Synthesized Classifiers for Zero-Shot Learning

Soravit Changpinyo*[1], Wei-Lun Chao*[1], Boqing Gong[2], and Fei Sha[3]

[1]U. of Southern California, [2]U. of Central Florida, [3]U. of California, Los Angeles

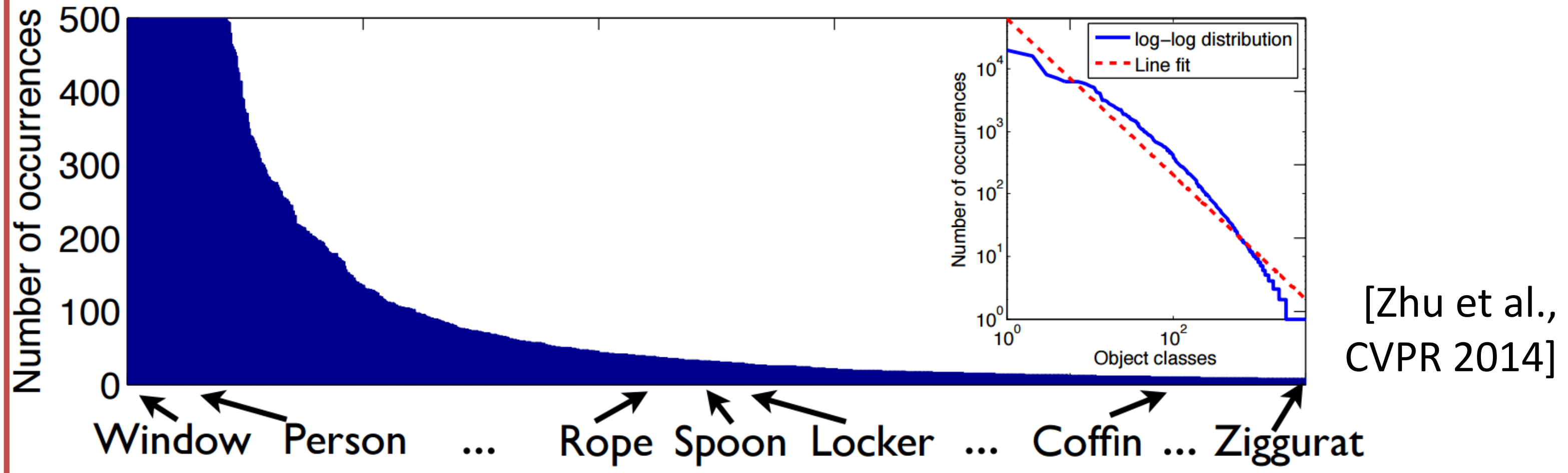
B.G. is supported by NSF IIS-1566511. Others are supported by USC Provost and Annenberg Graduate Fellowship, NSF IIS-1065243, 1451412, 1513966, 1208500, CCF-1139148, a Google Research Award, an Alfred. P. Sloan Research Fellowship and ARO # W911NF-12-1-0241 and W911NF-15-1-0484.

Highlights

- **Propose to align the *semantic space* to the *visual model space* via manifold learning**
- **Introduce and learn *"phantom"* classes to bridge the two spaces**
- **Attain state-of-the-art results on four benchmark datasets, including the full *ImageNet Fall 2011* with over 20,000 unseen classes**

Introduction

- **Ultimate Goal:** recognize objects in the wild
- **Challenges:** long-tailed distribution (SUN dataset)



[Zhu et al., CVPR 2014]

- **Zero-shot learning:** expand classifiers and the labeling space from *seen* classes to *unseen* classes
  - *How to relate seen & unseen?*

semantic info. (e.g., attributes)

seen

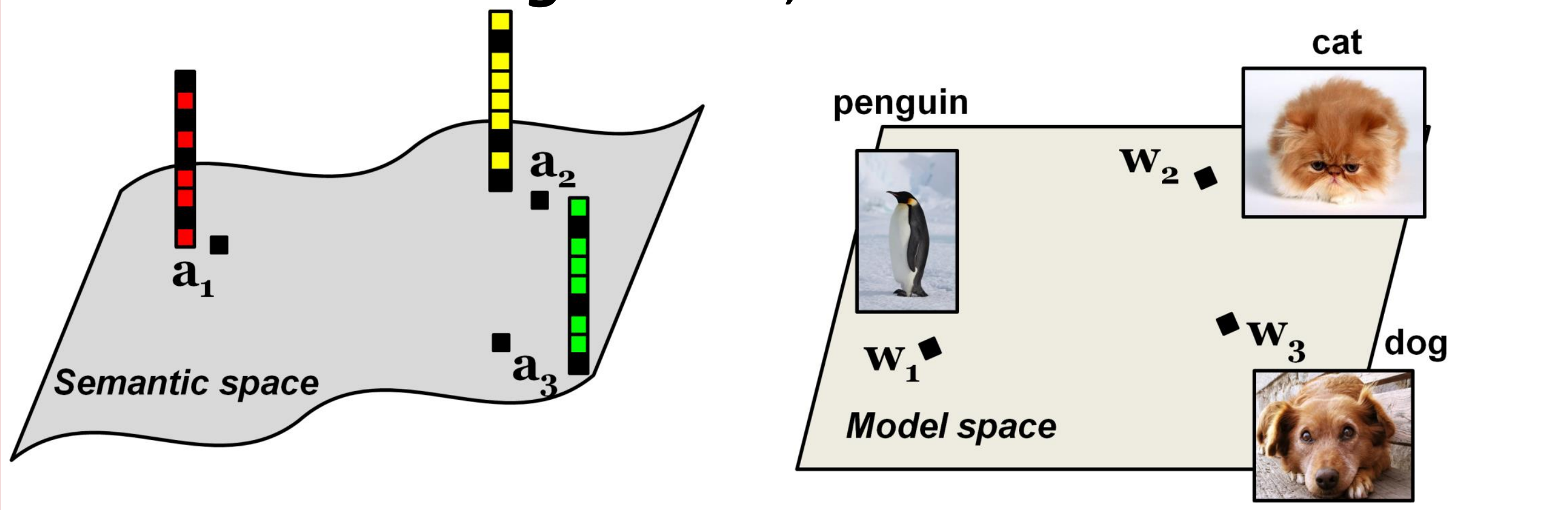
unseen

stripes mane snout stripes, mane, snout

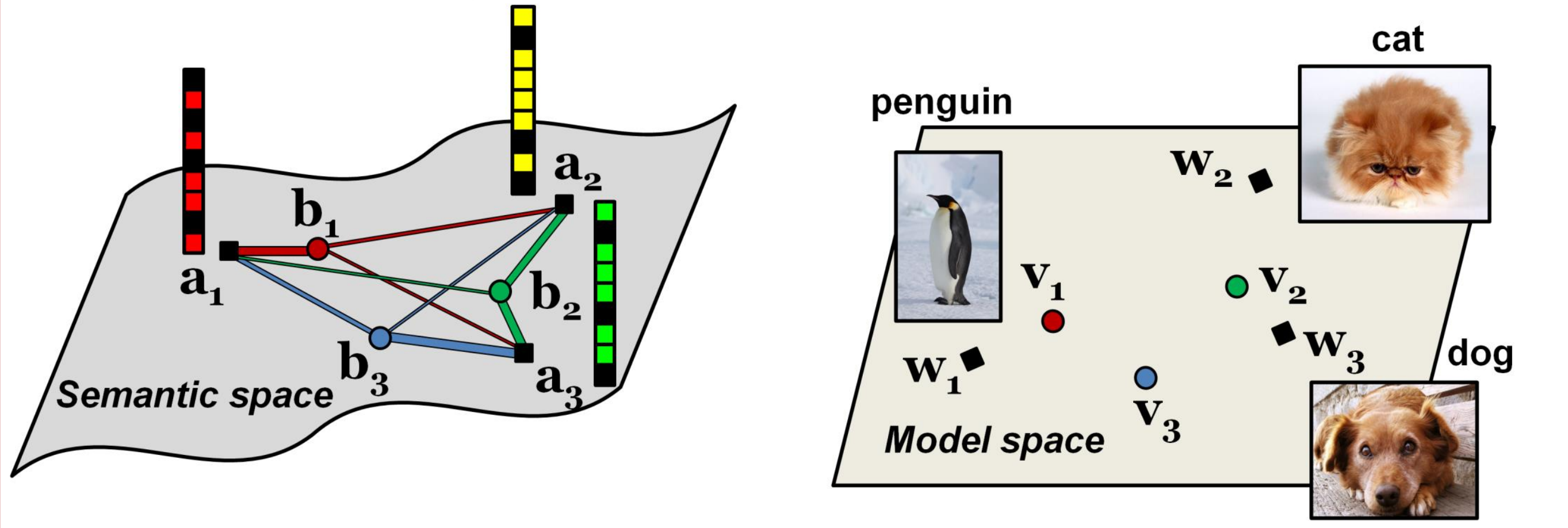
- *How to attain discriminative power?* ***our paper*** ☺

Approach

- Object classes live in both ***semantic*** and ***model*** spaces
  - If we can ***align*** them, we can construct the classifier for ***ANY*** class given its semantic info. (attributes, word vectors, etc.)

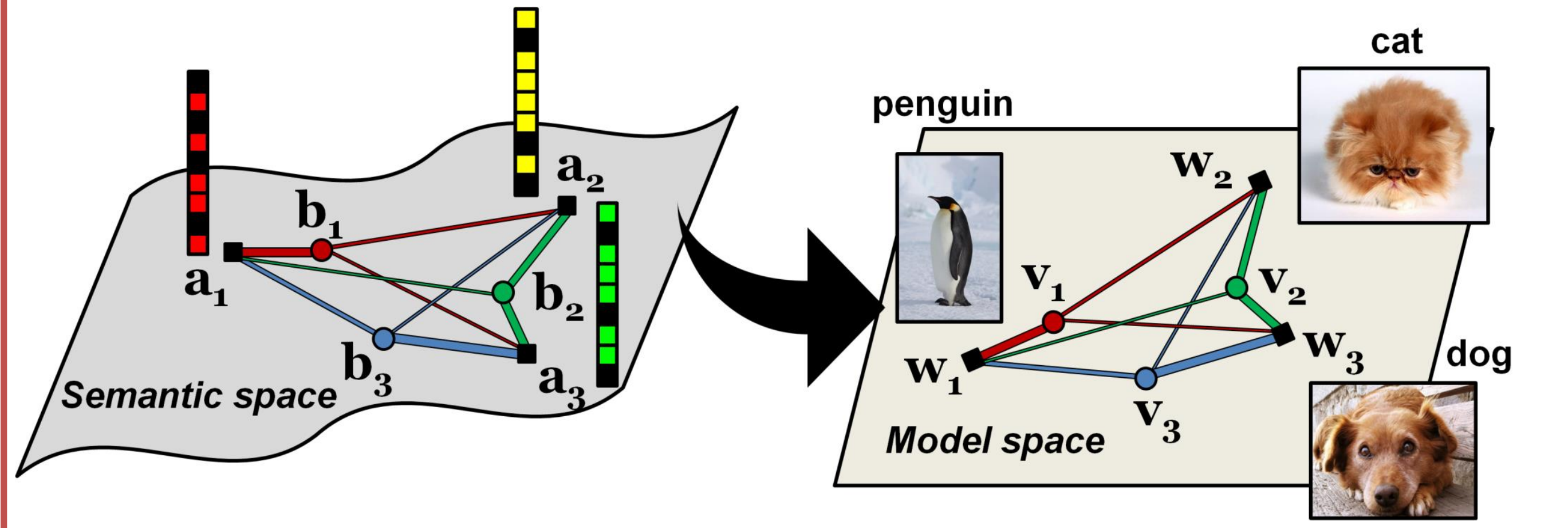


- Introduce ***"phantom classes"*** with coordinates {***b***, ***v***} in both spaces



Weighted graph $s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r'=1}^{R}\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_{r'})\}}$

- View the model space as the ***embedding*** of the weighted graph



Structure preserving

$$\min_{\boldsymbol{w}_c, \boldsymbol{v}_r} \left\| \boldsymbol{w}_c - \sum_{r=1}^{R} s_{cr} \boldsymbol{v}_r \right\|_2^2$$

- Classifier synthesis formula:

$$\boldsymbol{w}_c = \sum_{r=1}^{R} s_{cr} \boldsymbol{v}_r$$

- Learning the coordinates (i.e., ***b*** and ***v***) for optimal discrimination and generalization performance
- Class-wise cross validation: simulating zero-shot learning on training set for model selection

Experiments

- Datasets

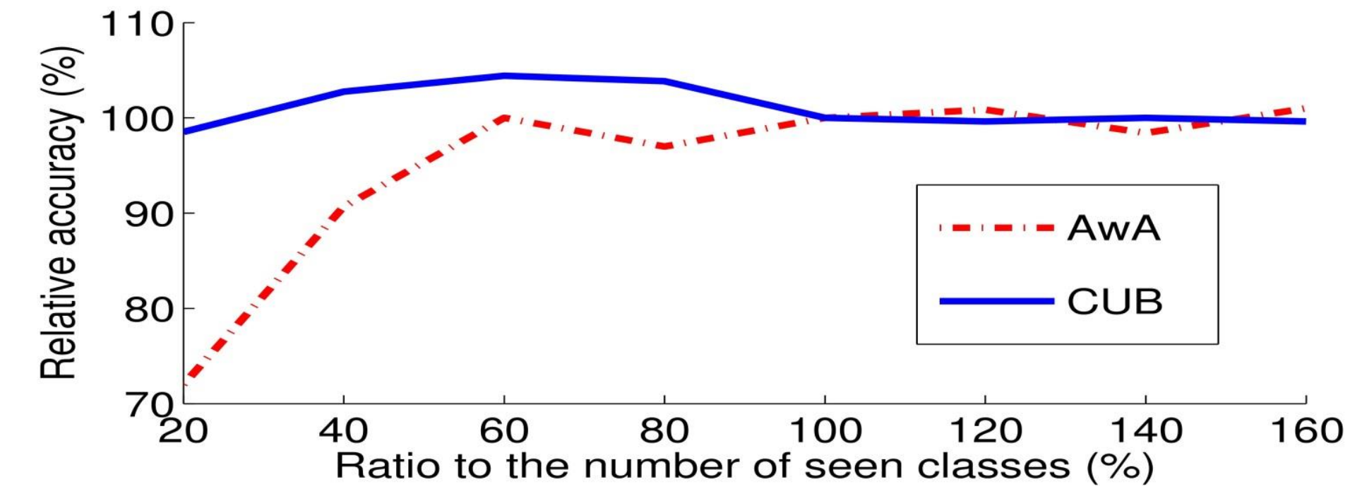
	AwA (animals)	CUB (birds)	SUN (scenes)	ImageNet
# of seen classes	40	150	645/646	1,000
# of unseen classes	10	50	72/71	20,842
Total # of images	30,475	11,788	14,340	14,197,122

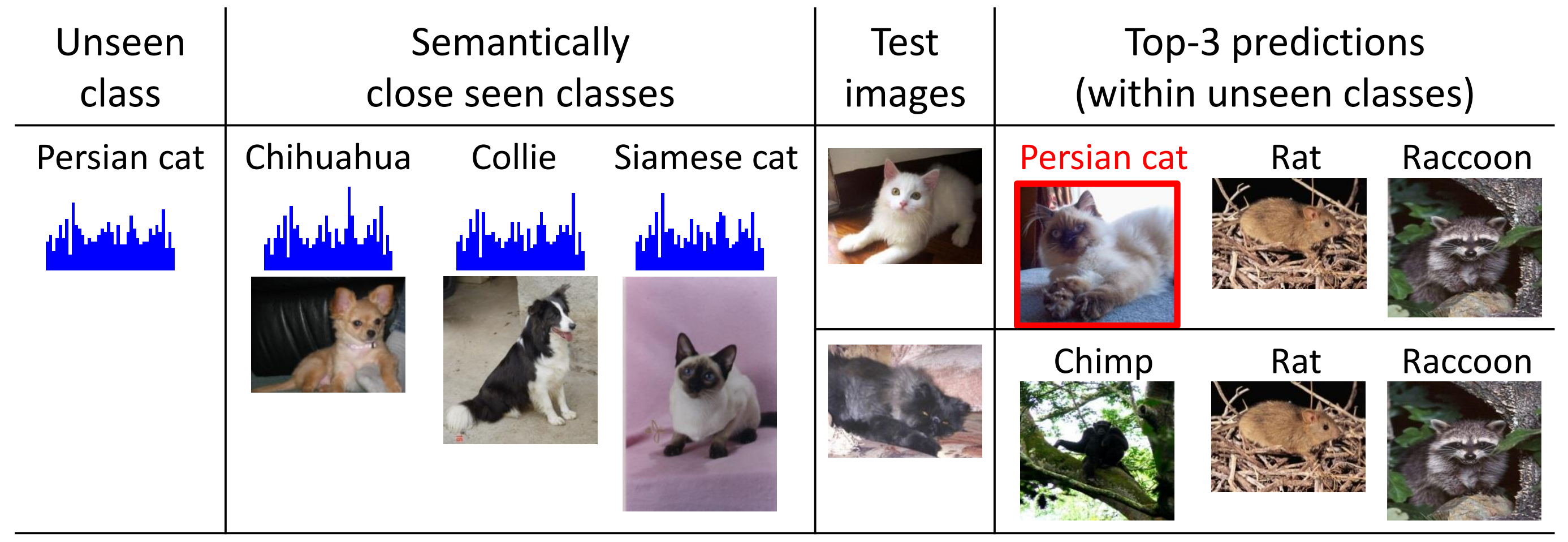
- Semantic space: attributes (85/312/102 for AwA/CUB/SUN), word2vec (500-dim for ImageNet)
- Visual features: 1,024-dim GoogLeNet features
- Evaluation: Top-K (Flat Hit@K) classification accuracy **among unseen classes**

[Top-1 results on AwA/CUB/SUN]

Methods	AwA	CUB	SUN
DAP [Lampert '14]	60.5	39.1	44.5
SJE [Akata '15]	66.7	50.1	56.1
ESZSL [Romera-Paredes '15]	64.5	44.0	18.7
ConSE [Norouzi '14]	63.3	36.2	51.9
COSTA [Mensink '14]	61.8	40.8	47.9
$\text{SynC}^{\text{o-vs-o}}$ (R, $\boldsymbol{b}_r$ fixed)	69.7	53.4	62.8
$\text{SynC}^{\text{struct}}$ (R, $\boldsymbol{b}_r$ fixed)	72.9	54.5	62.7
$\text{SynC}^{\text{o-vs-o}}$ (R fixed, $\boldsymbol{b}_r$ learned)	71.1	54.2	63.3

[Varying the number of phantom classes *R*]

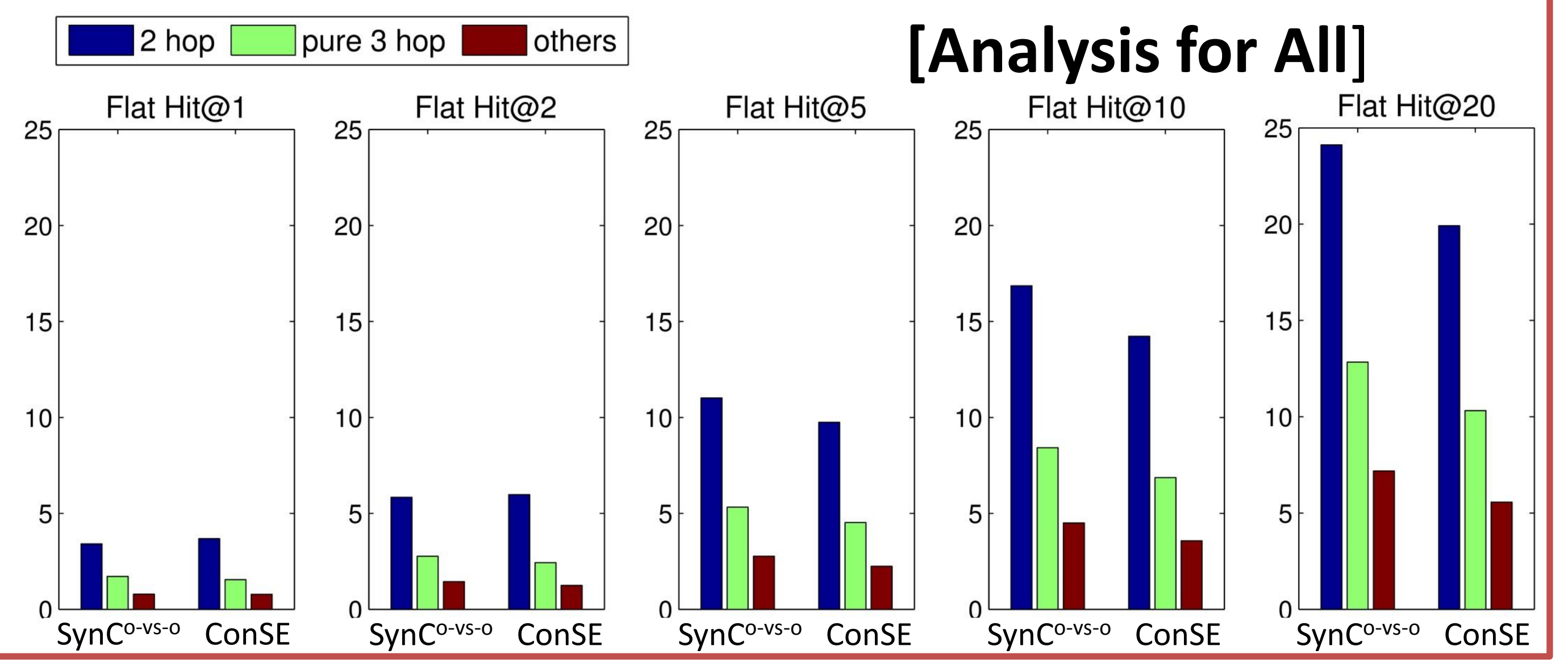




[Large-scale ZSL on ImageNet]

Scenarios	Methods	Top-1	Top-5	Top-10
2-hop (1,509)	ConSE	8.3	21.8	30.9
	$\text{SynC}^{\text{o-vs-o}}$	10.5	28.6	40.1
All (20,345)	ConSE	1.3	3.8	5.8
	$\text{SynC}^{\text{o-vs-o}}$	1.4	4.5	7.1

[Analysis for All]


2 hop
pure 3 hop
others
Flat Hit@1
Flat Hit@2
Flat Hit@5
Flat Hit@10
Flat Hit@20
SynC^o-vs-o
ConSE