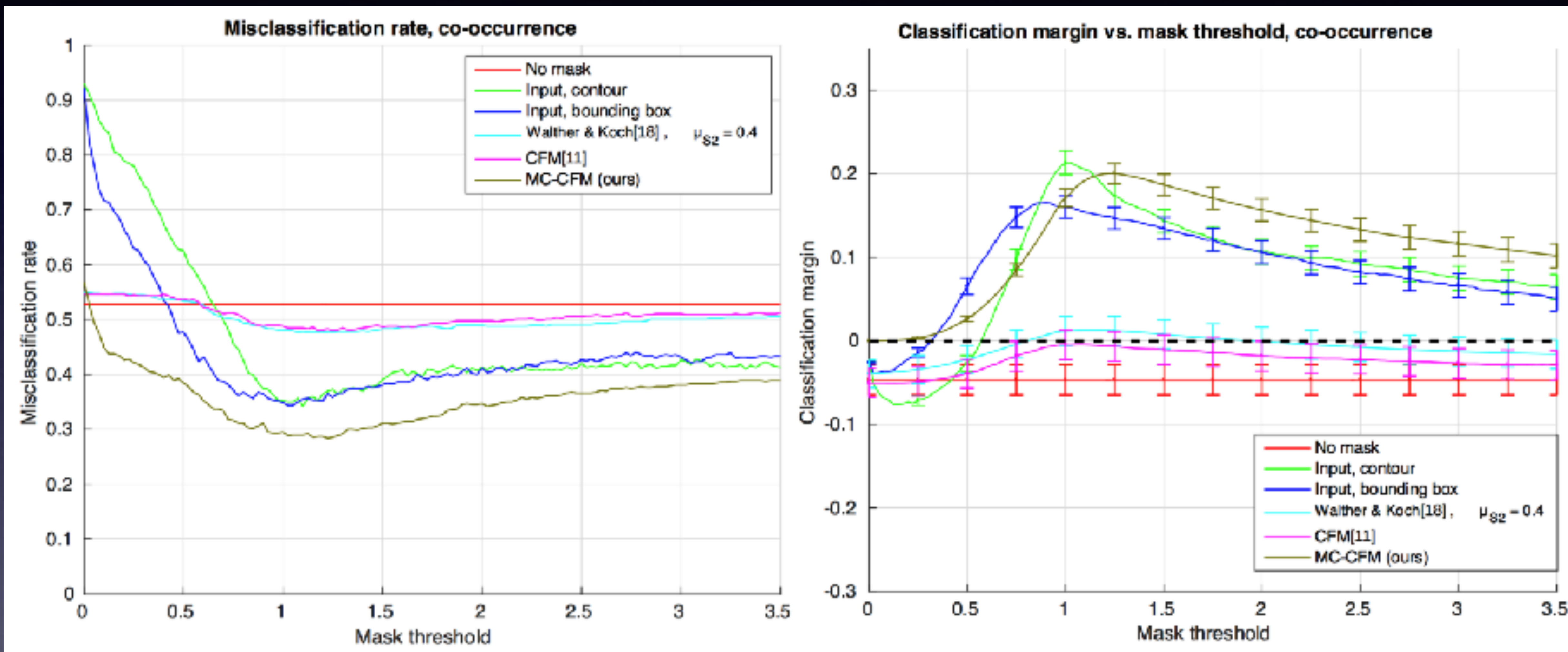
- A subset of 11 PASCAL-S classes
  with high co-occurrence are used:
  horse, cup, chair, dog, bird, tree, bottle,
  motorbike, bicycle, car and person

- Classification FC layers are trained
  on PASCAL-Context (excluding masks, and PASCAL-S)

- Mask blend weights are trained using NLSQ on top 1 % of co-occurrences (”crowded” scenes)

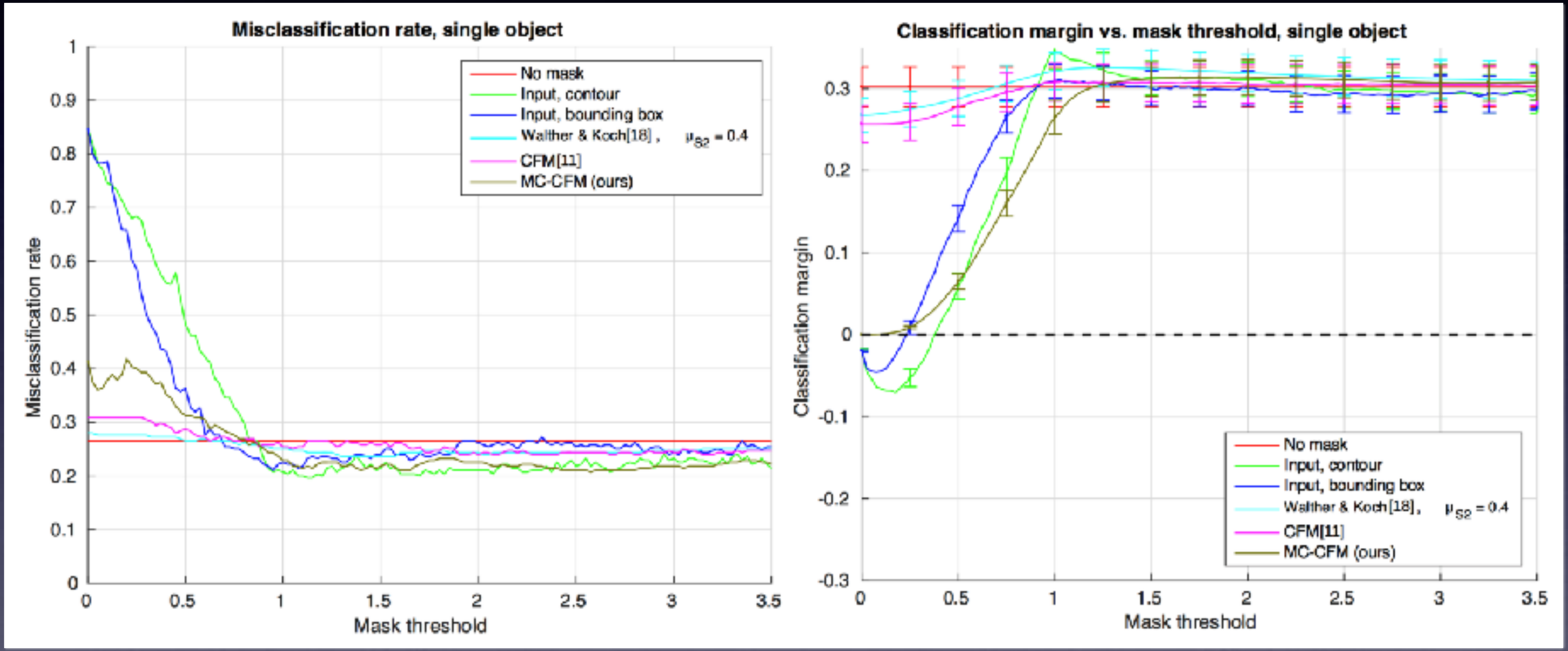
- Evaluation uses the remainder of PASCAL-S containing our classes.

Attentional masking under co-occurrence



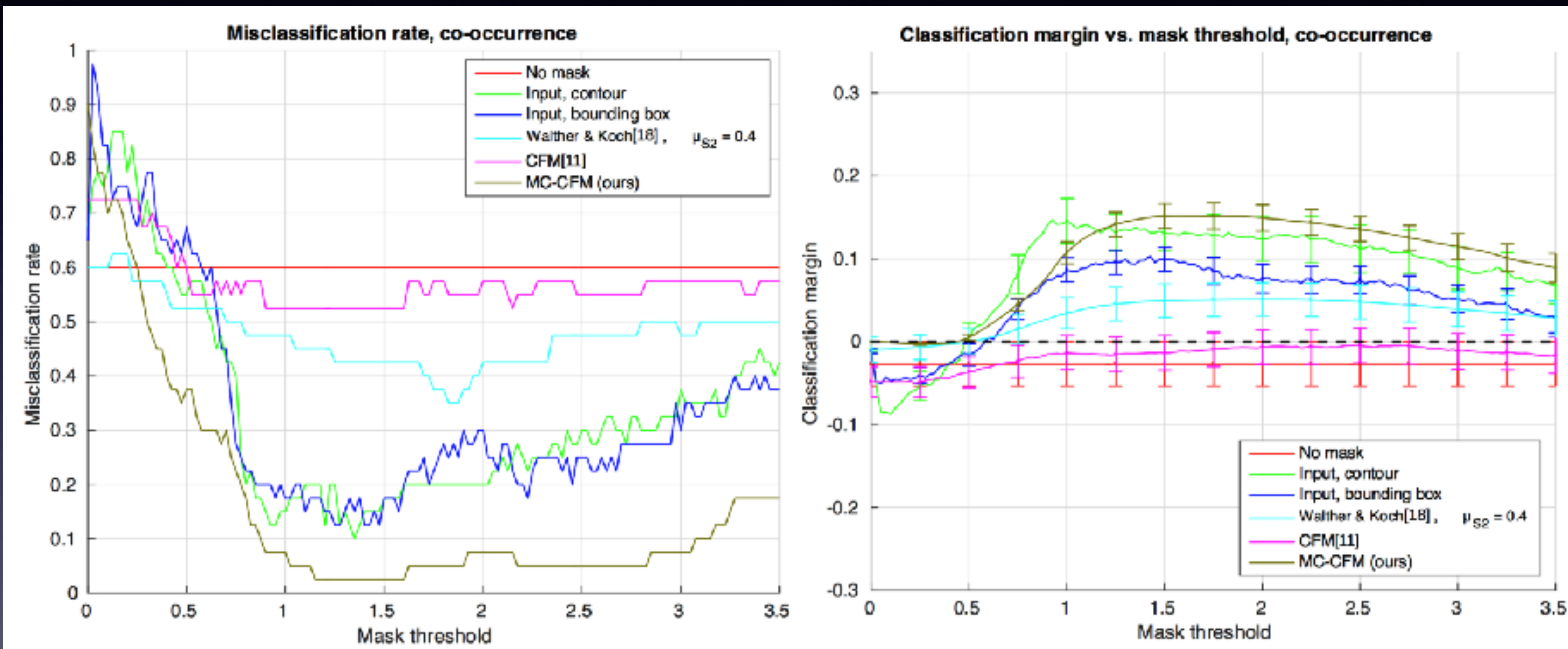
Results on PASCAL-S dataset, co-occurrence

Attentional masking w/o co-occurrence



Results on PASCAL-S dataset, single object

Results on robot with co-occurrence



Results on robot with co-occurrence

Summary

- The choice of attentional masking matters!
- By using multi-level masking we can reduce error rates substantially compared to using input masking
- On our robot 4% vs 17%
- On PASCAL-S 28% vs 35%