Learning the right thing with visual attributes

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Chao-Yeh Chen, Aron Yu,
and Dinesh Jayaraman

THE UNIVERSITY OF TEXAS AT AUSTIN

Beyond image labels

What does it mean to understand an image?

Labels

Cow
Tree
Grass

vs.

The story of an image

A <u>lone</u> cow grazes in a <u>bright green</u> pasture near an <u>old</u> tree, probably in the Scottish Highlands.

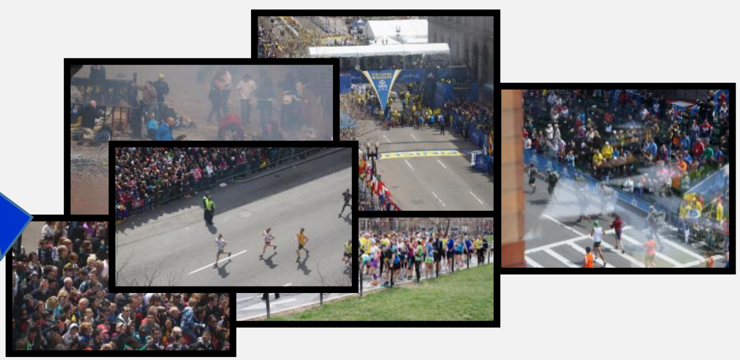
Attributes

- Mid-level semantic properties shared by objects
- Human-understandable *and* machine-detectable

[Ferrari & Zisserman 2007, Kumar et al. 2008, Farhadi et al. 2009, Lampert et al. 2009, Endres et al. 2010, Wang & Mori 2010, Berg et al. 2010, Parikh & Grauman 2011, …]

Using attributes: Visual search

Suspect #1: Male, sunglasses, black and white hat, blue shirt



Person search

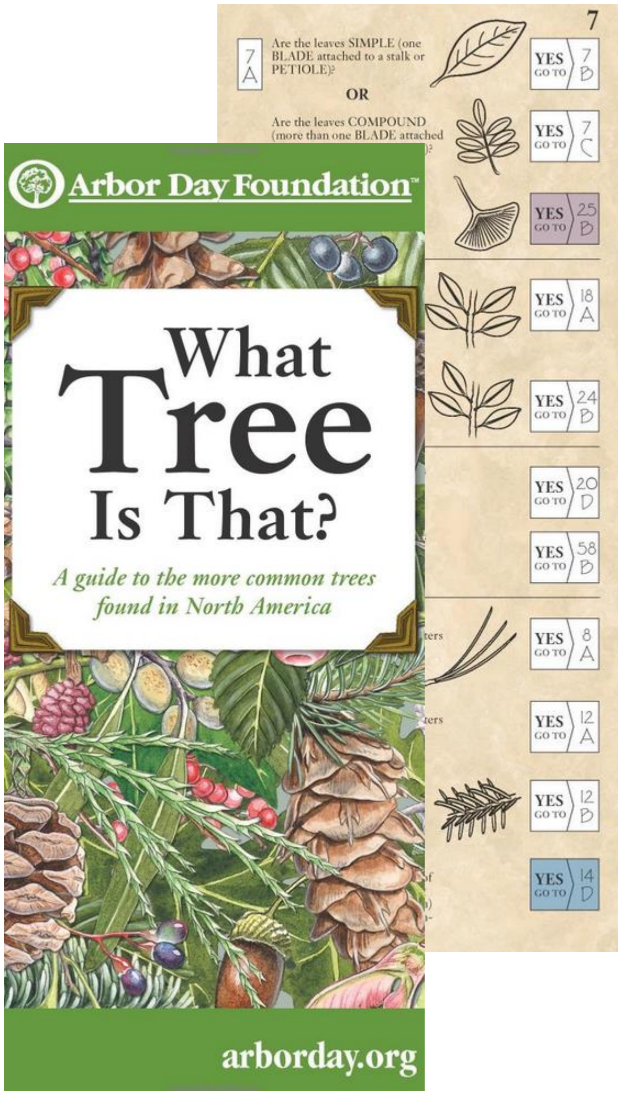
[Kumar et al. 2008, Feris et al. 2013]

"Like this…but *more ornate*"

Relative feedback

[Kovashka et al. 2012]

Using attributes: Interactive recognition



Computer Vision

↓

Cone-shaped beak?

↓ yes

Computer Vision

↓

American Goldfinch?

[Branson et al. 2010, 2013]

Using attributes: Semantic supervision

Band-tailed pigeons:

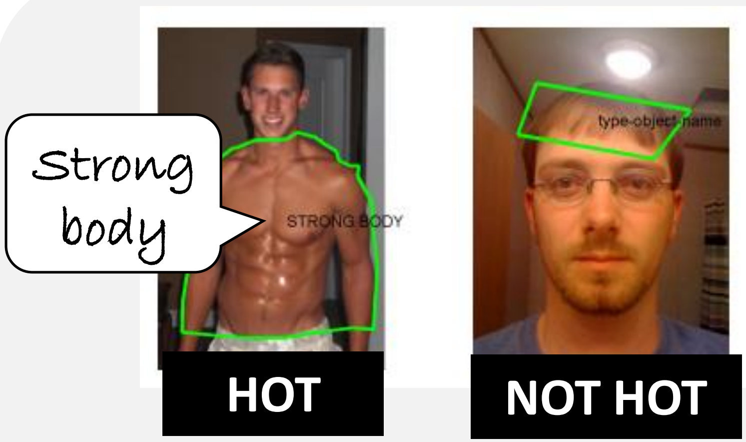
- ✓ White collar
- ✓ Yellow feet
- ✓ Yellow bill
- ✗ Red breast

Zero-shot learning
[Lampert et al. 2009]

Mules:

- ✓ Shorter legs than donkeys
- ✓ Shorter tails than horses

Training with relative descriptions
[Parikh & Grauman 2011, Shrivastava & Gupta 2012]



Annotator rationales
[Donahue & Grauman 2011]

Problem

With attributes, it’s easy to learn the wrong thing.

- Incidental correlations
- Spatially overlapping properties
- Subtle visual differences
- Partially category-dependent
- Variance in human-perceived definitions

…yet applications demand that correct meaning be captured!

Goal

Learn the right thing.

- How to decorrelate attributes that often occur simultaneously?
- Are attributes really class-independent?
- How to detect fine-grained attribute differences?

The curse of correlation

What will be learned from this training set?

Object Learning

Cat

The curse of correlation

What will be learned from this training set?

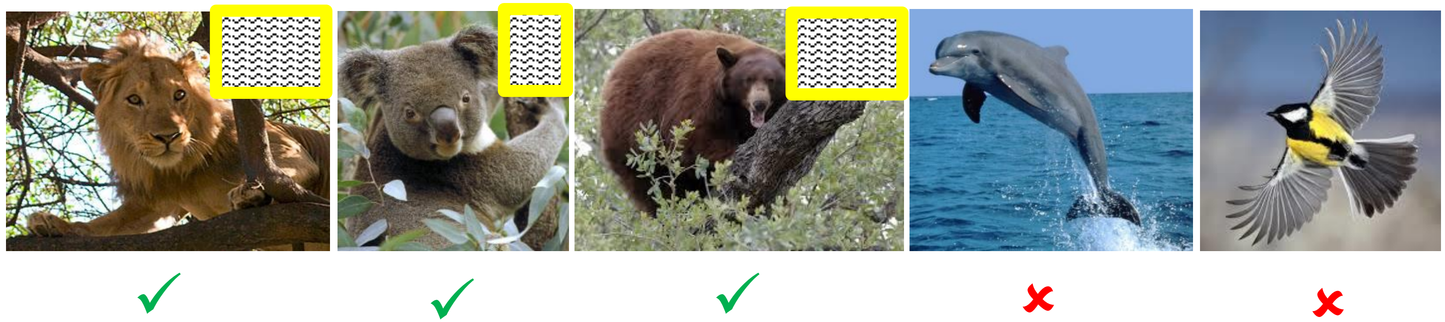
Attribute Learning

Forest animal? Brown? Has ears? Combinations?

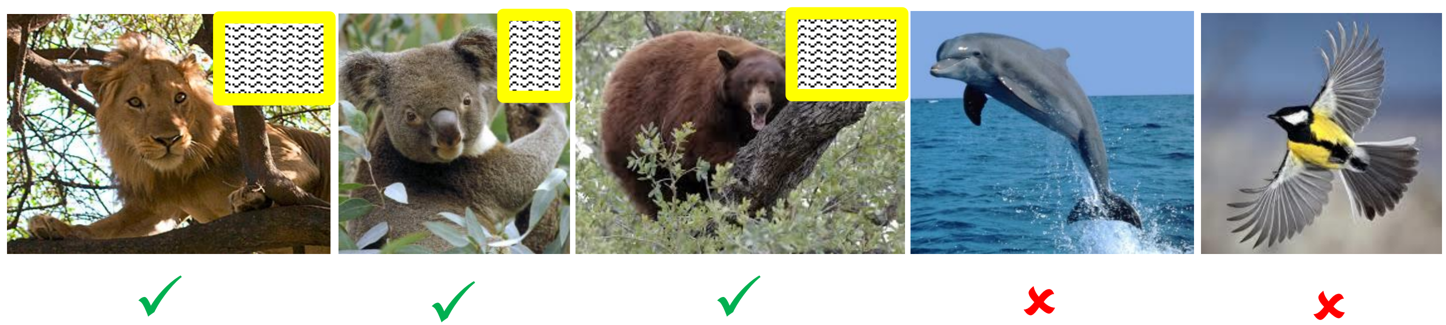
Problem: Attributes that often co-occur cannot be distinguished by the learner

The curse of correlation

Forest animal

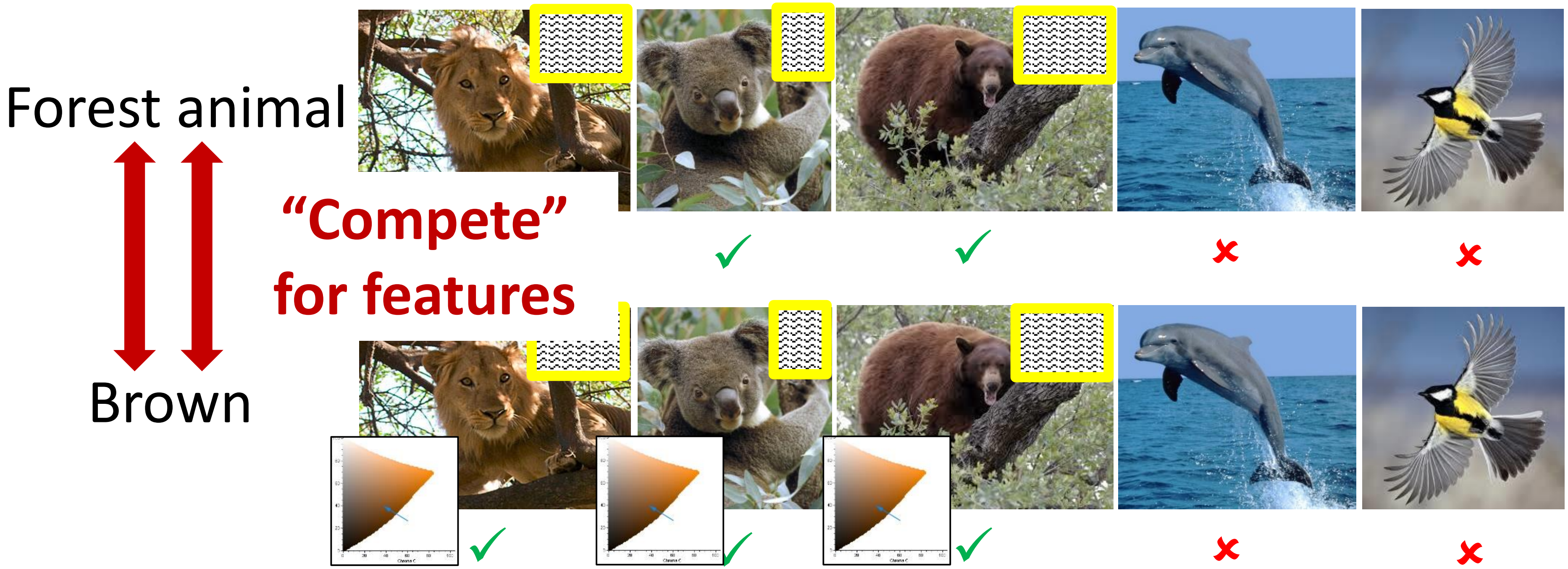


Brown



Problem: Attributes that often co-occur cannot be distinguished by the learner

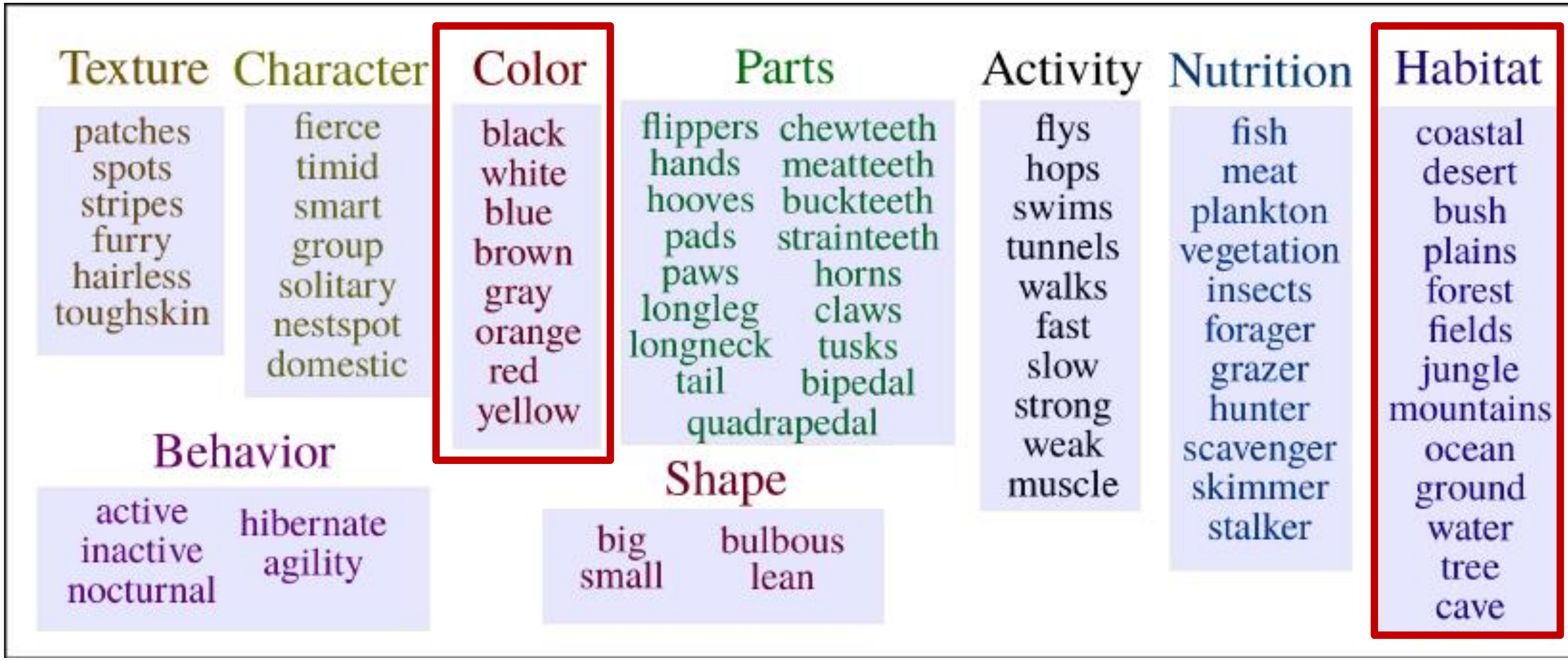
Idea: Resist the urge to share



Problem: Attributes that often co-occur cannot be distinguished by the learner

Semantic attribute groups

- Closely related attributes *may* share features
- Assume attribute "groups" from external knowledge.



Standard approach: learning separately

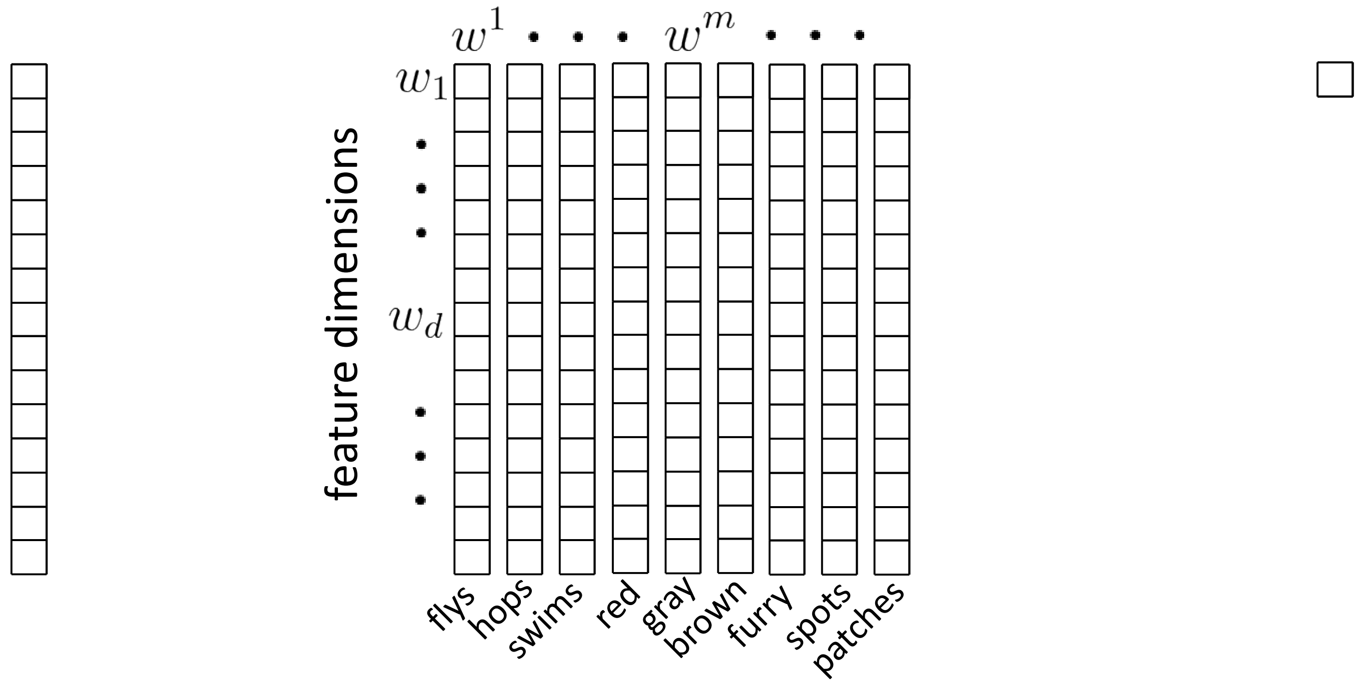
Loss function: $L(W|X,Y) = \sum_{m} \sum_{(x,y)\in \text{samples}} \log\left(1 + e^{-y^m (x^T w^m)}\right)$

m : attribute index

x : feature vector

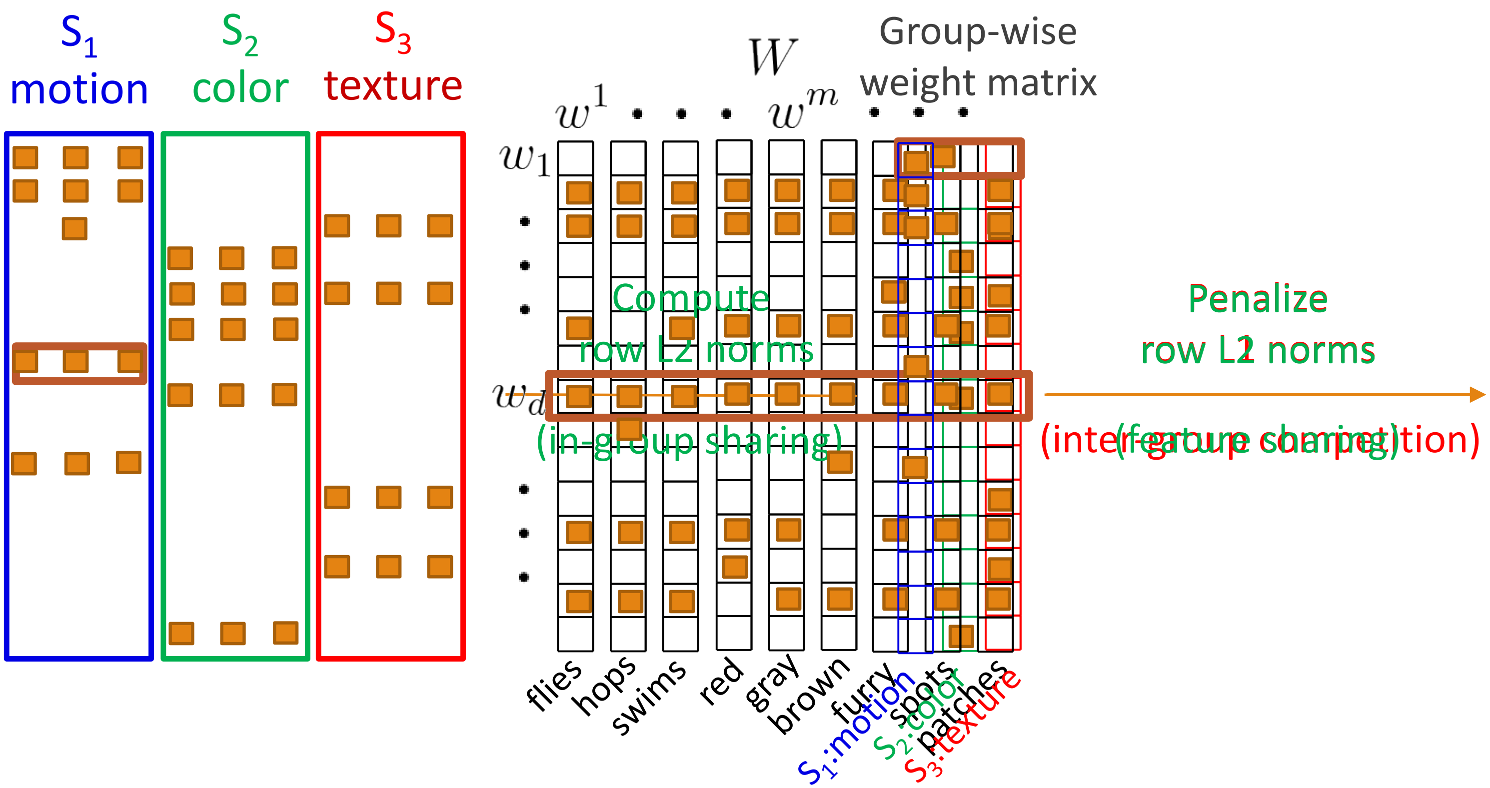
w^m : learned weights

y^m : label (± 1)



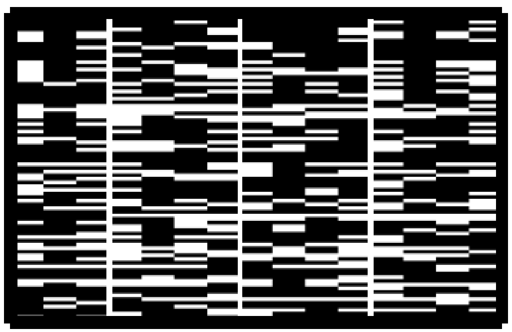
Proposed group-based formulation

$$\underset{W}{\operatorname{argmin}}\; L(W|X,Y) + \sum_d \sum_l \|w_d^{S_l}\|_2$$

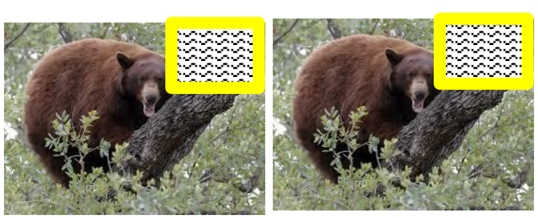


Formulation effect

Sparse features
(no relationships among attributes)

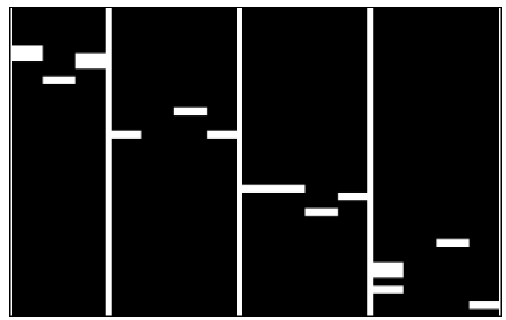


$$\sum_i \sum_j |w_i^j|$$



Forest animal Brown

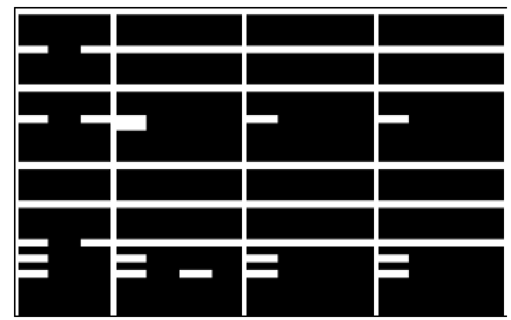
Ours
(inter-group competition, in-group sharing)



$$\sum_d \sum_l \|w_d^{S_l}\|_2$$

Forest animal Brown

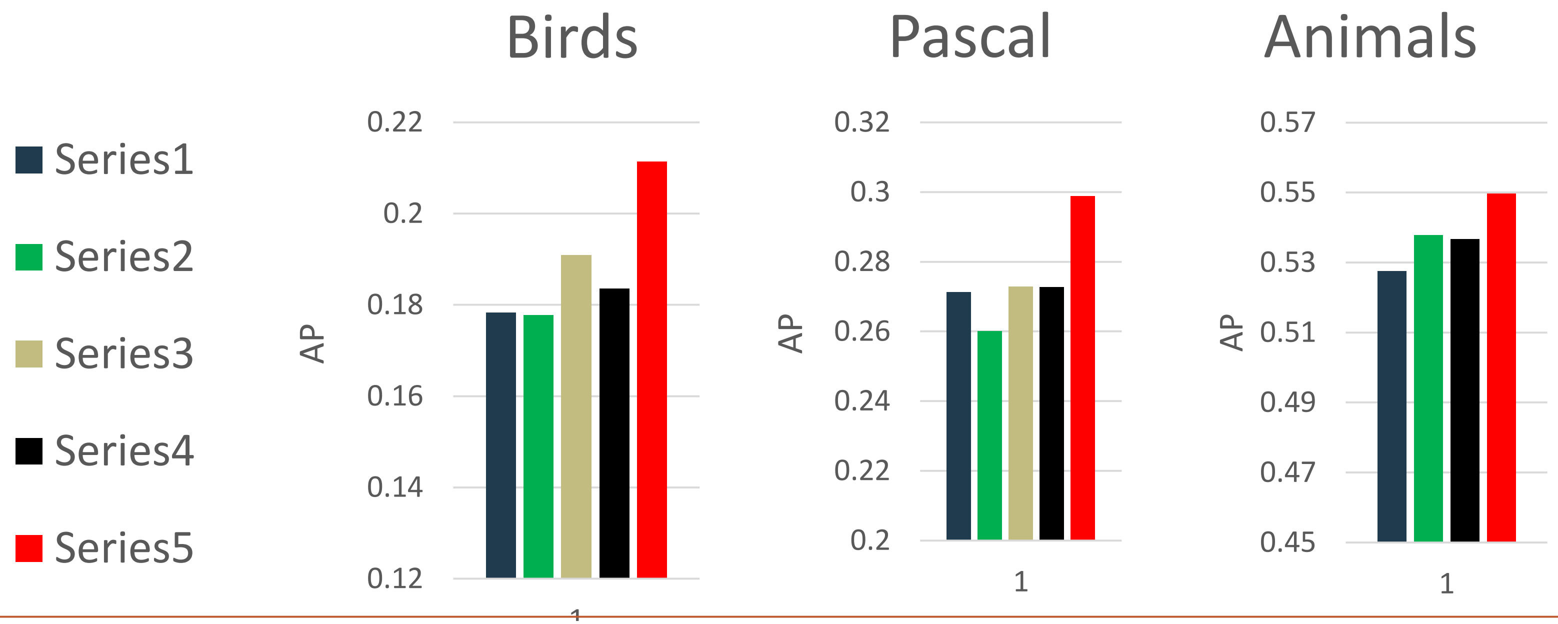
Standard multi-task learning
(sharing and conflation across groups)



$$\sum_d \|w_d\|_2$$

Forest animal Brown

Results – Attribute detection



By decorrelating attributes, our attribute detectors generalize much better to novel unseen categories.

(*) Argyriou et al, Multi-task Feature Learning, NIPS 2007
(~) Farhadi et al, Describing Objects by Their Attributes, CVPR 2009

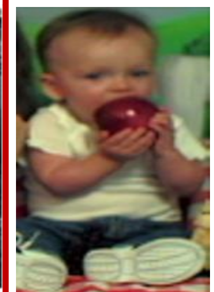
Attribute detection example

Success cases

Not brown underparts

No eye

Not boxy



No mouth

No ear

Failure cases

No feather

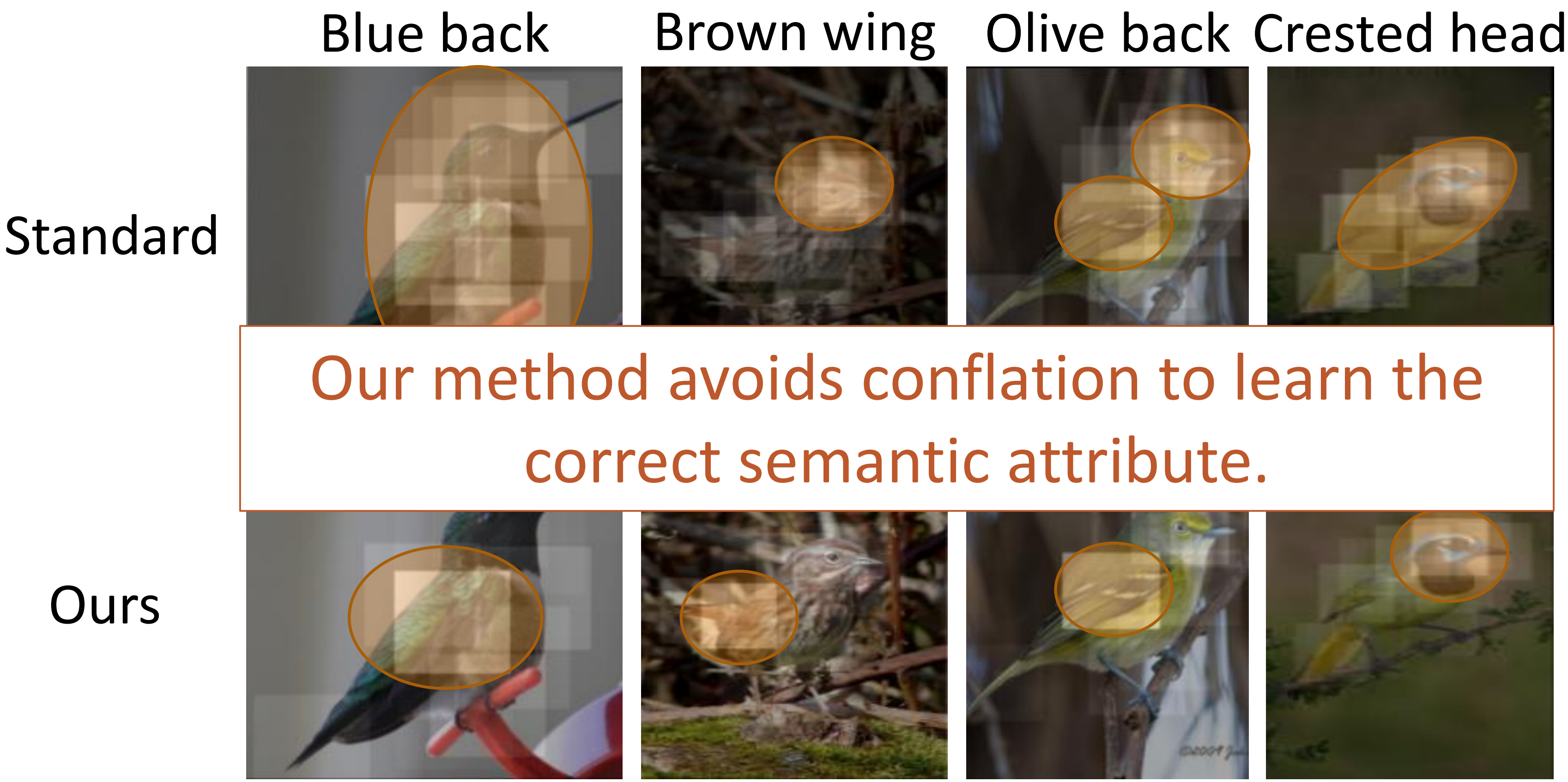
Not furry

Eyeline

Black breast

Not vegetation

Attribute localization examples



Our method avoids conflation to learn the correct semantic attribute.

Goal

Learn the right thing.

- How to decorrelate attributes that often occur simultaneously?
- Are attributes really class-independent?
- How to detect fine-grained attribute differences?

Problem

Are attributes really category-independent?

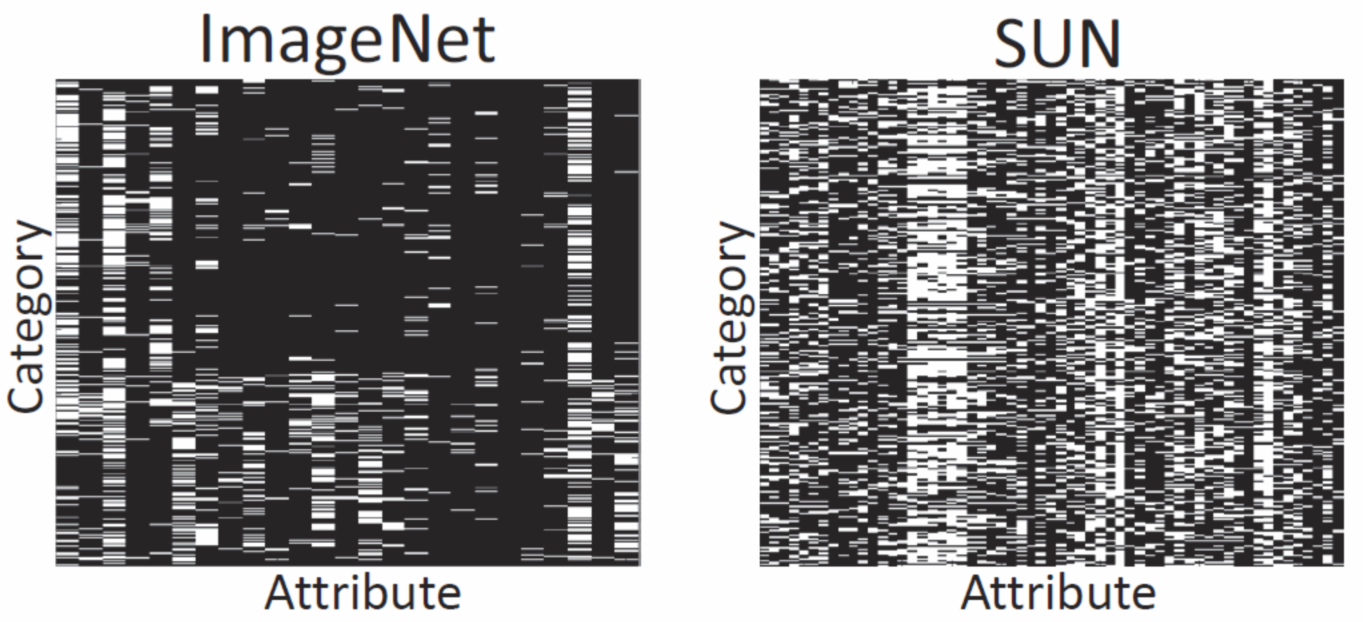
Fluffy dog

? =

Fluffy towel

An intuitive but impractical solution

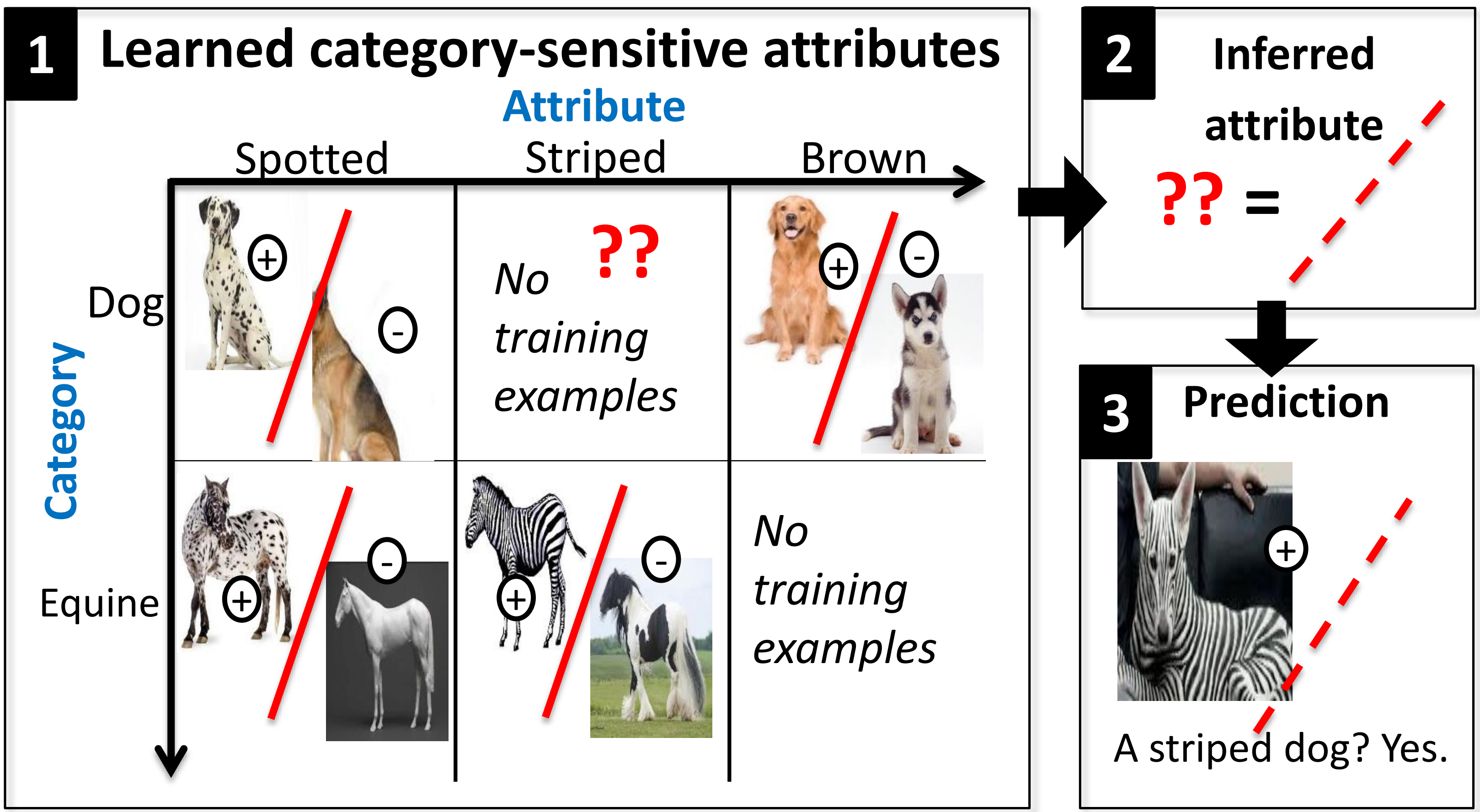
- Learn category-specific attributes?



Impractical!
Would need examples for **all** category-attribute combinations…

Idea: Analogous attributes

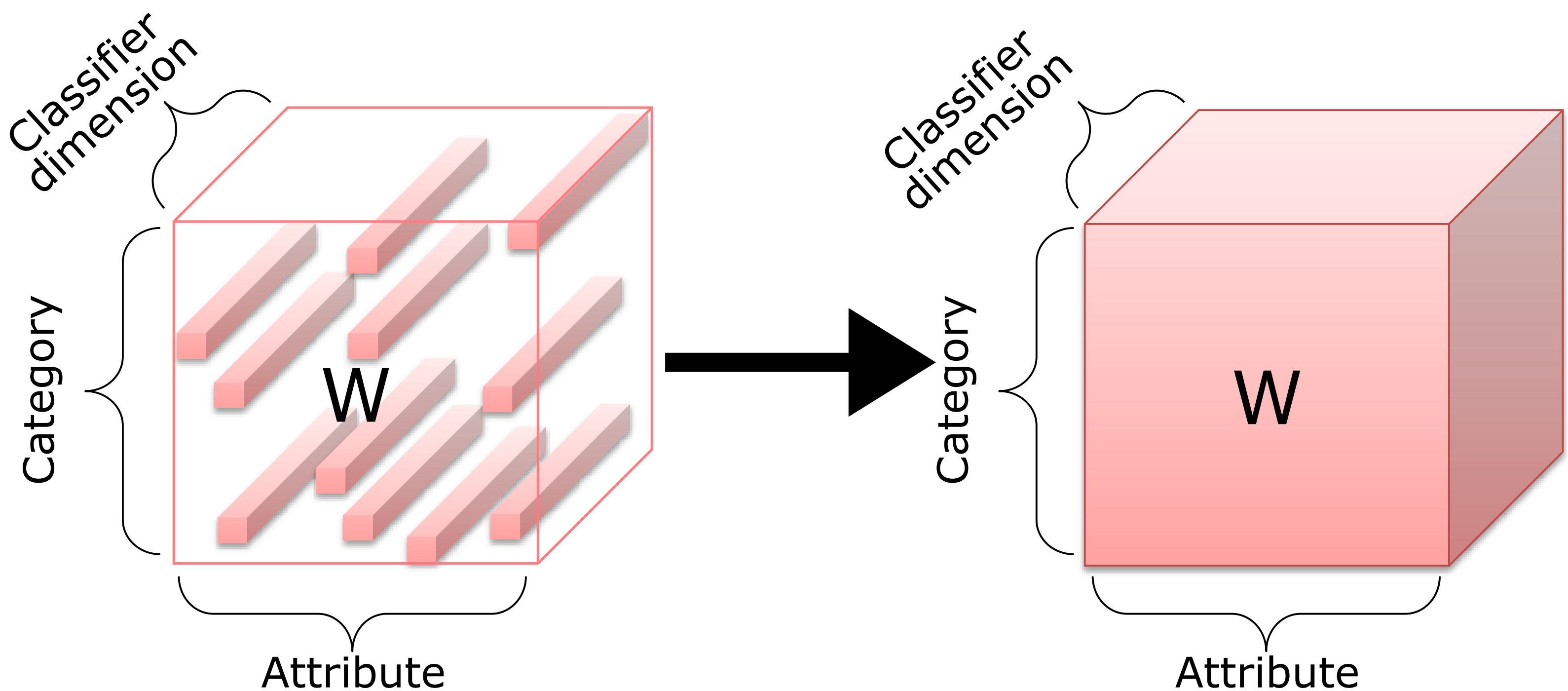
- Given sparse set of category-specific models, infer “missing” analogous attribute classifiers



Chen & Grauman, CVPR 2014

Transfer via tensor completion

Construct sparse object-attribute classifier tensor → Discover low-d latent factors and infer missing classifiers (the analogous attributes)



$$\mathbf{W} \approx \sum_{k=1}^{K} O^k \circ A^k \circ C^k$$

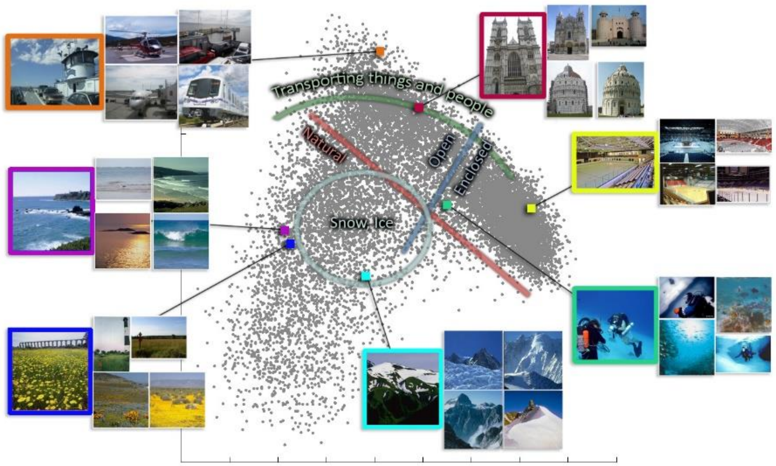
Bayesian probabilistic tensor factorization [Xiong et al., SDM 2010].

Datasets

- **ImageNet attributes**
  - 9600 images
  - 384 object categories
  - 25 attributes
  - 1498 object-attribute pairs available

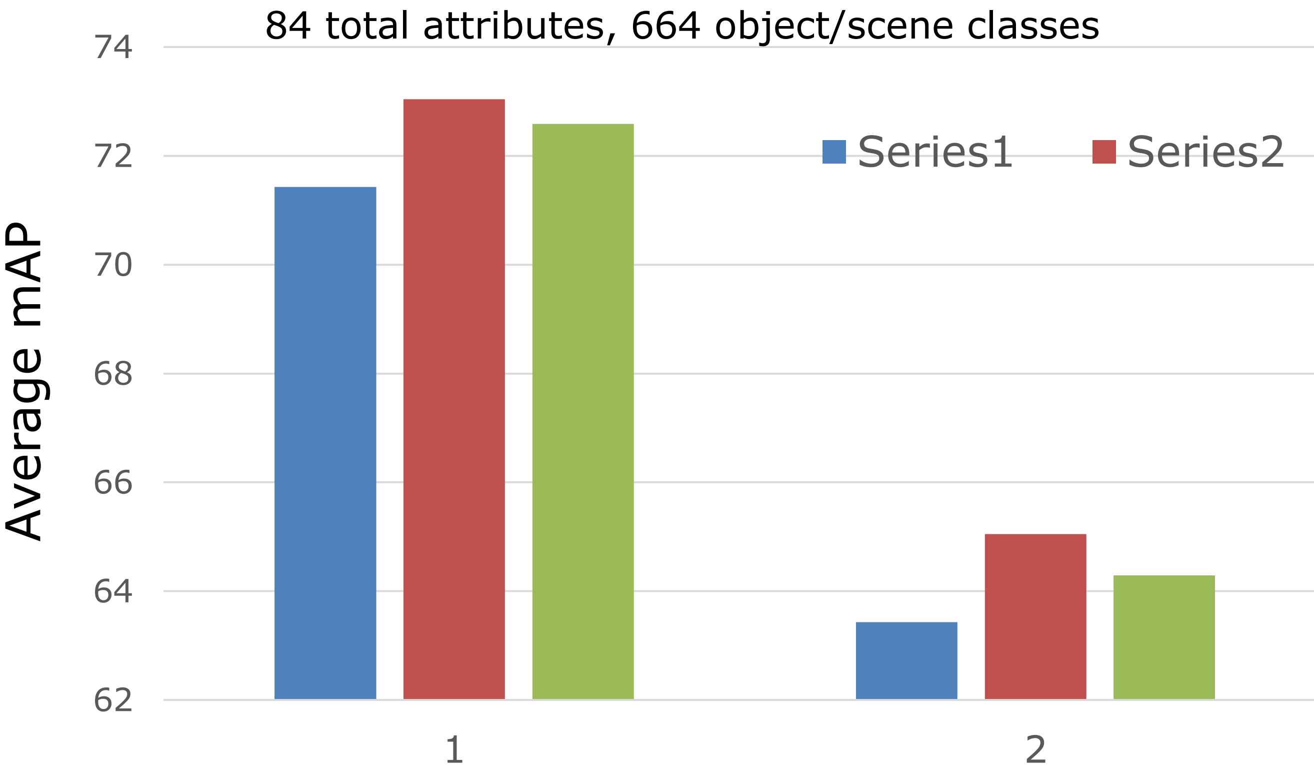
[Russakovsky & Fei-Fei 2010]

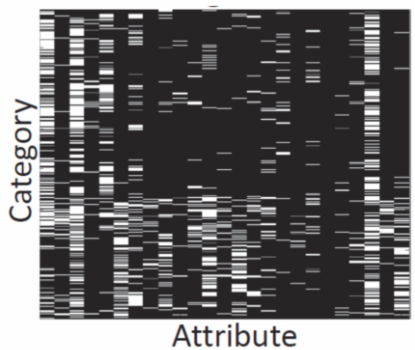
- **SUN attributes**
  - 14340 images
  - 280 object categories
  - 59 attributes
  - 6118 object-attribute pairs available

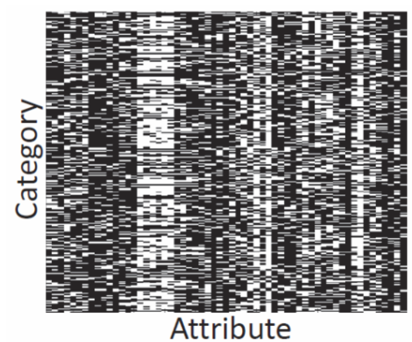


[Patterson & Hays 2012]

Inferring class-sensitive attributes





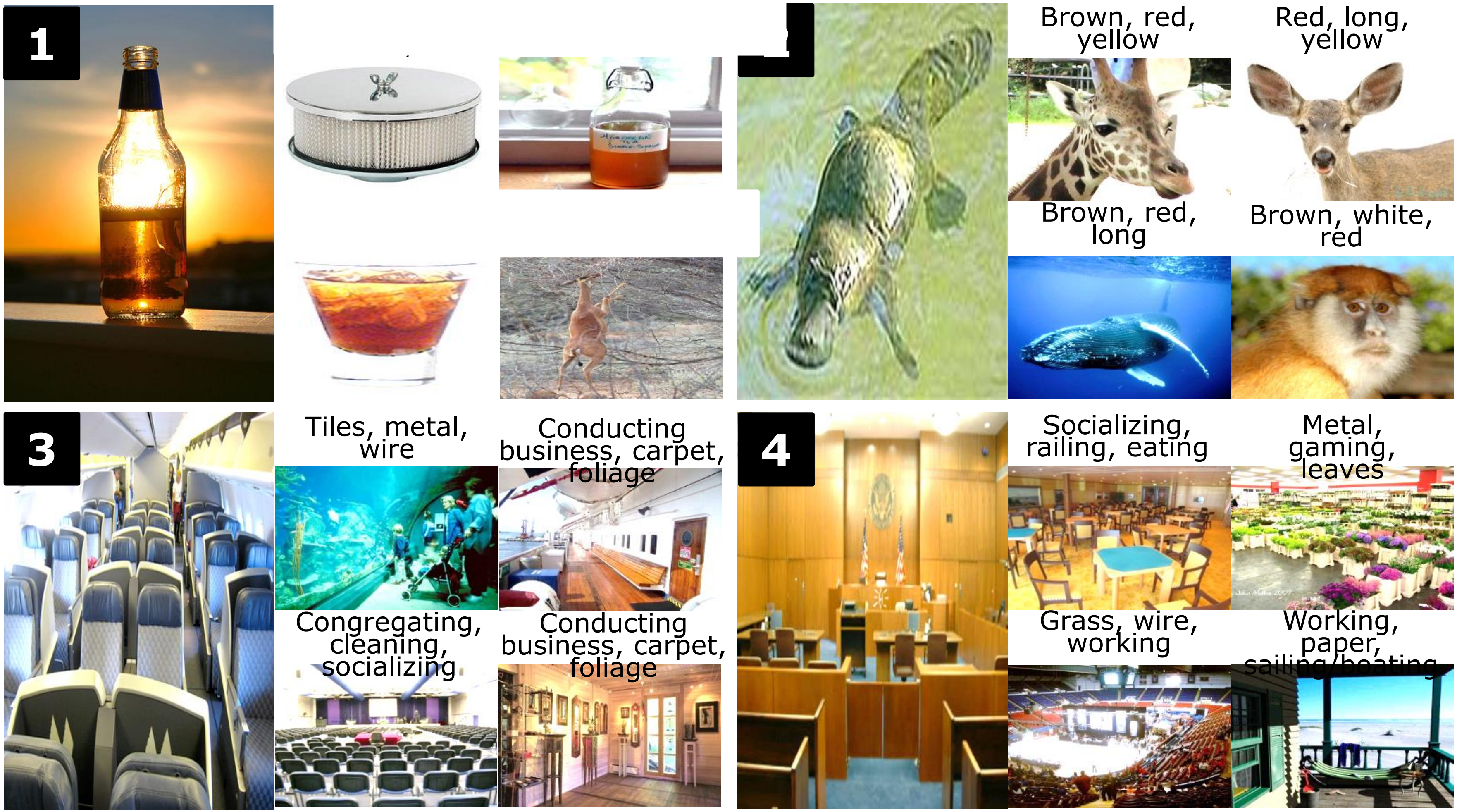


Our approach infers all 18K "missing" classifiers → savings of 348K labeled images

Category-sensitiv
outperforms status
76% of the time
average gain of 1
points in AP

Chen & Grauman, CVPR 2014

Which attributes are analogous?



Chen & Grauman, CVPR 2014

Goal

Learn the right thing.

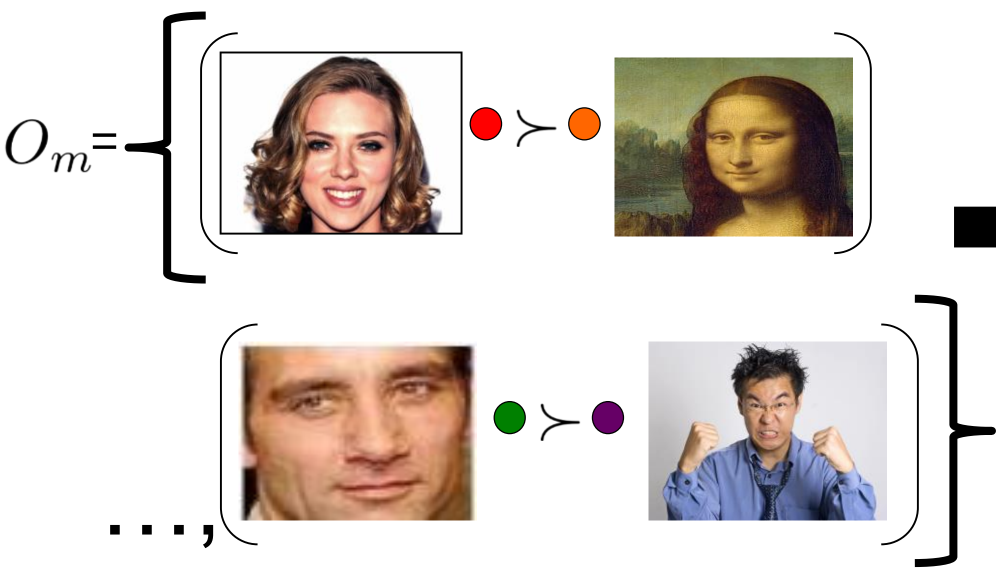
- How to decorrelate attributes that often occur simultaneously?
- Are attributes really class-independent?
- How to detect fine-grained attribute differences?

Problem: Fine-grained attribute comparisons

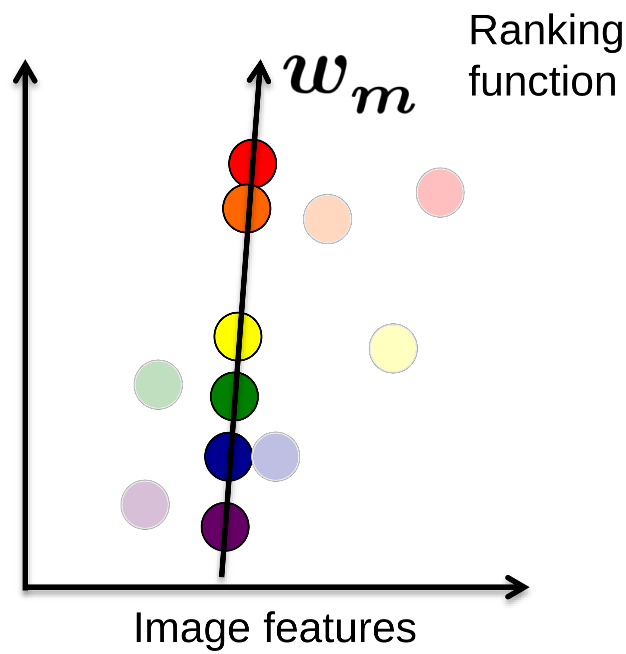
Which is ***more comfortable***?

Relative attributes

Use ordered image pairs to train a ranking function:



"smiling more than"



$$w_m^T x_i > w_m^T x_j$$

$$\forall (i,j) \in O_m$$

[Parikh & Grauman, ICCV 2011; Joachims 2002]

Relative attributes

Rather than simply label images with their properties,

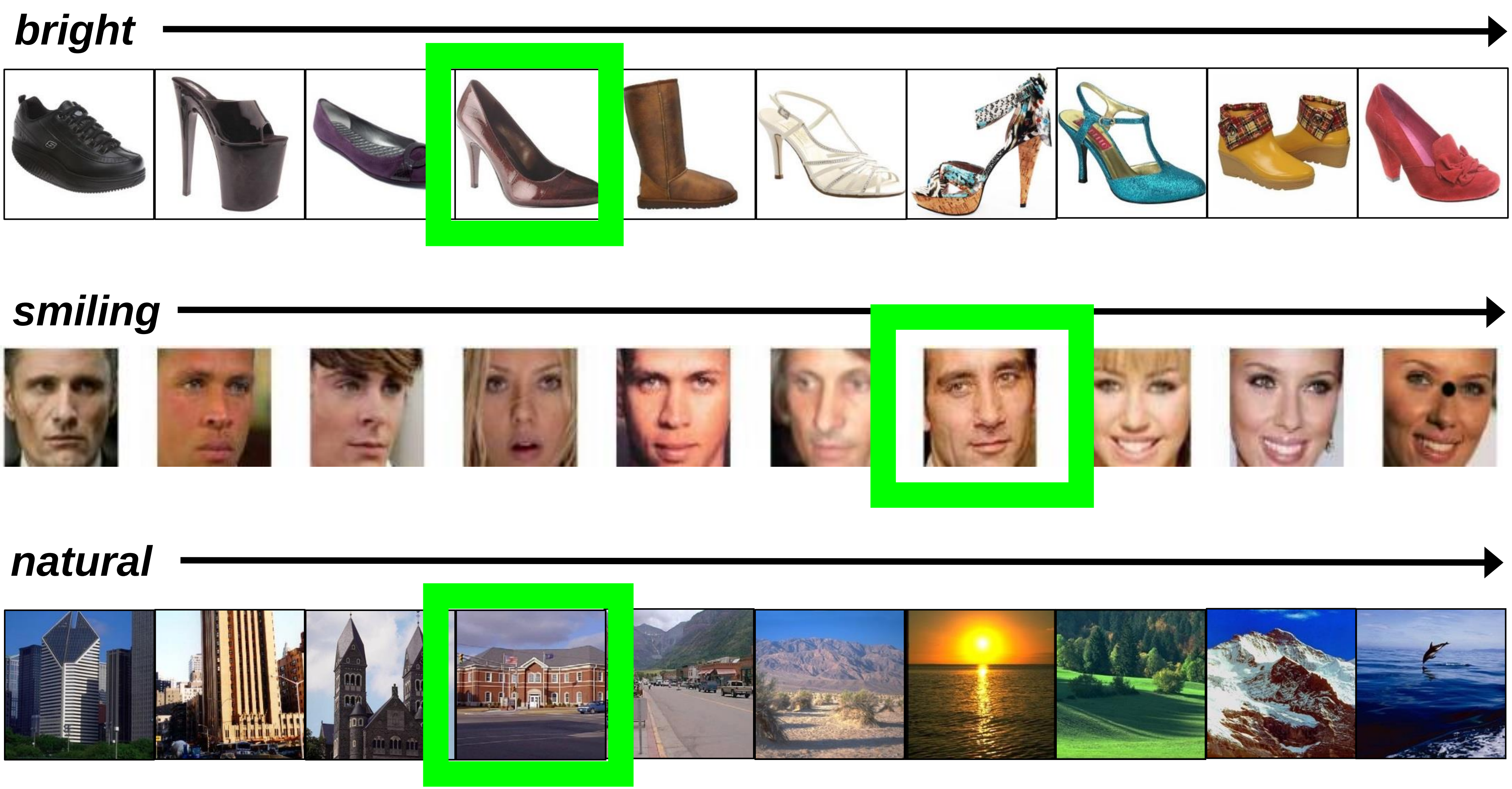
Not bright

Smiling

Not natural

Relative attributes

We can compare images by attribute's "strength"

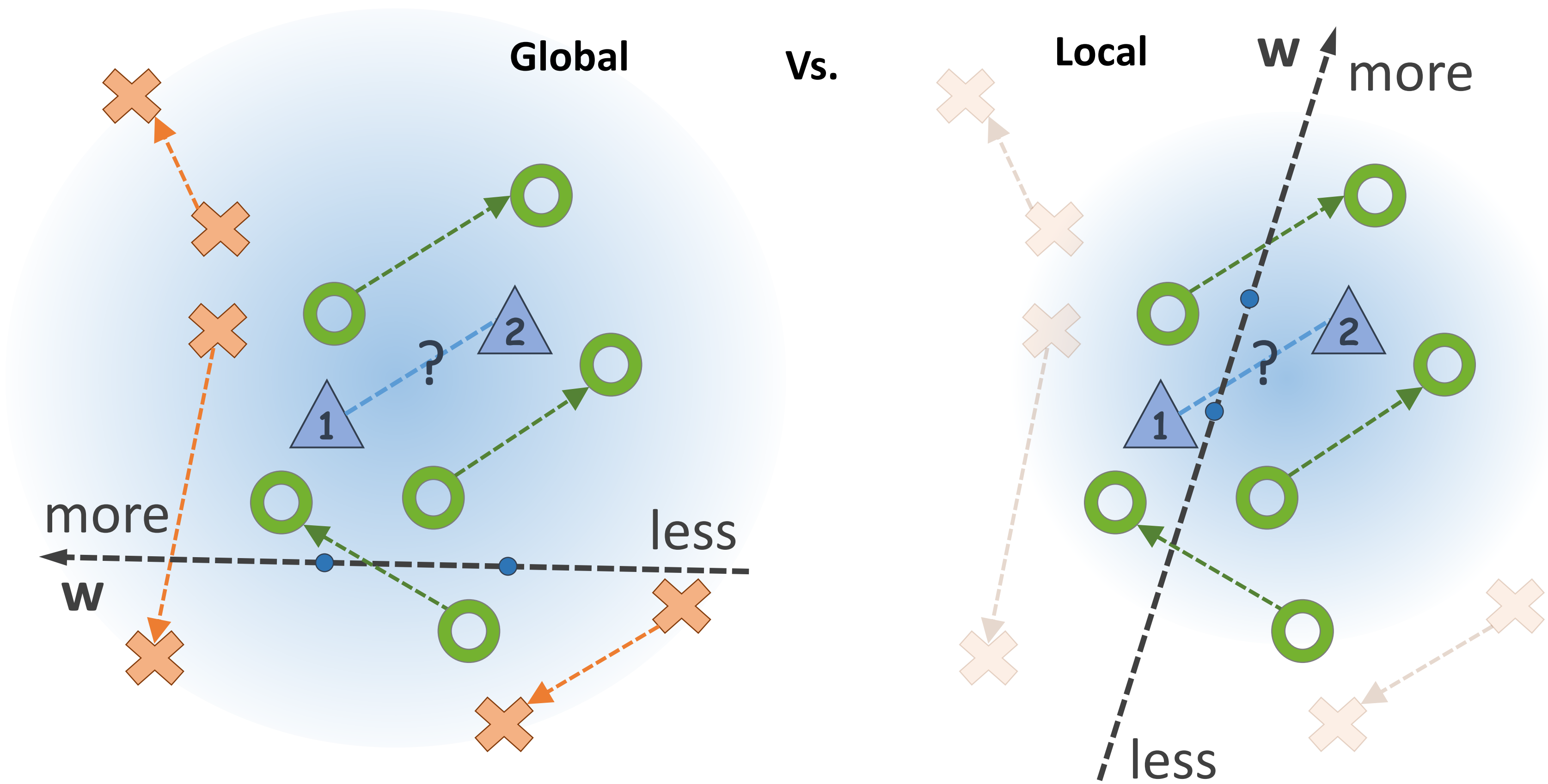


Idea: Local learning for fine-grained relative attributes

- Lazy learning: train query-specific model on the fly.
- Local: use only pairs that are similar/relevant to test case.

Yu & Grauman, CVPR 2014

Idea: Local learning for fine-grained relative attributes



Yu & Grauman, CVPR 2014

UT Zappos50K Dataset

Large shoe dataset, consisting of 50,025 catalog images from Zappos.com

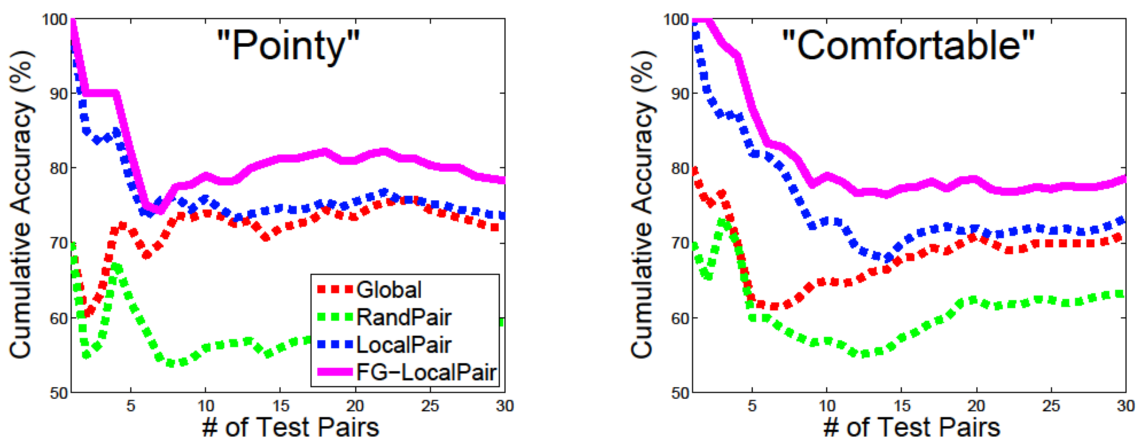
- 4 relative attributes
- High quality pairwise labels from mTurk workers
- 6,751 ordered labels + 4,612 "equal" labels
- 4,334 twice-labeled fine-grained labels (no "equal" option)

Yu & Grauman, CVPR 2014

Results: Fine-grained attributes

Accuracy of comparisons – all attributes

	Zap50K-1	Zap50K-2	OSR	PubFig
RelTree [2]	–	–	90.41	83.37
Global [3]	89.57	61.62	88.80	80.56
RandPair	84.34	57.98	86.93	72.46
FG-LocalPair	**91.64**	**66.43**	**92.37**	**89.72**

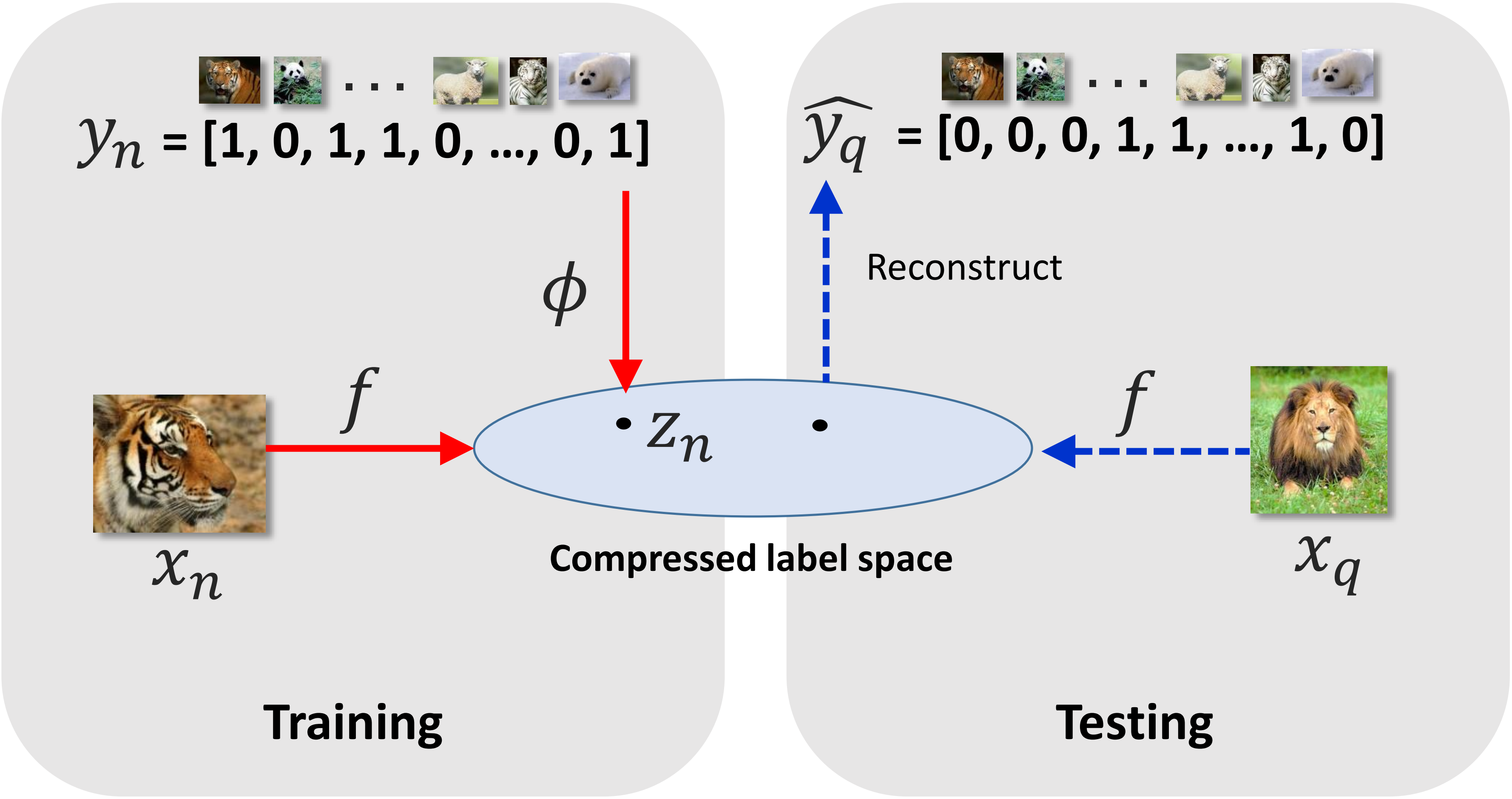


Accuracy on the 30 hardest test pairs

Yu & Grauman, CVPR 2014

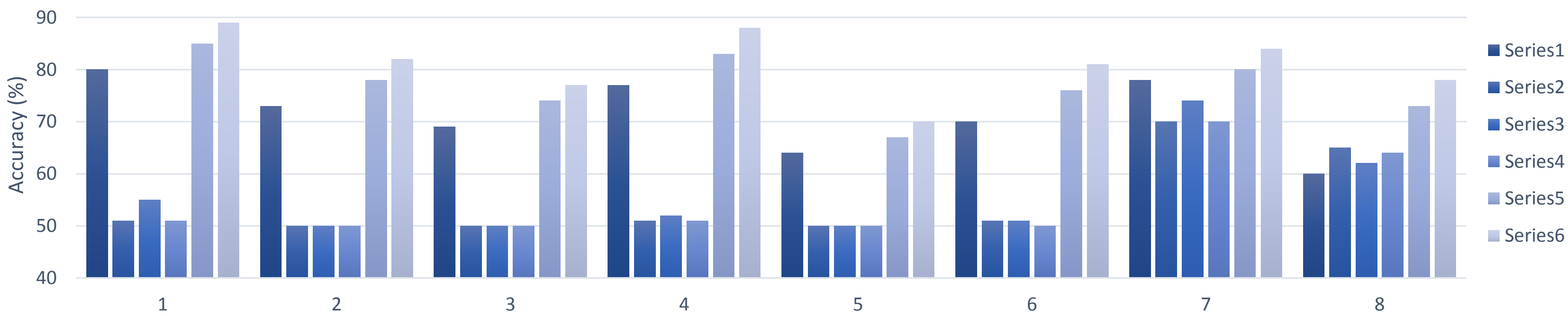
Predicting useful neighborhoods

- Most relevant points = most similar points?
- Pose as large-scale multi-label classification problem

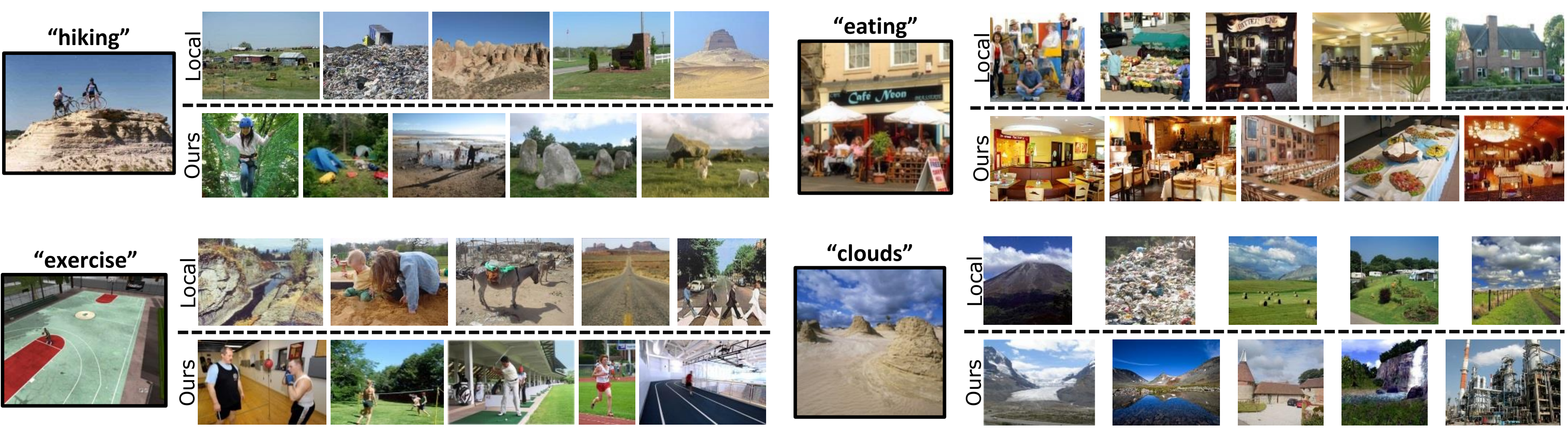


[Yu & Grauman NIPS 2014]

Predicting useful neighborhoods



SUN Attribute Dataset: 14,340 images, 707 classes



Yu & Grauman, NIPS 2014

Summary

- Attribute learning is more nuanced than object learning
- Essential that language and visual concepts align

- Ideas:
  - Explicitly decorrelate attribute classifiers
  - Transfer between analogous attribute-object models
  - Fine-grained comparisons via lazy local learning