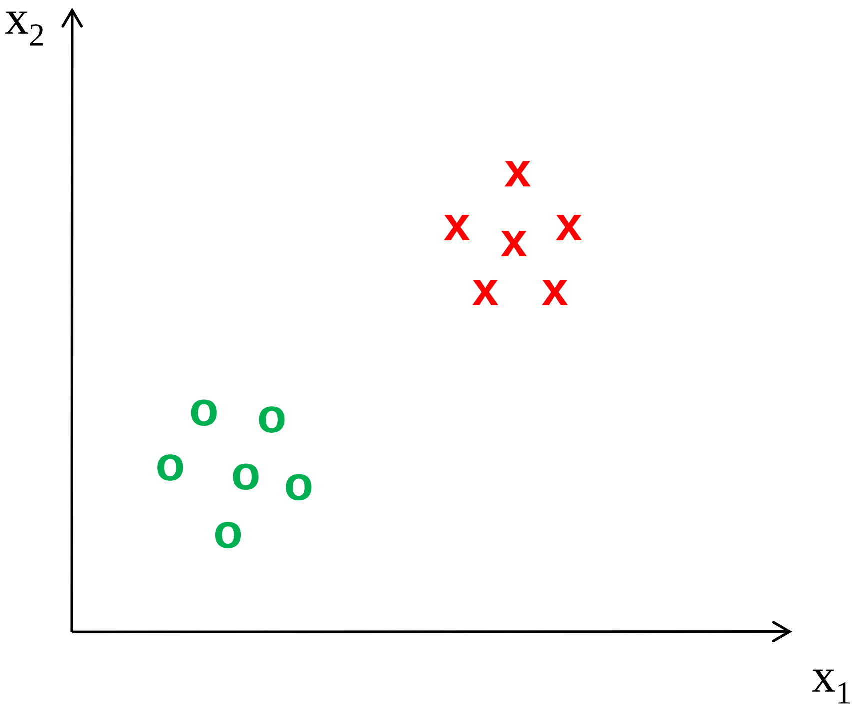
ECE 6554: Advanced Computer Vision
Spring 2017

Self-supervision or Unsupervised Learning of Visual Representation

Badour AlBahar

Supervised Learning

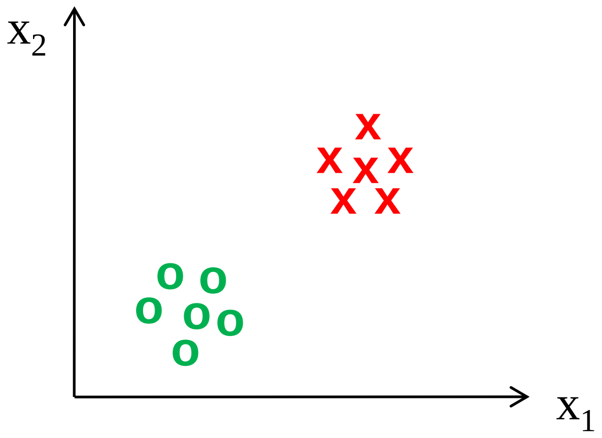


Data is labeled

Its goal is to learn to produce the correct output given a new input.

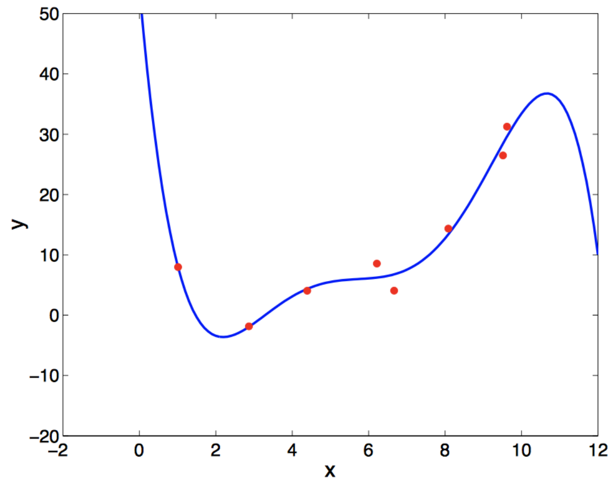
http://mlg.eng.cam.ac.uk/zoubin/course05/lect1.pdf

Supervised Learning



Classification

Output:
discrete class labels
Goal:
classify new inputs correctly

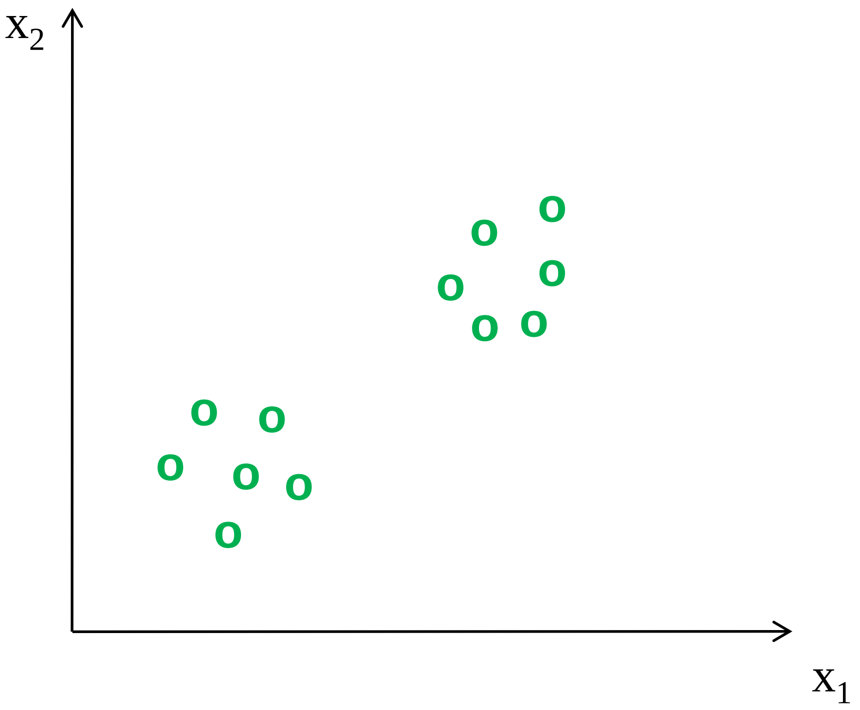


Regression

Output:
continuous values
Goal:
predict the output accurately for new inputs

http://mlg.eng.cam.ac.uk/zoubin/course05/lect1.pdf

Unsupervised Learning



Data is unlabeled

Its goal is to build a model that can be used for reasoning, decision making, predicting things, communicating, etc.

For example:

- finding clusters
- dimensionality reduction

http://mlg.eng.cam.ac.uk/zoubin/course05/lect1.pdf

Motivation and Strengths:

- Unsupervised learning is not expensive and time consuming like supervised learning.
- Unsupervised learning requires no human intervention.
- Unlabeled data is easy to find with large quantities, unlike labeled data which is scarce.

Weaknesses:

More difficult than supervised learning because there is **NO**:

- ❖ Gold standard (like an outcome variable)
- ❖ Single objective (like test set accuracy)

Unsupervised Visual Representation Learning by Context Prediction

C. Doersch, A. Gupta, A. A. Efros

ICCV 2015

- Semantic labels from humans are expensive.

Do we need semantic labels in order to learn a useful representation?

Or is there some other ***"Less Expensive"*** *pretext task that will learn something similar?*

Slide: Carl Doersch

Context Prediction

Given a pair of patches from one image.
Can you say where they go relative to one another?

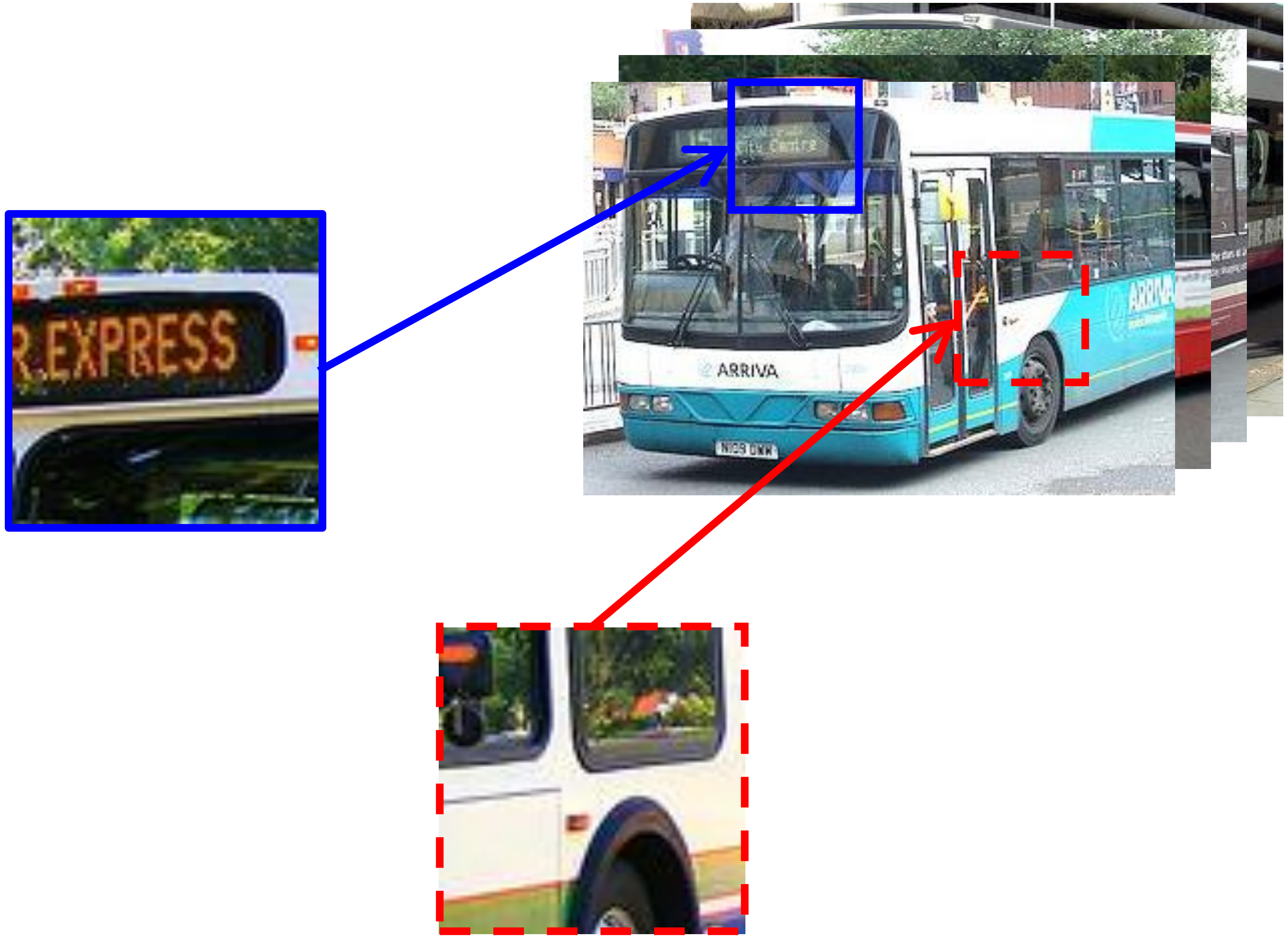
A

B

Slide: Carl Doersch

Context Prediction for Images

Semantics from a non-semantic task

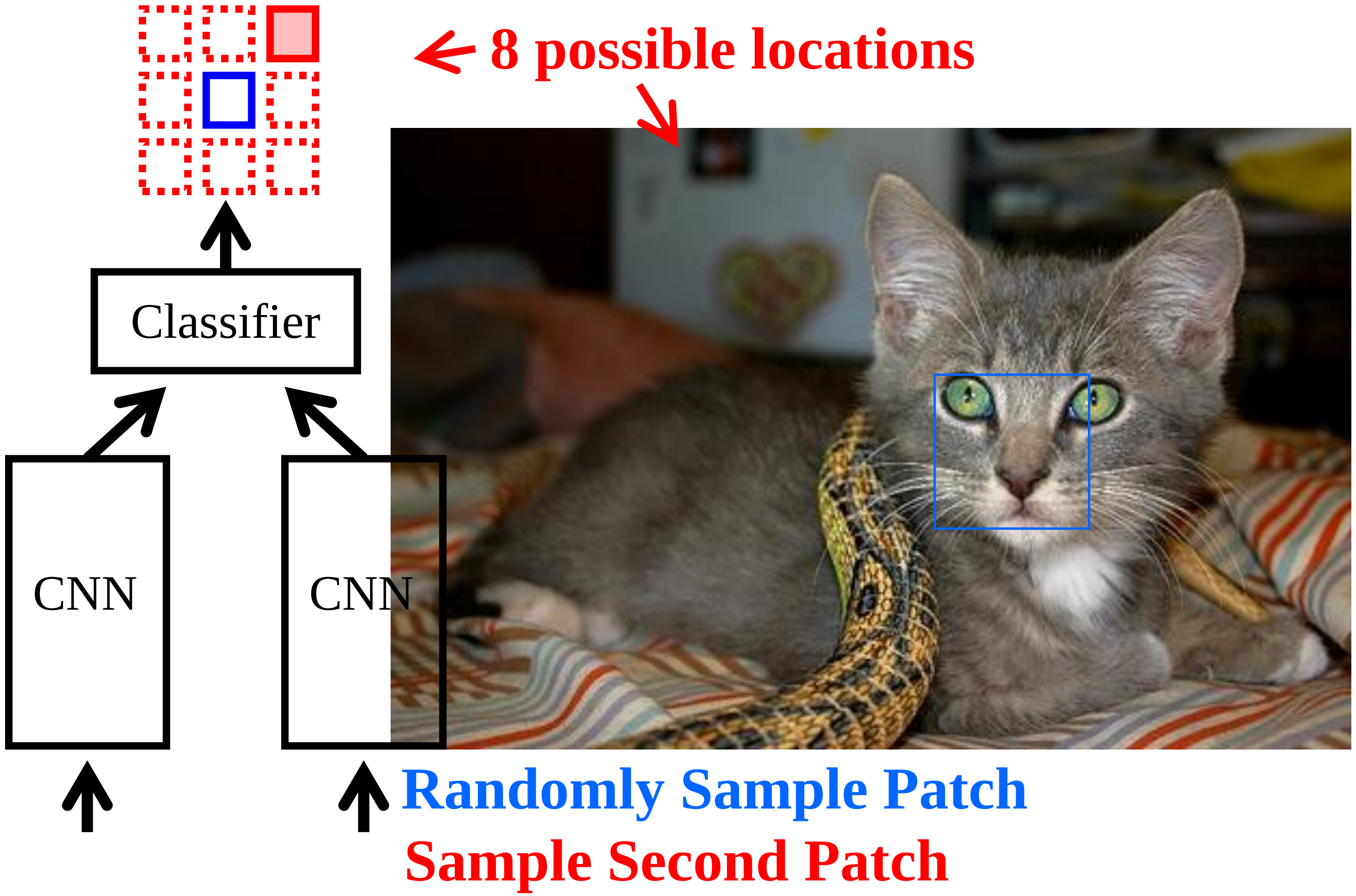


Slide: Carl Doersch

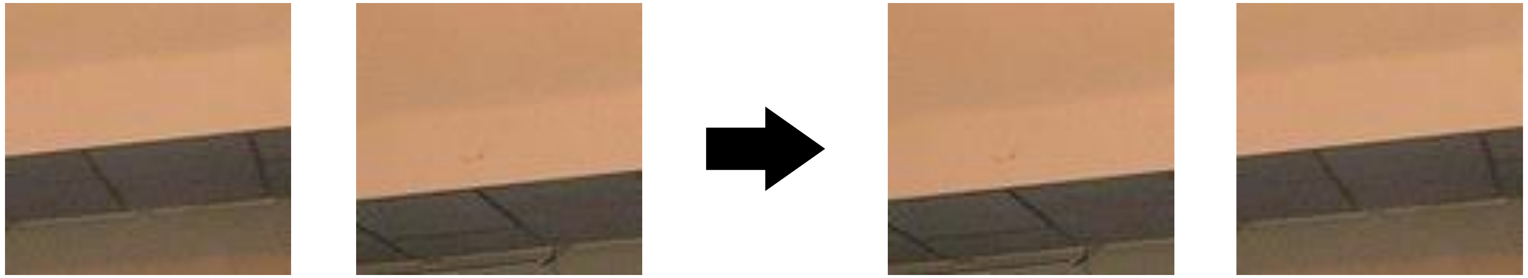
Relative Position Task

unlabeled image

Relative Position Task



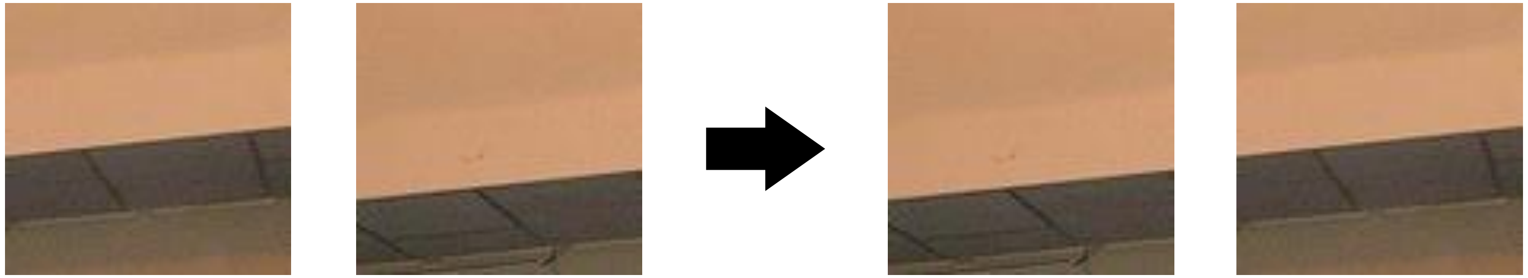
Avoiding Trivial Shortcuts

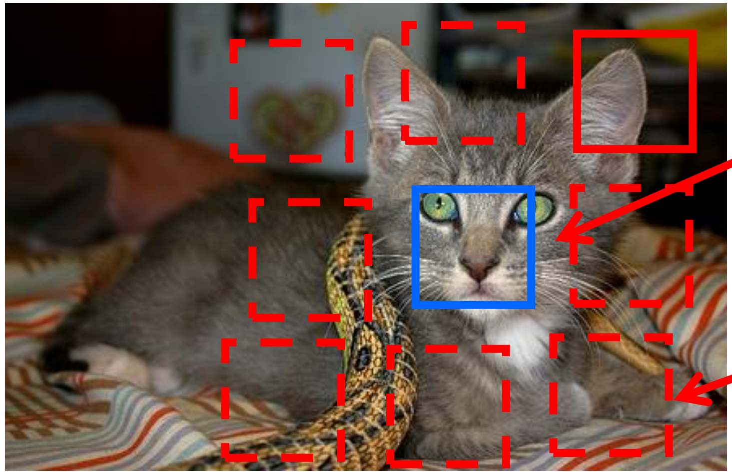


Ways that the network can solve the problem without really extracting the semantics that we're after.

Slide: Carl Doersch

Avoiding Trivial Shortcuts





Include a gap

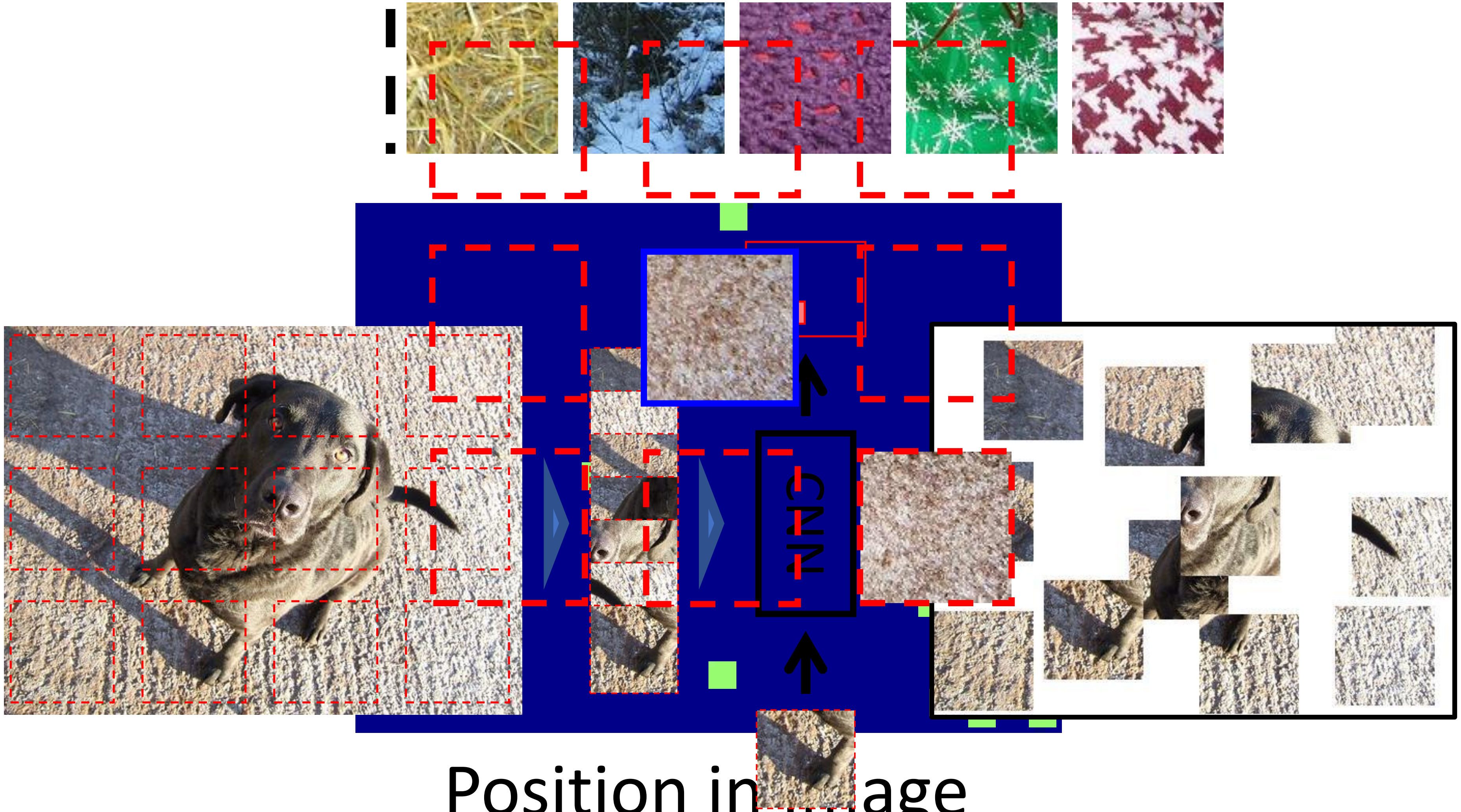
makes it less likely that low-level properties cross both patches

Jitter the patch locations

makes it harder to match straight lines between two patches

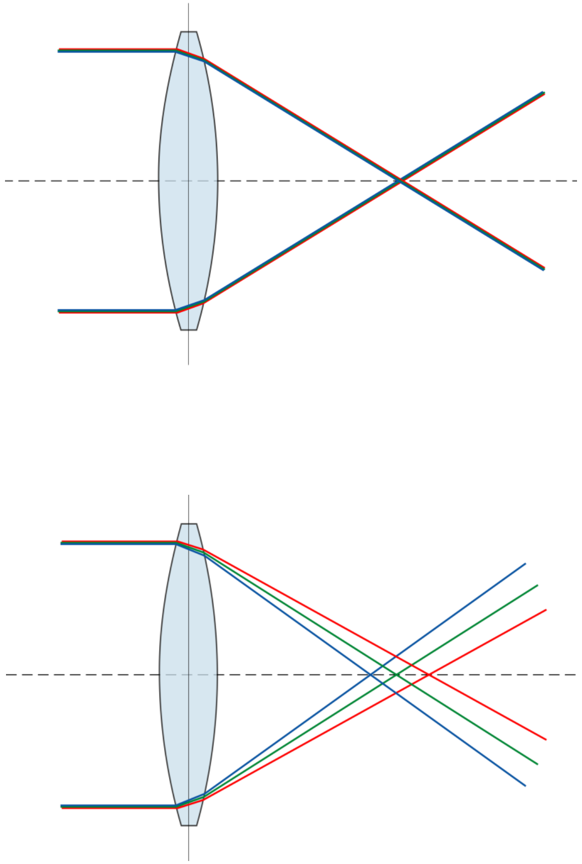
Slide: Carl Doersch

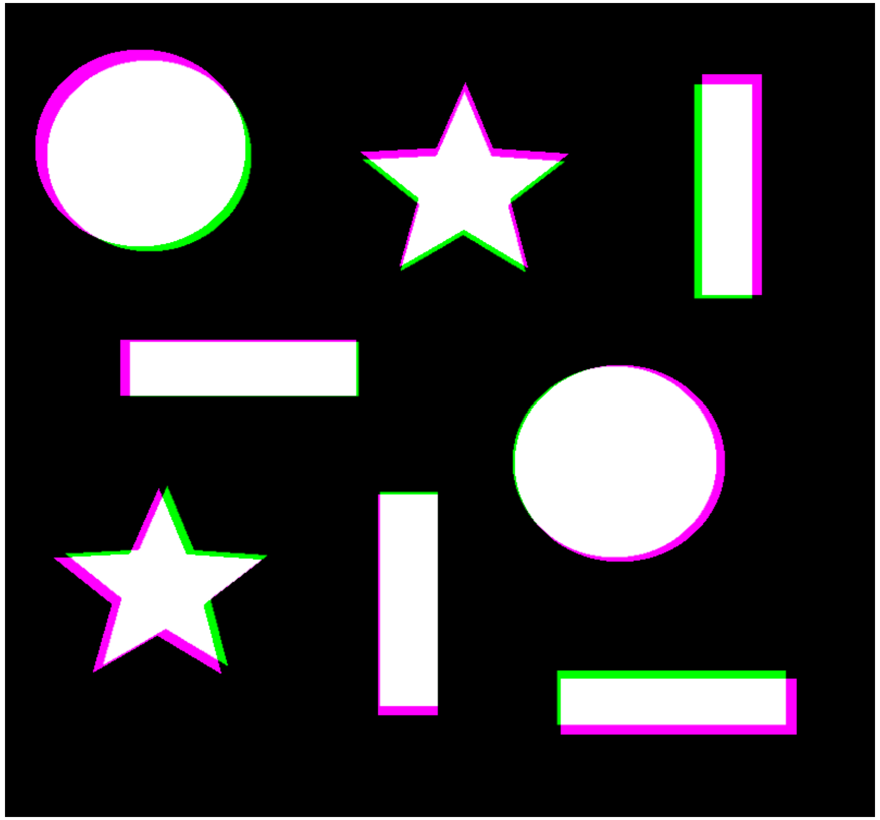
A Not-So “Trivial” Shortcut



Position in image

Chromatic Aberration

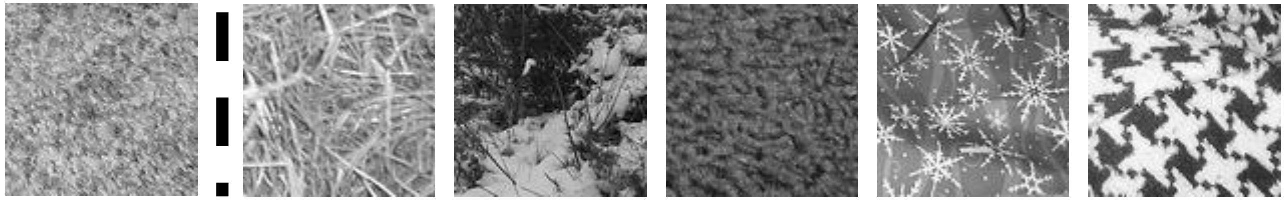




- Chromatic aberration happens when a lens bends different wavelengths at different amount
- Deep nets can detect this subtle shift, which tells the net where a patch is with respect to the lens, and gives away the answer to the relative-position task.
- For common lenses (specifically, the achromatic doublet), the green color channel is shrunk a little bit toward the image center relative to red and blue

Slide: Carl Doersch

Solution

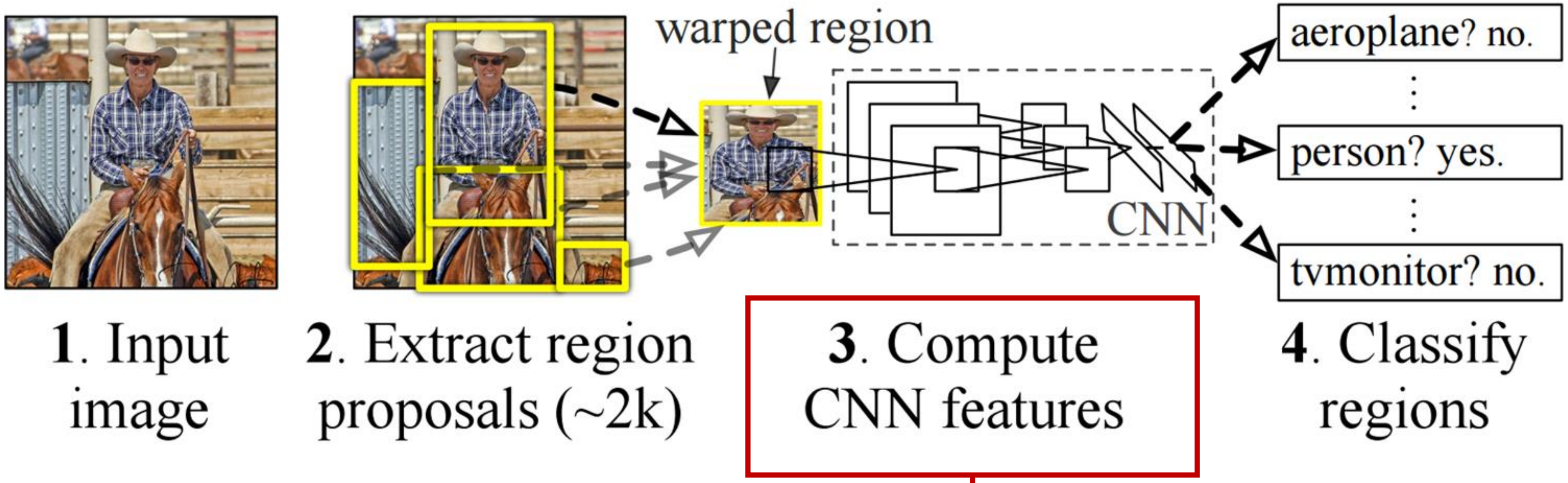


Removing color
In this paper, 2 of the 3 color channels are randomly dropped.

Important lesson:
Deep nets are kind-of lazy. If there's a way to solve a problem without learning semantics, they may learn to do that instead.

Slide: Carl Doersch

Pre-Training for R-CNN

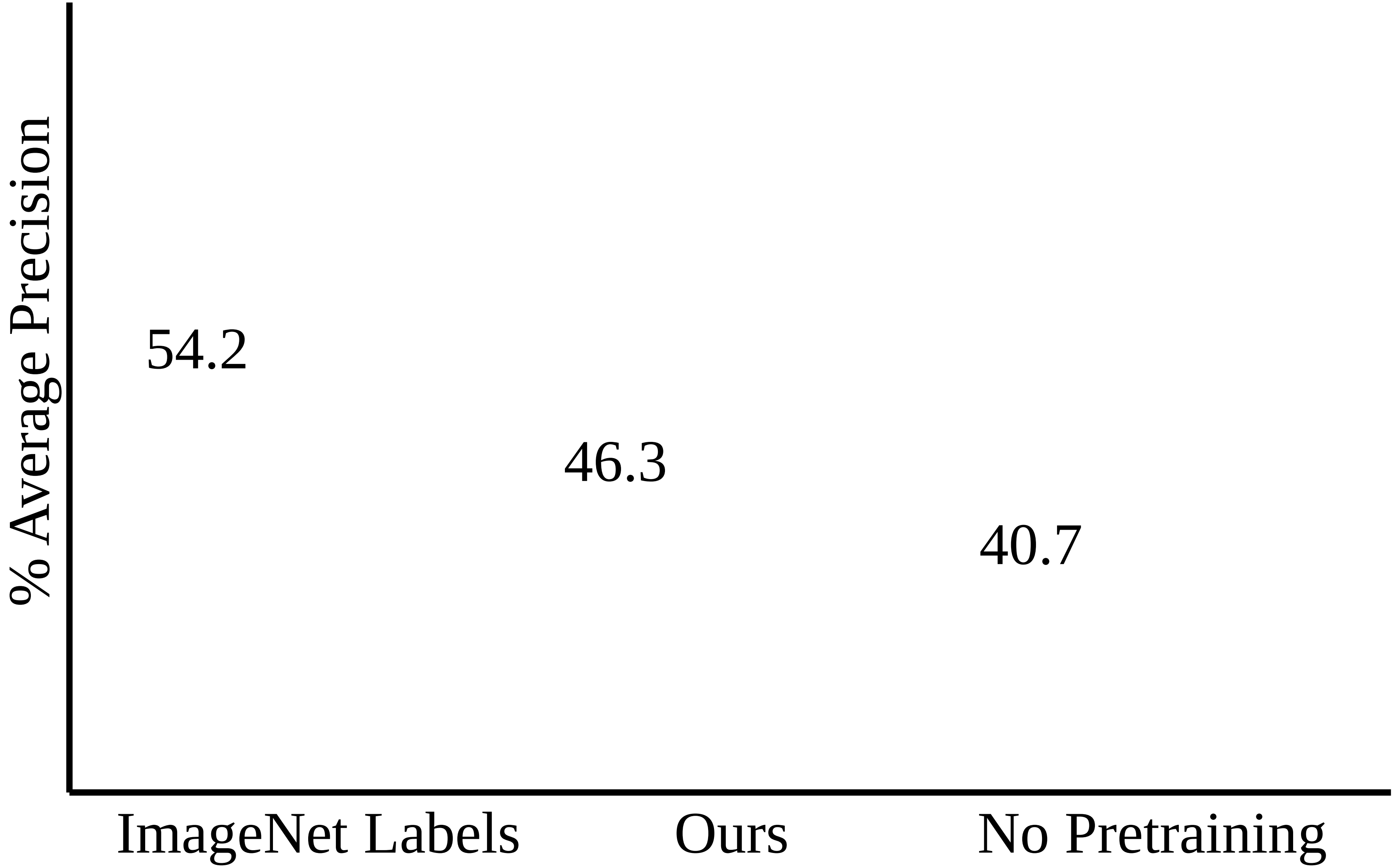


Pre-train on relative-position task, w/o labels

[Girshick et al. 2014]

Pascal Object Detection: VOC 2007 Performance

(pretraining for R-CNN)



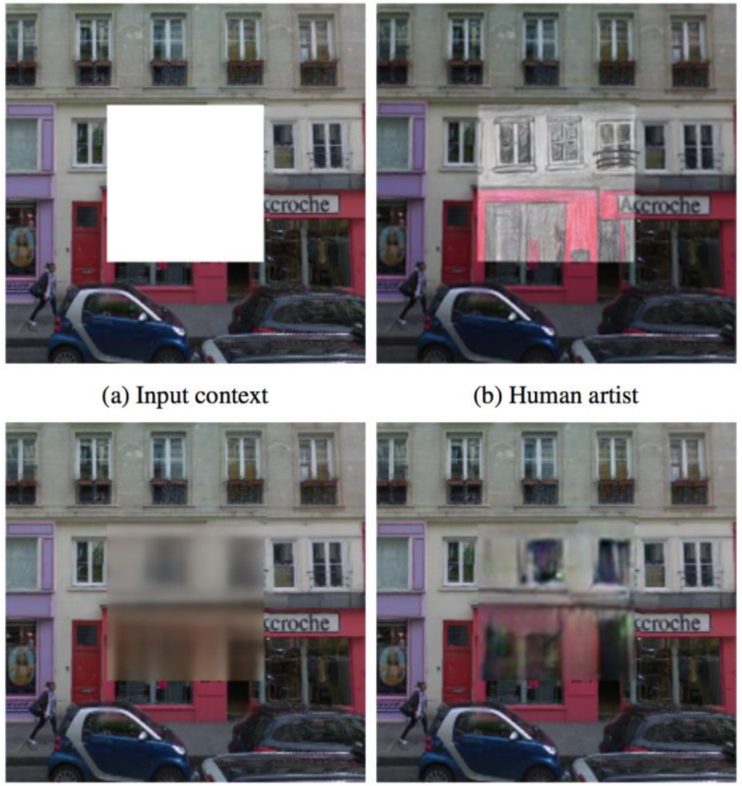
Slide: Carl Doersch

Context Encoders: Feature Learning by Inpainting

D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A. A. Efros, CVPR 2016

Inpainting:

The art of restoring missing parts of image.



(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

Context Encoders: Feature Learning by Inpainting

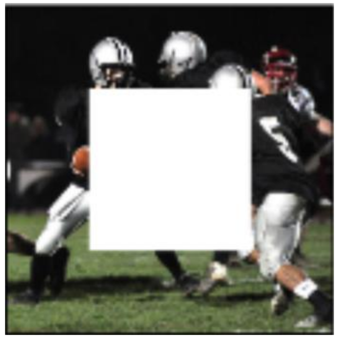
Classical inpainting or texture synthesis approaches are

local non-semantic methods

Hence, they cannot handle large missing region.

Context Encoders: Feature Learning by Inpainting

- Unsupervised semantic visual feature learning
- Semantic inpainting



<u>Input:</u>

an image with a missing region

<u>Output:</u>

the missing region

Context Encoders: Feature Learning by Inpainting

Output: a latent feature representation of that image

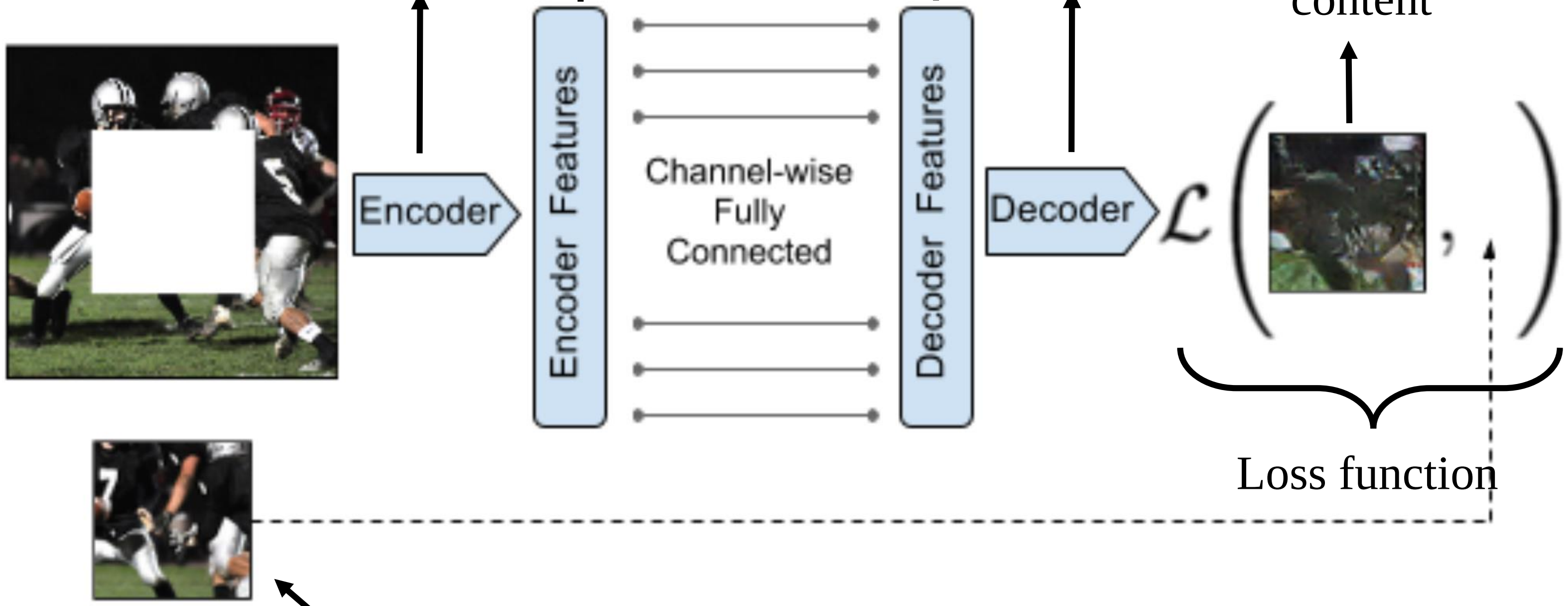
Input: a latent feature representation of that image

Encoder: captures the context of an image into a compact latent feature representation

Decoder: fill in realistic image content

Output: the missing image content

Input: image with the missing region



Loss function

Ground truth

Loss function

- **Standard pixel-wise reconstruction loss (L2):**

Tries to minimize the distance between the predicted missing region and the ground truth.

<u>produces blurry results</u>

- **Reconstruction plus an adversarial loss:**

Tries to make the predicted missing region as realistic as possible.

<u>Produces much sharper results</u>

Input

L2 loss

L2 + adversarial loss

Results

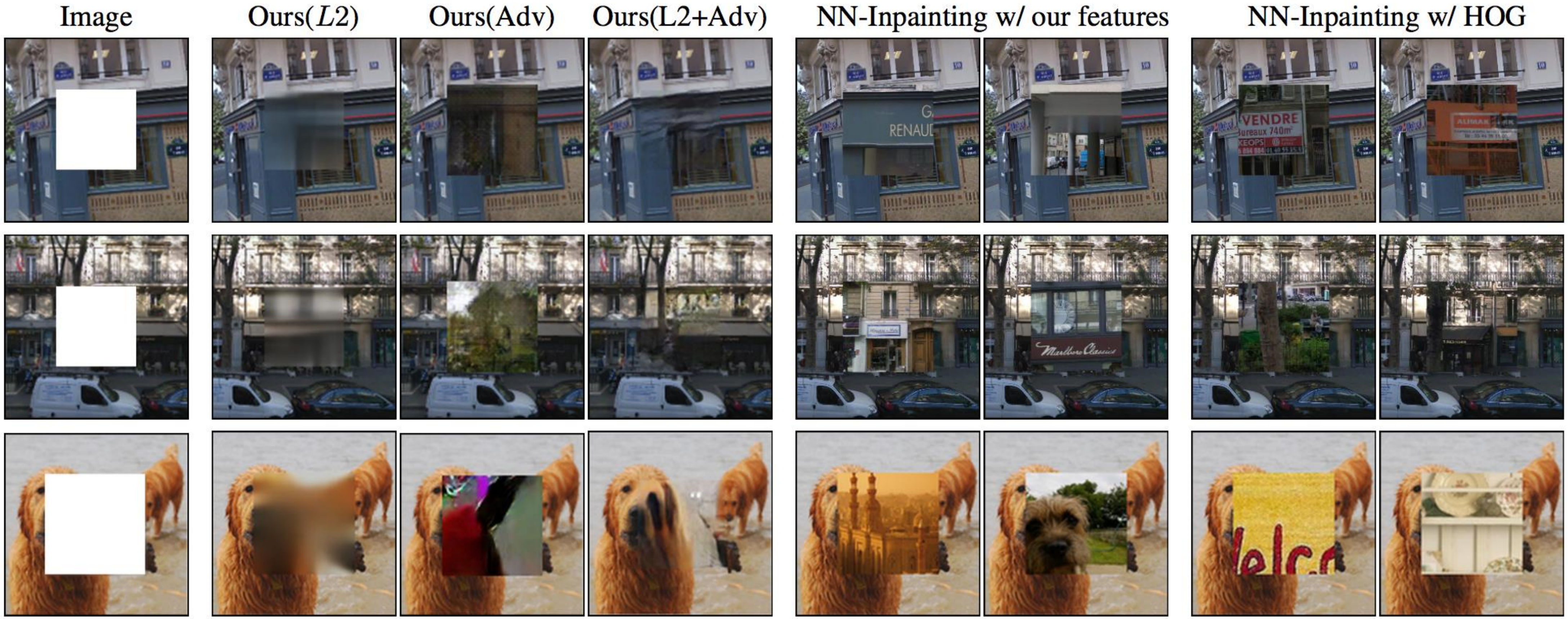


Image
Ours($L2$)
Ours(Adv)
Ours(L2+Adv)
NN-Inpainting w/ our features
NN-Inpainting w/ HOG

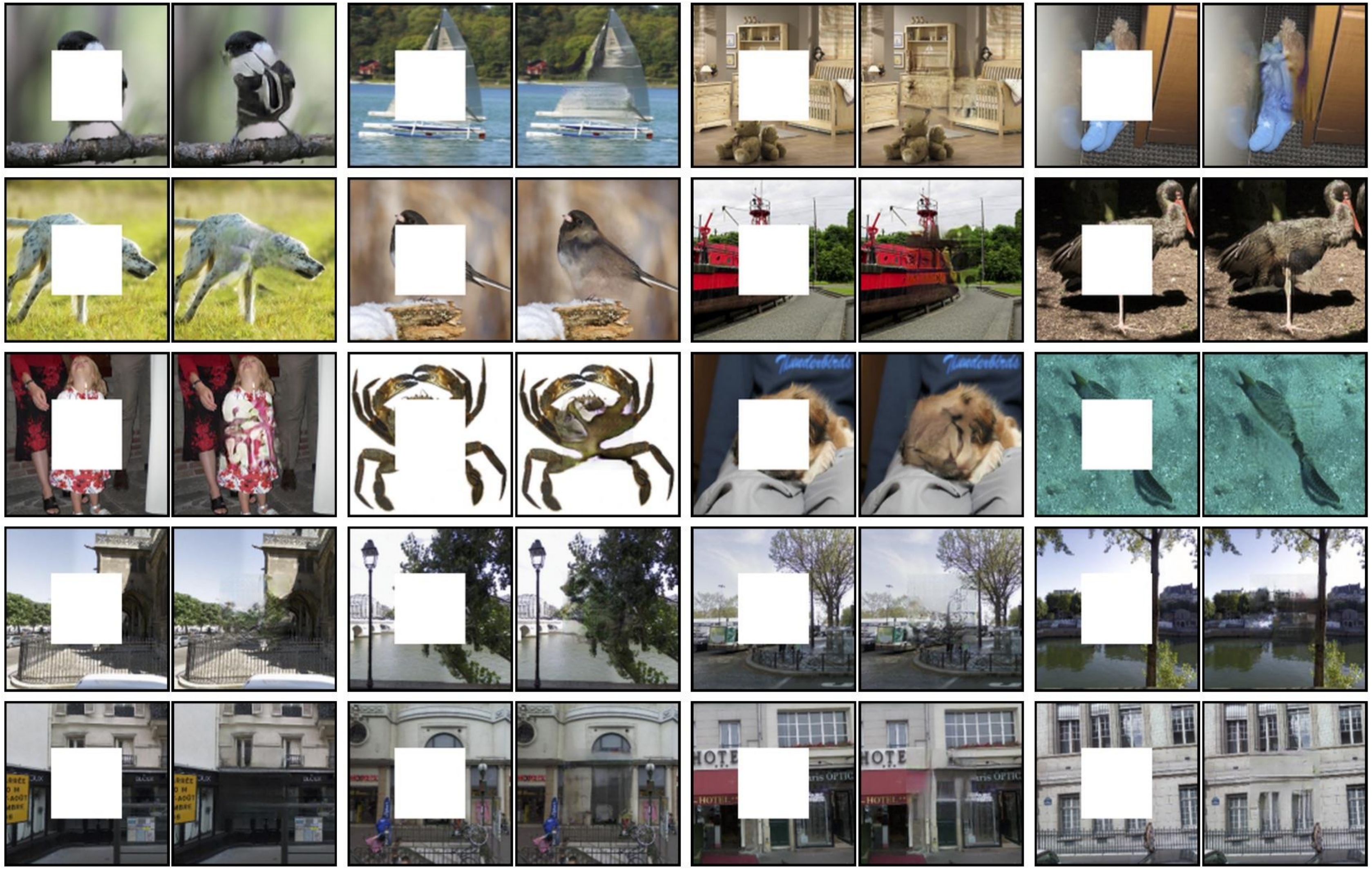
ECE 6554: Advanced Computer Vision
Spring 2017

EXPERIMENT

Context Encoders: Feature Learning by Inpainting

Badour AlBahar

Results:



Results:

Original **Input** **Output**

Results:

Original

Input

Output

Results:

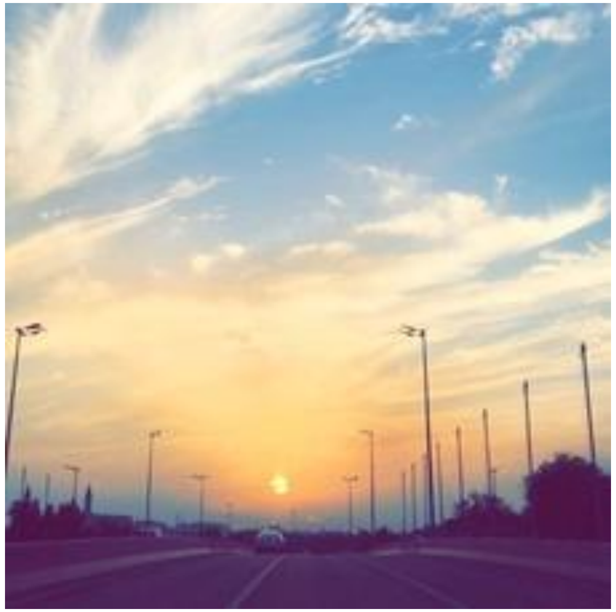
Original

Input

Output

Results:

Original



Input

Output

Results:

Original

Input

Output

Results:

Original

Input

Output

Results:

Original

Input

Output

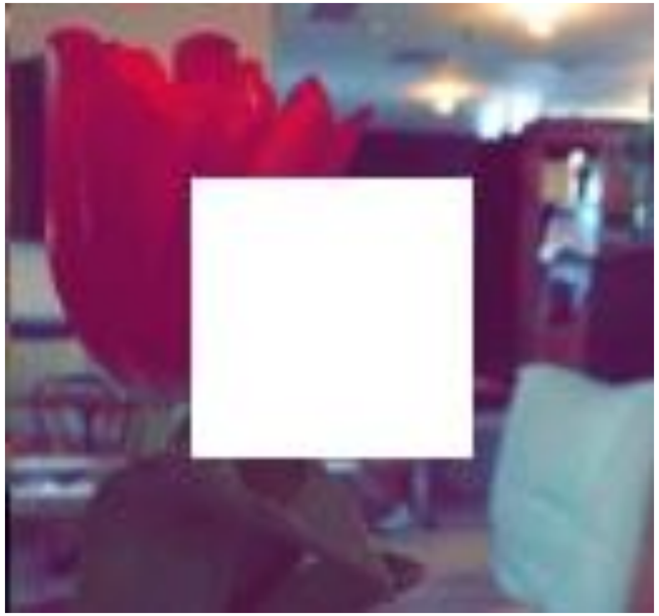
Results:

Original	Input	Output
		

Results:

Original

Input



Output

Results:

Original

Input

Output

Results:

Original

Input

Output

Results:

Original

Input

Output

Results:

Original

Input

Output

Results:

Original

Input

Output

Results:

Original

Input

Output

Thank you!