Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros, Abhinav Gupta

Presented by Maheen Rashid for ECS 289G

Motivation

- How can we scale to billions rather than millions of images?
    - Imagenet trained on ~1.2 million images
- Unsupervised learning
    - Problem - What should be represented?

Inspiration - Context

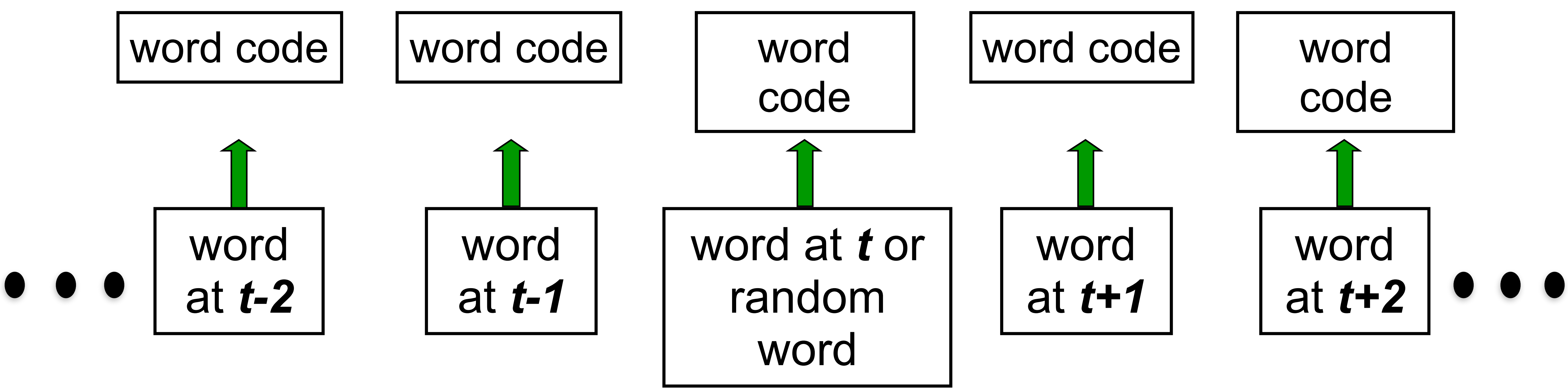
- Similar words appear in similar contexts
- Learn to relate a given word to its surrounding words
- Context prediction becomes a ‘pretext’ task

A simple way to learn feature vectors for words (Collobert and Weston, 2008)

• • • | word at ***t-2*** | word at ***t-1*** | word at ***t*** or random word | word at ***t+1*** | word at ***t+2*** | • • •

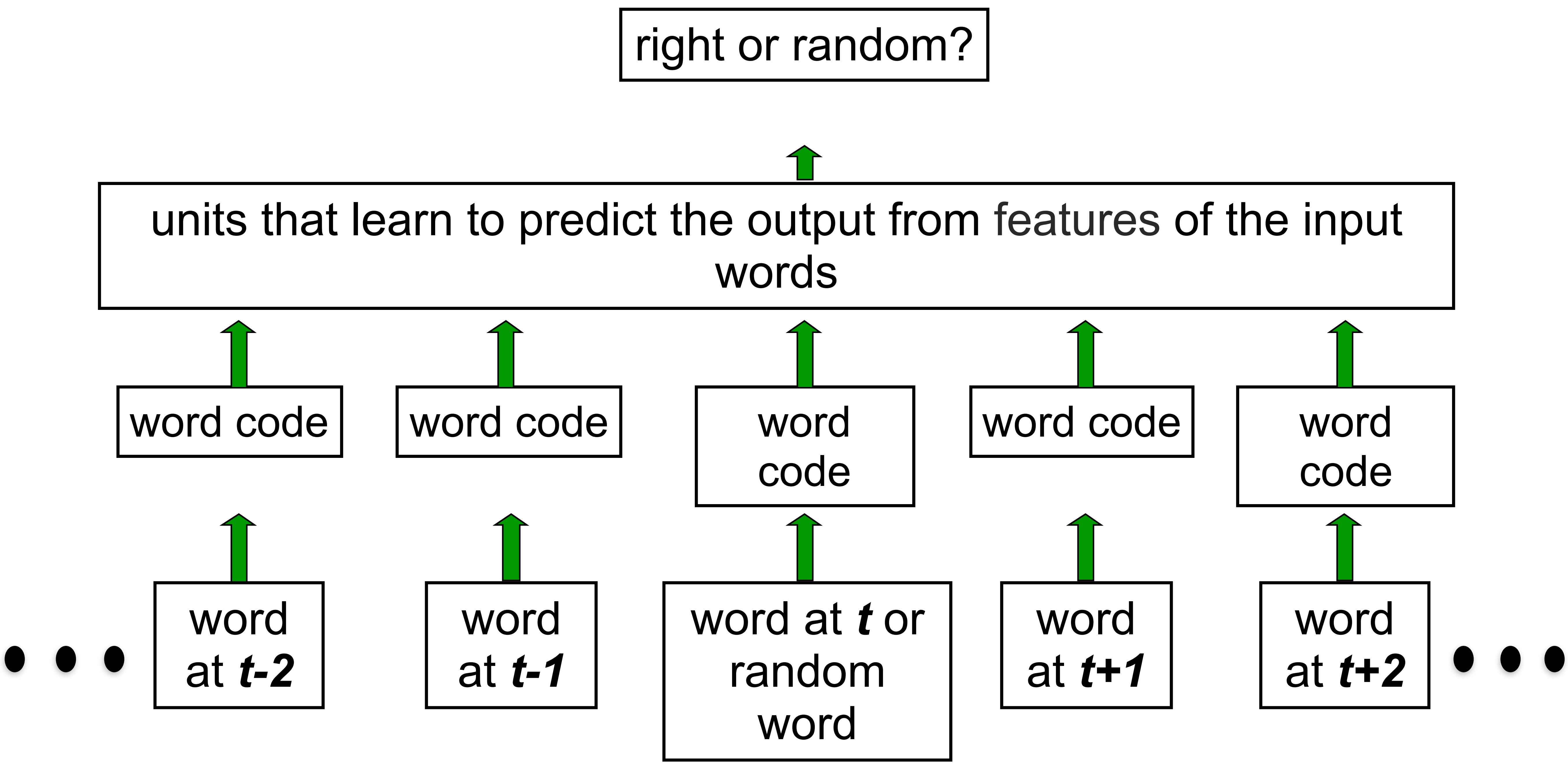
Slide from Geoff Hinton

A simple way to learn feature vectors for words (Collobert and Weston, 2008)



Slide from Geoff Hinton

A simple way to learn feature vectors for words (Collobert and Weston, 2008)



Slide from Geoff Hinton

Right or Random for Images?

Slide from Carl Doersch

Right or Random for Images?

Slide from Carl Doersch

A

B

Slide from Carl Doersch

A

B

Can you tell where B goes relative to A?

Slide from Carl Doersch

Answer:

Slide from Carl Doersch

Answer:

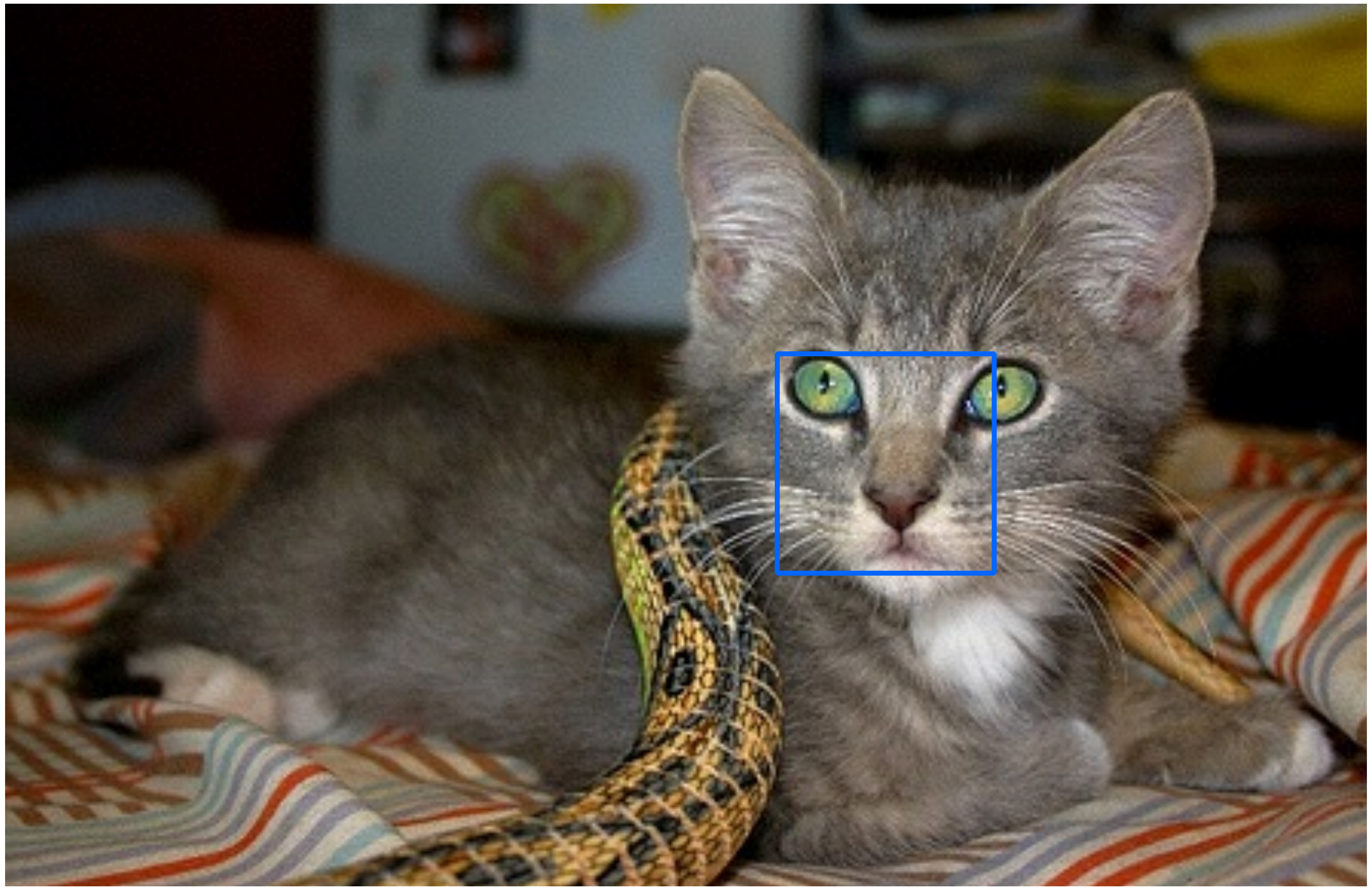
Doing this requires recognizing semantics!

Slide from Carl Doersch

Unlabeled training image

Slide from Carl Doersch

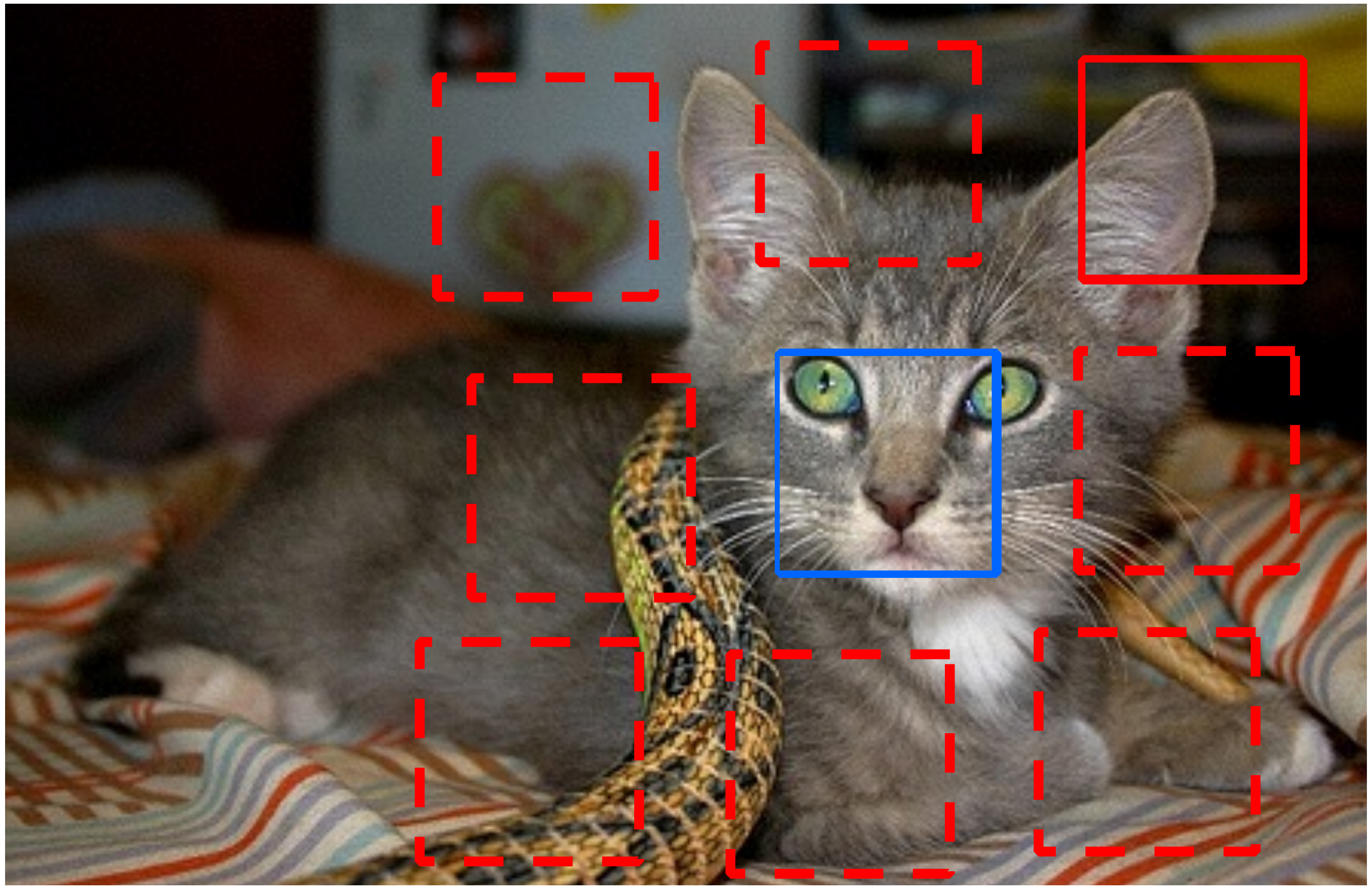
Unlabeled training image



Randomly Sample Patch

Slide from Carl Doersch

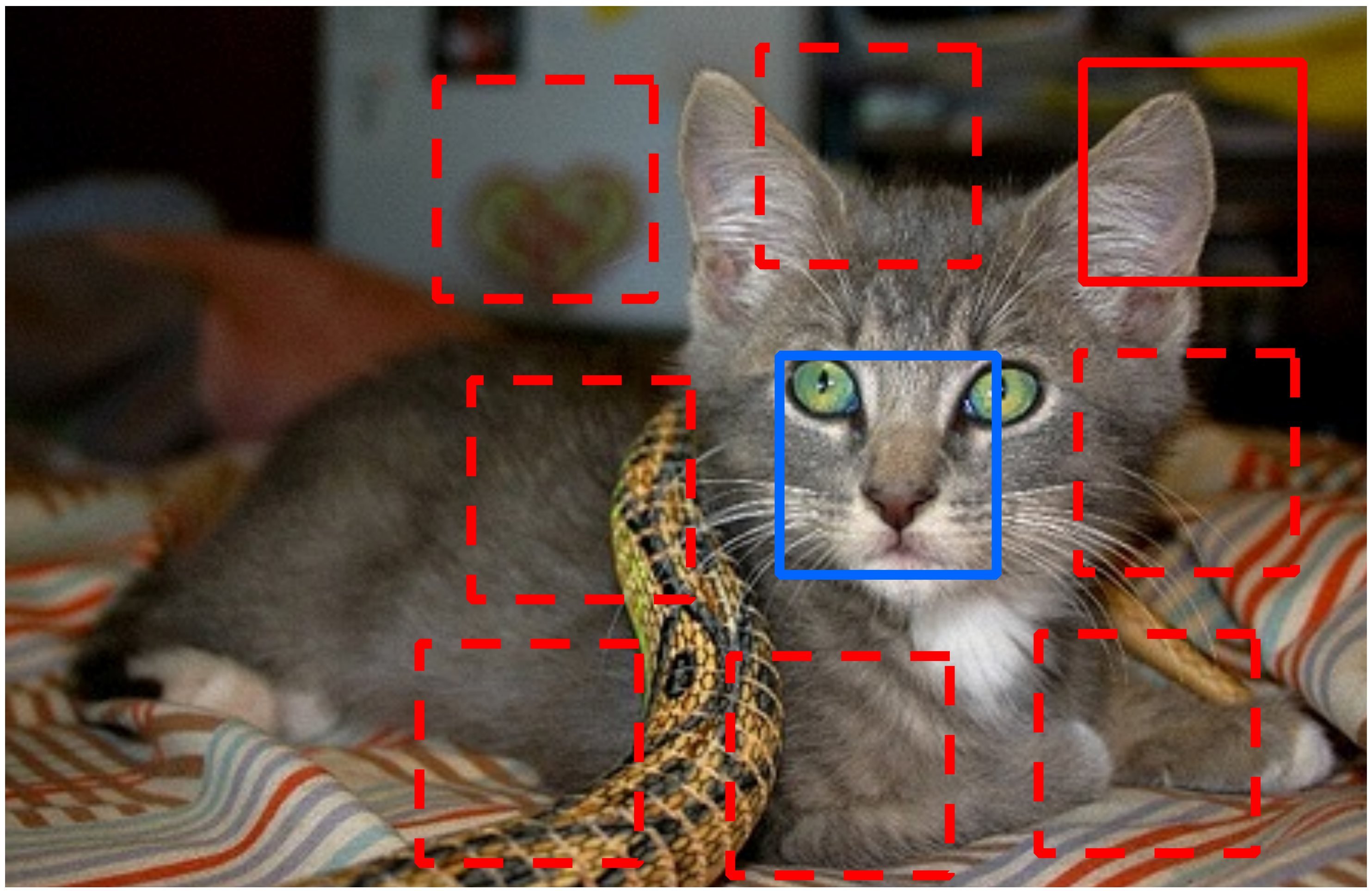
Unlabeled training image

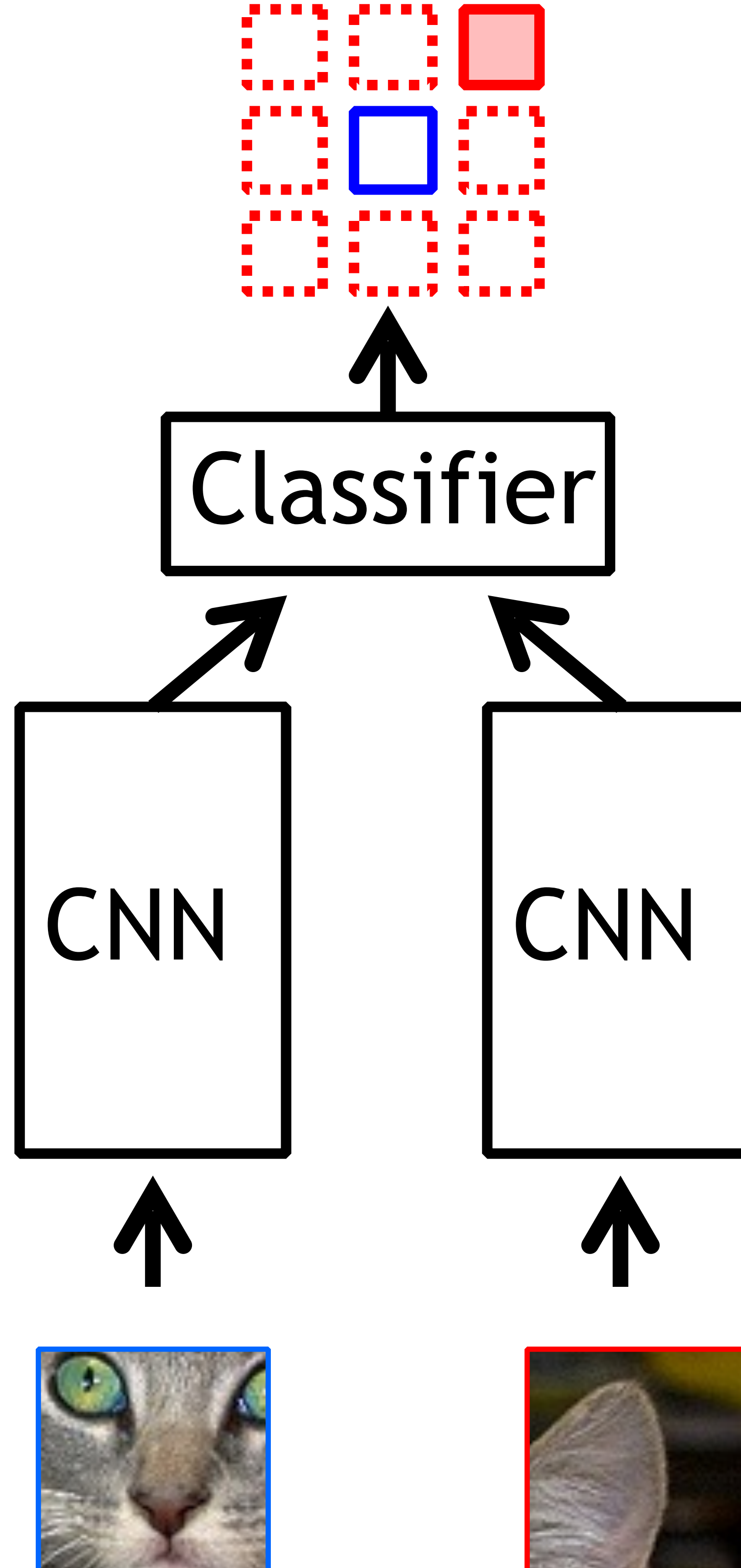


Randomly Sample Patch
Sample Second Patch

Slide from Carl Doersch

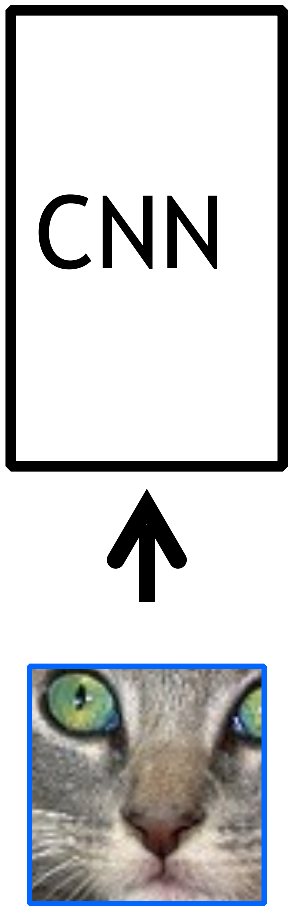
Unlabeled training image



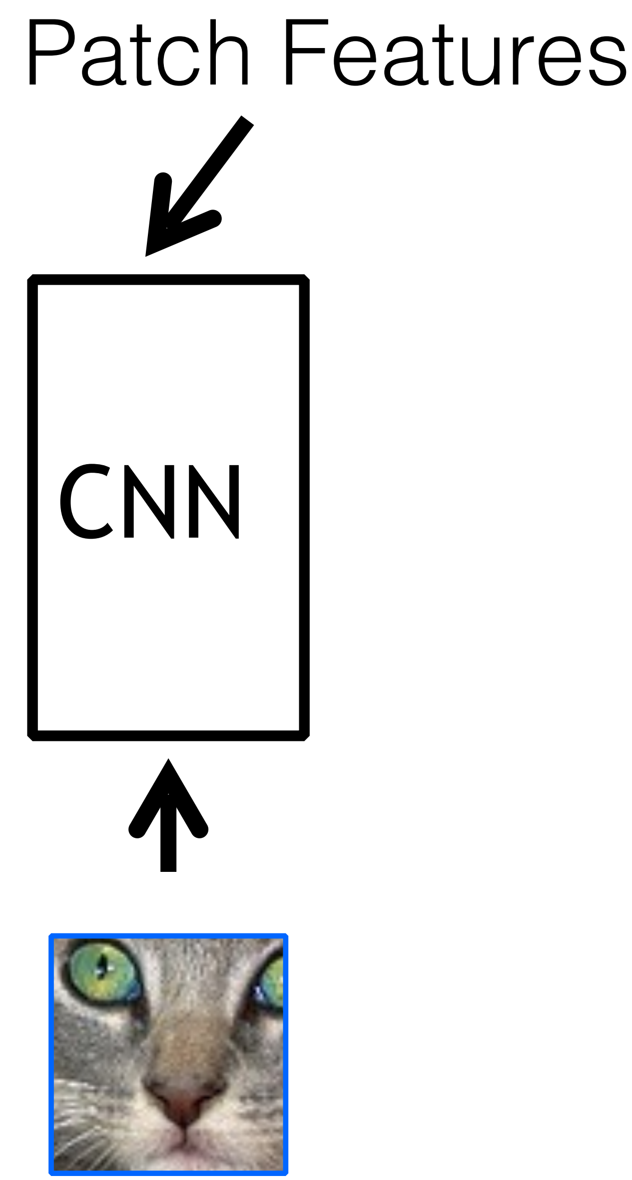


Train Deep Net to recover relative position

Slide from Carl Doersch



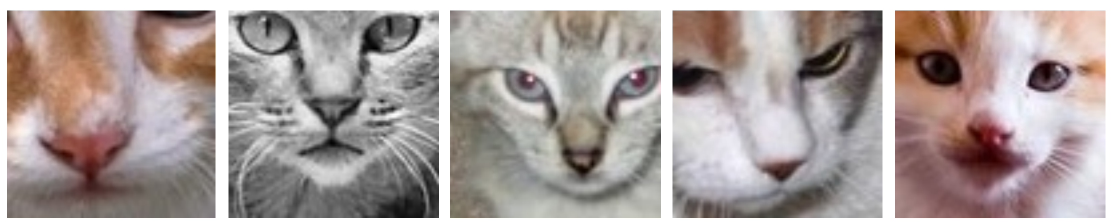
Slide from Carl Doersch



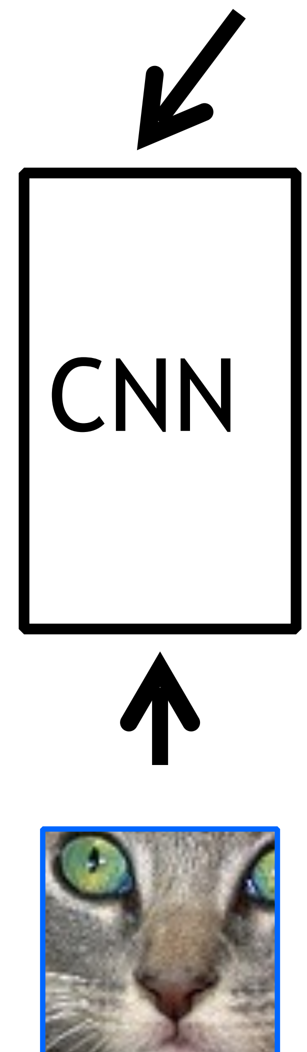
Slide from Carl Doersch

Input

Nearest Neighbors

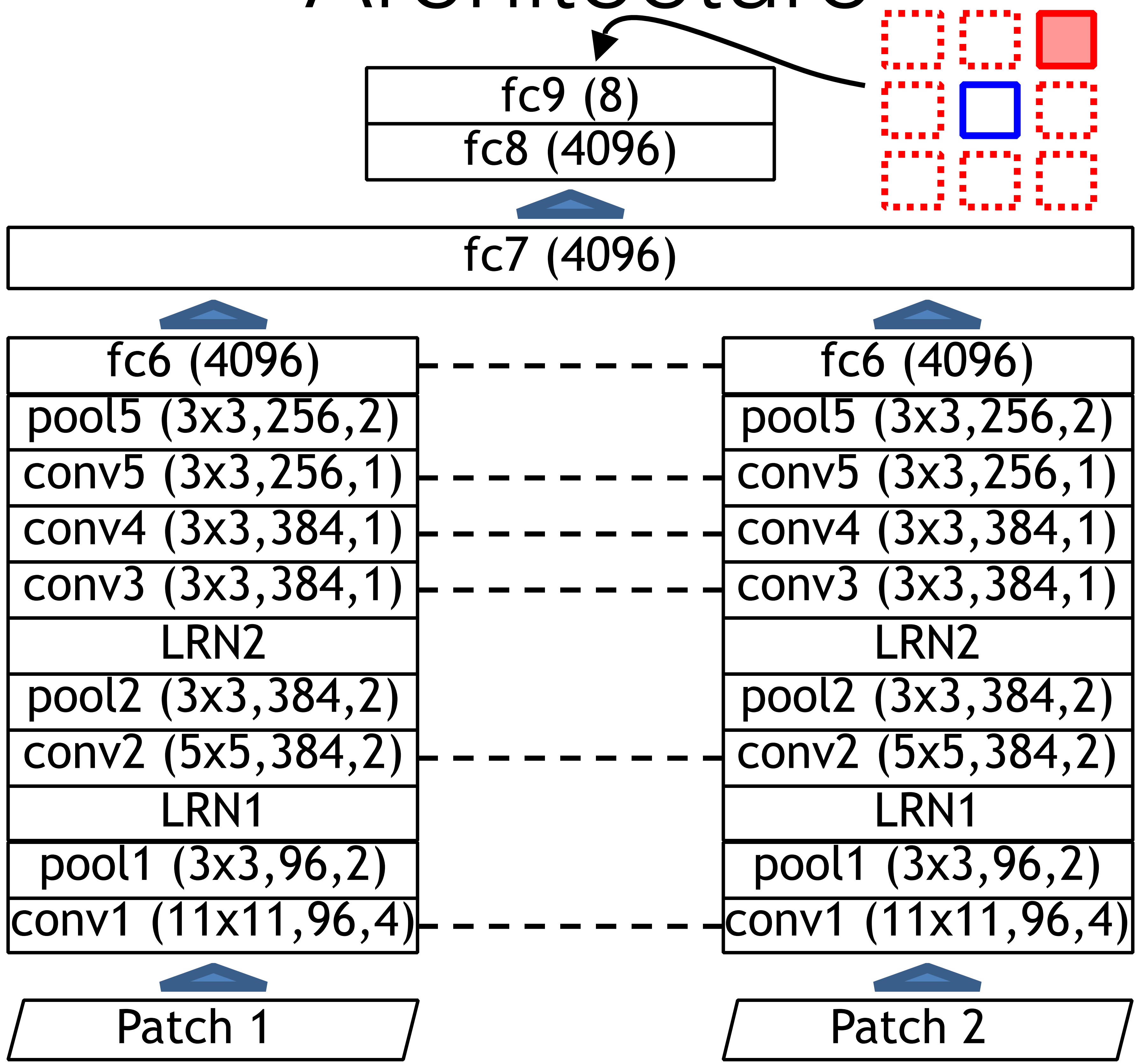


Patch Features



Slide from Carl Doersch

Architecture



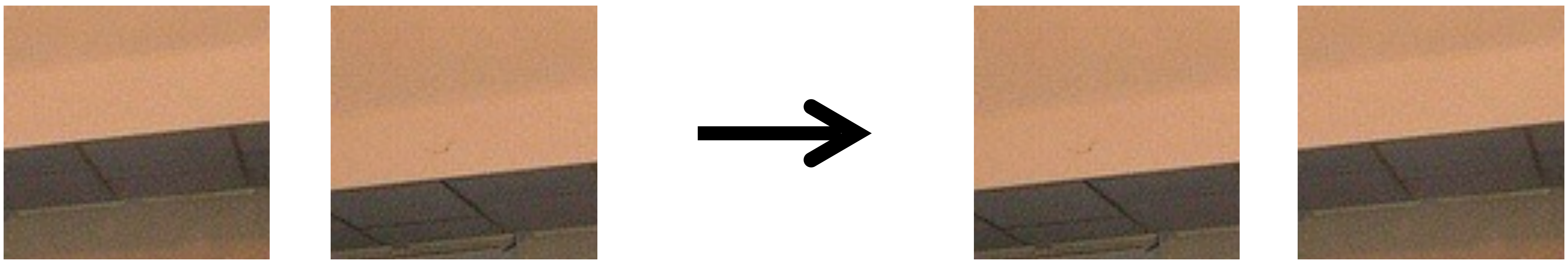
Slide from Carl Doersch

How to sample patches

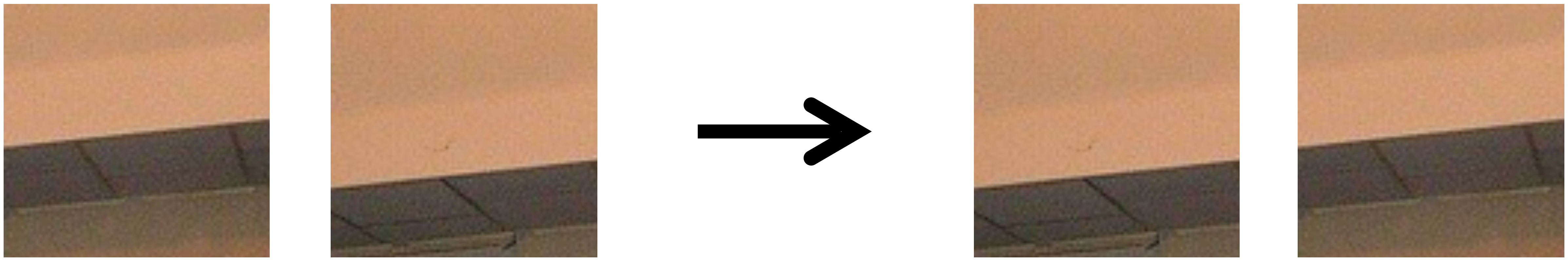
Slide from Carl Doersch

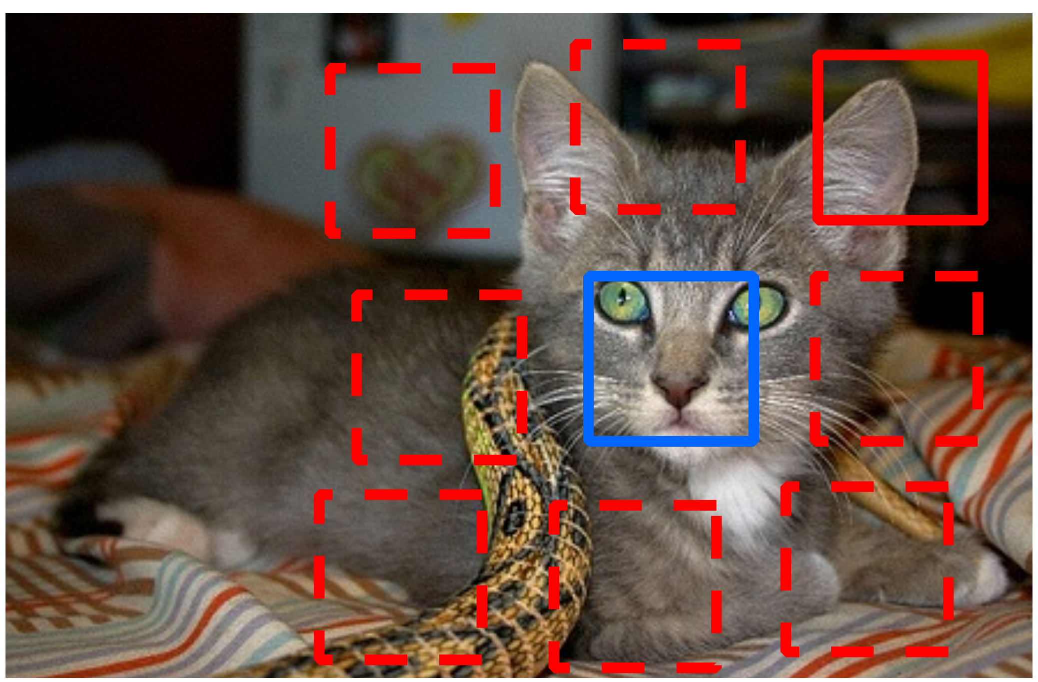
How to sample patches



Slide from Carl Doersch

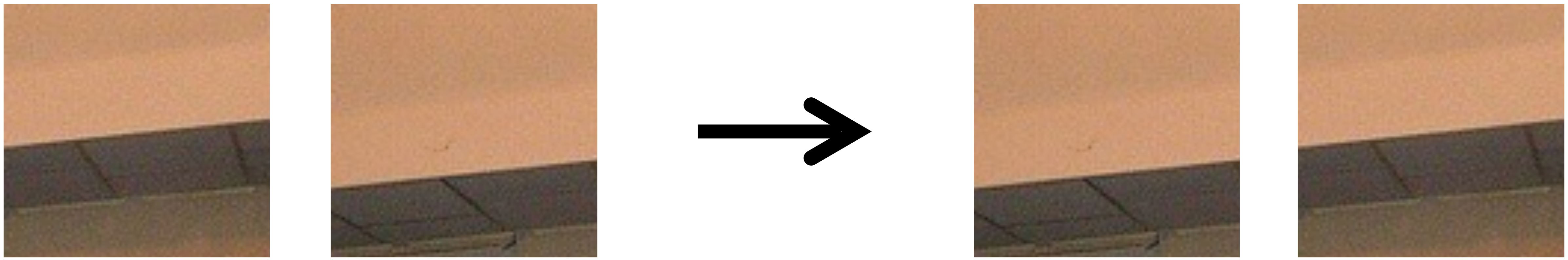
How to sample patches

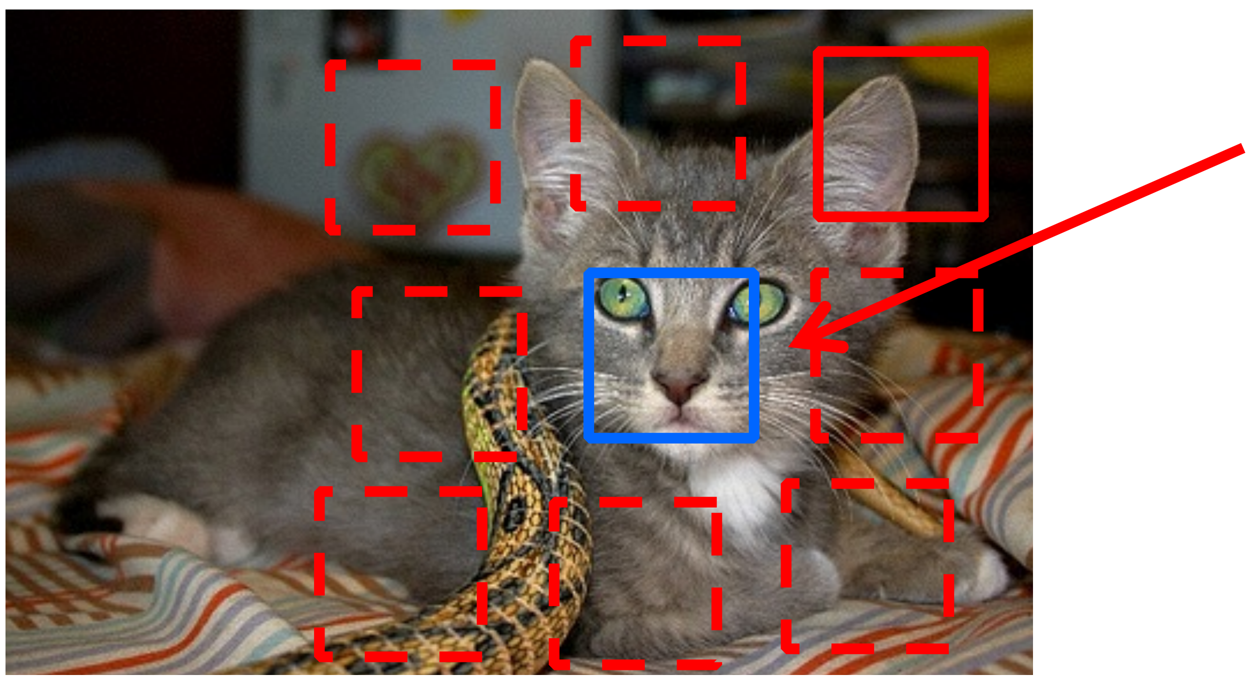




Slide from Carl Doersch

How to sample patches

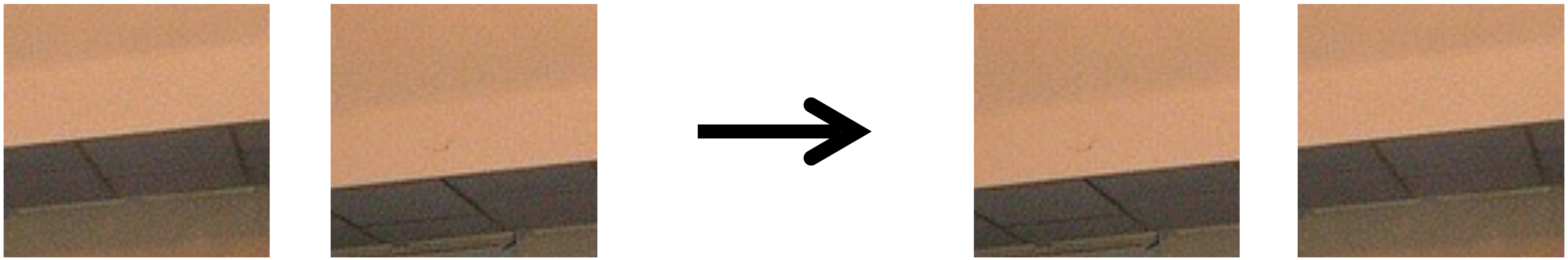


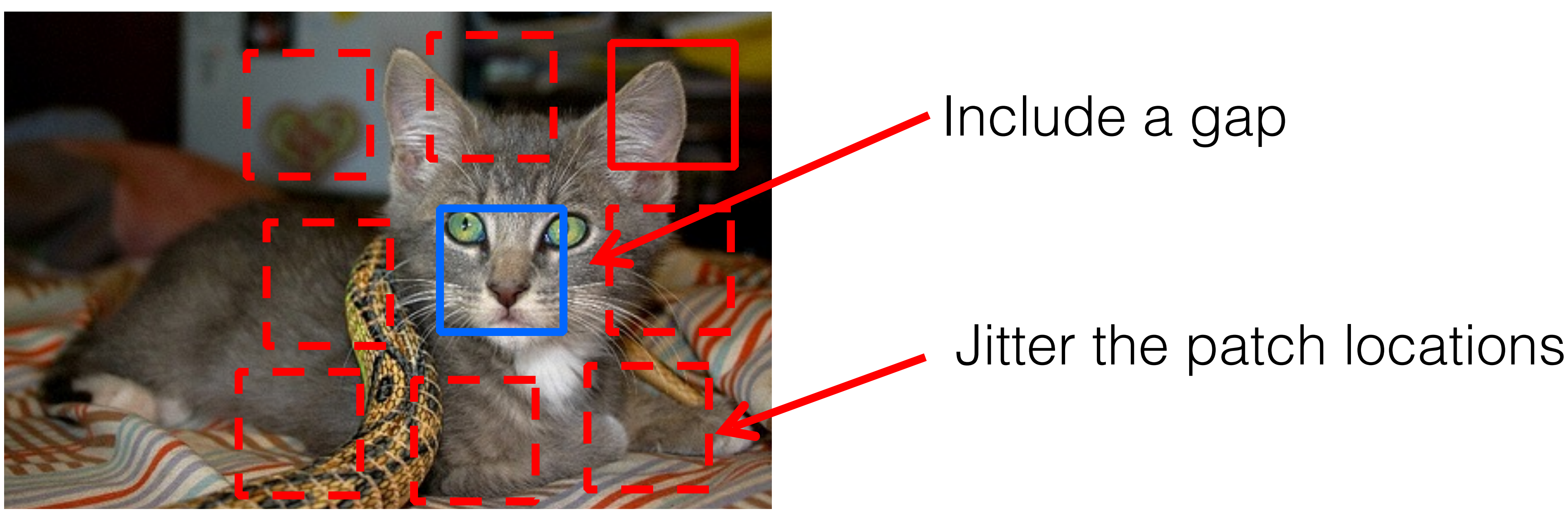


Include a gap

Slide from Carl Doersch

How to sample patches





Slide from Carl Doersch

Another trivial shortcut

- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training

Slide from Carl Doersch

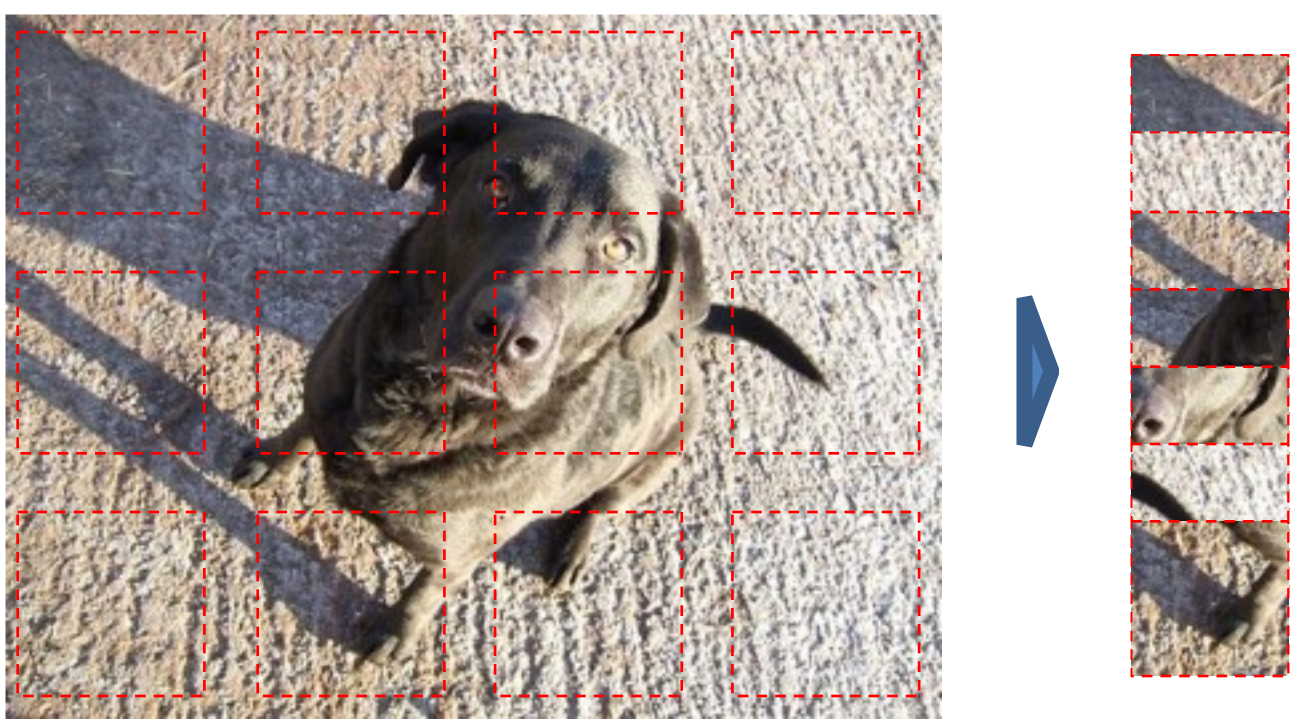
Another trivial shortcut

- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training

Slide from Carl Doersch

Another trivial shortcut

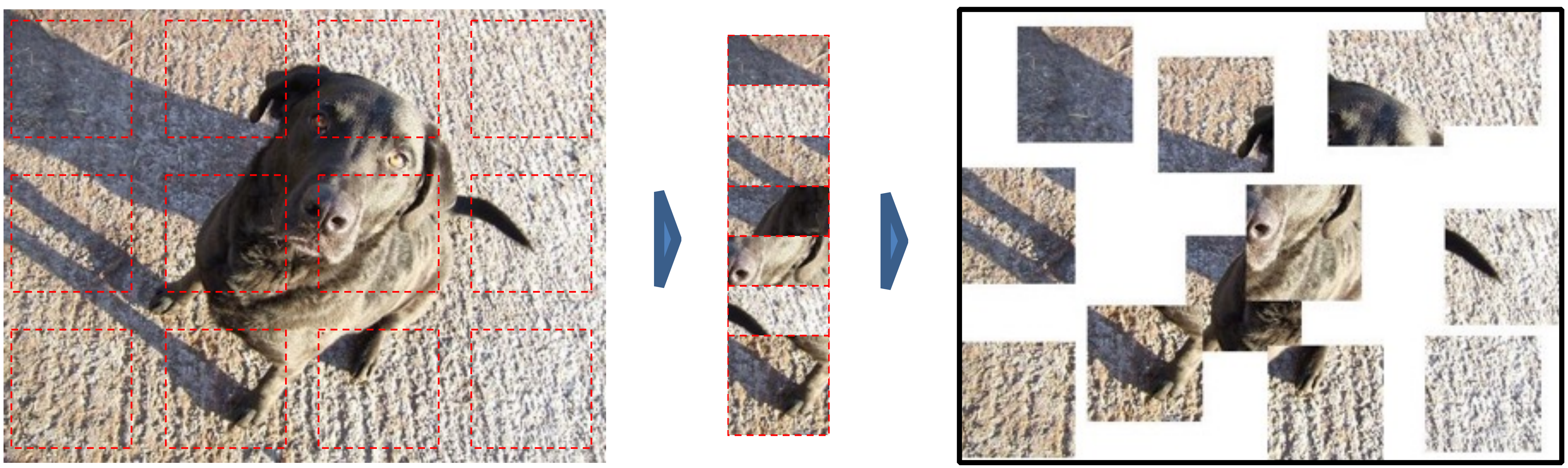
- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training



Slide from Carl Doersch

Another trivial shortcut

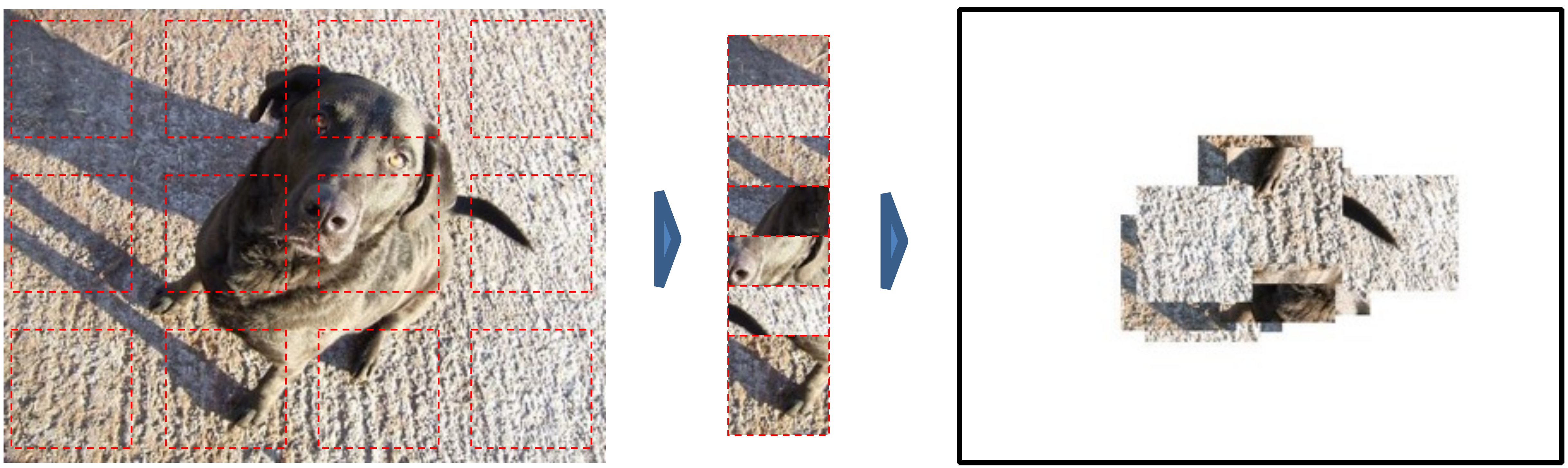
- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training



Slide from Carl Doersch

Another trivial shortcut

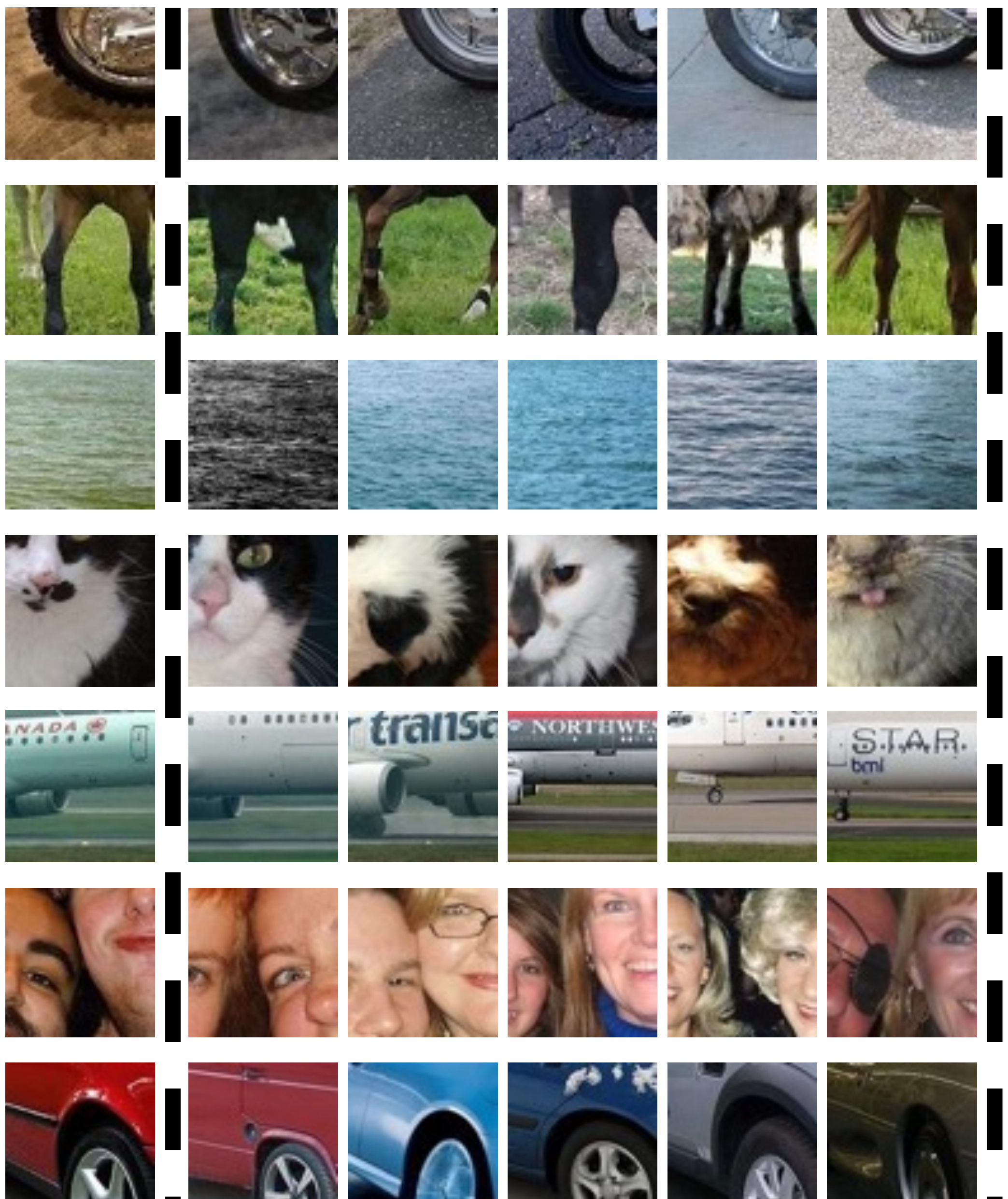
- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training



Slide from Carl Doersch

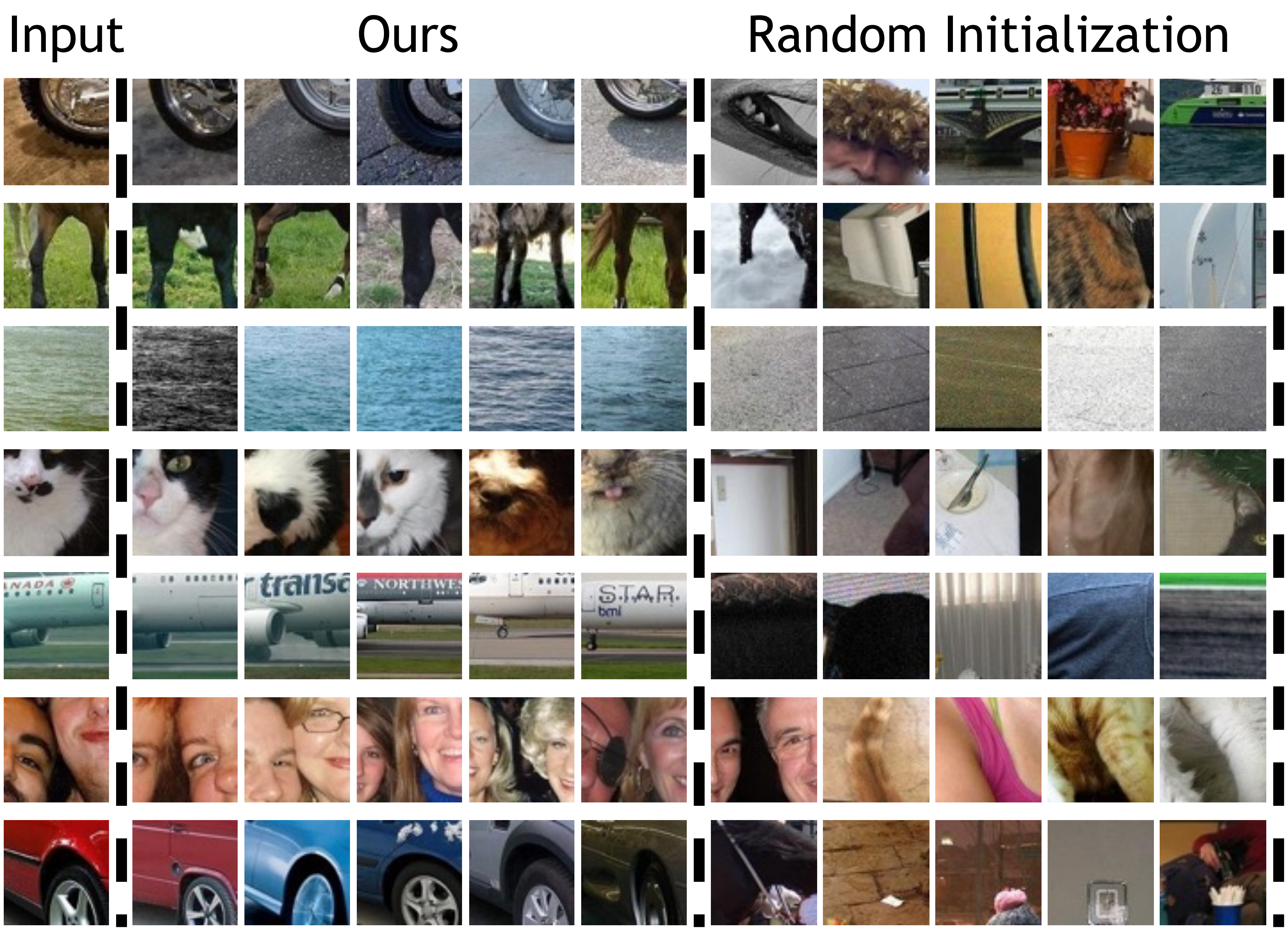
What is learned?

Input Ours



Slide from Carl Doersch

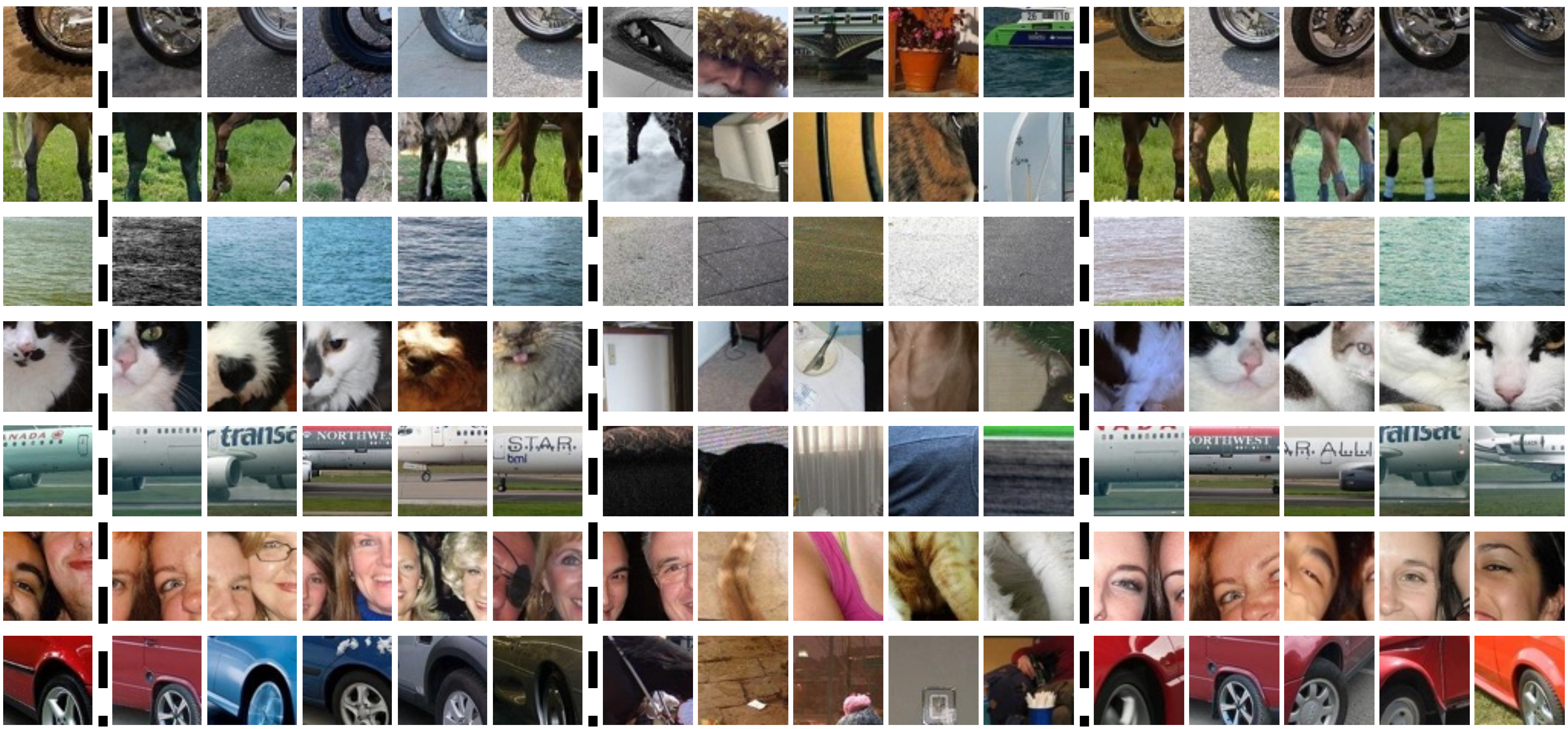
What is learned?



Slide from Carl Doersch

What is learned?

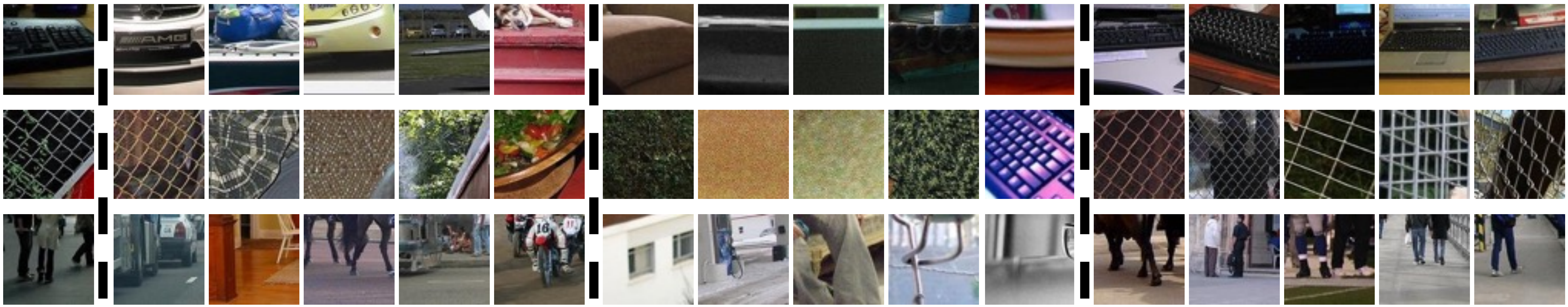
Input | Ours | Random Initialization | ImageNet AlexNet



Slide from Carl Doersch

Still don't capture everything

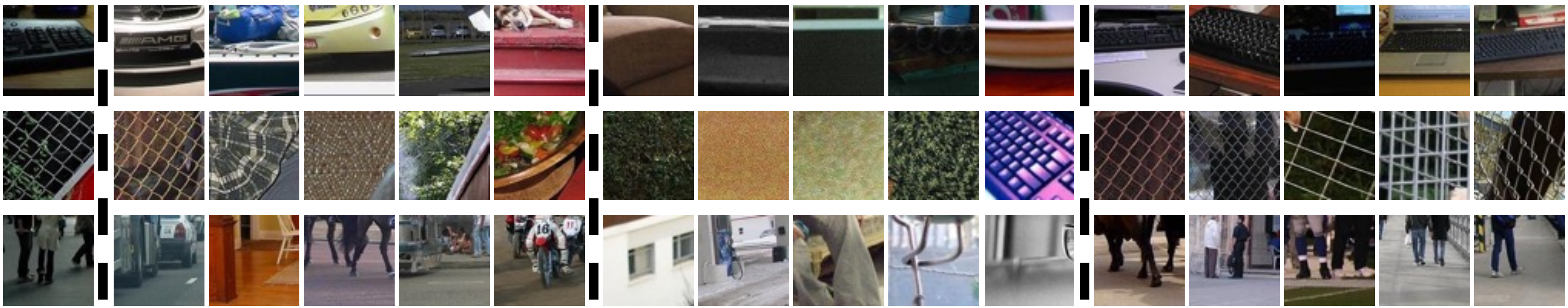
Input | Ours | Random Initialization | ImageNet AlexNet



Slide from Carl Doersch

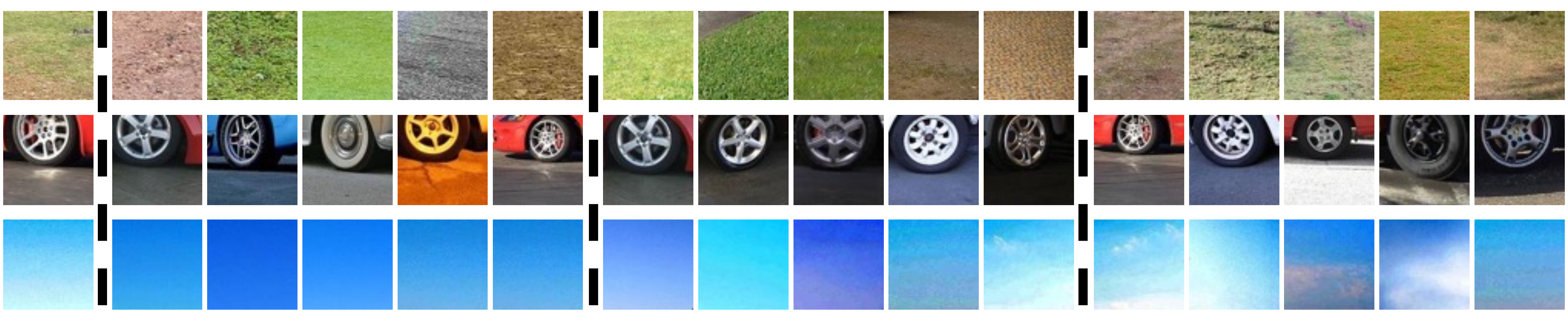
Still don't capture everything

Input Ours Random Initialization ImageNet AlexNet



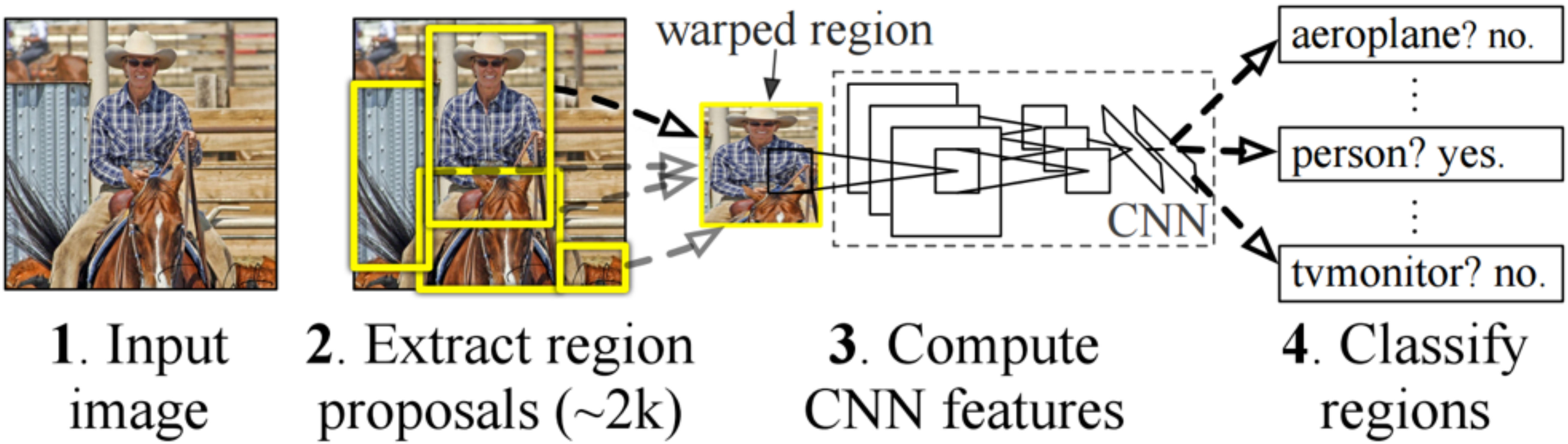
You don't always need to learn!

Input Ours Random Initialization ImageNet AlexNet



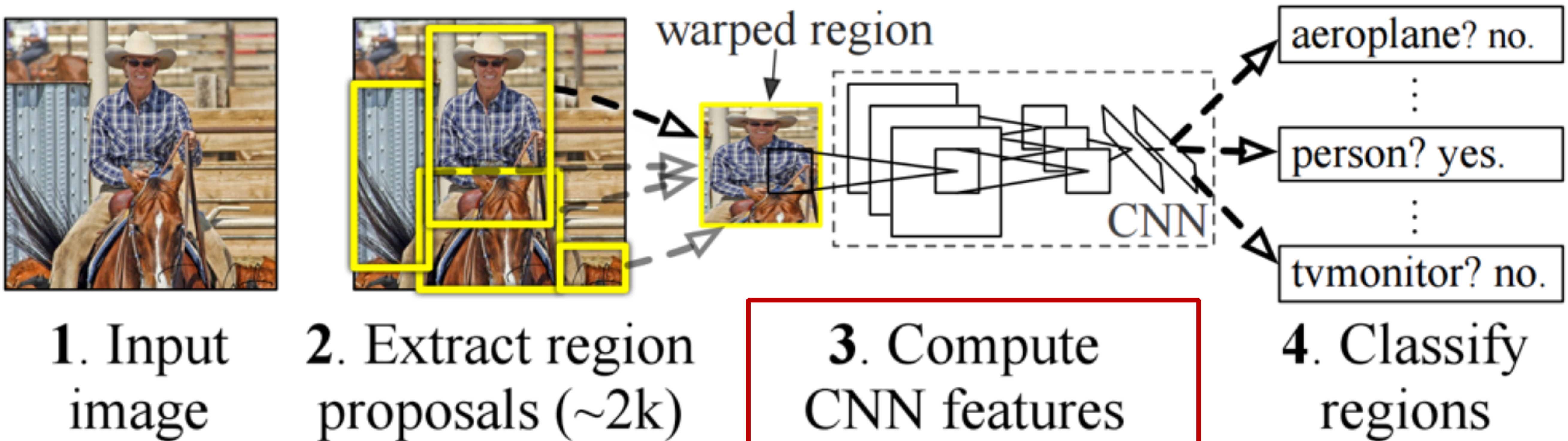
Slide from Carl Doersch

Pre-Training for R-CNN



Slide from Carl Doersch

Pre-Training for R-CNN

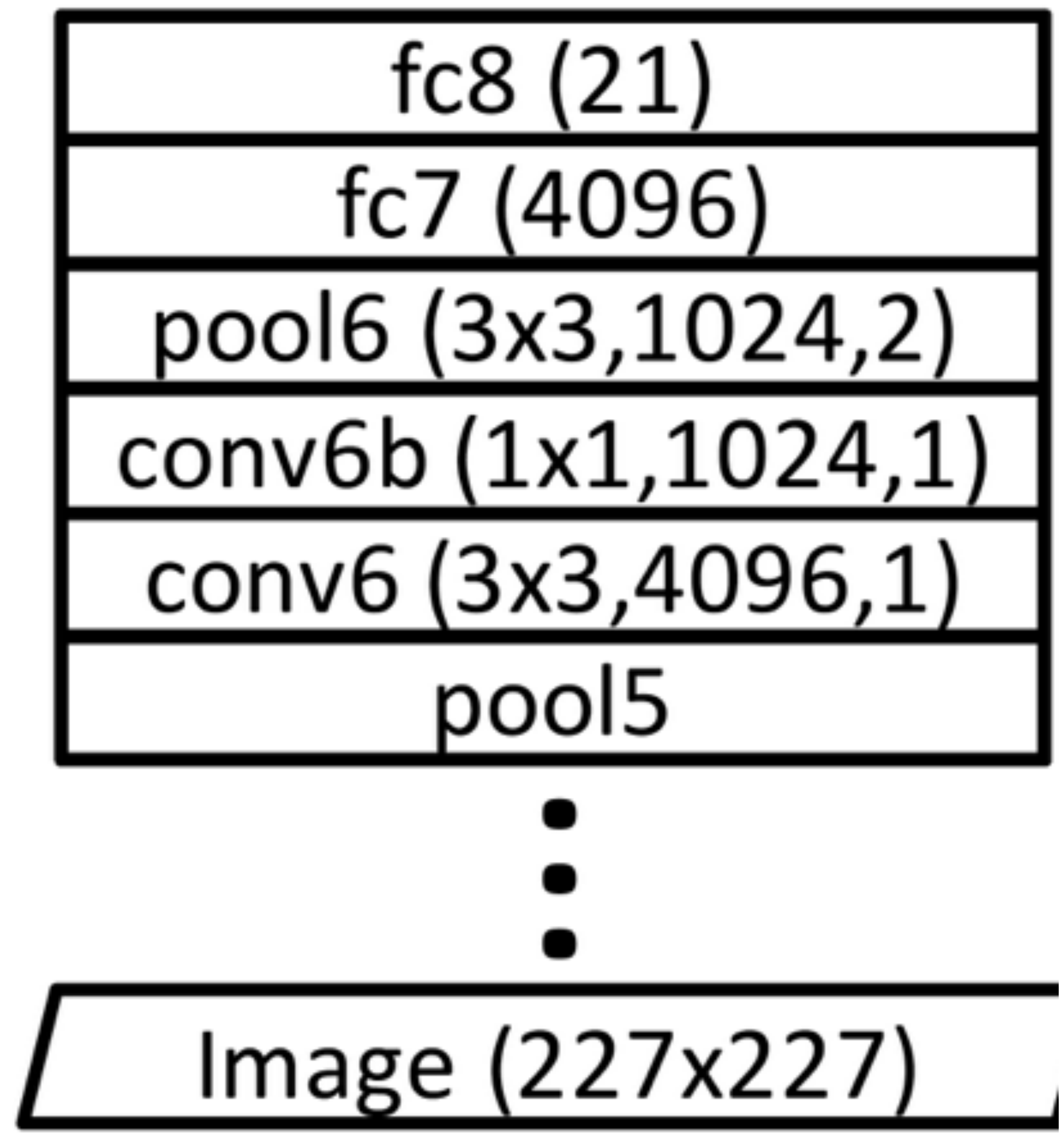


Pre-train on relative-position task, w/o labels

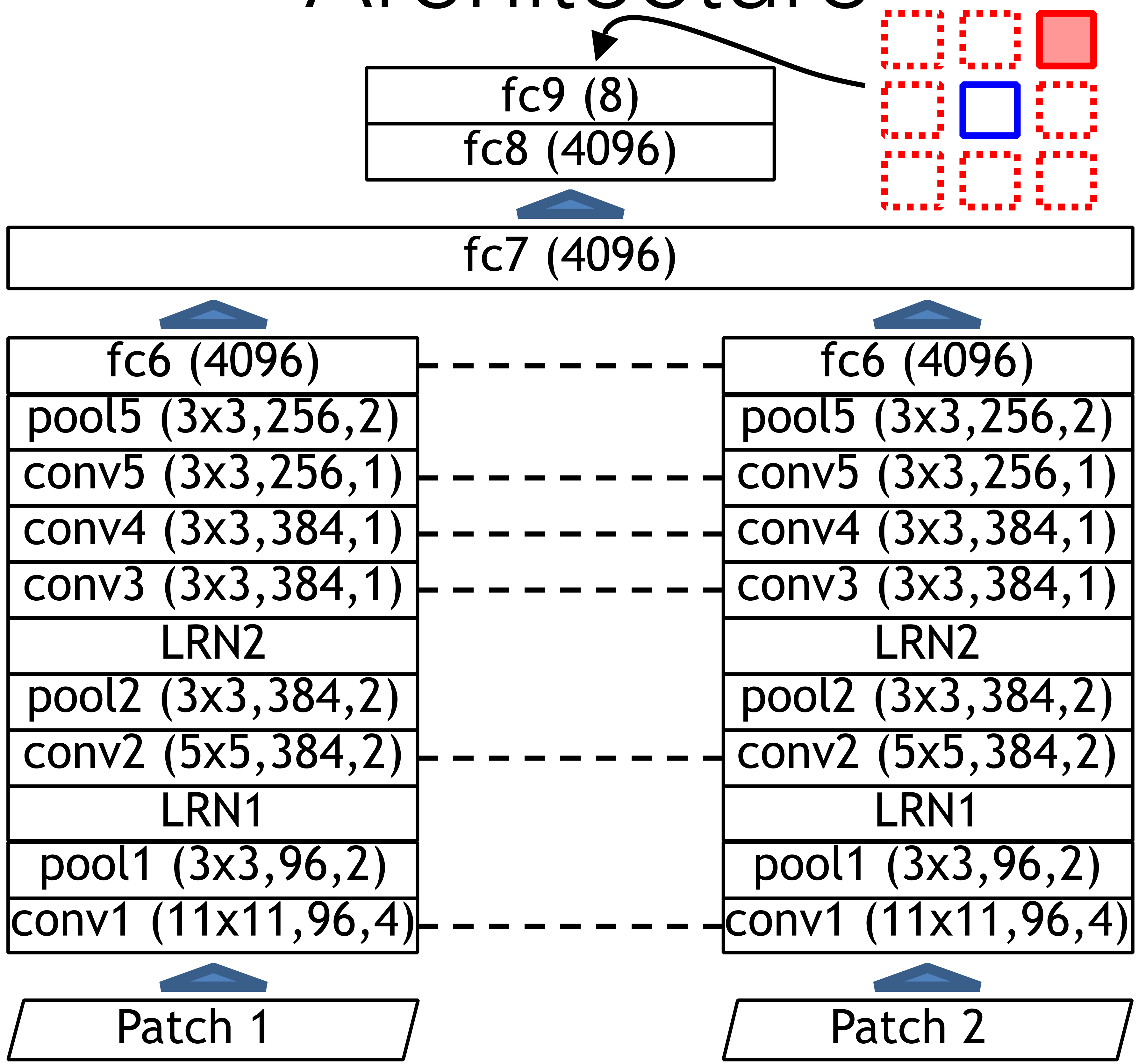
Slide from Carl Doersch

Details

- Use stack from patch context predictor before pool5
- Resize convolution layers to work on 227x227 instead of 96x96
- Use FC7 as the final representation



Architecture



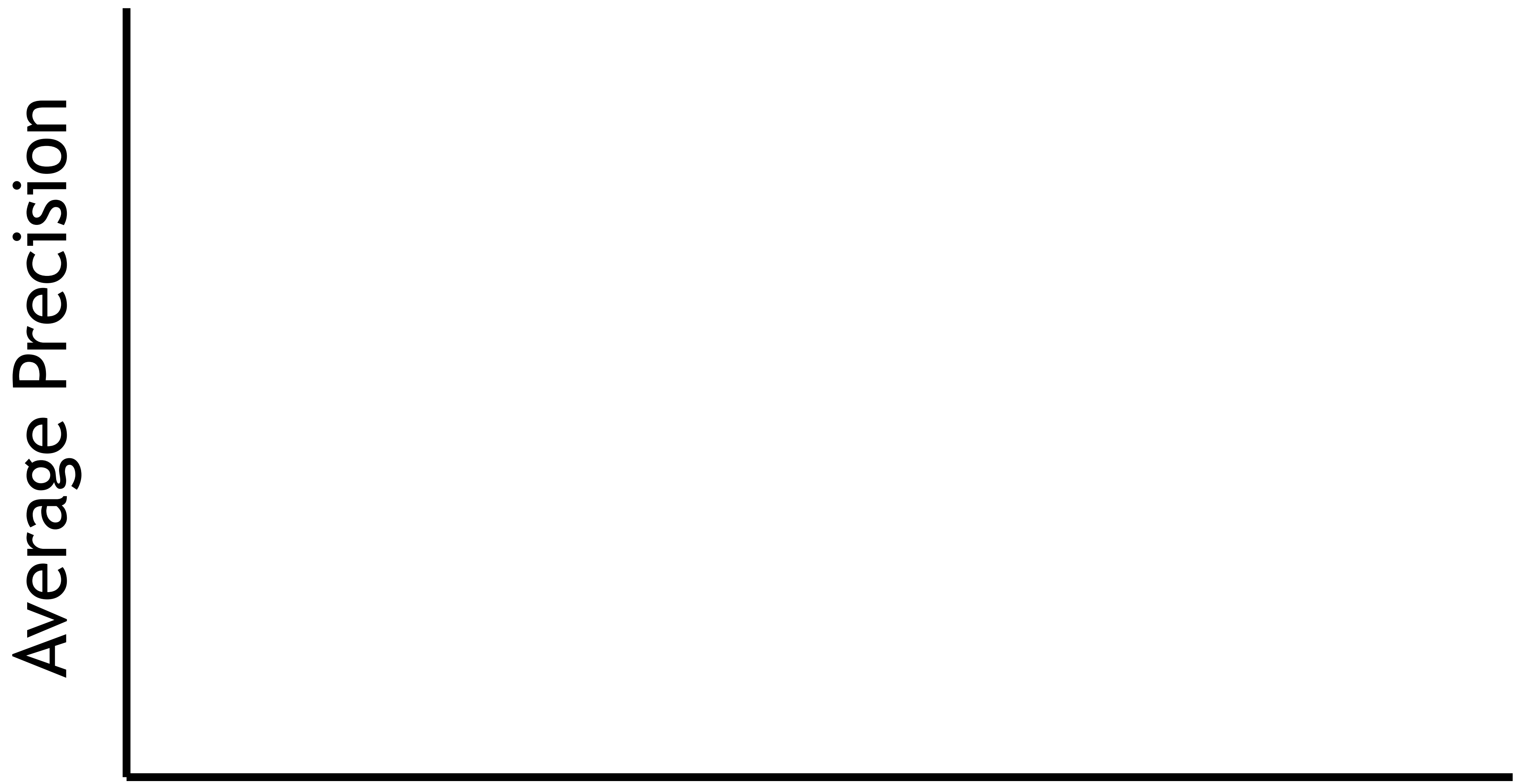
Slide from Carl Doersch

VOC 2007 Performance
(pretraining for R-CNN)

Slide from Carl Doersch

VOC 2007 Performance

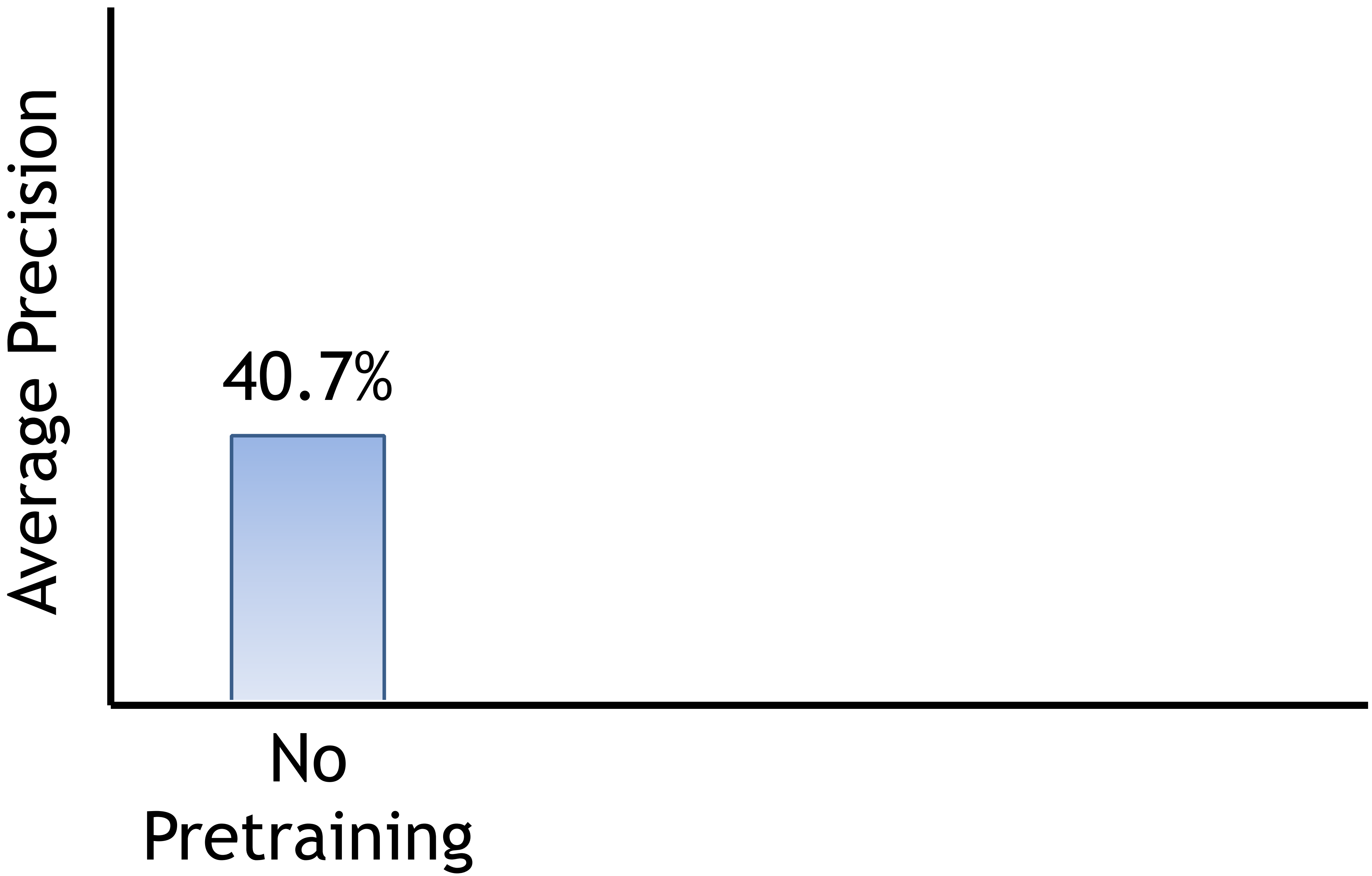
(pretraining for R-CNN)



Slide from Carl Doersch

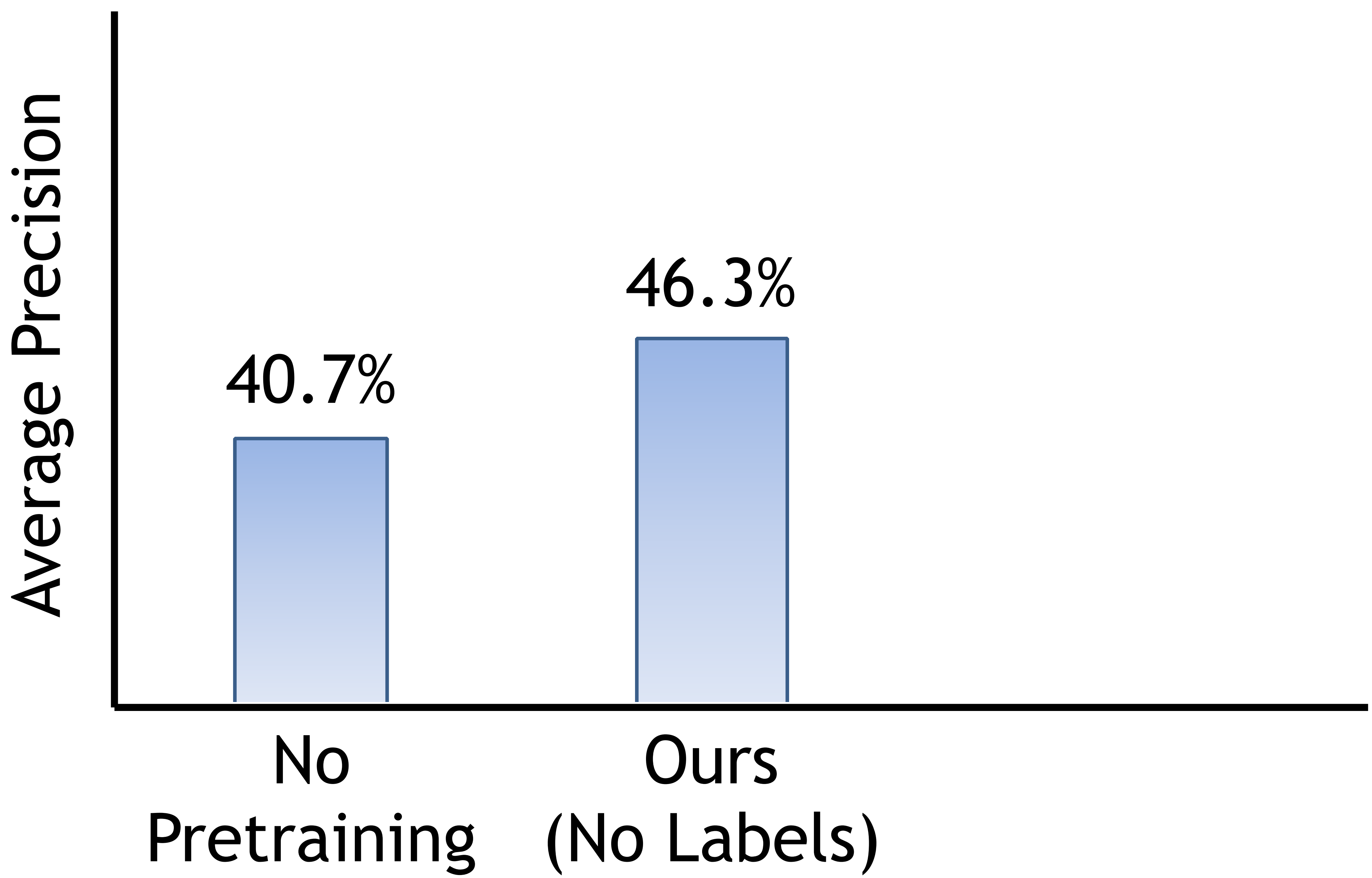
VOC 2007 Performance

(pretraining for R-CNN)



Slide from Carl Doersch

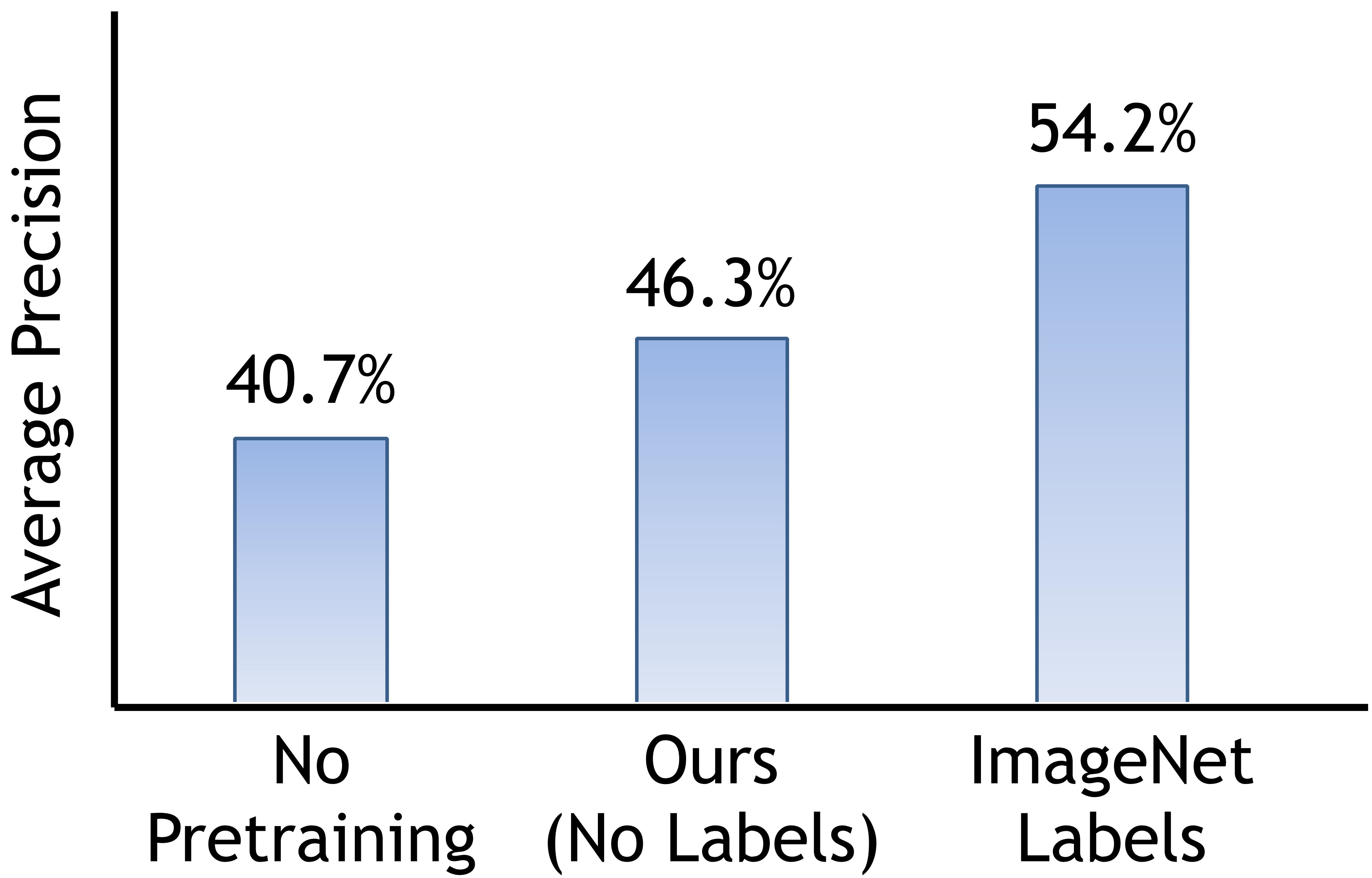
VOC 2007 Performance
(pretraining for R-CNN)



Slide from Carl Doersch

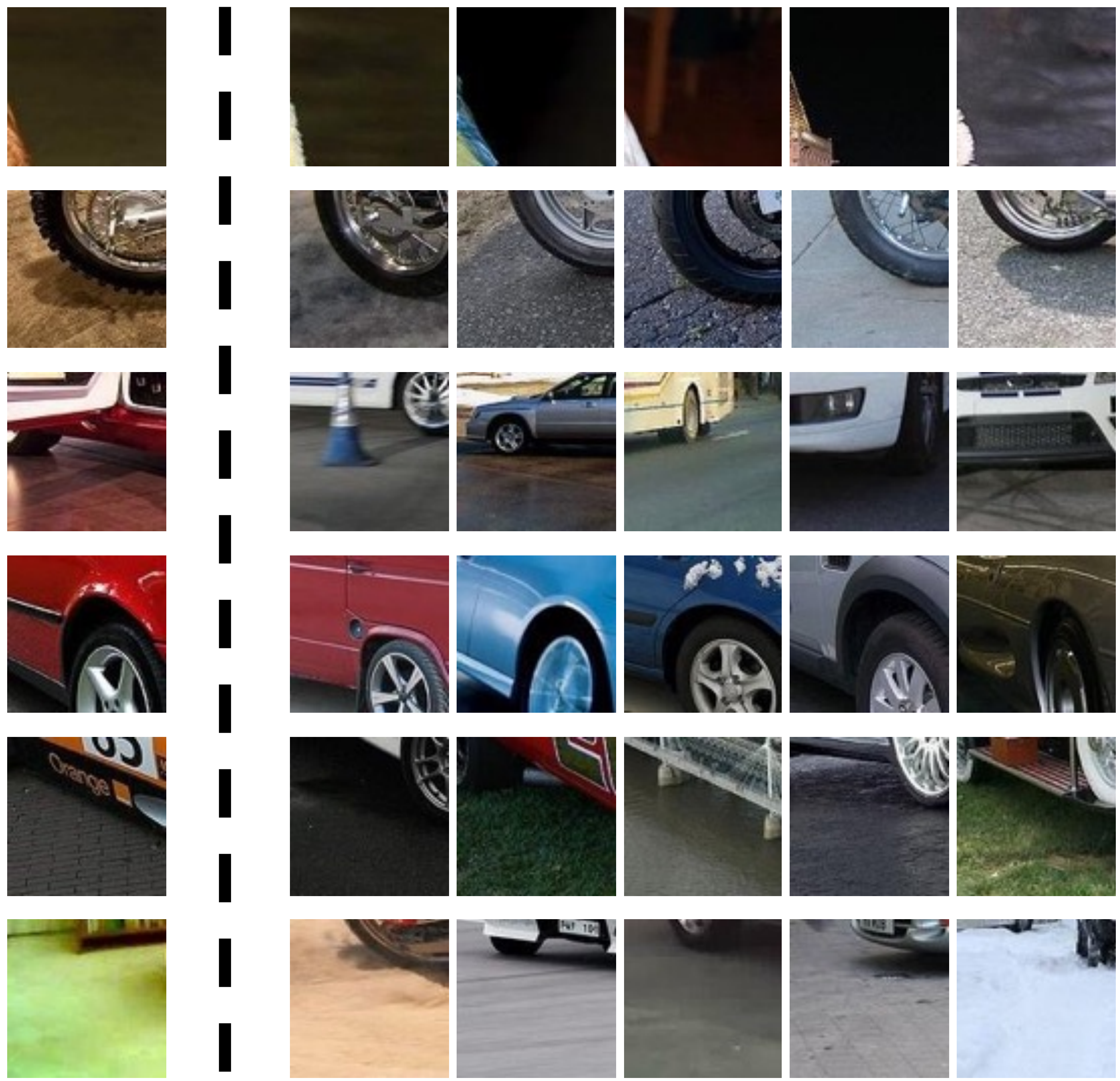
VOC 2007 Performance

(pretraining for R-CNN)



Slide from Carl Doersch

Unsupervised Object Discovery?



Slide from Carl Doersch

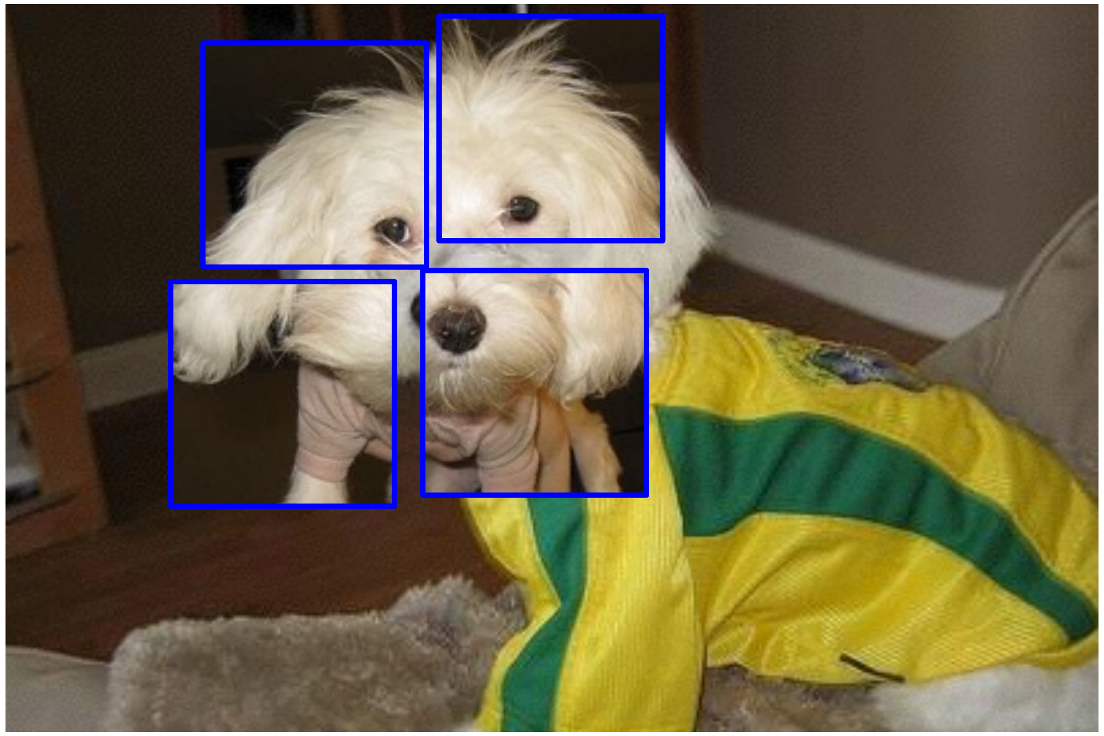
Unsupervised Object Discovery

Slide from Carl Doersch

Unsupervised Object Discovery

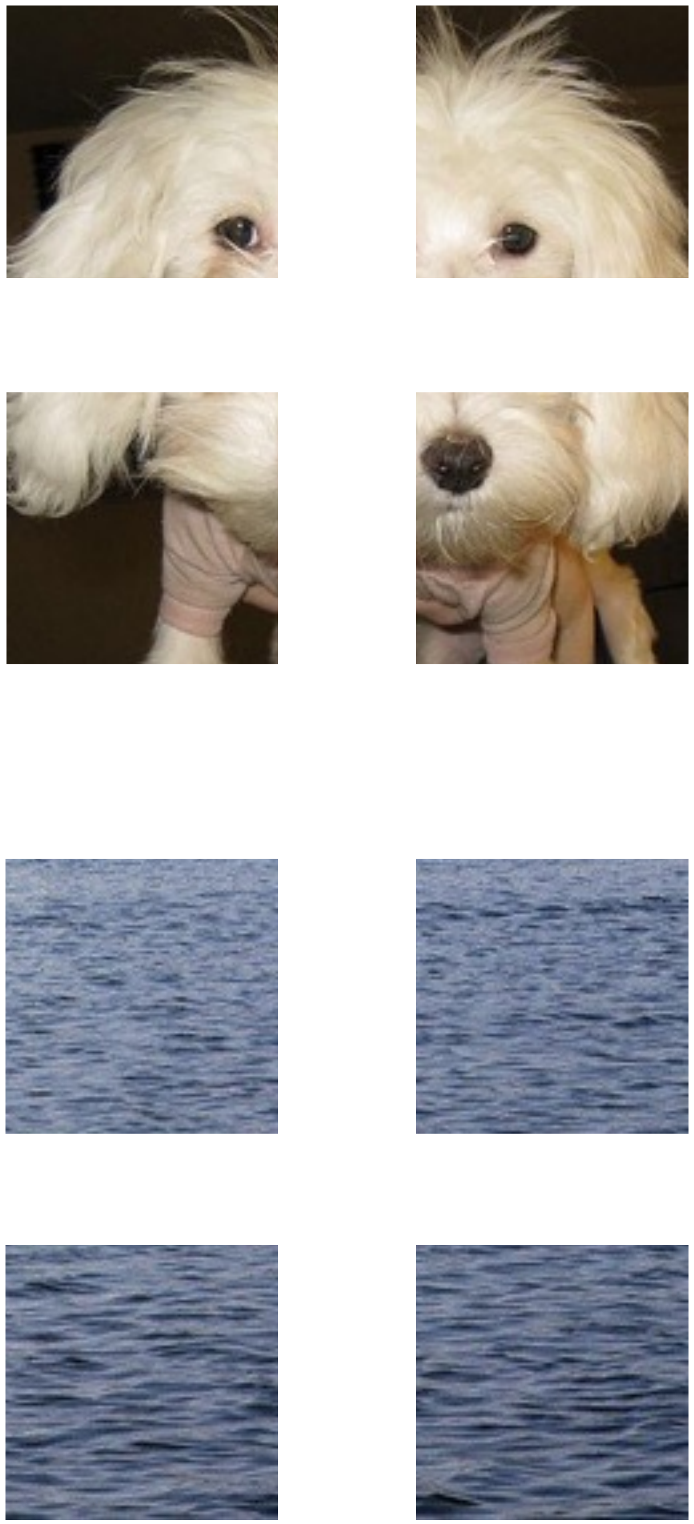
Slide from Carl Doersch

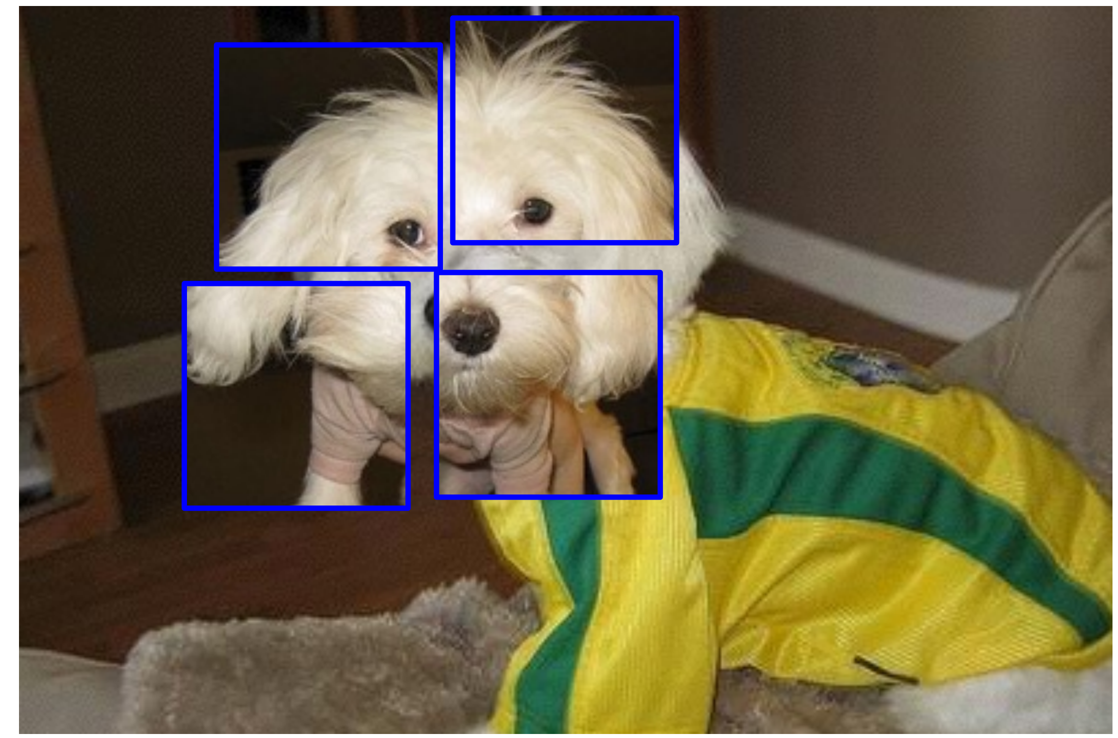
Unsupervised Object Discovery



Slide from Carl Doersch

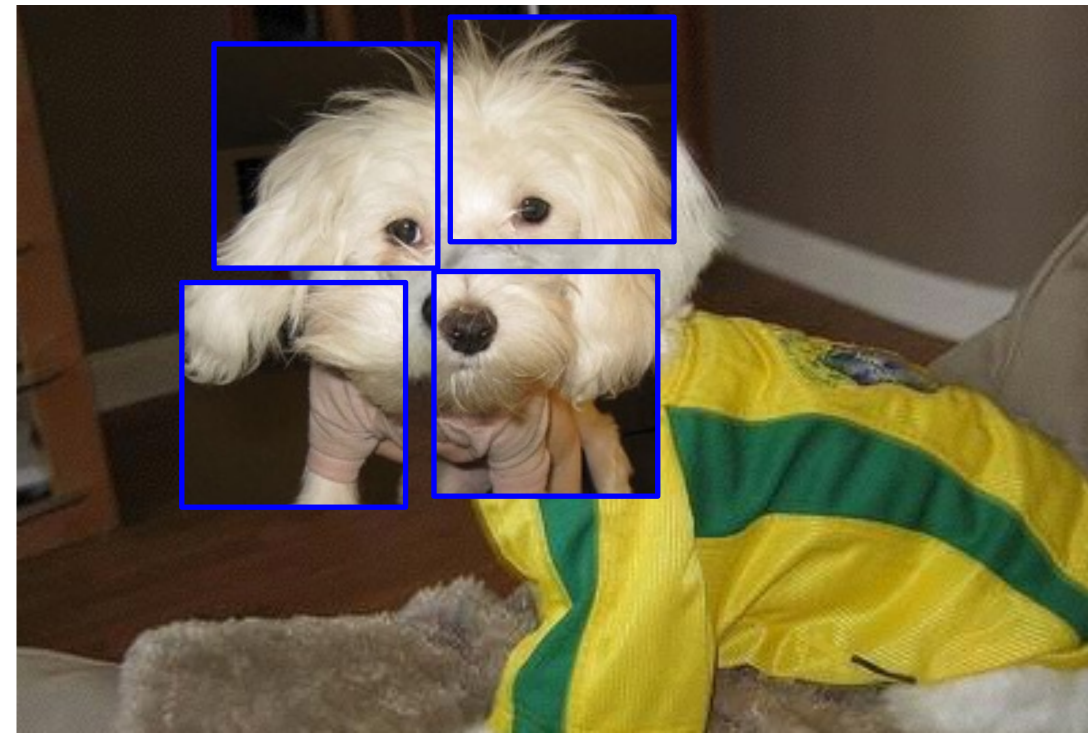
Unsupervised Object Discovery

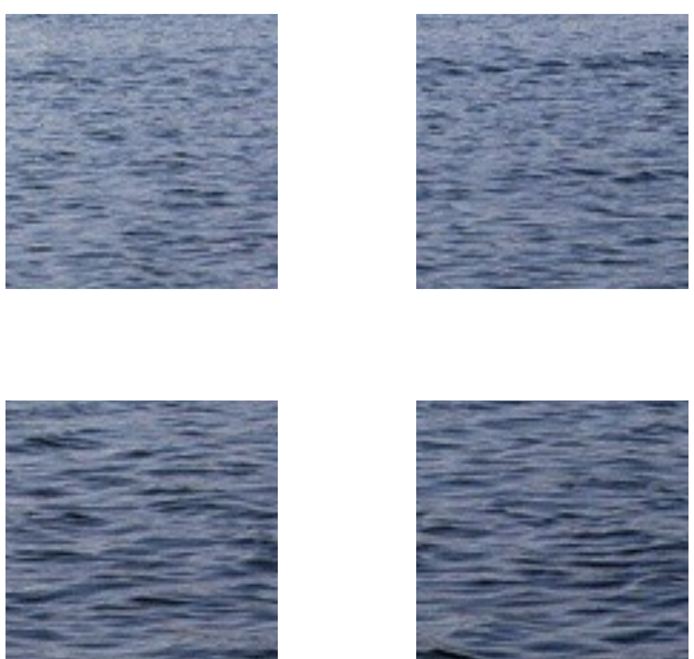


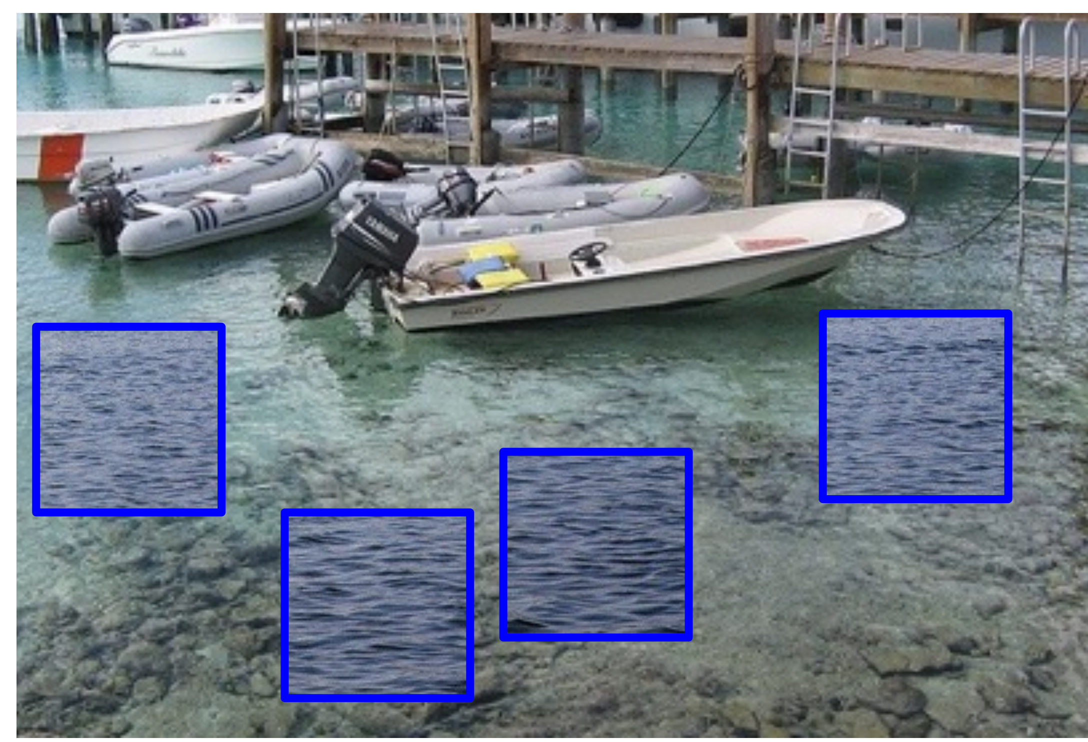


Slide from Carl Doersch

Unsupervised Object Discovery

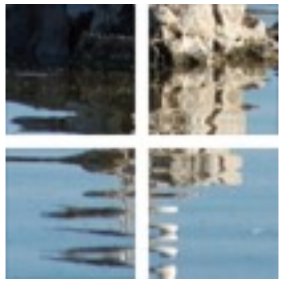






Slide from Carl Doersch

Algorithm



⋮

Slide from Carl Doersch

Algorithm

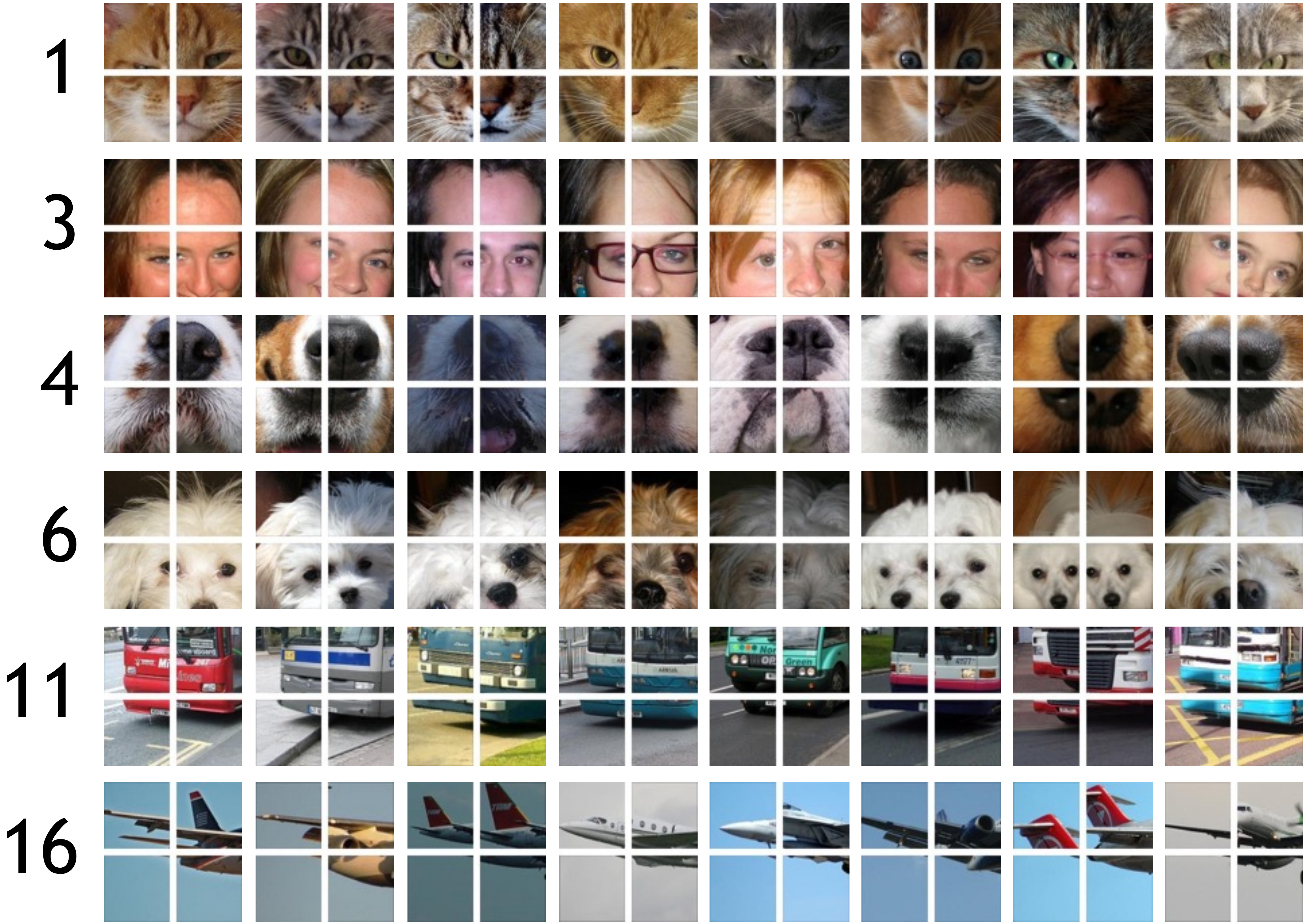
Slide from Carl Doersch

Algorithm

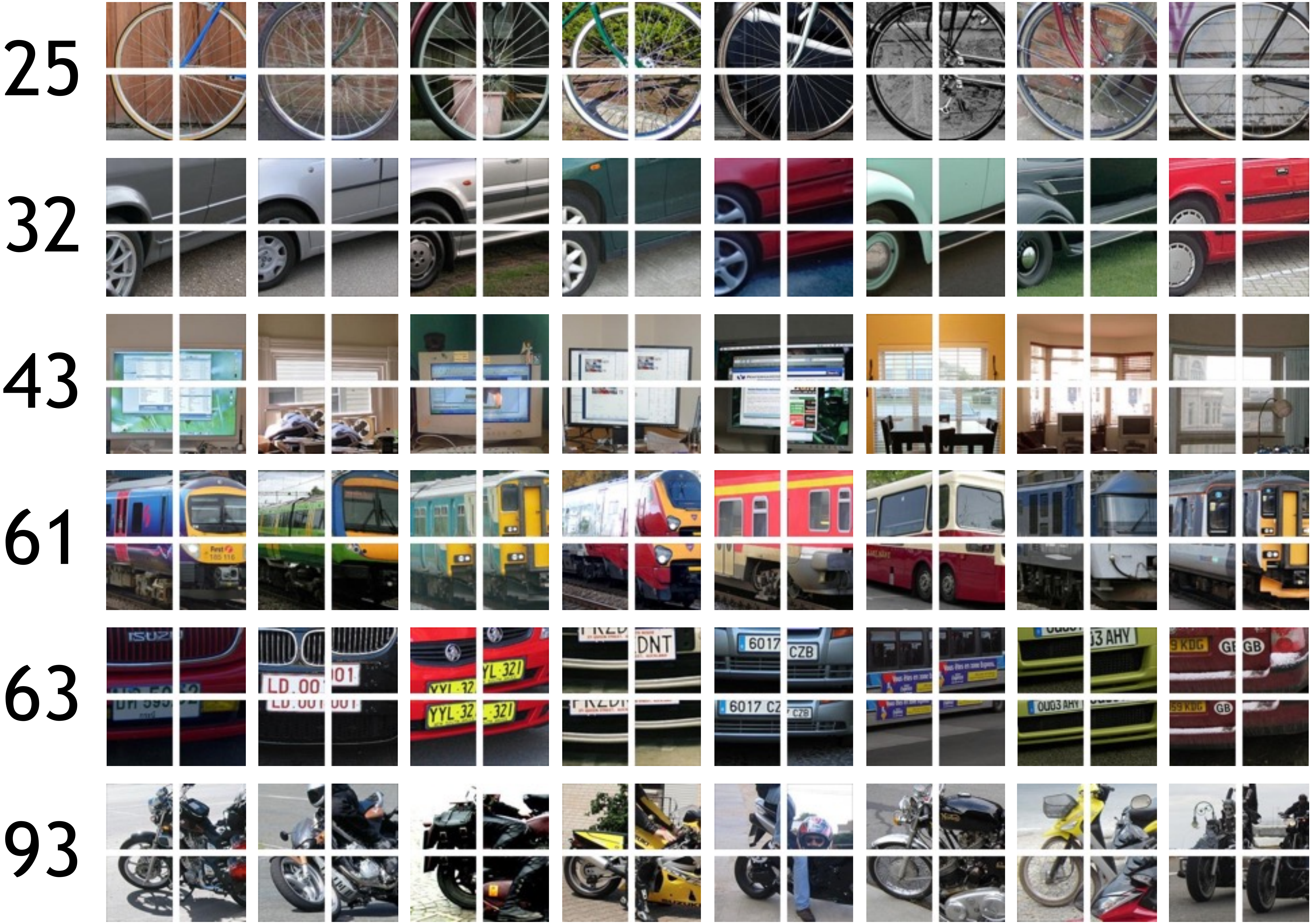
Slide from Carl Doersch

Algorithm

Slide from Carl Doersch



Slide from Carl Doersch



Slide from Carl Doersch

119

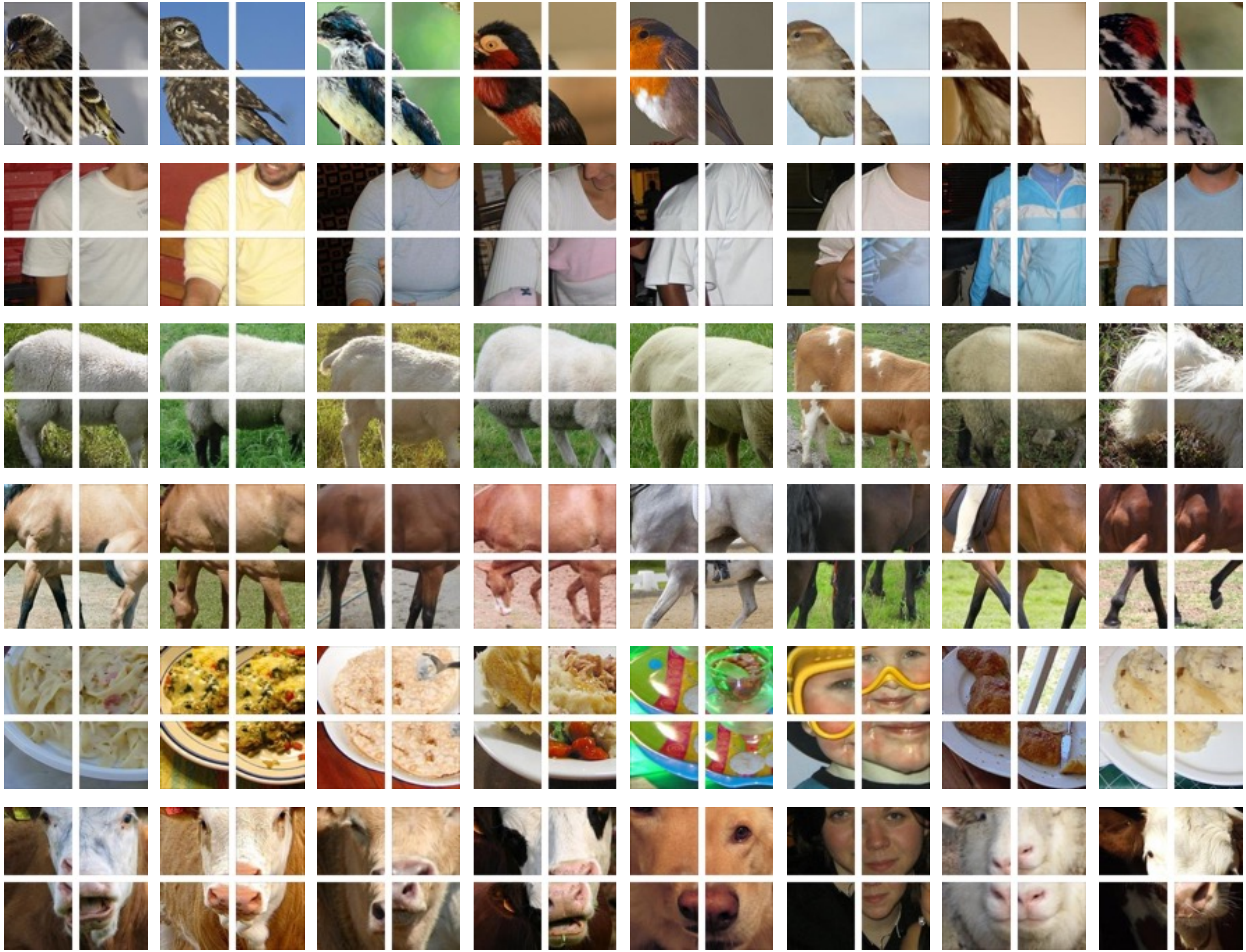
152

169

182

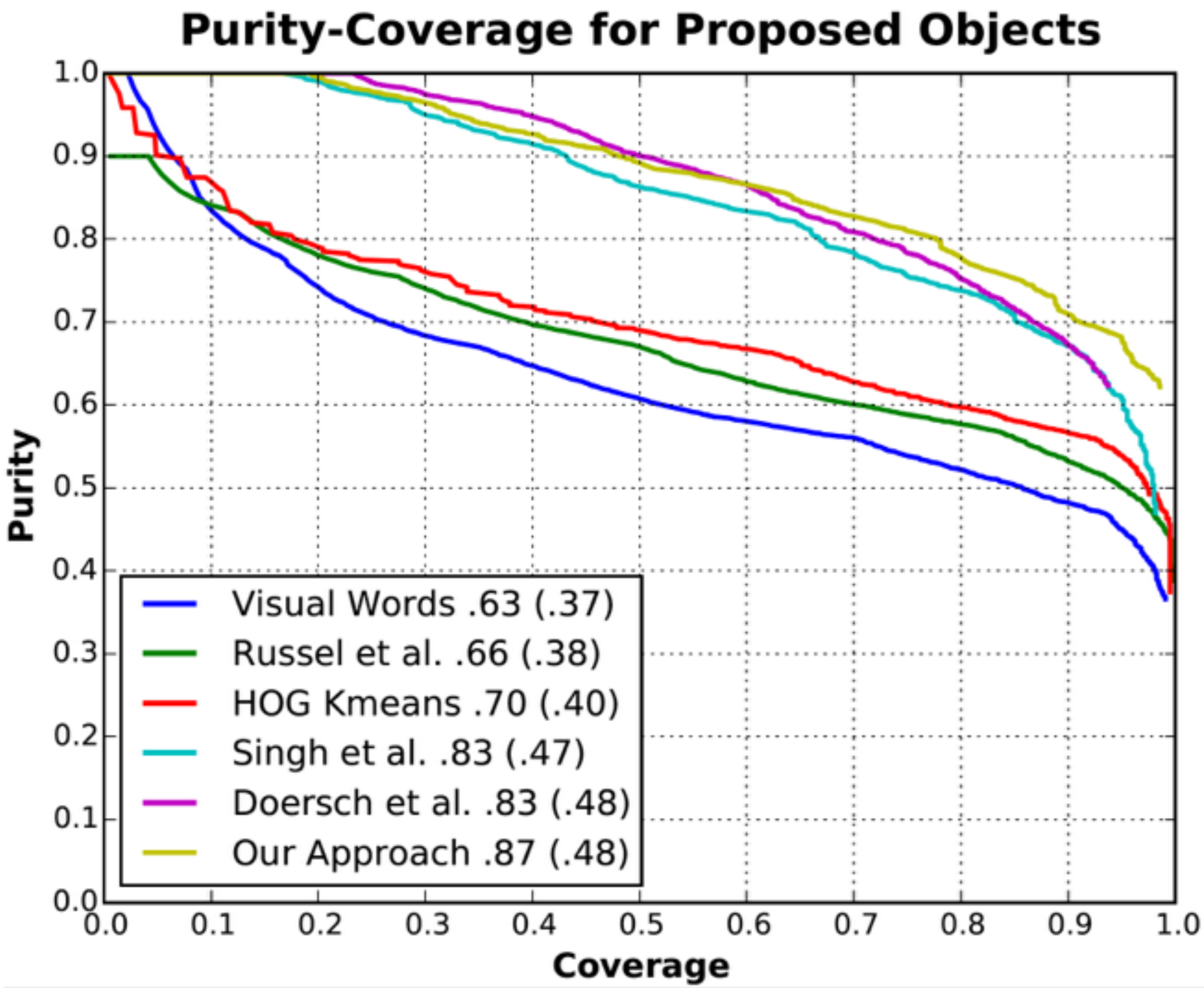
395

457



Slide from Carl Doersch

Purity vs Coverage



Pretext Task

- Performance on Pascal VOC is 38.4% (Chance is 12.5%)

- On ImageNet accuracy is 39.5% on training set, and 40.5% on validation

- On GT box patches - similar performance. 39.2% overall with 45.6% on cars

Questions?