From Visual Subcategory to Webly-supervised Visual Recognition

Moin Nabi

Santosh Divalla (Allen Institute for Artificial Intelligence)

Ali Farhadi (University of Washington)

Massimiliano Pontil (UCL)

Vittorio Murino (Italian Institute of Technology)

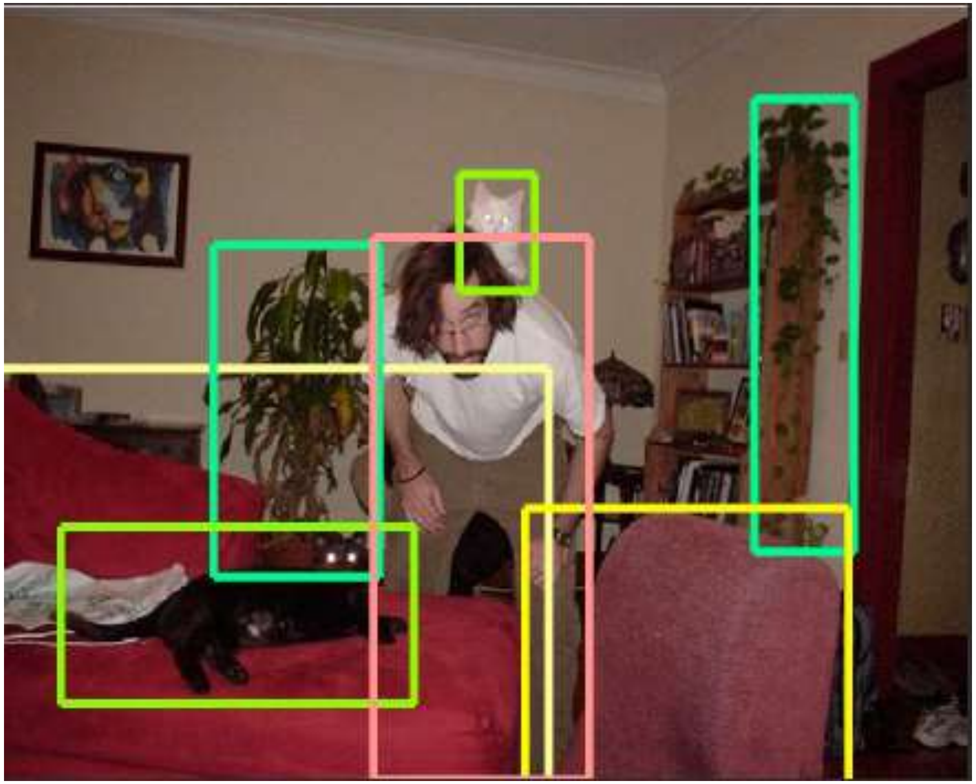
Outline

- Introduction on object detection
- Visual subcategory
- Evolution of DPM
- web-supervised visual recognition
- Webly-supervised discriminative patch
- Subcategory and dataset bias

Question

What objects are where?

Goal: detecting objects in cluttered images



person, plant, cat, dog chair, sofa, car, bicycle, motorbike, table, plane, …

Application

- Applications
  - Semantic image and video search
  - Human-computer interaction (*e.g.*, Kinect)
  - Automotive safety
  - Camera focus-by-detection
  - Surveillance
  - Semantic image and video editing
  - Assistive technologies
  - Medical imaging

Challenges

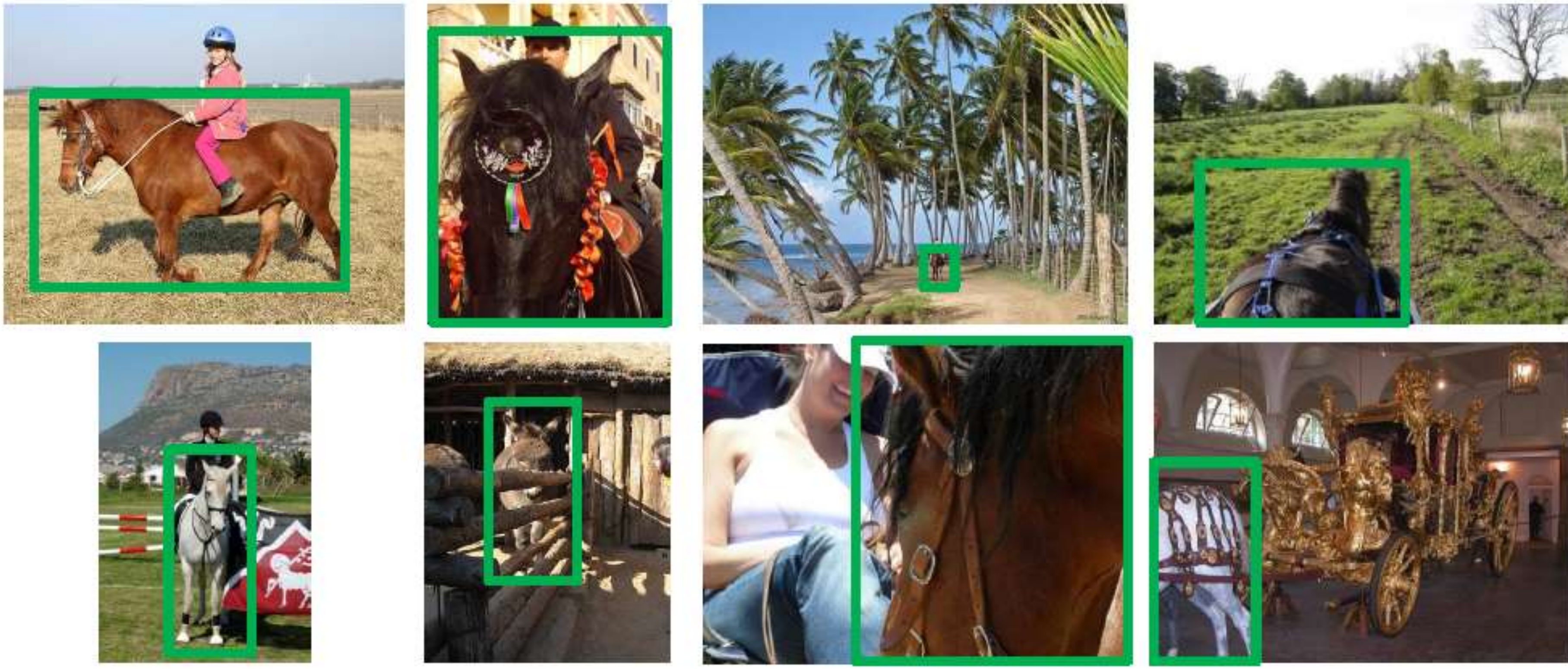
Variation in illumination

Variation in pose, viewpoint

Variation in appearance

Occlusion and clutter

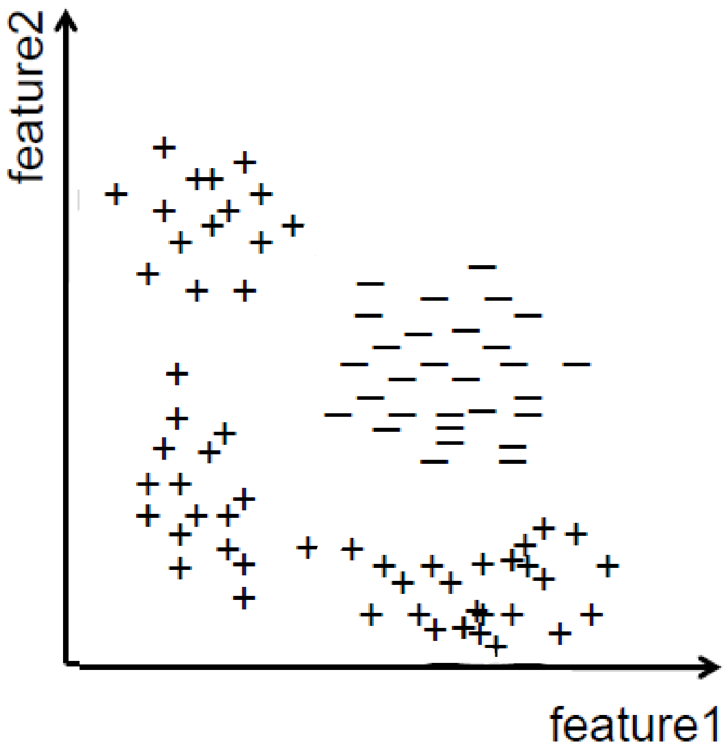
Intra Category Diversity



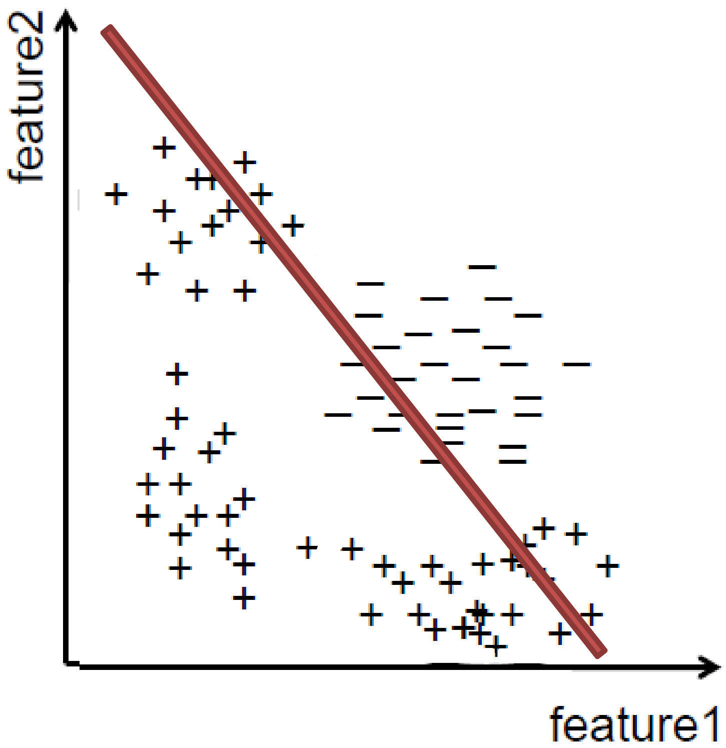
Example images for “Horse” from PASCAL VOC

Variation due to change in camera viewpoint, object pose, and occlusion

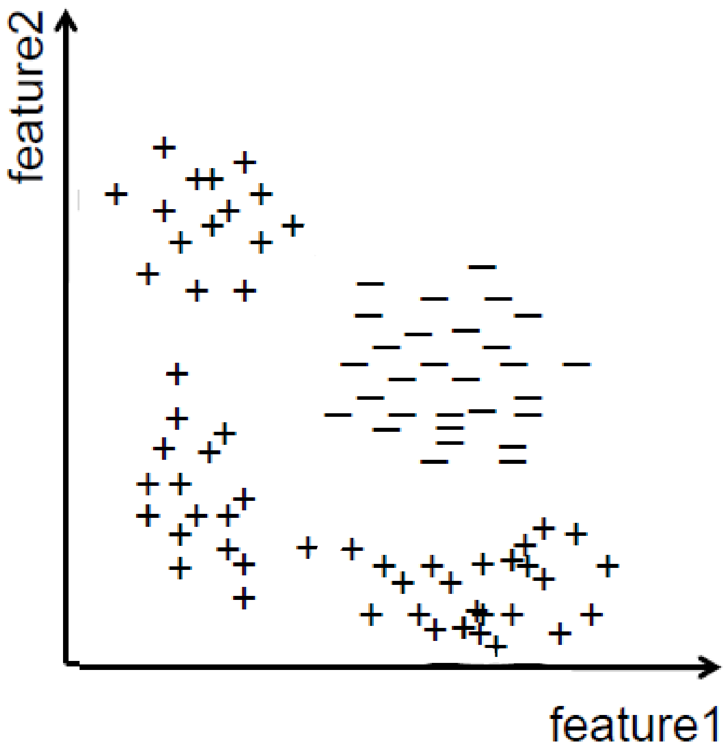
Subcategories



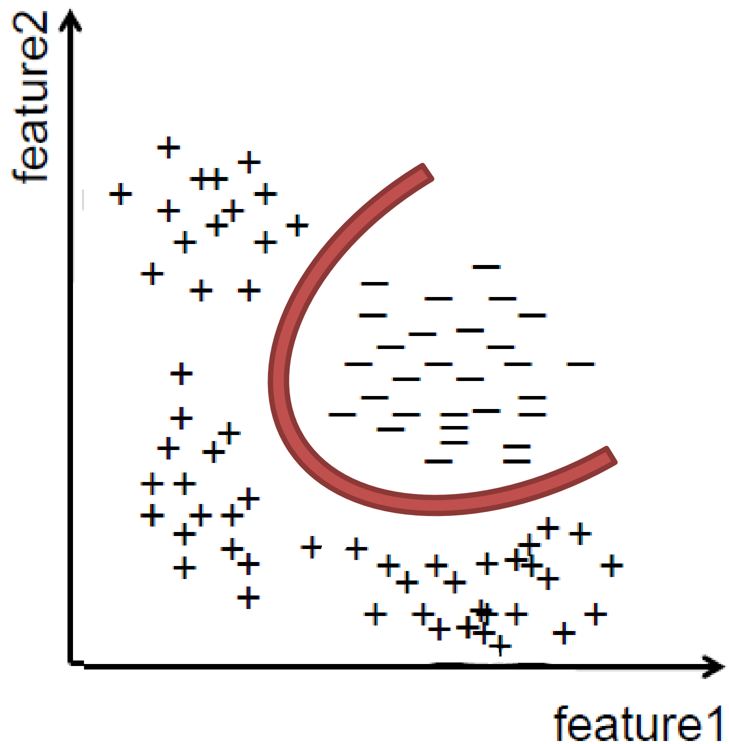
Subcategories



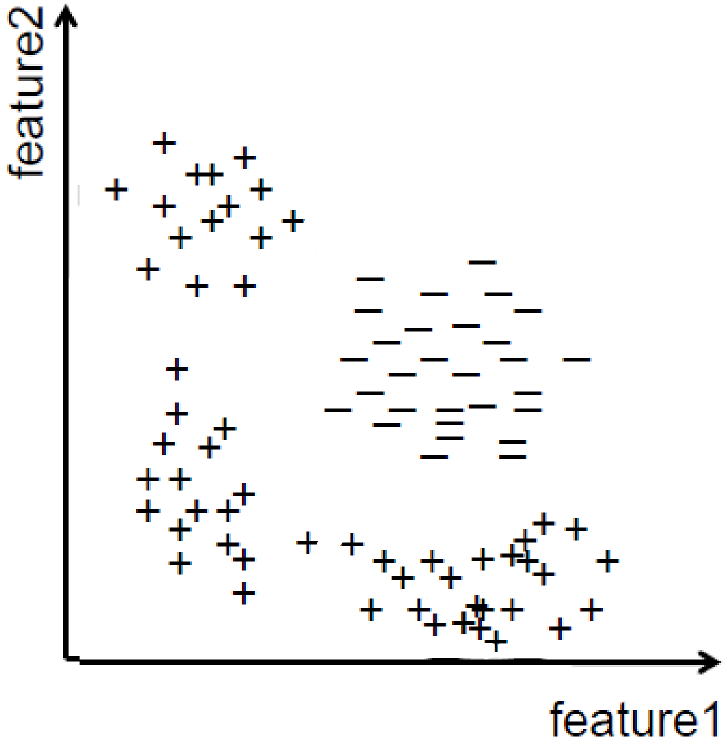
Subcategories



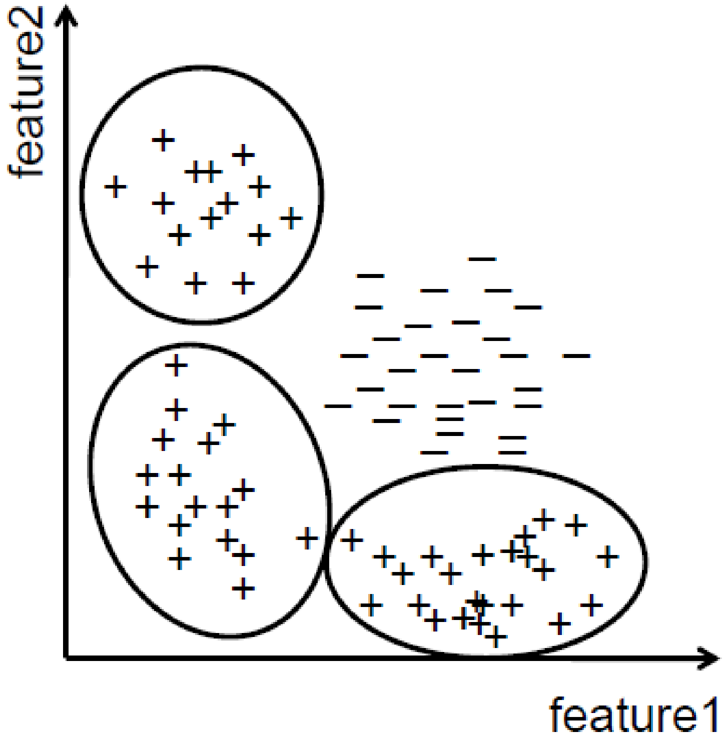
Subcategories



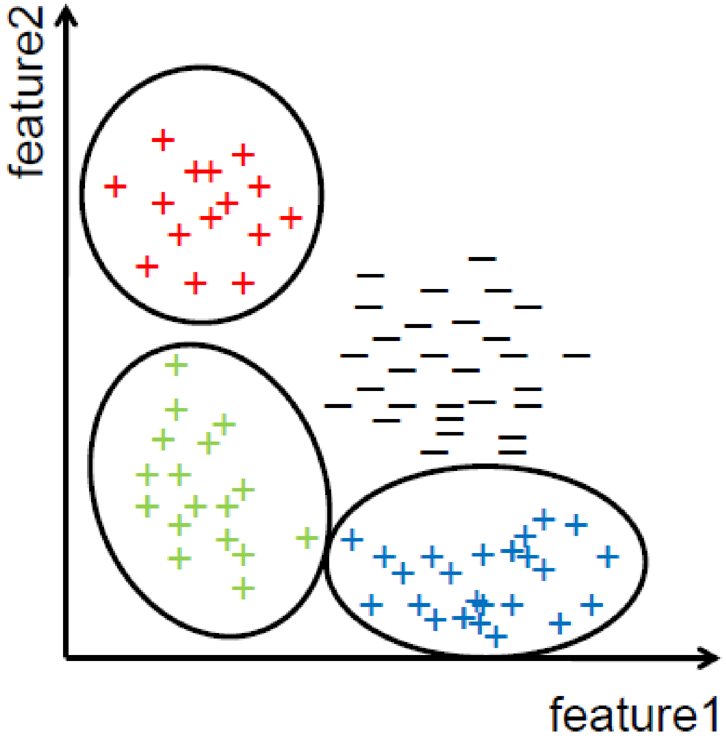
Subcategories



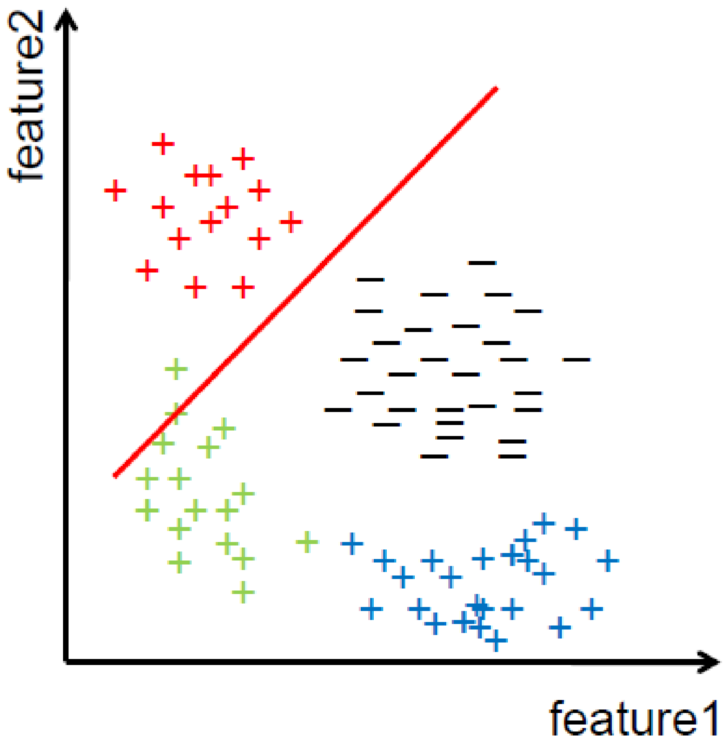
Subcategories



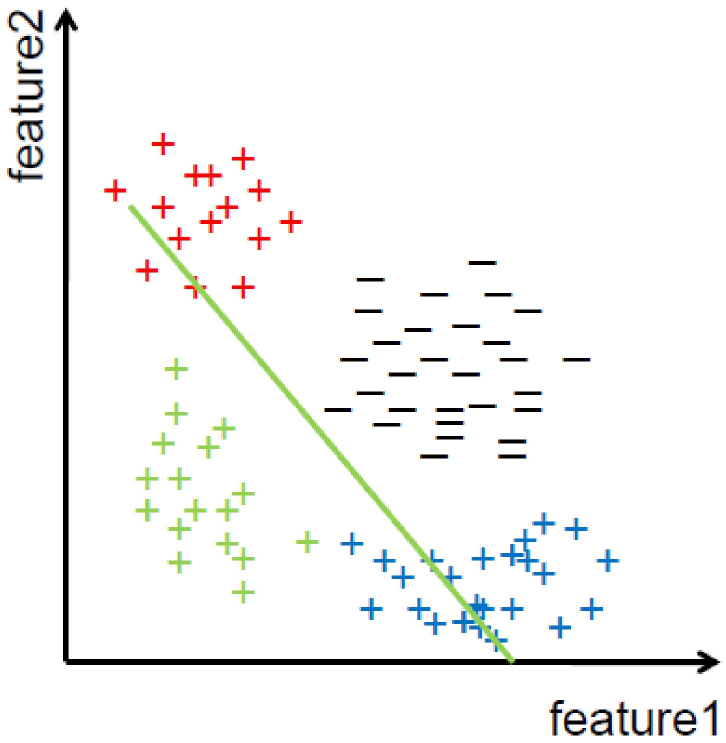
Subcategories



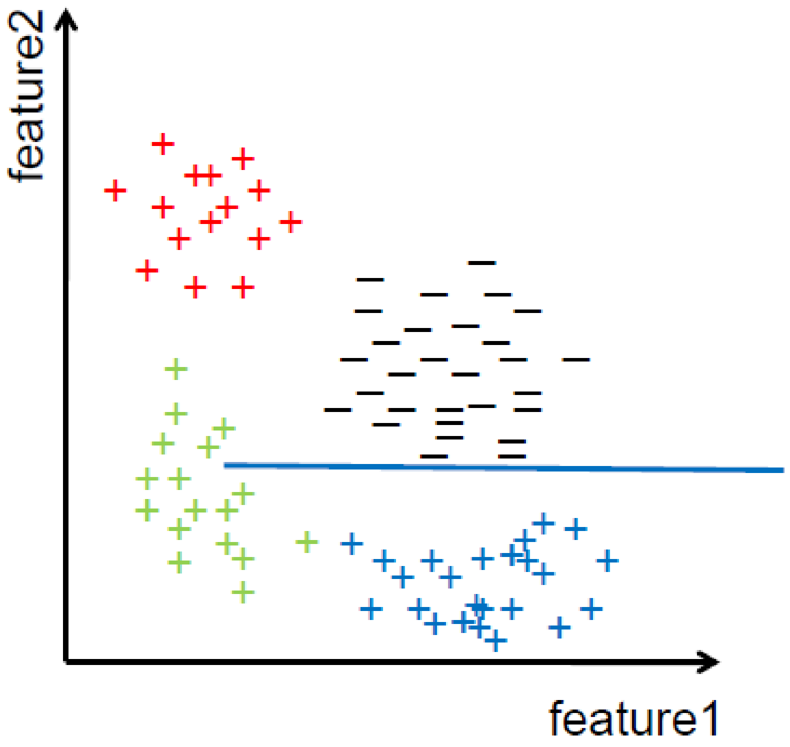
Subcategories



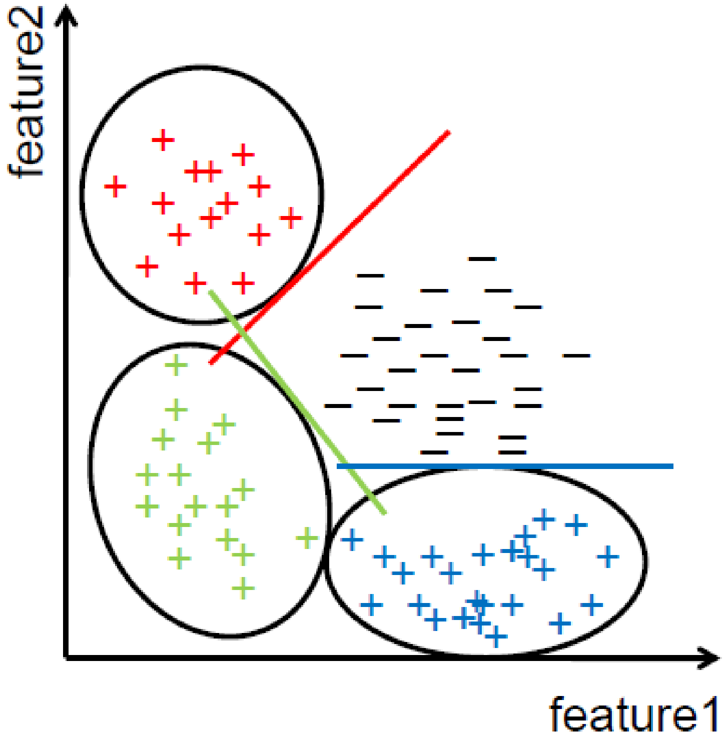
Subcategories



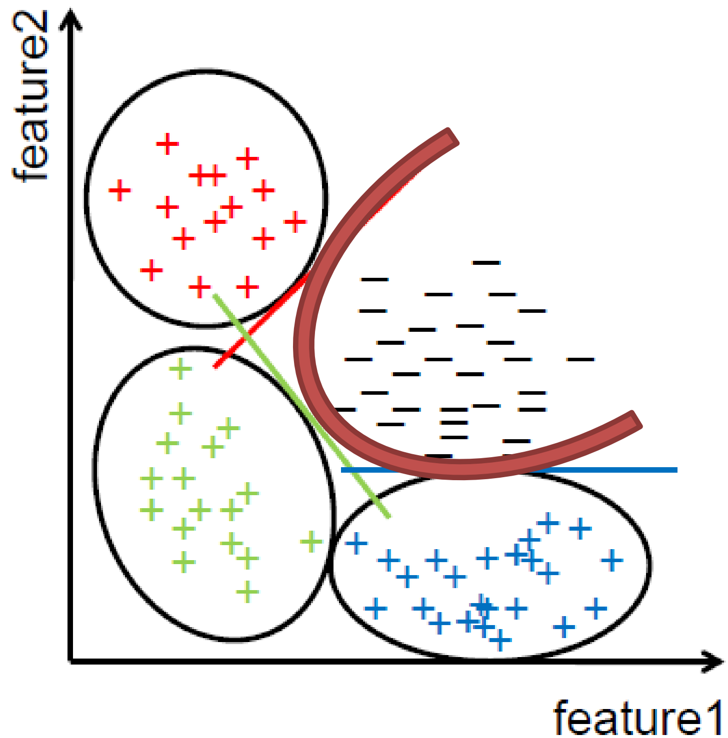
Subcategories



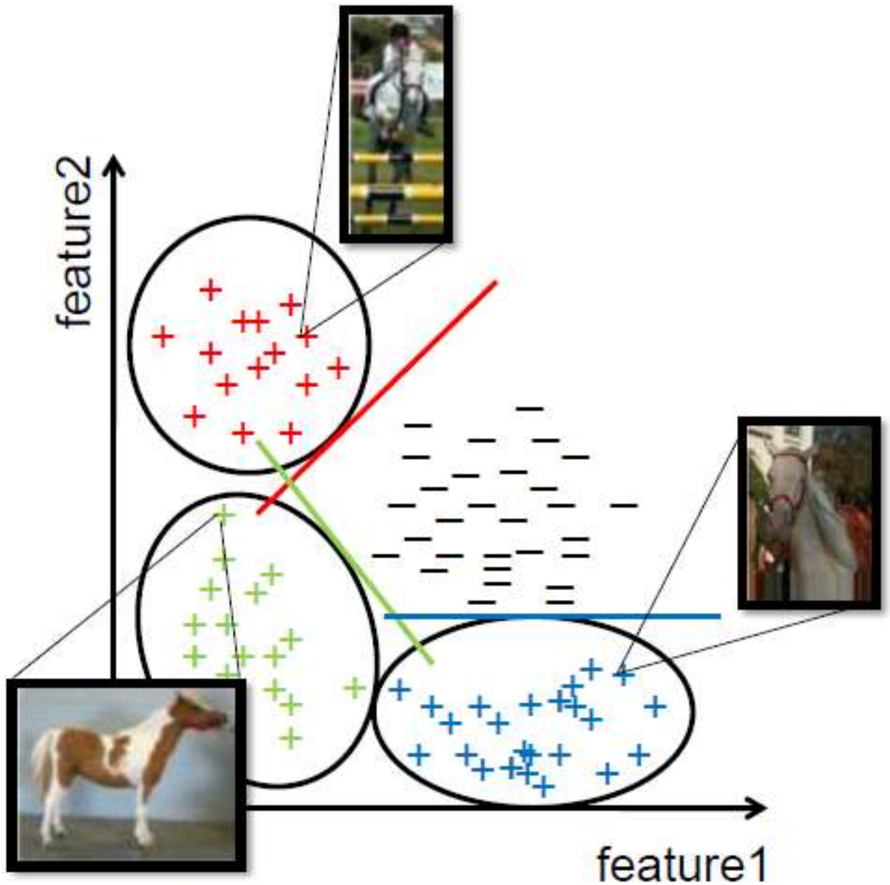
Subcategories



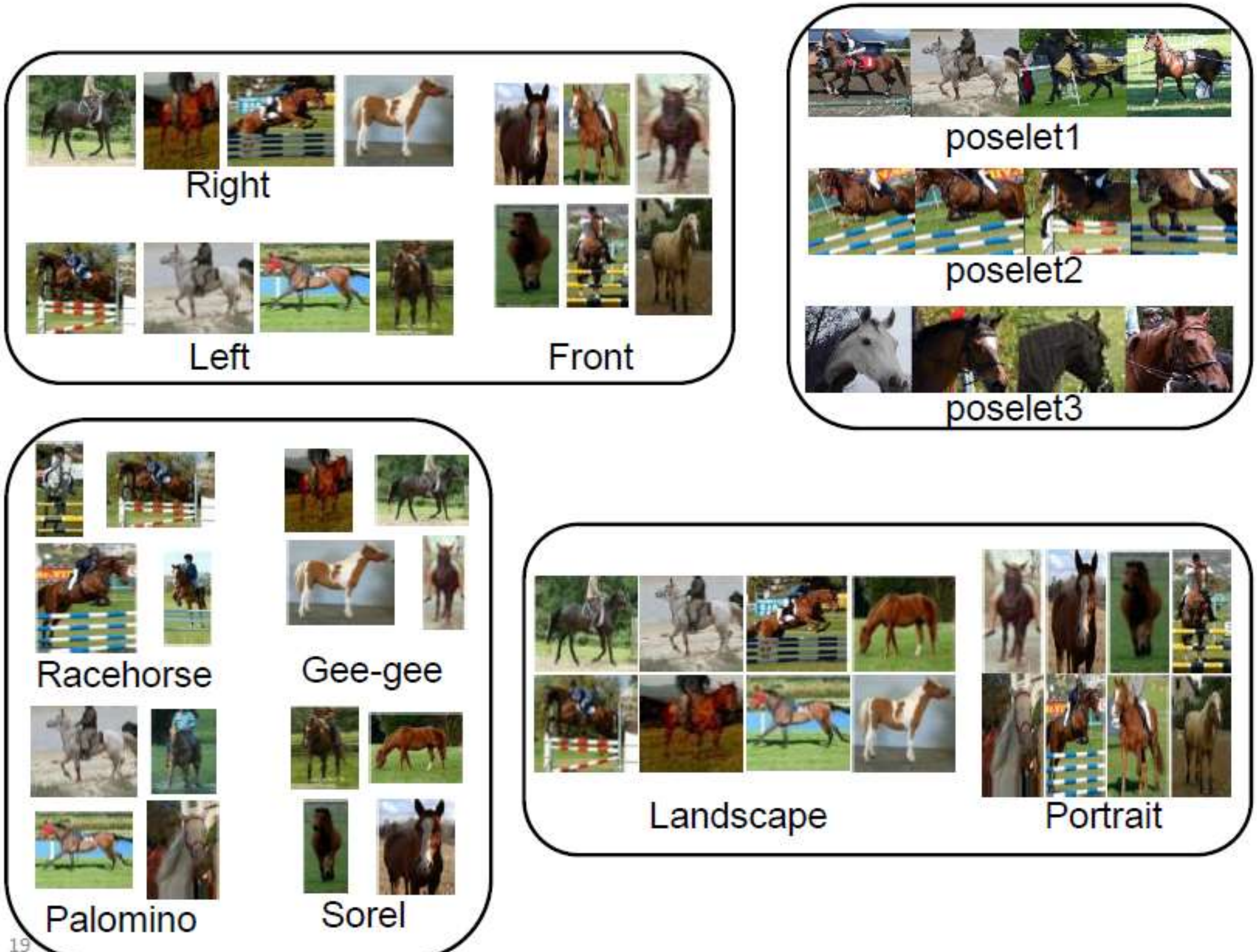
Subcategories



Subcategories



Subcategories



The Evolution of DPM

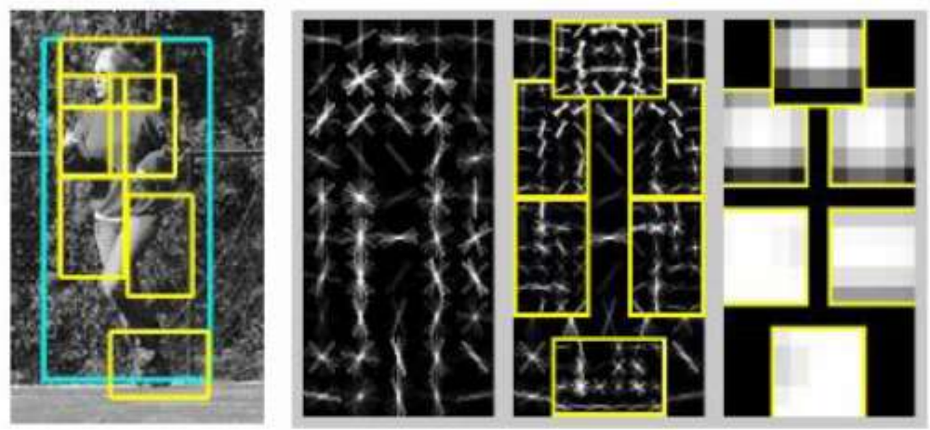
K=1 no Parts [DT'05]
mAP = 0.17

HOG (Histogram of
Oriented Gradients)

The Evolution of DPM

K=1 6 Parts [PFF'08]
mAP = 0.21

K=1 no Parts [DT'05]
mAP = 0.17



Image

Root filter (Coarse resolution)

Part filters (Fine resolution)

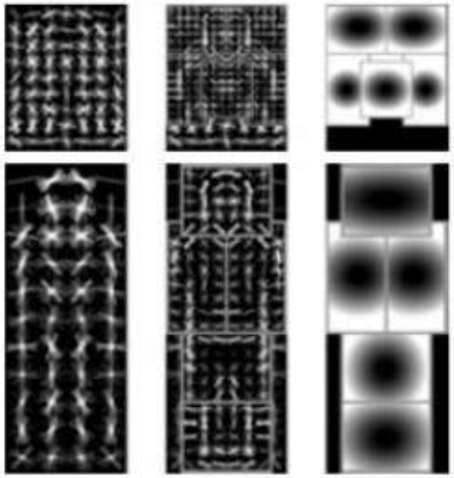
Deformation Models

The Evolution of DPM

K=1 no Parts [DT'05]
mAP = 0.17

K=1 6 Parts [PFF'08]
mAP = 0.21

K=2 (ar) 6 Parts [PFF'10]
mAP = 0.26



2 DPM (PAMI10)

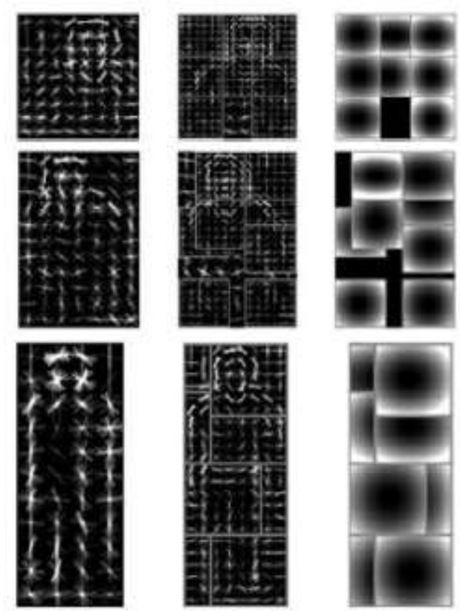
The Evolution of DPM

K=1 no Parts [DT'05]
mAP = 0.17

K=1 6 Parts [PFF'08]
mAP = 0.21

K=2 (ar) 6 Parts [PFF'10]
mAP = 0.26

K=6 (ar) 8 Parts [PFF'11]
mAP = 0.32



6 DPM (voc-release4)

Santosh's Experiment

K=1 6 Parts [PFF'08]	K=2 (ar) 6 Parts [PFF'10]	K=6 (ar) 8 Parts [PFF'11]	K=15 (app) 8 Parts [This work]
mAP = 0.21	mAP = 0.26	mAP = 0.32	mAP = 0.35

K=1 no Parts [DT'05]
mAP = 0.17

S. Divalla, E. Efros, M. Hebert **"How important are "Deformable Parts" in the Deformable Parts Model?"**, in *ECCV* 2012.

Santosh's Experiment

K=1
6 Parts
[PFF'08]
mAP = 0.21

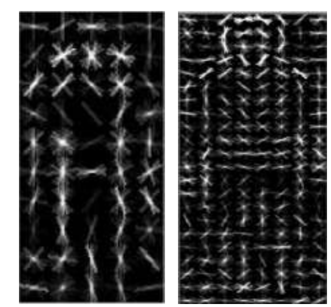
K=2 (ar)
6 Parts
[PFF'10]
mAP = 0.26

K=6 (ar)
8 Parts
[PFF'11]
mAP = 0.32

K=15 (app)
8 Parts
[This work]
mAP = 0.35

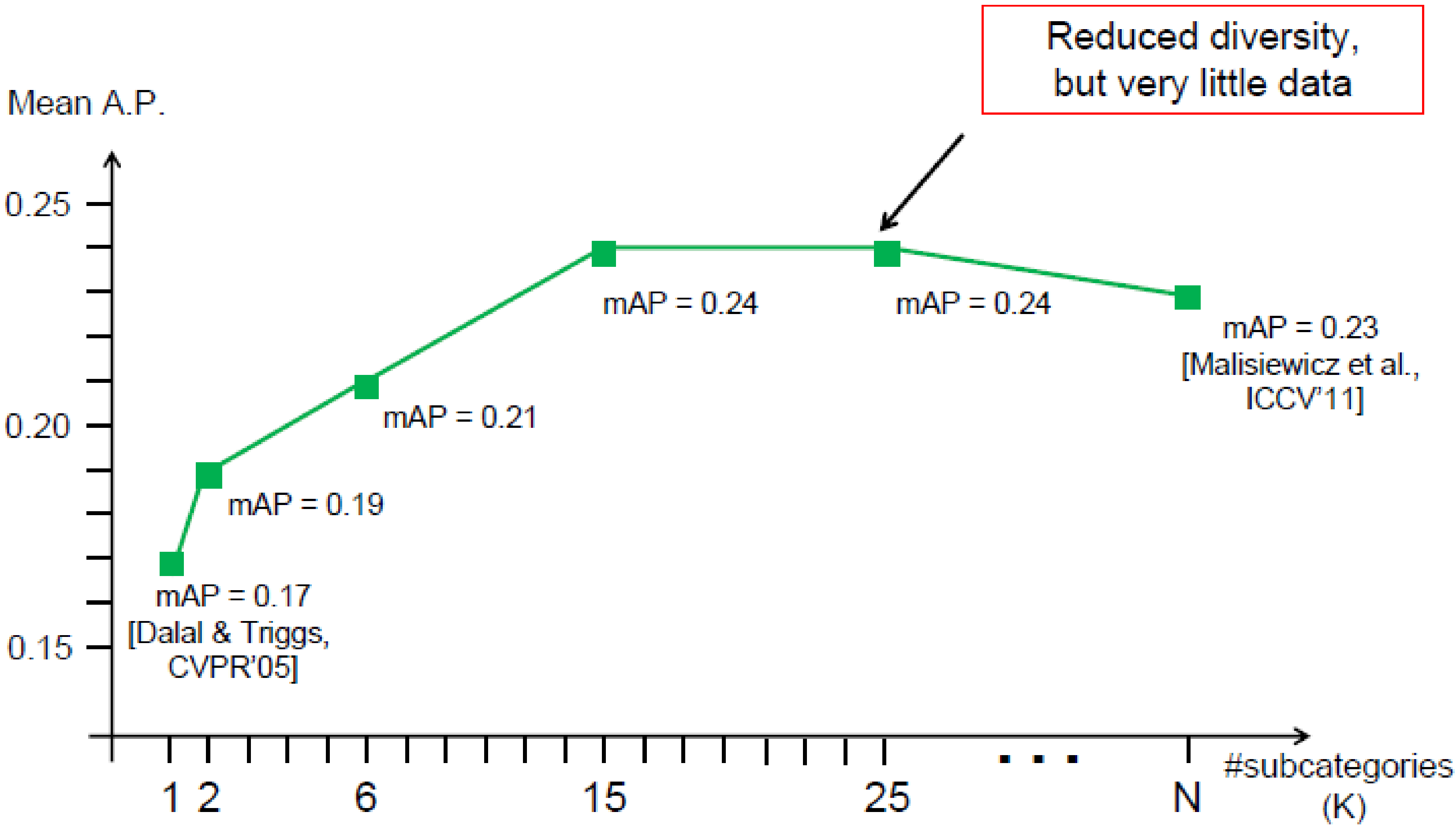
K=1
no Parts
[DT'05]
mAP = 0.17

K=15 (app)
no Parts
[This work]
mAP = 0.31



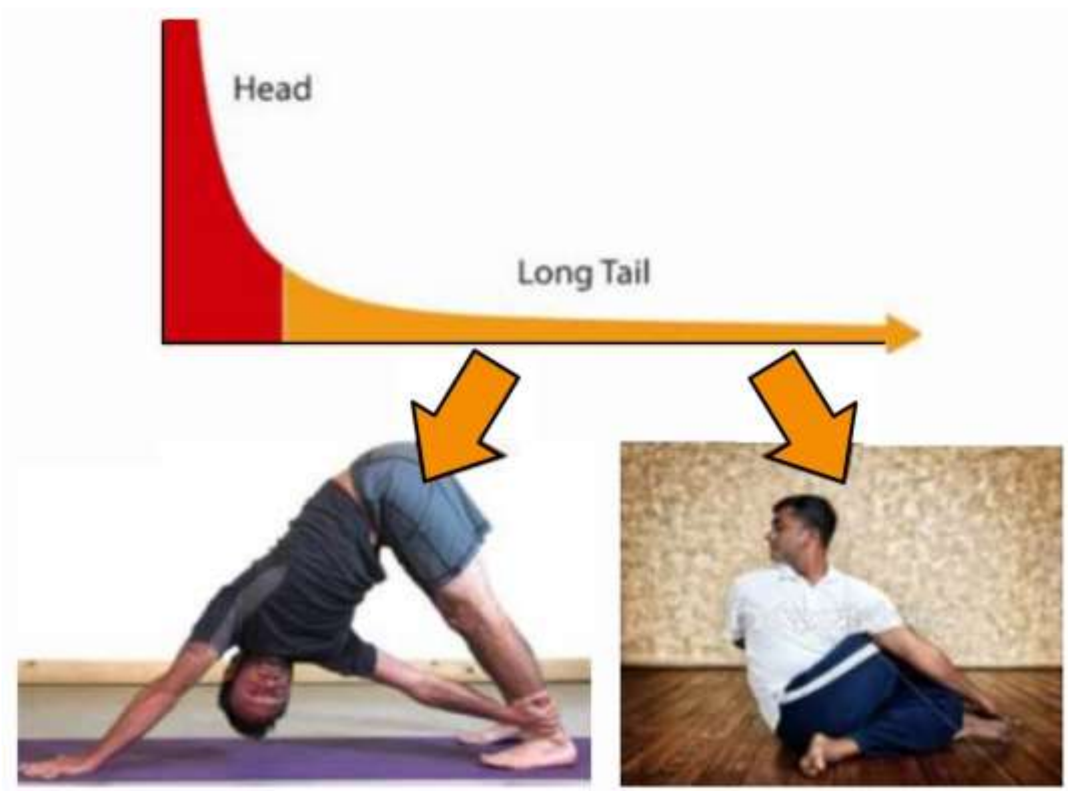
S. Divalla, E. Efros, M. Hebert **"How important are "Deformable Parts" in the Deformable Parts Model?"**, in *ECCV* 2012.

Effect of varying # subcategories



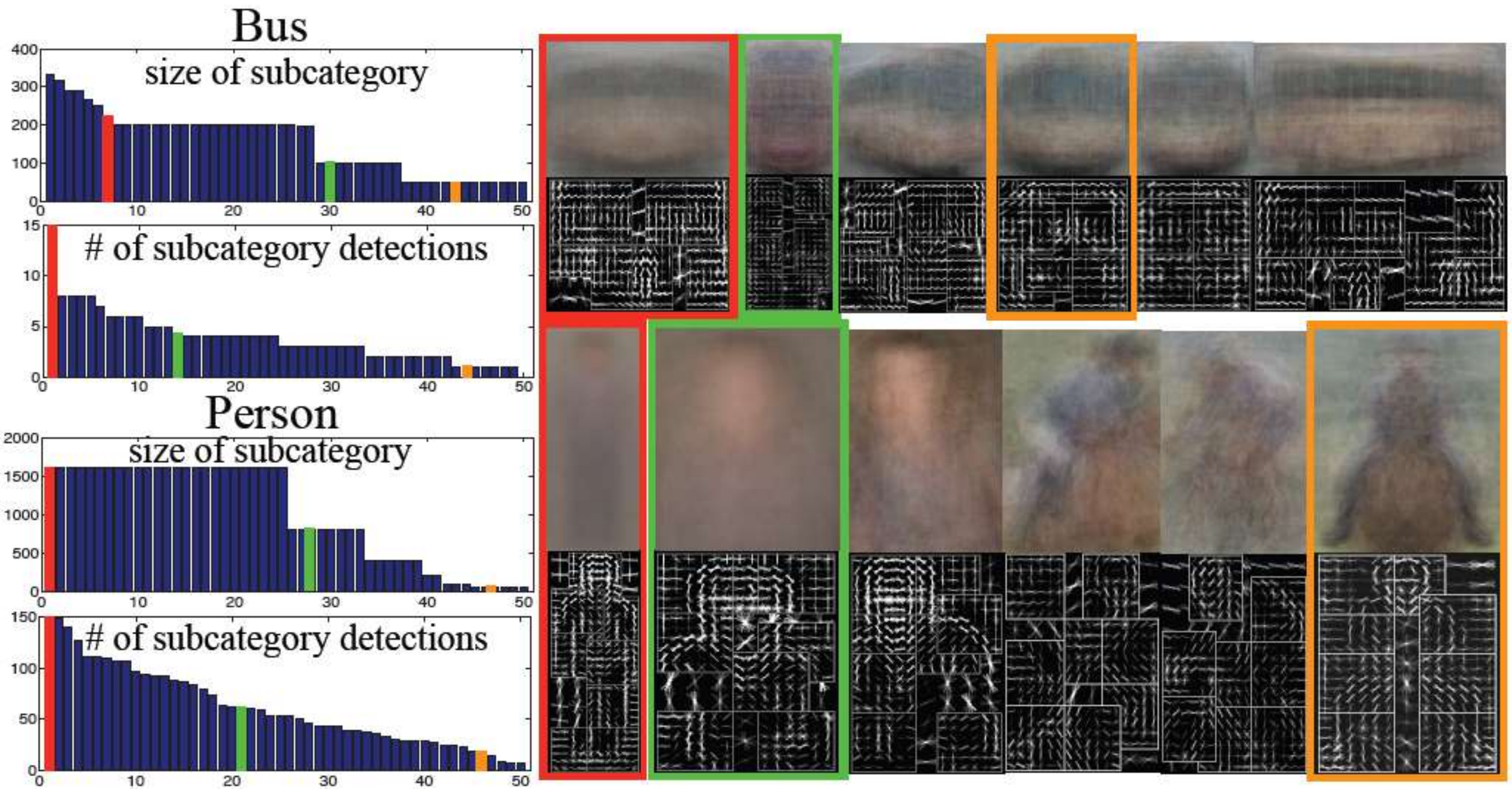
* S. Divvala, A. Effros, M. Hebert **"Object Instance Sharing by Enhanced Bounding Box Correspondence"** , *BMVC'*12.

Long-tail distribution



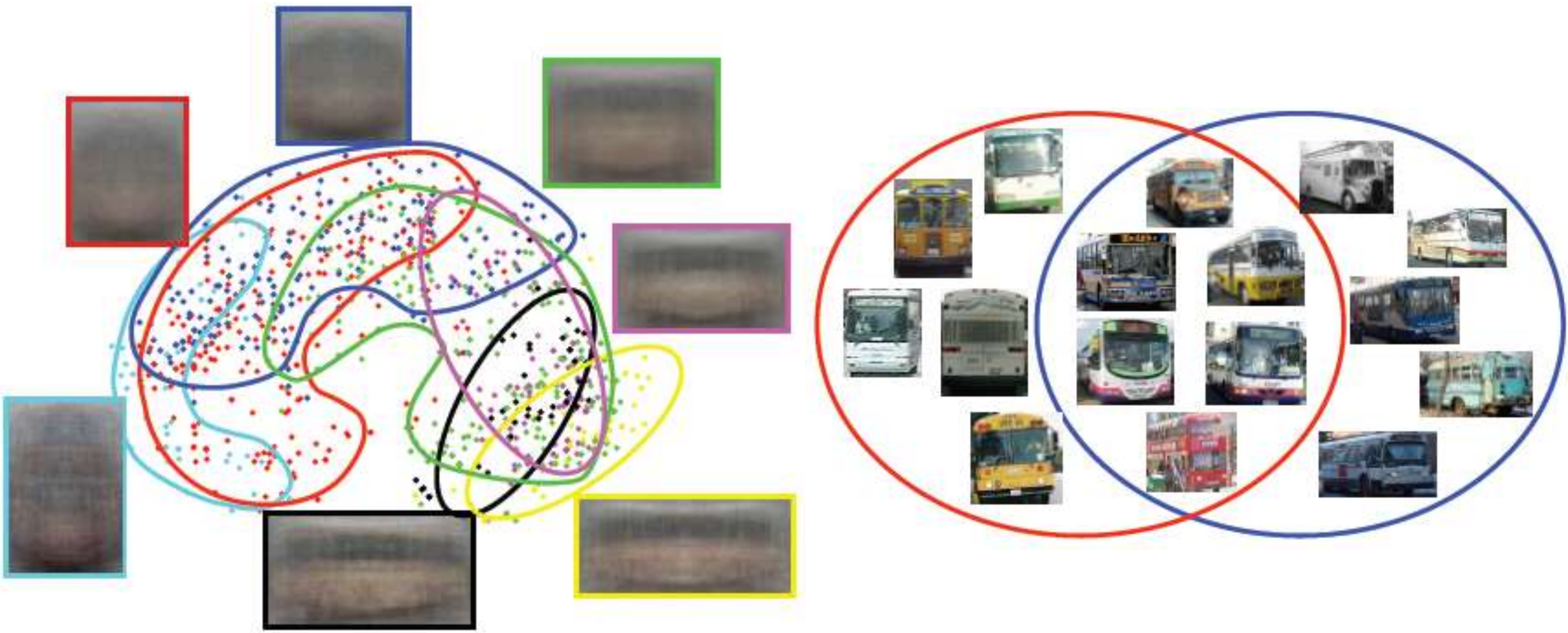
We need lots of templates, have little data of *'yoga twist'* poses

Long-tail distribution



* X. Zhu, D. Ramanan, "**Capturing long-tail distribution of object subcategories**", In *CVPR* 2014.

Instance sharing across subcategories



* X. Zhu, D. Ramanan, "**Capturing long-tail distribution of object subcategories**", In *CVPR* 2014.

Web-supervision

Web-supervision

- **GOAL:** Use Internet contents instead of explicit **human supervision.**

- Use internet contents (texts/images) for :
  - *Subcategory discovery:*
    - leveraging vast resources of **online books** to discover the vocabulary of variance.
  - *Enrichment of poor subcategories:*
    - Using gigantic amount of **unlabeled images** on Internet.

* S. Divvala, "**Learning everything about anything**", In *CVPR* 2014.

The PASCAL Visual Object Classes (VOC) Challenge

Mark Everingham · Luc Van Gool ·
Christopher K. I. Williams · John Winn ·
Andrew Zisserman

Received: 30 July 2008 / Accepted: 16 July 2009 / Published online: 9 September 2009

Abstract The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. Organised annually from 2005 to present, the challenge and its associated dataset has become accepted as *the* benchmark for object detection.

This paper describes the dataset and evaluation procedure. We review the state-of-the-art in evaluated methods for both classification and detection, analyse whether the methods are statistically different, what they are learning from the images (e.g. the object or its context), and what the methods find easy or confuse. The paper concludes with lessons learnt in the three year history of the challenge, and proposes directions for future improvement and extension.

Keywords Database · Benchmark · Object recognition · Object detection

M. Everingham (✉)
University of Leeds, Leeds, UK
e-mail: m.everingham@leeds.ac.uk

L. Van Gool
KU Leuven, Leuven, Belgium

C.K.I. Williams
University of Edinburgh, Edinburgh, UK

J. Winn
Microsoft Research, Cambridge, UK

A. Zisserman
University of Oxford, Oxford, UK

1 Introduction

The PASCAL[1] Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available *dataset* of images and annotation, together with standardised evaluation software; and (ii) an annual *competition* and workshop. The VOC2007 dataset consists of annotated consumer photographs collected from the flickr[2] photo-sharing web-site. A new dataset with ground truth annotation has been released each year since 2006. There are two principal challenges: *classification*—"does the image contain any instances of a particular object class?" (where the object classes include cars, people, dogs, etc.), and *detection*—"where are the instances of a particular object class in the image (if any)?". In addition, there are two subsidiary challenges ("tasters") on pixel-level segmentation—assign each pixel a class label, and "person layout"—localise the head, hands and feet of people in the image. The challenges are issued with deadlines each year, and a workshop held to compare and discuss that year's results and methods. The datasets and associated annotation and software are subsequently released and available for use at any time.

The objectives of the VOC challenge are twofold: first to provide challenging images and high quality annotation, together with a standard evaluation methodology—a "plug and play" training and testing harness so that performance of algorithms can be compared (the dataset component); and second to measure the state of the art each year (the competition component).

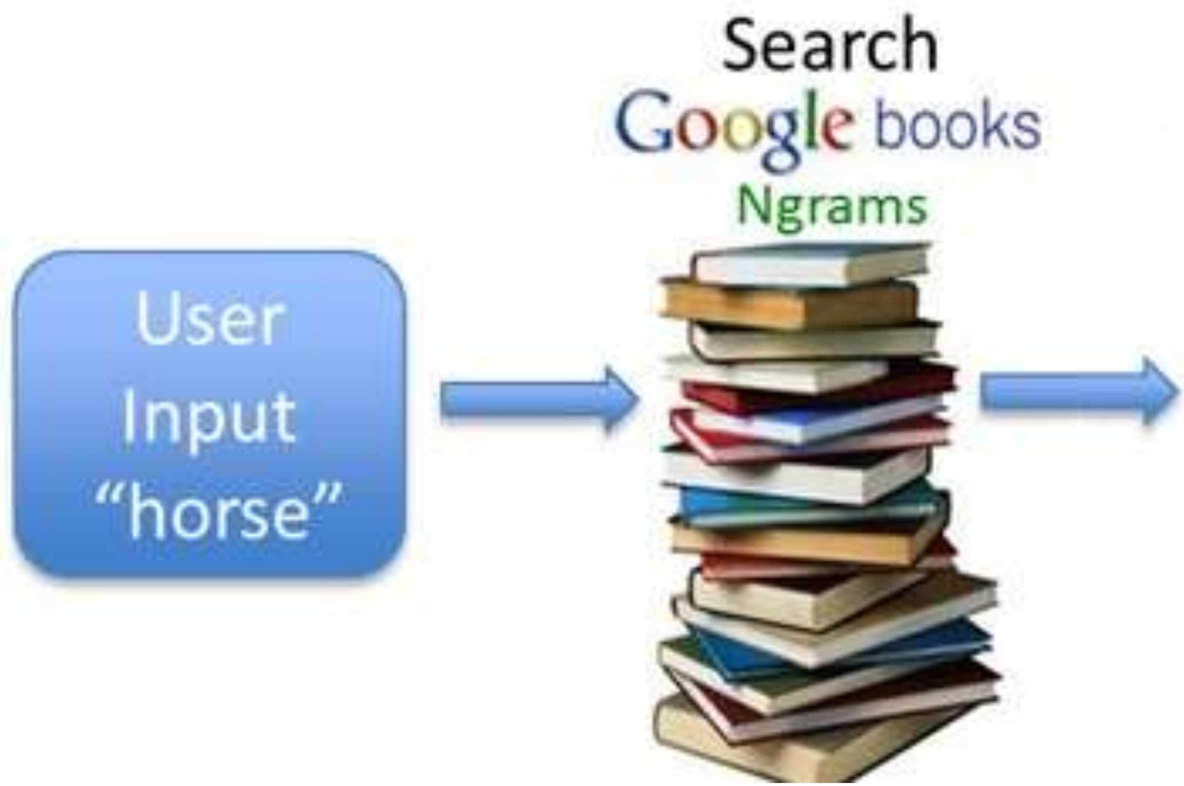
[1] PASCAL stands for pattern analysis, statistical modelling and computational learning. It is an EU Network of Excellence funded under the IST Programme of the European Union.

[2] http://www.flickr.com/

Table 1 Queries used to retrieve images from flickr. Words in bold show the "targeted" class. Note that the query terms are quite general—including the class name, synonyms and scenes or situations where the class is likely to occur

- **horse**, gallop, jump, buck, equine, foal, cavalry, saddle, canter, buggy, mare, neigh, dressage, trial, racehorse, steeplechase, thoroughbred, cart, equestrian, paddock, stable, farrier
- **motorbike**, motorcycle, minibike, moped, dirt, pillion, biker, trials, motorcycling, motorcyclist, engine, motocross, scramble, sidecar, scooter, trail
- **person**, people, family, father, mother, brother, sister, aunt, uncle, grandmother, grandma, grandfather, grandpa, grandson, granddaughter, niece, nephew, cousin
- **sheep**, ram, fold, fleece, shear, baa, bleat, lamb, ewe, wool, flock
- **sofa**, chesterfield, settee, divan, couch, bolster
- **table**, dining, cafe, restaurant, kitchen, banquet, party, meal
- **potted plant**, pot plant, plant, patio, windowsill, window sill, yard, greenhouse, glass house, basket, cutting, pot, cooking, grow
- **train**, express, locomotive, freight, commuter, platform, subway, underground, steam, railway, railroad, rail, tube, underground, track, carriage, coach, metro, sleeper, railcar, buffet, cabin, level crossing
- **tv/monitor**, television, plasma, flatscreen, flat screen, lcd, crt, watching, dvd, desktop, computer, computer monitor, PC, console, game

Query expansion



- horse
- grazing horse
- jumping horse
- rolling horse
- reining horse
- sledge horse
- fighting horse
- crazy horse
- horse ears
- last horse
- sleigh horse
- eating horse
- horse head
- …

+20K variations

- horse
- grazing horse
- jumping horse
- rolling horse
- reining horse
- sledge horse
- fighting horse
- crazy horse
- horse ears
- last horse
- sleigh horse
- eating horse
- horse head
- …

+20K variations

- horse
- grazing horse
- jumping horse
- rolling horse
- reining horse
- sledge horse
- fighting horse
- crazy horse
- horse ears
- last horse
- sleigh horse
- eating horse
- horse head
- …

+20K variations

Pruning non visual ngrams

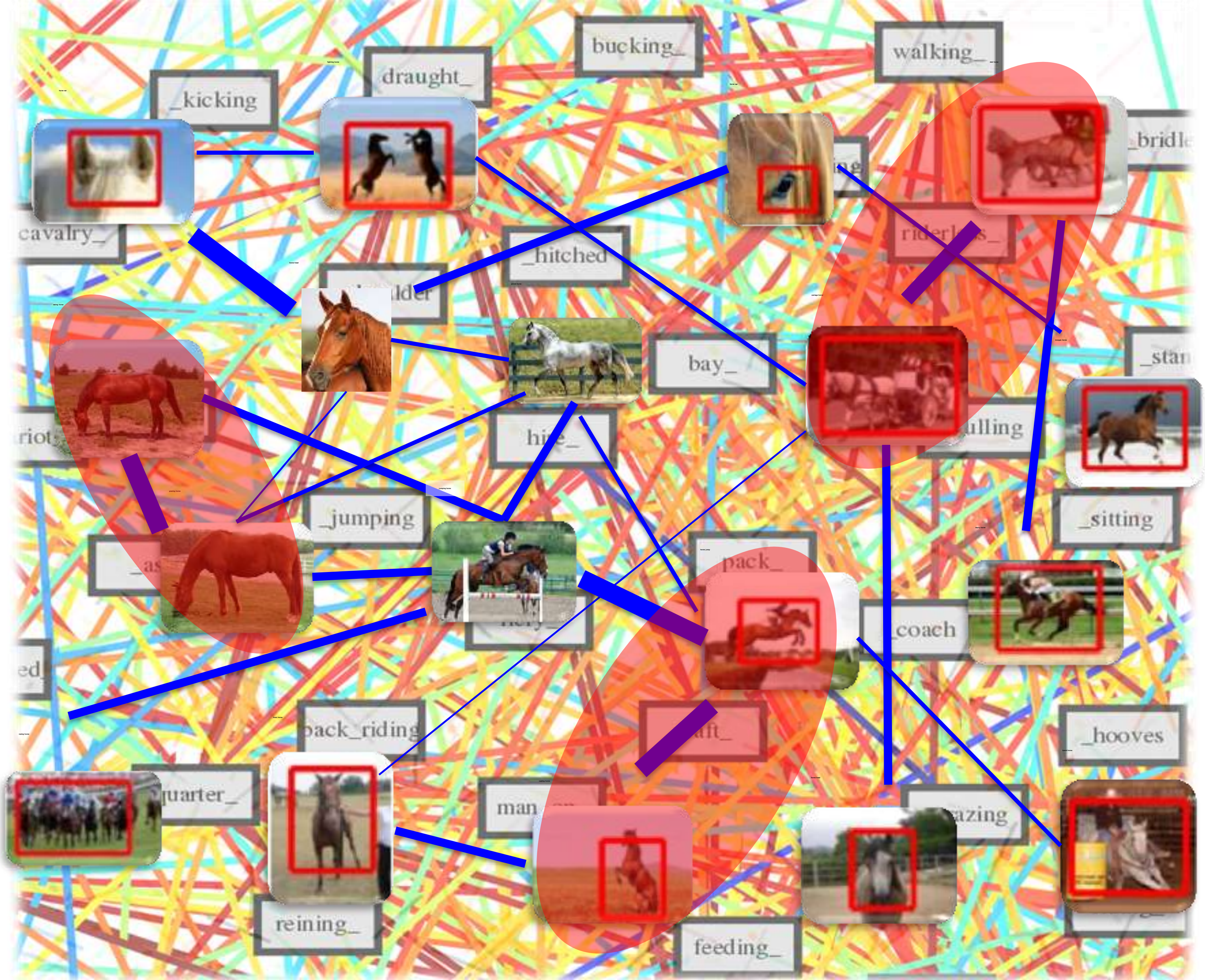
- horse
- grazing horse
- jumping horse
- rolling horse
- reining horse
- sledge horse
- fighting horse
- crazy horse
- horse ears
- last horse
- sleigh horse
- eating horse
- horse head
- …

+20K variations

Identifying Visual Synonyms

- horse
- grazing horse
- jumping horse
- rolling horse
- reining horse
- sledge horse
- fighting horse
- ~~crazy horse~~
- horse ears
- ~~last horse~~
- sleigh horse
- eating horse
- horse head
- …

+20K variations



bucking_
walking_
draught_
_kicking
_bridle
cavalry_
_hitched
bay_
_stan
_jumping
_sitting
pack_
_coach
back_riding
_hooves
quarter_
reining_
feeding_

Taming Intra-class Variance

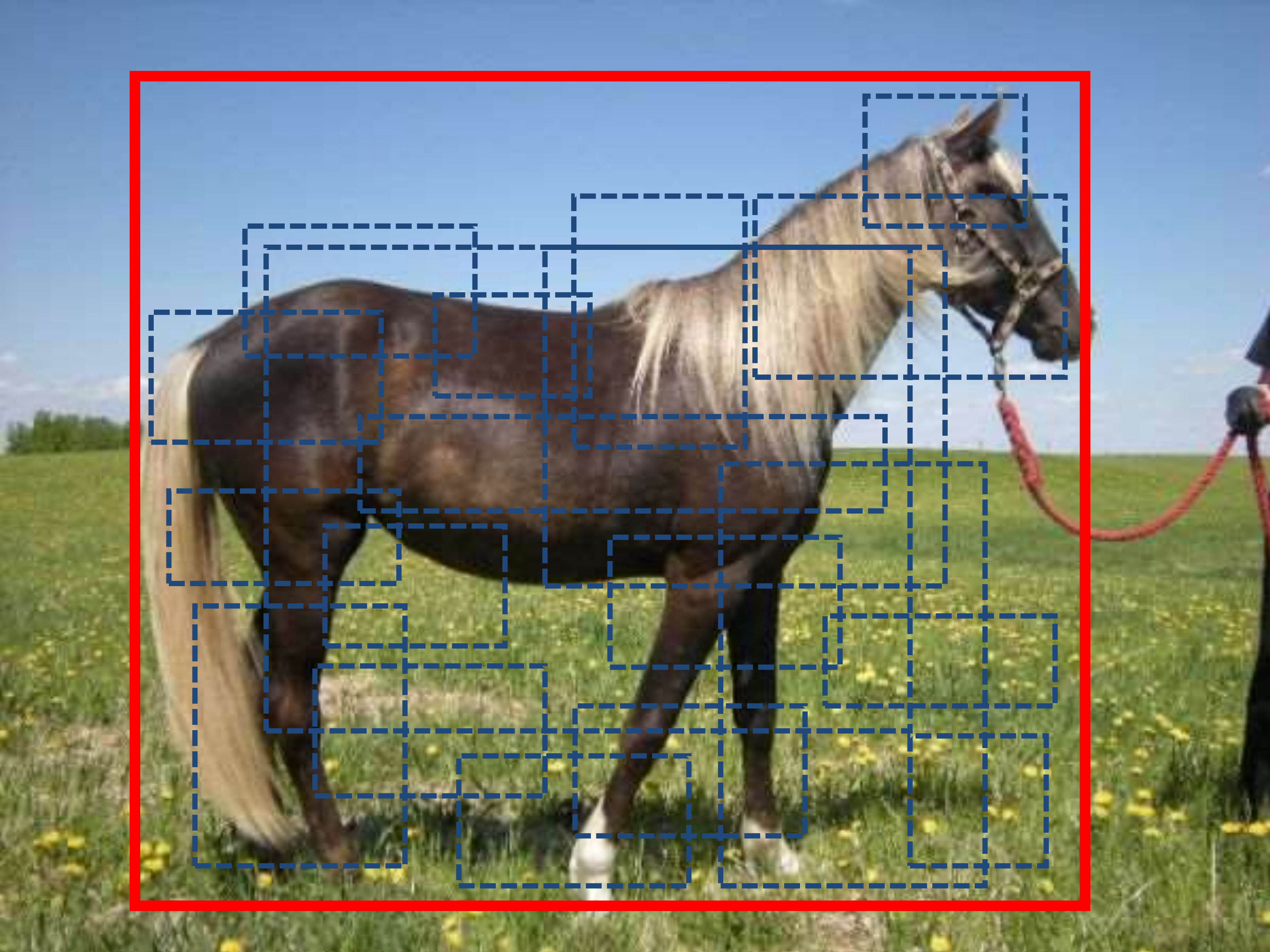
M. Pandey and S. Lazebnik, **"Weakly Supervised Object Localization with DPM"**, ICCV 2011.

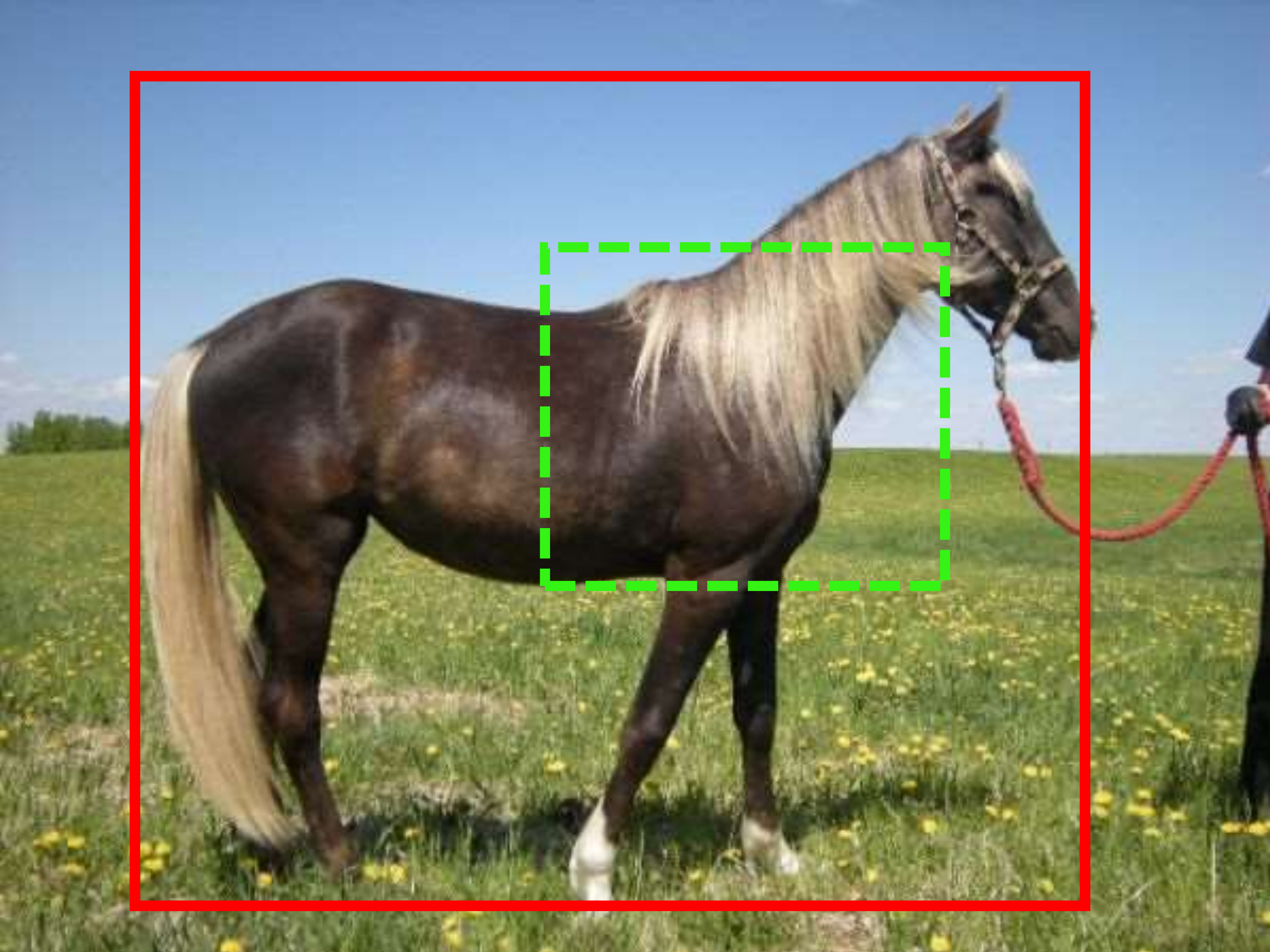

IMAGENET
Equine, equid
Horse, Equus caballus
Saddle
Racehorse
Pony
Male
Pony
Harness
Stalking-hor
Eohippus
Stepper
Pacer
Steeplechaser
Hack
Polo
Workhorse
Wild
Liver
Sorrel
Bay
Stablemate
Roan
Hack
Chestnut
Mare
Gee-gee
Pinto
Post
Palomino
NATIONAL

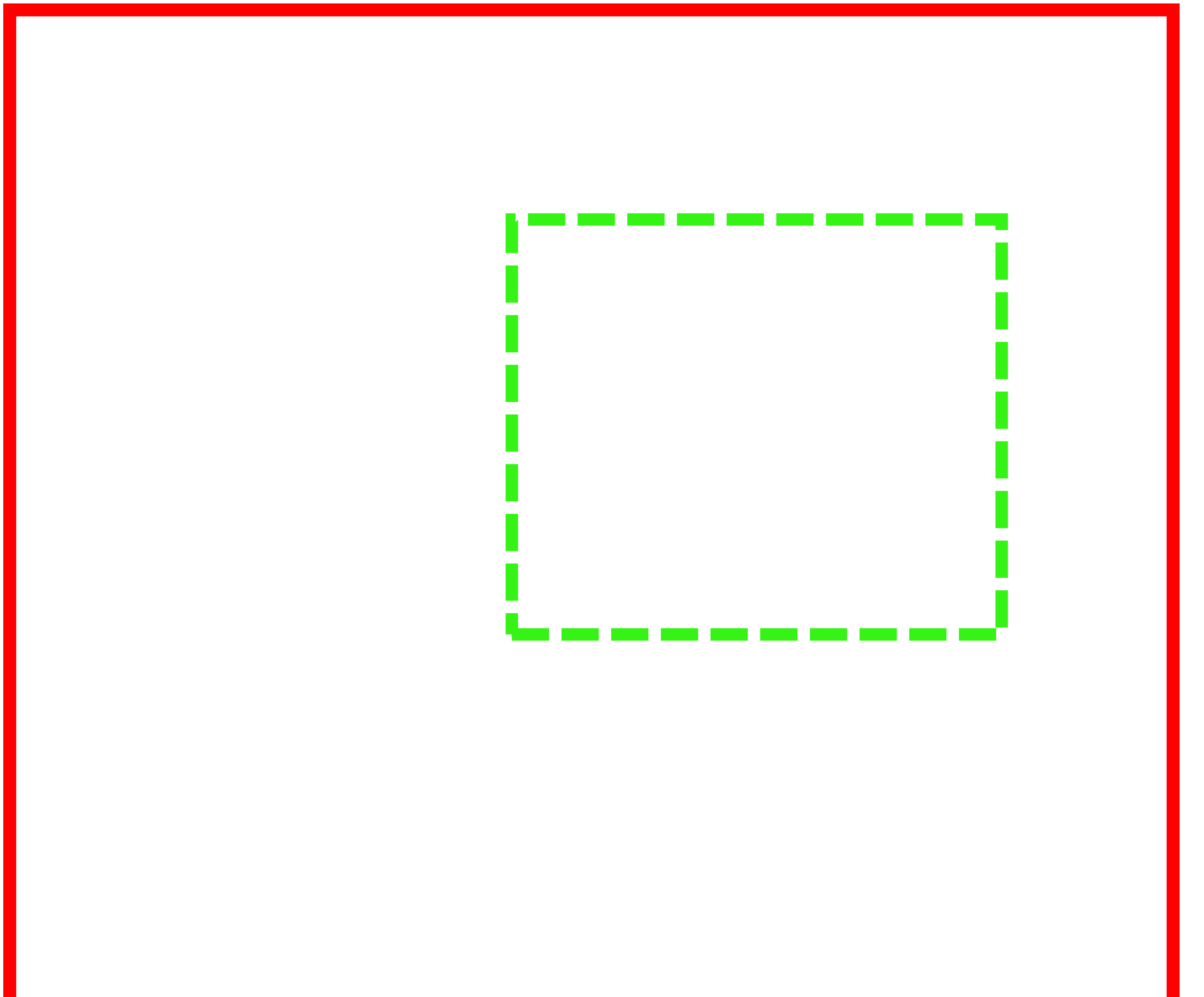
Swimming Horse
Rolling Horse
Rearing Horse
Fighting Horse
Bucking Horse
Reining Horse
Barrel Horse
Horse Tram
Horse Eye
Front Horse
Bridled Horse
Jennet horse

Webly-supervised discriminative patch

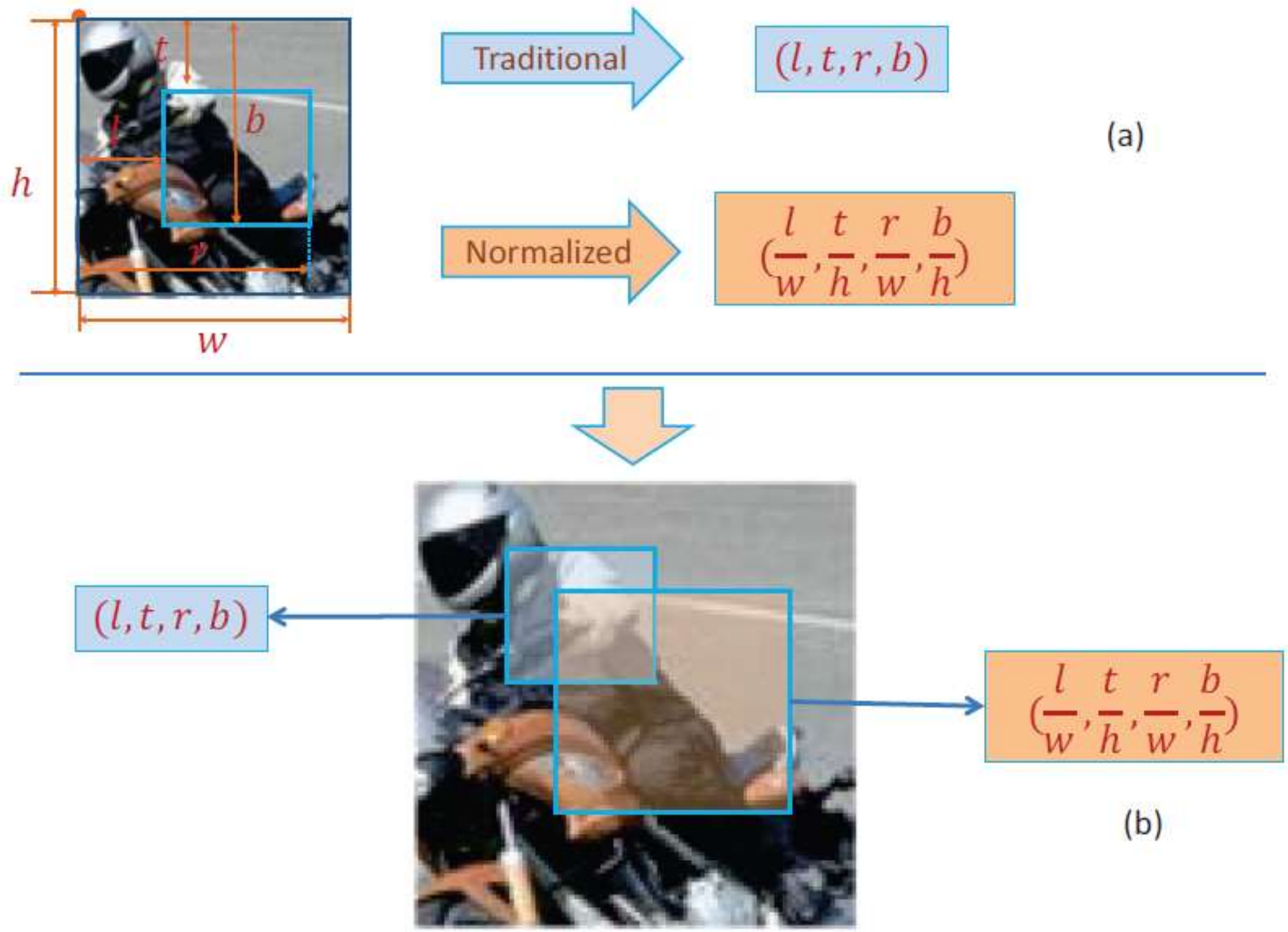
- Can we go **inside the box** and find the discriminative patch?
- **Subcategory-aware** discriminative patches.
- **Fixed-position** patches.



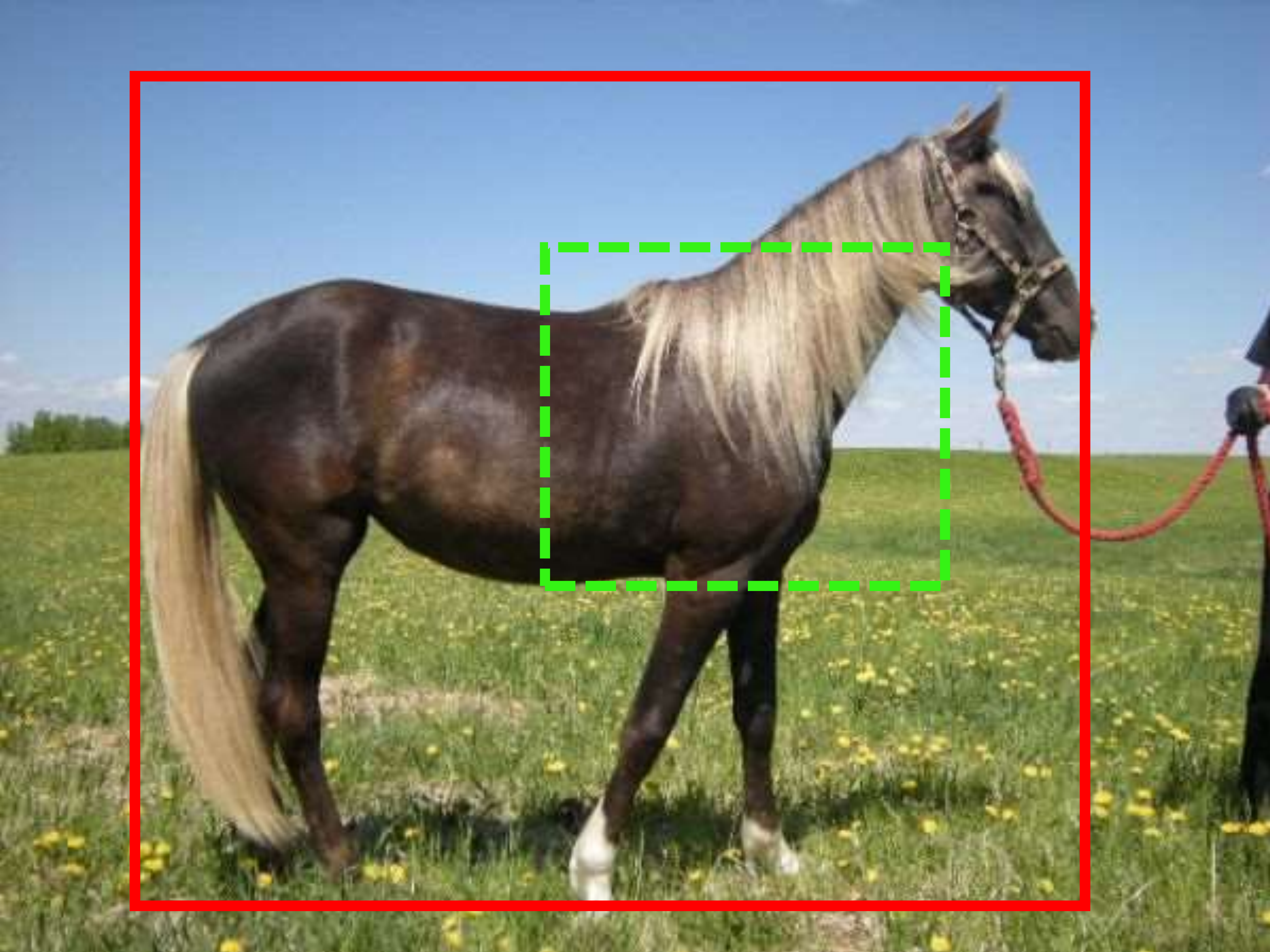




Relative Normalized Position



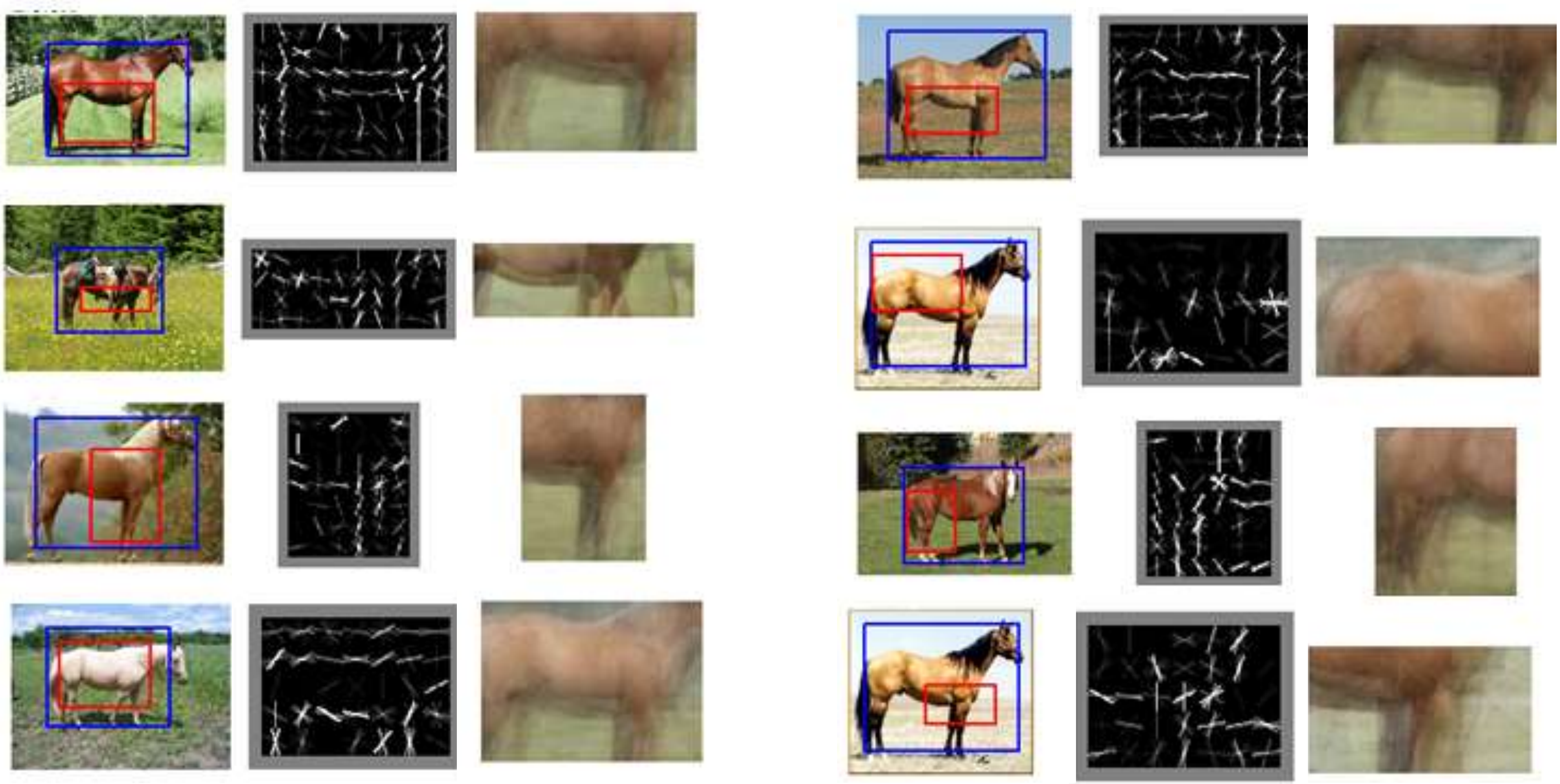
Reference: X. Wang, M. Yang, S. Zhu, Y. Lin, Regionlets for Generic Object Detection, ICCV 2013.



Train Exemplar-SVM

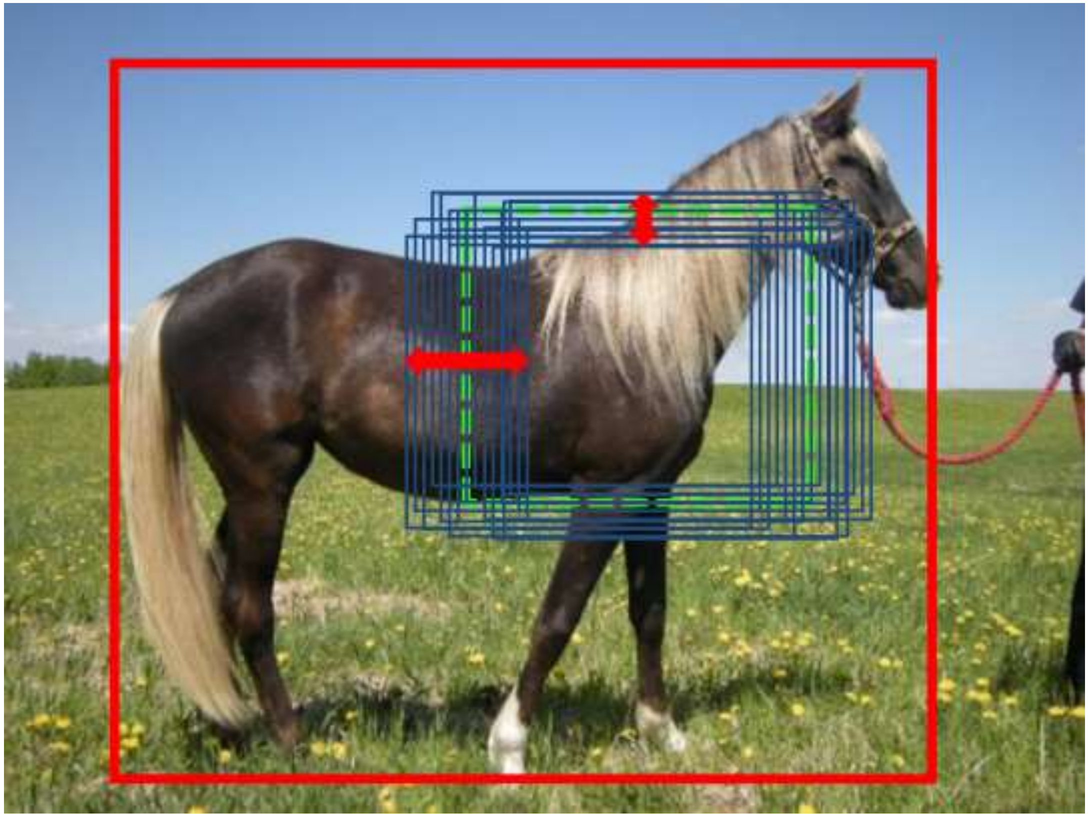
Hariharan et.al. **"Discriminative Decorrelation for Clustering and Classification"**, ECCV 2012.

Initial Patch Models



Patch Deformability

- Patch should NOT be fully fixed-position
  - Use NMS to find deformation of the patch

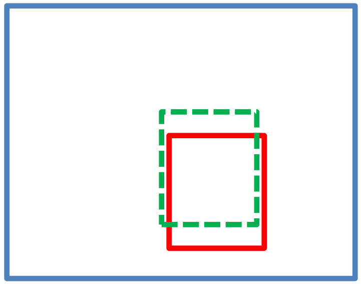


Patch Selection

- What are good patch?
  - Appearance consistency score
    - Repetitive visual pattern
    - Confidence score of E-LDA patch model

  - Spatial consistency score
    - Spatially consistent
    - Patch activation

$$Activation(a, b) = \frac{(a \cap b)}{(a \cup b)}$$



Patch Selection

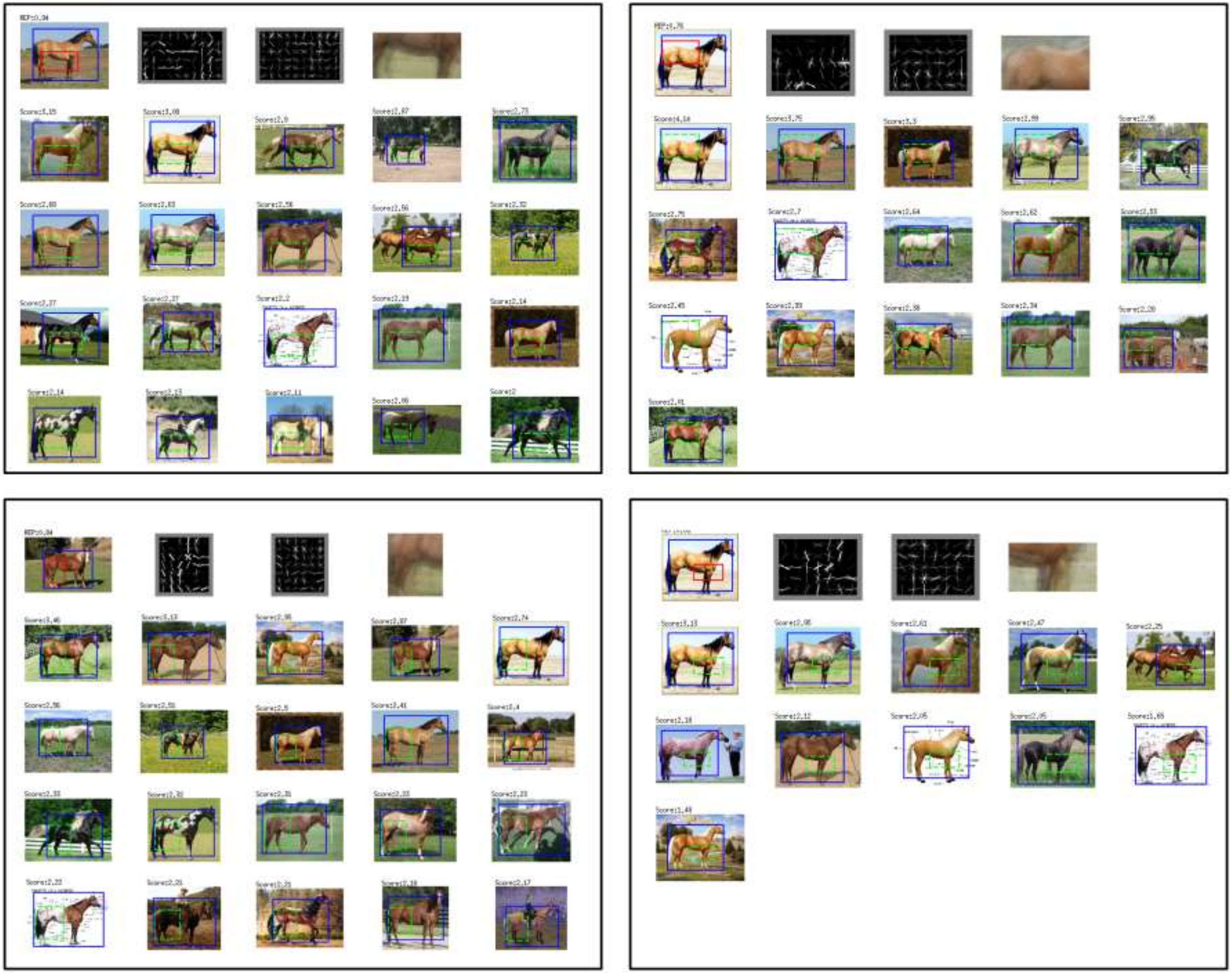
- Representativeness criteria
  - Intra-subcategory consistency

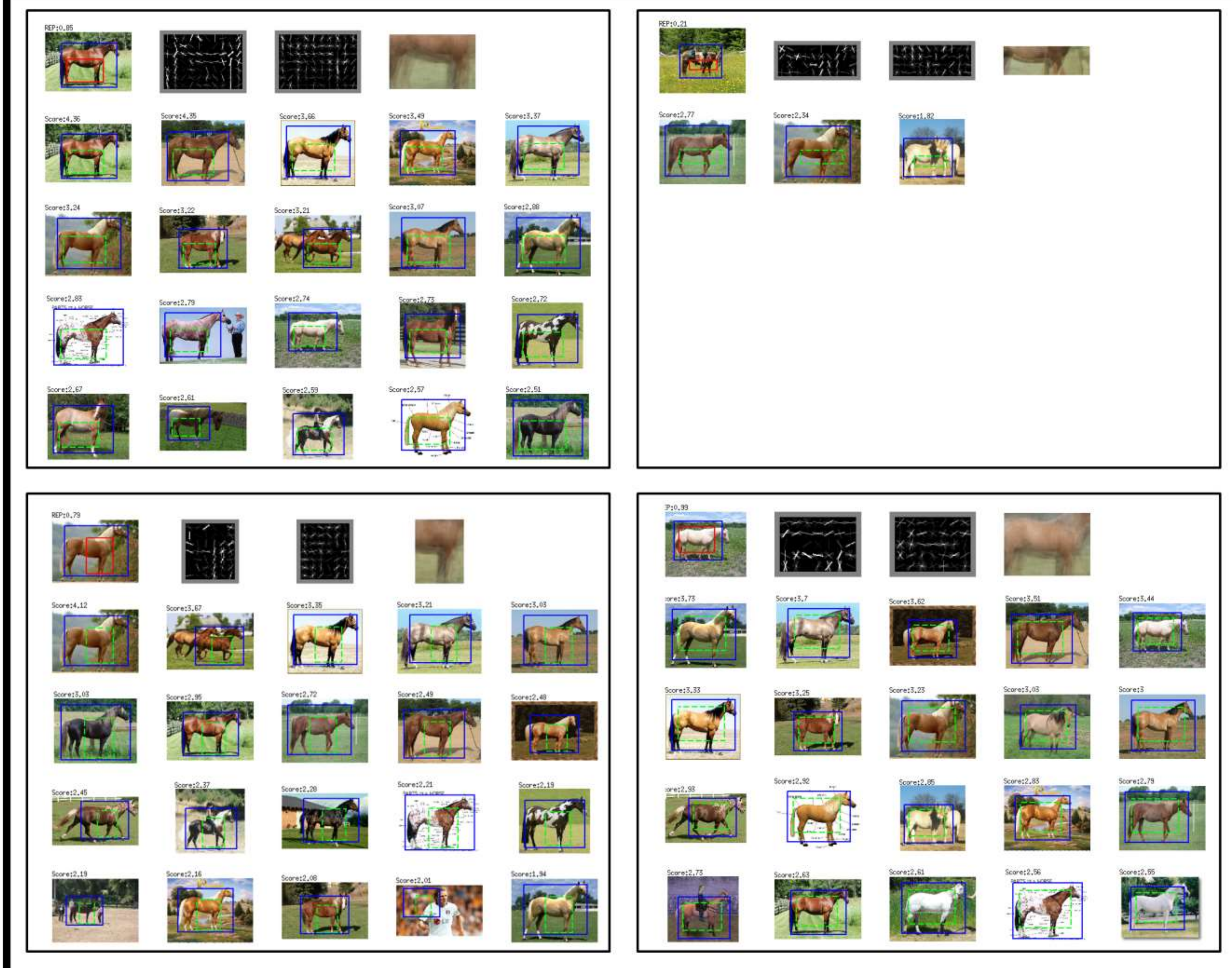
$$rep(p, I_s) = \frac{1}{|I_s|} \sum_{i=1}^{|I_s|} score(x_i, p)$$

- Discrimination criteria
  - Inter-Category discriminativity
  - Normalized median rank on a mixed set of subcategory images and a huge PASCAL negative set.

$$disc(p, I_s, \bar{I}) = \frac{median(rank(p, I_s, \bar{I}))}{|I_s|}$$

$$rank(p, I_s, \bar{I}) : \mathbb{R}^{|I_s \cup \bar{I}|} \longmapsto \mathbb{N}^{|I_s|}$$

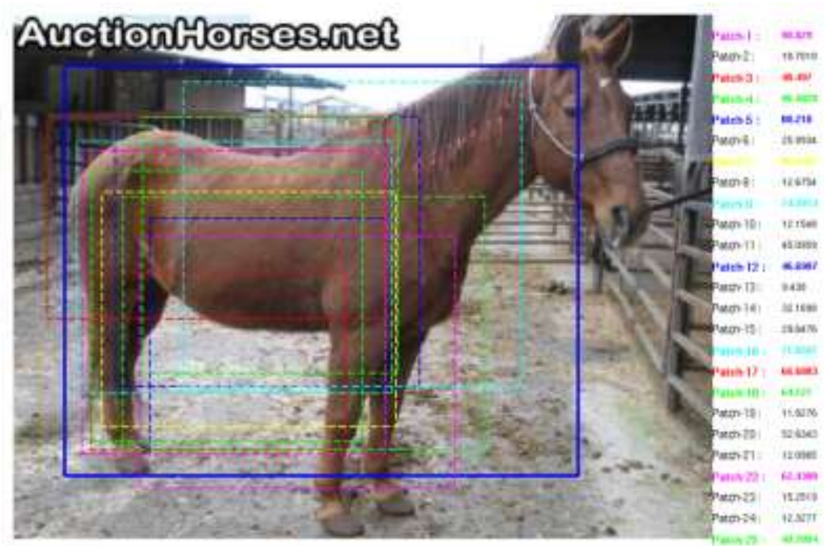


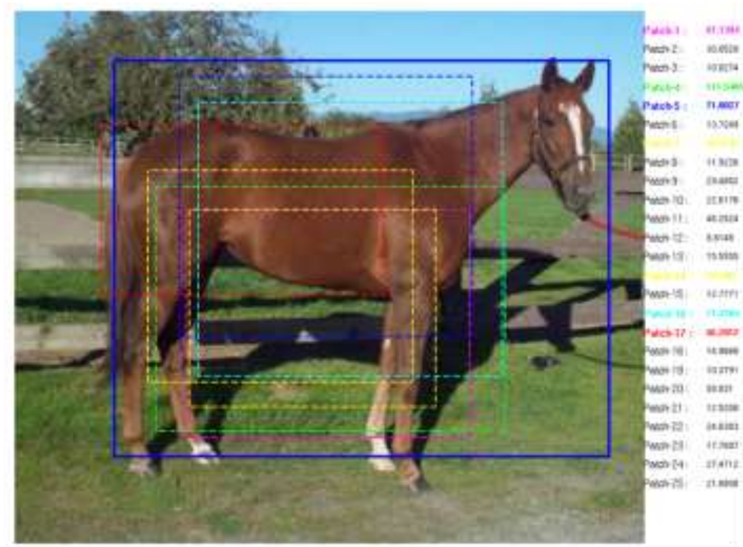

REP:0.85
Score:4.36
Score:4.35
Score:3.66
Score:3.49
Score:3.37
Score:3.24
Score:3.22
Score:3.21
Score:3.07
Score:2.88
Score:2.83
Score:2.79
Score:2.74
Score:2.73
Score:2.72
Score:2.67
Score:2.61
Score:2.59
Score:2.57
Score:2.51
REP:0.21
Score:2.77
Score:2.34
Score:1.82
REP:0.79
Score:4.12
Score:3.67
Score:3.35
Score:3.21
Score:3.03
Score:3.03
Score:2.95
Score:2.72
Score:2.49
Score:2.48
Score:2.45
Score:2.37
Score:2.28
Score:2.21
Score:2.19
Score:2.19
Score:2.16
Score:2.08
Score:2.01
Score:1.94
P:0.93
ore:3.73
Score:3.7
Score:3.62
Score:3.51
Score:3.44
Score:3.33
Score:3.25
Score:3.23
Score:3.03
Score:3
ore:2.93
Score:2.92
Score:2.85
Score:2.83
Score:2.79
Score:2.73
Score:2.63
Score:2.61
Score:2.56
Score:2.55

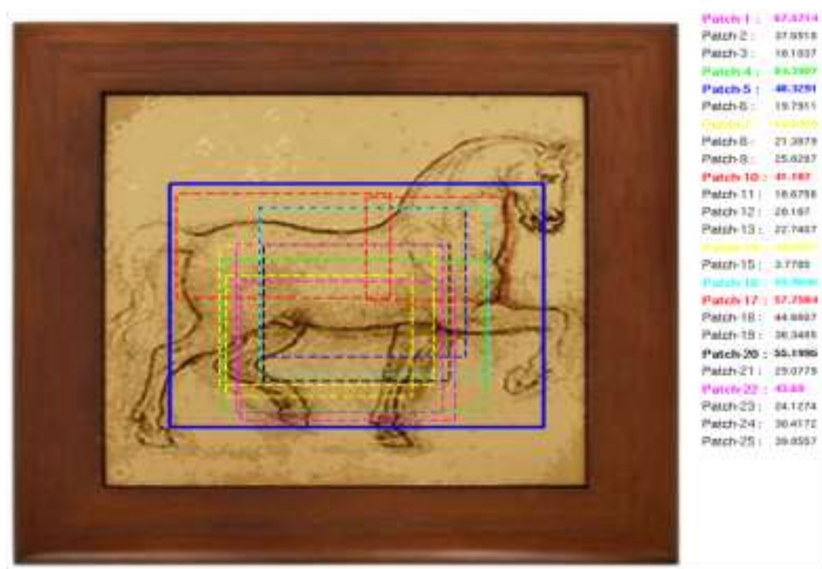
Patch re-training

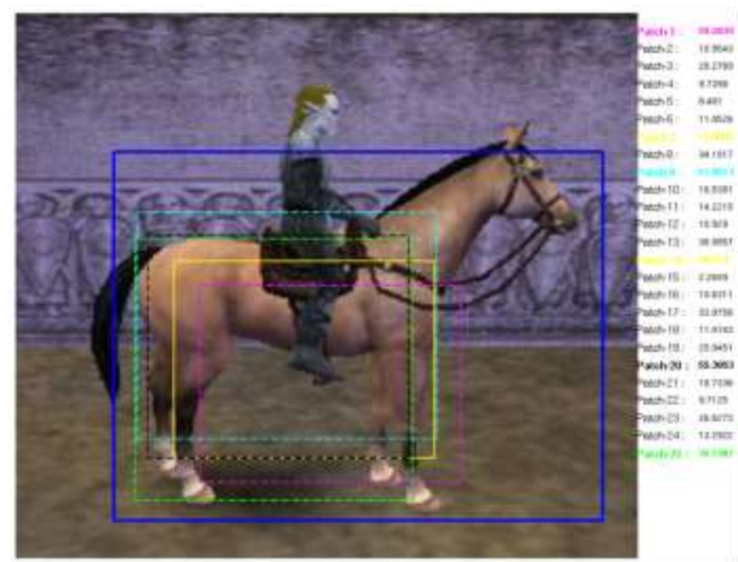
- **Why?**
  - E-LDA are **shallow models**
  - LDA and SVM models are **highly correlated**
- **Patch retraining:**
  - Train initial patch models with LDA
  - Patch selection
  - Example selection
  - Then re-train the expensive models only for the selected Patches using Latent-LDA

Select good example

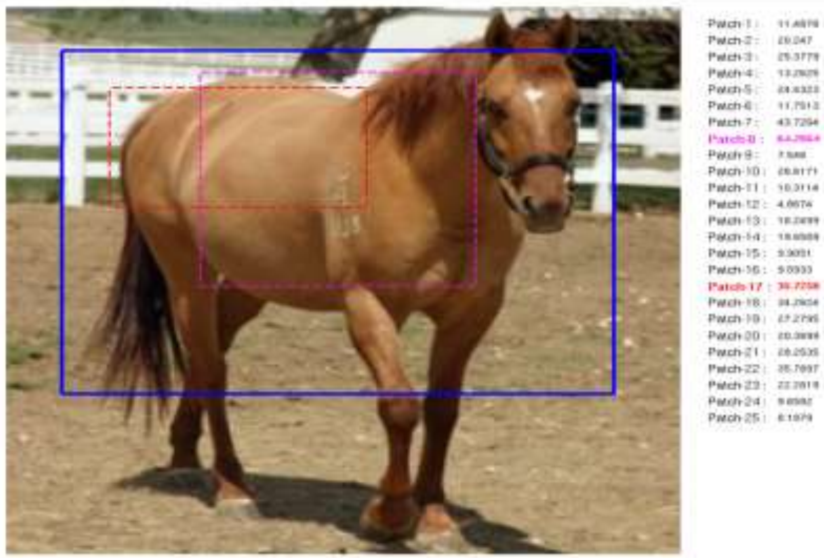


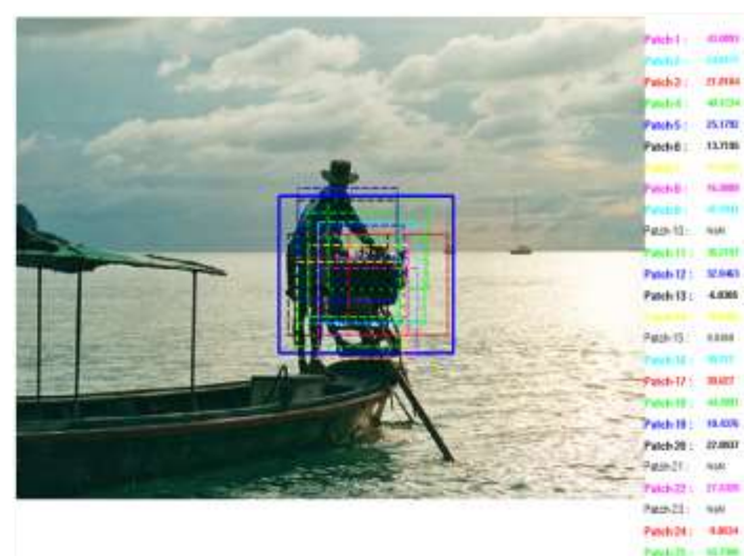




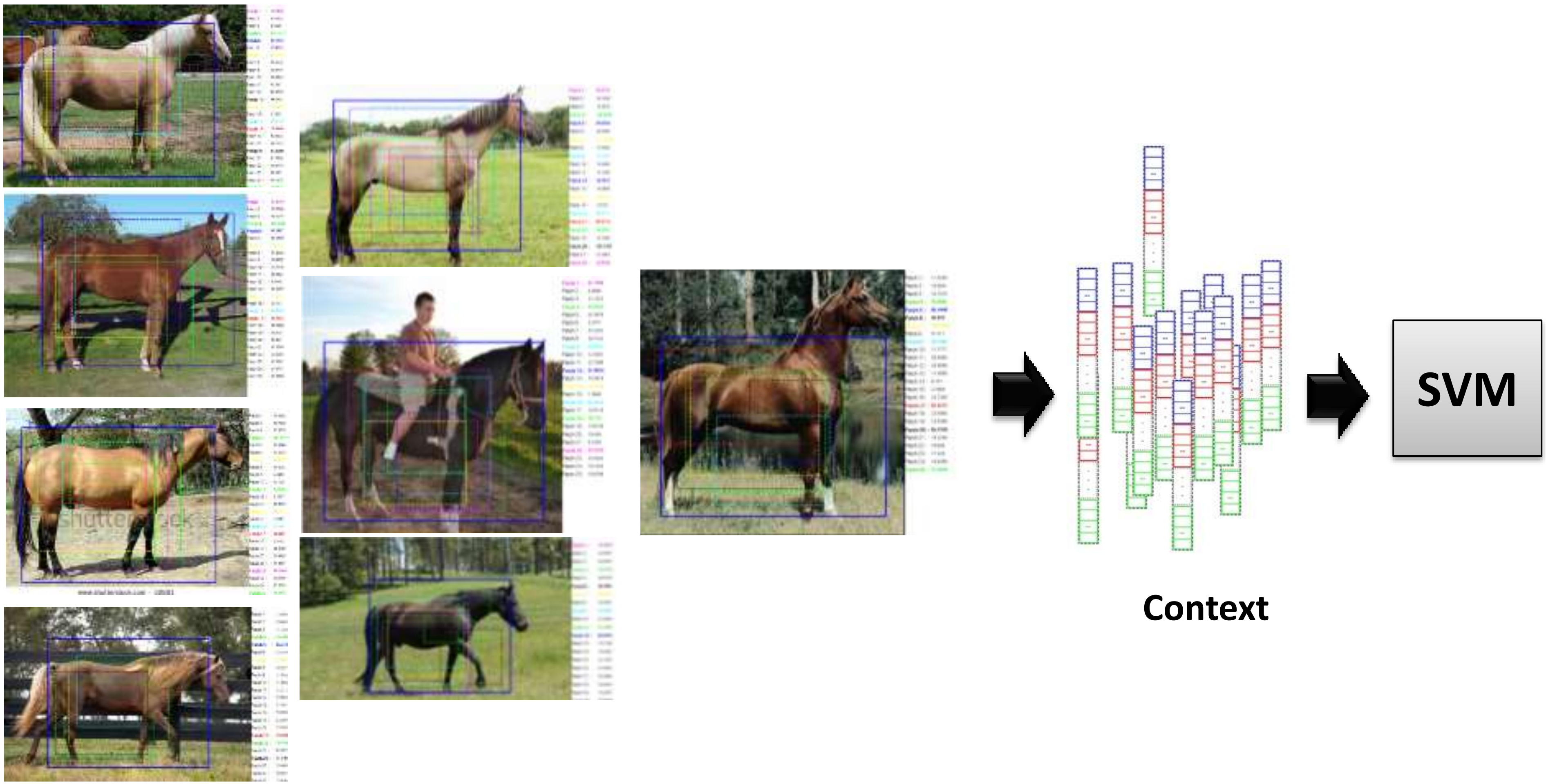


Prune noisy images





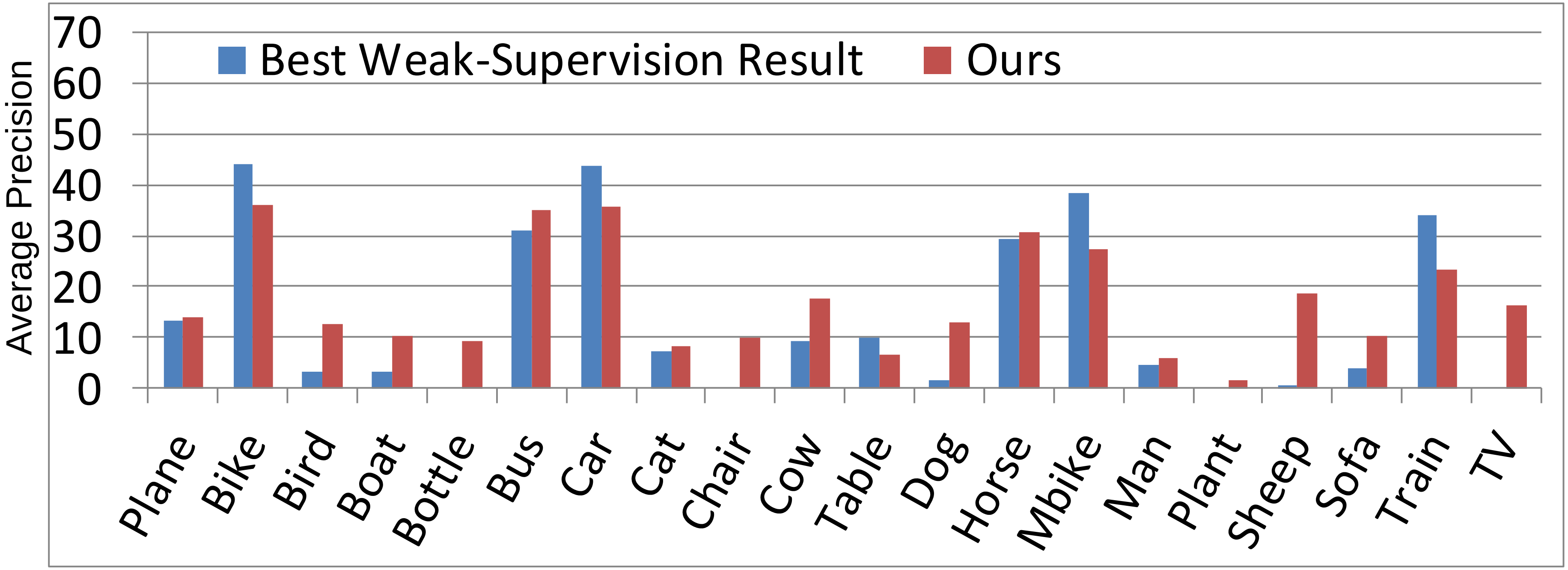
Patch Calibration



Patch-based detection

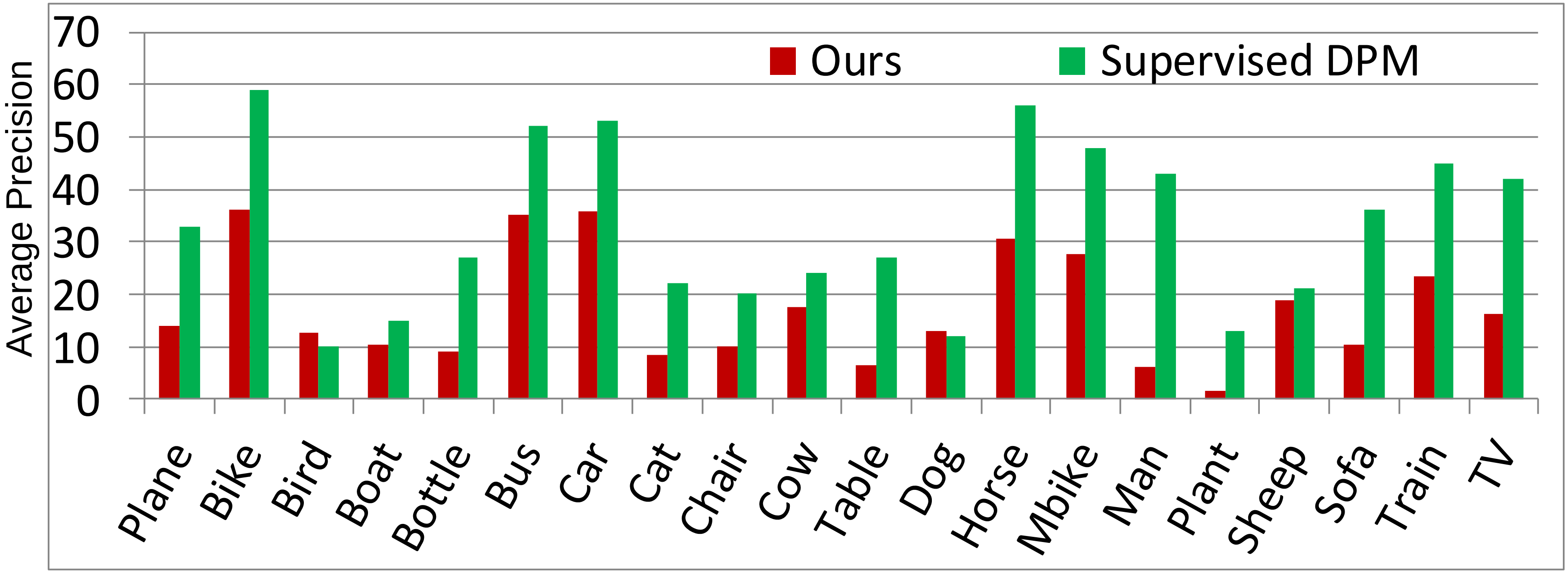
Results

PASCAL VOC 2007 Object Detection

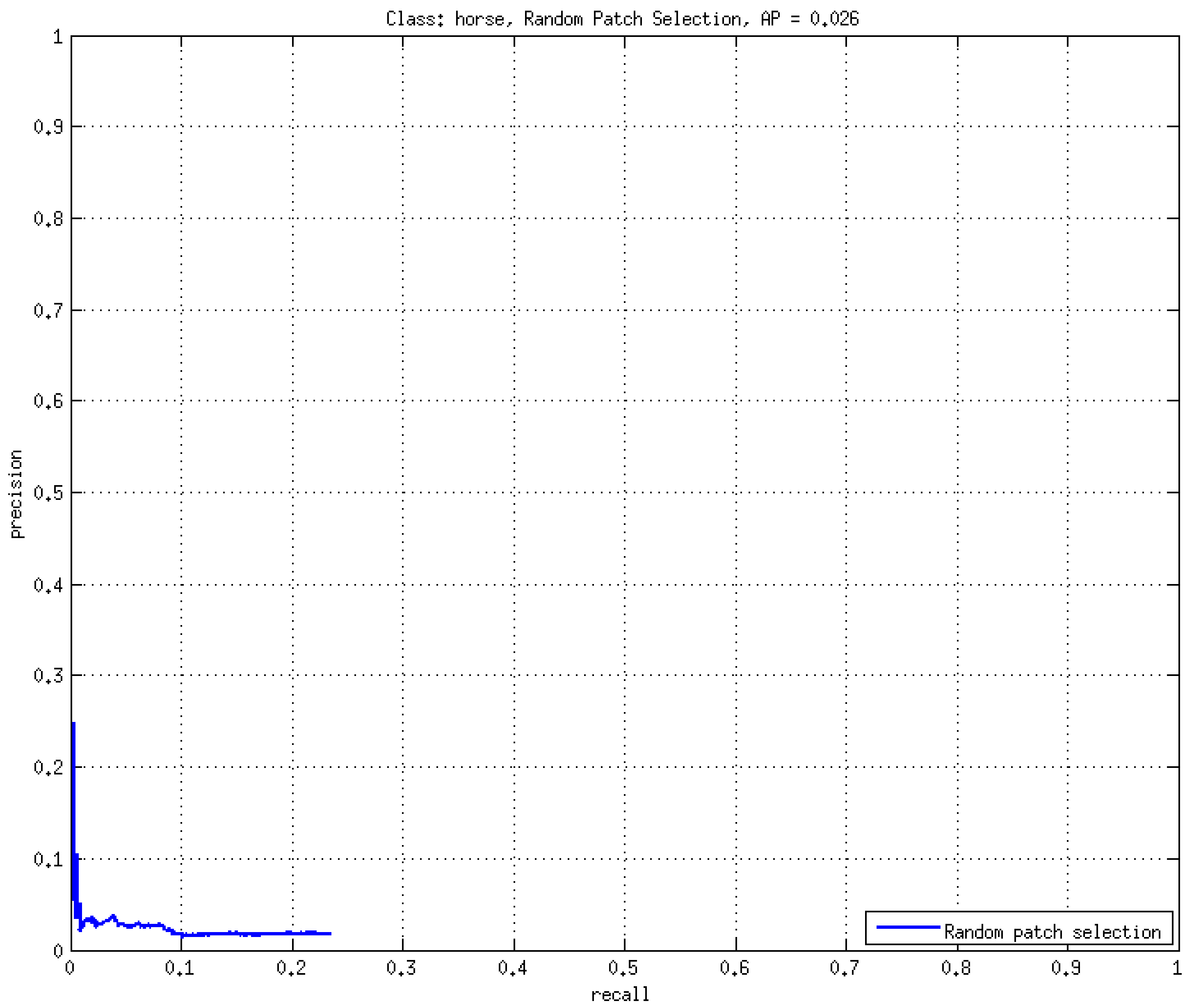


- Holistic model beats the previous best on 13 classes
- Previous best [Siva_ICCV11, Prest_CVPR12] uses weak human-supervision i.e., image/video labels, and Objectness

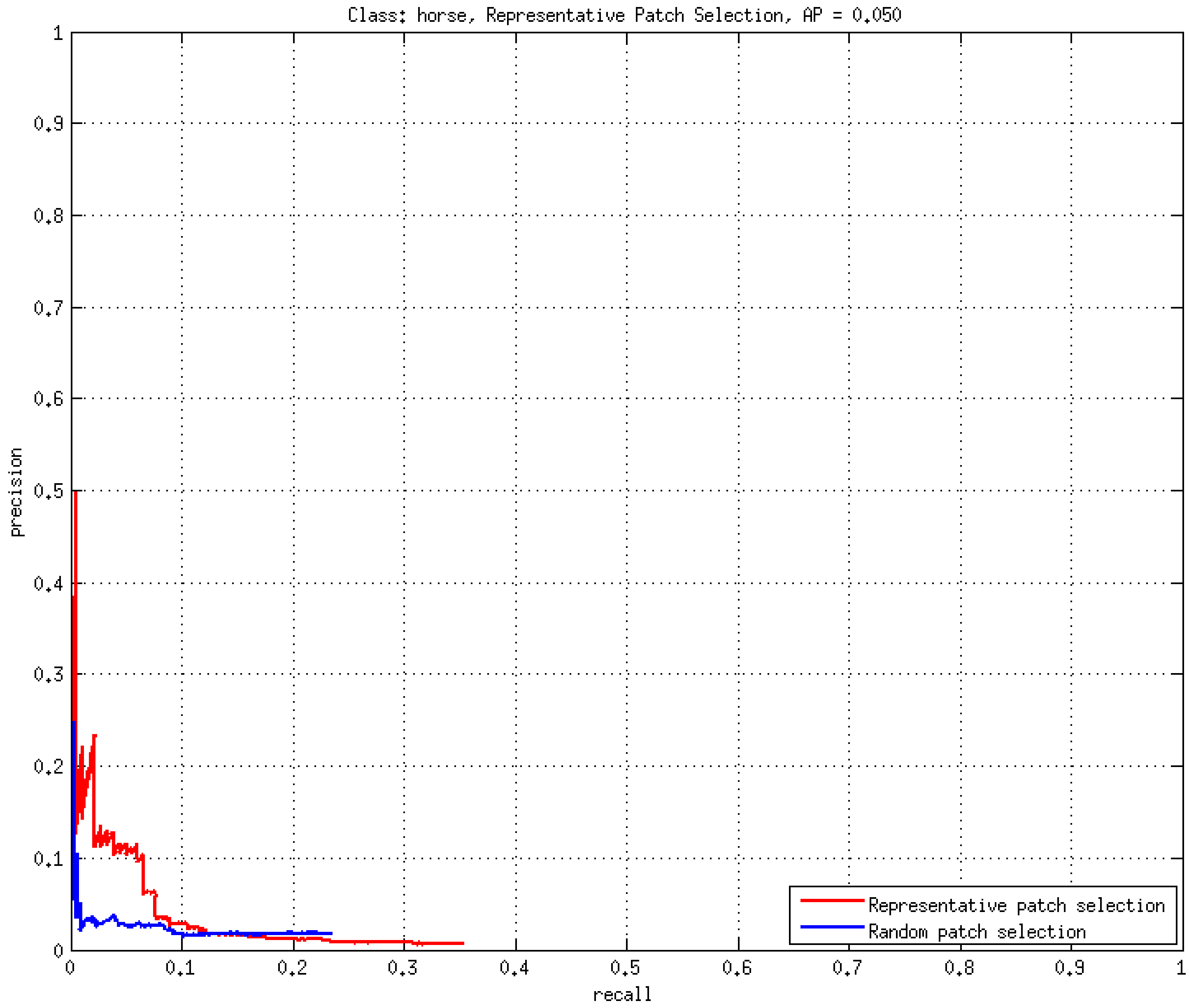
PASCAL VOC 2007 Object Detection



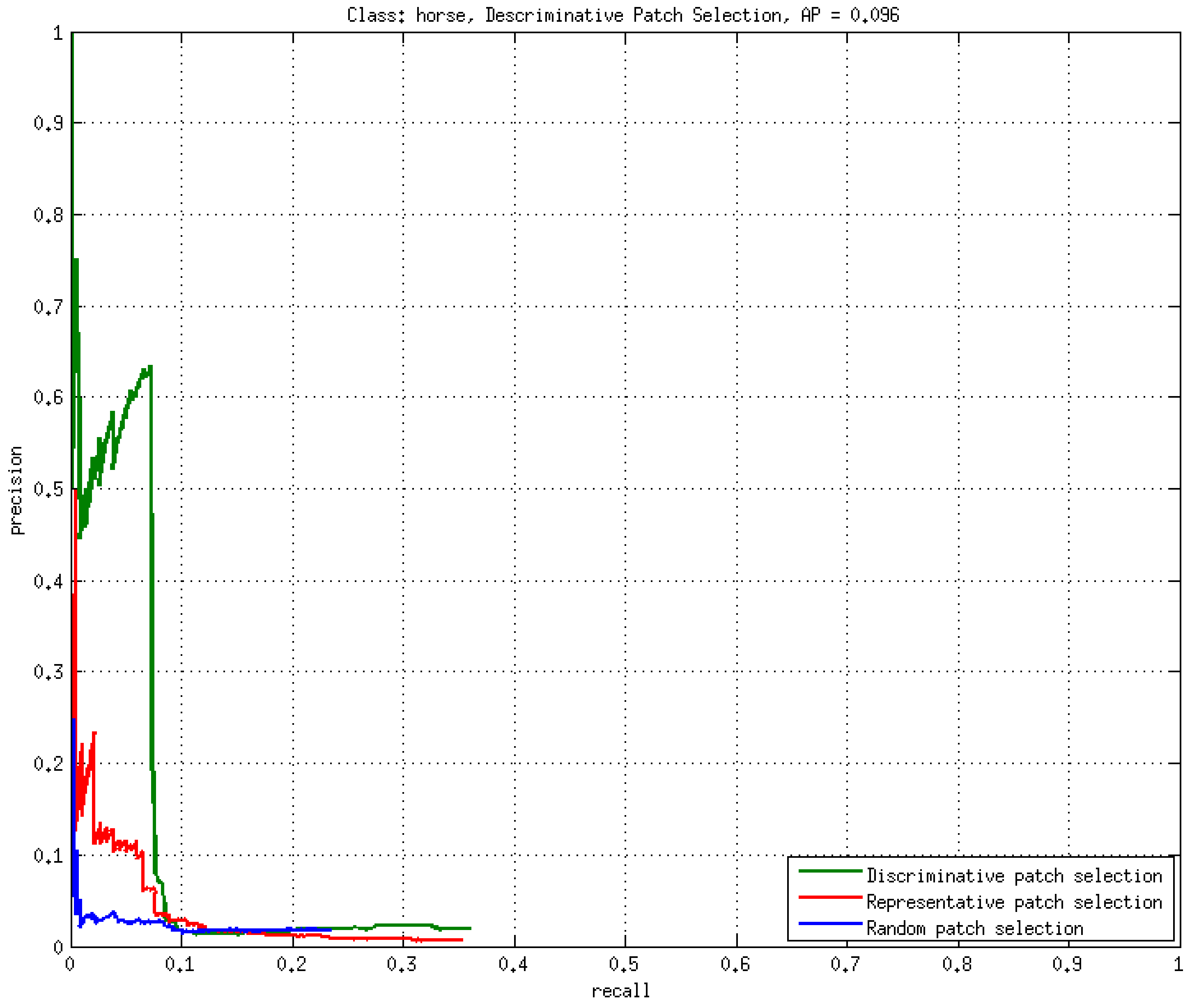
Holistic webly-supervised model is almost on par with supervised DPM on 4 classes



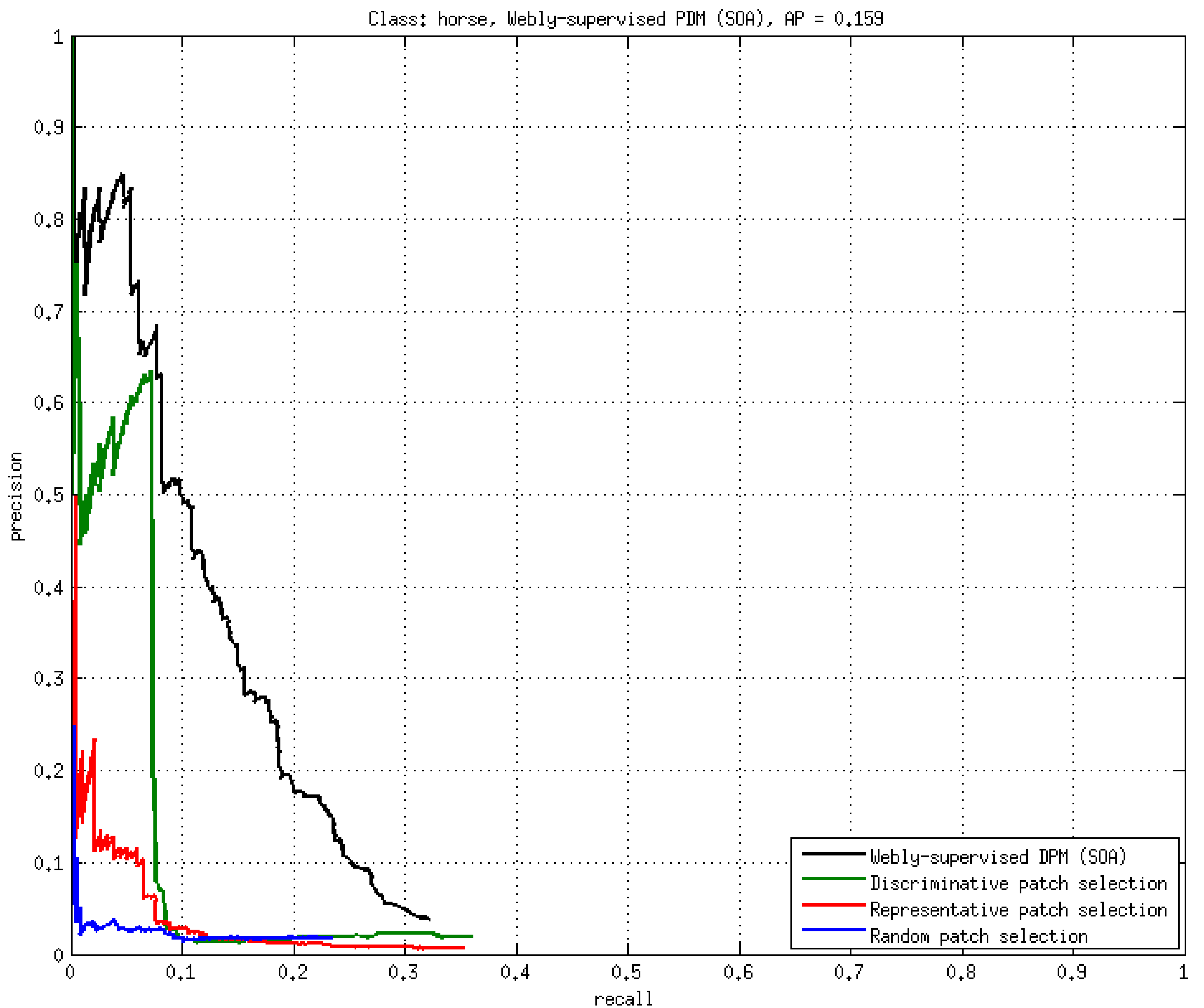
Class: horse, Random Patch Selection, AP = 0.026
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Random patch selection



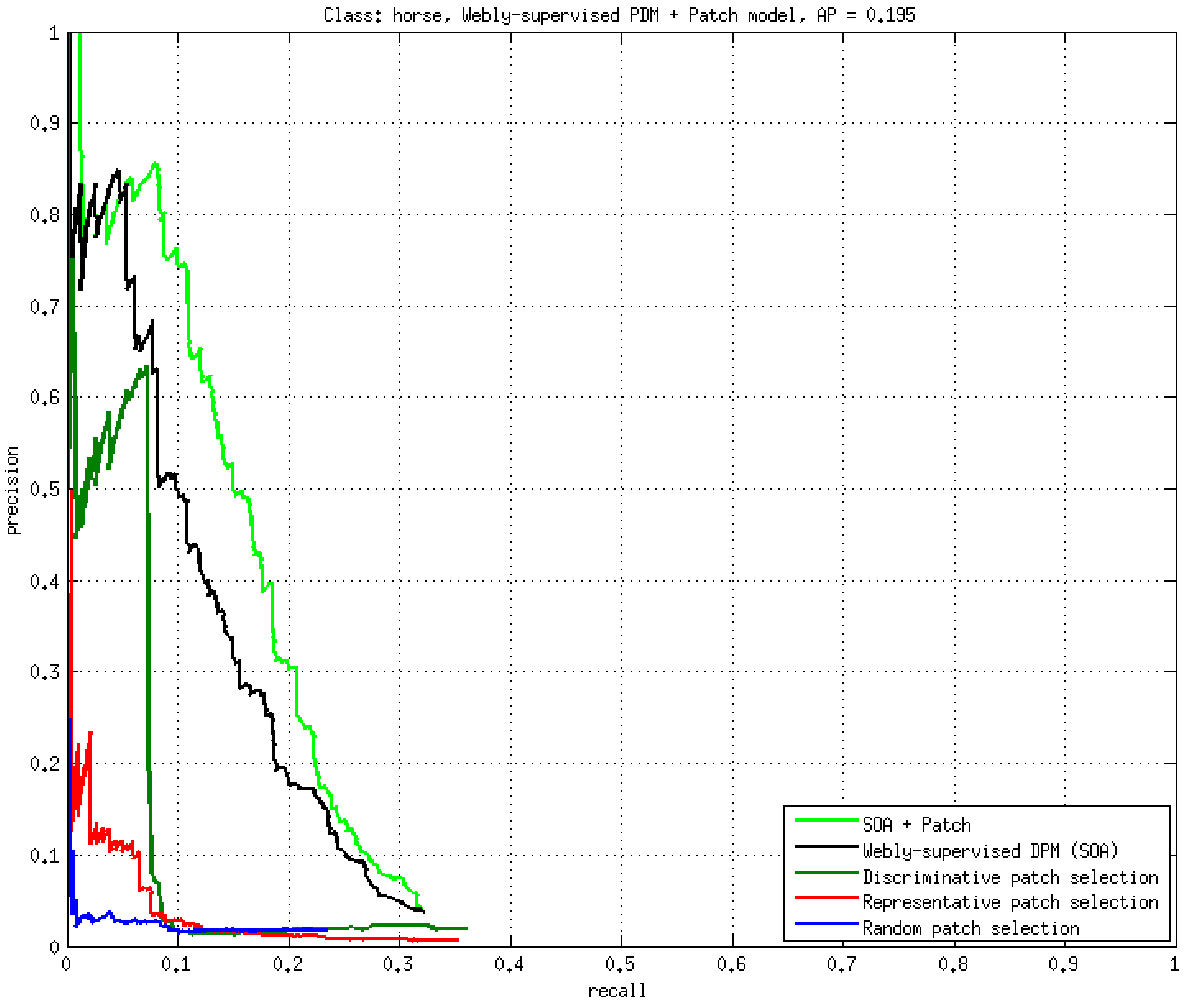
Class: horse, Representative Patch Selection, AP = 0.050
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Representative patch selection
Random patch selection



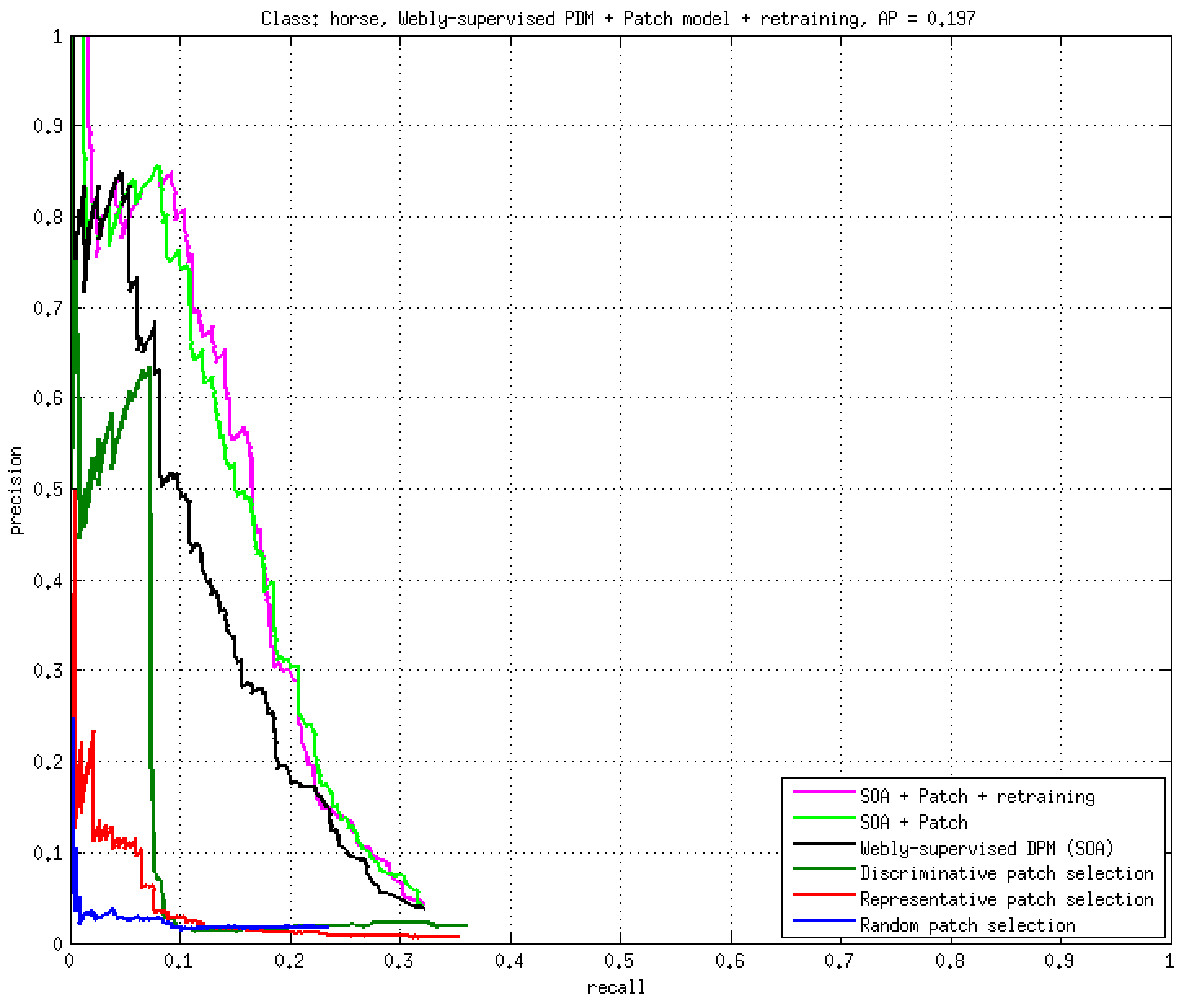
Class: horse, Descriminative Patch Selection, AP = 0.096
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Discriminative patch selection
Representative patch selection
Random patch selection



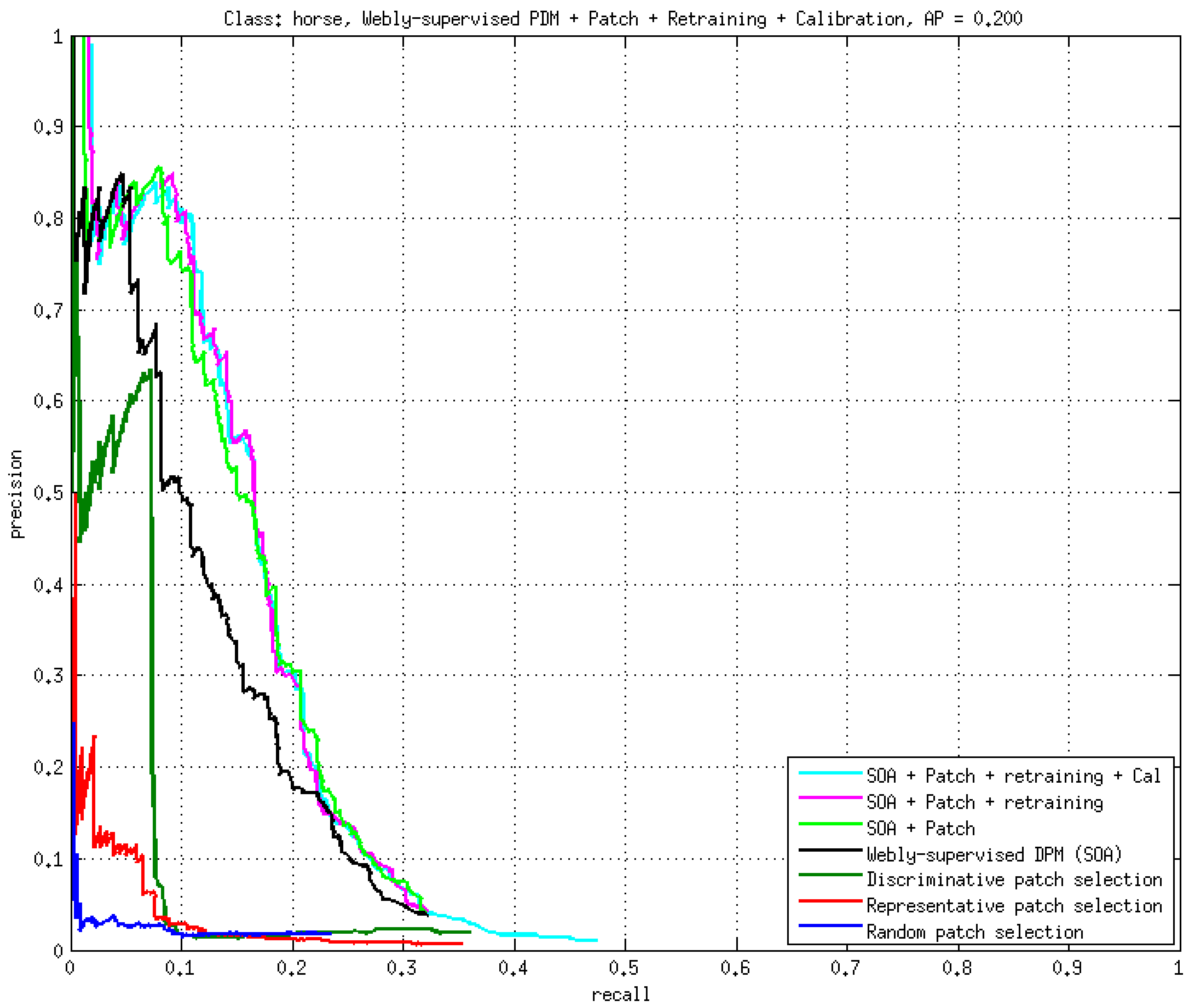
Class: horse, Webly-supervised PDM (SOA), AP = 0.159
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Webly-supervised DPM (SOA)
Discriminative patch selection
Representative patch selection
Random patch selection



Class: horse, Webly-supervised PDM + Patch model, AP = 0.195
precision
recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
SOA + Patch
Webly-supervised DPM (SOA)
Discriminative patch selection
Representative patch selection
Random patch selection

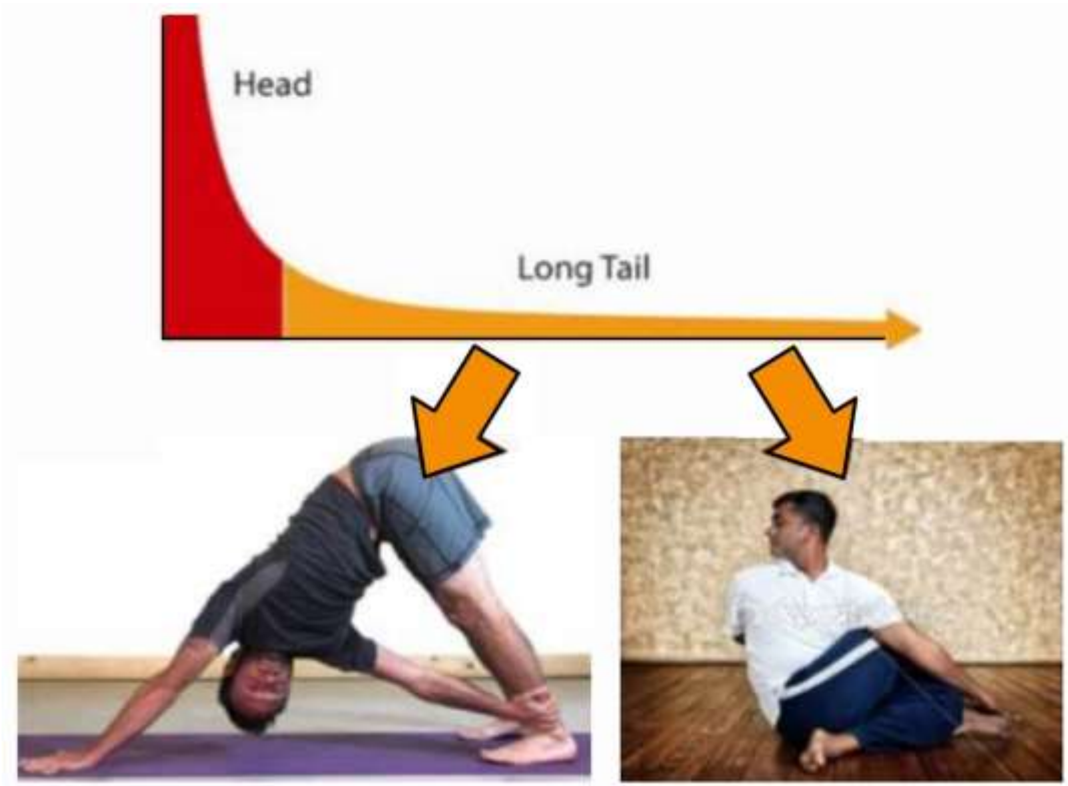


Class: horse, Webly-supervised PDM + Patch model + retraining, AP = 0.197
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
SOA + Patch + retraining
SOA + Patch
Webly-supervised DPM (SOA)
Discriminative patch selection
Representative patch selection
Random patch selection



Class: horse, Webly-supervised PDM + Patch + Retraining + Calibration, AP = 0.200
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
SOA + Patch + retraining + Cal
SOA + Patch + retraining
SOA + Patch
Webly-supervised DPM (SOA)
Discriminative patch selection
Representative patch selection
Random patch selection

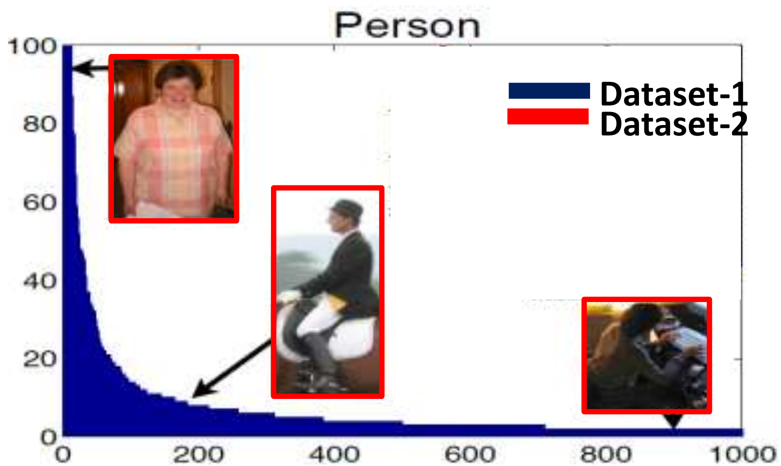
Long-tail distribution

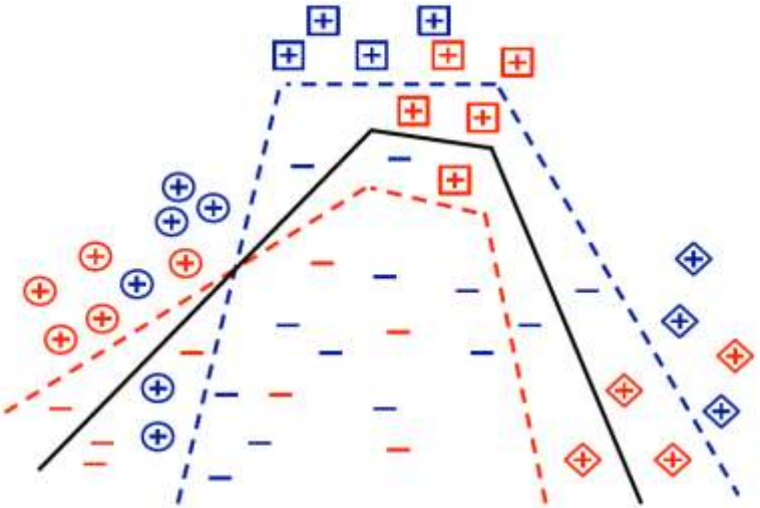


We need lots of templates, have little data of *'yoga twist'* poses

Sharing across datasets

- Solution: Enrich poor subcategory models with *statistical strength* borrowed from other datasets



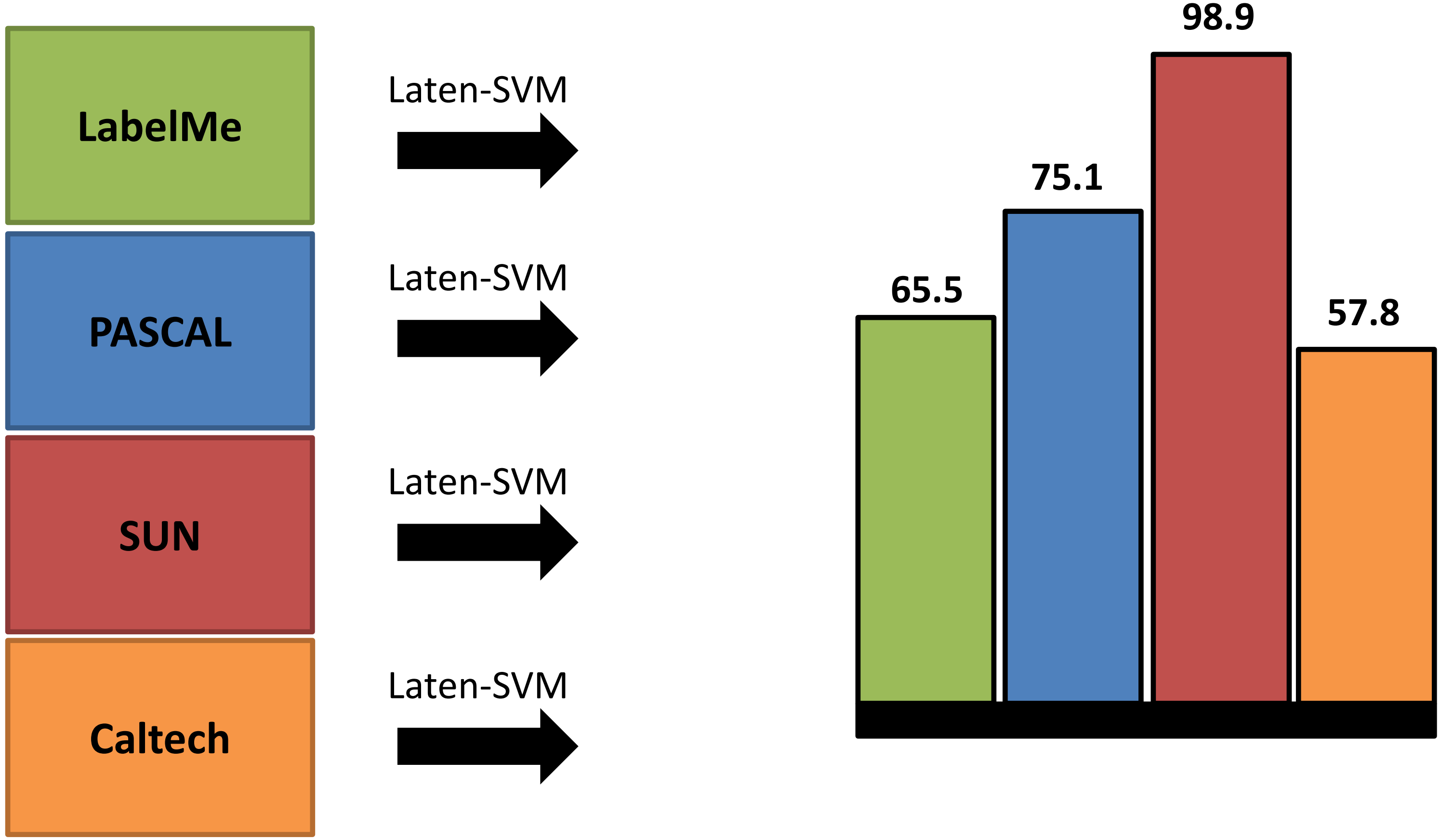


○ **Sample** sharing

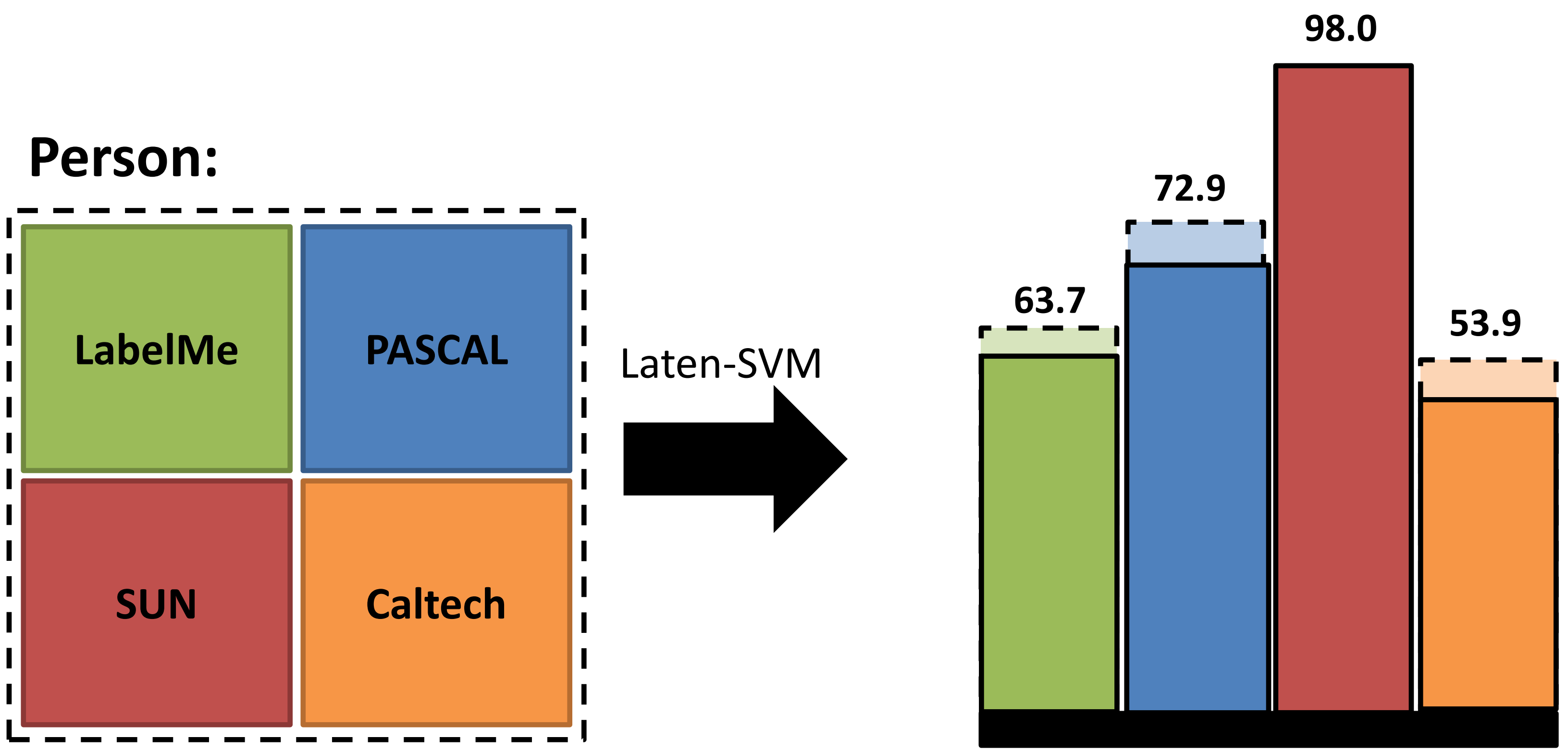
● **Parameter** sharing

Training on dataset individually

Person:



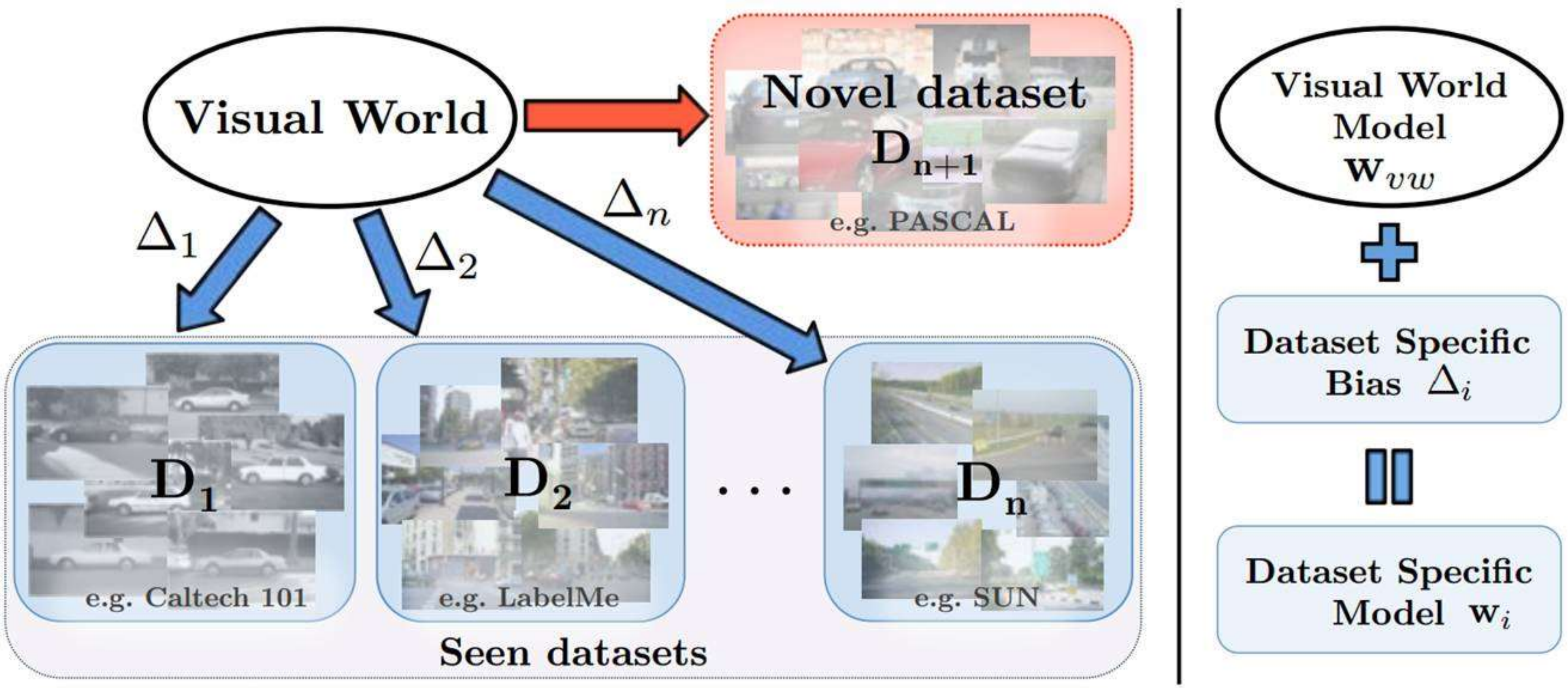
Concatenation of Datasets



* A. Torralba and A. Efross, "**Unbiased look at dataset bias**", In *CVPR* 2011.

Subcategory-based undoing bias

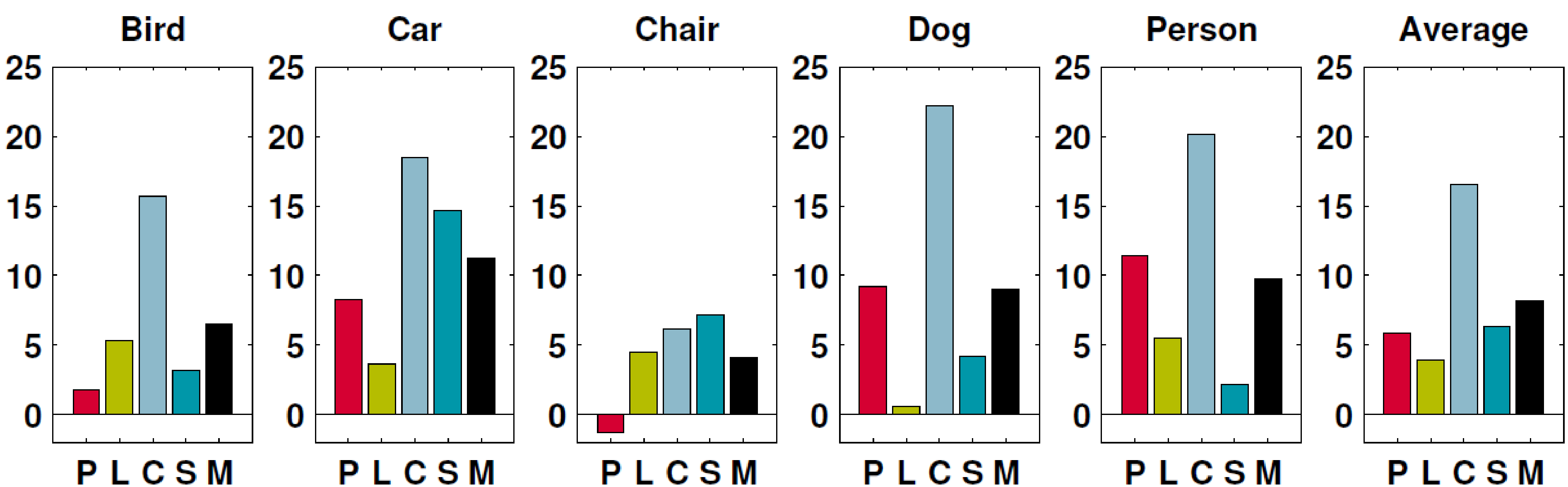
Extend the regularized multi-task learning framework* to the Latent Subcategory setting



* A. Khosla, "**Undoing the damage of dataset bias**", In *ECCV* 2012.

Experiments

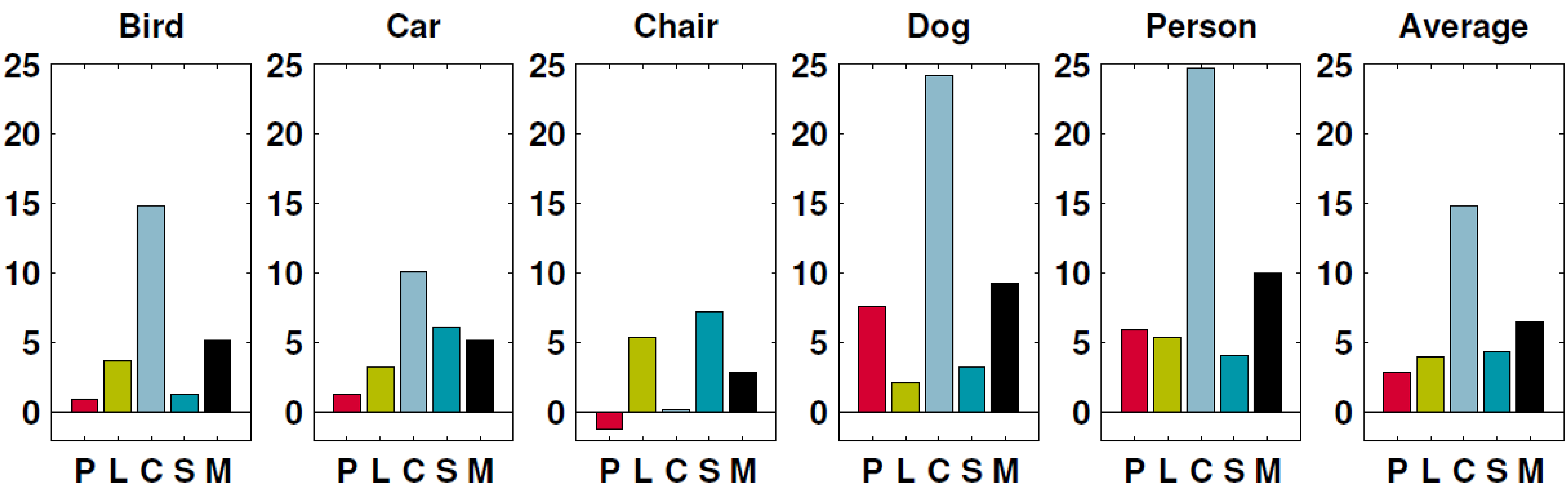
- **Leave-one-dataset-out**
  - **Compare to LSVM on concatenated dataset**



Average improvement: 8.5%

Experiments

- **Leave-one-dataset-out**
  - **Compare to SVM-based undo bias (SOA)**



Average improvement: 6.5%


LabelMe
PASCAL
SUN
Caltech-101