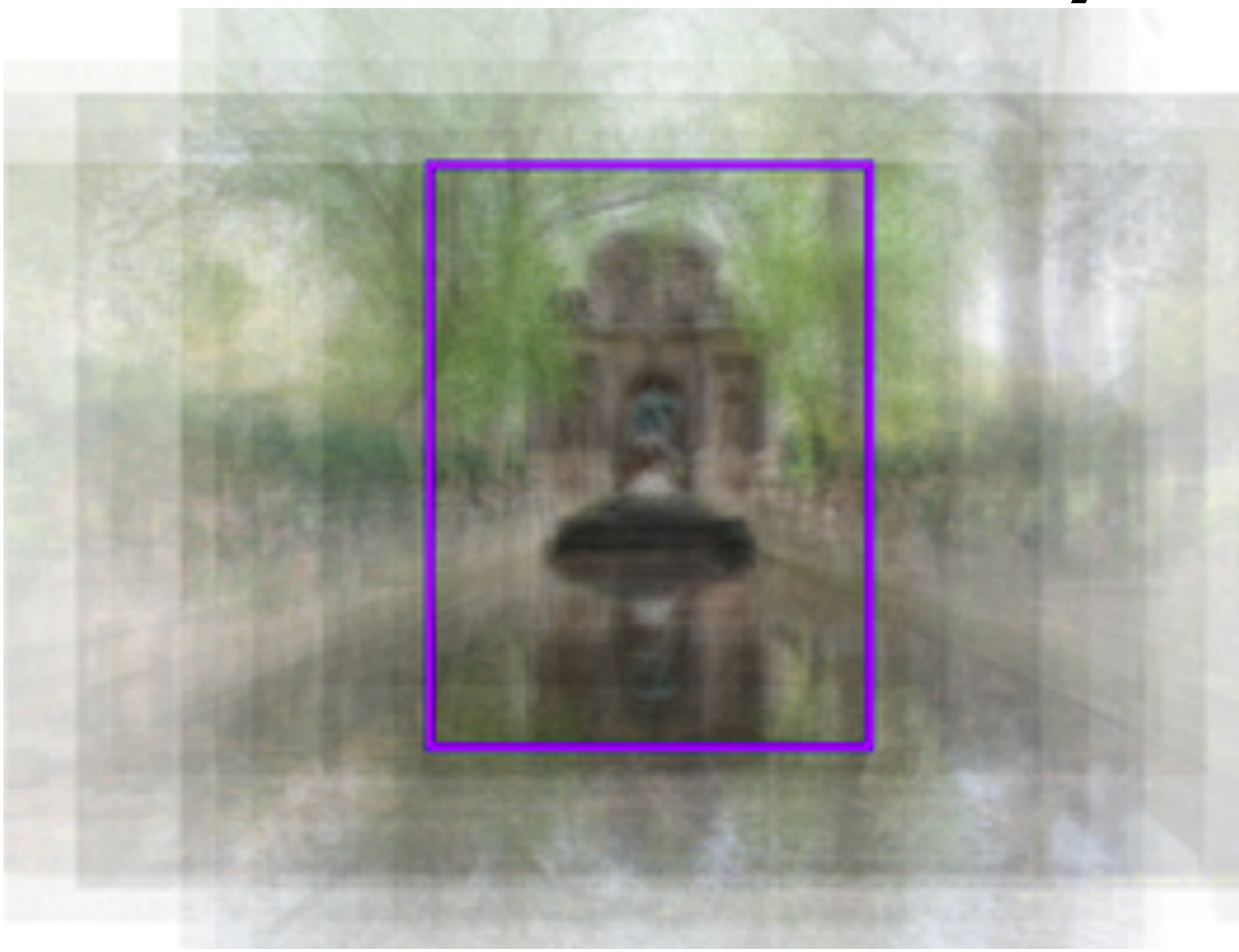
Per-exemplar learning: Object Detection and Beyond

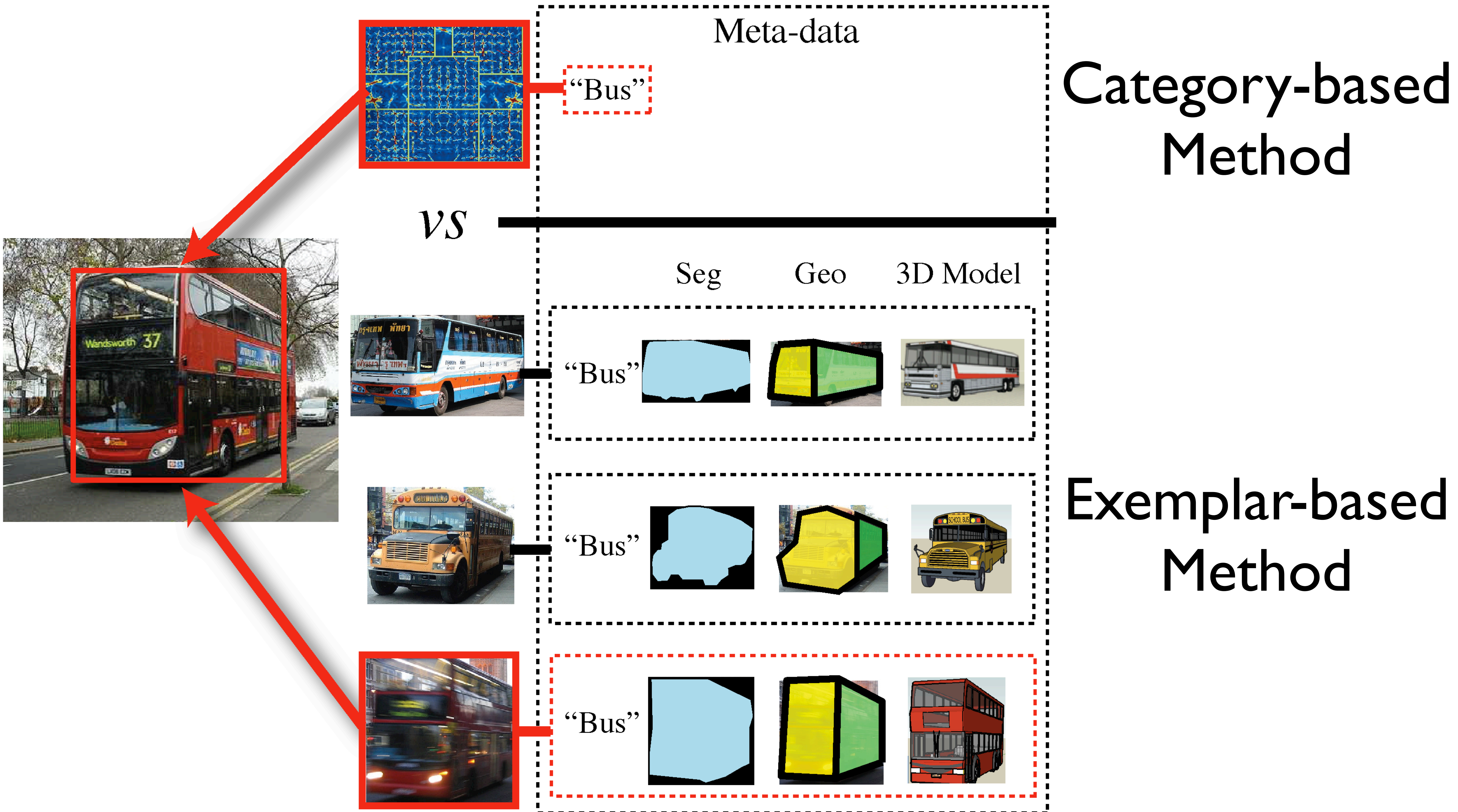


CarnegieMellon
THE ROBOTICS INSTITUTE

Tomasz Malisiewicz
tomasz@csail.mit.edu

Workshop on Kernels and Distances @ICCV 2011 Barcelona, Spain

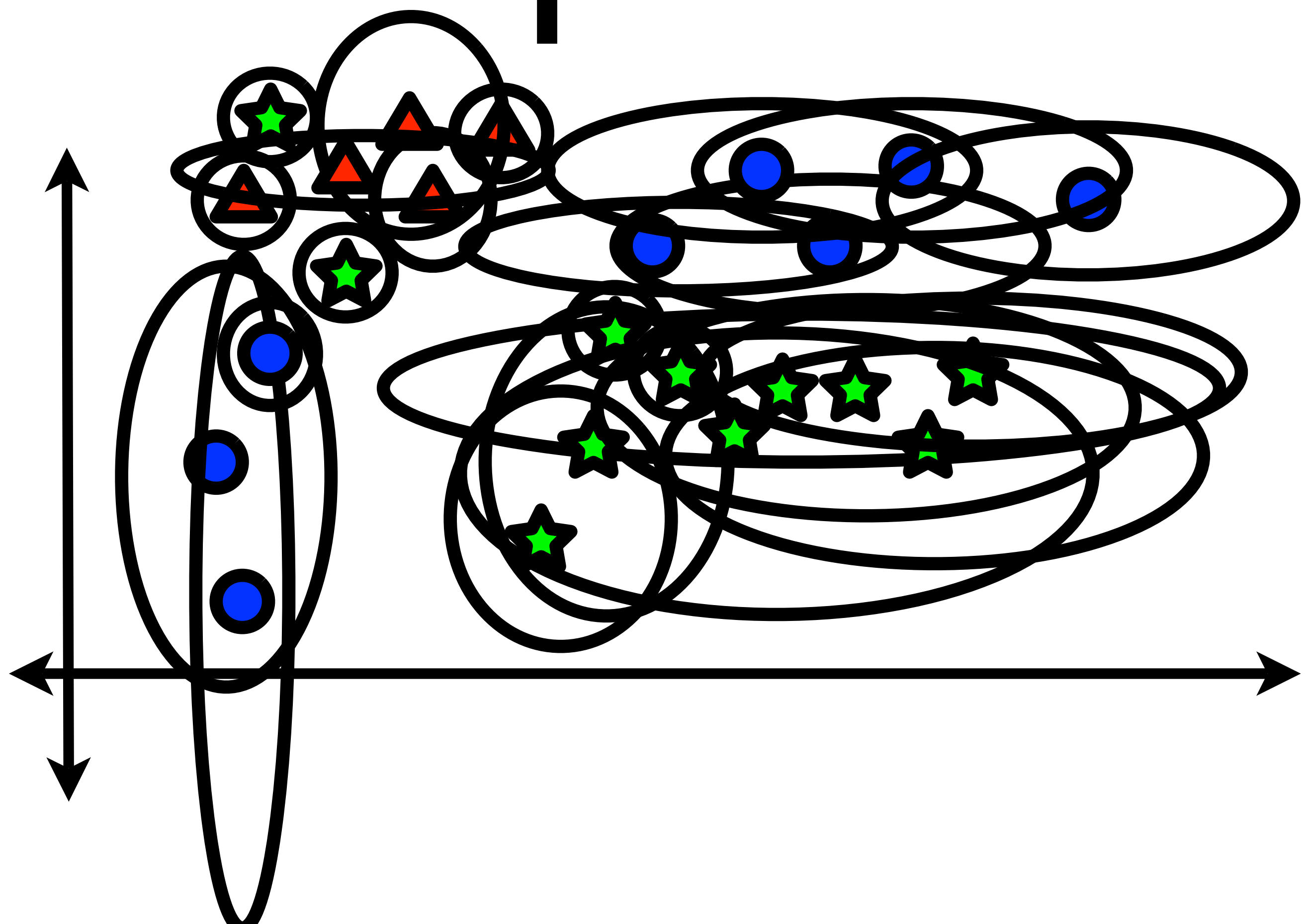
Why Nearest Neighbors Matter



Overview

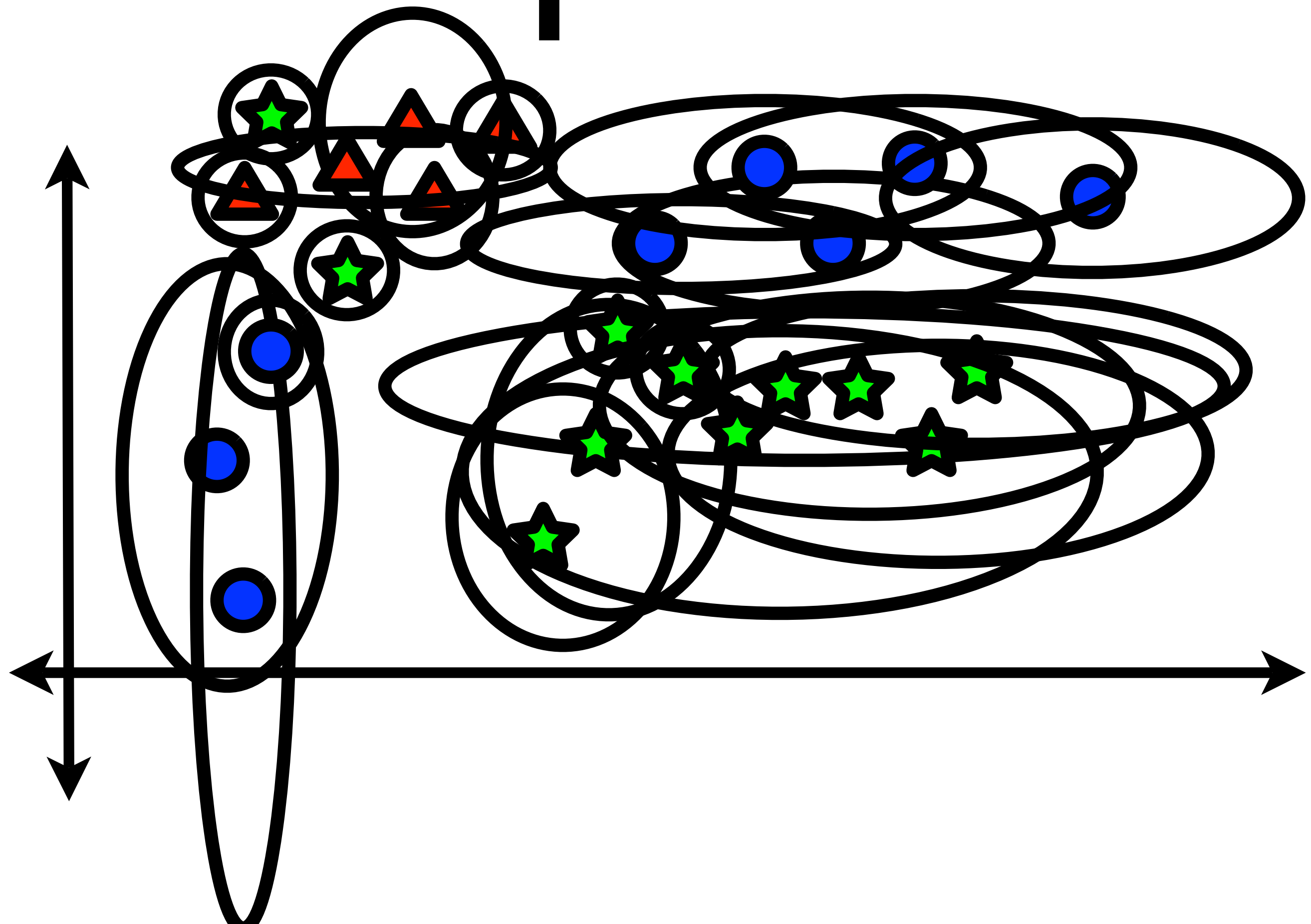
- Learning Per-exemplar Distance Functions
- **ExemplarSVMs**: Coping with large scale detection problems
  - PASCAL VOC Object Detection and Meta-data Transfer
  - Cross-domain Image Matching
- Concluding Remarks and Open Problems

Per-Exemplar Learning



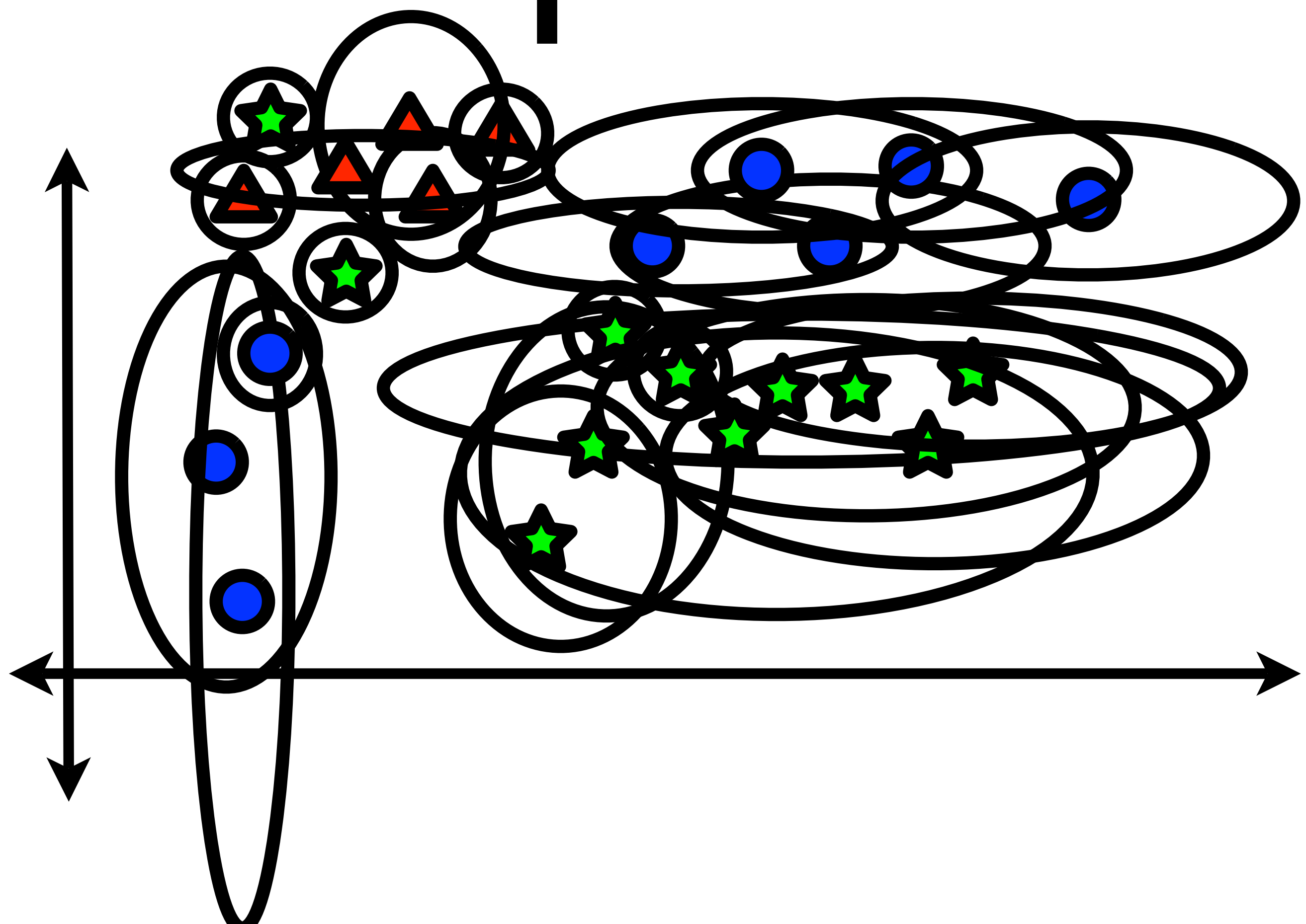
- NN-method, where each exemplar has its own distance "similarity" function

Per-Exemplar Learning



- NN-method, where each exemplar has its own distance "similarity" function
- Introduced for Image Classification by Frome et al., NIPS 2007

Per-Exemplar Learning



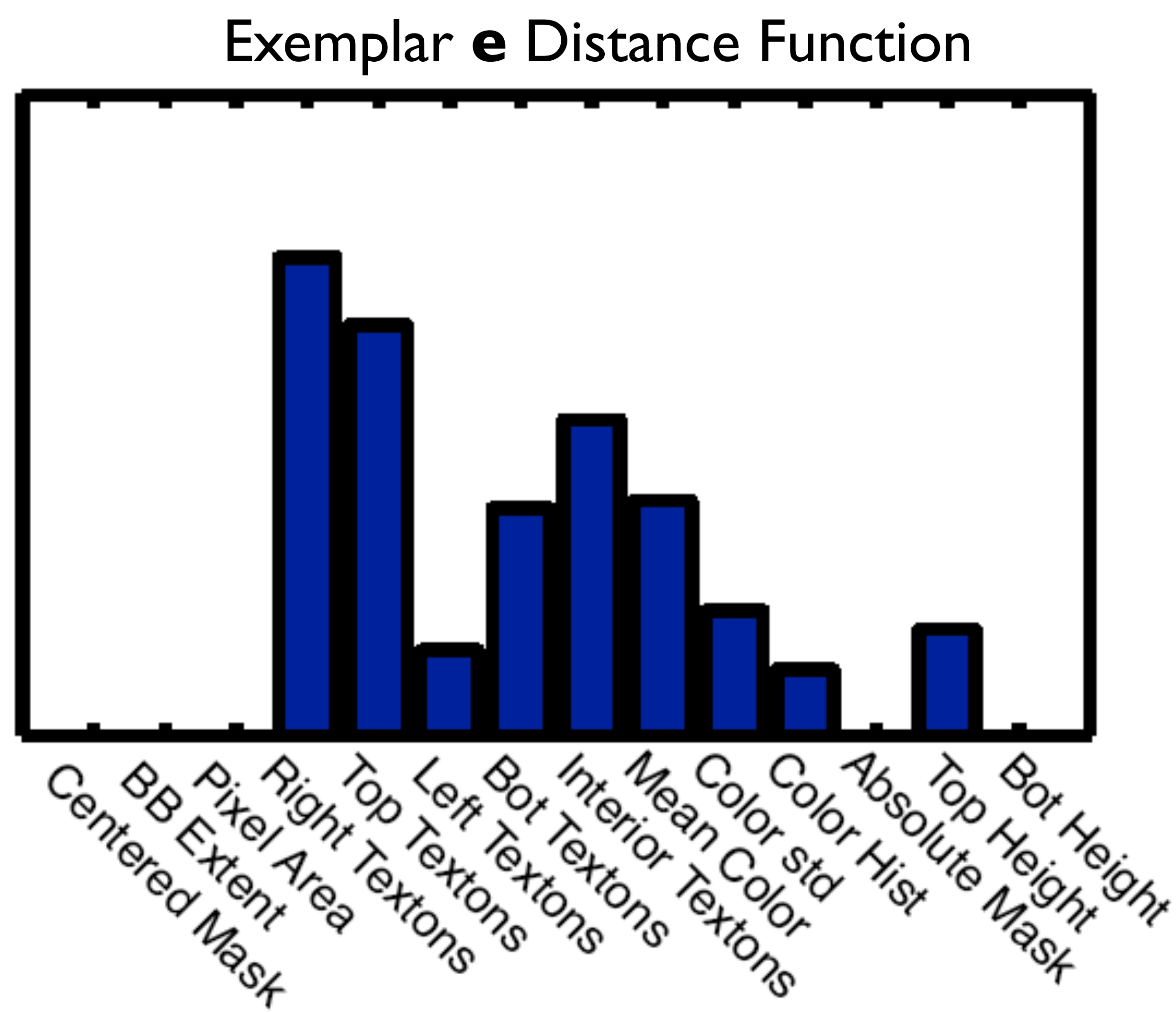
- NN-method, where each exemplar has its own distance "similarity" function
- Introduced for Image Classification by Frome et al., NIPS 2007
- Extended to Segmentation-based detection Malisiewicz et al., CVPR 2008

Per-Exemplar
Distance "Similarity" Functions

- Positive linear combination of elementary distances

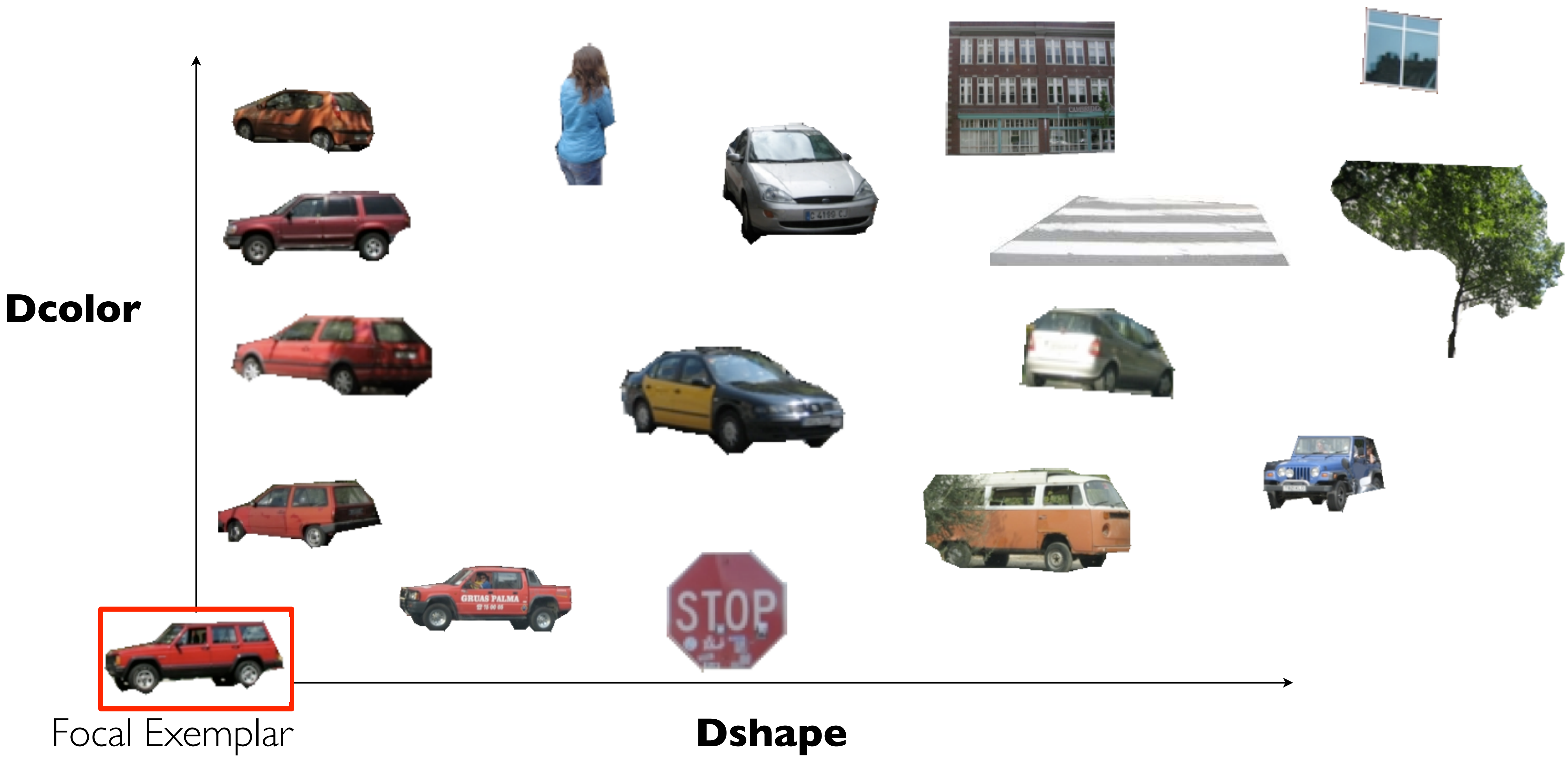
$$D_e(z) = \mathbf{w}_e \cdot \mathbf{d}_{ez}$$

Exemplar **e**

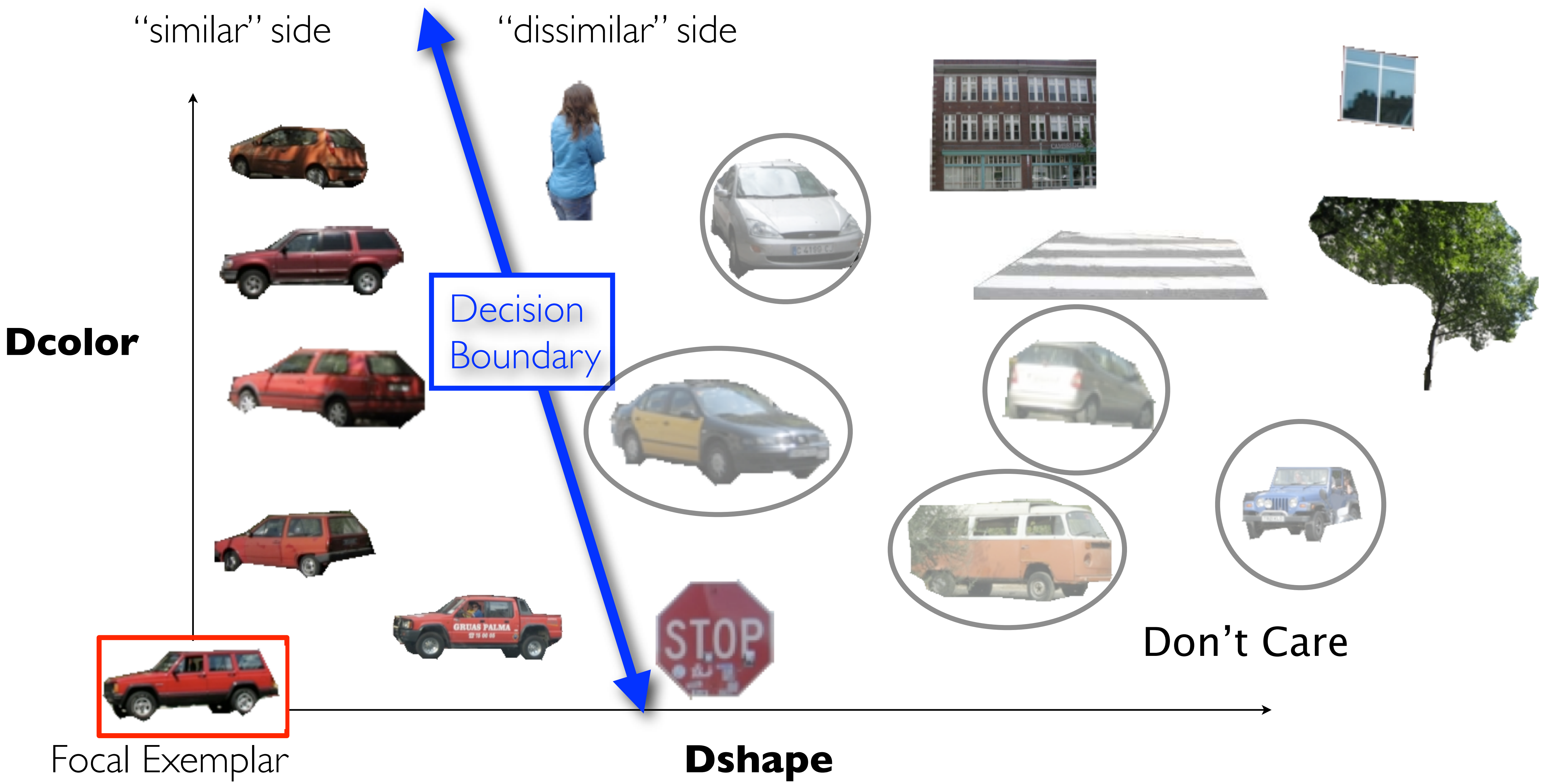


Frome et al. NIPS 2007, Malisiewicz et al. CVPR 2008

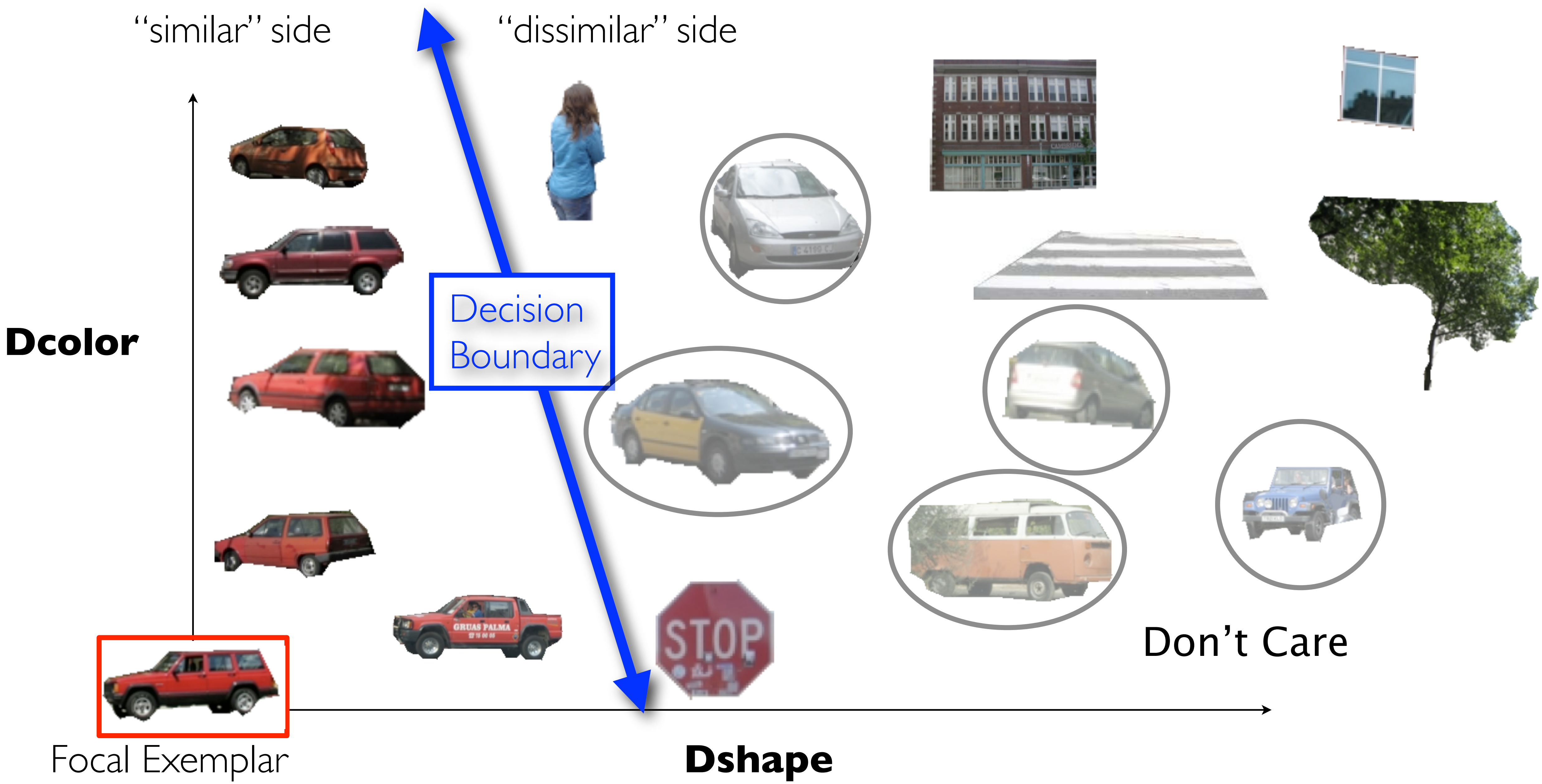
Learning Distance Function



Learning Distance Function

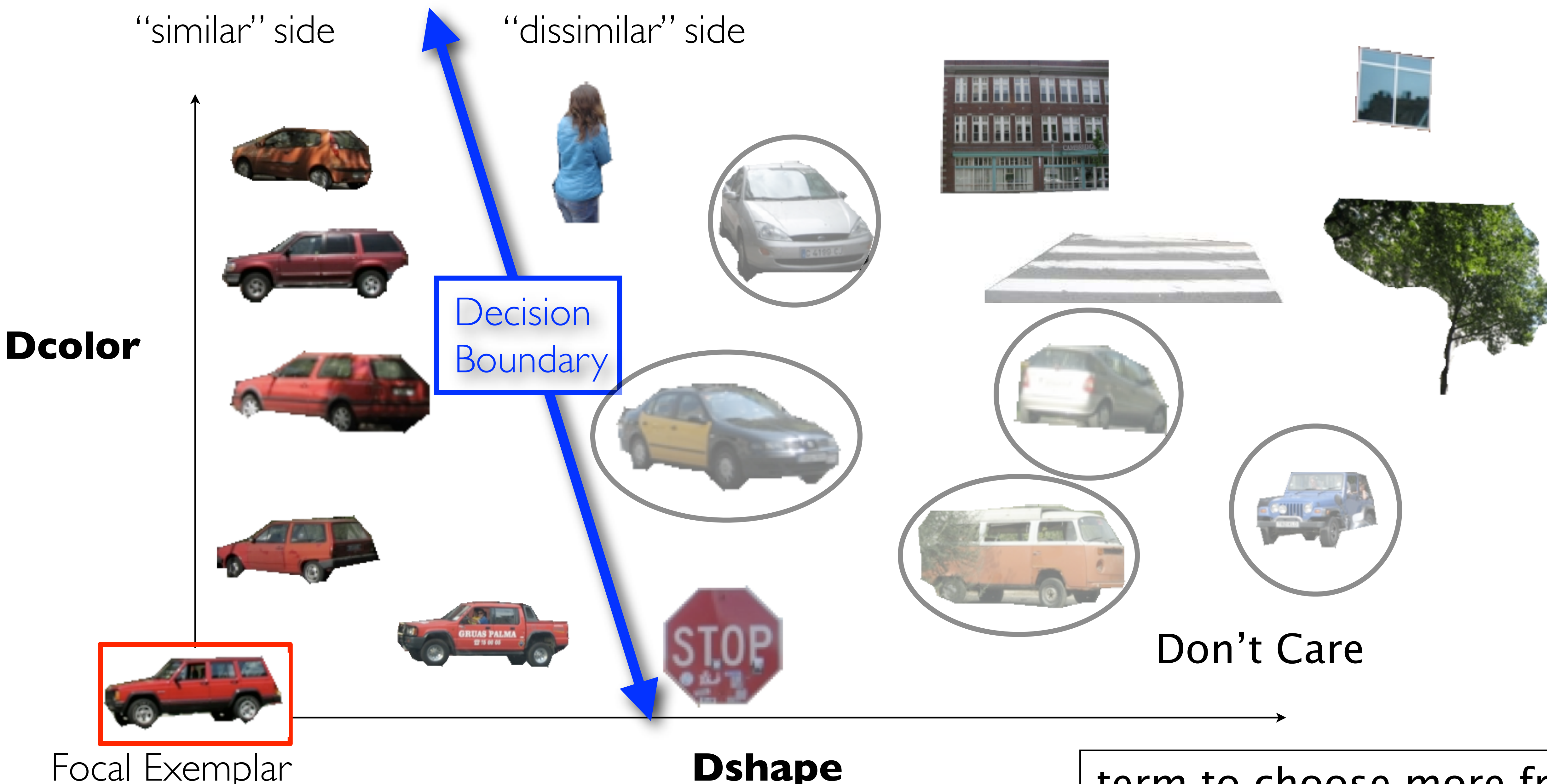


Learning Distance Function



$$f(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_{i \in C} \boldsymbol{\alpha}_i \, L(-(\mathbf{w}^T \mathbf{d}_i + b)) + \sum_{i \notin C} L(\mathbf{w}^T \mathbf{d}_i + b) - \sigma|| \, \boldsymbol{\alpha} \, ||^2$$

Learning Distance Function



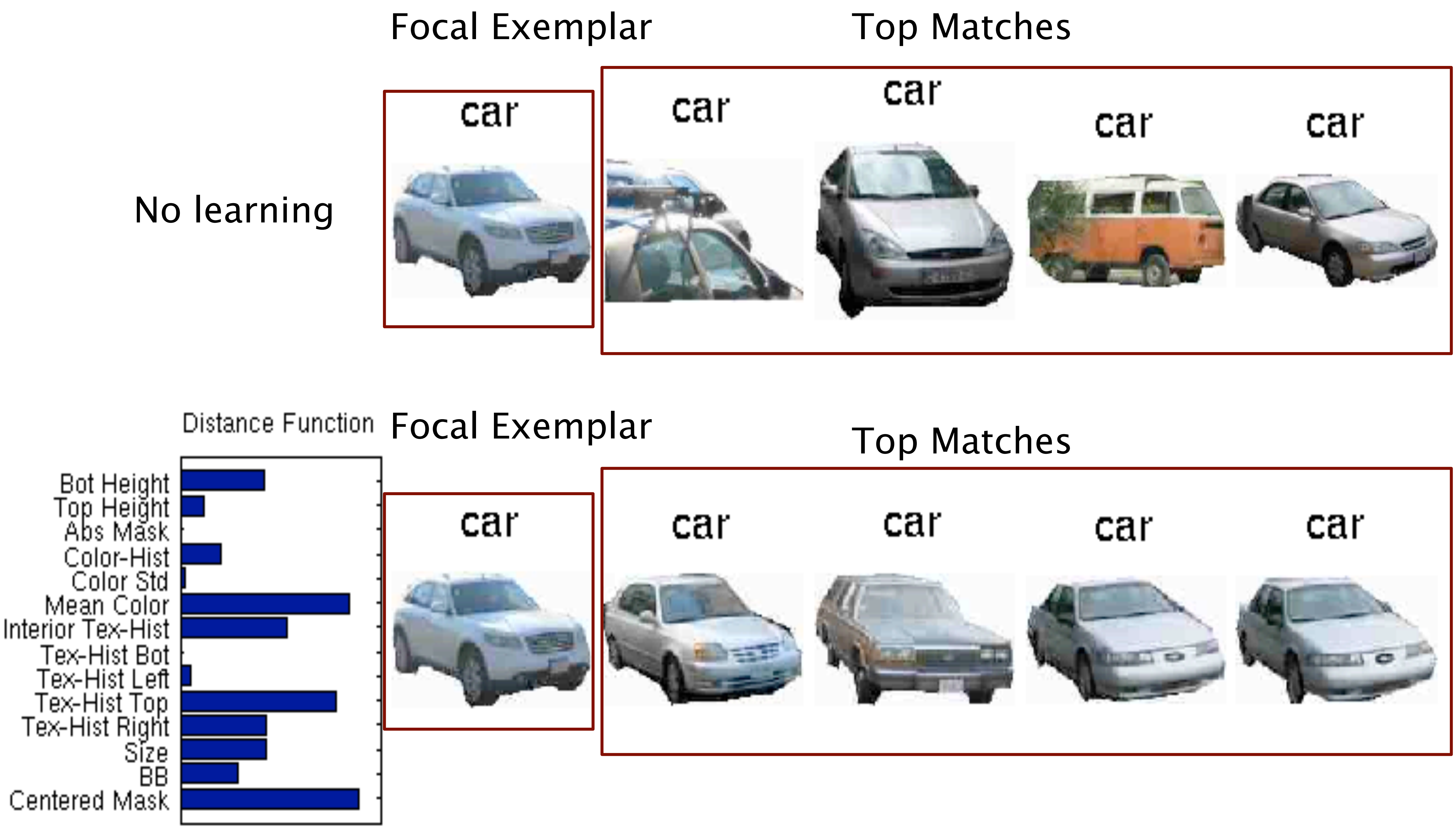
term to choose more friends

$$f(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_{i \in C} \boldsymbol{\alpha}_i L(-(\mathbf{w}^T \mathbf{d}_i + b)) + \sum_{i \notin C} L(\mathbf{w}^T \mathbf{d}_i + b) - \sigma || \boldsymbol{\alpha} ||^2$$

distance function

binary "friend" selection variable

A Learned Distance Function



Segment-then-recognize

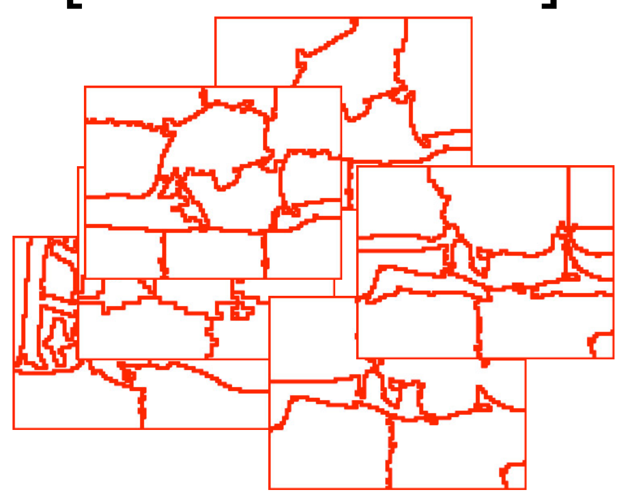
Input Image

Malisiewicz et al. CVPR 2008

Segment-then-recognize

Input Image

Multiple Segmentations
[Hoiem et al. 2005]



Malisiewicz et al. CVPR 2008

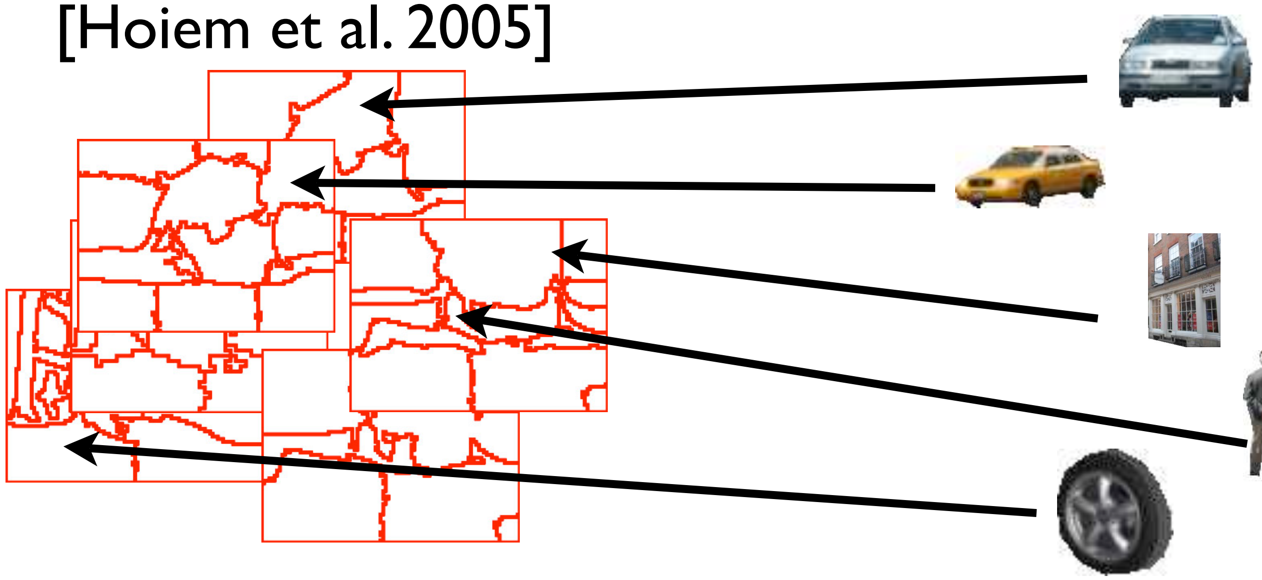
Segment-then-recognize

Input Image

Multiple Segmentations

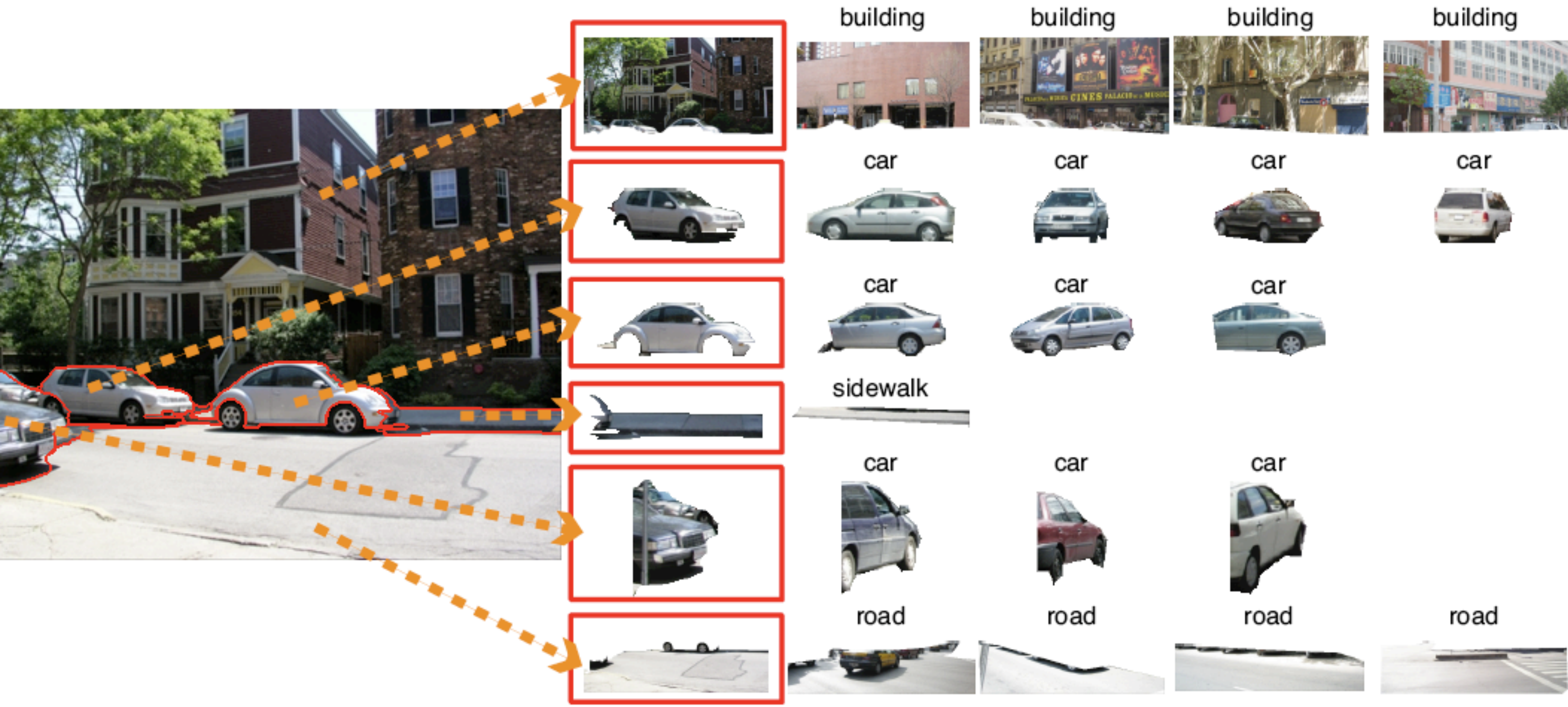
[Hoiem et al. 2005]

Exemplars + Distance Funs



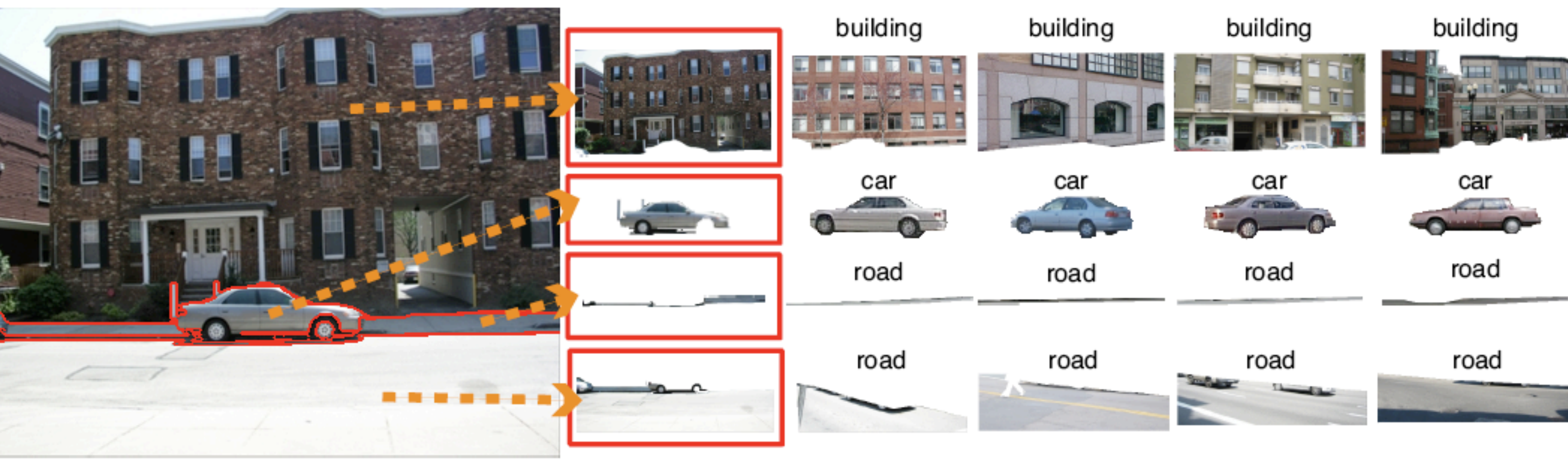
Malisiewicz et al. CVPR 2008

Segment-then-recognize Results



Malisiewicz et al. CVPR 2008

Segment-then-recognize Results



Malisiewicz et al. CVPR 2008

Limits of distance function learning

- Learning focuses on objects, but in **object detection** there are many more non-objects than objects

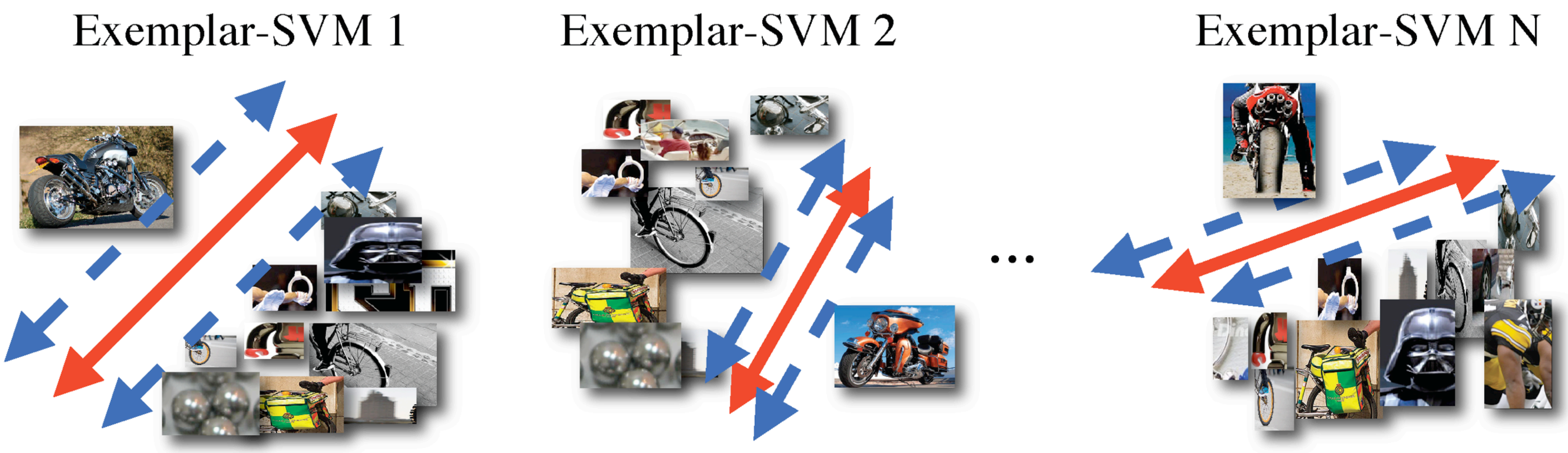
Limits of distance function learning

- Learning focuses on objects, but in **object detection** there are many more non-objects than objects
- Need to potentially cope with millions of negatives during learning

Limits of distance function learning

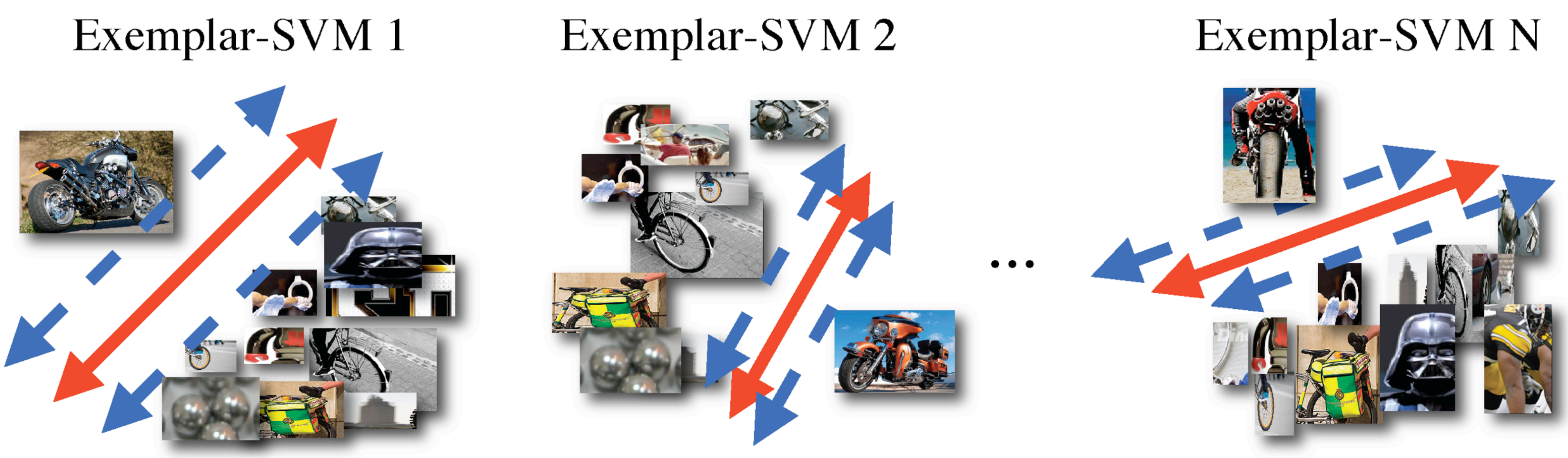
- Learning focuses on objects, but in **object detection** there are many more non-objects than objects
- Need to potentially cope with millions of negatives during learning
- State-of-the-art object detectors deal with negative data by **hard negative mining** [Dalal-Triggs 2005, Felzenszwalb et al. 2008]

Exemplar-SVMs



Malisiewicz et al. ICCV 2011

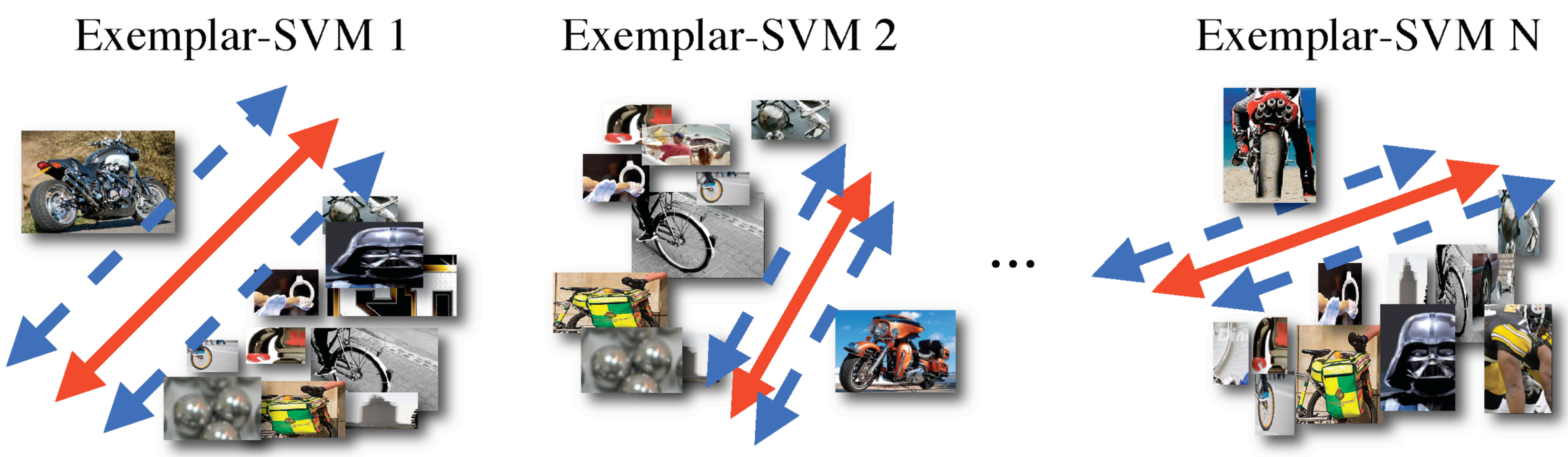
Exemplar-SVMs



- Best of both worlds:
    - Effectiveness of discriminatively-trained object detectors
    - Explicit correspondence of Nearest Neighbor approaches

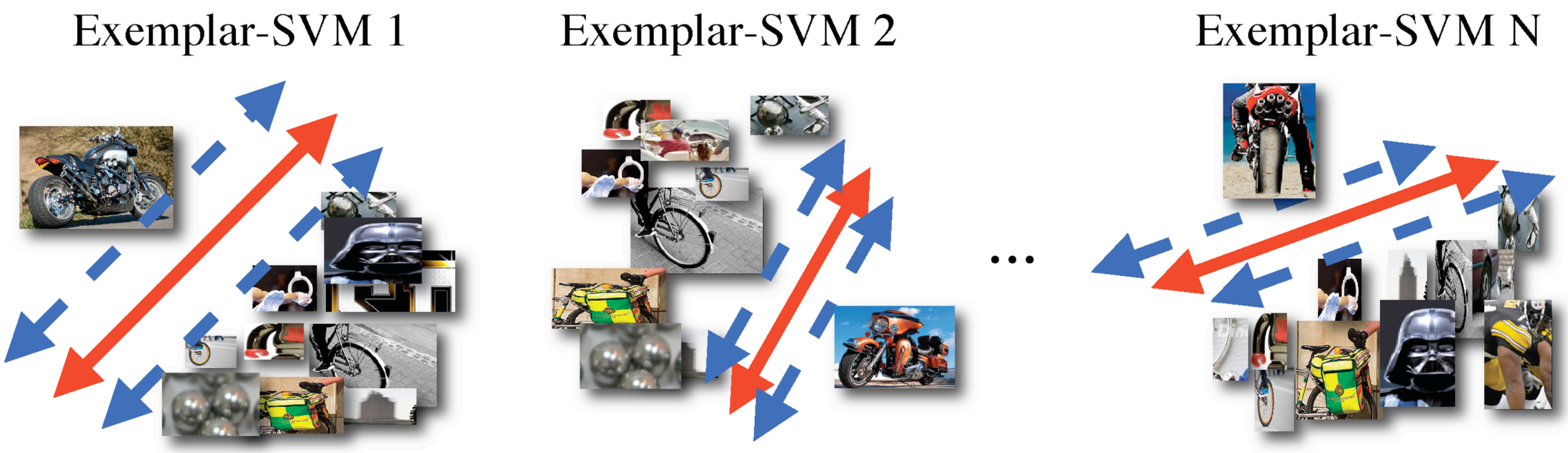
Malisiewicz et al. ICCV 2011

Exemplar-SVMs



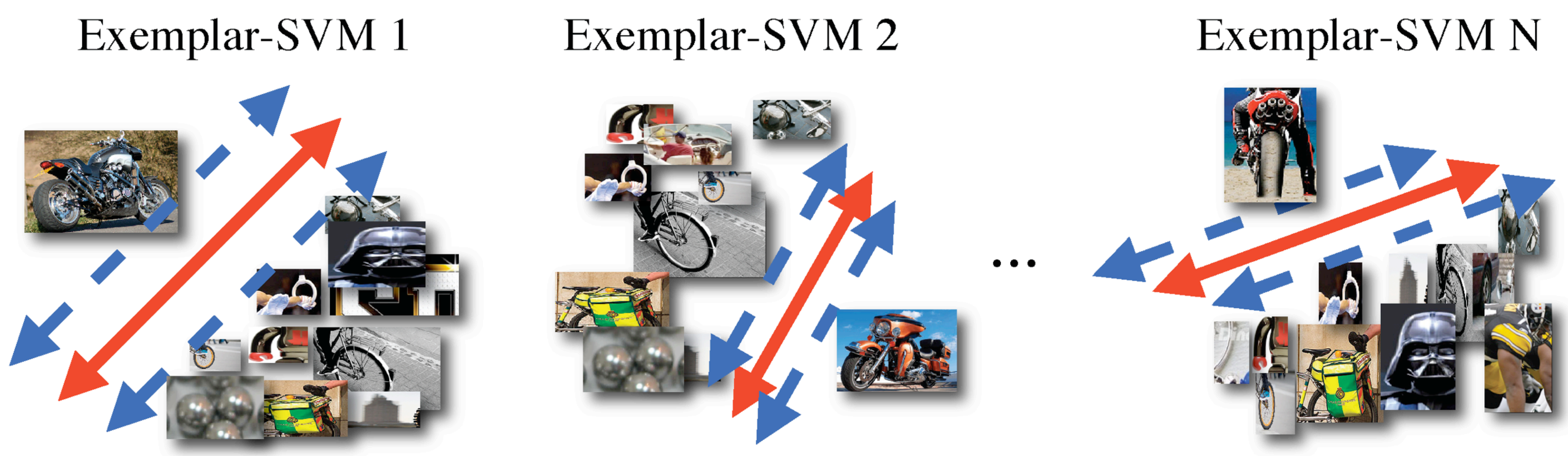
- Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)

Exemplar-SVMs



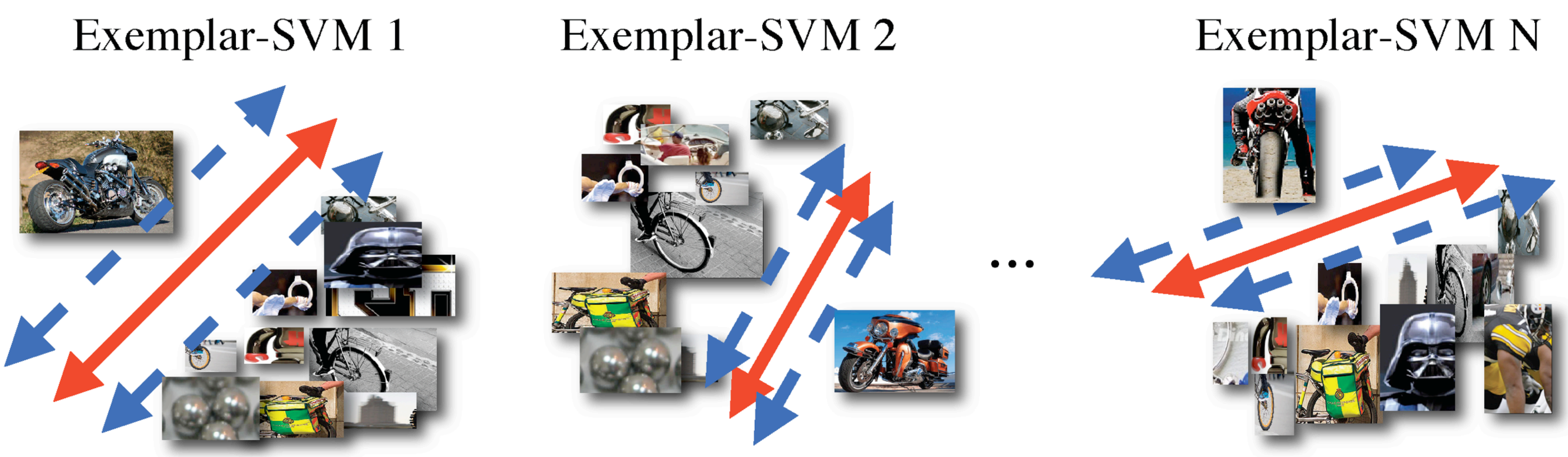
- Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)
- Each Exemplar-SVM is trained with a **single** positive instance and **millions** of negatives

Exemplar-SVMs

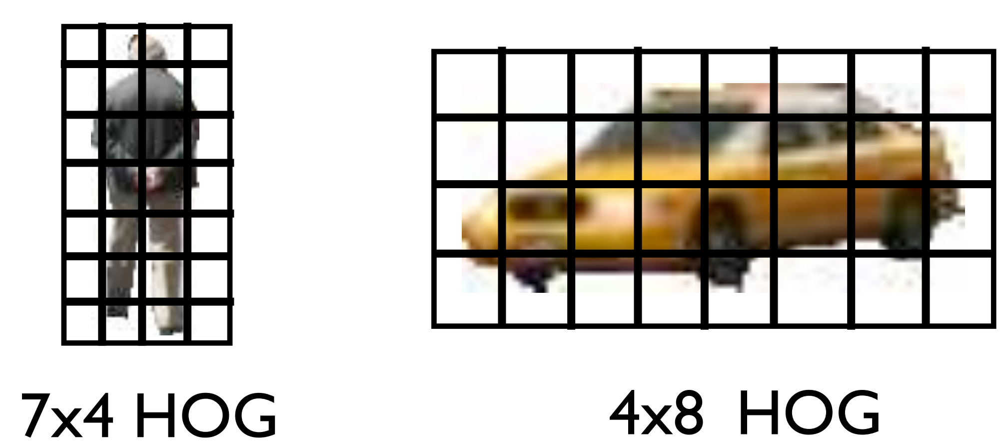


- Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)
- Each Exemplar-SVM is trained with a **single** positive instance and **millions** of negatives
- Each Exemplar-SVM is more defined by "*what it is not*" vs. "*what it is similar to*"

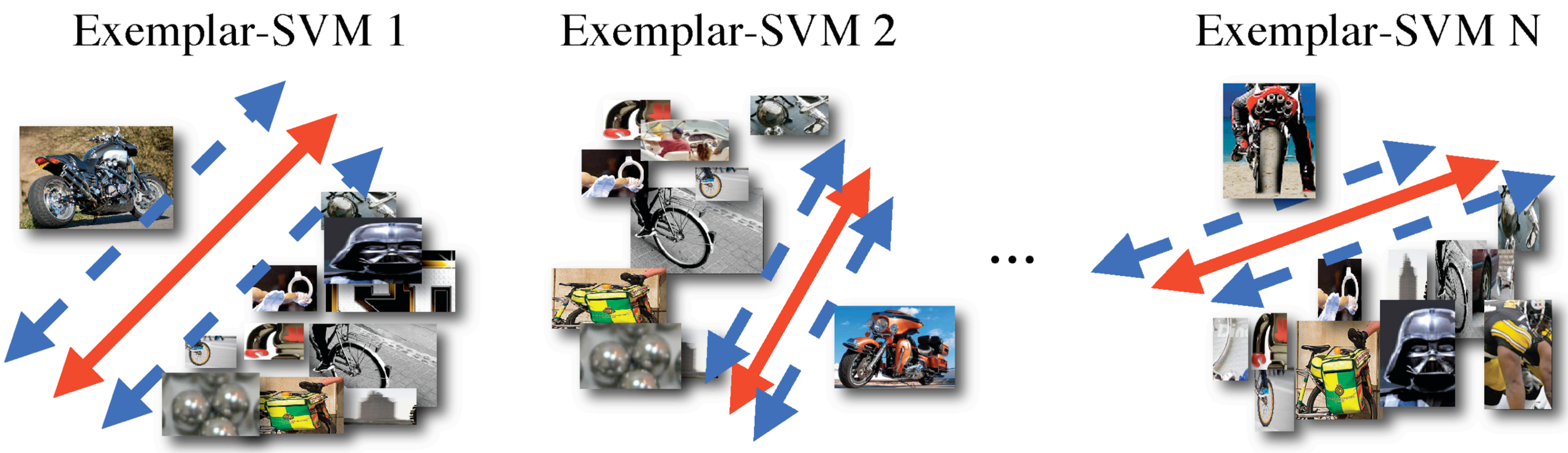
Exemplar-SVMs



- Because each Exemplar-SVM is defined by a **single** positive instance, we can use different features for each exemplar

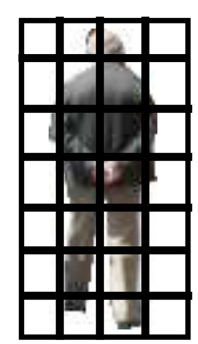


Exemplar-SVMs



- Because each Exemplar-SVM is defined by a **single** positive instance, we can use different features for each exemplar

- Apply each Exemplar-SVM to test image in a sliding-window fashion



7x4 HOG

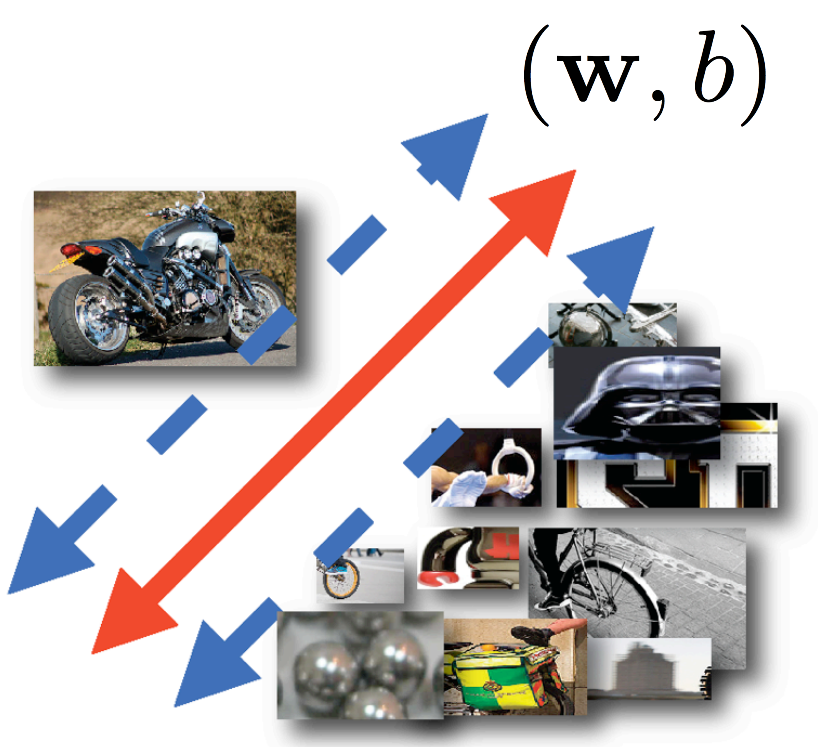
4x8 HOG

Exemplar-SVMs

Exemplar E's Objective Function:

$$\Omega_E(\mathbf{w}, b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T \mathbf{x}_E + b) + C_2 \sum_{\mathbf{x} \in \mathcal{N}_E} h(-\mathbf{w}^T \mathbf{x} - b)$$

h(x) = max(1-x,0) “hinge-loss”

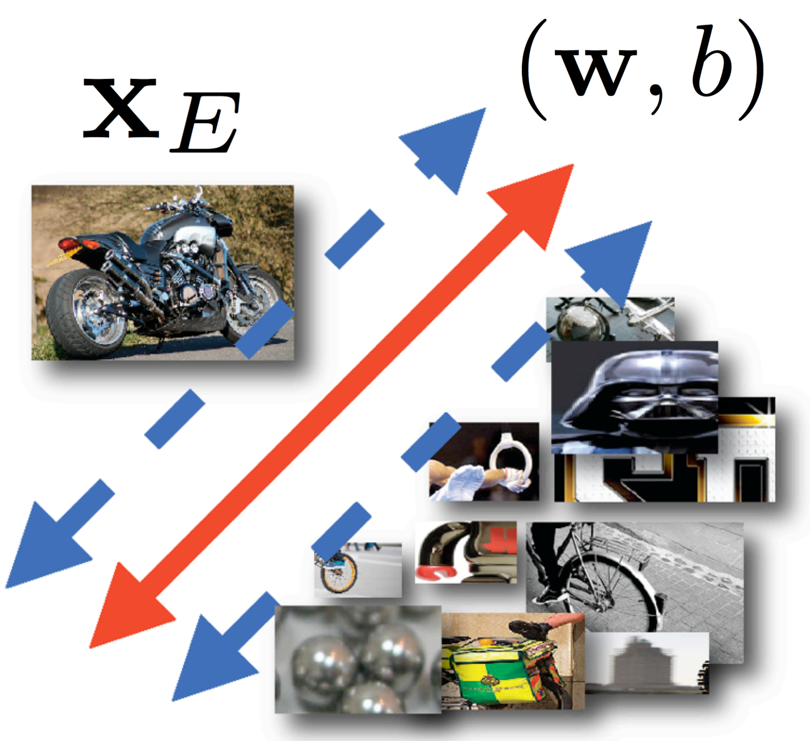


Exemplar-SVMs

Exemplar E's Objective Function:

$$\Omega_E(\mathbf{w},b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T\mathbf{x}_E + b) + C_2 \sum_{\mathbf{x}\in\mathcal{N}_E} h(-\mathbf{w}^T\mathbf{x} - b)$$

h(x) = max(1-x,0) “hinge-loss”



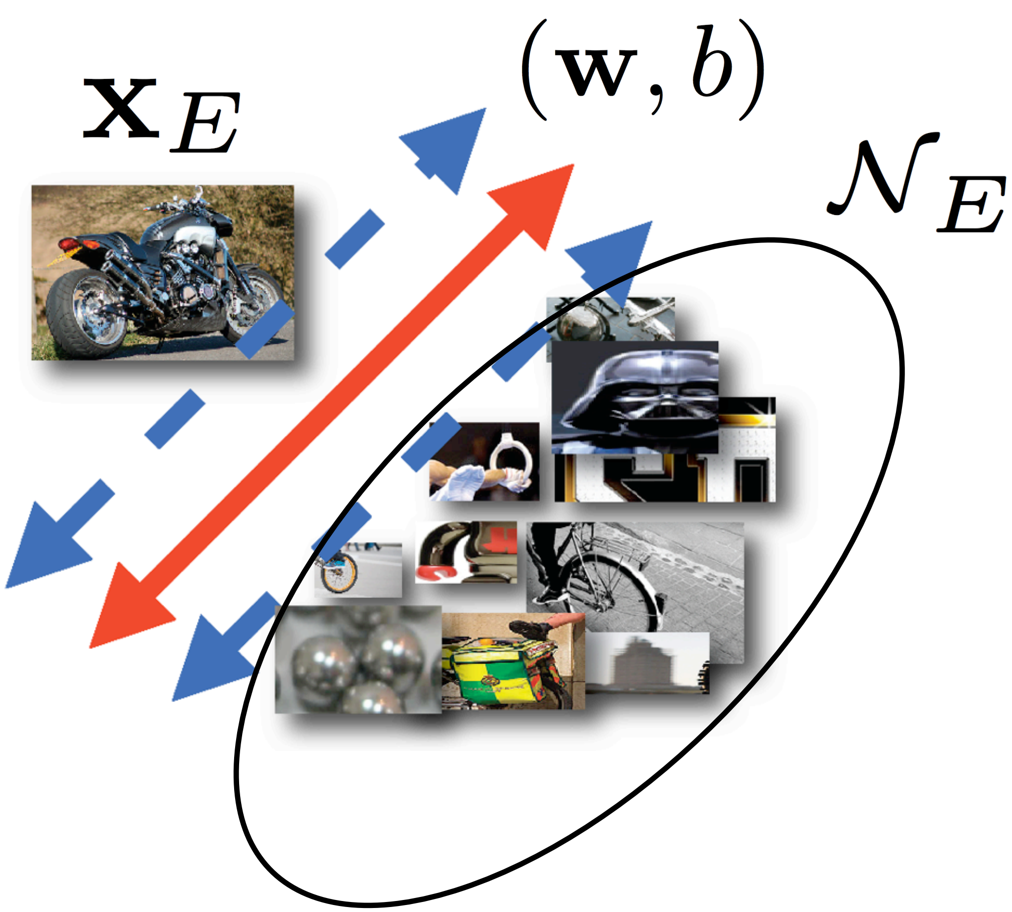
$\mathbf{x}_E$ Exemplar represented by ~100 HOG Cells (~3,100 features)

Exemplar-SVMs

Exemplar E's Objective Function:

$$\Omega_E(\mathbf{w}, b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T \mathbf{x}_E + b) + C_2 \sum_{\mathbf{x} \in \mathcal{N}_E} h(-\mathbf{w}^T \mathbf{x} - b)$$

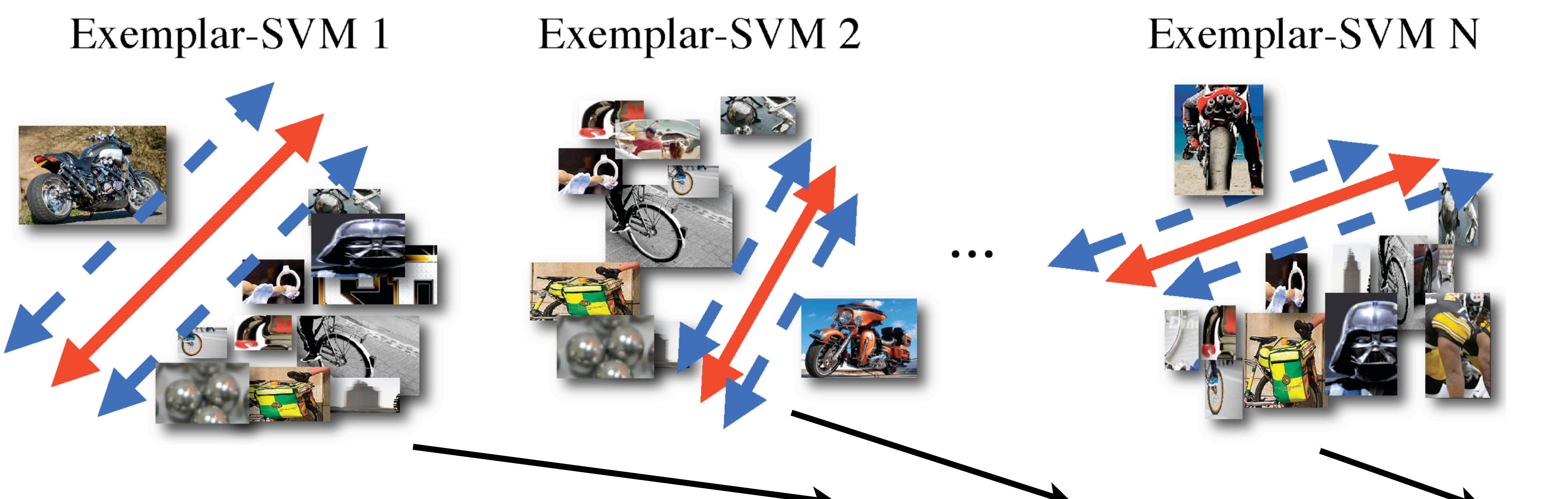
h(x) = max(1-x,0) "hinge-loss"

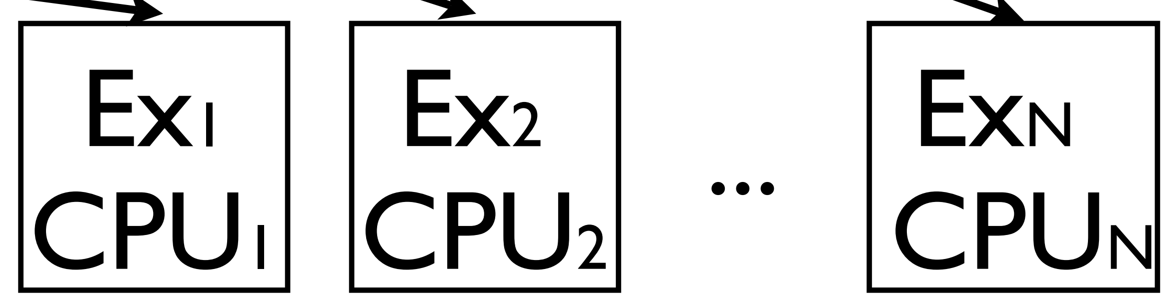


$\mathbf{x}_E$ Exemplar represented by ~100 HOG Cells (~3,100 features)

$\mathcal{N}_E$ Windows from images not containing any in-class instances (~2,000 images x ~10,000 windows/image = ~2M negatives)

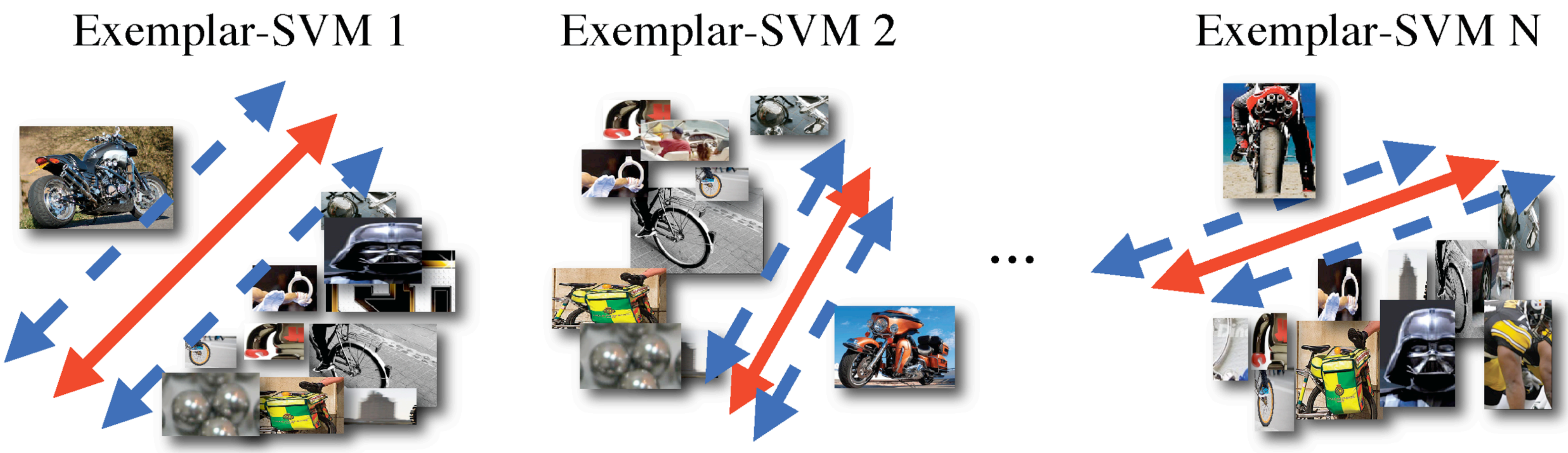
Large-scale training





- Each exemplar performs its own hard negative mining
- Solve many convex learning problems
- Parallel training on cluster

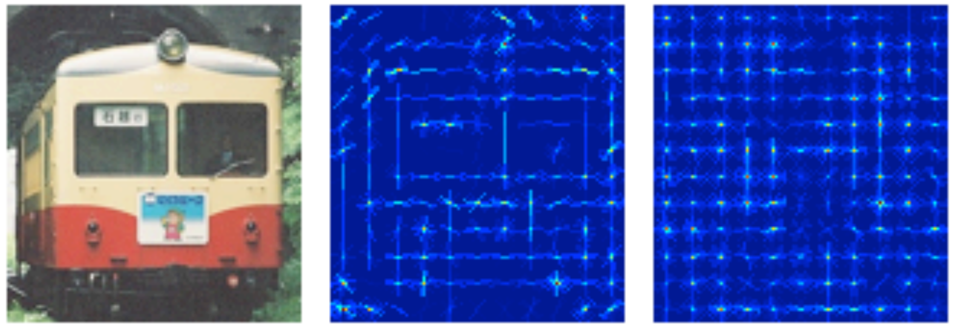
Interpreting Exemplar-SVMs



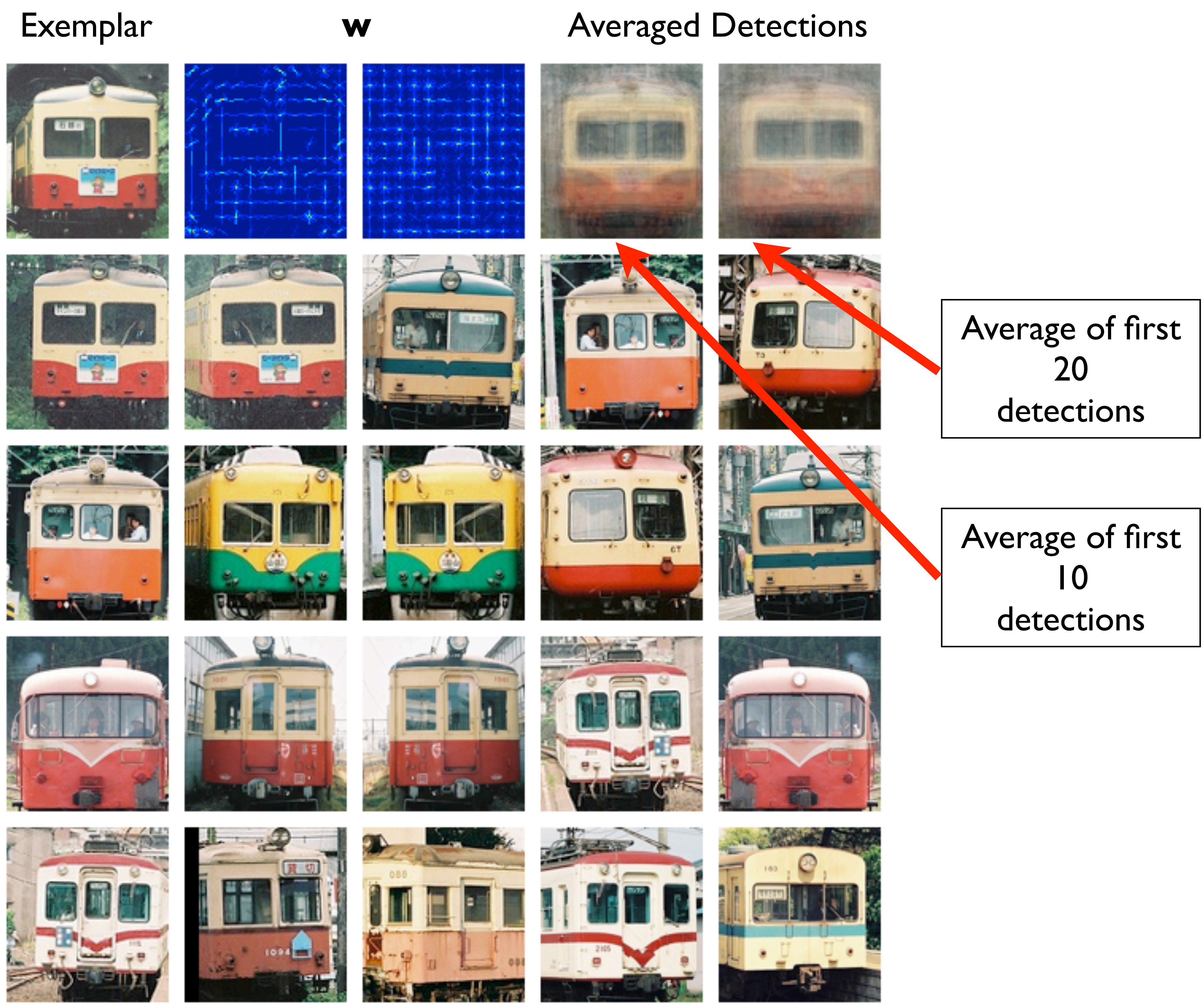
- Each exemplar defines its own single-instance "category"
- Each Exemplar-SVM acts as a "distance function" but without the exemplar at origin constraint
- As a linear classifier, Exemplar-SVMs operate as a simple dot product in feature space

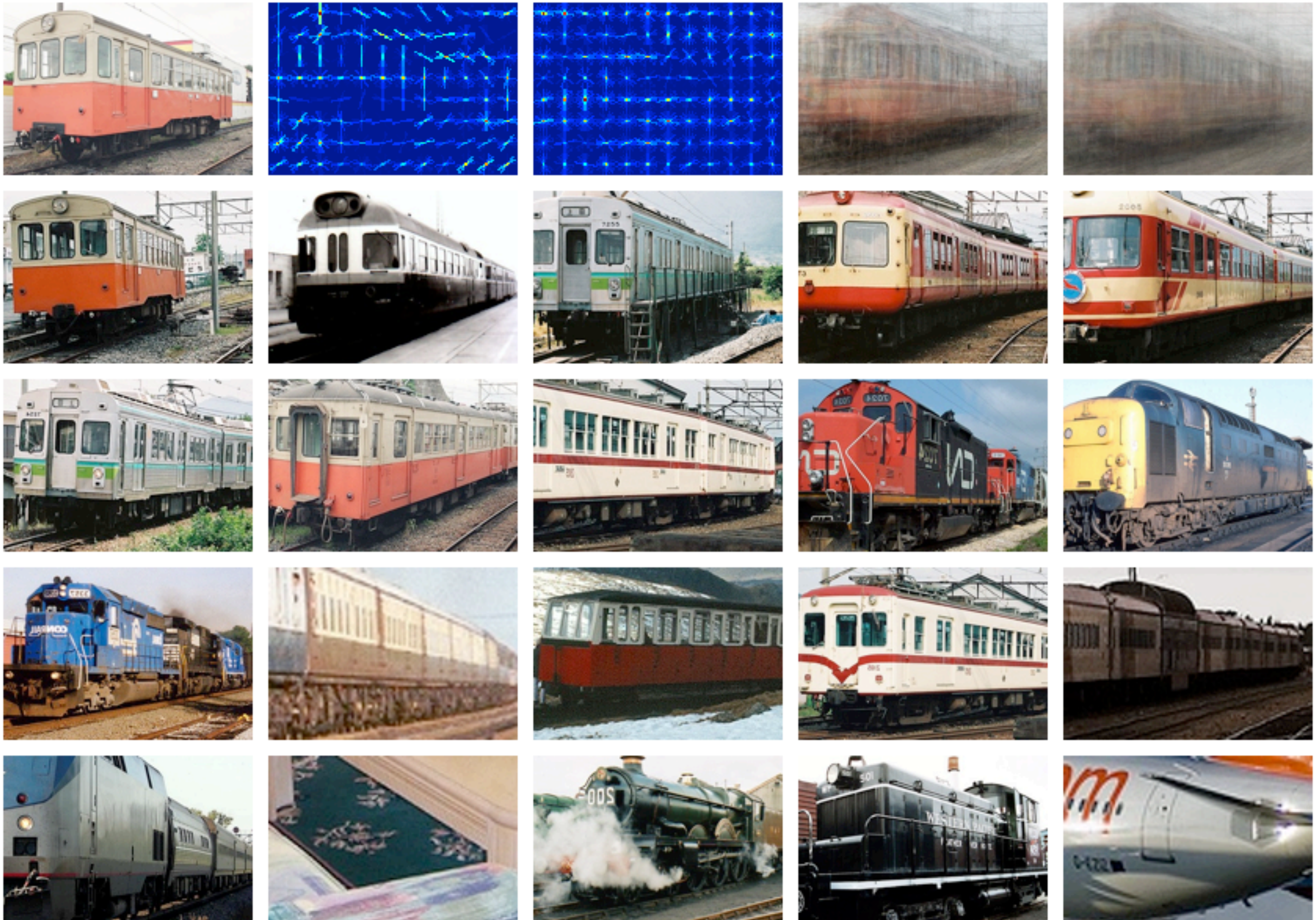
Visualizing Exemplar-SVMs

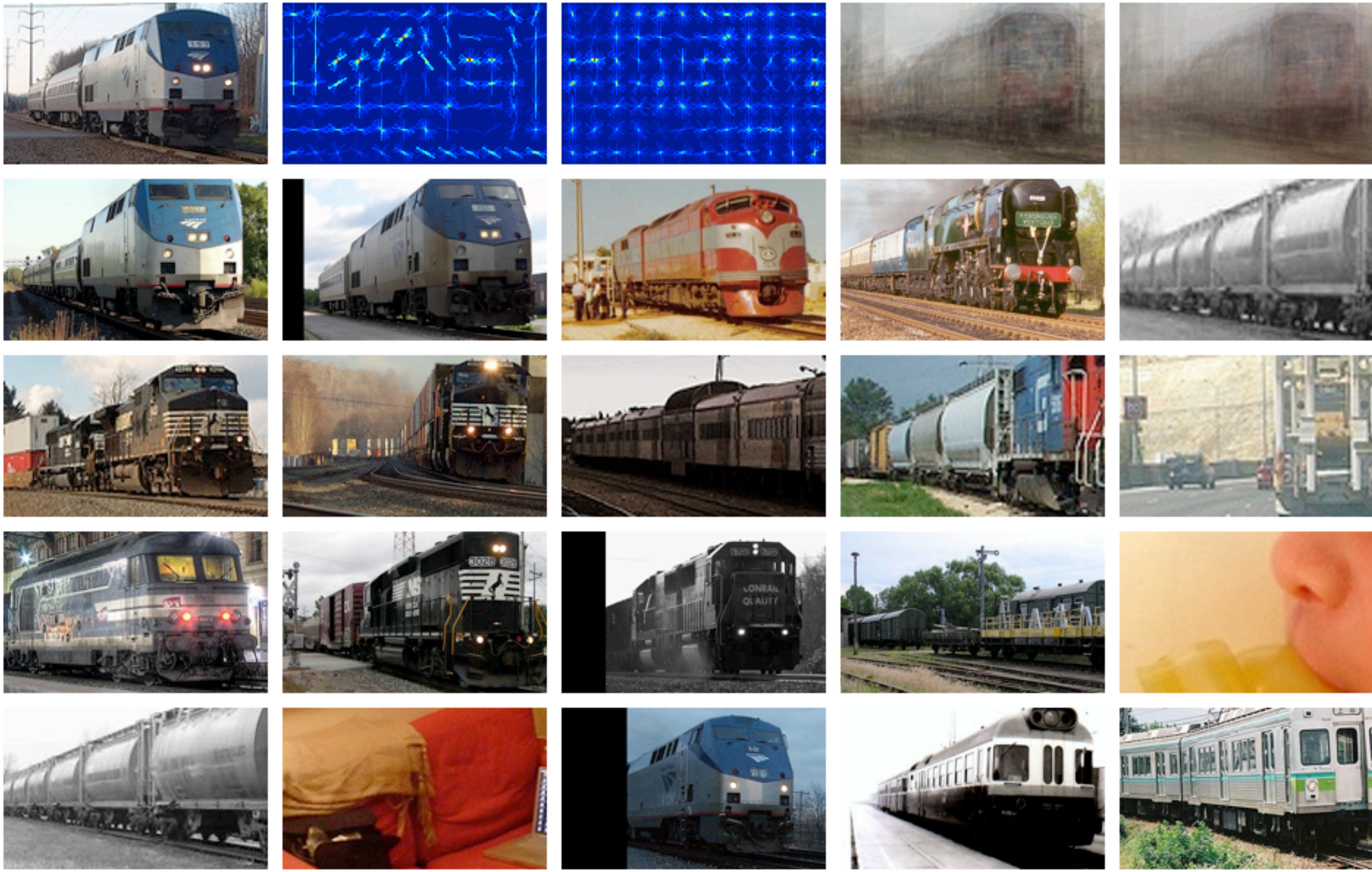
Exemplar **w**



Exemplar **w**


Exemplar
w
Averaged Detections
Average of first 20 detections
Average of first 10 detections



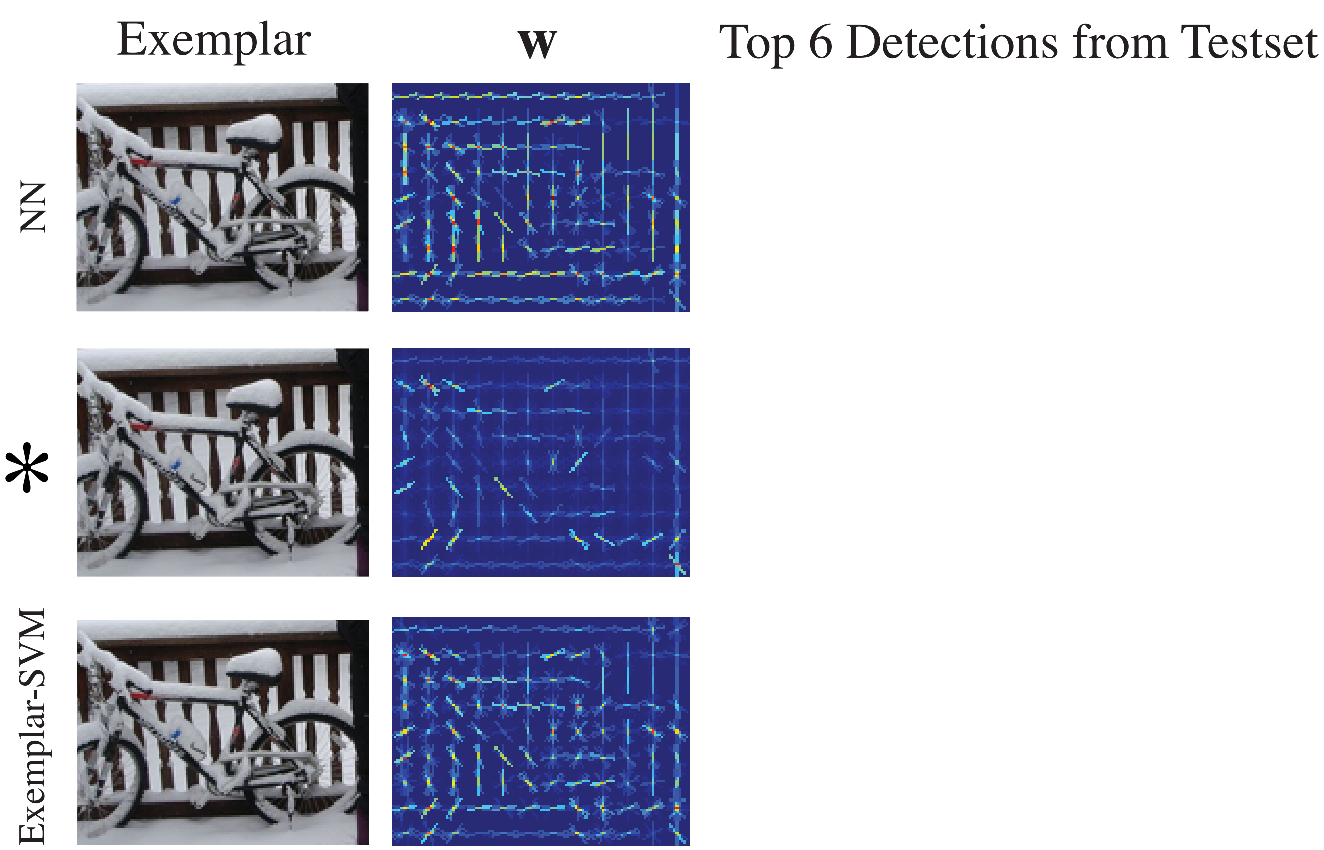


Pan Am Railways
NREX
8006
BNSF
CN
PRIVATE PROPERTY
NO TRESPASSING
DANGER
CONRAIL

Understanding Exemplar-SVMs

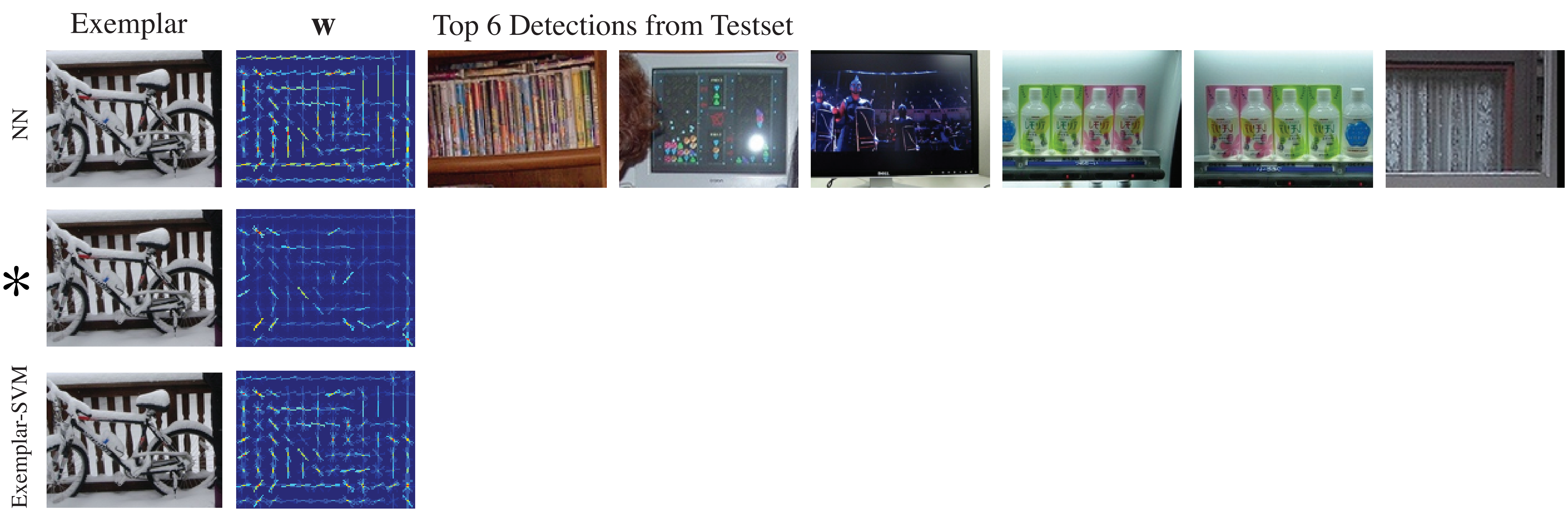
- **Nearest Neighbor**
  - No Learning
- **Per-Exemplar Distance Functions**
  - Learning in distance-to-exemplar space [Malisiewicz et al. 2008]
- **Exemplar-SVMs**

Comparison of 3 methods



*Learned Distance Function

Comparison of 3 methods



*Learned Distance Function

Comparison of 3 methods

*Learned Distance Function

Comparison of 3 methods

*Learned Distance Function

PASCAL VOC 2007 Object Category Detection Results

Approach	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
NN	.006	.094	.000	.005	.000	.006	.010	.092	.001	.092	.001	.004	.096	.094	.005	.018	.009	.008	.096	.144	.039
NN+Cal	.056	.293	.012	.034	.009	.207	.261	.017	.094	.111	.004	.033	.243	.188	.114	.020	.129	.003	.183	.195	.110
DFUN+Cal	.162	.364	.008	.096	.097	.316	.366	.092	.098	.107	.002	.093	.234	.223	.109	.037	.117	.016	.271	.293	.155
E-SVM+Cal	.204	.407	.093	.100	.103	.310	.401	.096	.104	.147	.023	.097	.384	.320	.192	.096	.167	.110	.291	.315	.198
E-SVM+Co-occ	.208	.480	.077	.143	.131	.397	.411	.052	.116	.186	.111	.031	.447	.394	.169	.112	.226	.170	.369	.300	.227
CZ [6]	.262	.409	–	–	–	.393	.432	–	–	–	–	–	–	.375	–	–	–	–	.334	–	–
DT [7]	.127	.253	.005	.015	.107	.205	.230	.005	.021	.128	.014	.004	.122	.103	.101	.022	.056	.050	.120	.248	.097
LDPM [9]	.287	.510	.006	.145	.265	.397	.502	.163	.165	.166	.245	.050	.452	.383	.362	.090	.174	.228	.341	.384	.266

Table 1. **PASCAL VOC 2007 object detection results.** We compare our full system (ESVM+Co-occ) to four different exemplar based baselines including NN (Nearest Neighbor), NN+Cal (Nearest Neighbor with calibration), DFUN+Cal (learned distance function with calibration) and ESVM+Cal (Exemplar-SVM with calibration). We also compare our approach against global methods including our implementation of Dalal-Triggs (learning a single global template), LDPM [9] (Latent deformable part model), and Chum et al. [6]'s exemplar-based method. [The NN, NN+Cal and DFUN+Cal results for person category are obtained using 1250 exemplars]

PASCAL VOC 2007 Object Category Detection Results

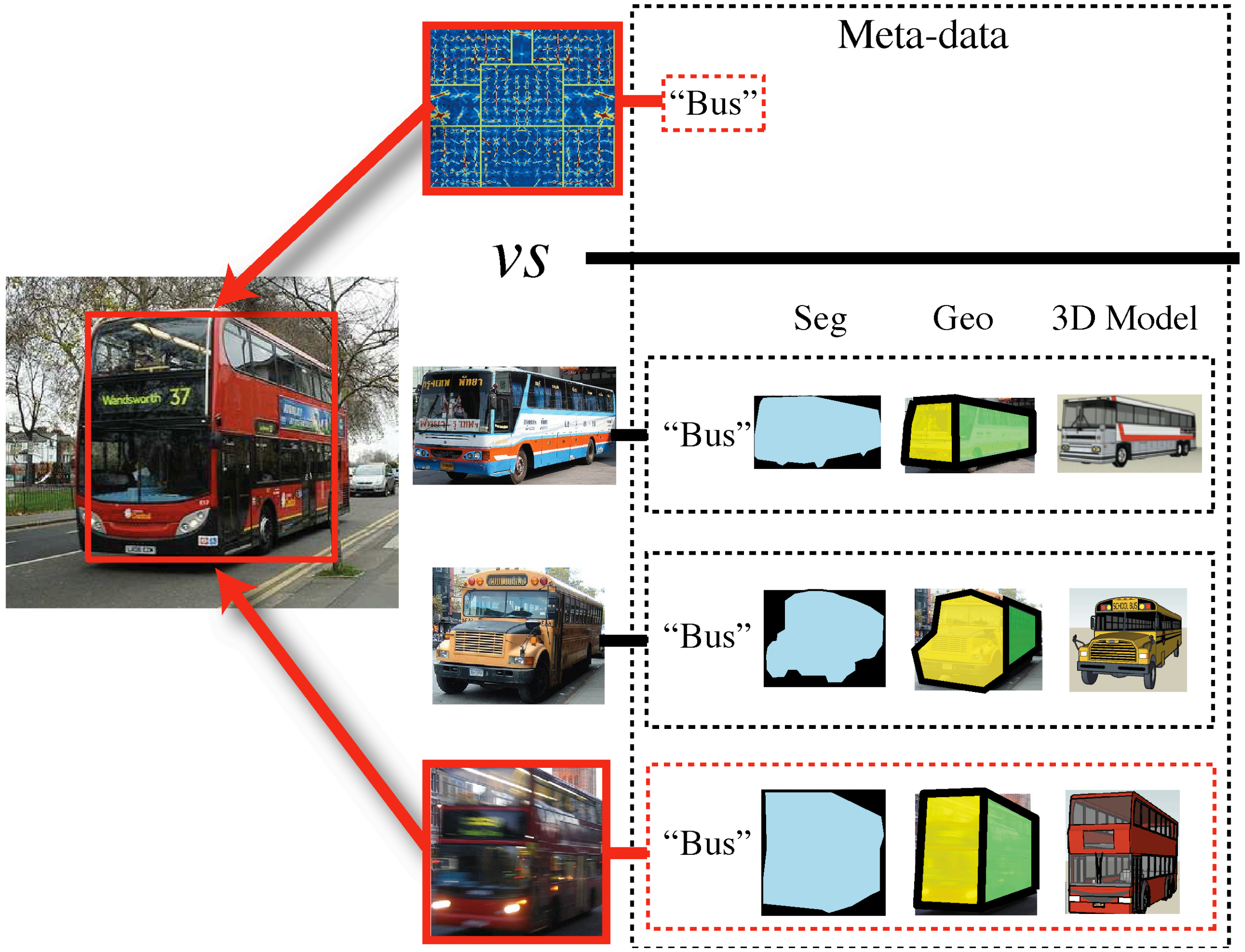
Approach	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
NN	.006	.094	.000	.005	.000	.006	.010	.092	.001	.092	.001	.004	.096	.094	.005	.018	.009	.008	.096	.144	.039
NN+Cal	.056	.293	.012	.034	.009	.207	.261	.017	.094	.111	.004	.033	.243	.188	.114	.020	.129	.003	.183	.195	.110
DFUN+Cal	.162	.364	.008	.096	.097	.316	.366	.092	.098	.107	.002	.093	.234	.223	.109	.037	.117	.016	.271	.293	.155
E-SVM+Cal	.204	.407	.093	.100	.103	.310	.401	.096	.104	.147	.023	.097	.384	.320	.192	.096	.167	.110	.291	.315	.198
E-SVM+Co-occ	.208	.480	.077	.143	.131	.397	.411	.052	.116	.186	.111	.031	.447	.394	.169	.112	.226	.170	.369	.300	.227
CZ [6]	.262	.409	–	–	–	.393	.432	–	–	–	–	–	–	.375	–	–	–	–	.334	–	–
DT [7]	.127	.253	.005	.015	.107	.205	.230	.005	.021	.128	.014	.004	.122	.103	.101	.022	.056	.050	.120	.248	.097
LDPM [9]	.287	.510	.006	.145	.265	.397	.502	.163	.165	.166	.245	.050	.452	.383	.362	.090	.174	.228	.341	.384	.266

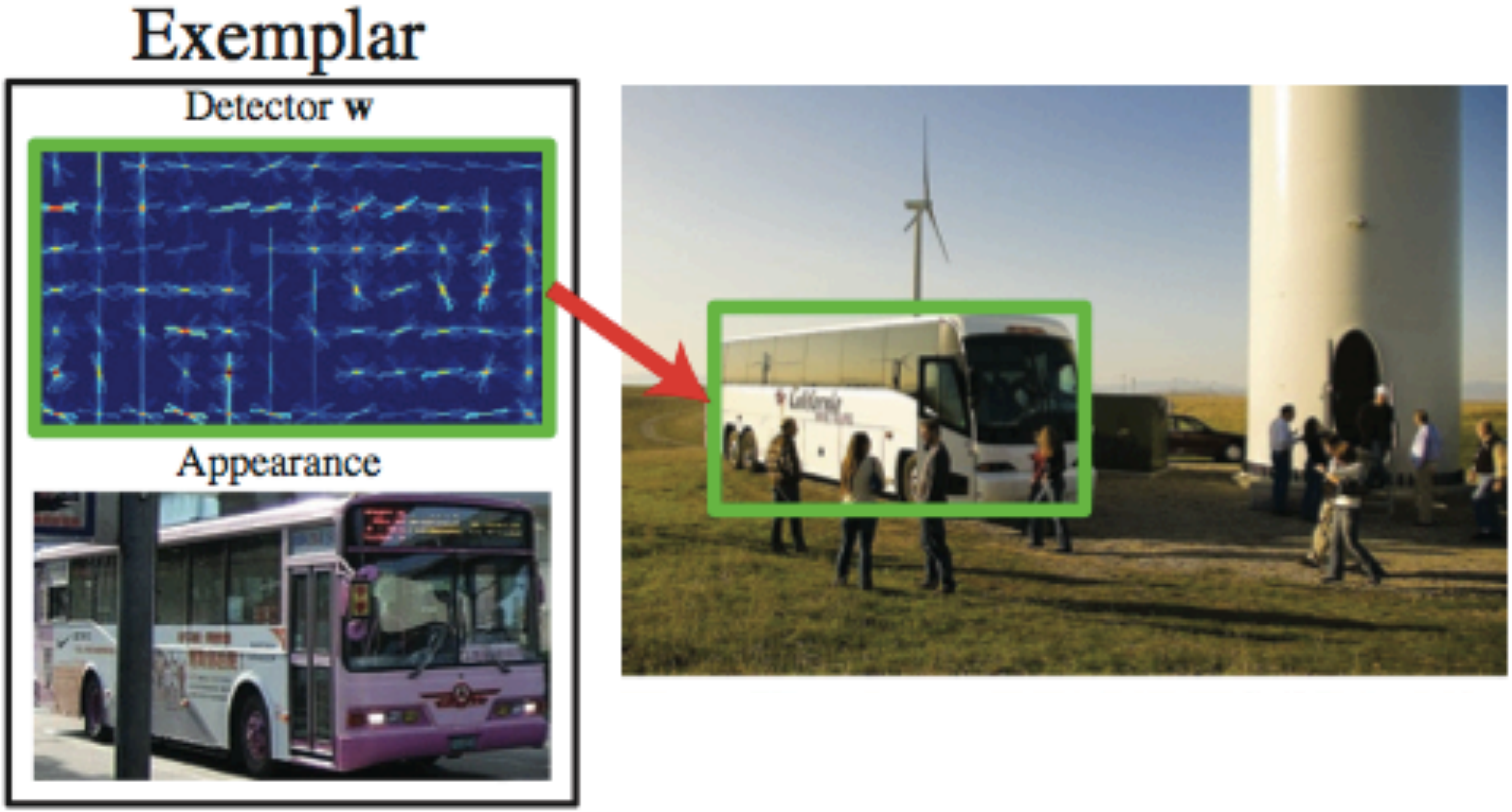
Table 1. **PASCAL VOC 2007 object detection results.** We compare our full system (ESVM+Co-occ) to four different exemplar based baselines including NN (Nearest Neighbor), NN+Cal (Nearest Neighbor with calibration), DFUN+Cal (learned distance function with calibration) and ESVM+Cal (Exemplar-SVM with calibration). We also compare our approach against global methods including our implementation of Dalal-Triggs (learning a single global template), LDPM [9] (Latent deformable part model), and Chum et al. [6]'s exemplar-based method. [The NN, NN+Cal and DFUN+Cal results for person category are obtained using 1250 exemplars]

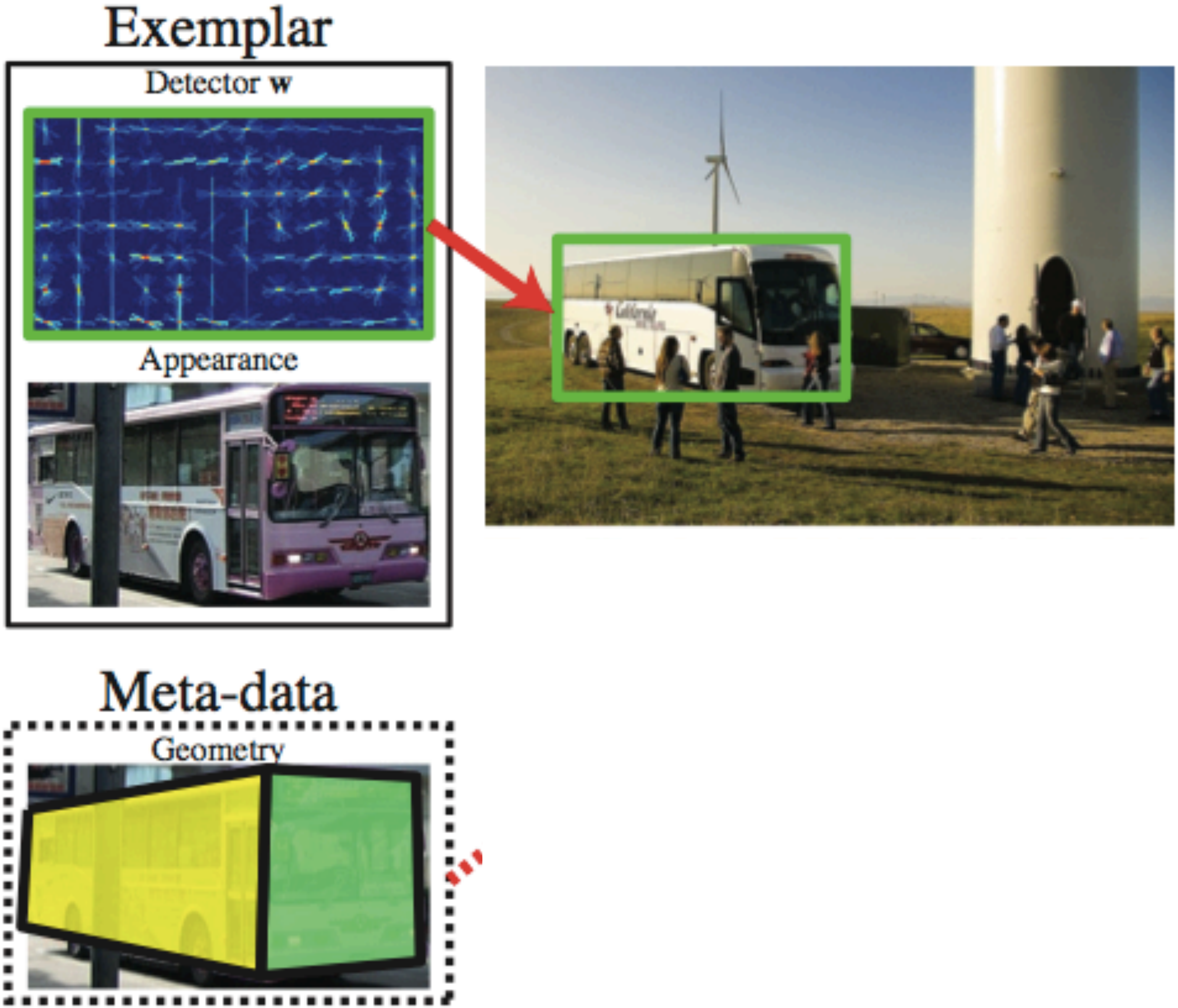
Equal or better in performance than Pedro Felzenszwalb's Latent Deformable Part-based Model in 7 PASCAL VOC 2007 categories.

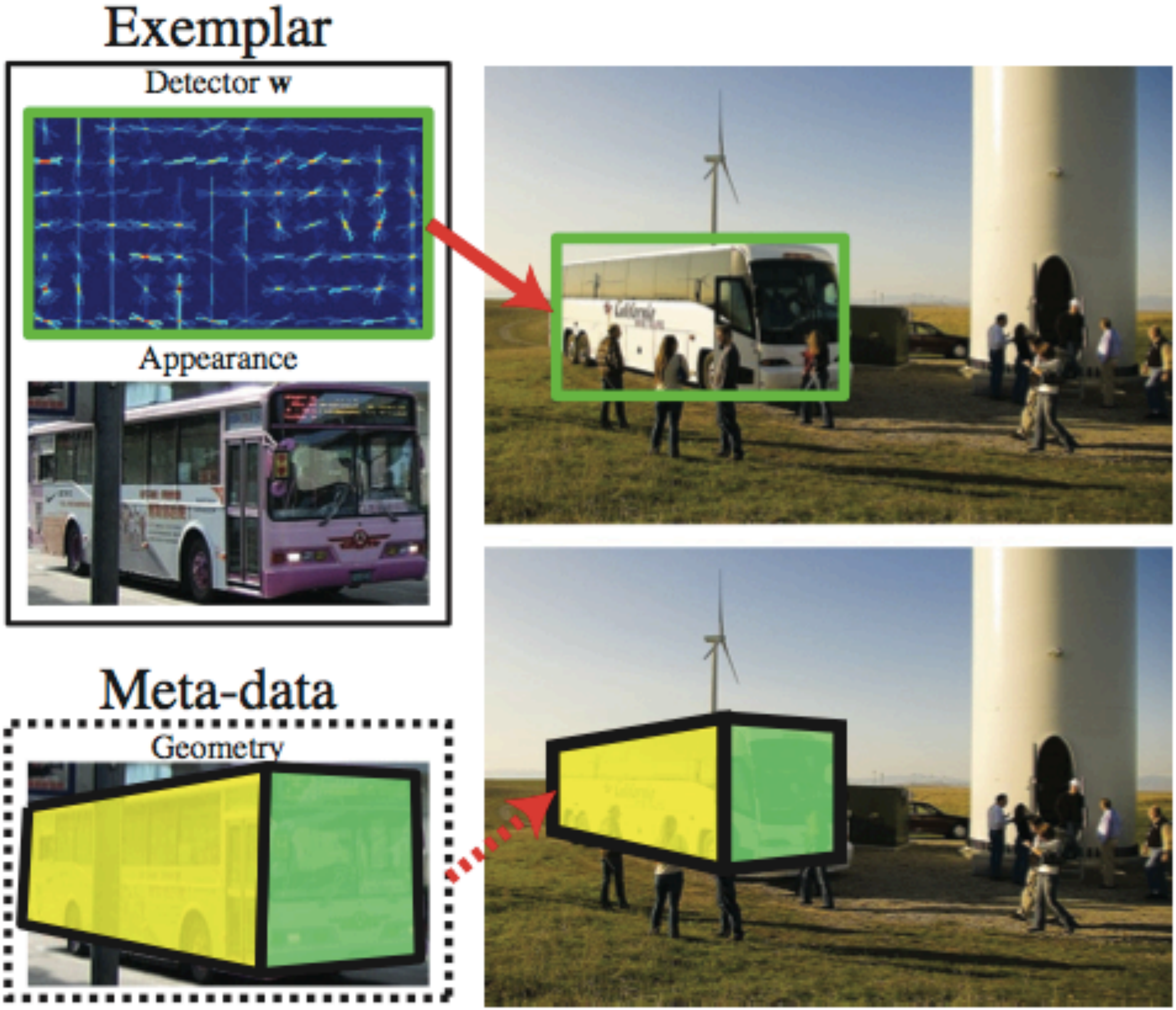
Meta-data transfer

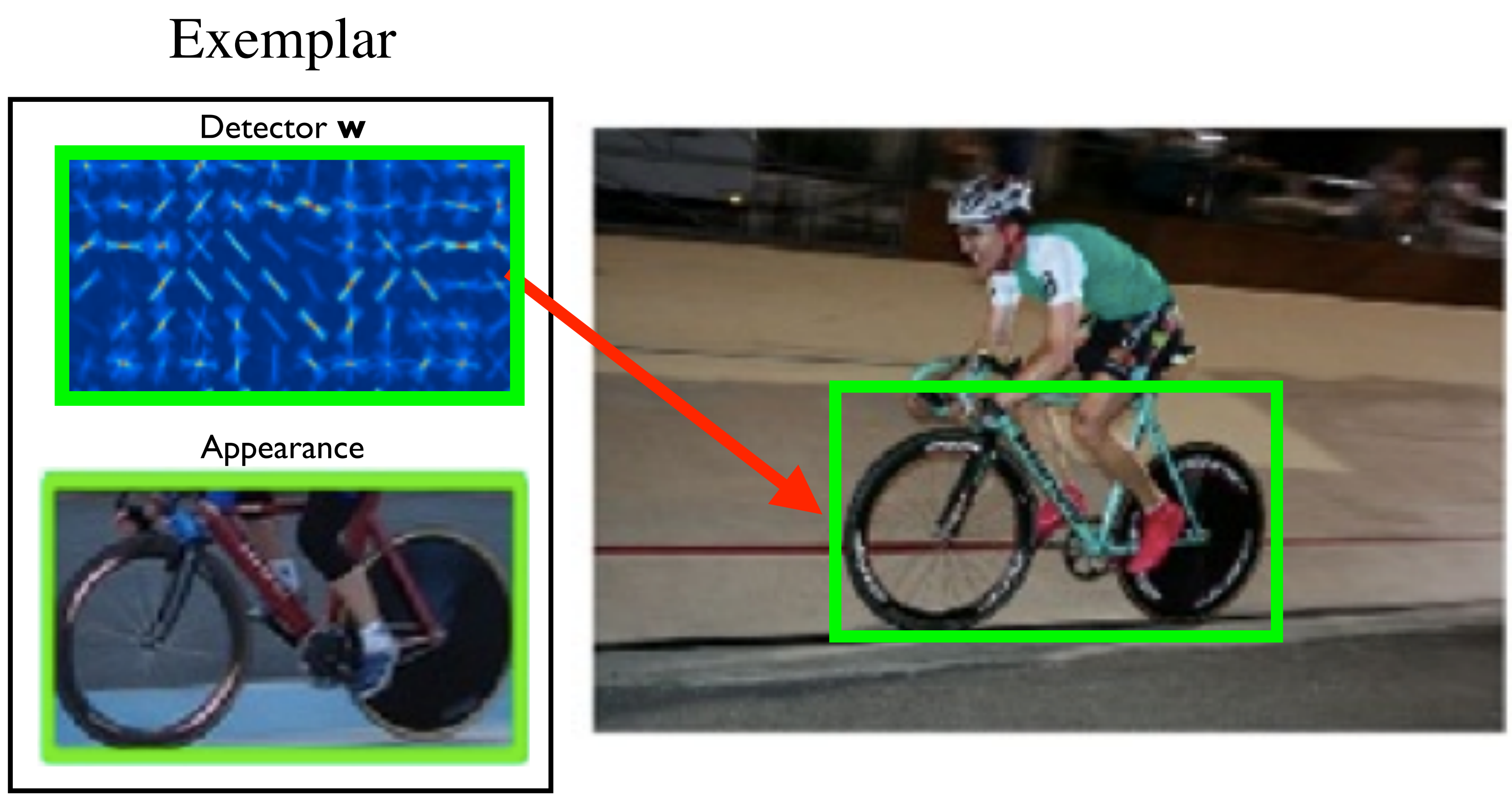
- Based on the idea of label transfer [Torralba et al], Exemplar-SVMs can be used for tasks which go beyond object category detection

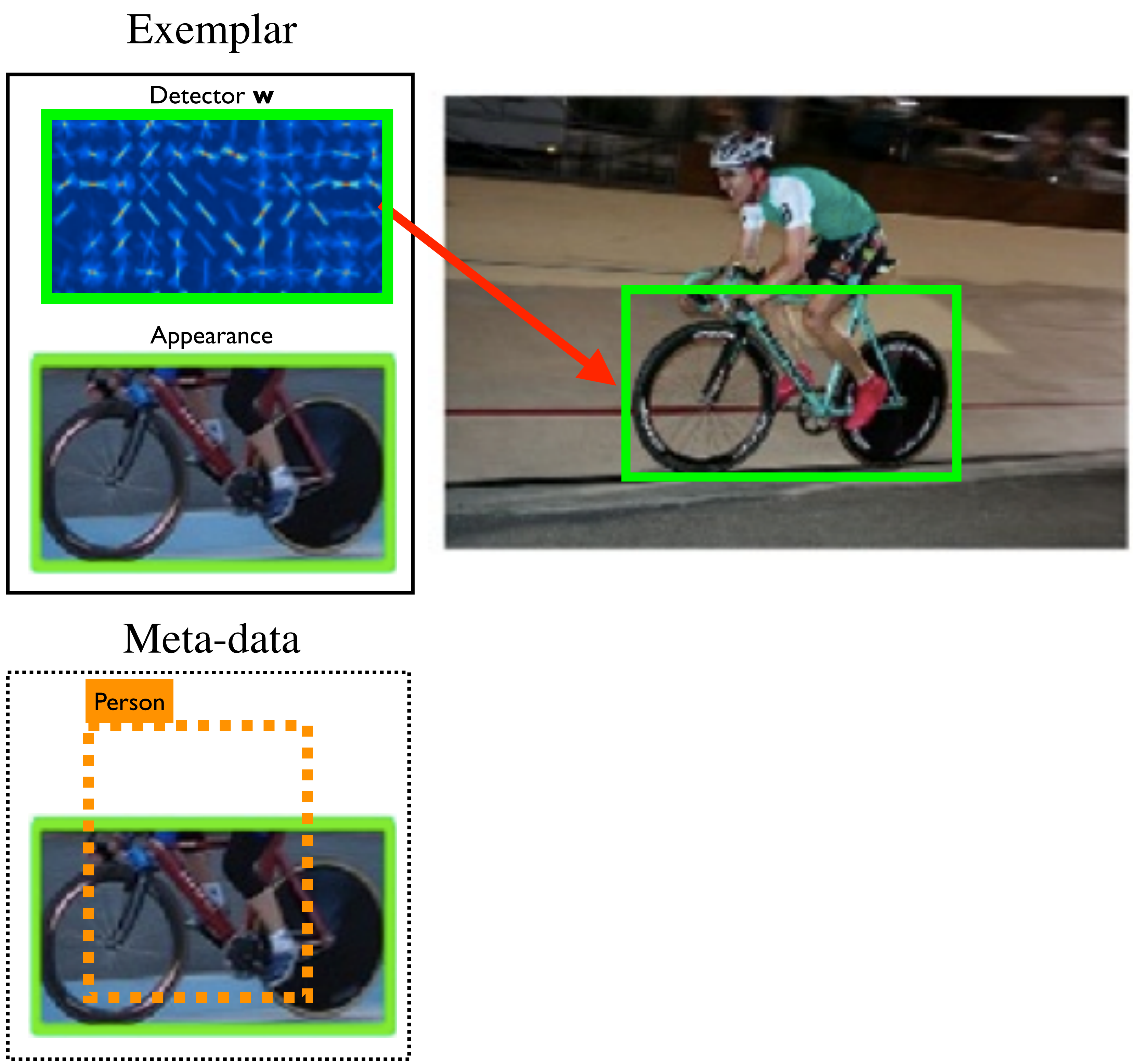


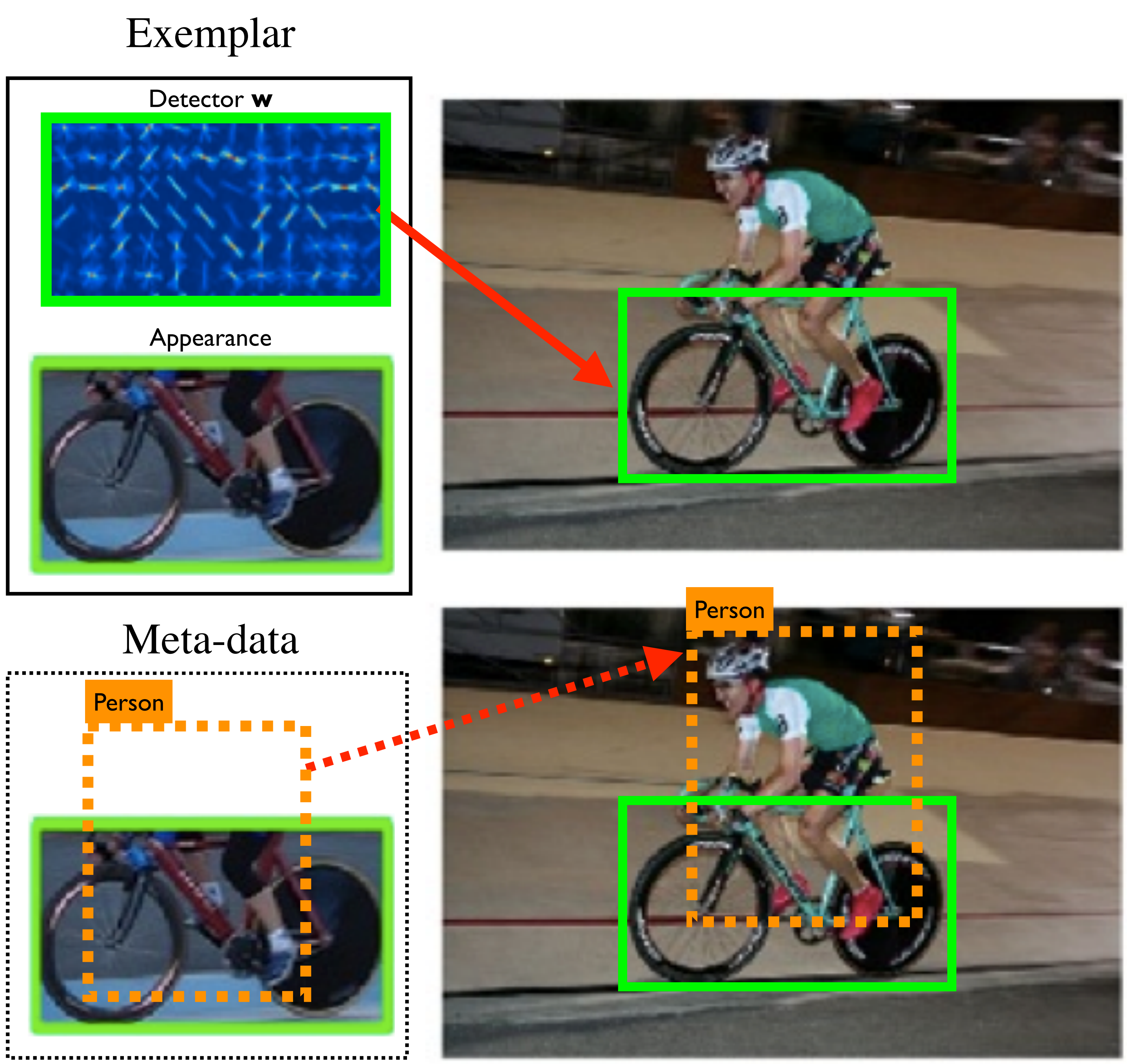

Exemplar
Detector w
Appearance

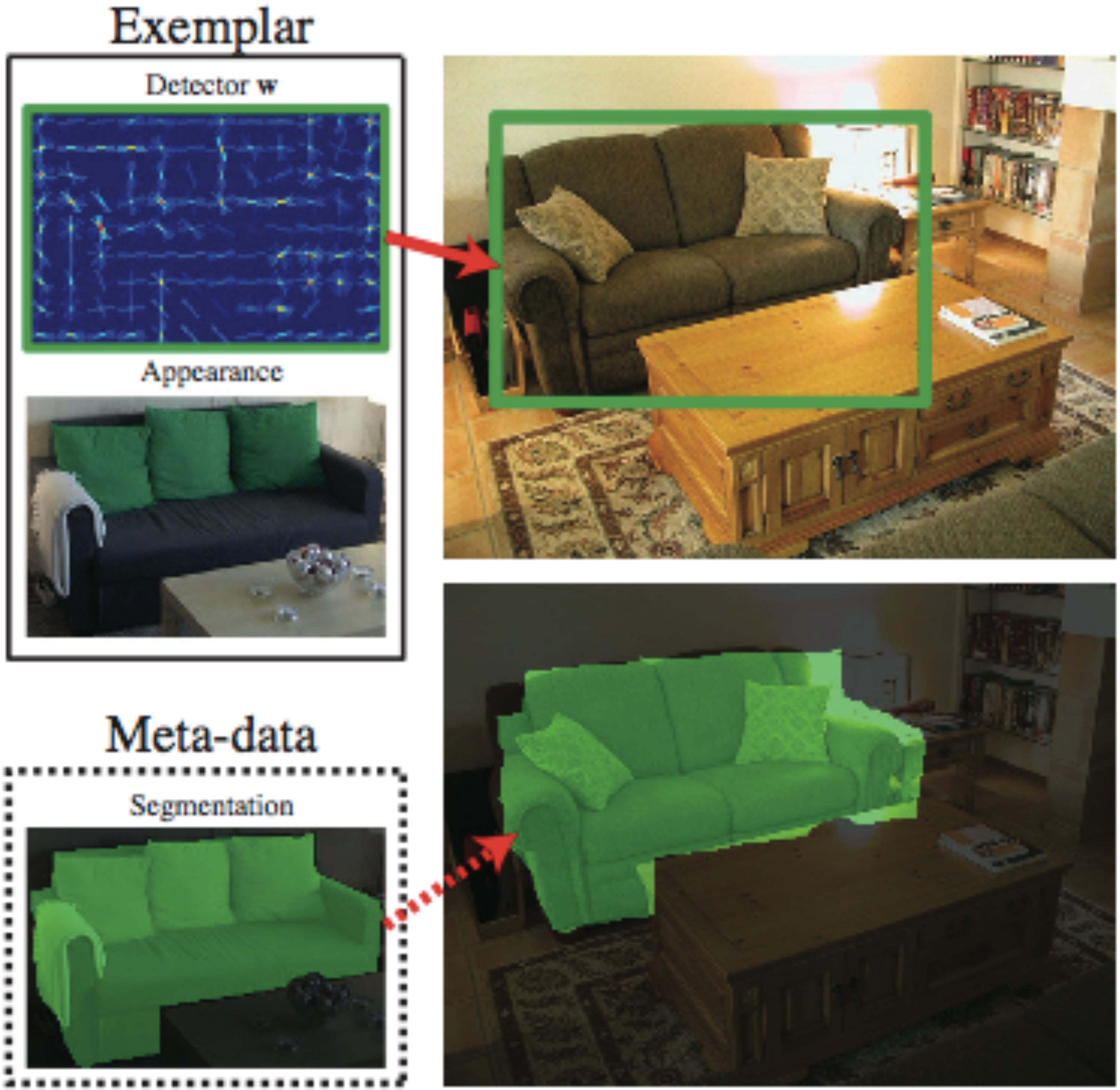

Exemplar
Detector w
Appearance
Meta-data
Geometry


Exemplar
Detector w
Appearance
Meta-data
Geometry


Exemplar
Detector w
Appearance

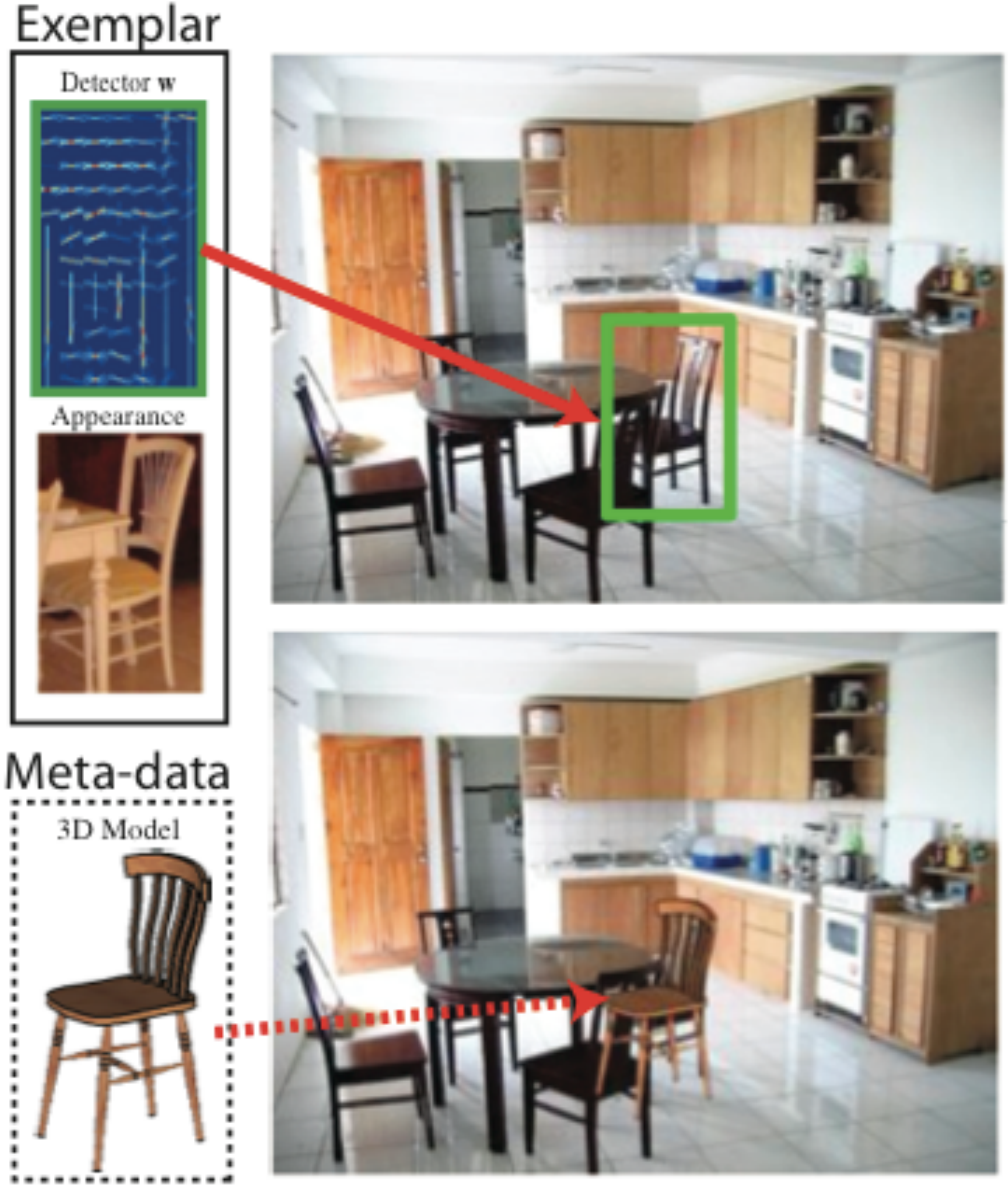

Exemplar
Detector w
Appearance
Meta-data
Person


Exemplar
Detector w
Appearance
Meta-data
Person
Person



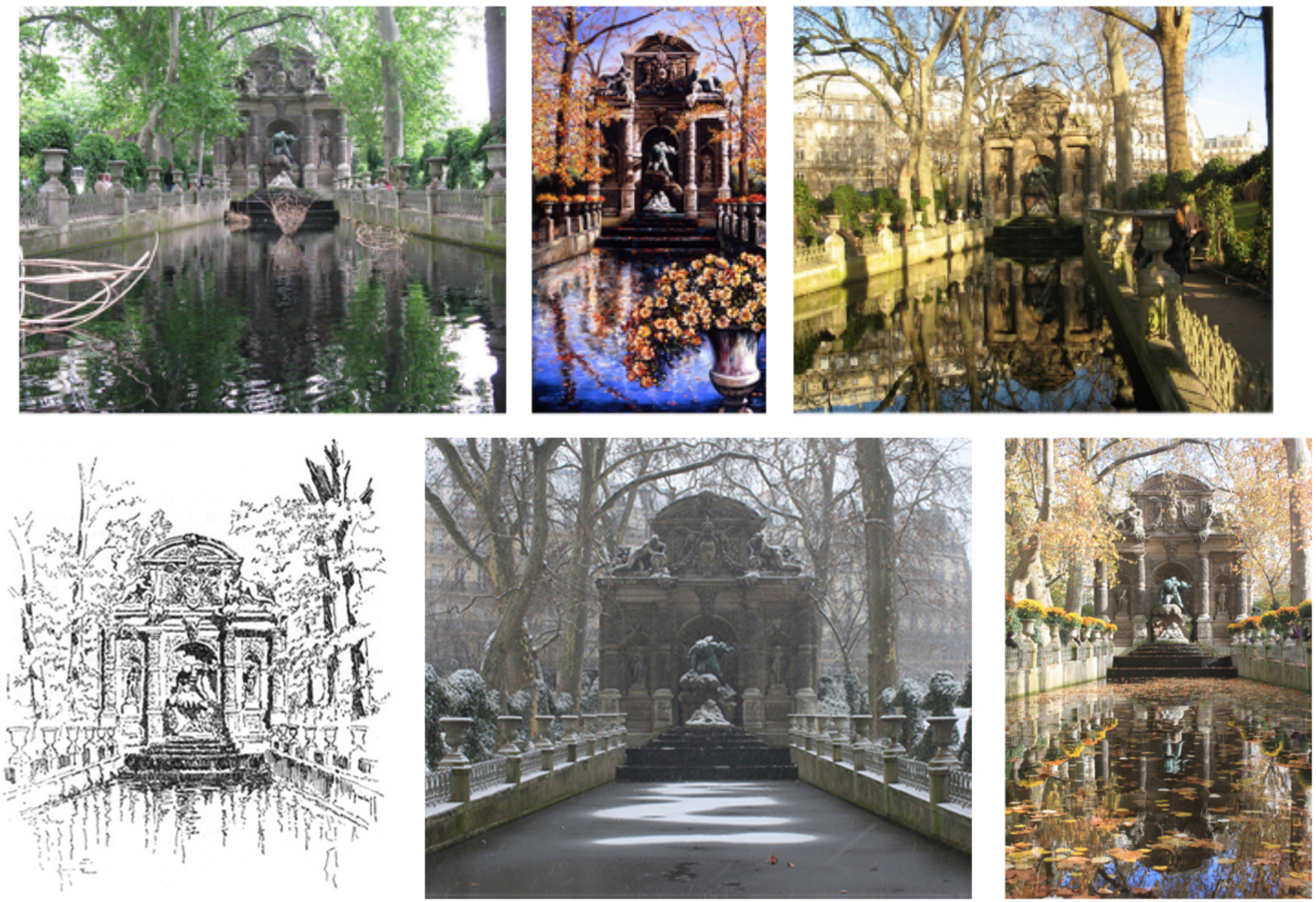
Exemplar
Detector w
Appearance
Meta-data
Segmentation


Exemplar
Detector w
Appearance
Meta-data
3D Model



Exemplar
Detector w
Appearance
Meta-data
3D Model

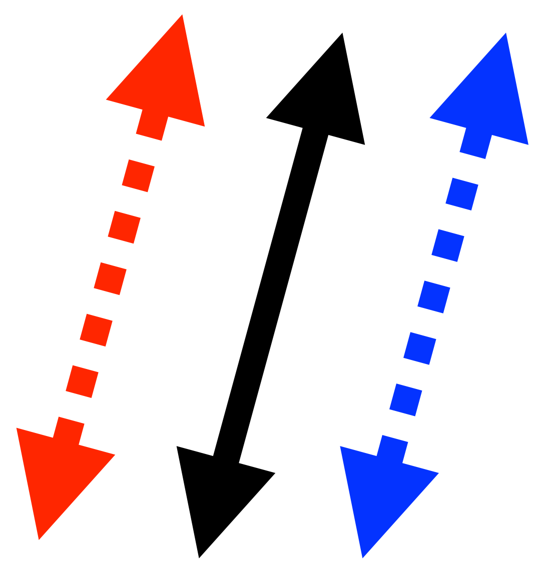
Cross-domain Image Matching



Shrivastava et al. SIGGRAPH ASIA 2011

Learn Exemplar-SVM for query image

Query Sketch

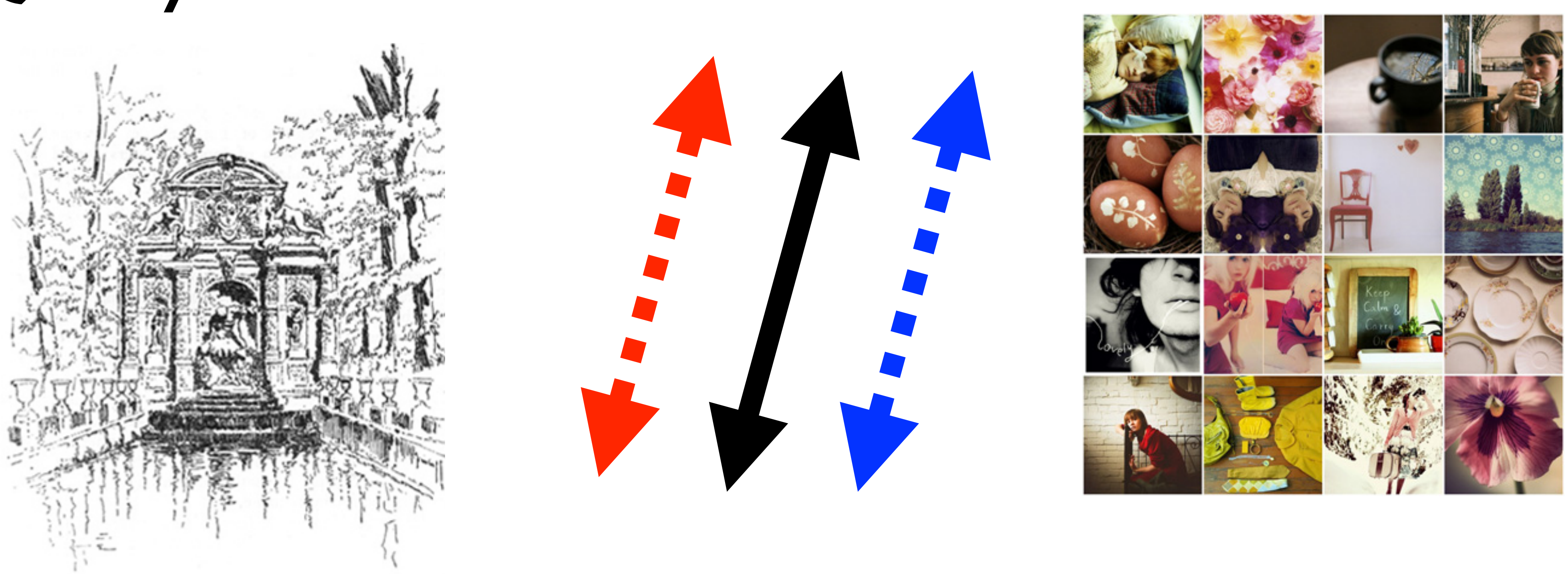


Negatives mined from random Flickr images

Learn Exemplar-SVM for query image

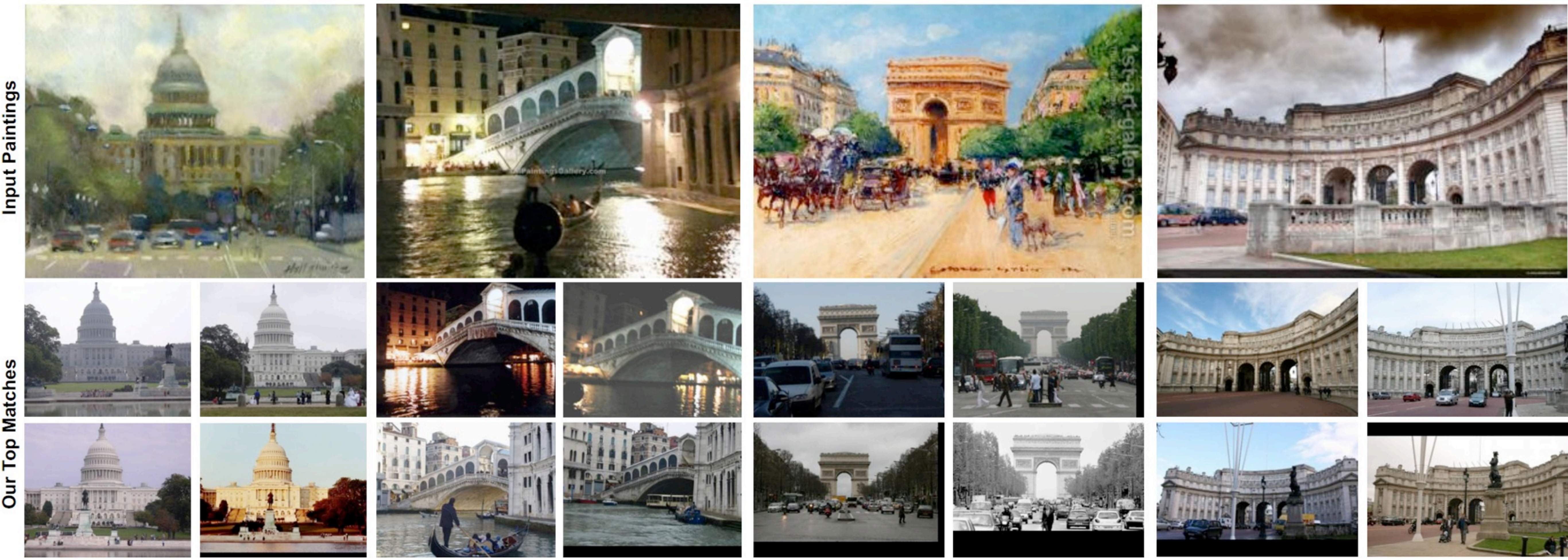
Query Sketch

Negatives mined from random Flickr images

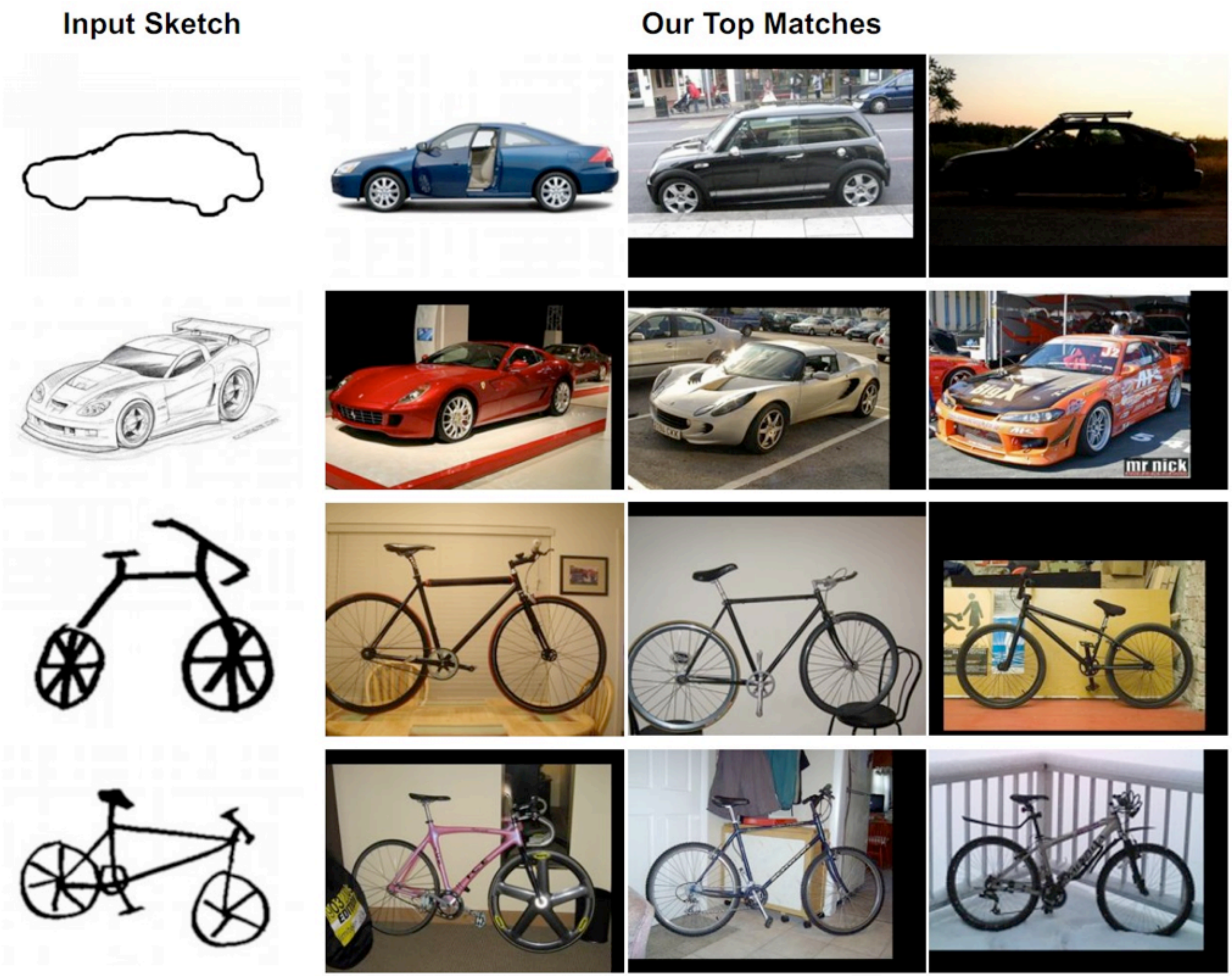


Then apply learned **w** to retrieval set of images in a sliding-window fashion

Painting to Image



Sketch to Image



Painting to GPS

IM2GPS: Hays et al. 2008

Open Problems

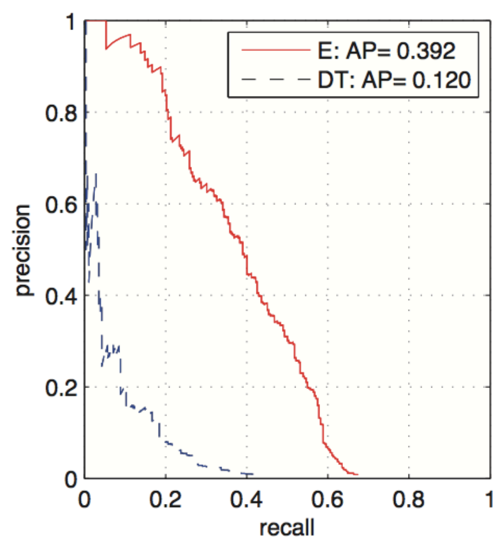
- 1. Learning for many exemplars is computationally expensive. Can cleverly reusing mined negatives help speed-up training?

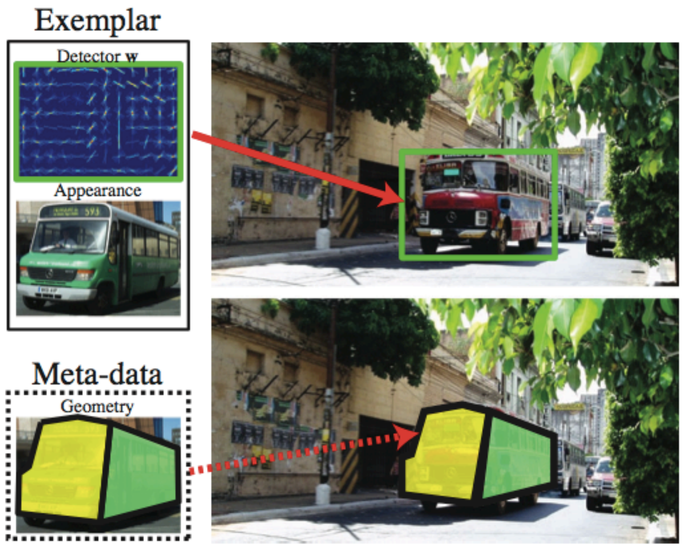
- 2. At test-time, applying N Exemplar-SVMs takes O(N) time. Can exemplar pruning or approximate matching algorithms help?

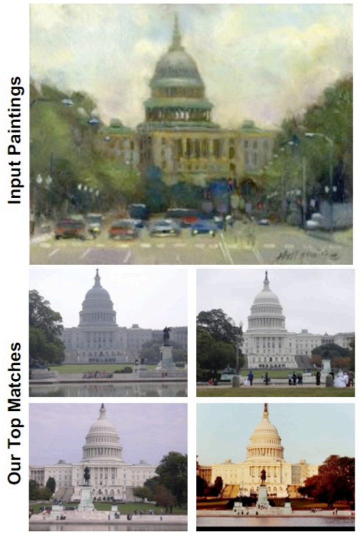
Concluding Remarks

Concluding Remarks

- Exemplar-SVMs can be used for detection, meta-data transfer, as well as cross-domain image matching

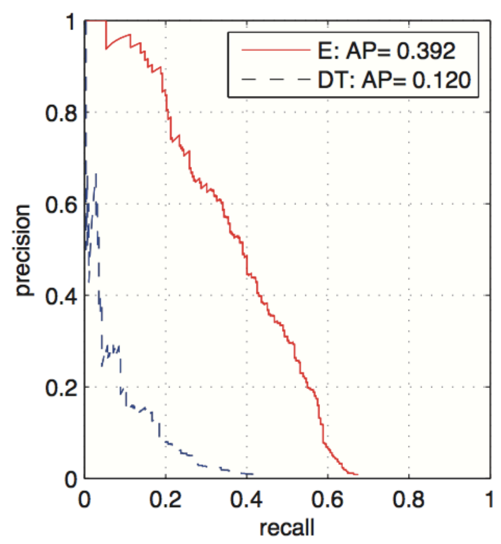


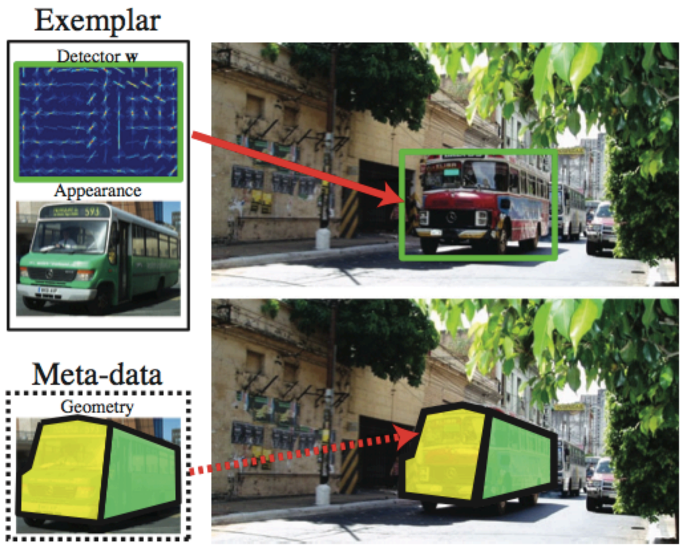


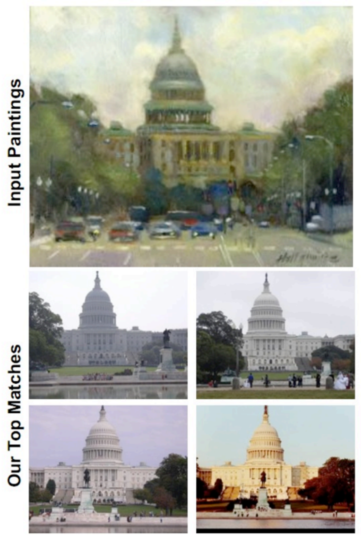


Concluding Remarks

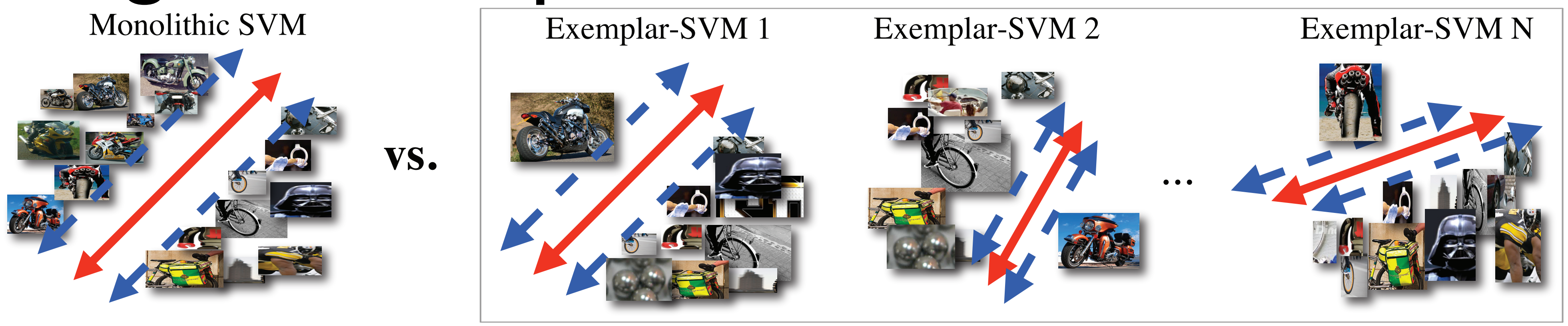
- Exemplar-SVMs can be used for detection, meta-data transfer, as well as cross-domain image matching







- Dealing with lots of data is the **key** to learning a good Exemplar-SVM



Thank you for listening

Abhinav Shrivastava

Tomasz Malisiewicz

Abhinav Gupta

Alyosha Efros

Thank you for listening


Abhinav Shrivastava


Tomasz Malisiewicz


Abhinav Gupta


Alyosha Efros

ExemplarSVMs

Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, Alexei A. Efros. **Data-driven Visual Similarity for Cross-domain Image Matching.** In SIGGRAPH ASIA, 2011.

Tomasz Malisiewicz. **Exemplar-based Representations for Object Detection, Association and Beyond.** CMU PhD Dissertation. August, 2011.

Tomasz Malisiewicz, Abhinav Gupta, Alexei A. Efros. **Ensemble of Exemplar-SVMs for Object Detection and Beyond.** In ICCV, 2011.

Per-Exemplar Distance Functions

Tomasz Malisiewicz, Alexei A. Efros. **Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships**. In NIPS, 2009.

Tomasz Malisiewicz, Alexei A. Efros. **Recognition by Association via Learning Per-exemplar Distances.** In CVPR, 2008.