CS 2770: Computer Vision

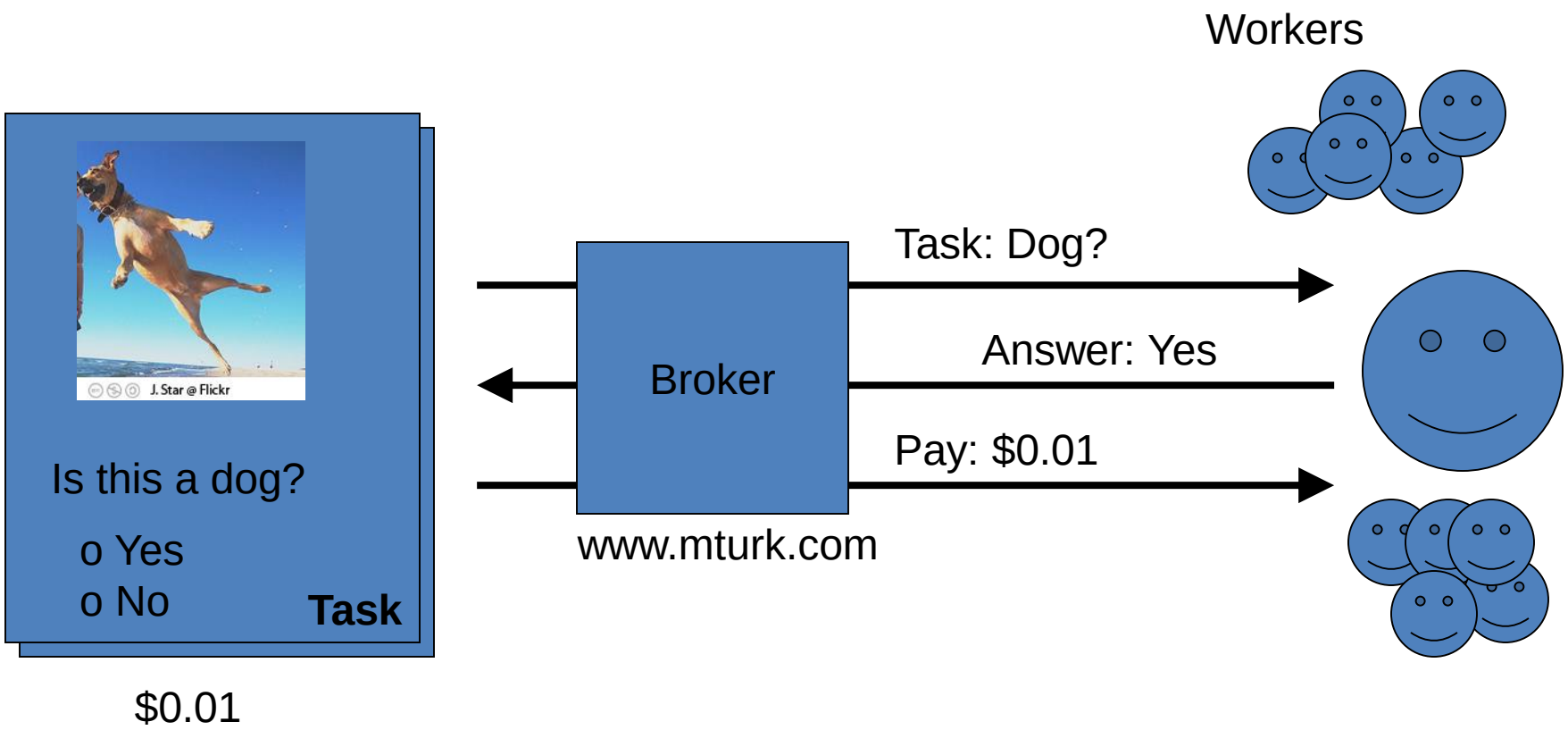
Self-Supervised Learning & Visual Discovery

Prof. Adriana Kovashka
University of Pittsburgh
April 10, 2017

Motivation

- So far we've assumed access to plentiful labeled data
- How is this data obtained?

Crowdsourcing



Alex Sorokin

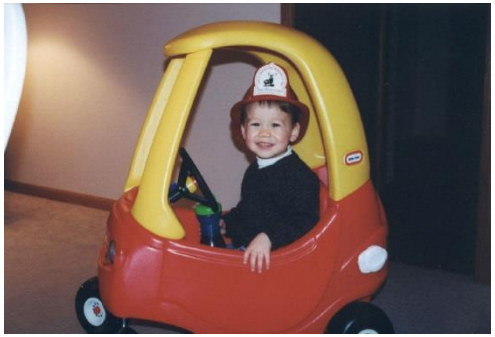
Crowdsourcing via games

- The ESP Game
  - Two-player online game
  - Partners don’t know each other and can’t communicate
  - Objective of the game: type the same word
  - The only thing in common is an image

von Ahn and Dabbish, “Labeling Images with a Computer Game,” CHI 2004

THE **ESP GAME**

PLAYER 1



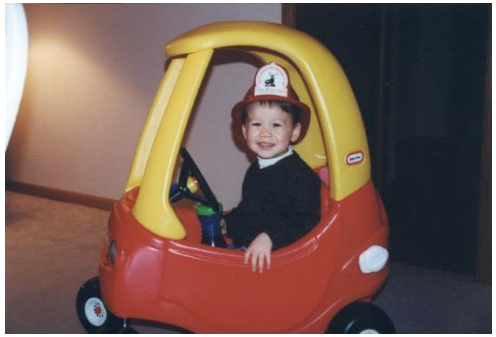
GUESSING: CAR

GUESSING: HAT

GUESSING: KID

SUCCESS!
YOU AGREE ON CAR

PLAYER 2



GUESSING: BOY

GUESSING: CAR

SUCCESS!
YOU AGREE ON CAR

von Ahn and Dabbish, “Labeling Images with a Computer Game,” CHI 2004

Motivation

- So far we've assumed access to plentiful labeled data
- **What if we have limited or no labeled data?**
- One approach: learn from unlabeled data (unsupervised learning)
  - Mine for interesting patterns (discovery)
  - Use supervision (labels) inherent in the data (self-supervised learning)
- Another approach (not discussed): carefully choose which data to label
  - Active learning, human-in-the-loop

Plan for this lecture

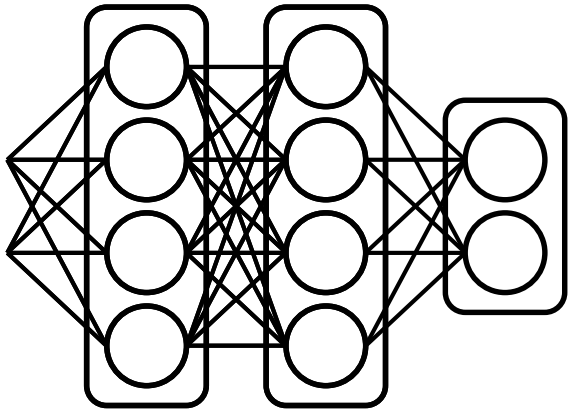
- Self-supervised learning
  - From surrounding regions (ICCV 2015)
  - From temporal order (ECCV 2016)
  - From ego-motion (ICCV 2015)
- Visual discovery
  - From surrounding objects (CVPR 2010)
  - Elements of architectural styles (SIGGRAPH 2012)
  - Gradual style changes in time/space (ICCV 2013)

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei Efros and Abhinav Gupta

ICCV 2015

ImageNet + Deep Learning

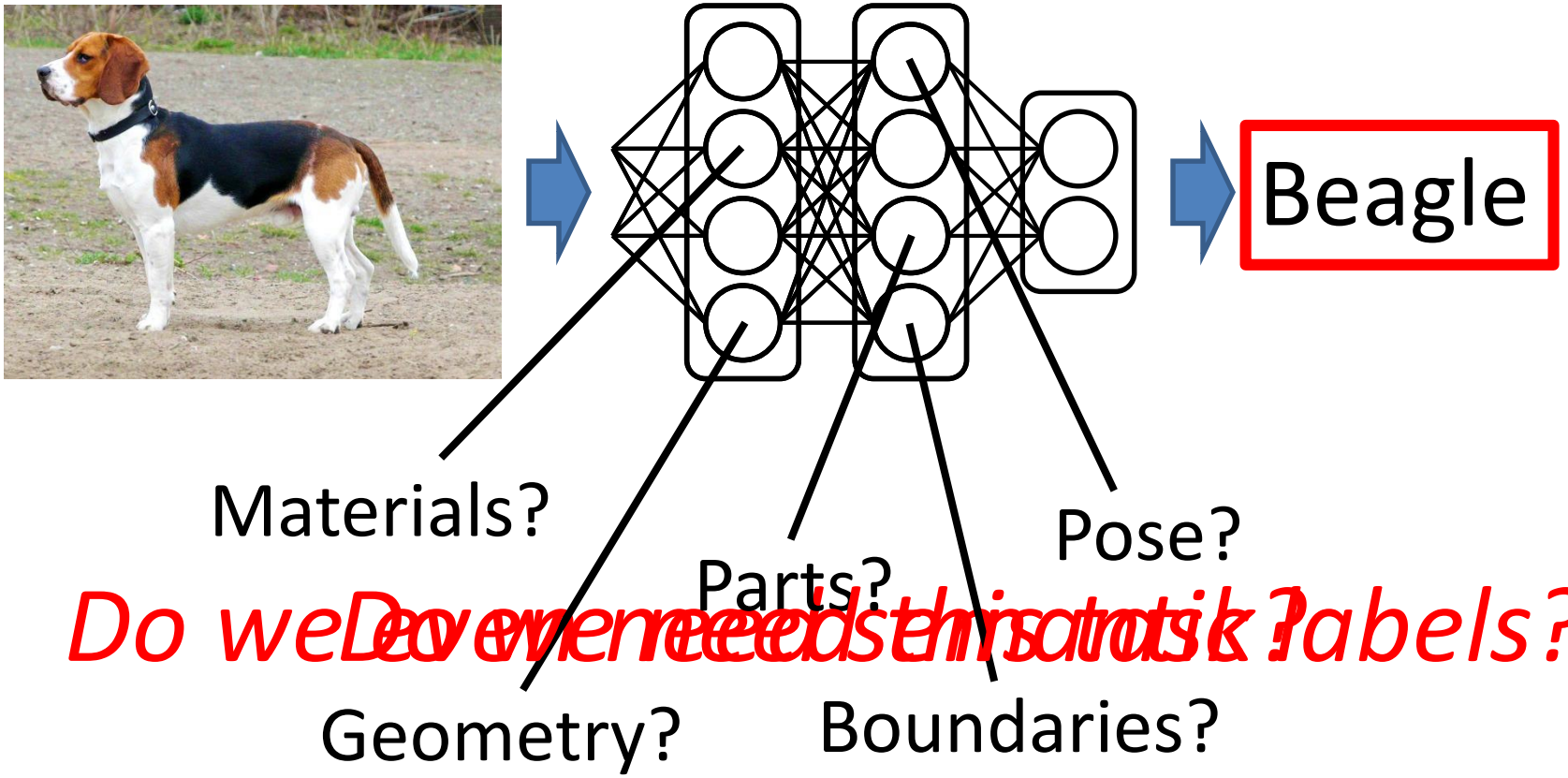


Beagle

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

Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

ImageNet + Deep Learning



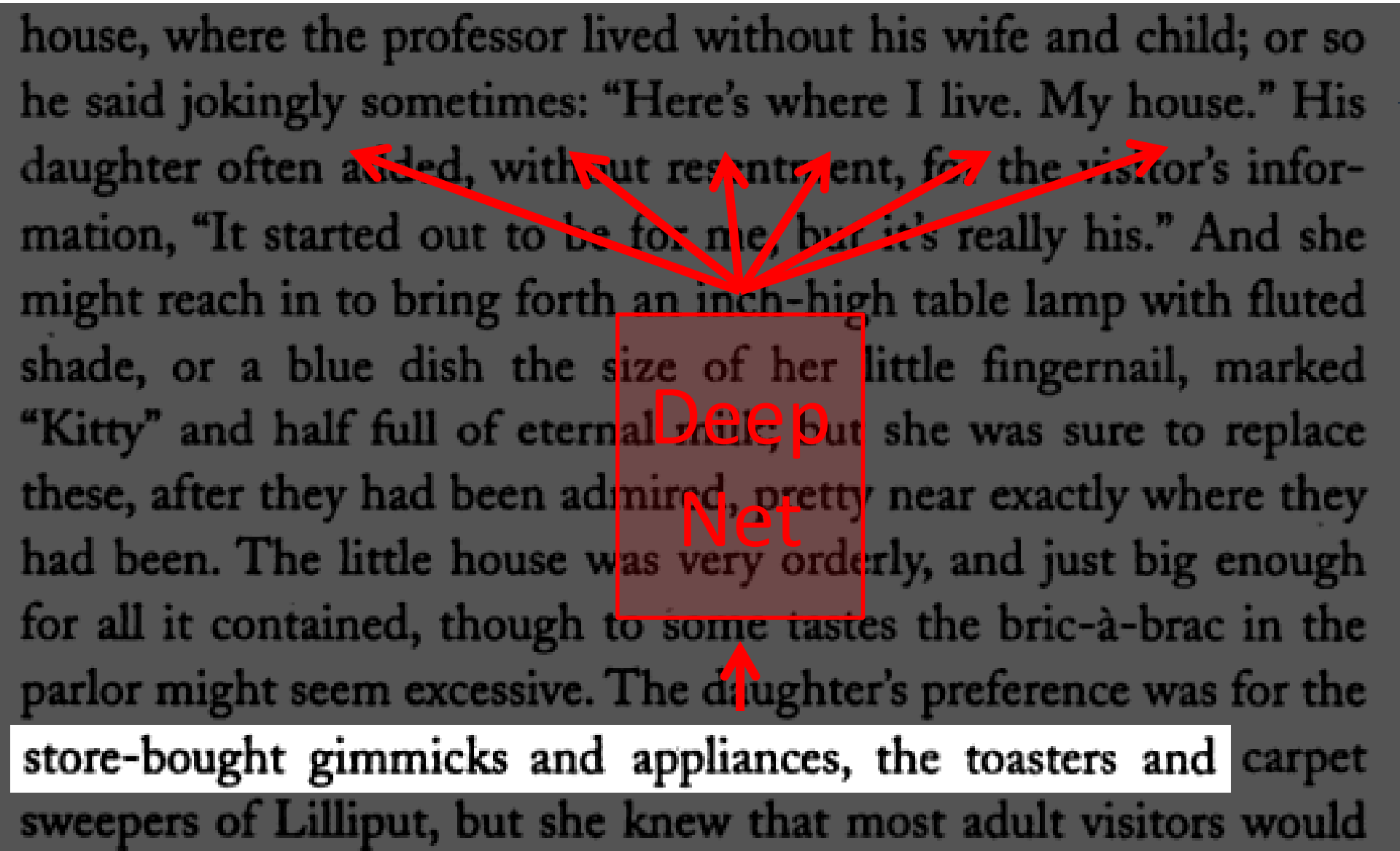
Do we need this task?

Do we need labels?

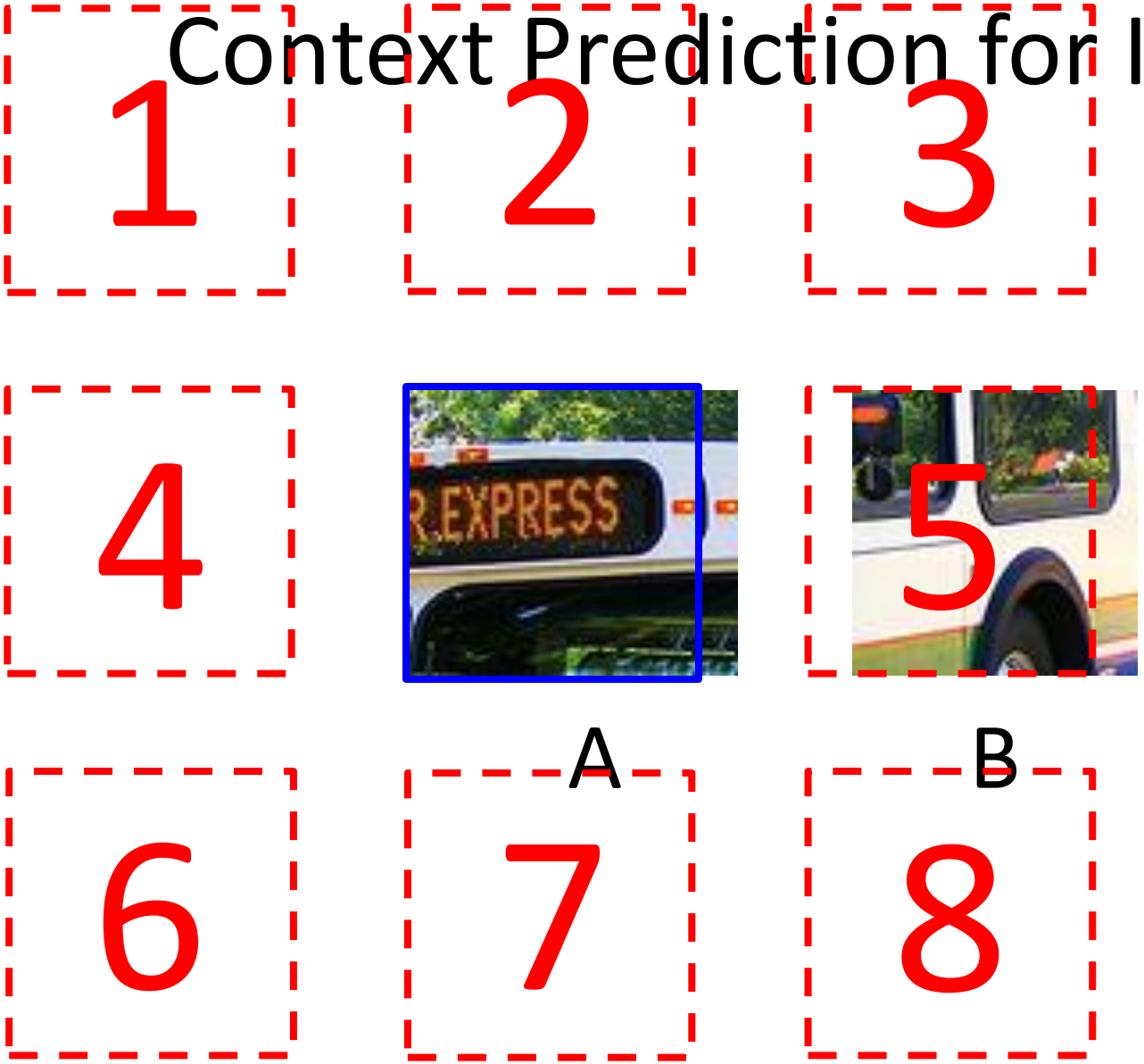
Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Context as Supervision

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

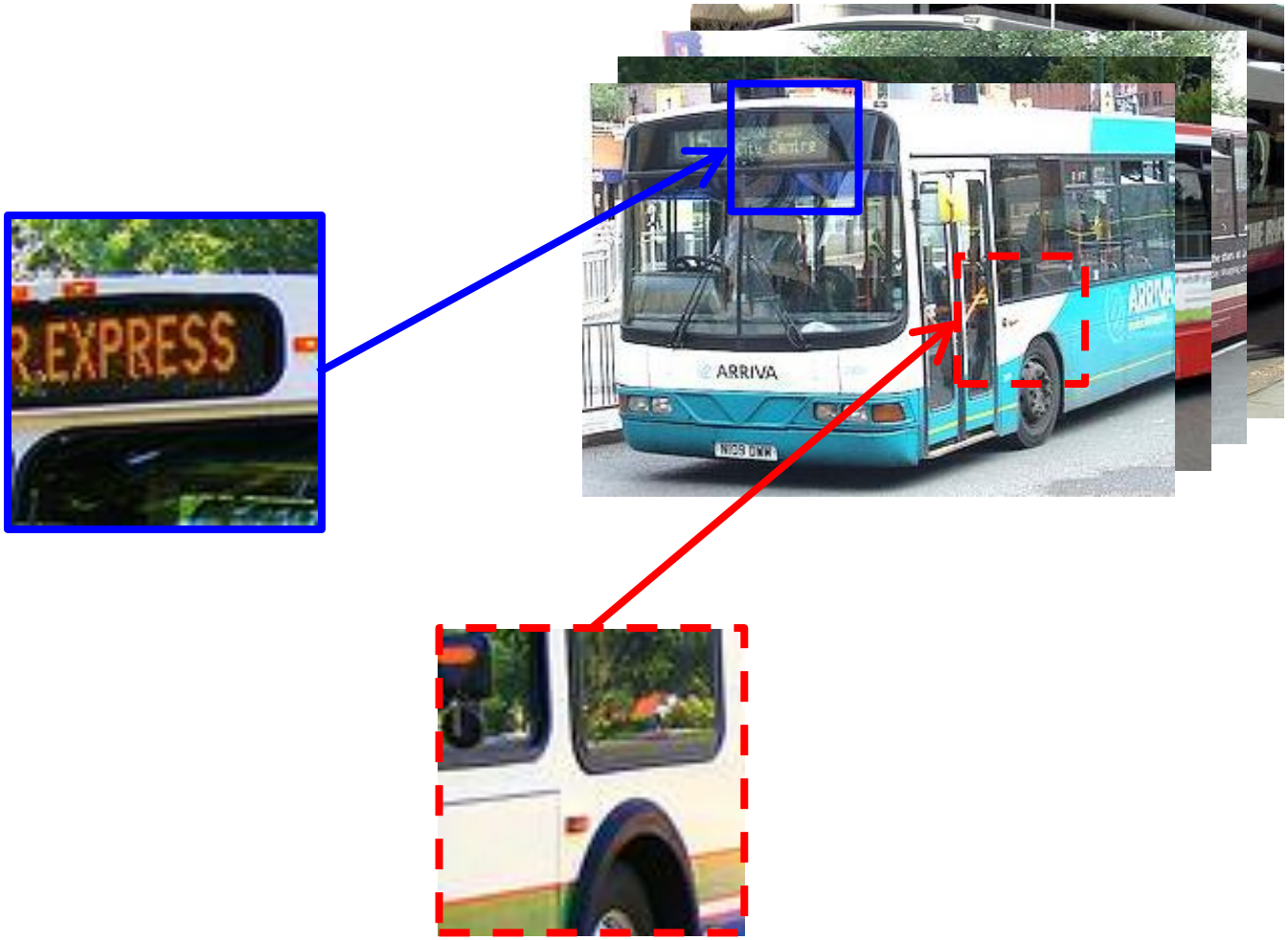


Context Prediction for Images



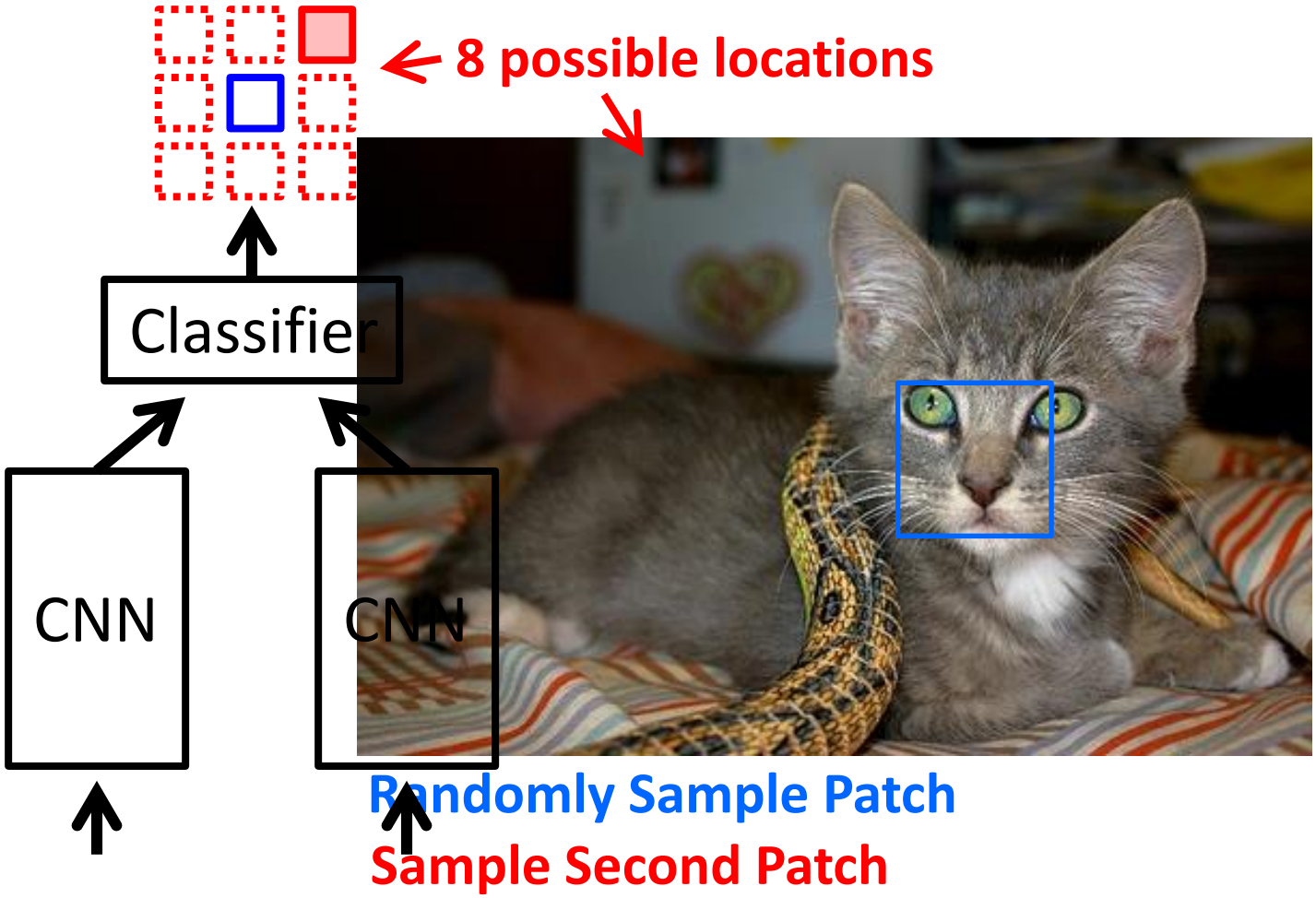
Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Semantics from a non-semantic task

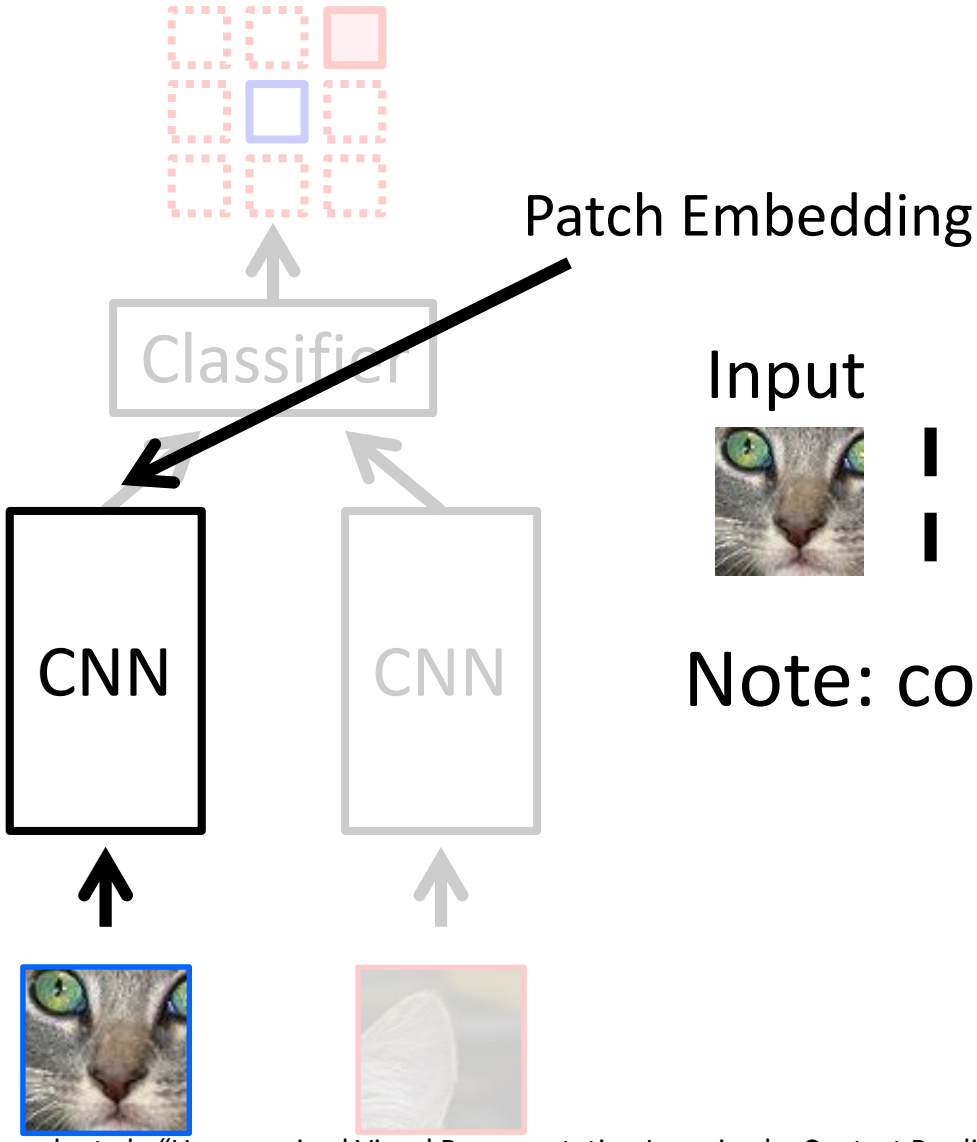


Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Relative Position Task



Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015



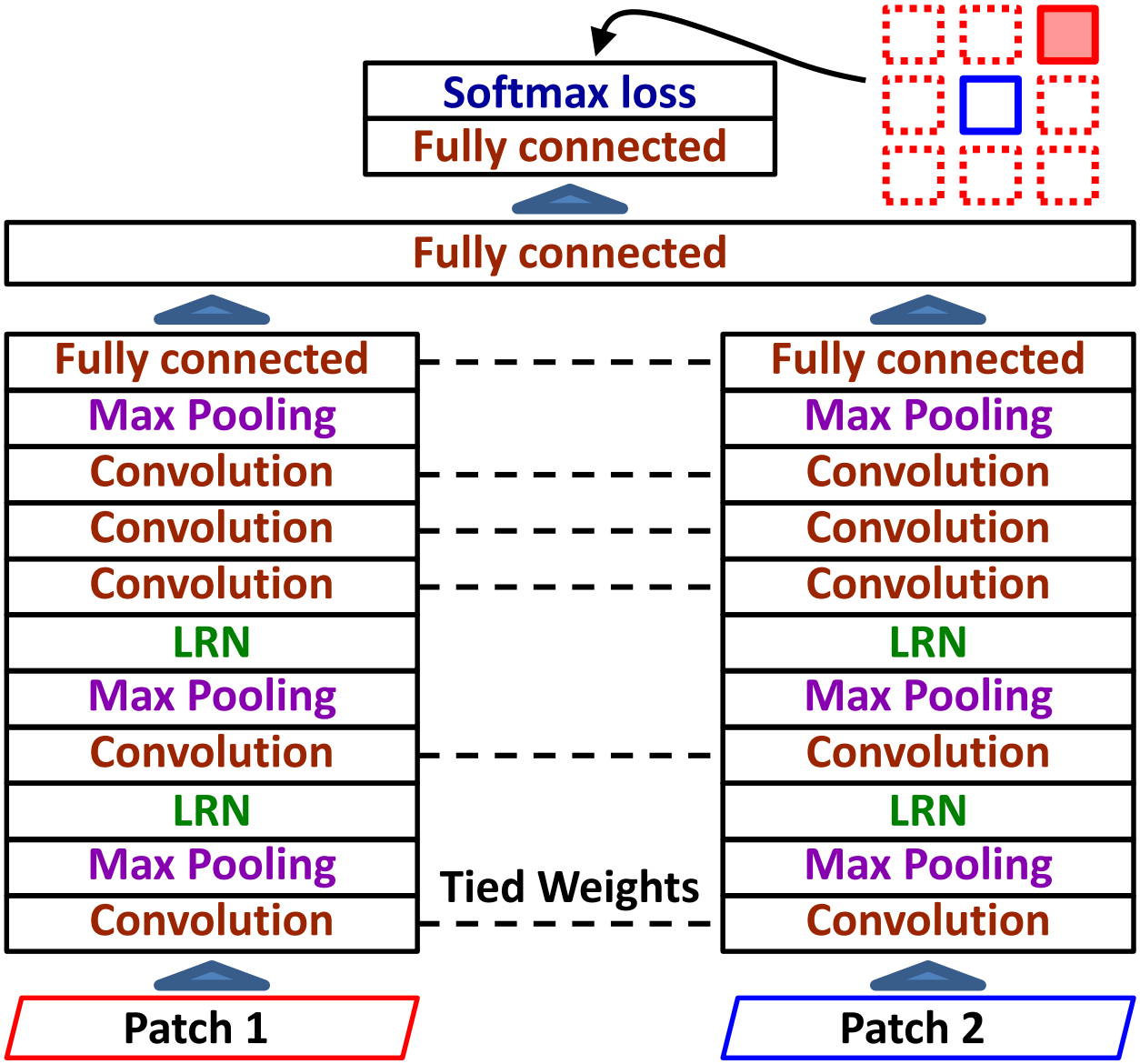
Input

Nearest Neighbors

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

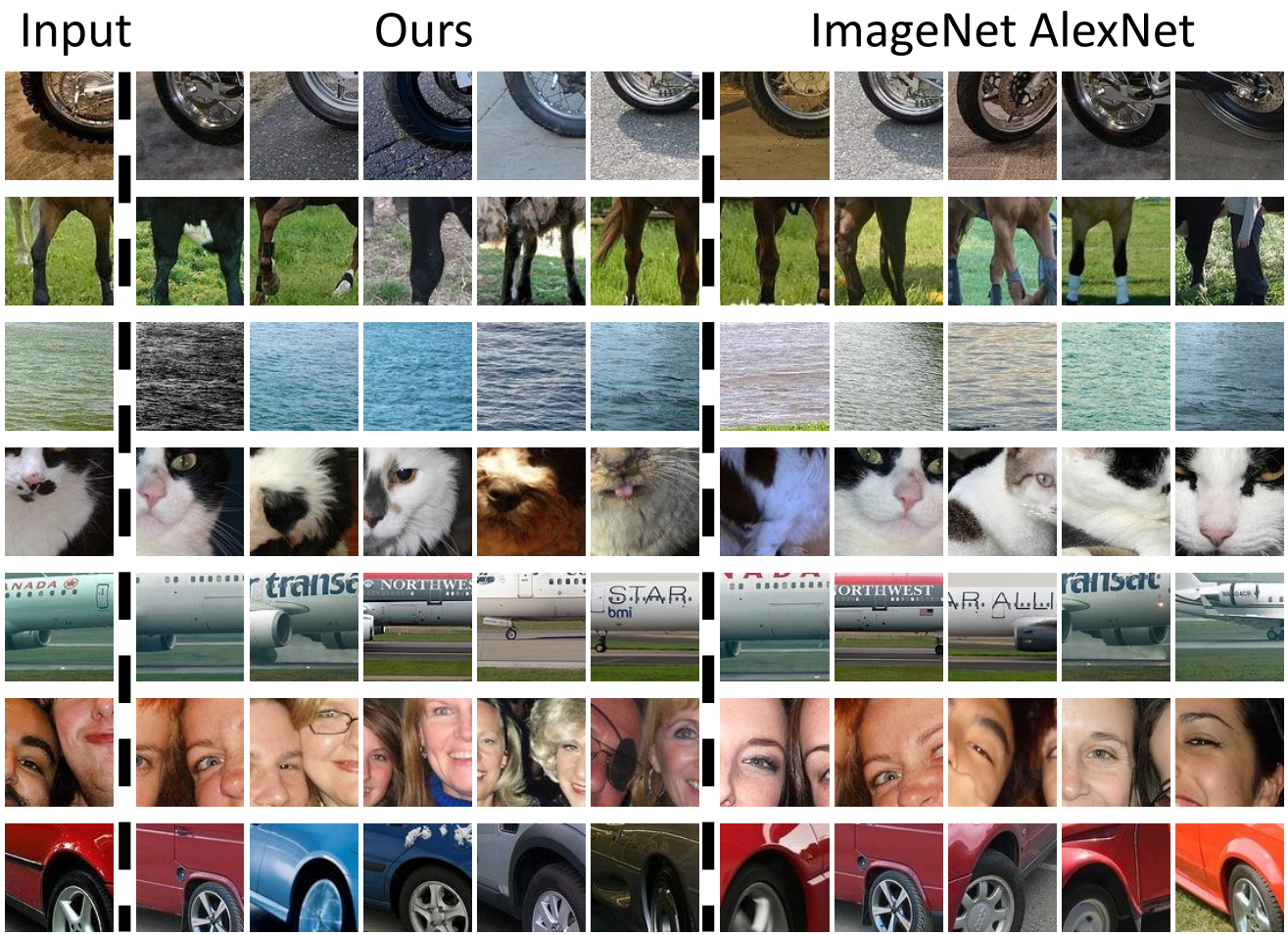
Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Architecture



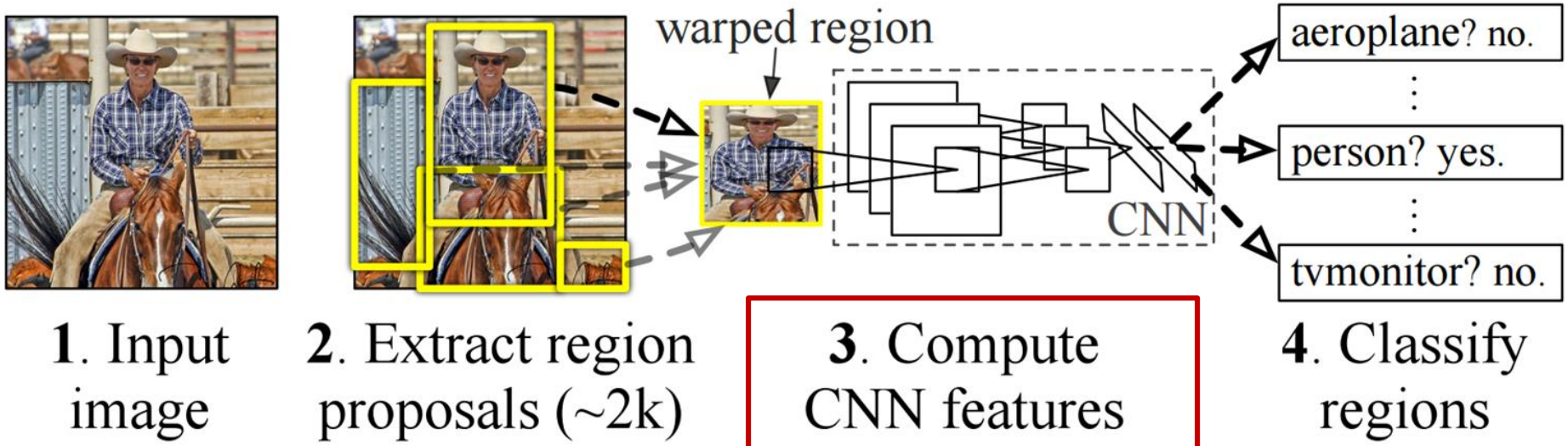
Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

What is learned?



Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Pre-Training for R-CNN



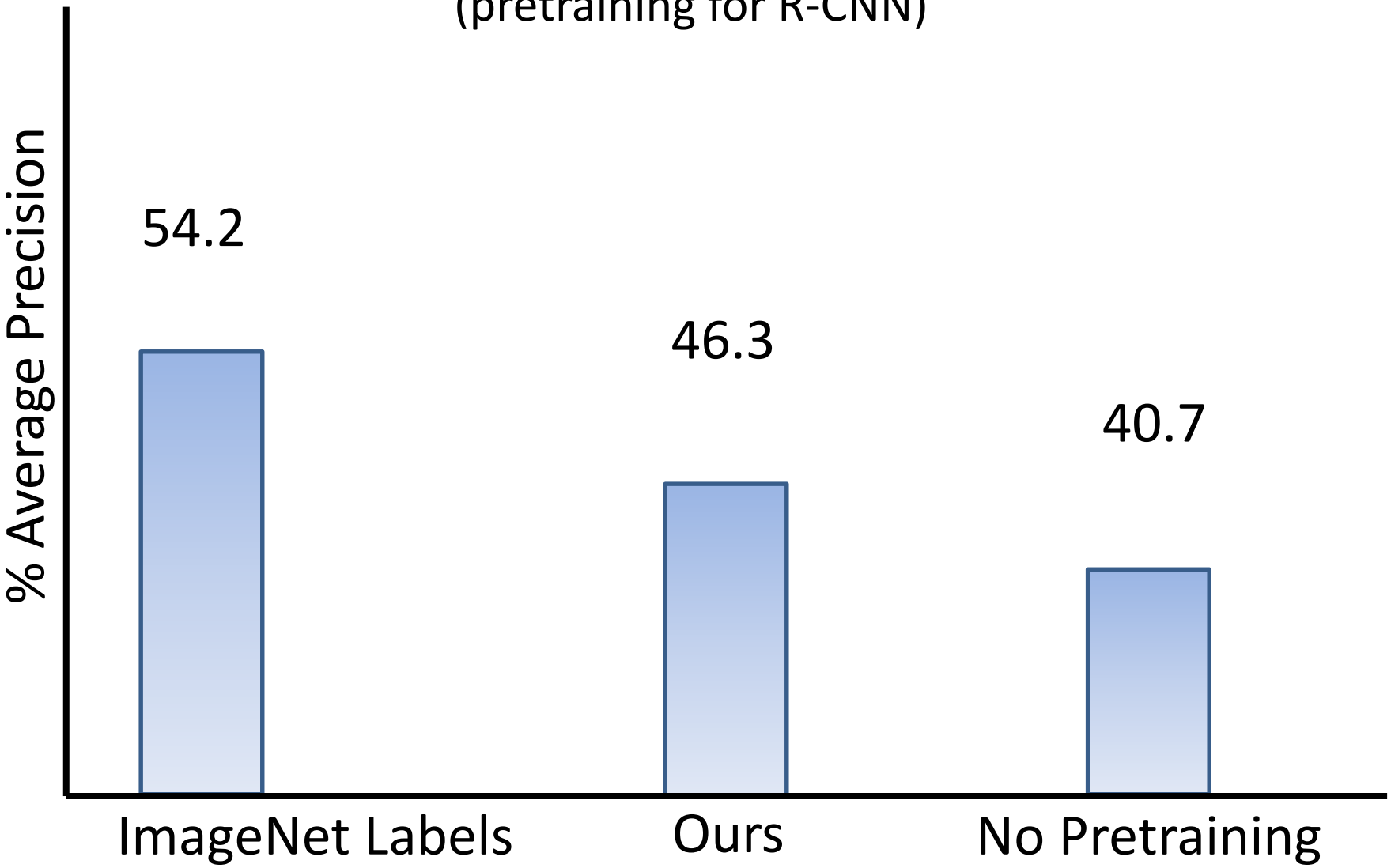
Pre-train on relative-position task, w/o labels

Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

[Girshick et al. 2014]

VOC 2007 Performance

(pretraining for R-CNN)



Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert

ECCV 2016

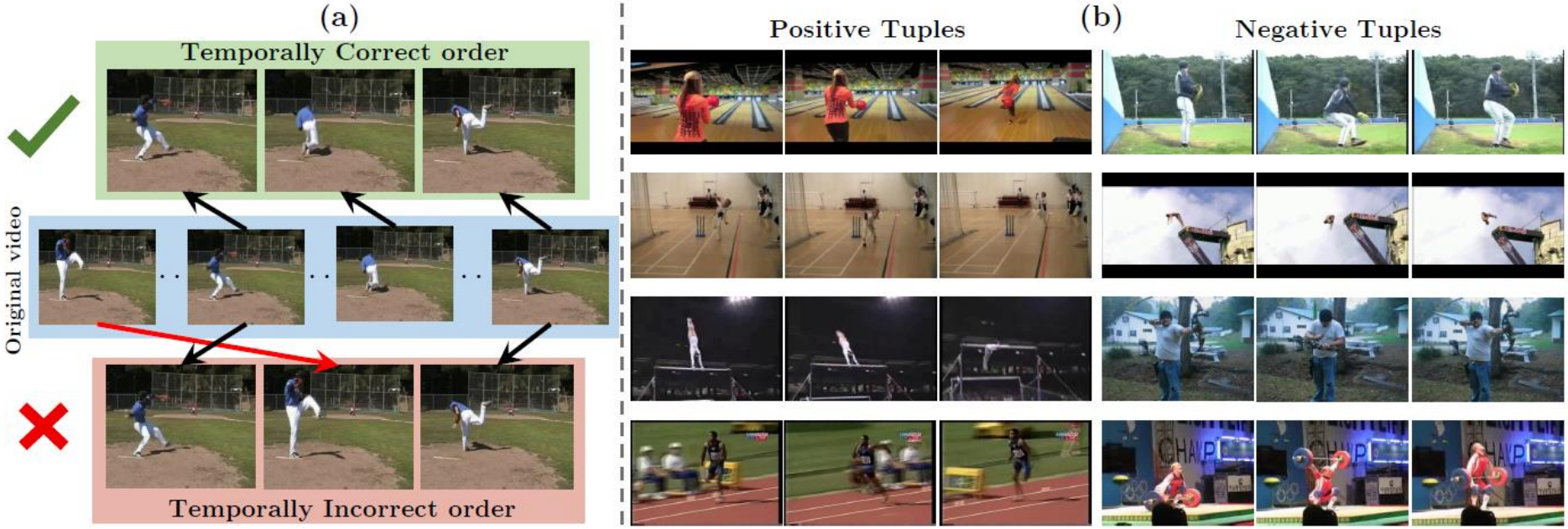


Fig. 1: **(a)** A video imposes a natural temporal structure for visual data. In many cases, one can easily verify whether frames are in the correct temporal order (shuffled or not). Such a simple sequential verification task captures important spatiotemporal signals in videos. We use this task for unsupervised pre-training of a Convolutional Neural Network (CNN). **(b)** Some examples of the automatically extracted positive and negative tuples used to formulate a classification task for a CNN.

Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", ECCV 2016

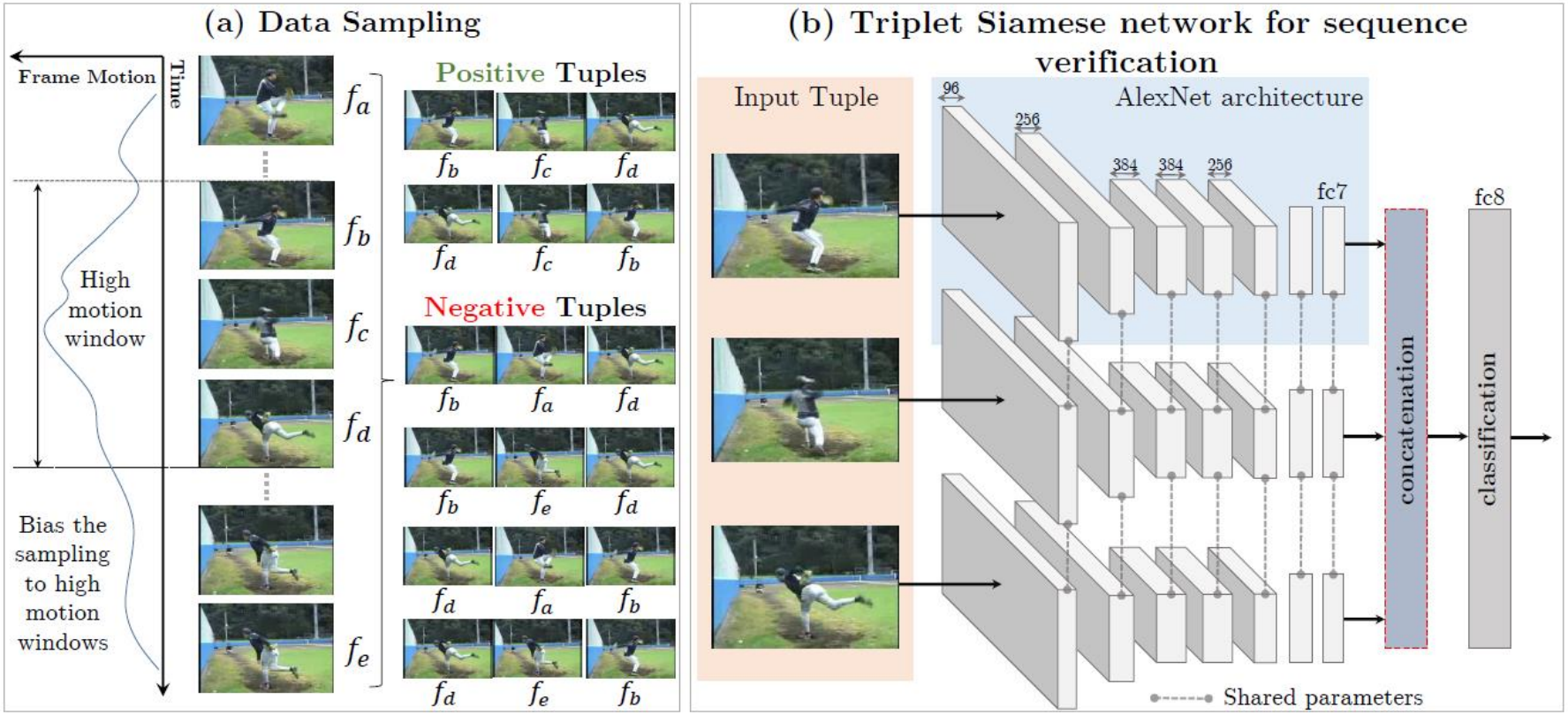


Fig. 2: **(a)** We sample tuples of frames from high motion windows in a video. We form positive and negative tuples based on whether the three input frames are in the correct temporal order. **(b)** Our triplet Siamese network architecture has three parallel network stacks with shared weights upto the `fc7` layer. Each stack takes a frame as input, and produces a representation at the `fc7` layer. The concatenated `fc7` representations are used to predict whether the input tuple is in the correct temporal order.

Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", ECCV 2016

Table 2: Mean classification accuracies over the 3 splits of UCF101 and HMDB51 datasets. We compare different initializations and finetune them for action recognition.

Dataset	Initialization	Mean Accuracy
UCF101	Random	38.6
	(Ours) Tuple verification	**50.2**
HMDB51	Random	13.3
	UCF Supervised	15.2
	(Ours) Tuple verification	**18.1**

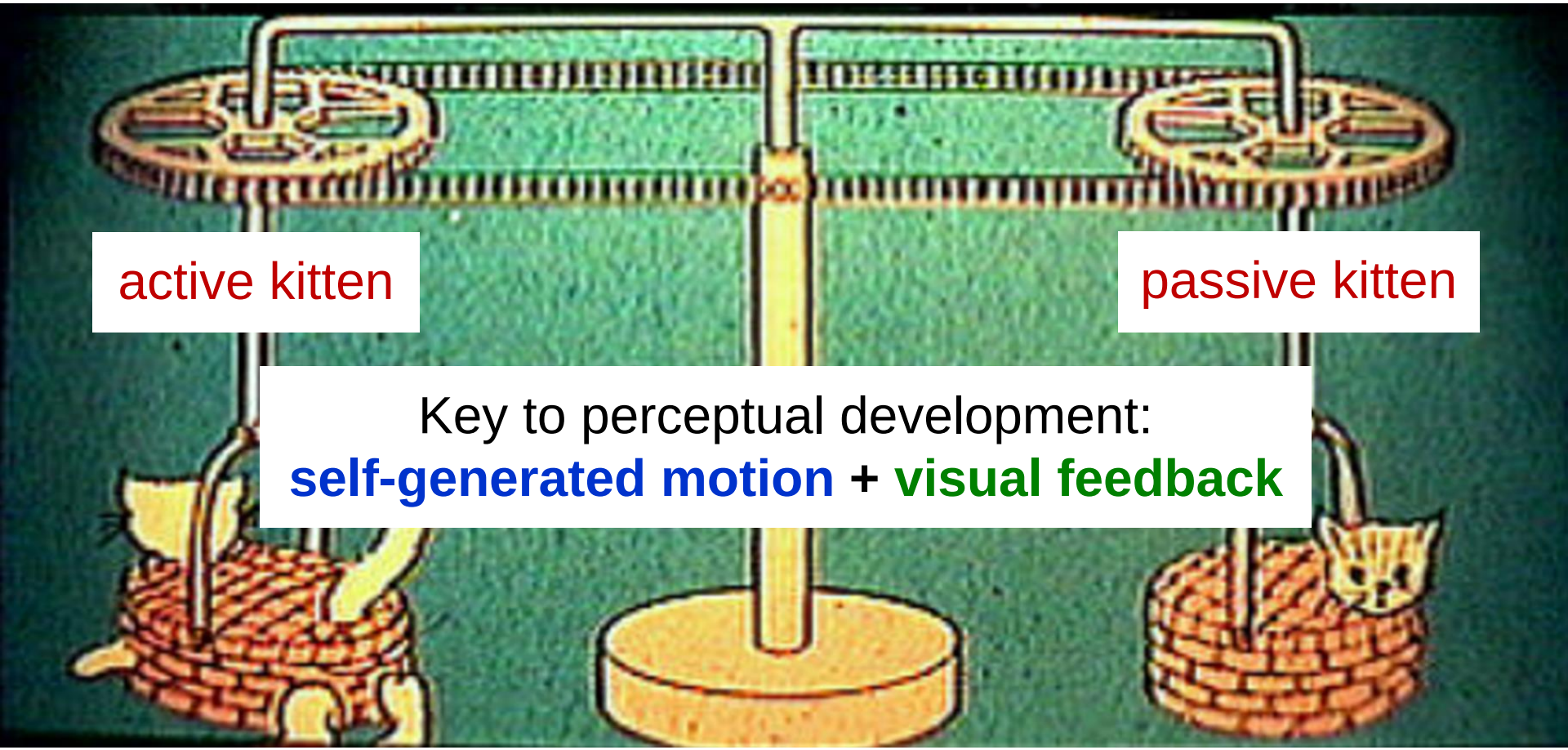
Misra et al., “Shuffle and Learn: Unsupervised Learning using Temporal Order Verification”, ECCV 2016

Learning image representations tied to ego-motion

Dinesh Jayaraman and Kristen Grauman

ICCV 2015

The kitten carousel experiment
[Held & Hein, 1963]



Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Problem with today's visual learning

Status quo: Learn from "disembodied" bag of labeled snapshots.

Our goal: Learn in the context of acting and moving in the world.

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Our idea: Ego-motion ↔ vision

Goal: Teach computer vision system the connection:
"how I move" ↔ "how my visual surroundings change"

Ego-motion motor signals

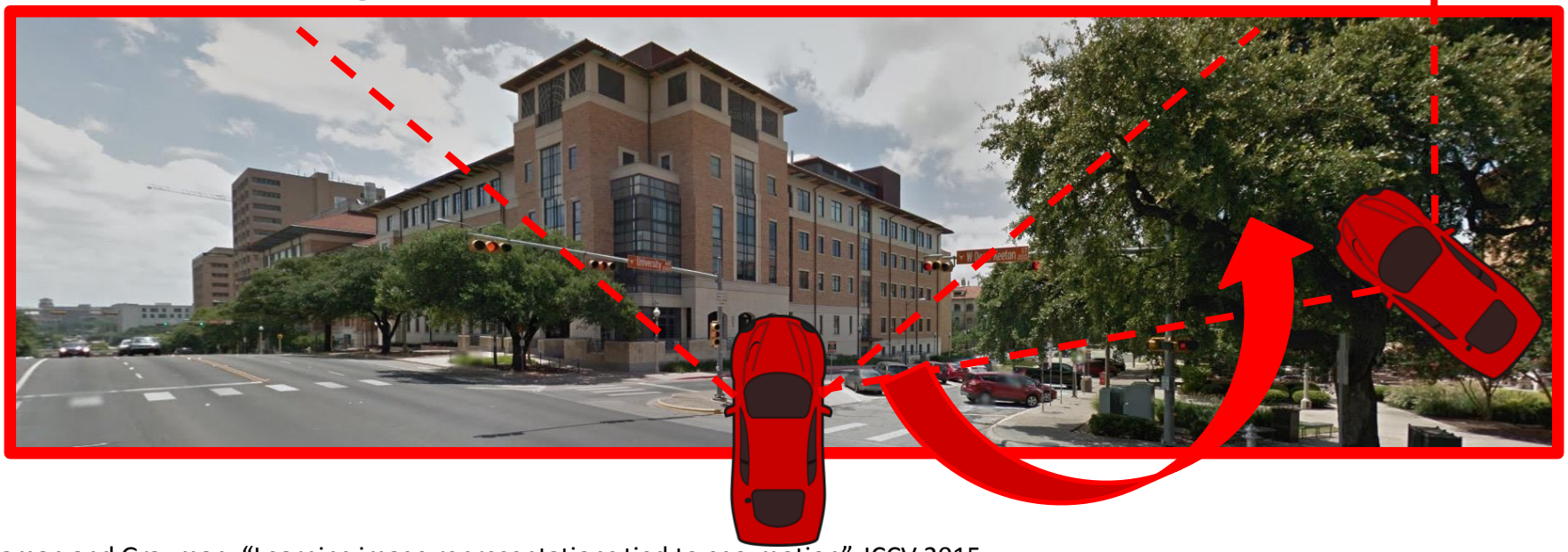
+

Unlabeled video

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Ego-motion ↔ vision: view prediction

After moving:



Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Ego-motion ↔ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

Also key to recognition!

Can be learned without manual labels!

Our approach: unsupervised feature learning using egocentric video + motor signals

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Approach idea: Ego-motion equivariance

Invariant features: unresponsive to some classes of transformations

$$\mathbf{z}(g\mathbf{x}) \approx \mathbf{z}(\mathbf{x})$$

Equivariant features : *predictably* responsive to some classes of transformations, through simple mappings (e.g., linear)

"equivariance map"

$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

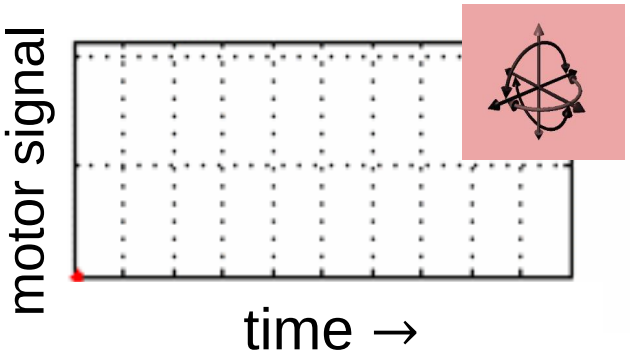
Invariance <u>*discards*</u> information;
equivariance <u>*organizes*</u> it.

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Approach idea: Ego-motion equivariance

Training data

Unlabeled video + motor signals



Learn

Equivariant embedding

organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Approach overview

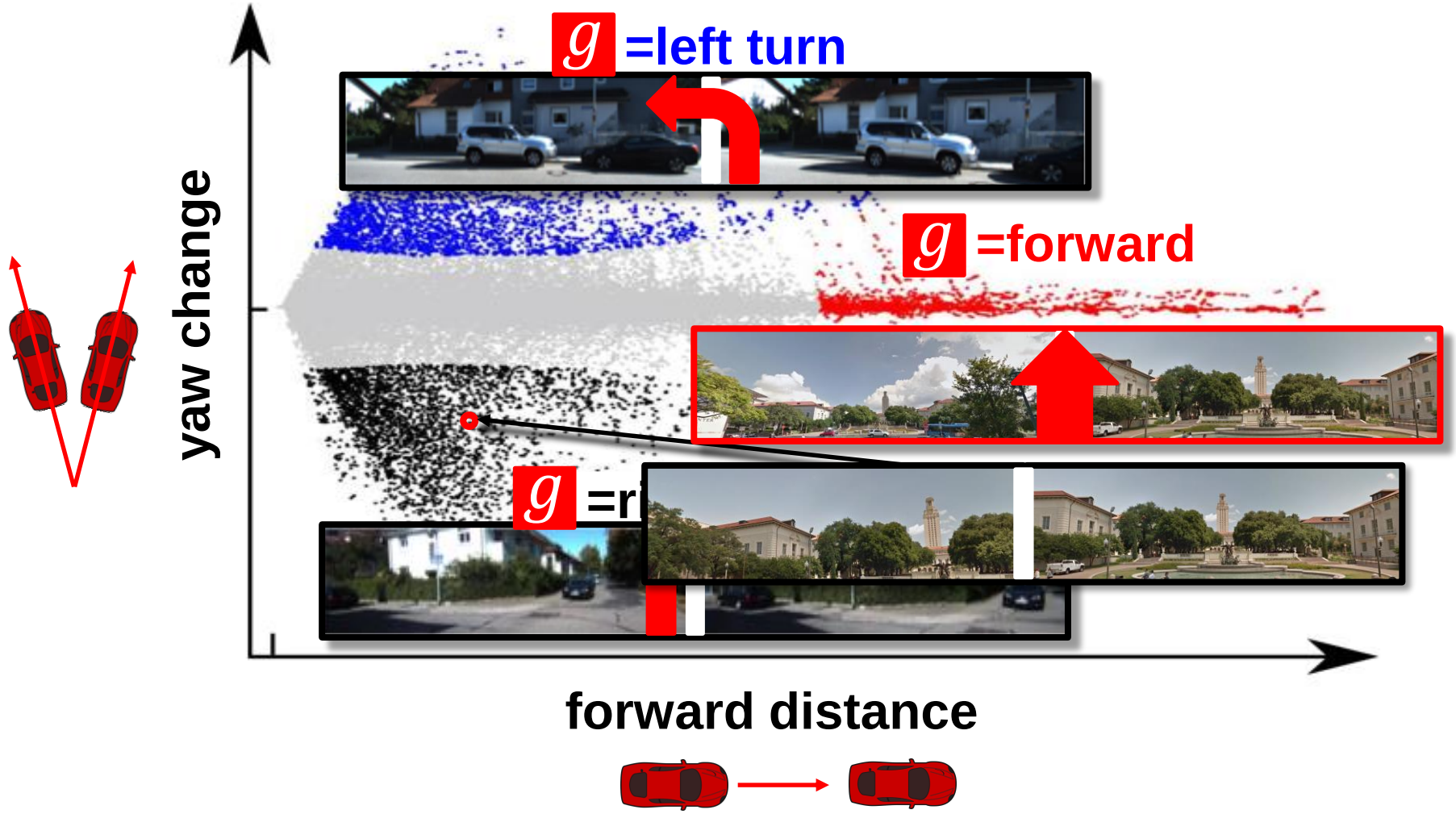
Our approach: unsupervised feature learning using egocentric video + motor signals

1. Extract training frame pairs from video
2. Learn ego-motion-equivariant image features
3. Train on target recognition task in parallel

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Training frame pair mining

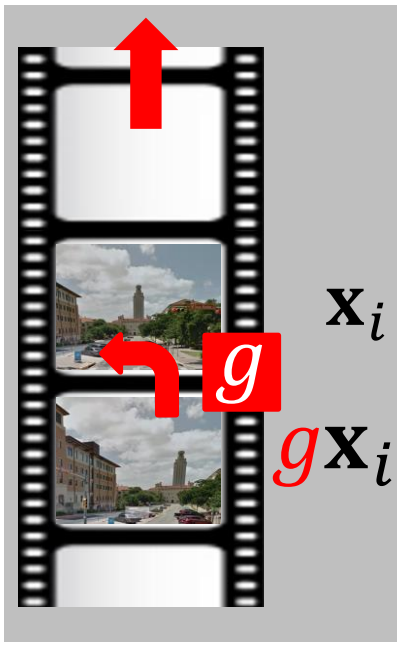
Discovery of ego-motion clusters



Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Ego-motion equivariant feature learning

Given:

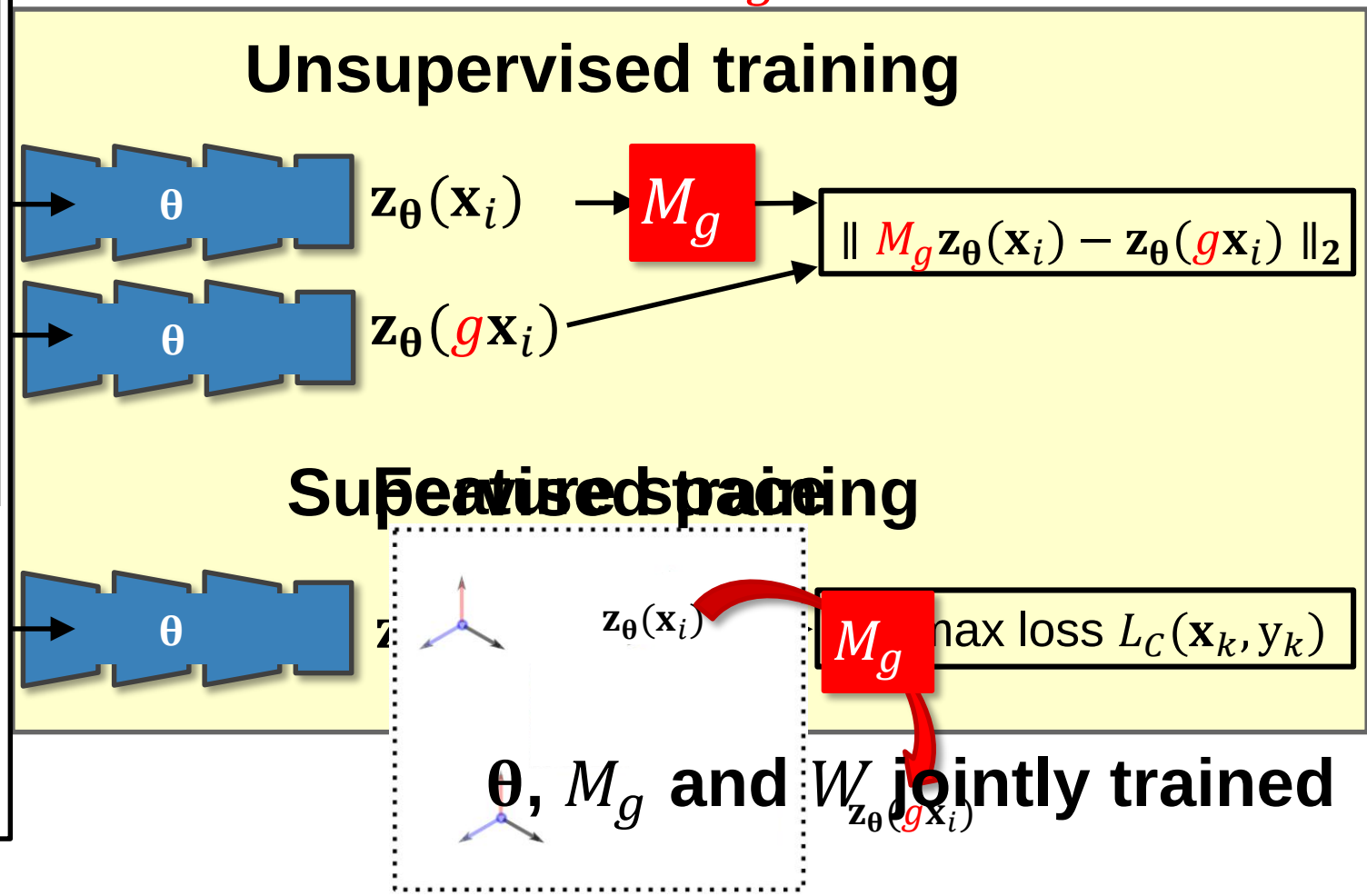


$\mathbf{x}_k$

class y_k

Desired: for all motions g and all images $\mathbf{x}$,

$$\mathbf{z}_{\boldsymbol{\theta}}(g\mathbf{x}) \approx M_g \mathbf{z}_{\boldsymbol{\theta}}(\mathbf{x})$$

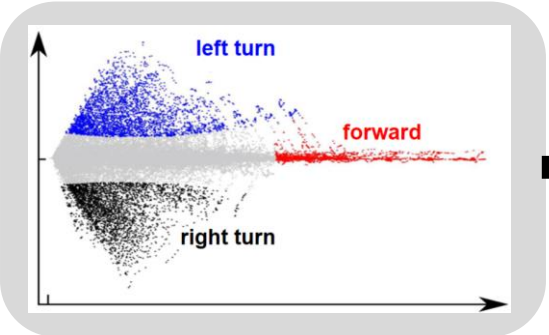


Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

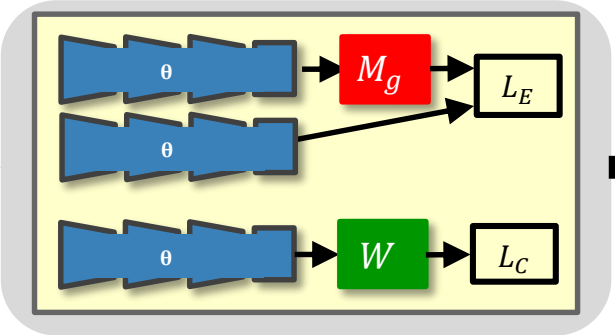
Summary

APPROACH

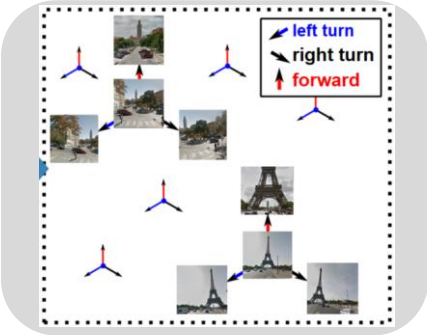
Ego-motion training pairs



Neural network training



Equivariant embedding



RESULTS

Scene and object recognition

Football field?
Pagoda?
Airport?
Cathedral?
Army base?

Next-best view selection

cup

frying pan

Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Results: Recognition

Learn from **unlabeled car video** (KITTI)

Geiger et al, IJRR ’13

Exploit features for **static scene classification** (SUN, 397 classes)

Apse | Window seat | Art school | Library | Auditorium | Bus interior | Cathedral | Freeway | Guardhouse

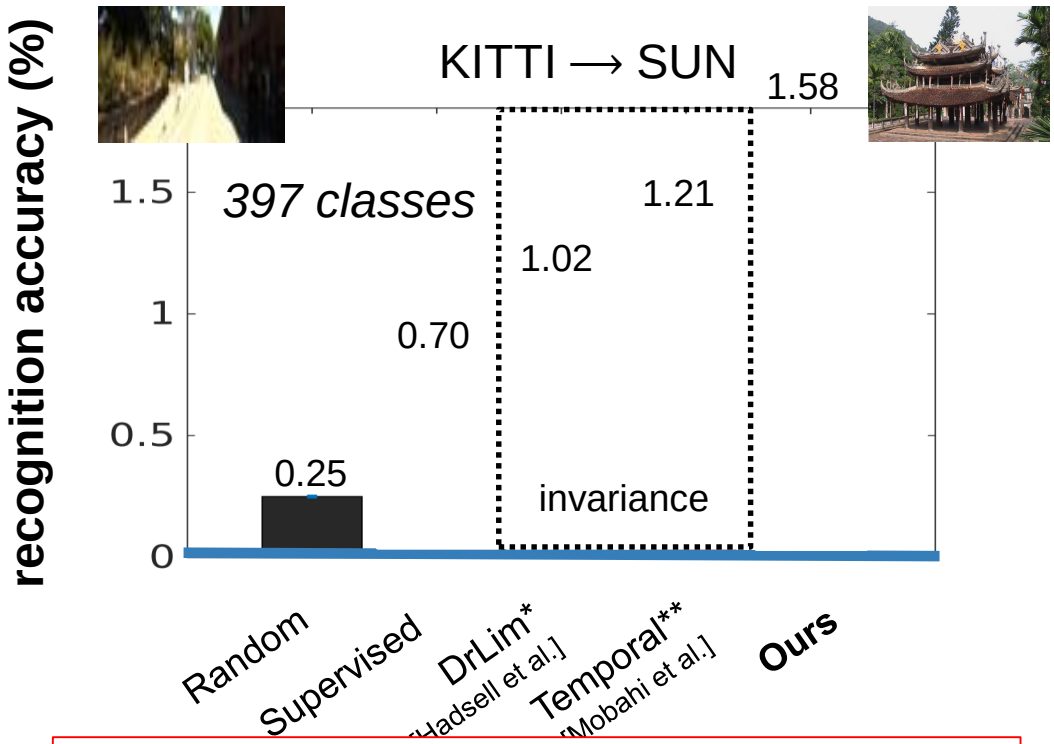
Xiao et al, CVPR ’10

Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Results: Recognition

Do ego-motion equivariant features improve recognition?

6 labeled training examples per class



Up to 30% accuracy increase over state of the art!

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Discussion

- Many types of supervision cues have been tried
- What other types of supervision “for free” can we use?
- How do we know if a certain supervision type would work?
- Can we make this type of learning perform on par with supervised learning?

Plan for next two lectures

- Self-supervised learning
  - From surrounding regions (ICCV 2015)
  - From temporal order (ECCV 2016)
  - From ego-motion (ICCV 2015)
- Visual discovery
  - From surrounding objects (CVPR 2010)
  - Elements of architectural styles (SIGGRAPH 2012)
  - Gradual style changes in time/space (ICCV 2013)

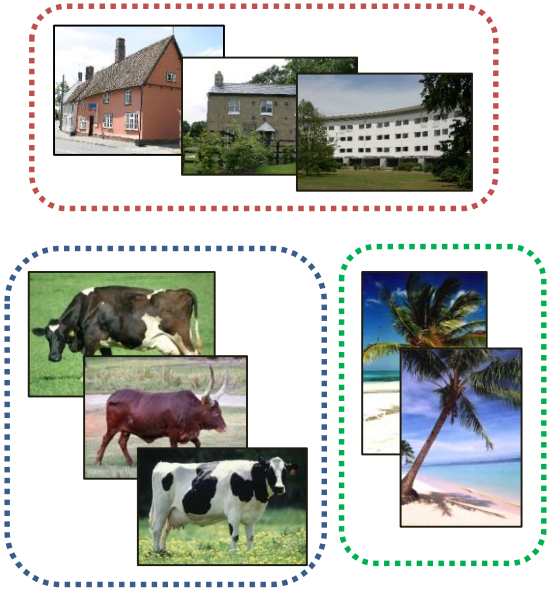
Object-Graphs for Context-Aware Category Discovery

Yong Jae Lee and Kristen Grauman

CVPR 2010

Goal

Unlabeled Image Data



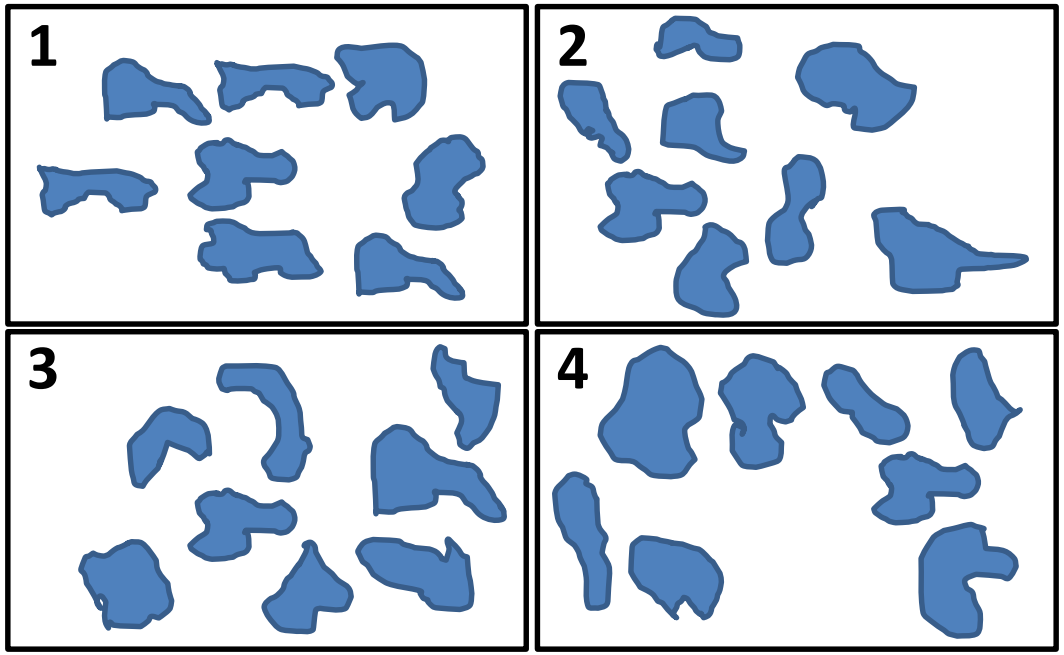
Discovered categories

- Discover *new* object categories, based on their relation to categories for which we have trained models

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010

Existing approaches

Previous work treats unsupervised visual discovery as an appearance-grouping problem.

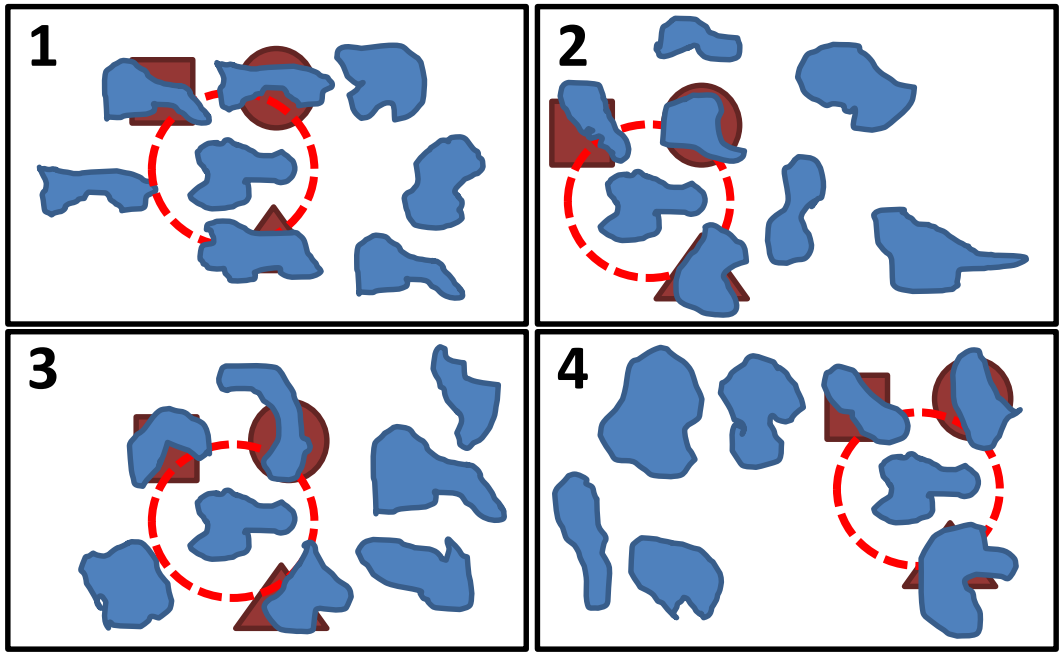


Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010

Our idea

How can seeing previously learned objects in novel images help to discover *new* categories?

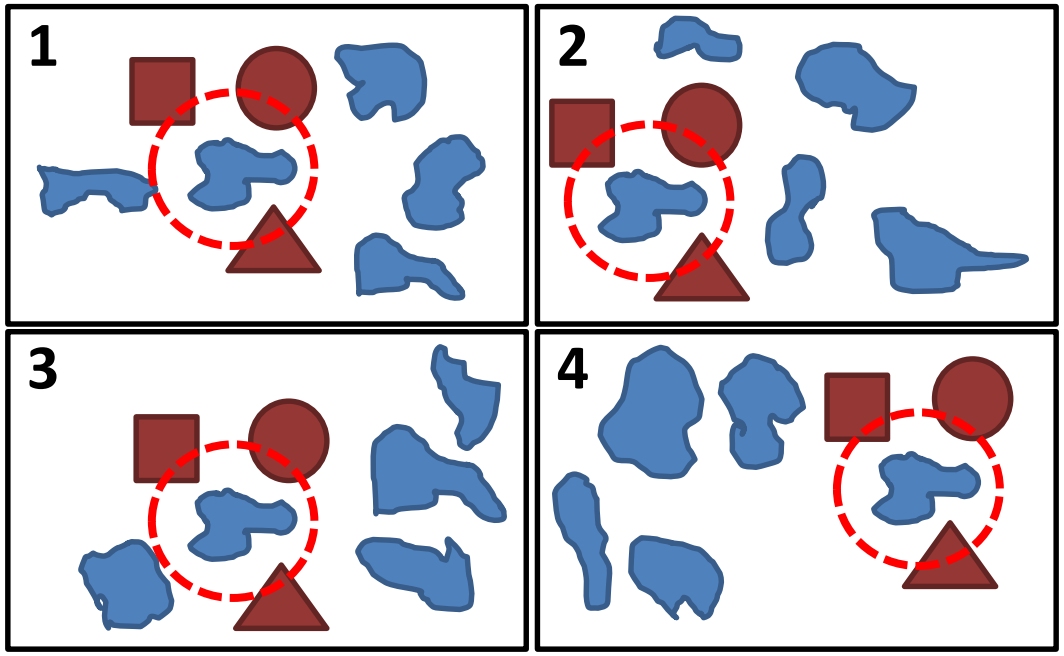


Can you identify the recurring pattern?

Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Our idea

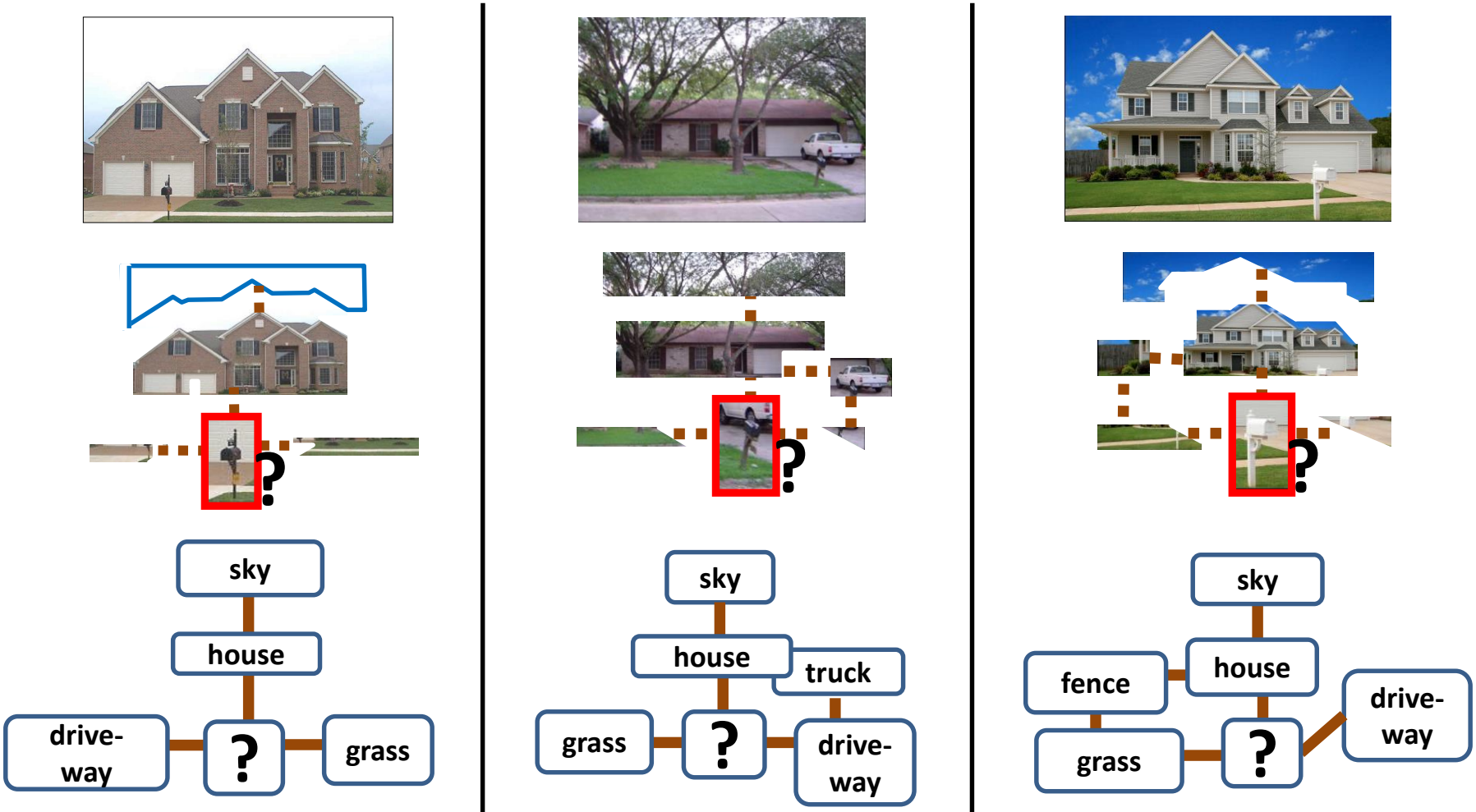
Discover visual categories within unlabeled images by modeling interactions between the unfamiliar regions and familiar objects.



Can you identify the recurring pattern?

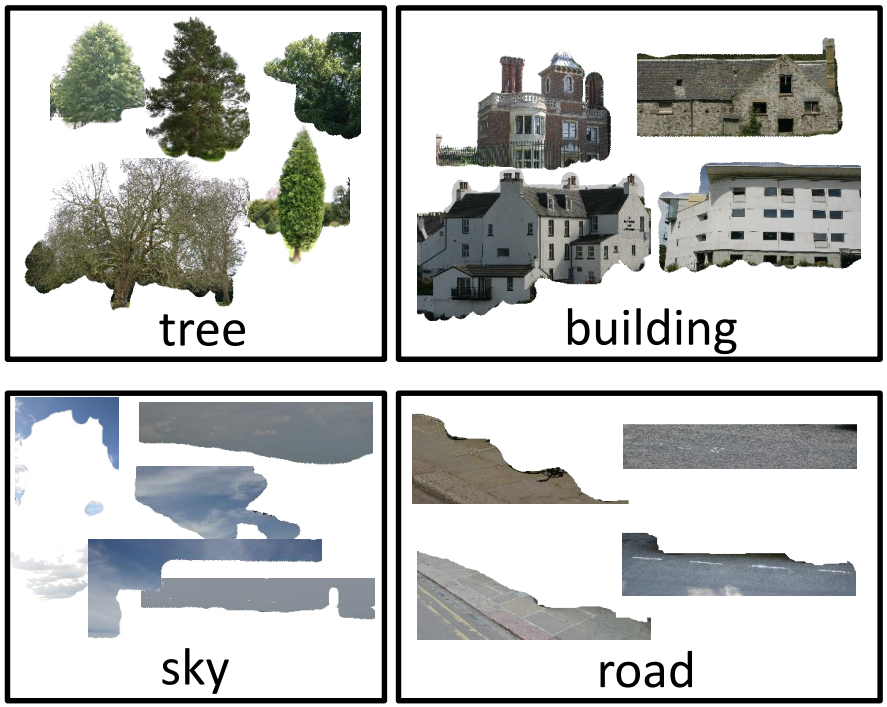
Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Context-aware visual discovery



Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010

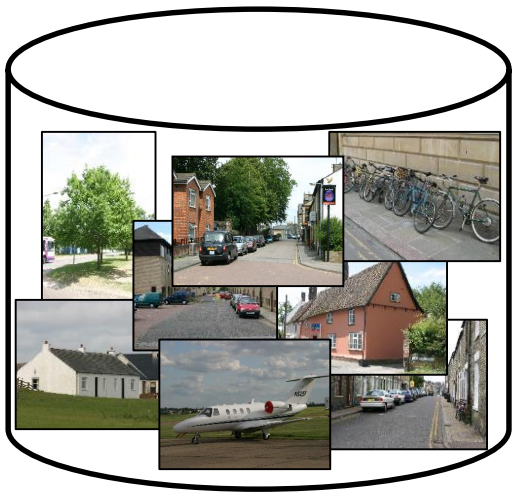
Learn "Known" Categories



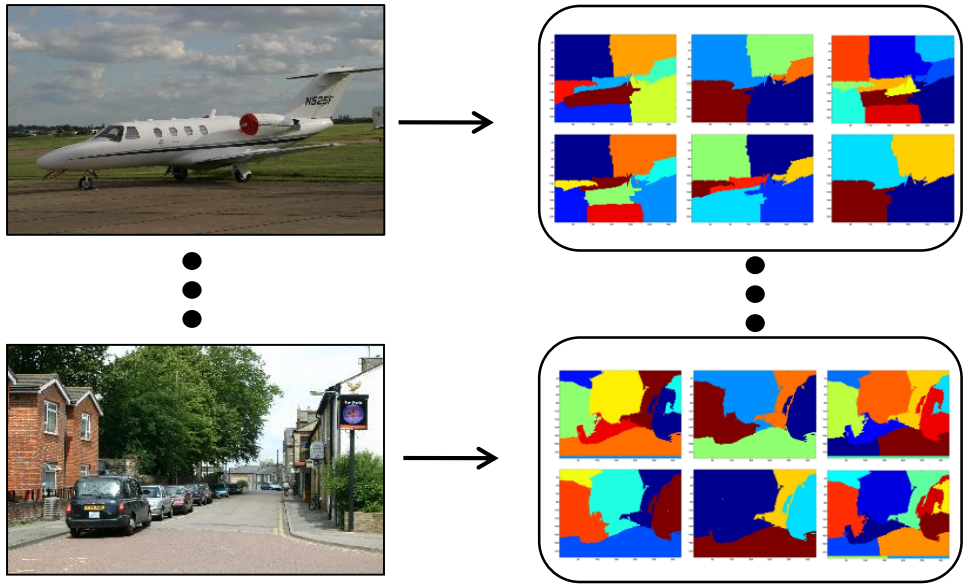
- Offline: Train region-based classifiers for *N* "known" categories using labeled training data.

Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Identifying Unknown Objects



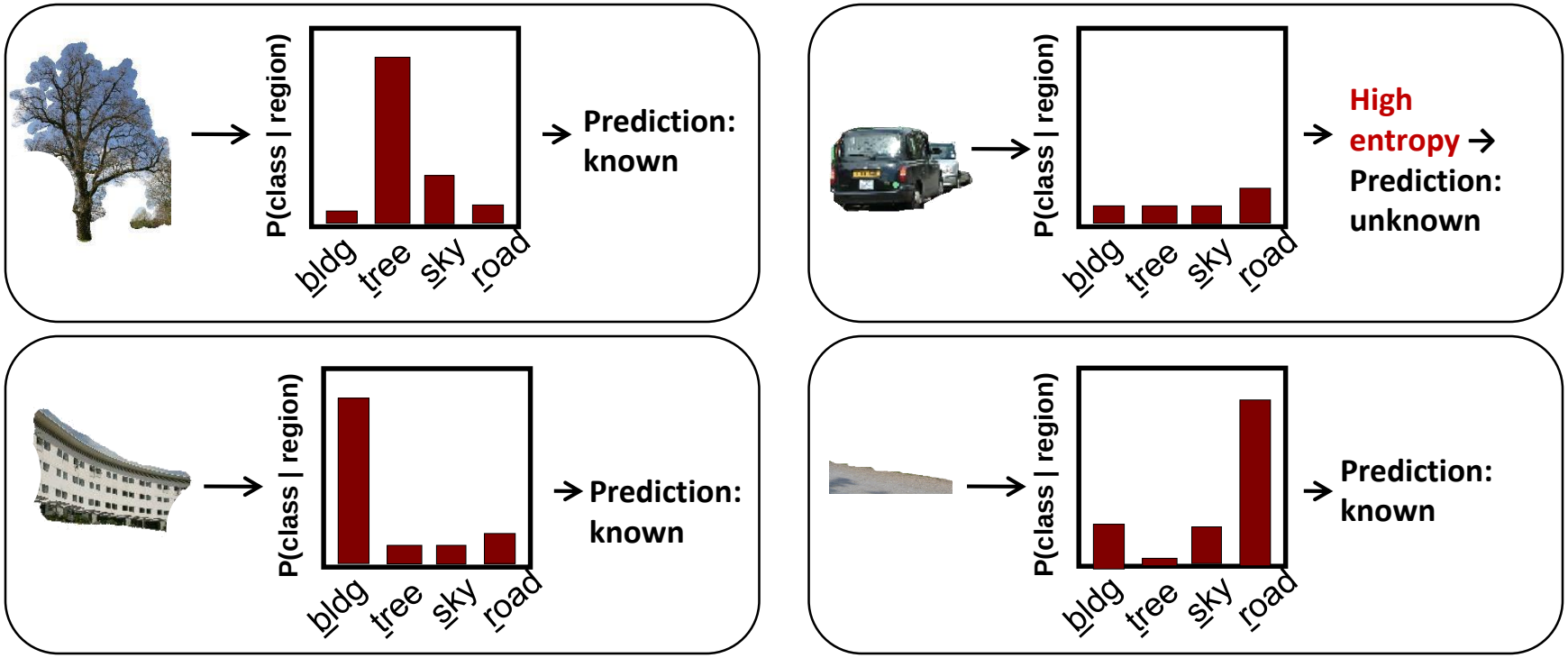
Input: unlabeled pool of novel images



Compute multiple-segmentations for each unlabeled image

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010

Identifying Unknown Objects

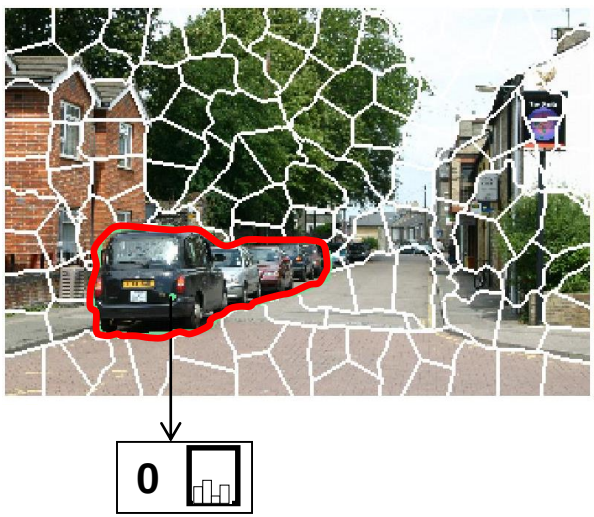


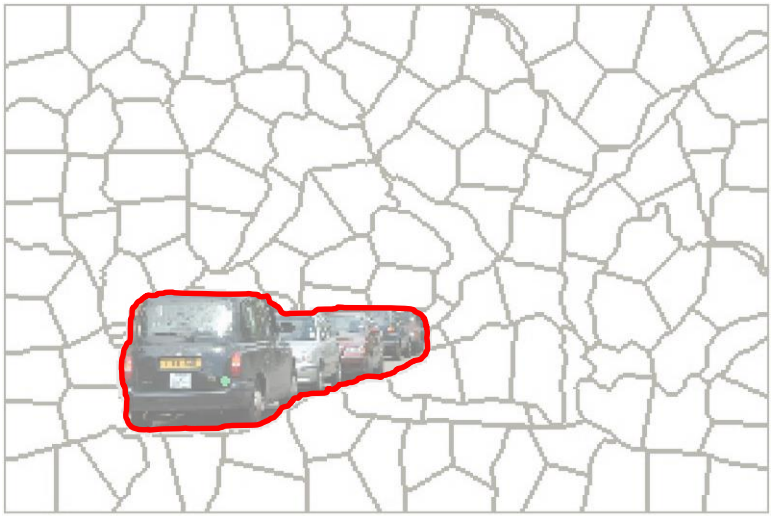
- For all segments, use classifiers to compute posteriors for the *N* "known" categories.
- Deem each segment as "known" or "unknown" based on resulting entropy.

Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Object-Graphs

An unknown region within an image



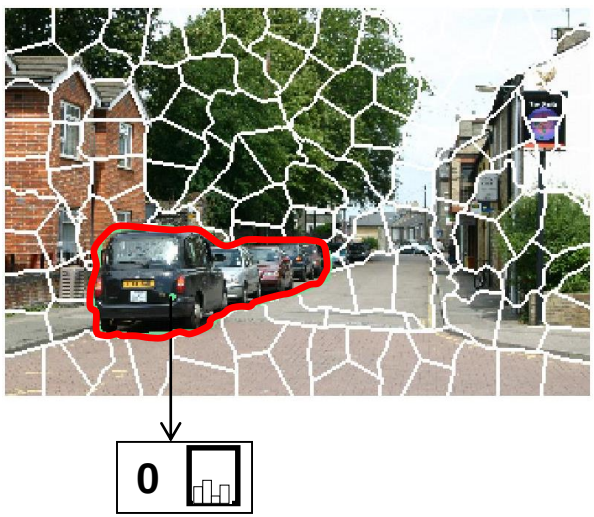


- Model the topology of category predictions relative to the unknown (unfamiliar) region.
- Incorporate uncertainty from classifiers.

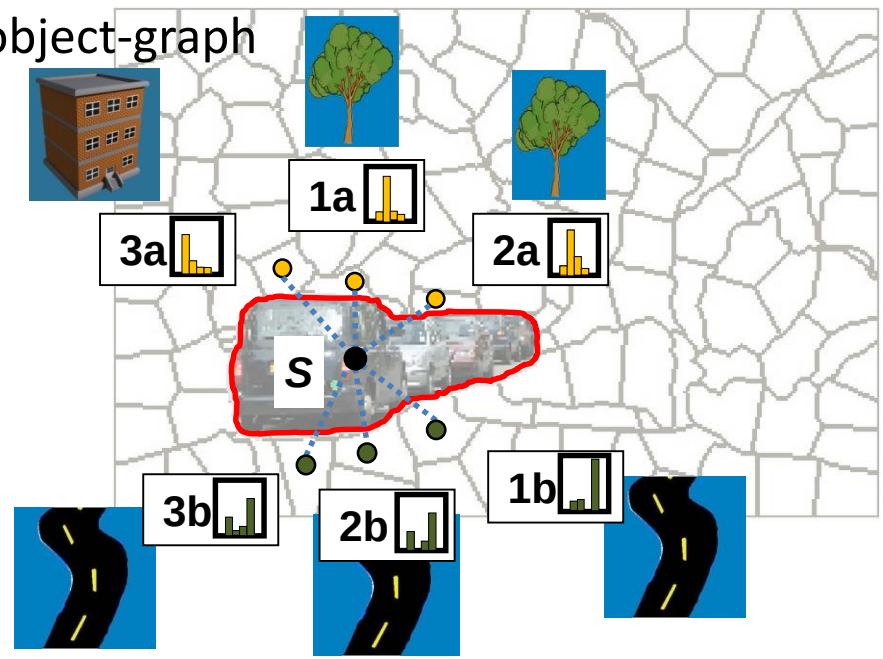
Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Object-Graphs

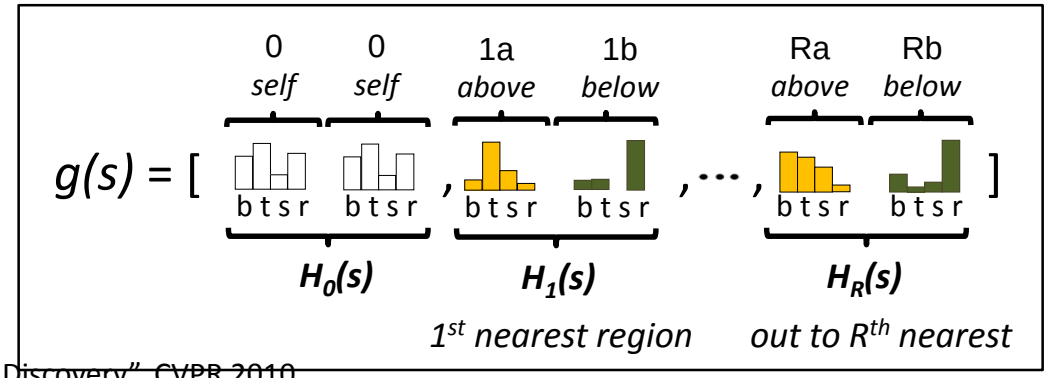
An unknown region within an image



Closest nodes in its object-graph



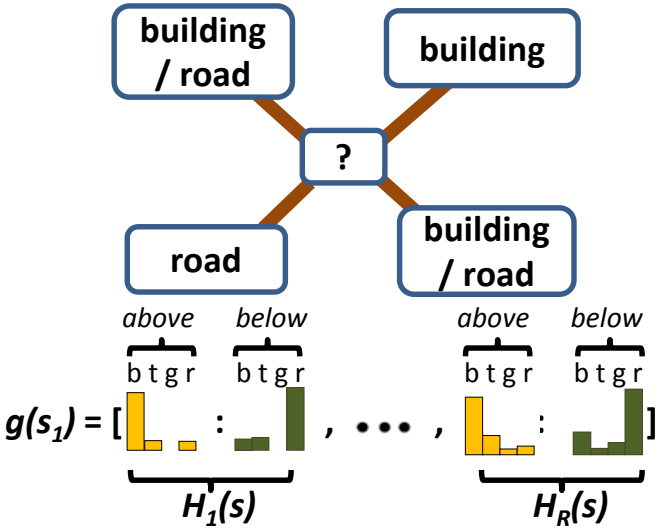
- Consider spatially near regions *above* and *below*, record distributions for each known class.

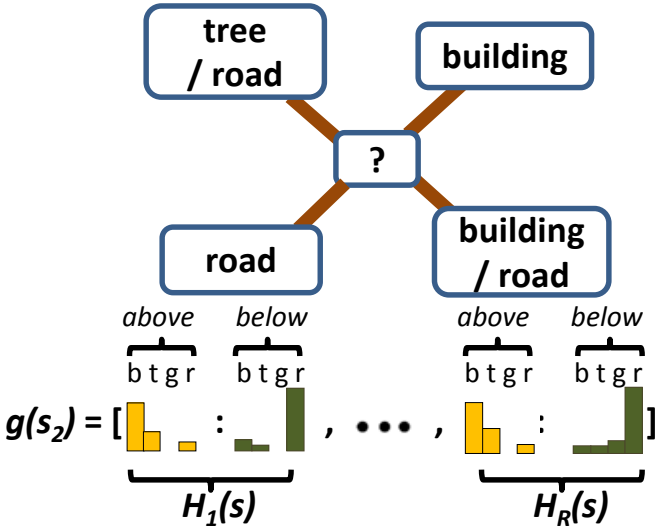


Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Object-Graph matching

Known classes
b: building
t: tree
g: grass
r: road

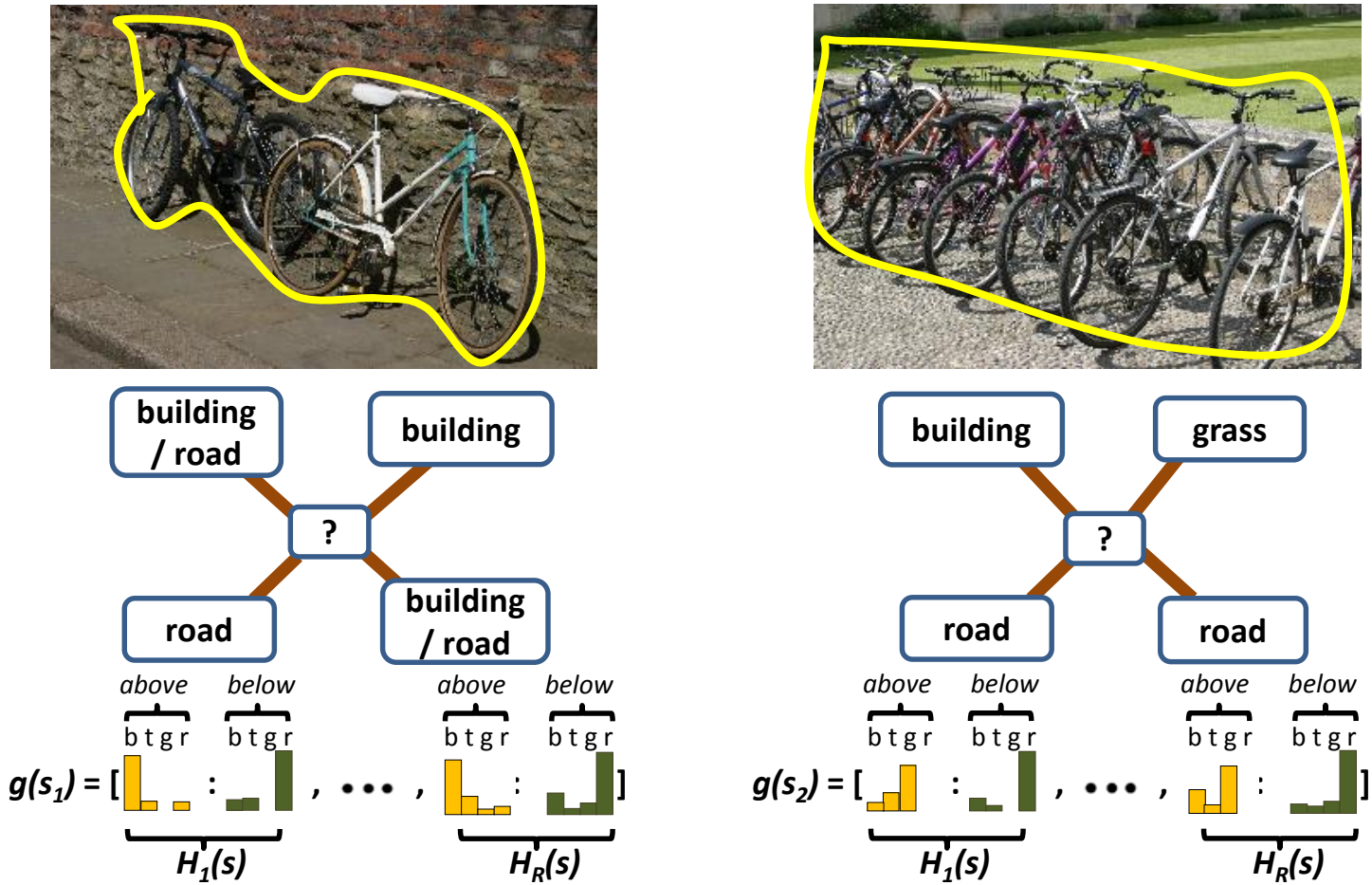




- Object-graphs are very similar → produces a strong match

Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

Object-Graph matching

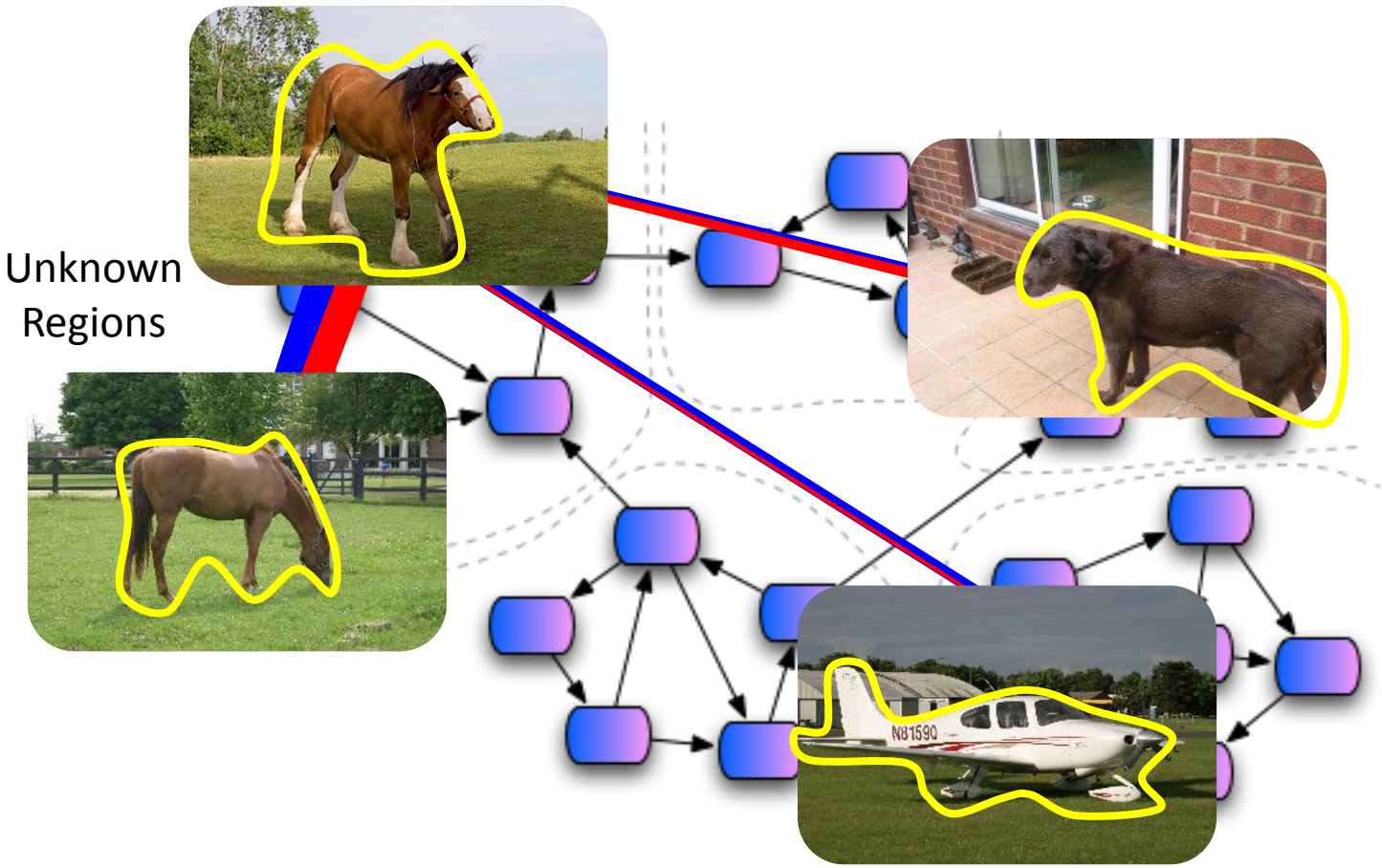


Known classes
b: building
t: tree
g: grass
r: road

- Object-graphs are partially similar → produces a fair match

Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

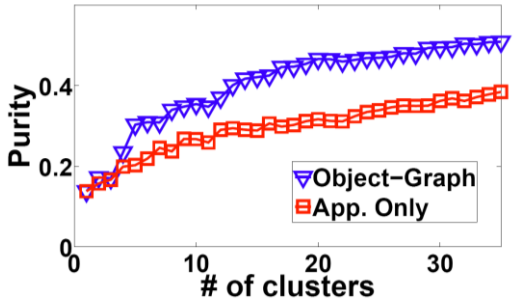
Clusters from region-region affinities



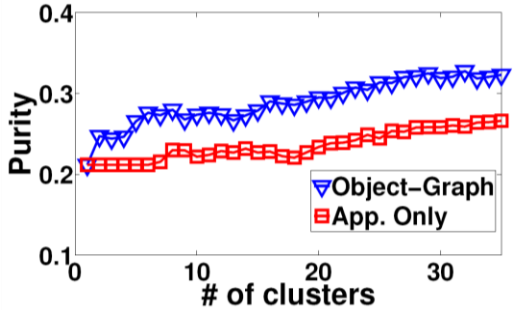
$$K(s_i, s_j) = K_{app}(s_i, s_j) + K_{obj-graph}(s_i, s_j)$$

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010

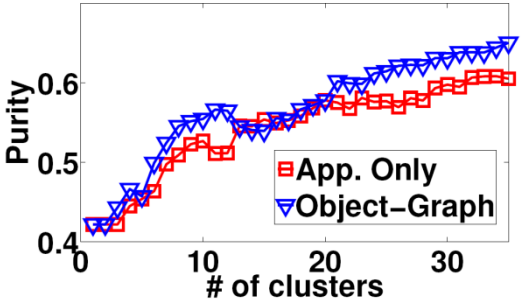
Object Discovery Accuracy

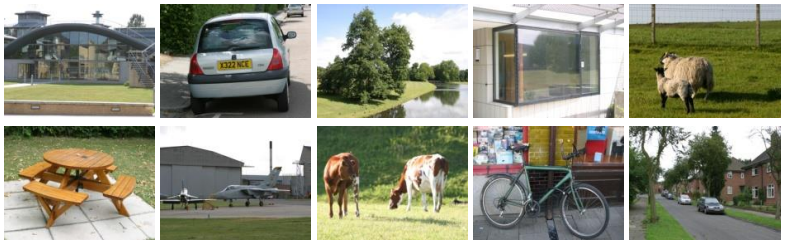


MSRC-v2

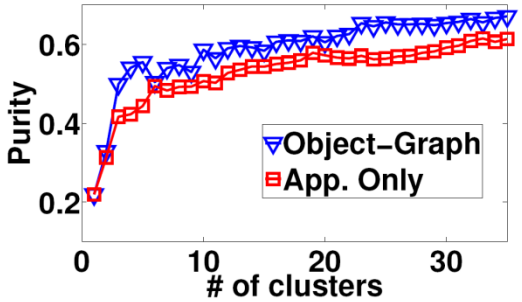


PASCAL 2008





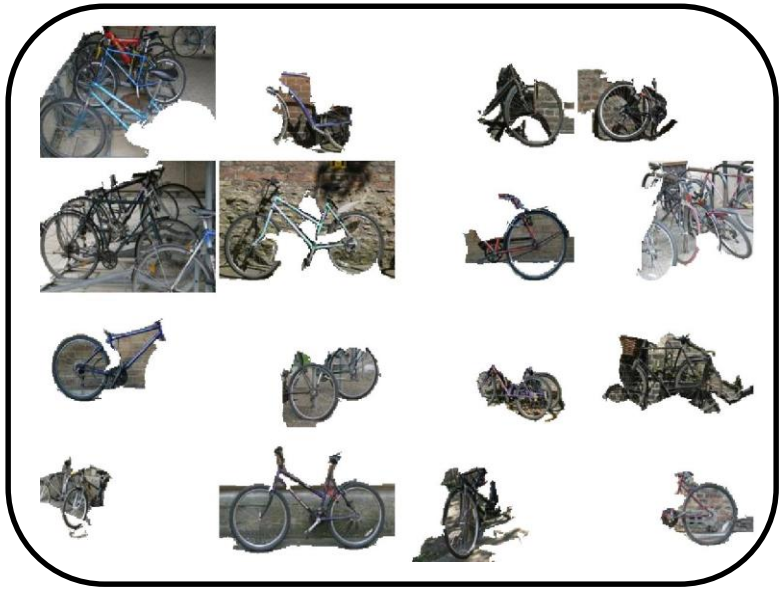
MSRC-v0



Corel

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010

Examples of Discovered Categories



Lee and Grauman, "Object-Graphs for Context-Aware Category Discovery", CVPR 2010

What Makes Paris Look like Paris?

Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, Alexei Efros

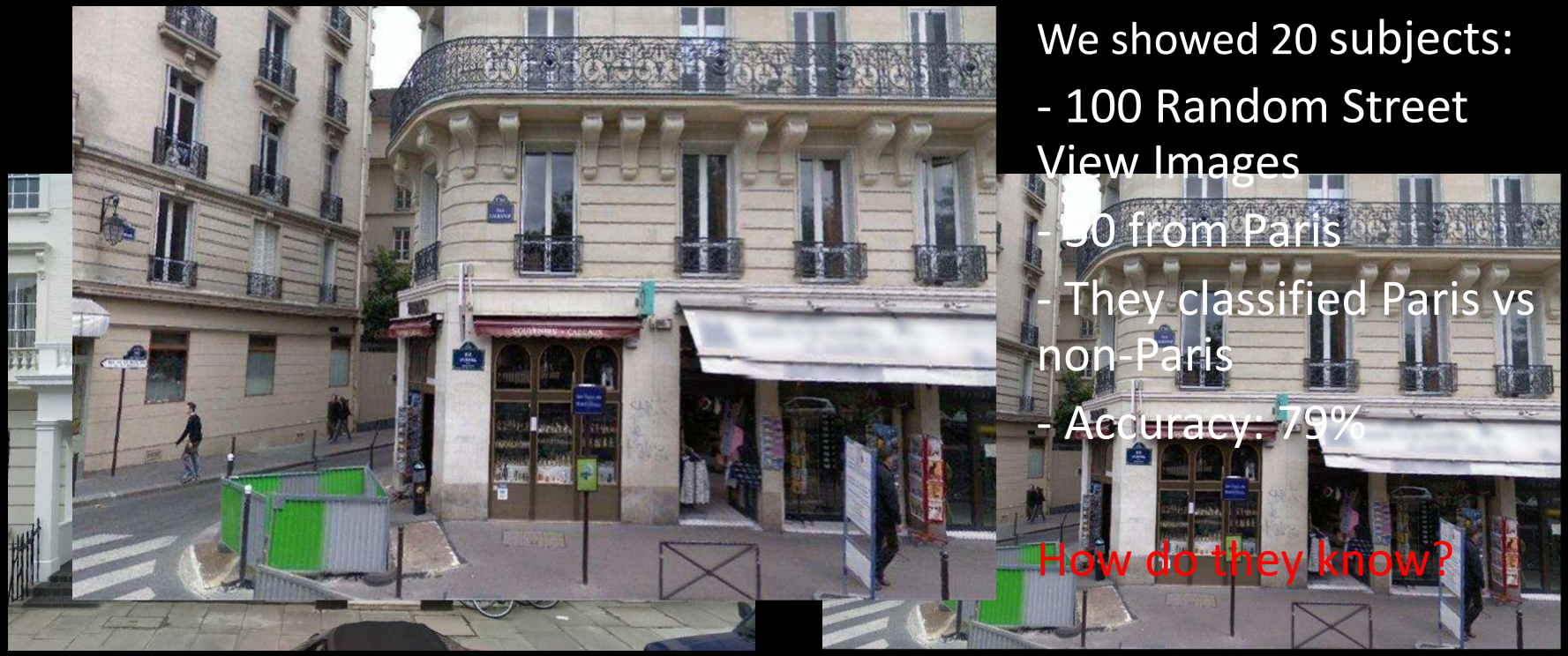
SIGGRAPH 2012

One of these is from Paris
Raise your hand if…

…this is Paris

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Raise your hand if…

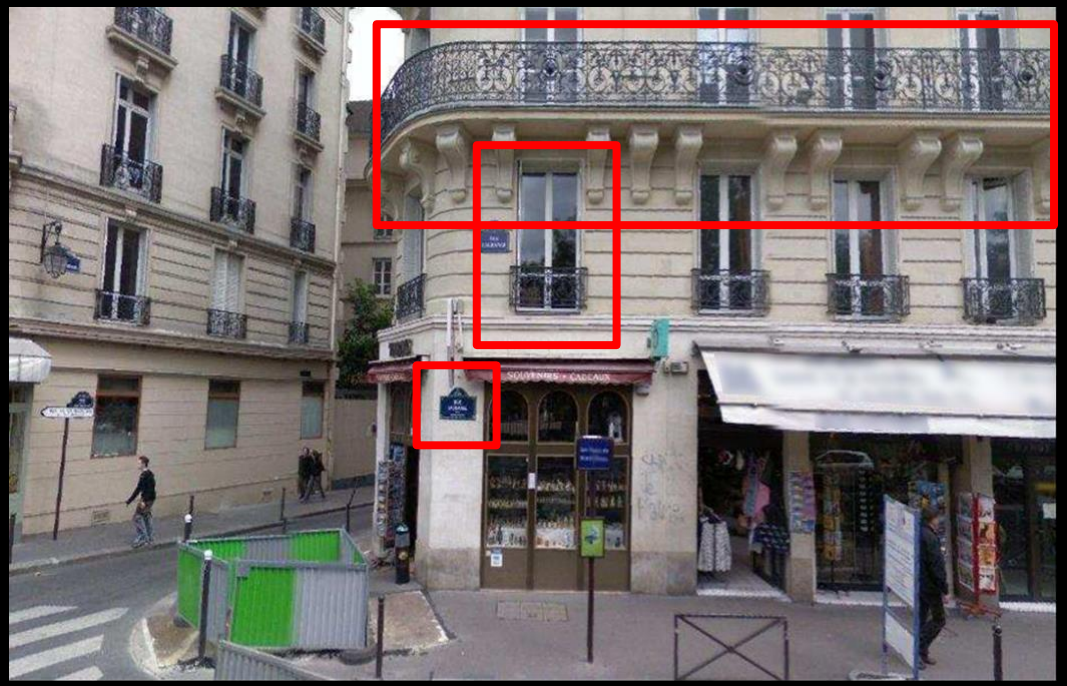


We showed 20 subjects:

- 100 Random Street View Images
- 50 from Paris
- They classified Paris vs non-Paris
- Accuracy: 79%

How do they know?

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012



We showed 20 subjects:

- 100 Random Street View Images
- 50 from Paris
- They classified Paris non-Paris
- Accuracy: 79%

How do they know?

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Our Goal:

Given a large geo-tagged image dataset, we automatically discover ***visual elements*** *that characterize a geographic location*

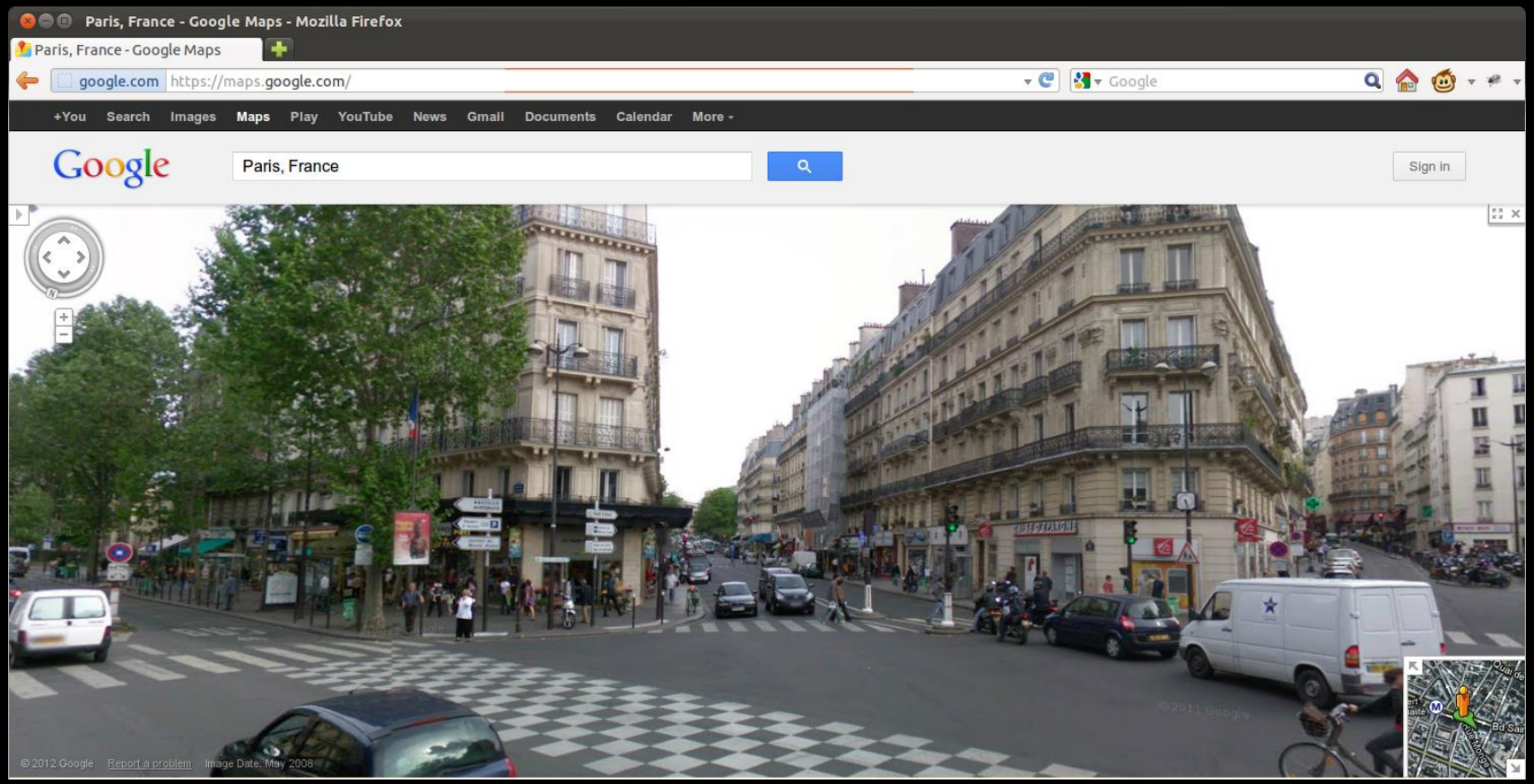
Why might this be a useful task?

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Our Hypothesis

- The visual elements that capture Paris:
  - Frequent: Occur often in Paris
  - Discriminative: Are not found outside Paris

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012



Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Step 1: Nearest Neighbors for Every Patch

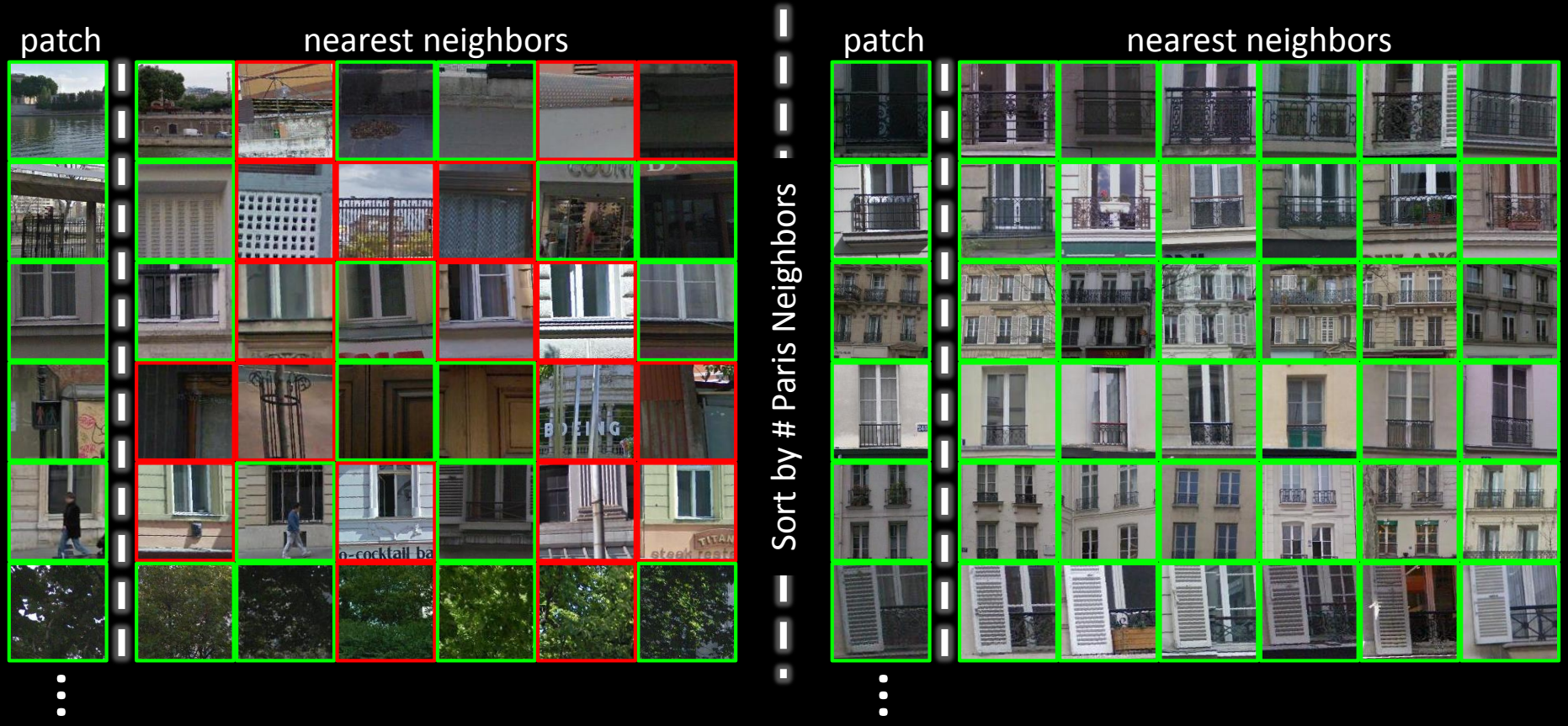
Using normalized correlation of HOG features as a distance metric

Paris

Not Paris

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

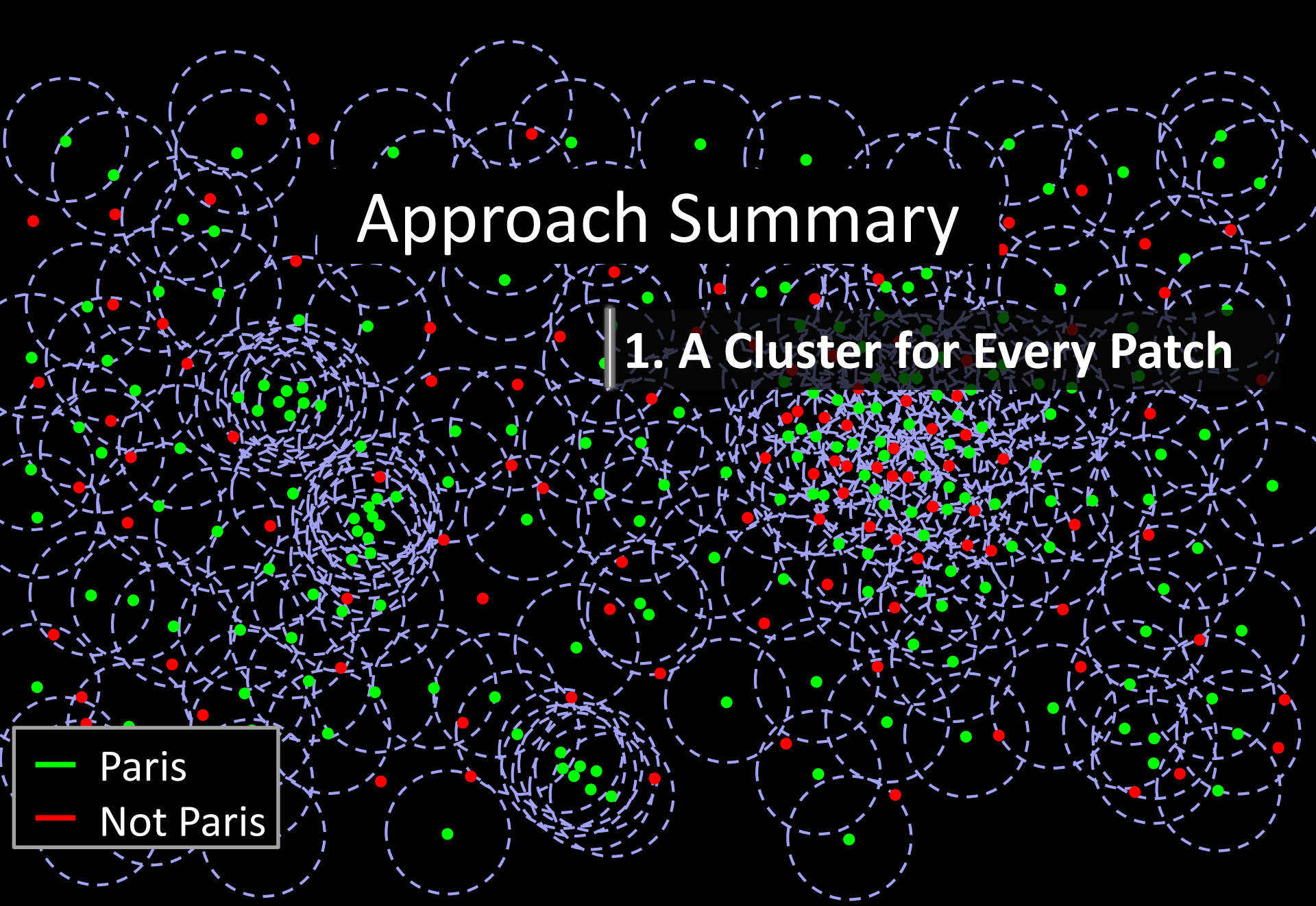
Step 2: Find the Parisian Clusters by Sorting



Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Approach Summary

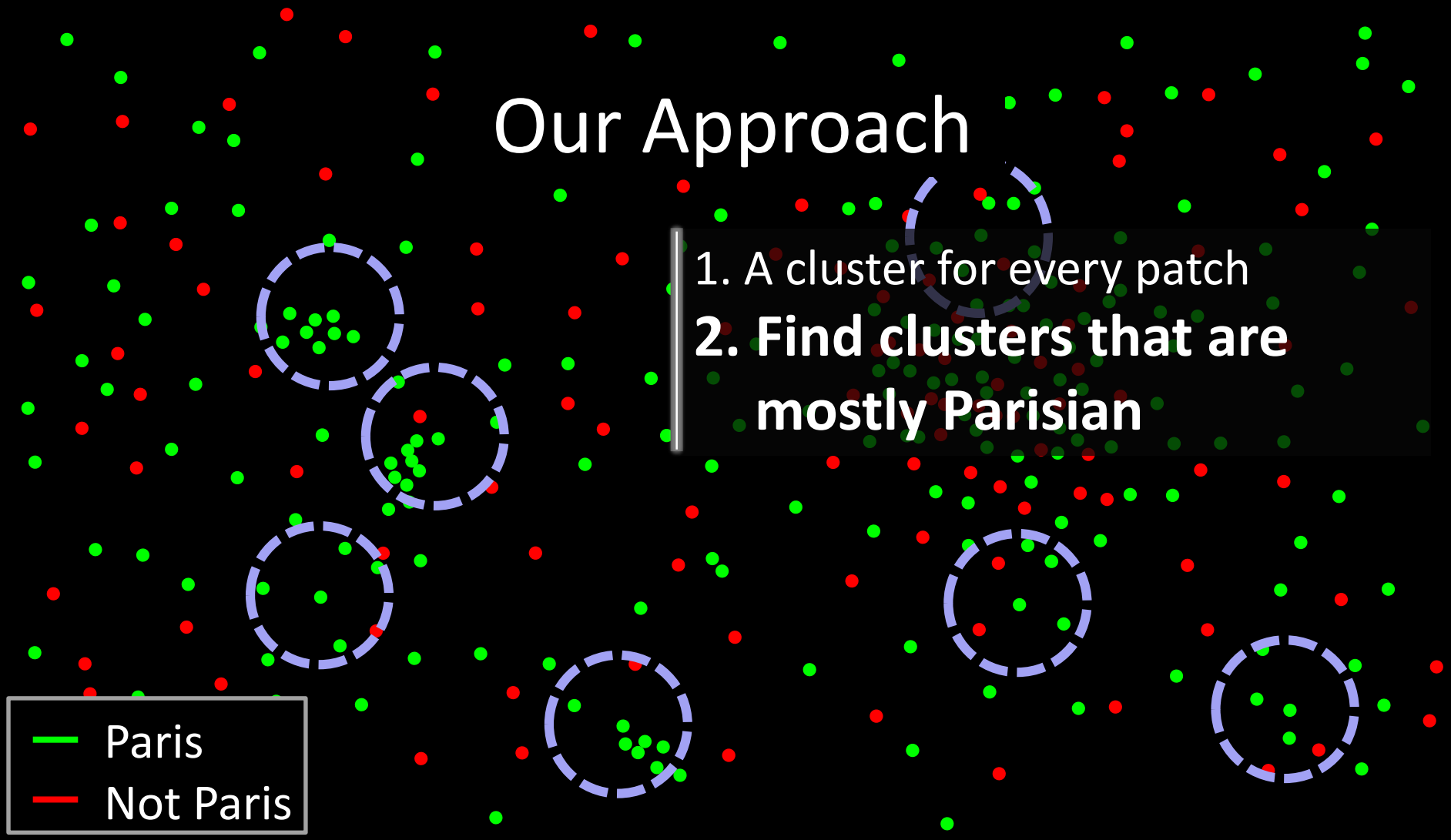
1. A Cluster for Every Patch



Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Our Approach

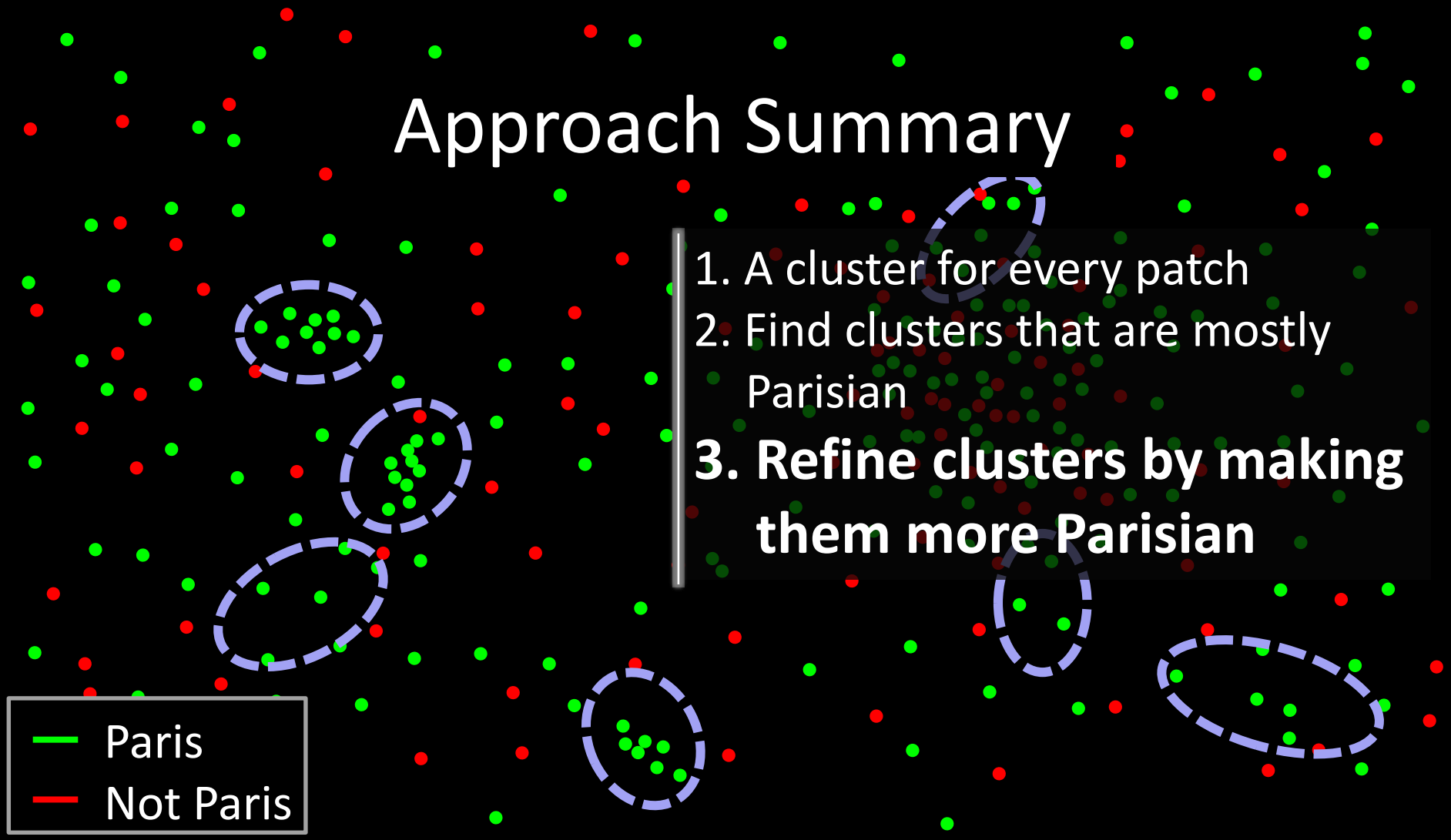
1. A cluster for every patch
2. **Find clusters that are mostly Parisian**



Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Approach Summary

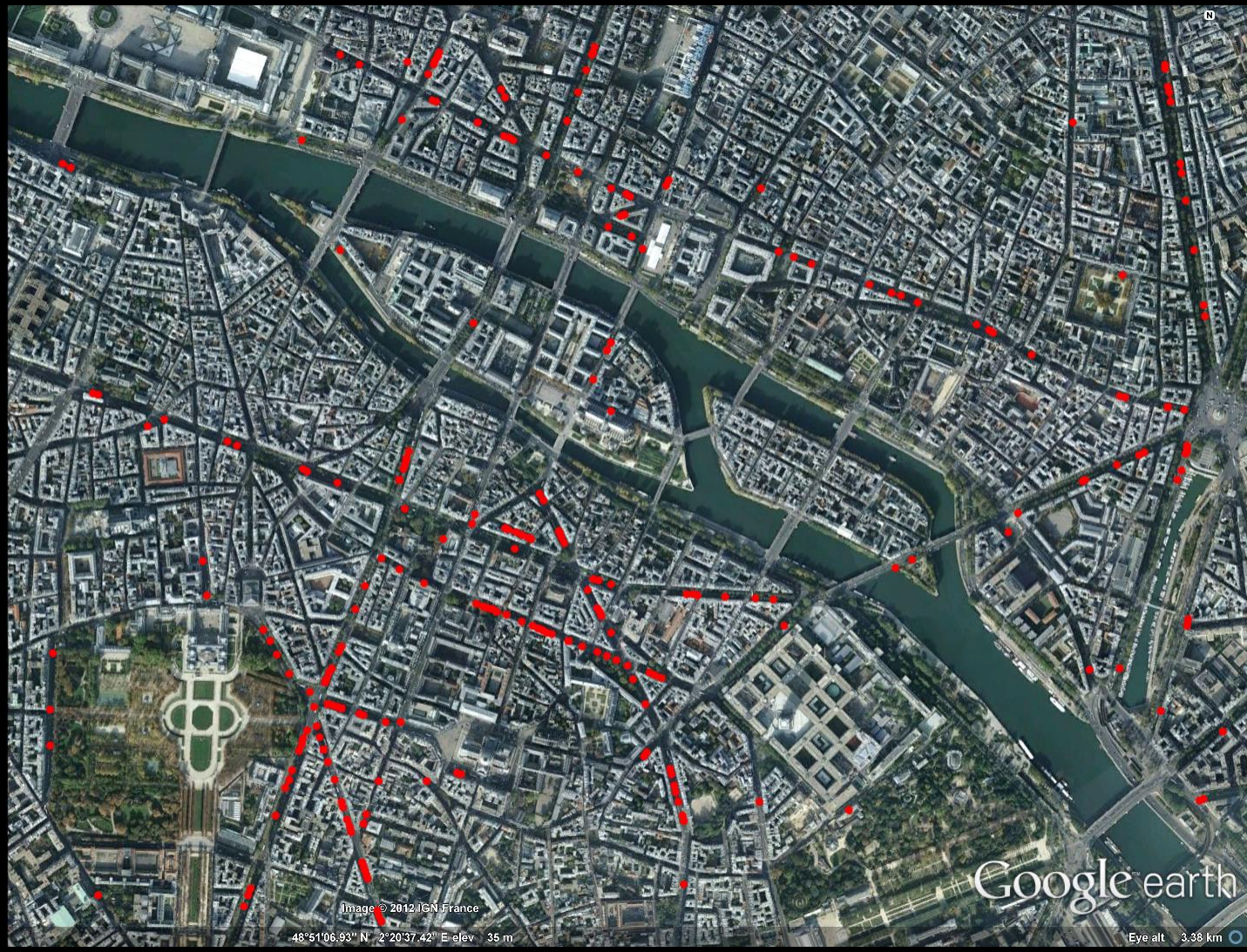
1. A cluster for every patch
2. Find clusters that are mostly Parisian
3. **Refine clusters by making them more Parisian**



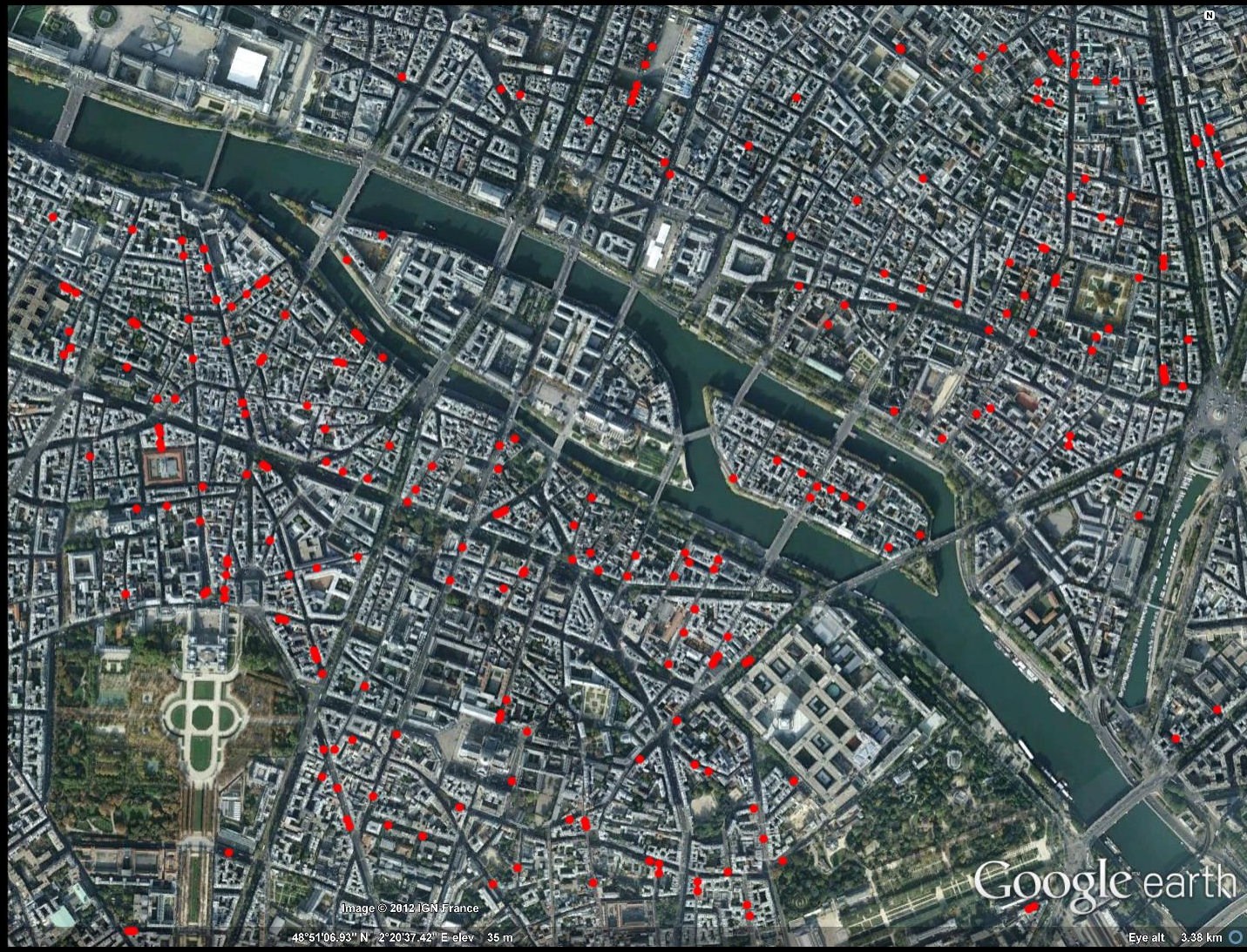
Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Paris: A Few Top Elements

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012



Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012



Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

Elements from Prague

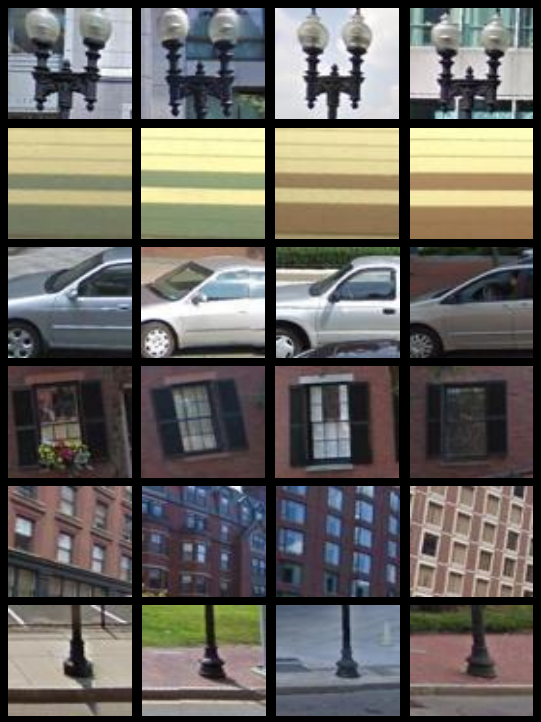
Elements from London

Elements from Barcelona

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

In the U.S.

Elements from San Francisco



Elements from Boston

Doersch et al., “What Makes Paris Look Like Paris?”, SIGGRAPH 2012

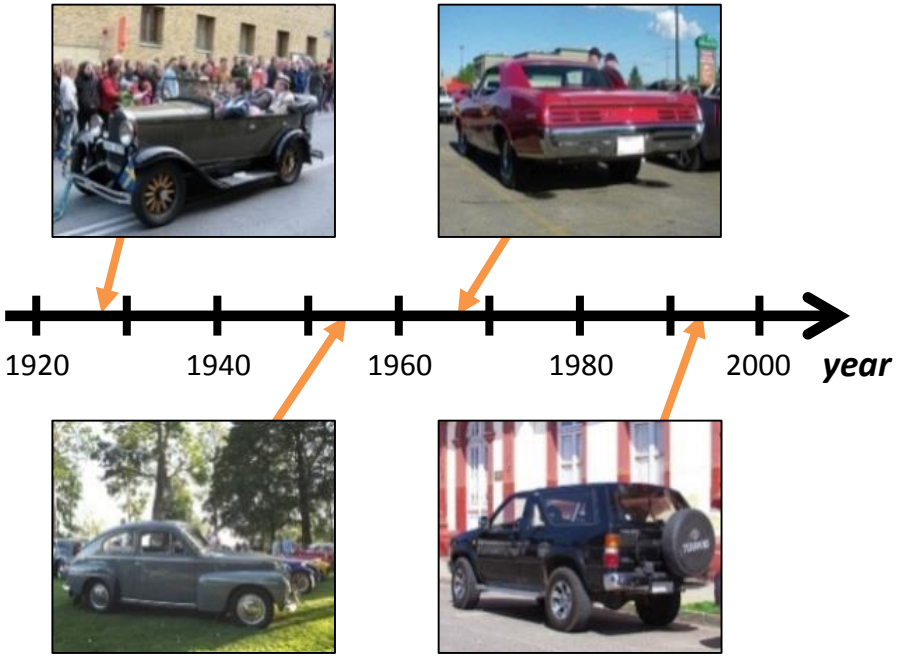
Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time

Yong Jae Lee, Alexei A. Efros, and Martial Hebert

ICCV 2013

Our Goal

- Mine mid-level visual elements in temporally- and spatially-varying data and model their "visual style"



when?

Historical dating of cars

where?

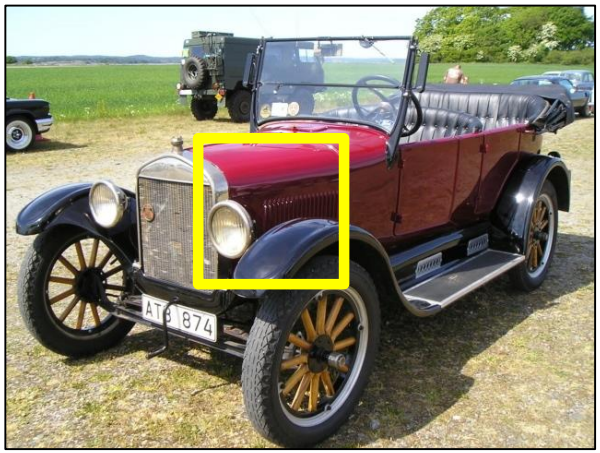
Geolocalization of StreetView images

Lee et al., "Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time", ICCV 2013

Key Idea

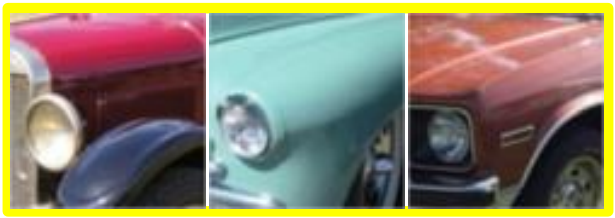
1) Establish connections

1926



1947

1975



1926 1947 1975

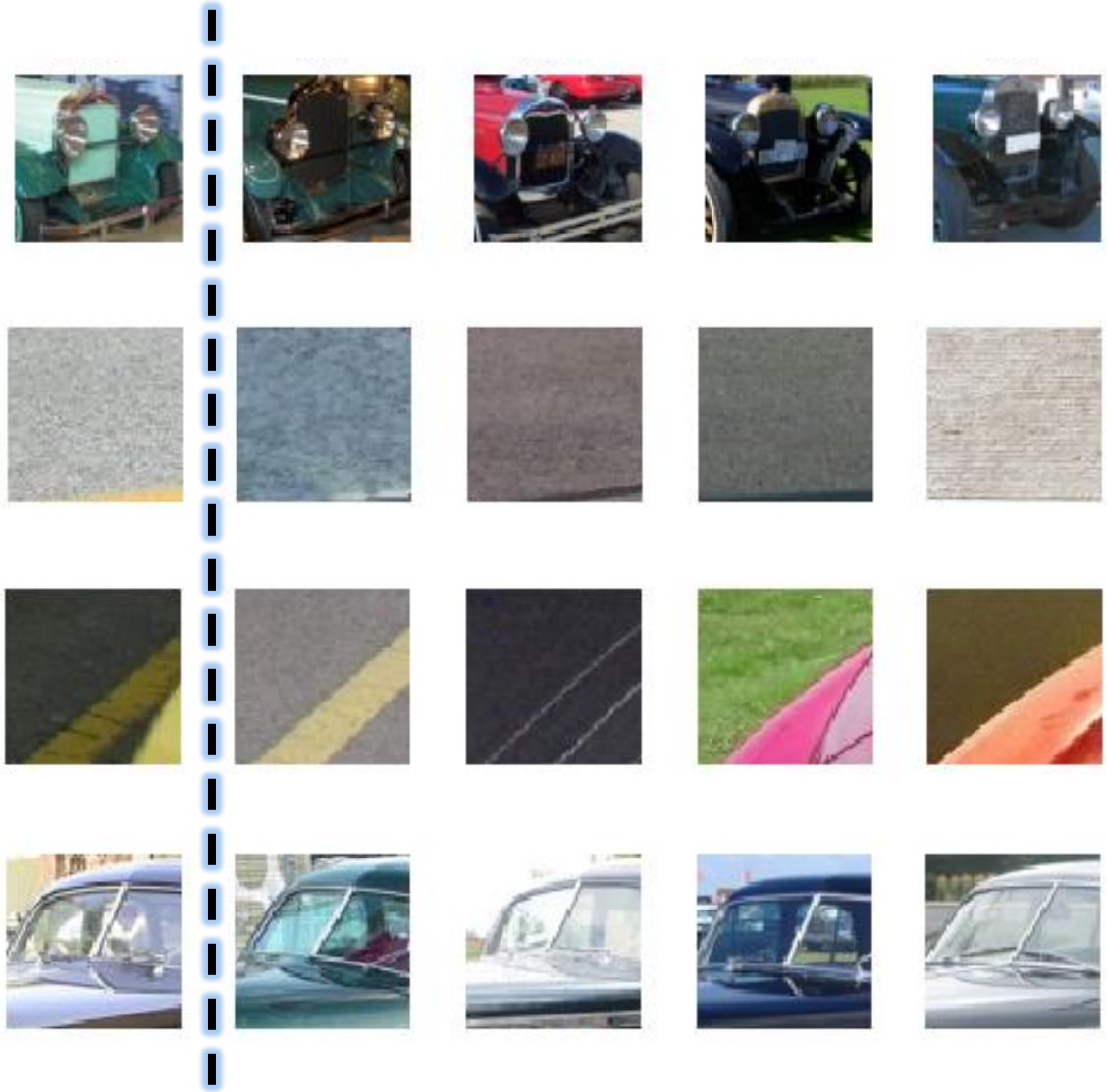
"closed-world"

2) Model style-specific differences

Lee et al., "Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time", ICCV 2013

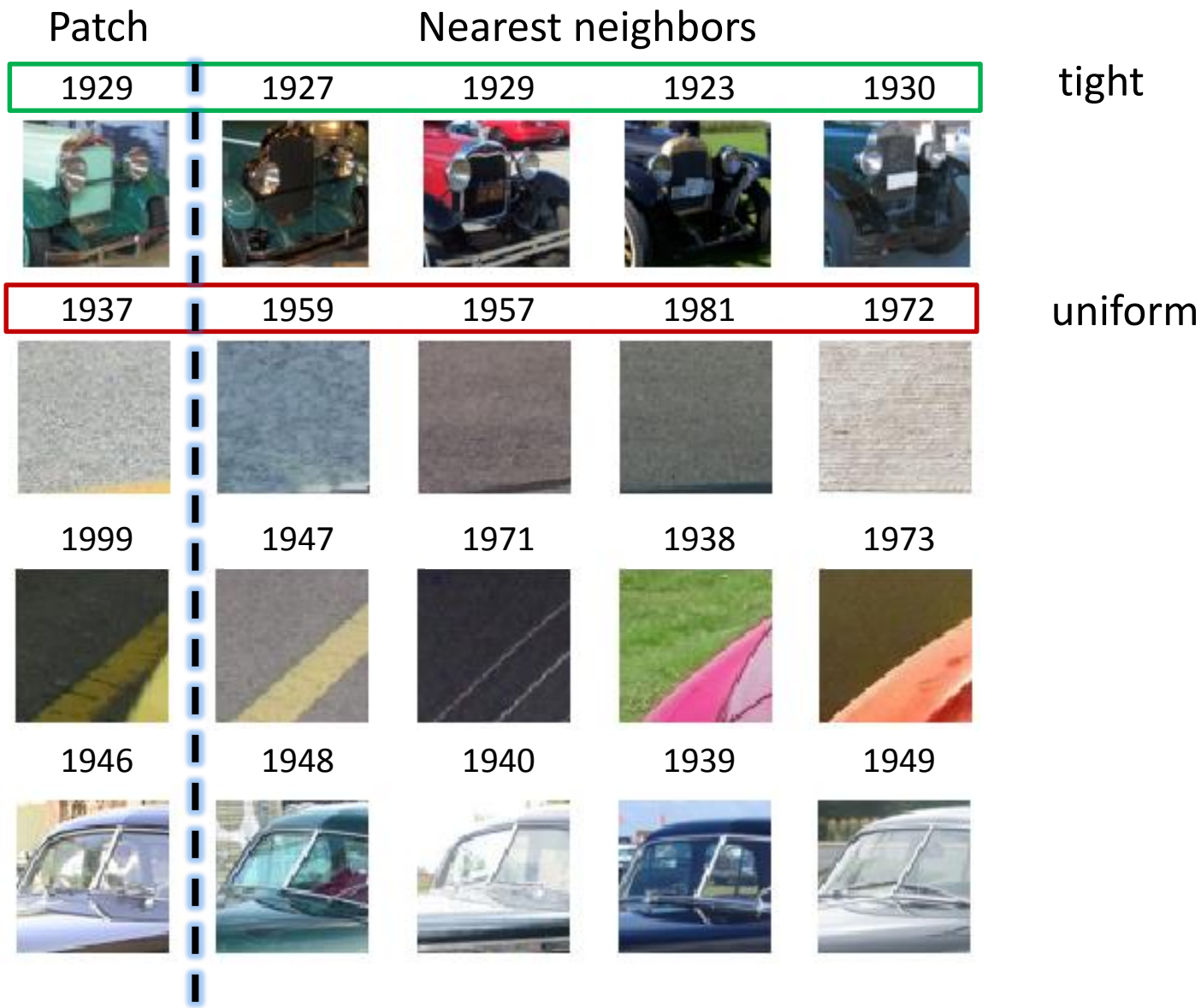
Mining style-sensitive elements

Patch | Nearest neighbors



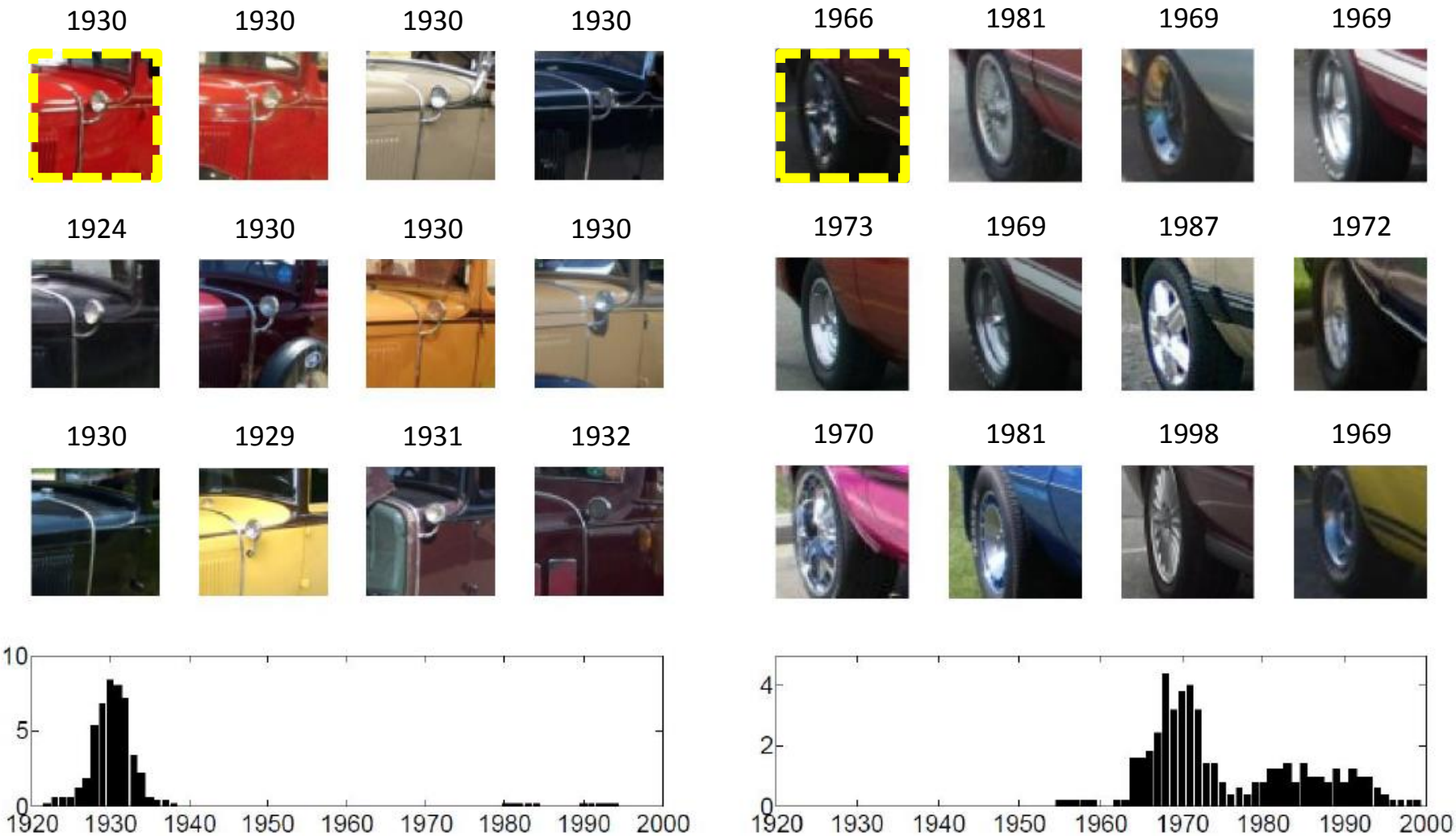
Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013

Mining style-sensitive elements



Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013

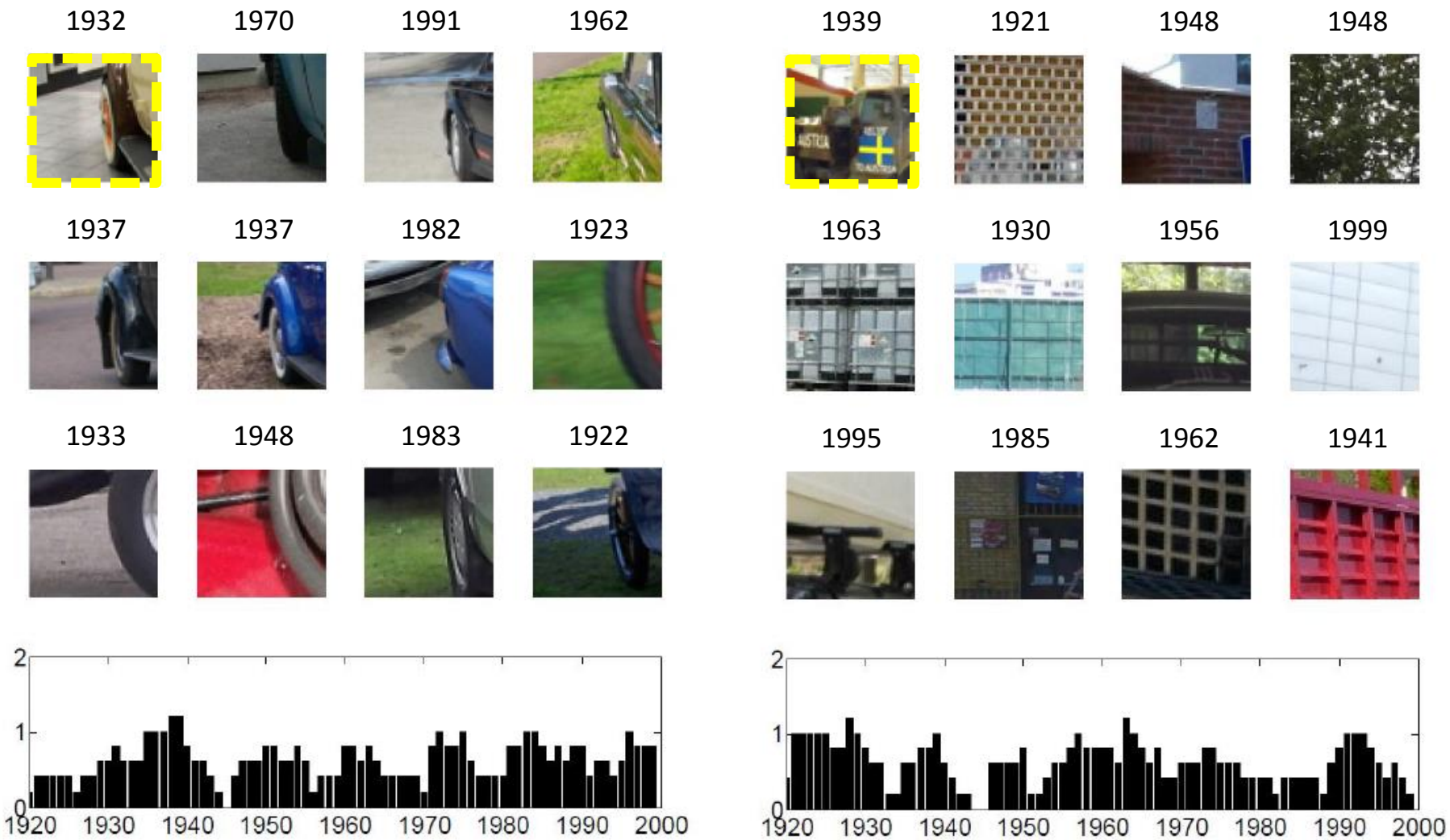
Style-sensitive elements



(a) Peaky (low-entropy) clusters

Lee et al., "Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time", ICCV 2013

Uninformative elements



(b) Uniform (high-entropy) clusters

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013