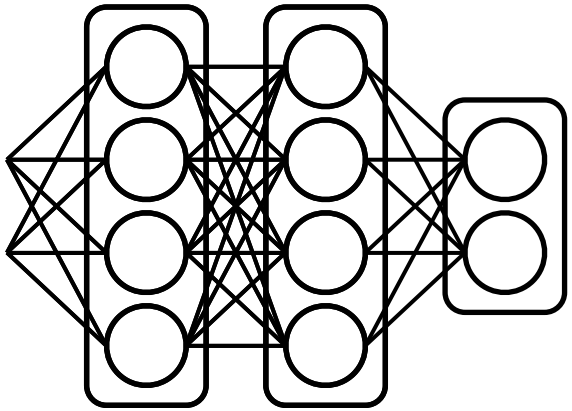
Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros & Abhinav Gupta

Present: Yisha Sun
11/01/2016

ImageNet + Deep Learning

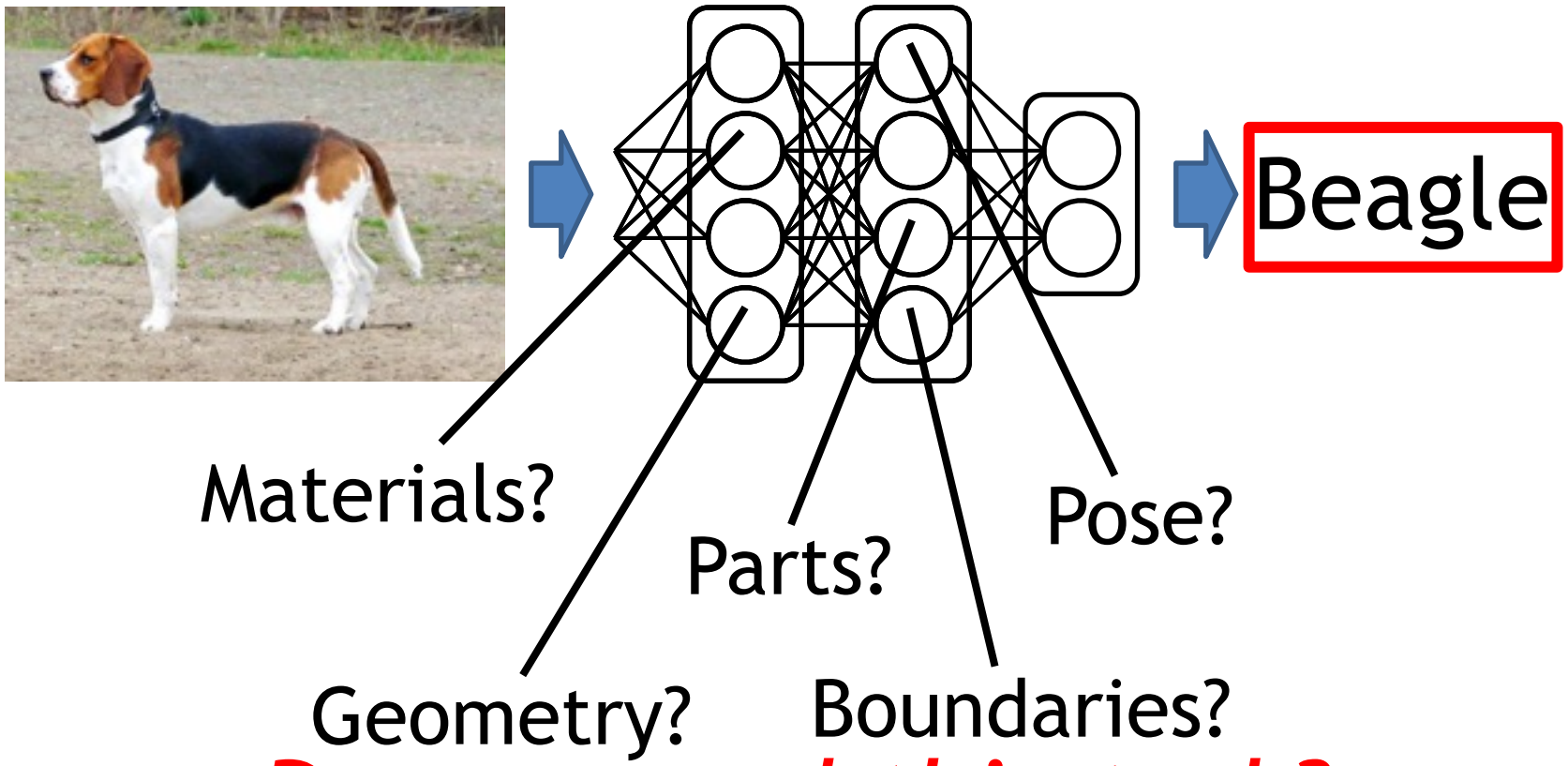


Beagle

- Image Retrieval
- Detection (RCNN)
- Segmentation (FCN
- Depth Estimation
- …

Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

ImageNet + Deep Learning



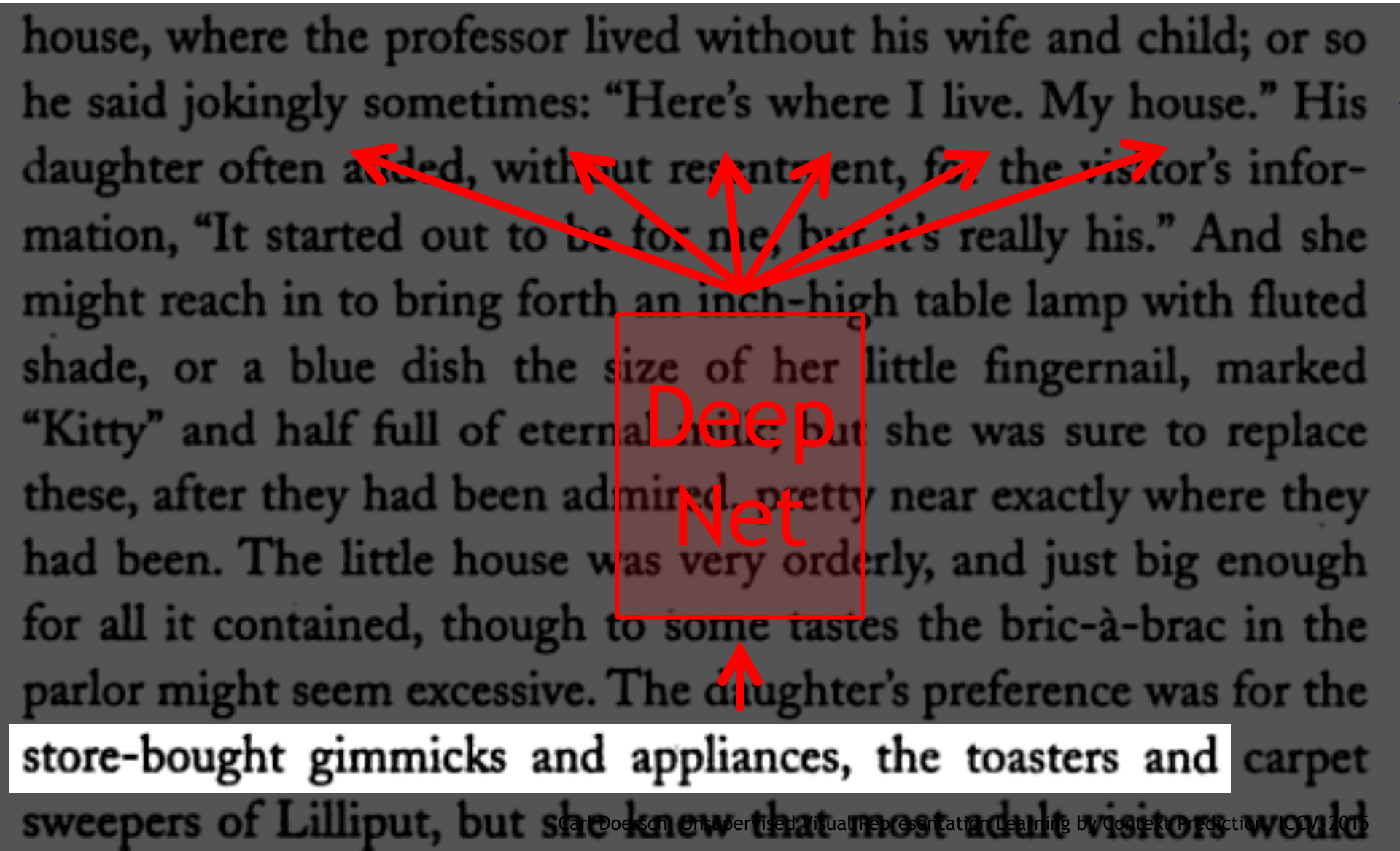
Do we need this task?

Do we even need semantic labels?

Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Context as Supervision

[Collobert & Weston 2008; Mikolov et al. 2013]

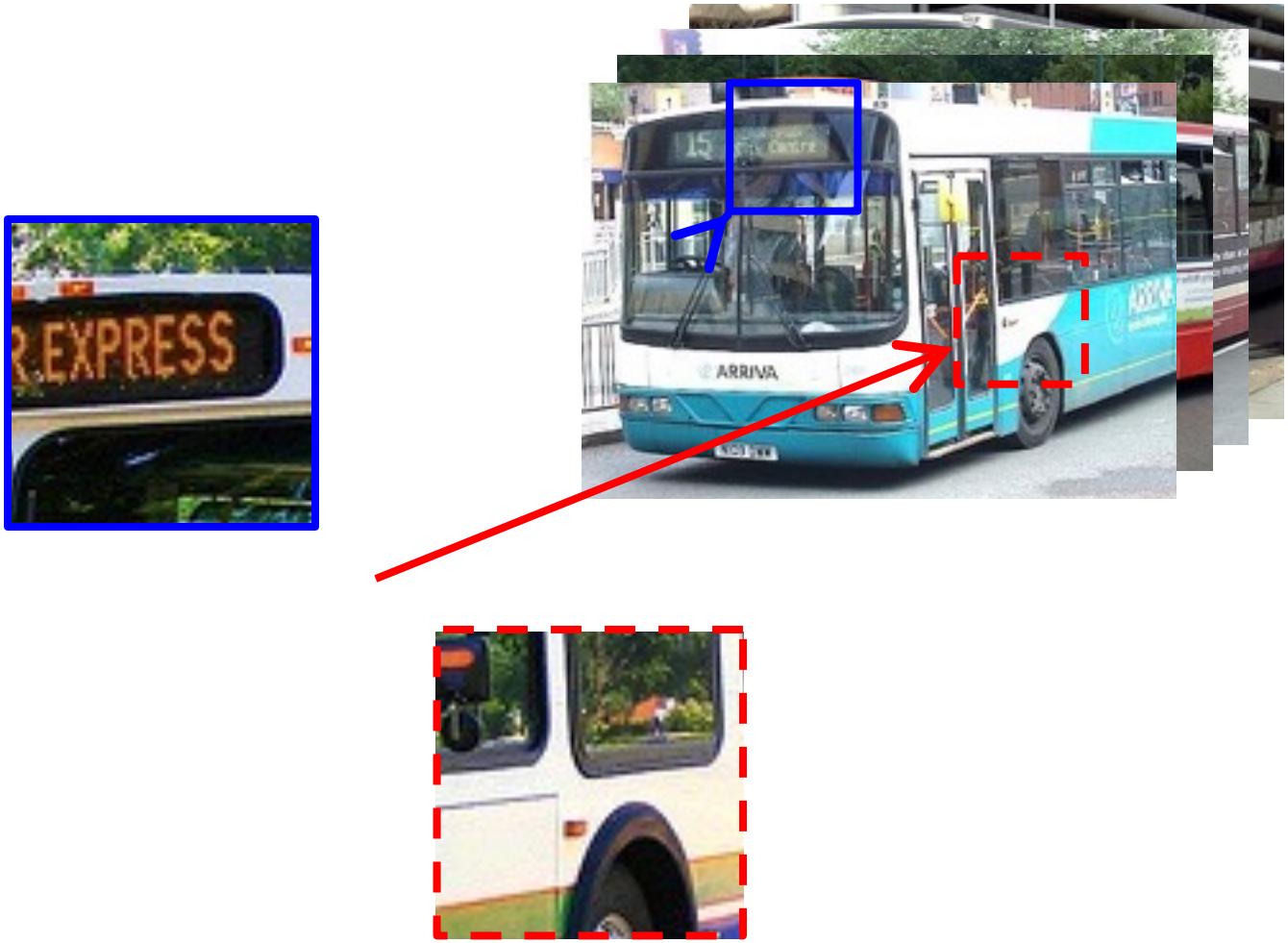


Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV 2015

Context Prediction for Images

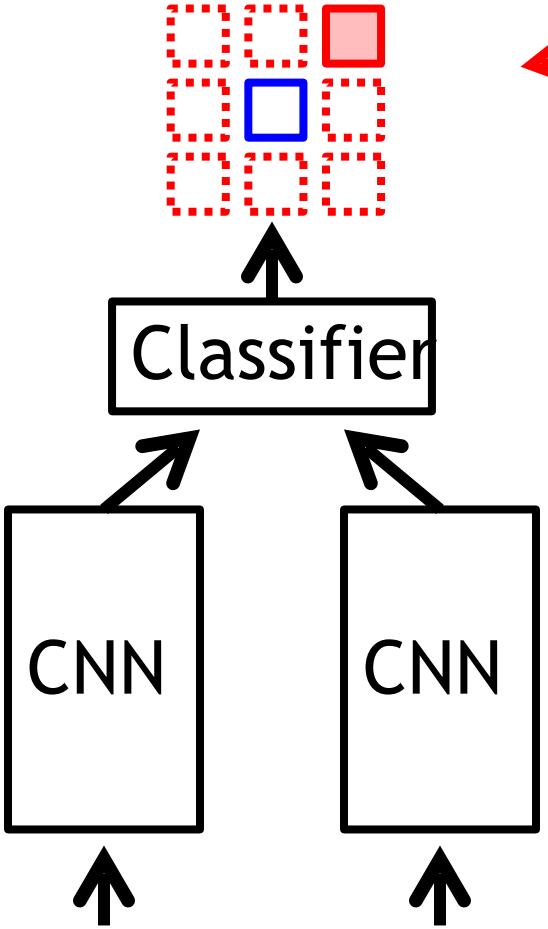
Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Semantics from a non-semantic task

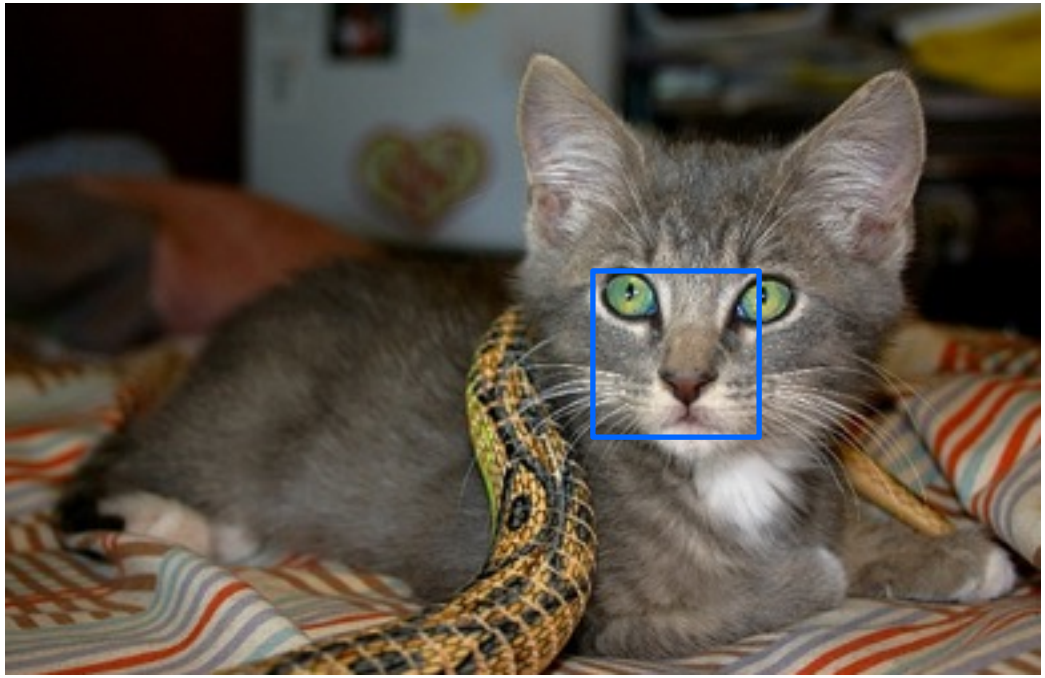


Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Relative Position Task



8 possible locations

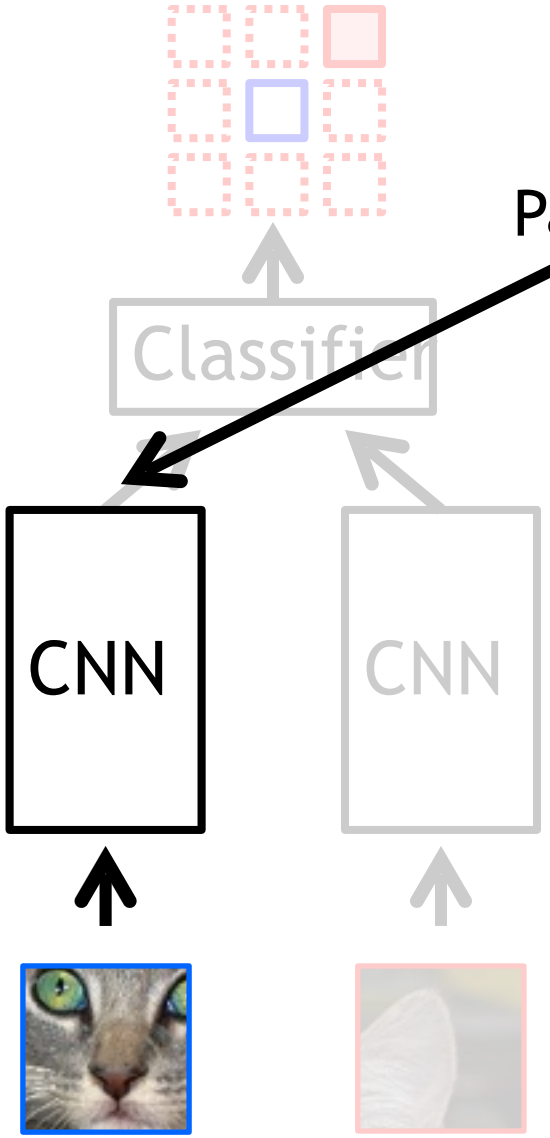


Randomly Sample Patch

Sample Second Patch

Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

- The ultimate goal is to learn a feature embedding for *individual* patches, such that patches which are visually similar (across different images) would be close in the embedding space.



Patch Embedding

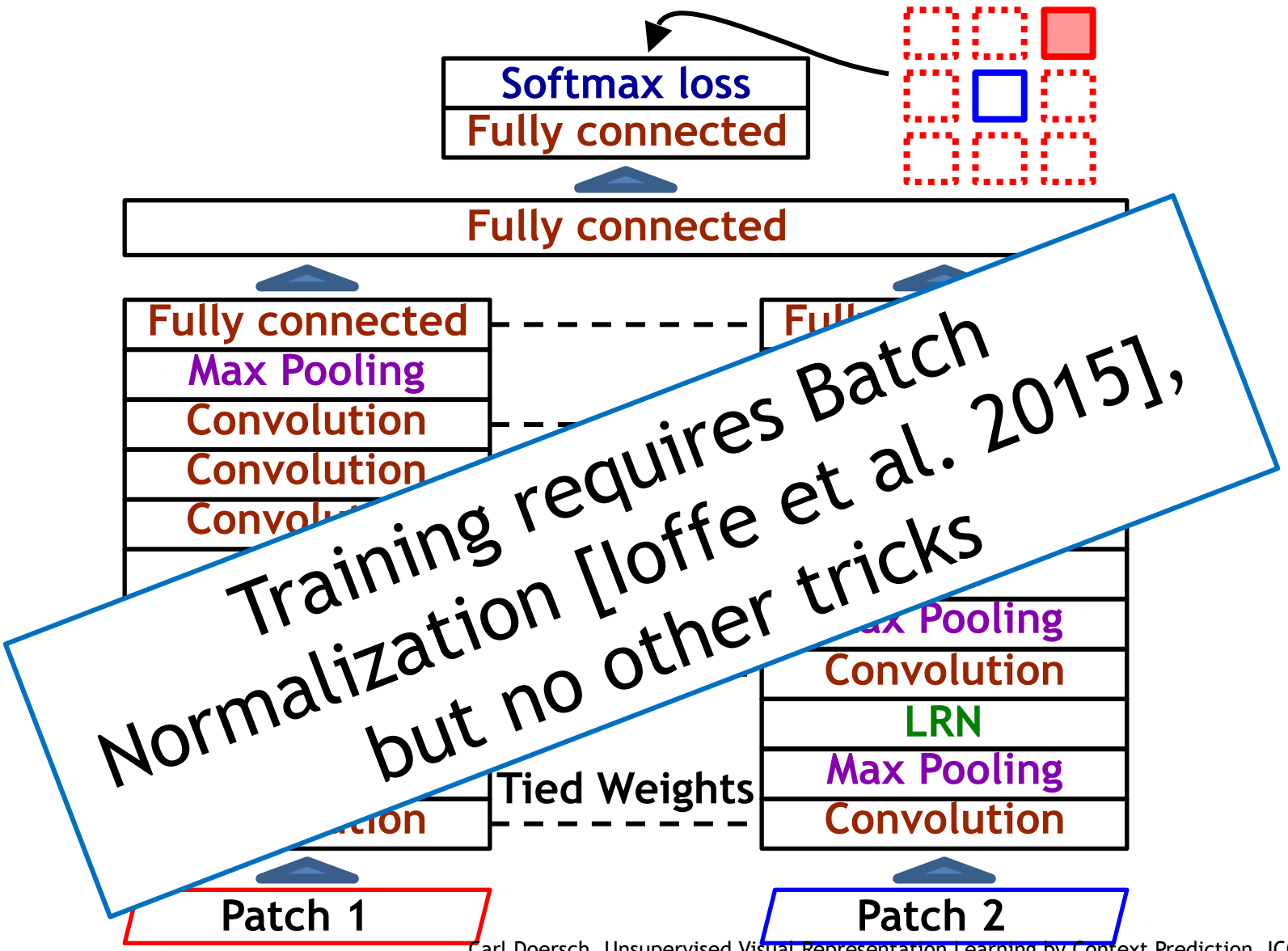
Input

Nearest Neighbors

Note: connects ***across*** instances!

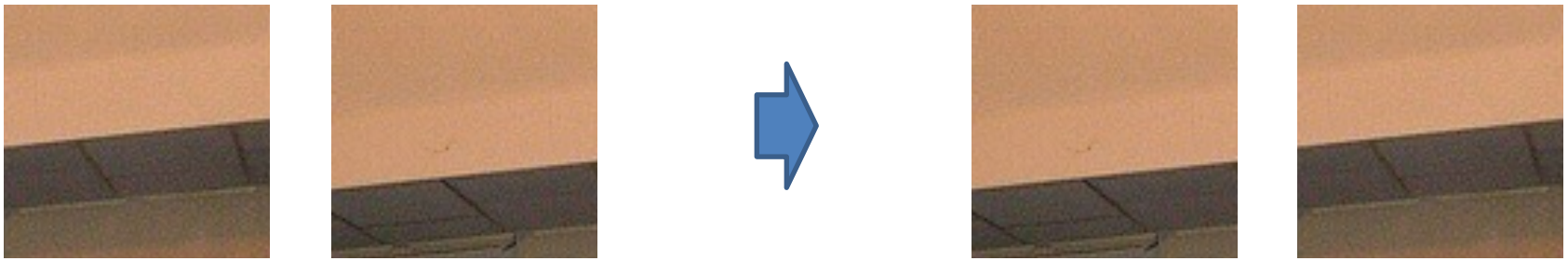
Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

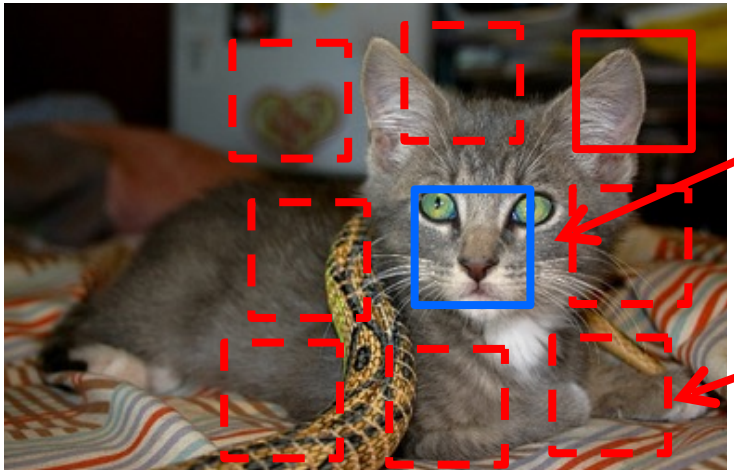
Architecture



Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Avoiding Trivial Shortcuts



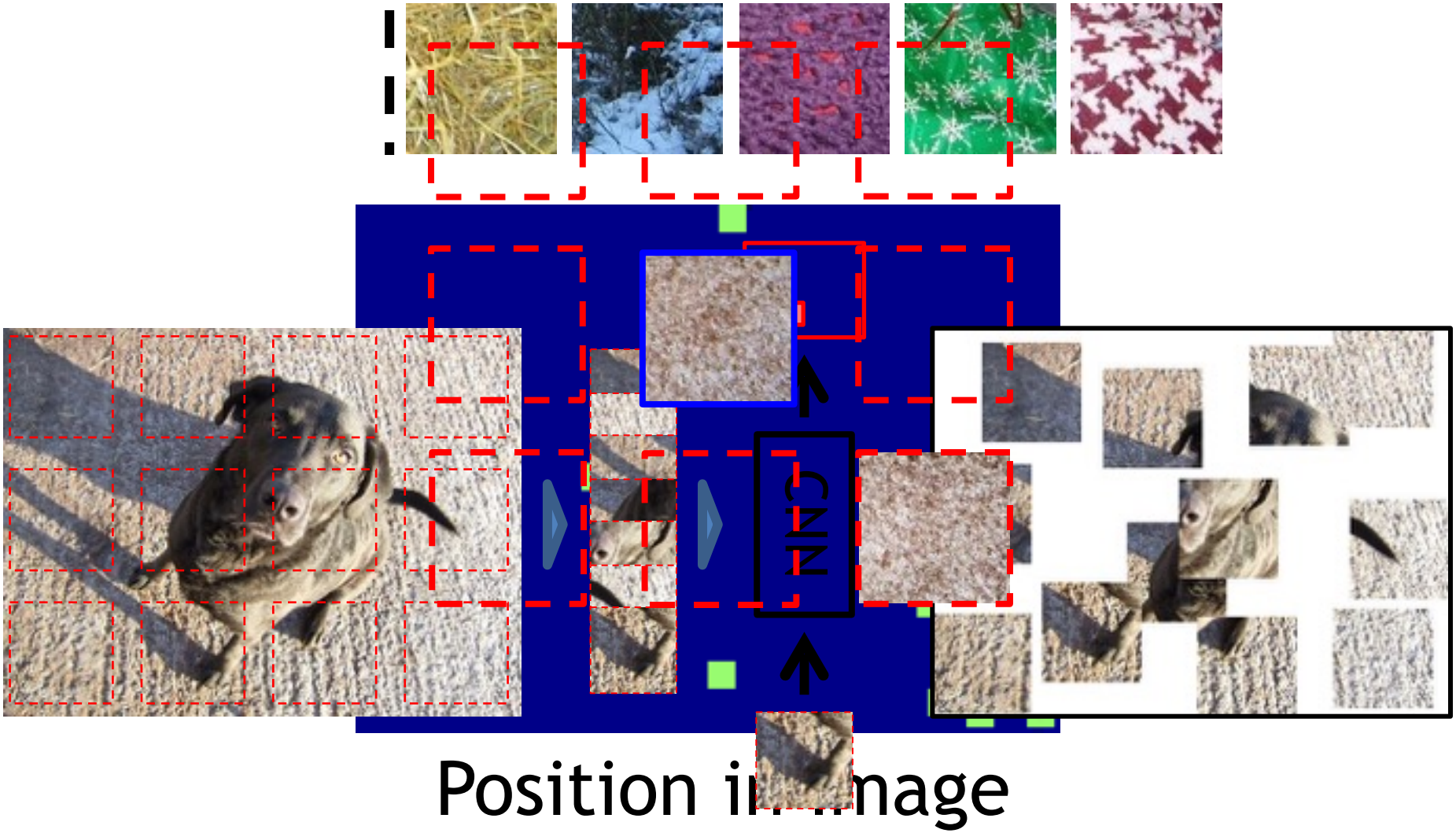


Include a gap

Jitter the patch locations

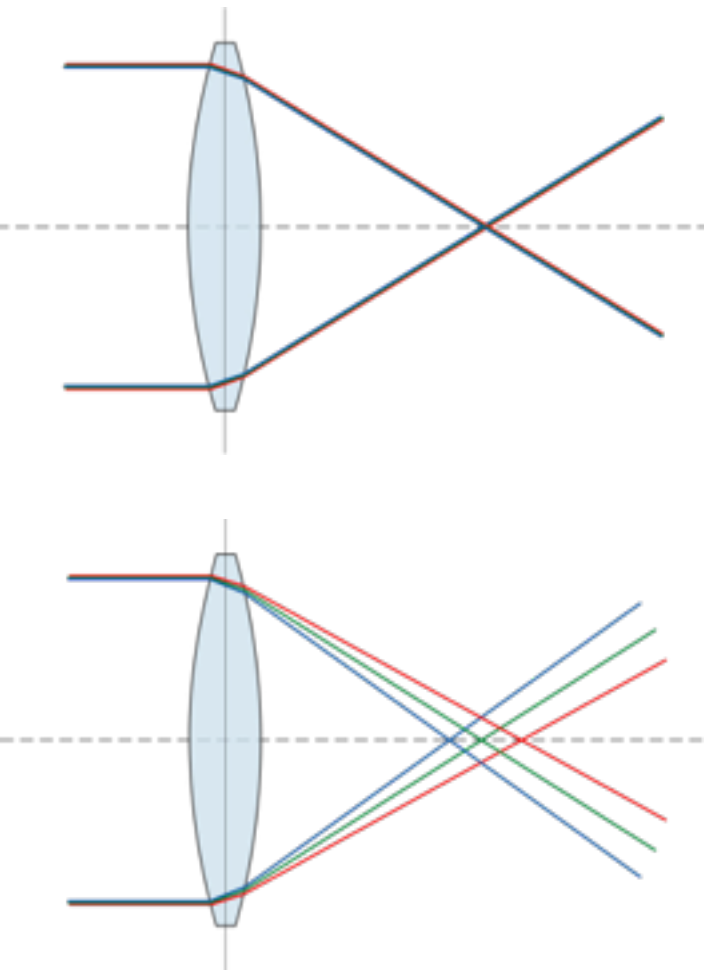
Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

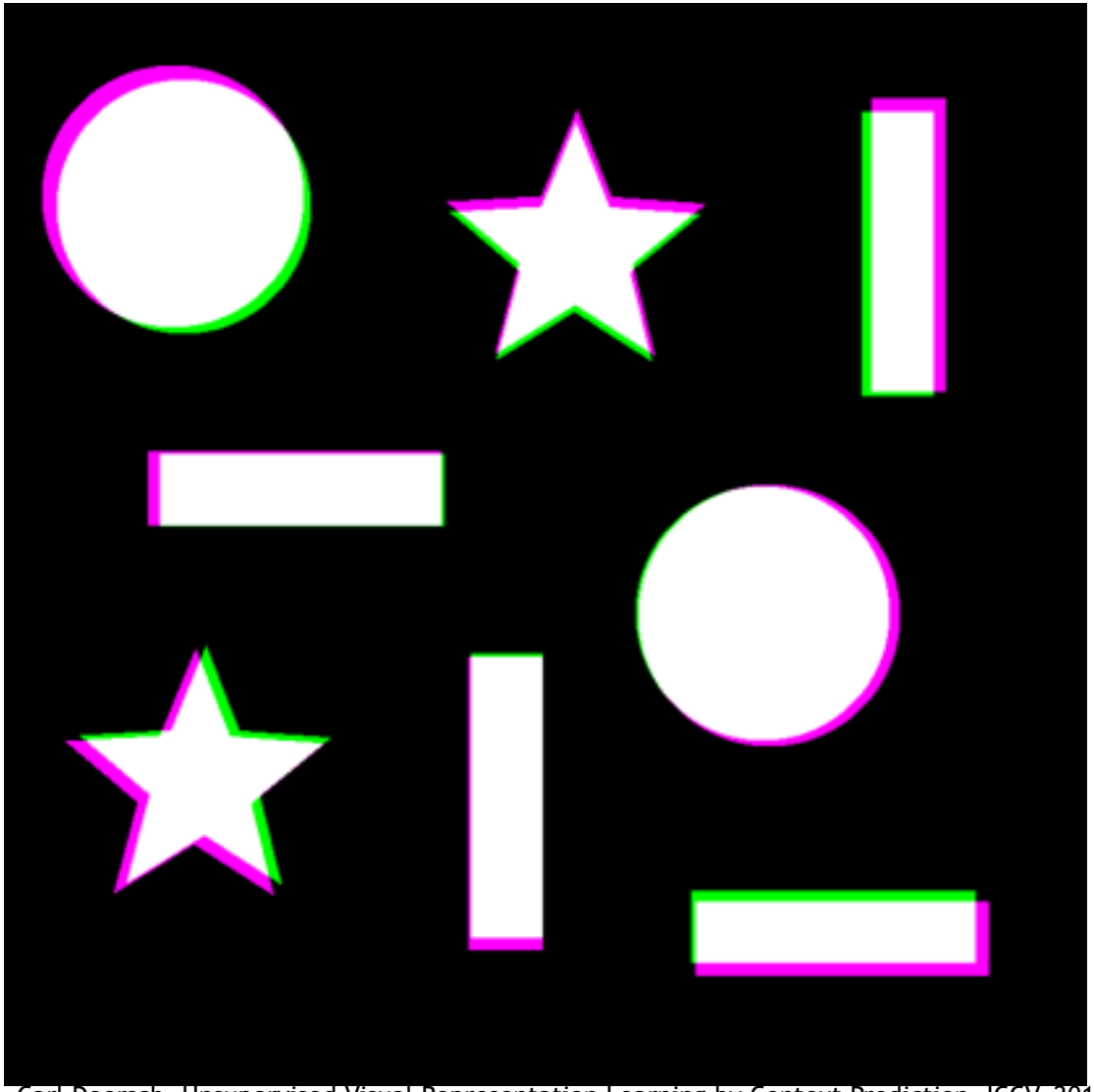
A Not-So “Trivial” Shortcut



Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

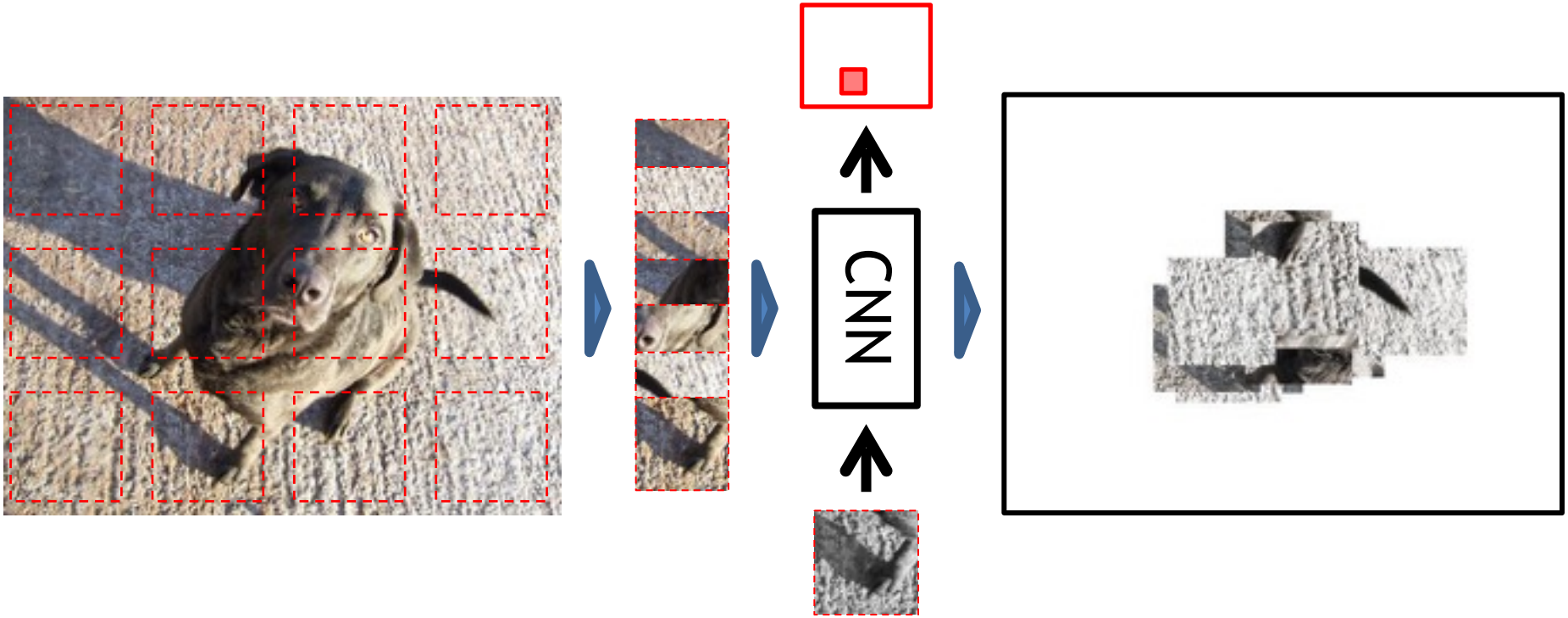
Chromatic Aberration





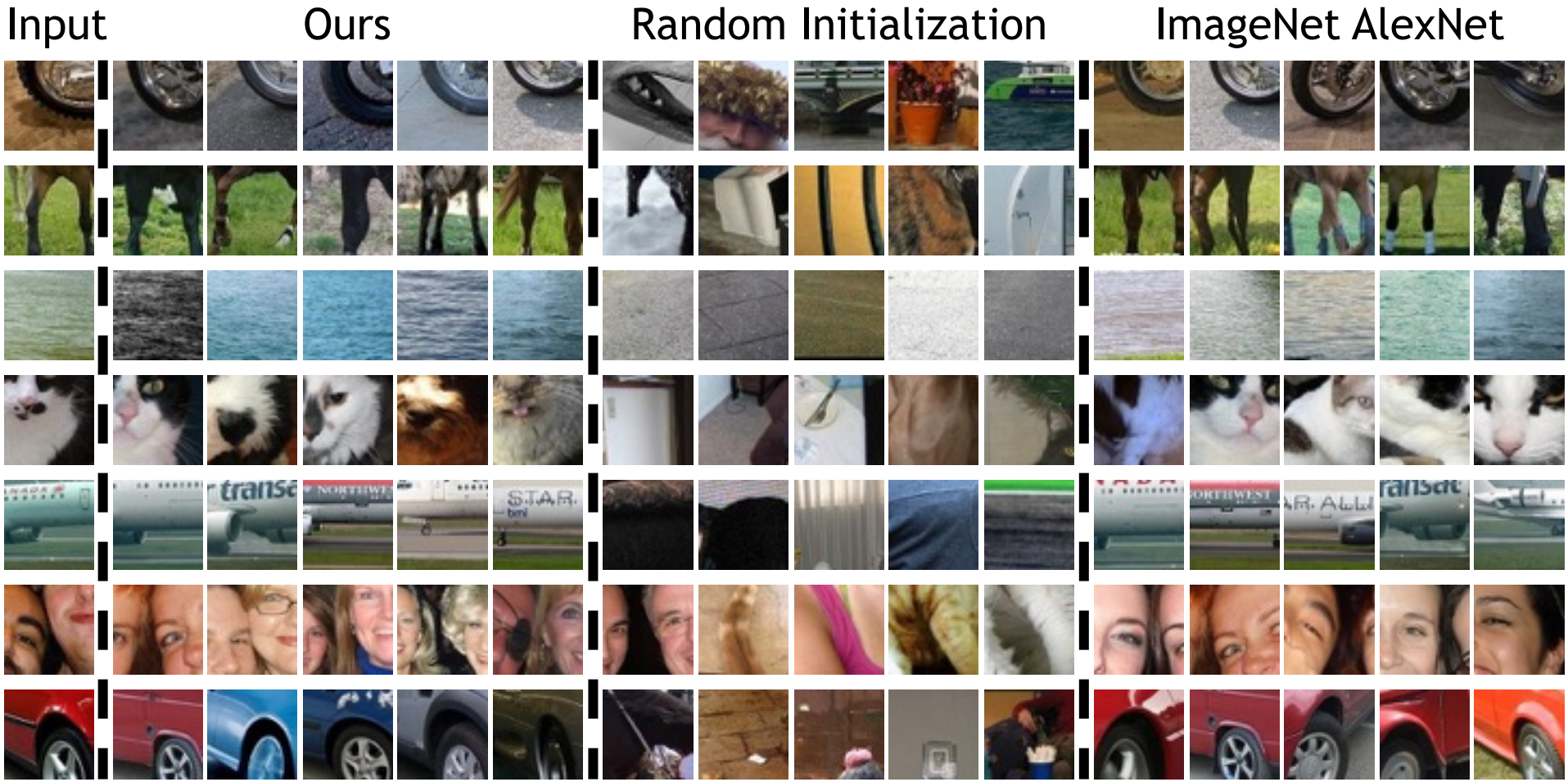
Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Chromatic Aberration



Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

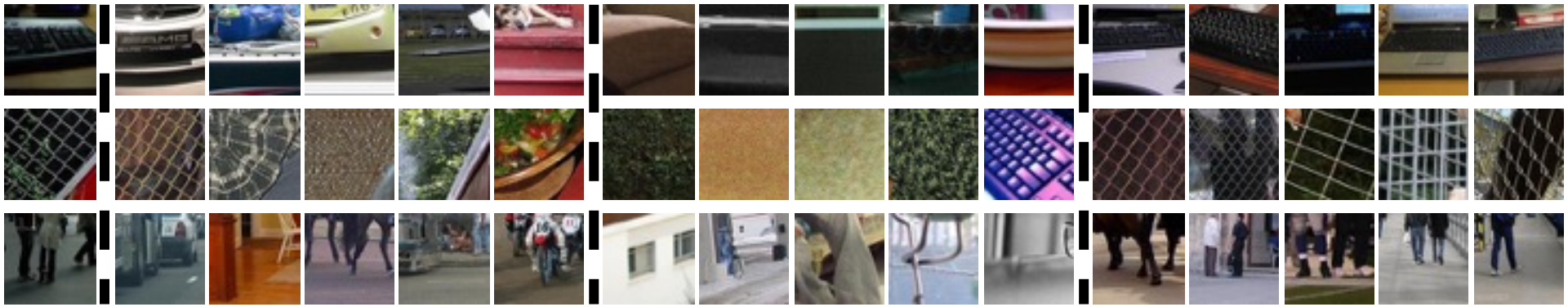
What is learned?



Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

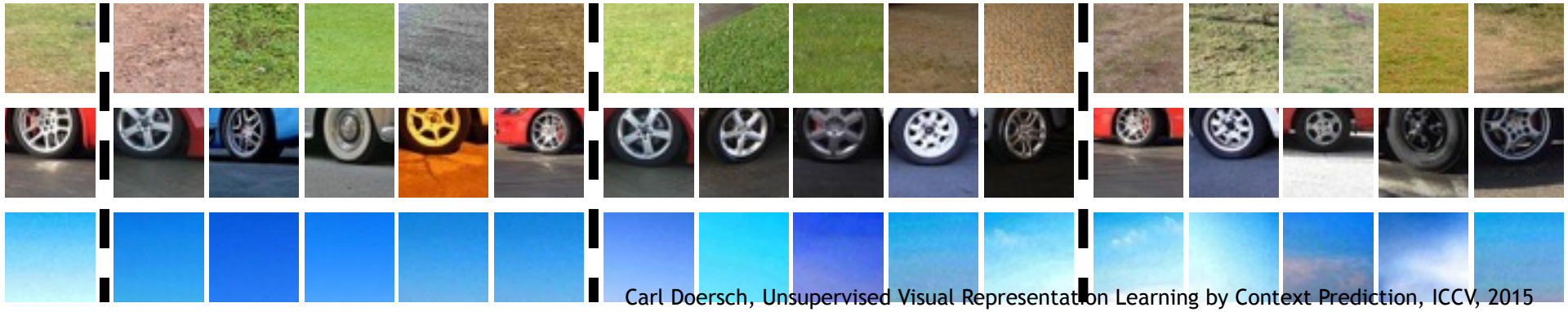
Still don’t capture everything

Input | Ours | Random Initialization | ImageNet AlexNet



You don’t always need to learn!

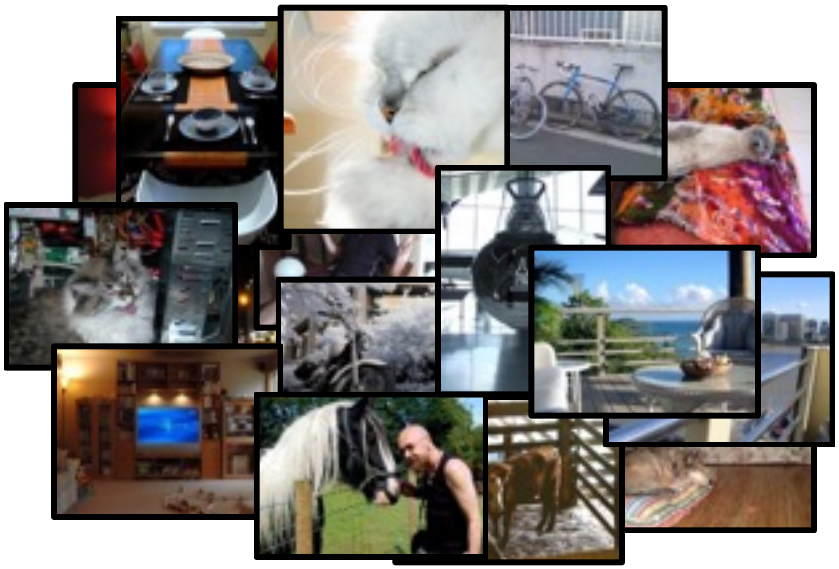
Input | Ours | Random Initialization | ImageNet AlexNet

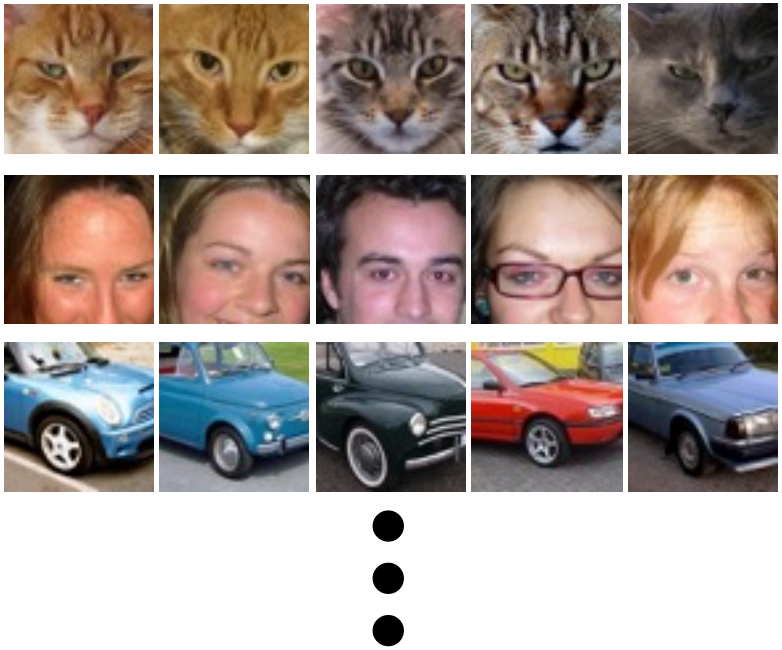


Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

- The trained network is applied to two domains:
  - Unsupervised object detection
    - Find sets of images or image segments in a large image collection which depict the same object
  - Used a pre-training for standard vision task
    - R-CNN

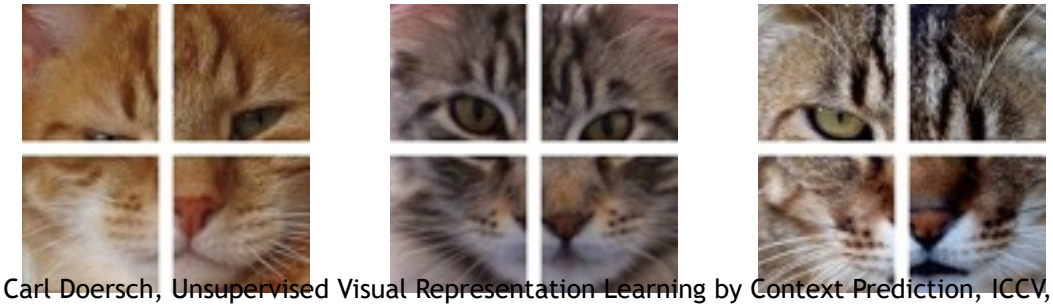
Visual Data Mining





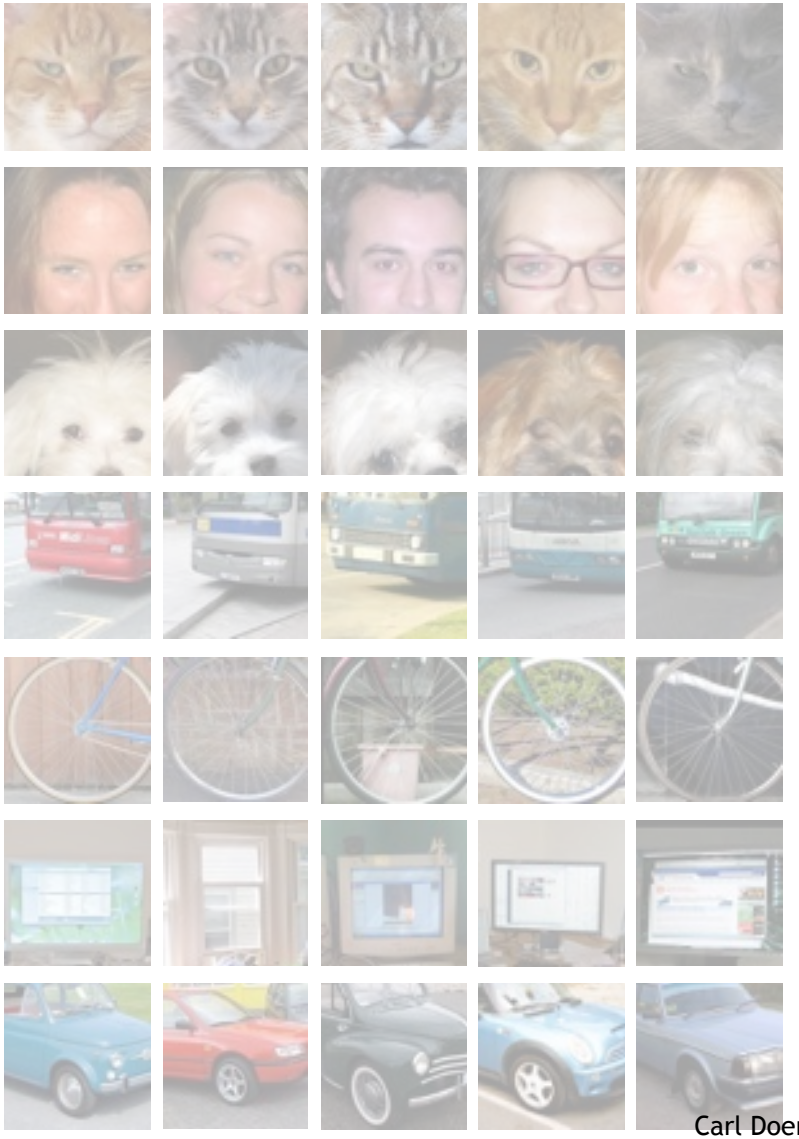
Via Geometric Verification

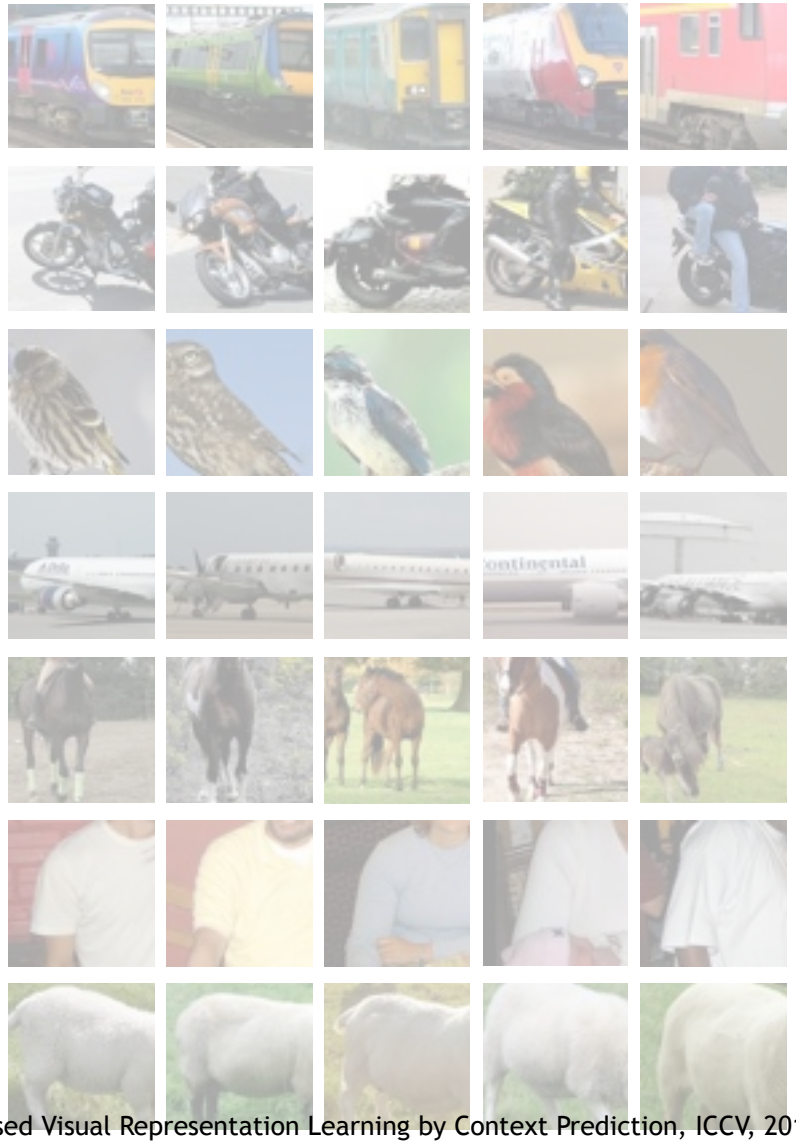
Simplified from [Chum et al 2007]



Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

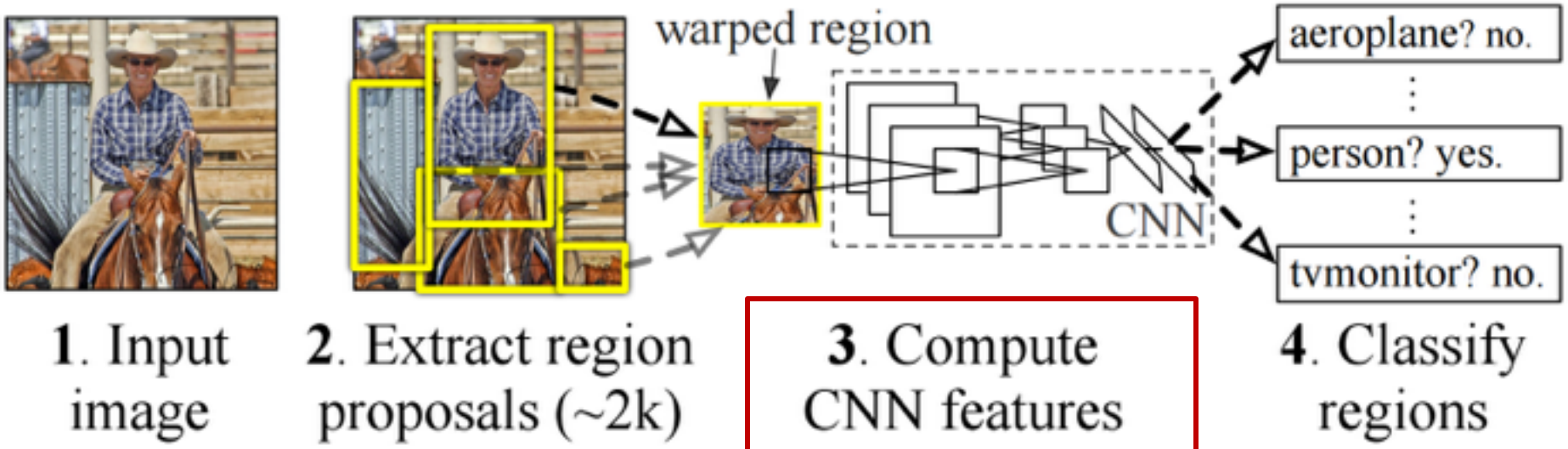
Mined from Pascal VOC2011





Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Pre-Training for R-CNN



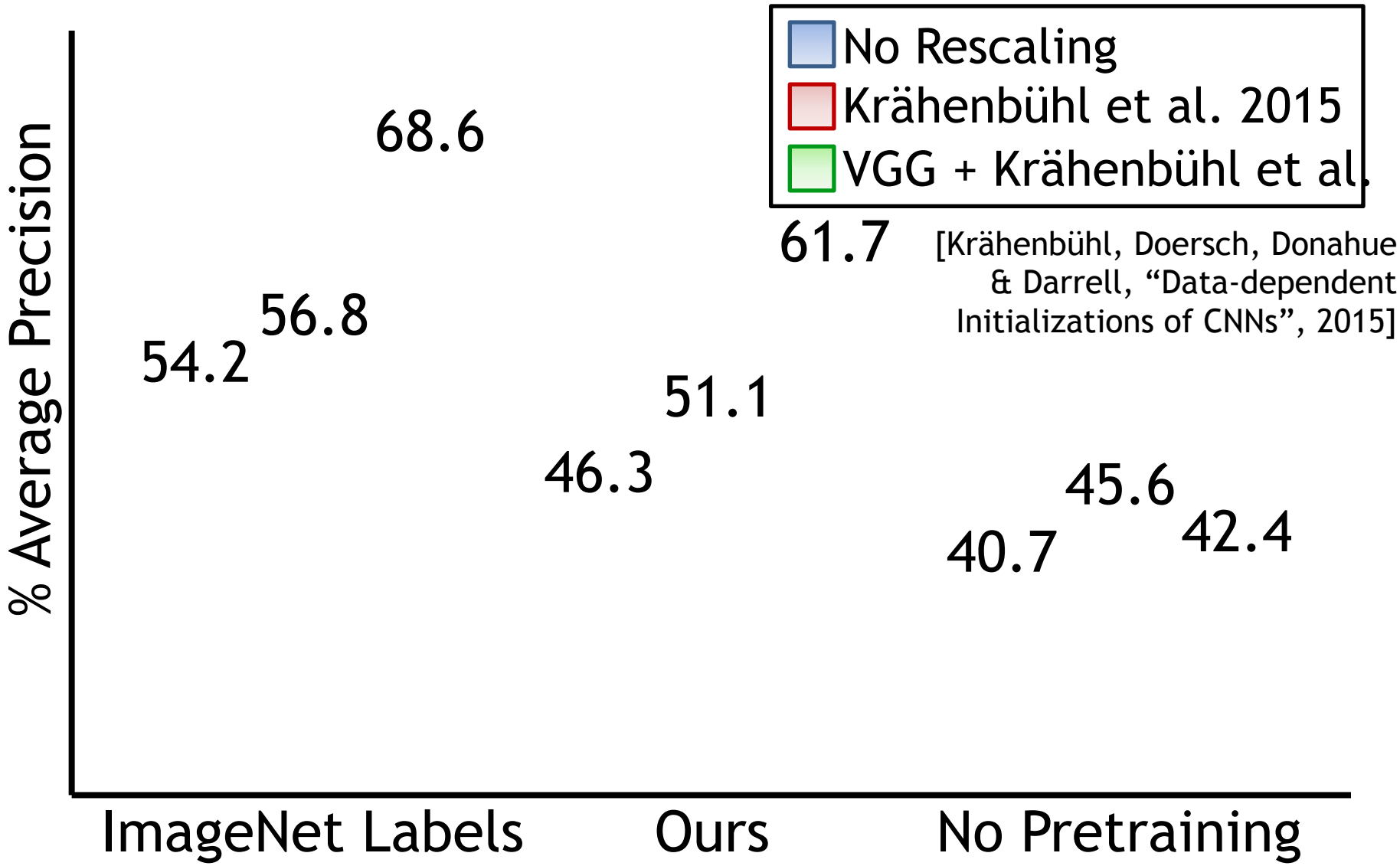
Pre-train on relative-position task, w/o labels

[Girshick et al. 2014]

Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

VOC 2007 Performance

(pretraining for R-CNN)

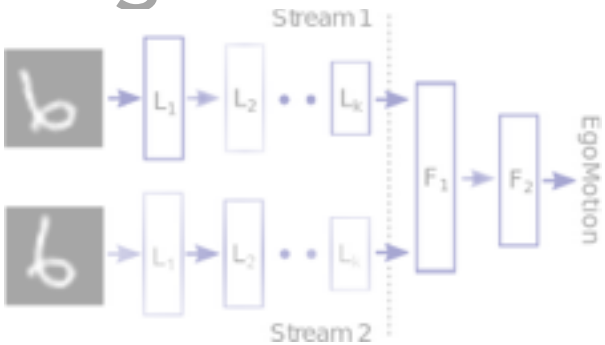


[Krähenbühl, Doersch, Donahue & Darrell, "Data-dependent Initializations of CNNs", 2015]

So, do we need semantic labels?

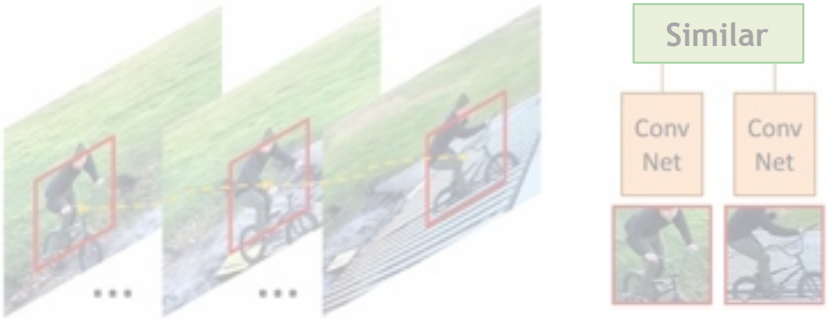
“Self-Supervision” and the Future

Ego-Motion



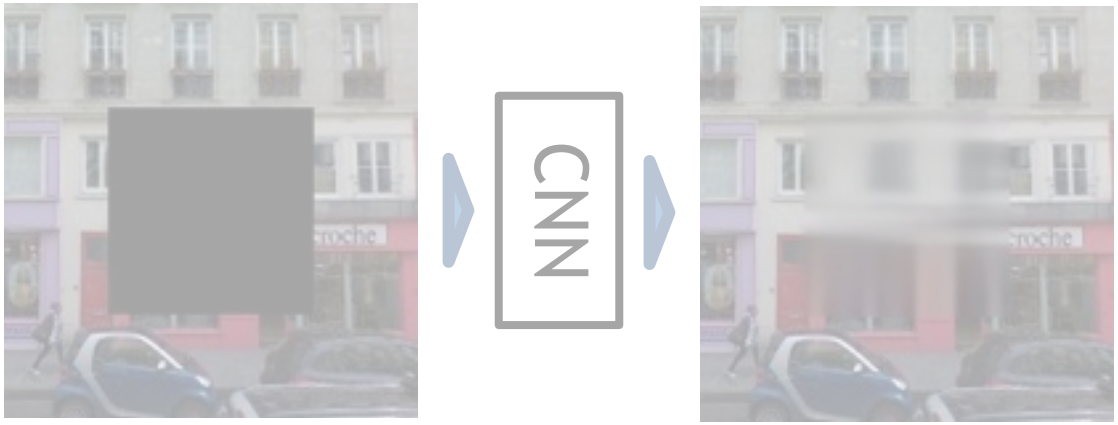
[Agrawal et al. 2015; Jayaraman et al. 2015]

Video



[Wang et al. 2015; Srivastava et al 2015; …]

Context



[Doersch et al. 2014; Pathak et al. 2015; Isola et al. 2015]

Carl Doersch, Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015

Strength and Weaknesses

- Unsupervised approach for feature learning
- High coverage rate of proposed objects, indicating learning invariances
- Used as pre-training, boost the R-CNN performance significantly
- Uncovers some connection of objects and respective surrounding

- Low performance on pretext task, while reason unclear

Possible Extensions

- Train on larger datasets
- Use DecovNets to uncover what is learned in each layer
- Trace the reasons for low performance by modifying the datasets, like the color aberration problem.

Thank you!