Sharing knowledge for large scale visual recognition

Lamberto Ballan
AI Lab, Stanford University

Standard computer vision paradigm

- Q: What objects are in the image?

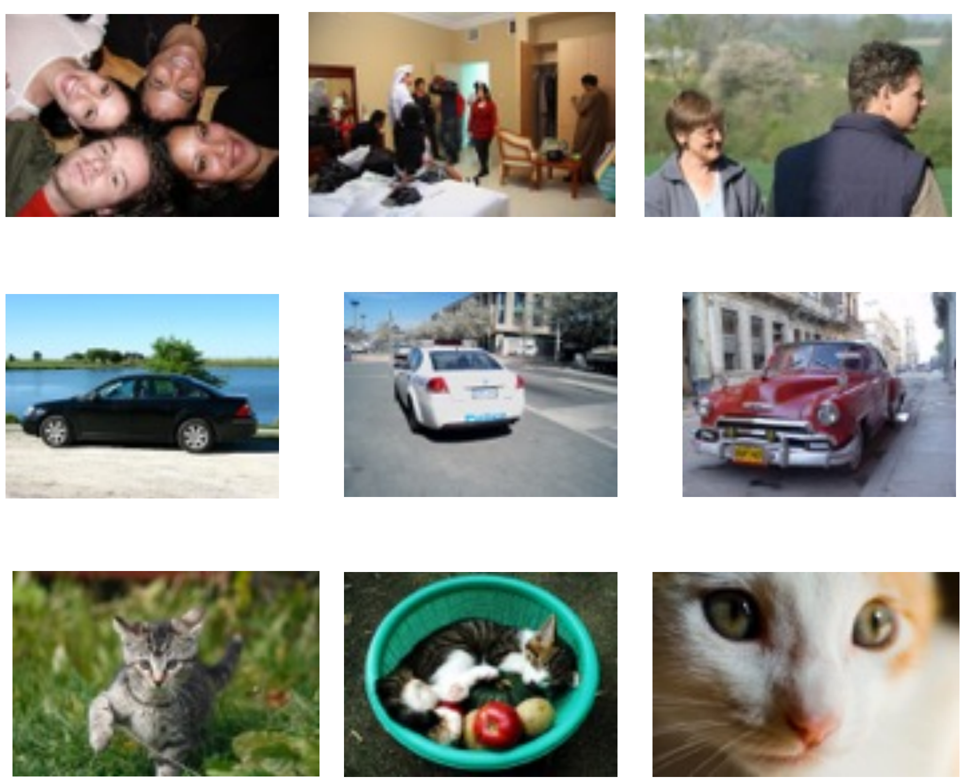
Test Data

Training Data

person

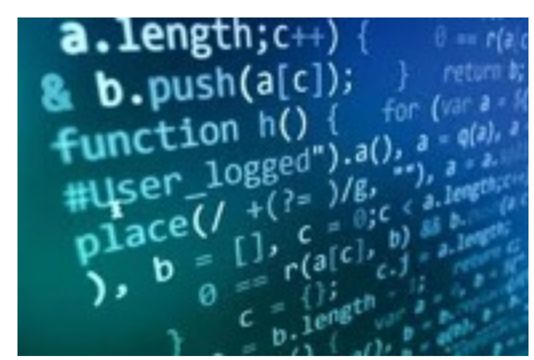
car

cat



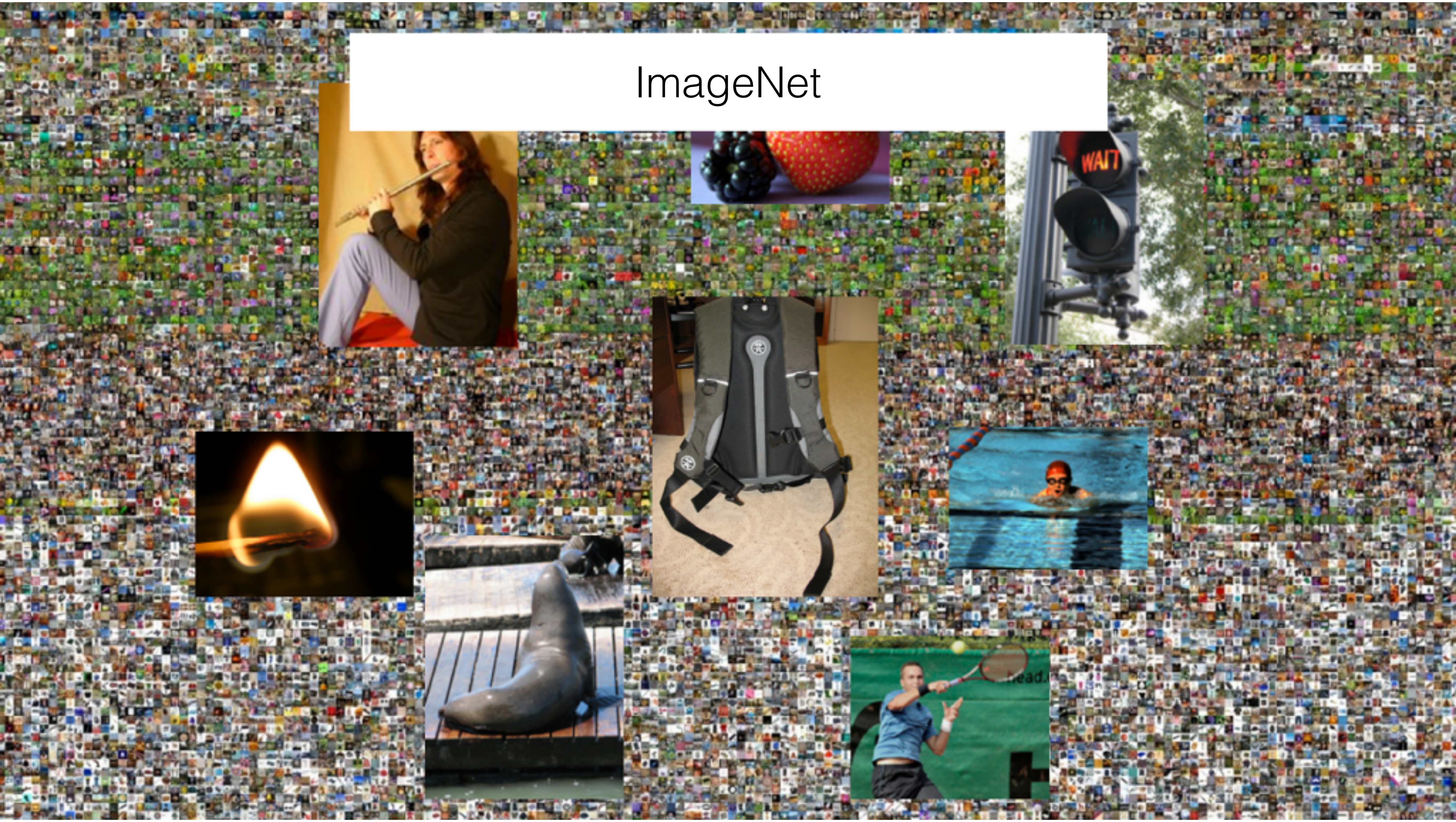
...

(ML) Algorithm



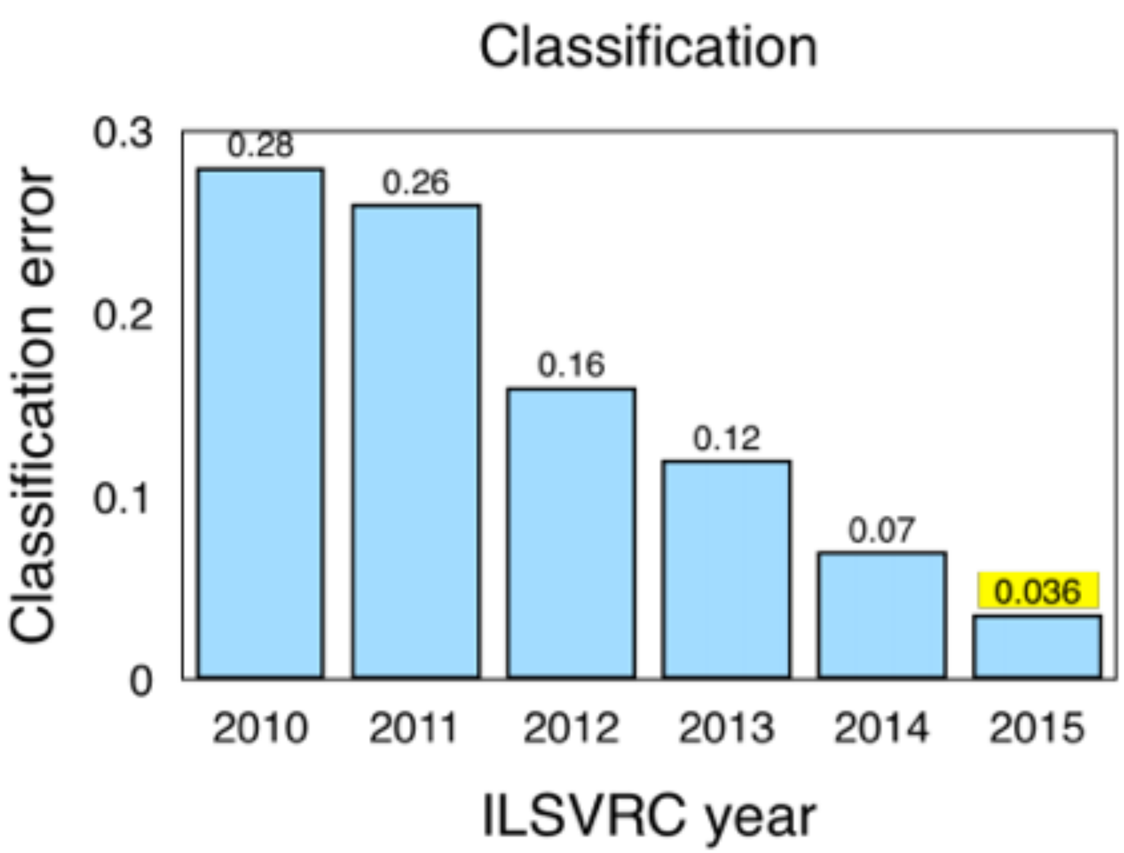
car
person

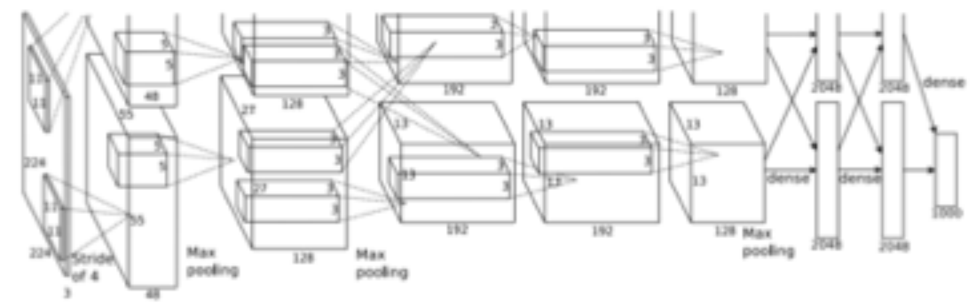
Datasets drive computer vision progress



ImageNet: ILSVRC results

- Result in ILSVRC (classification) over the years

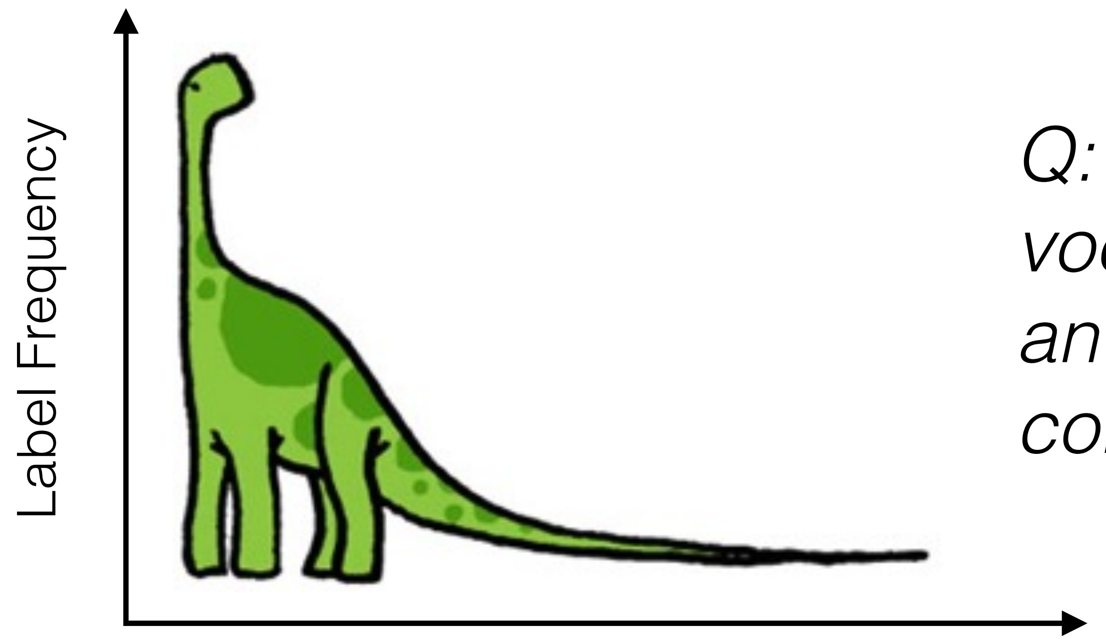




Team Name	Error (%)
MSRA	**3.57**
ReCeption	3.58
Trimps-Soushen	4.58
Qualcomm Research	4.87
VUNO	5.03
CIL	5.48
CUimage	5.86
MCG-ICT-CAS	6.31
HiVision	6.48

The long tail

- A small number off generic objects/entities/labels appear very often while most others appear rarely
- There are a few real-world scenarios in which we have access to 1M+ images uniformly belonging to a set of 1000+ classes



Q: How to scale up to very large vocabularies (infrequent labels) and a scenario where it is hard to collect ground truth data?

Automatic image annotation by exploiting image metadata and weak labels

Motivation

- Can you guess what's in the image?

petal?

fruit?

tentacle?

Motivation

- Let’s try to add more context…

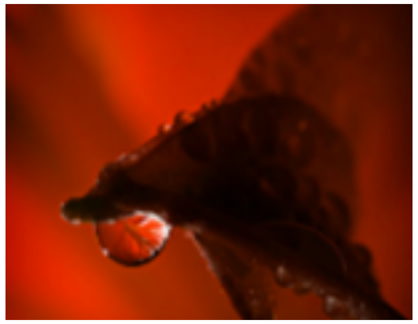
Tags:

flower
petal
closeup
water

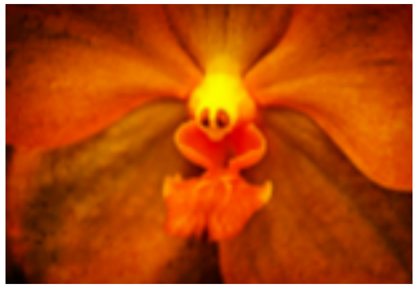
GPS
groups
…

Motivation

- In the context of images which share similar metadata it is easier to give the right answer





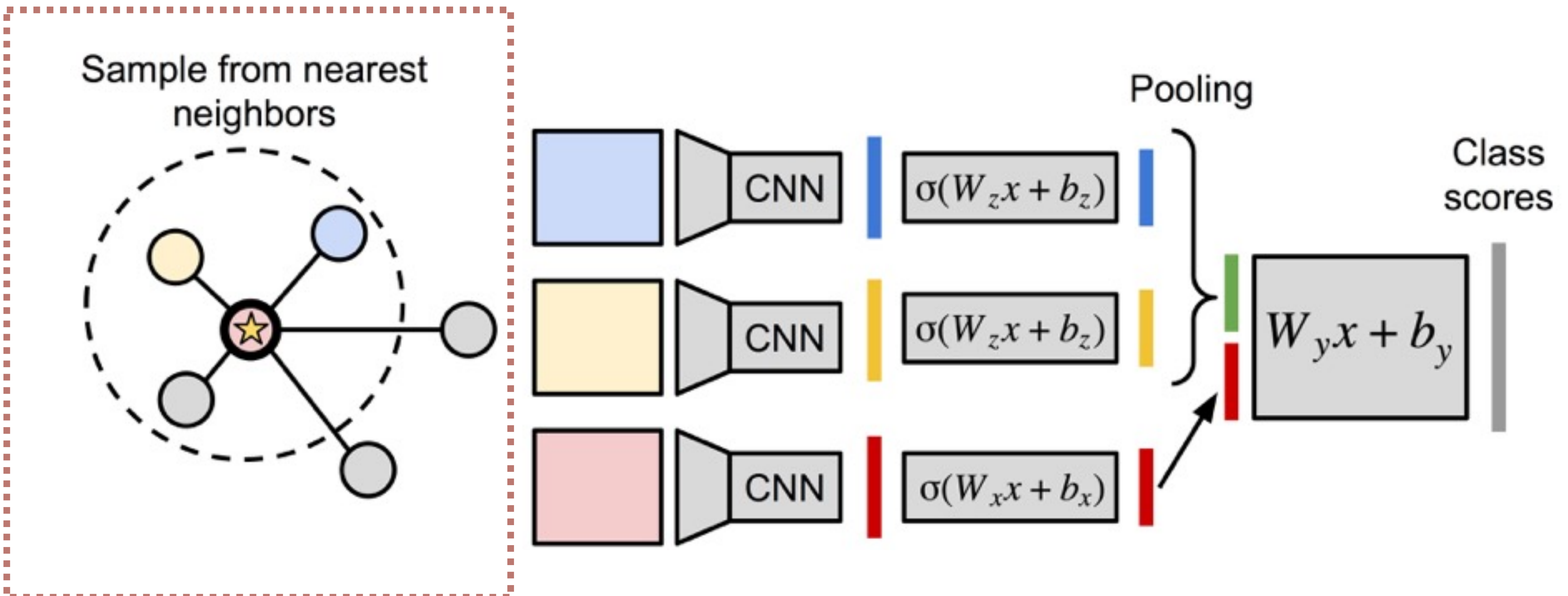




Approach

- For an image $x \in X$ and neighborhood $z \in Z_x$, we use a function f parameterized by w to predict labels
    - We compute hidden state representations for the image and its neighbors
    - Then we operate on the concatenation of these two representations to compute label scores
- We demonstrate that our model can:
    - handle *different types of image metadata*
    - adapt to *changing vocabularies*

Approach

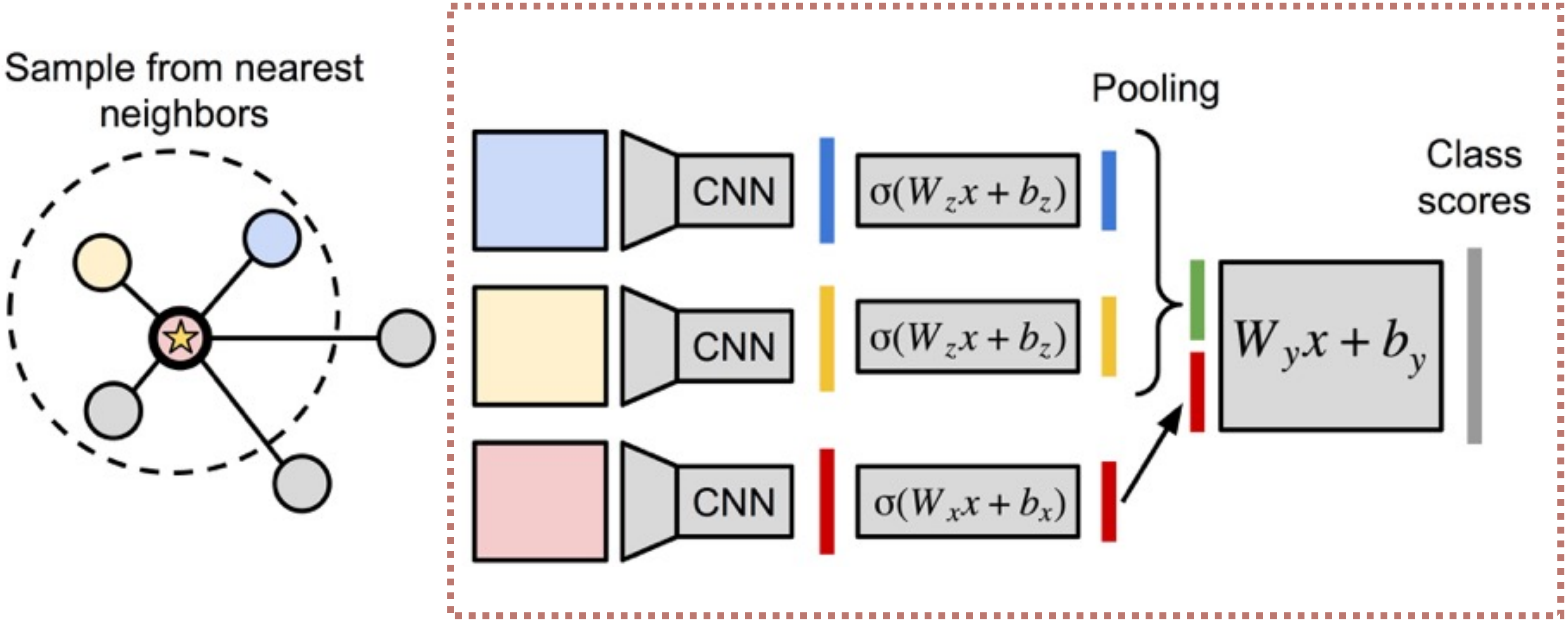
- (1) *non-parametric* step to build a neighborhood



[J.Johnson*, **L.Ballan***, L.Fei-Fei - ICCV 2015]

Approach

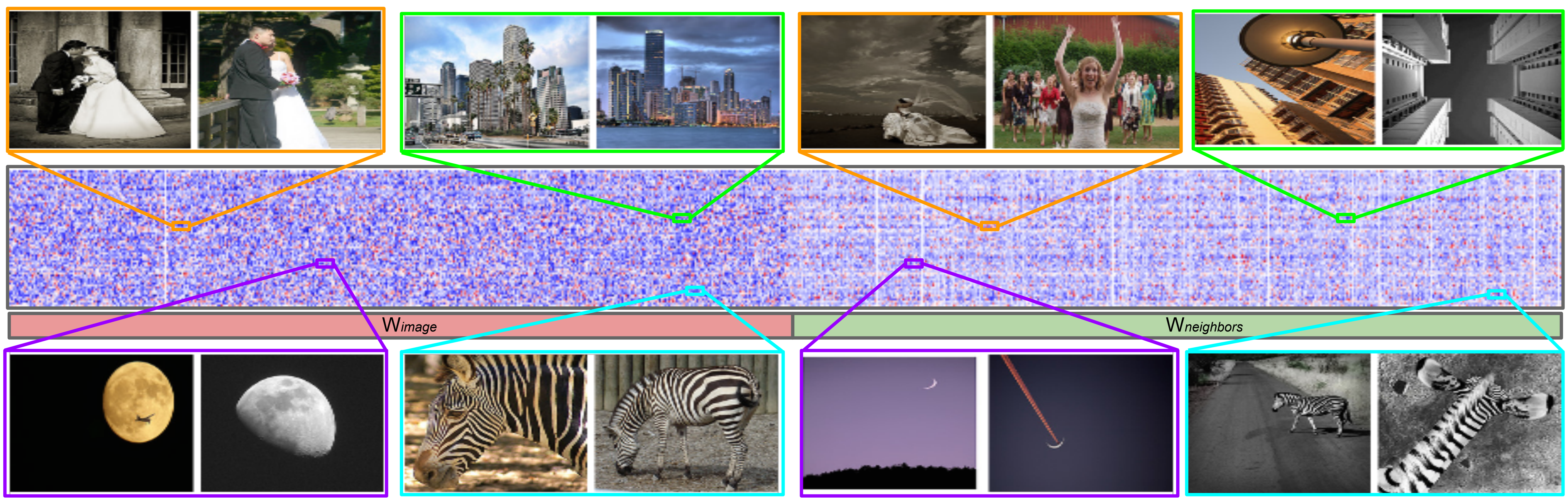
- (2) *deep neural network* to blend visual information from the image and its neighbors



[J.Johnson*, **L.Ballan***, L.Fei-Fei - ICCV 2015]

Approach

- In this way the model uses features from both the image and its neighbors



Results

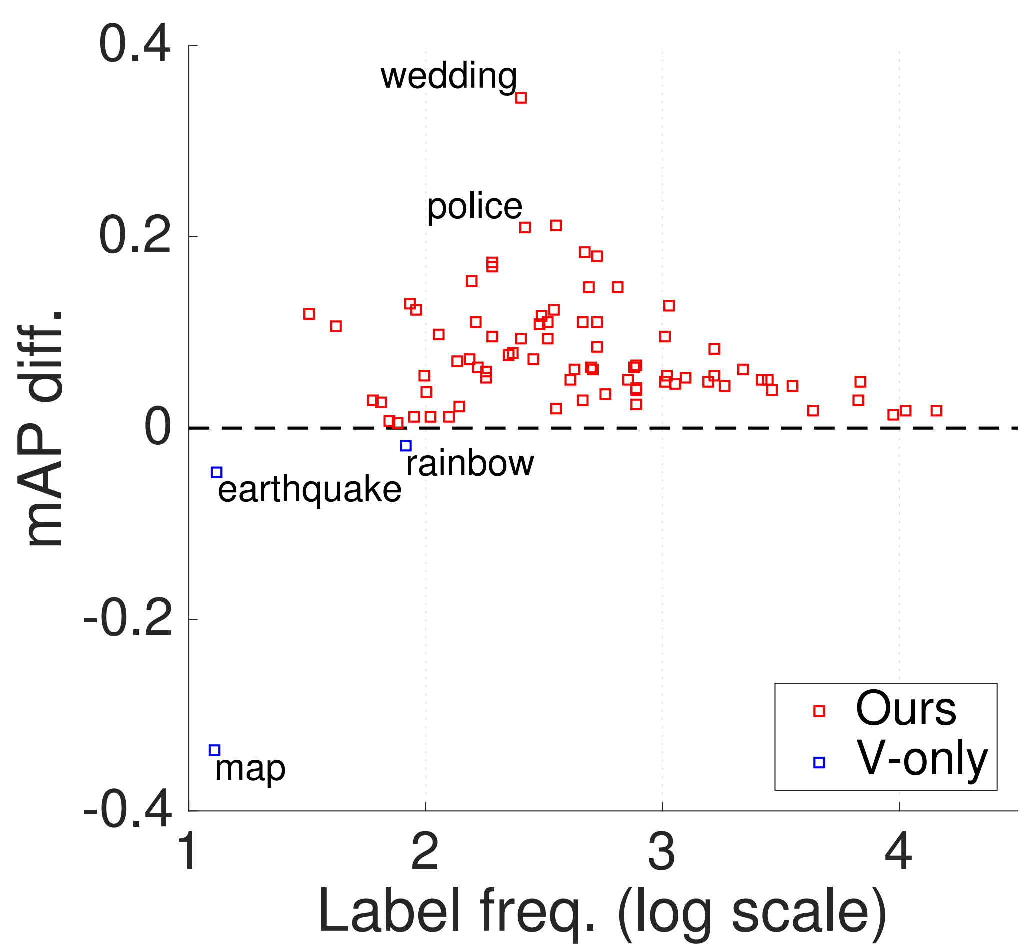
- Multi-label image annotation results on the NUS-WIDE dataset (~240K Flickr images)

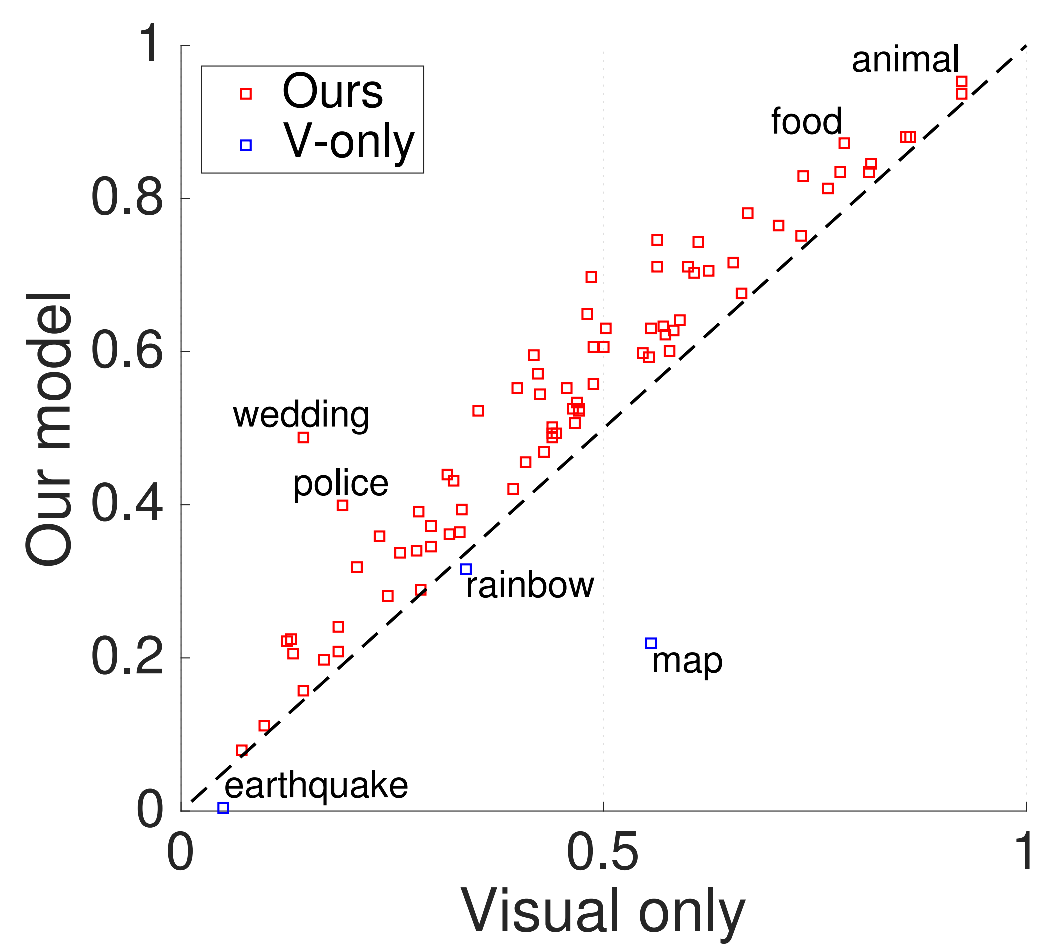
Method	mAP_L	mAP_I	Rec_L	$Prec_L$	Rec_I	$Prec_I$
Tag-only Model + linear SVM [37]	46.67	-	-	-	-	-
Graphical Model (all metadata) [37]	49.00	-	-	-	-	-
CNN + softmax [15]	-	-	31.22	31.68	59.52	47.82
CNN + ranking [15]	-	-	26.83	31.93	58.00	46.59
CNN + WARP [15]	-	-	35.60	31.65	60.49	48.59
Upper bound	100.00±0.00	100.00±0.00	68.52±0.35	60.68±1.32	92.09±0.10	66.83±0.12
Tag-only + logistic	43.88±0.32	77.06±0.14	47.52±2.59	46.83±0.89	71.34±0.16	51.18±0.16
CNN [27] + kNN-voting [36]	44.03±0.26	73.72±0.10	30.83±0.37	44.41±1.05	68.06±0.15	49.49±0.11
CNN [27] + logistic (visual-only)	45.78±0.18	77.15±0.11	43.12±0.39	40.90±0.39	71.60±0.19	51.56±0.11
Image neighborhoods + CNN-voting	50.40±0.23	77.86±0.15	34.52±0.47	**56.05**±1.47	72.12±0.21	51.91±0.20
Our model: tag neighbors	52.78±0.34	**80.34**±0.07	43.61±0.47	46.98±1.01	74.72±0.16	**53.69**±0.13
Our model: tag neighbors + tag vector	**61.88**±0.36	80.27±0.08	**57.30**±0.44	54.74±0.63	**75.10**±0.20	53.46±0.09

Table 2: Results on NUS-WIDE. Precision and recall are measured using $n = 3$ labels per image. Metrics are reported both per-label (mAP_L) and per-image (mAP_I). We run on 5 splits of the data and report mean and standard deviation.

Results: ours *vs* CNN baseline

- Experiment 1: evaluates AP for each label of our model vs the visual-only CNN baseline





Qualitative results

Qualitative results

Neighborhood

Qualitative results

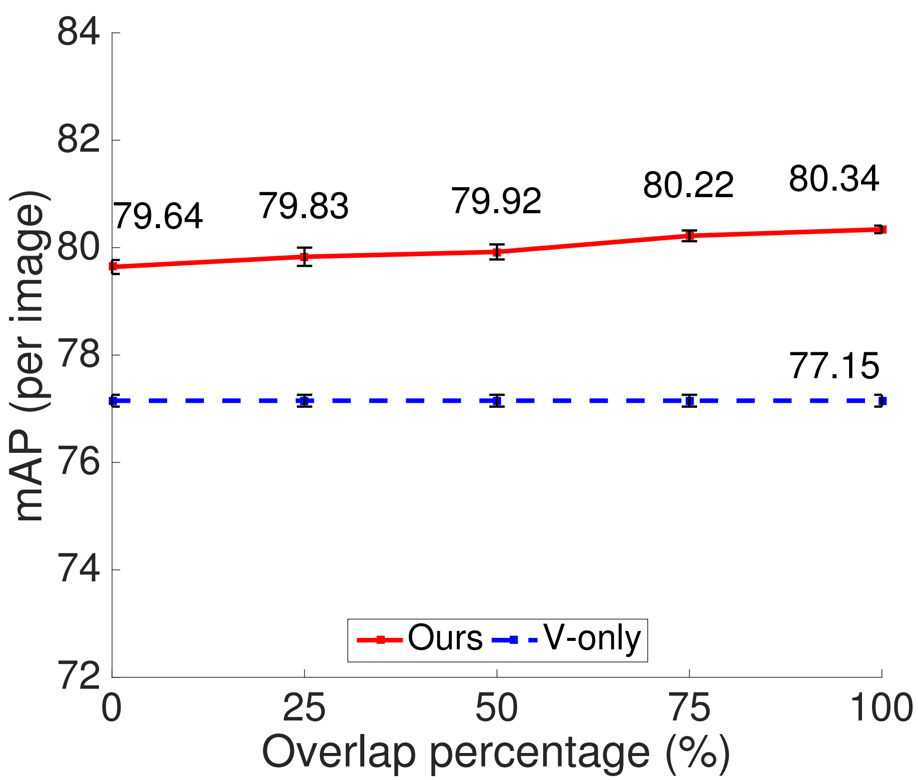
Neighborhood

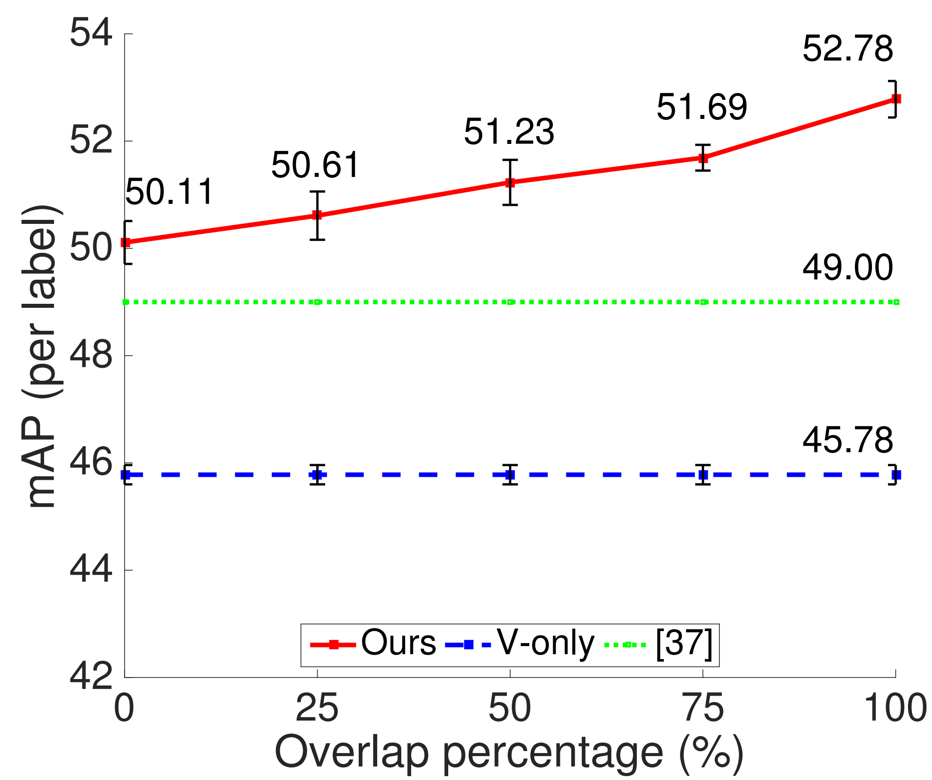
Qualitative results

Neighborhood

Results: generalization

- Experiment 2: vocabulary generalization





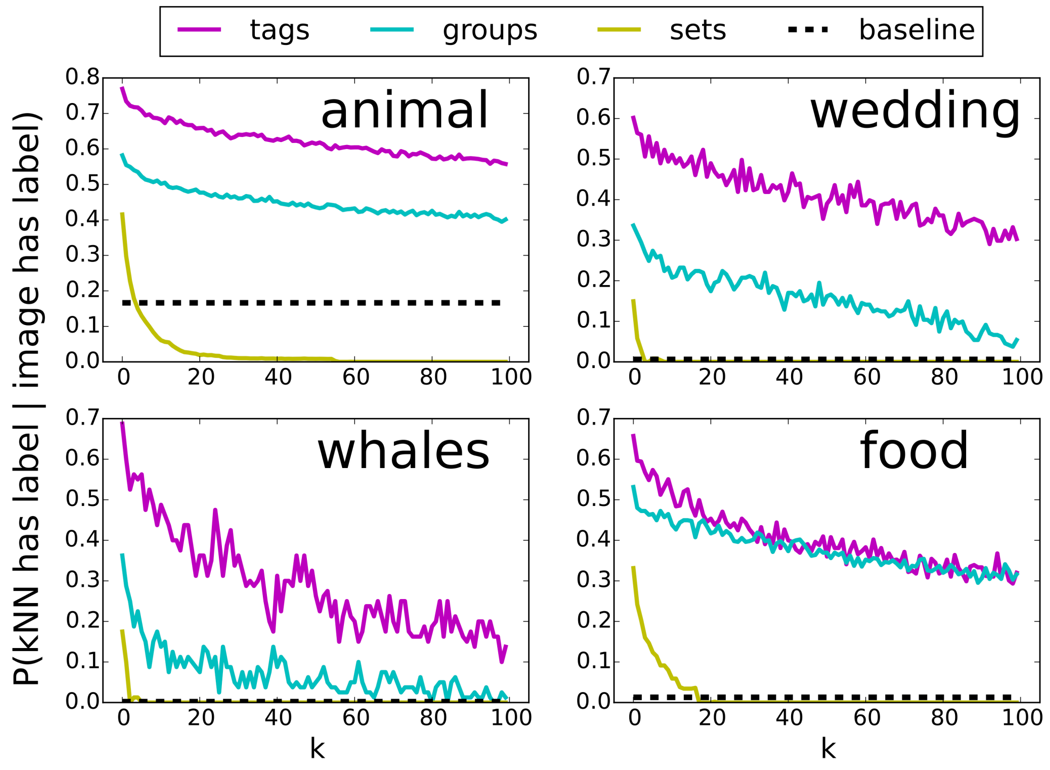
Performance as we vary overlap between tag vocabularies used for training and testing: strong results even in the case of disjoint vocabularies

Results: generalization

- Experiment 3: metadata generalization

Train: \ Test:	Tags	Sets	Groups
Tags	52.78±0.34	47.12±0.35	48.14±0.33
Sets	52.21±0.29	48.02±0.33	48.49±0.16
Groups	50.32±0.28	47.82±0.24	48.87±0.22

Results using different types of metadata for training and testing

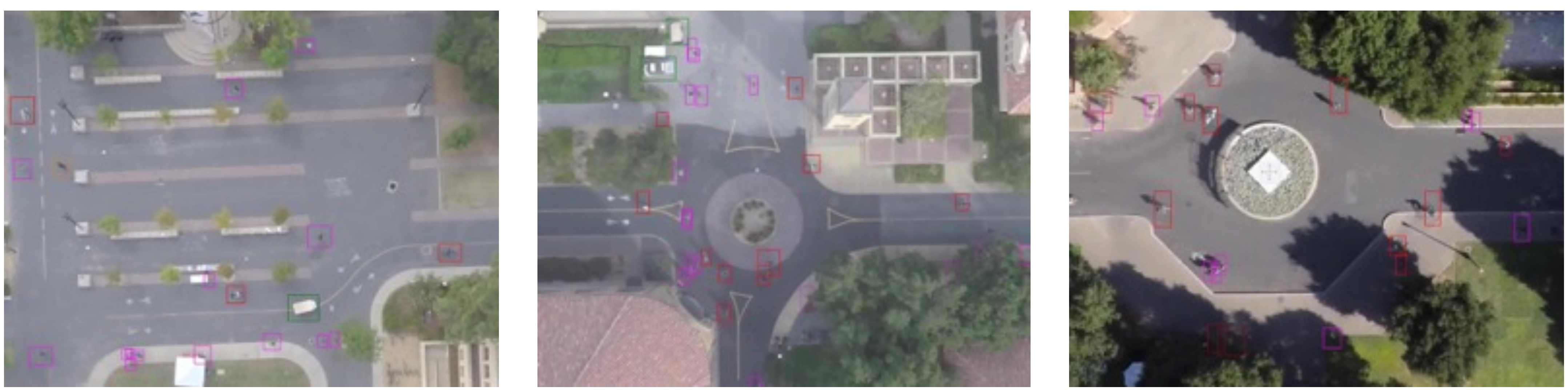


Probability that the k-th neighbor of an image has a label given that the image has the label

Knowledge transfer for scene-specific motion prediction

Humans in crowded spaces

- When humans navigate a crowed space their motion is influenced by the scene and the other active agents
- Stanford Campus Dataset: videos of various agents that navigate in a real world outdoor environment



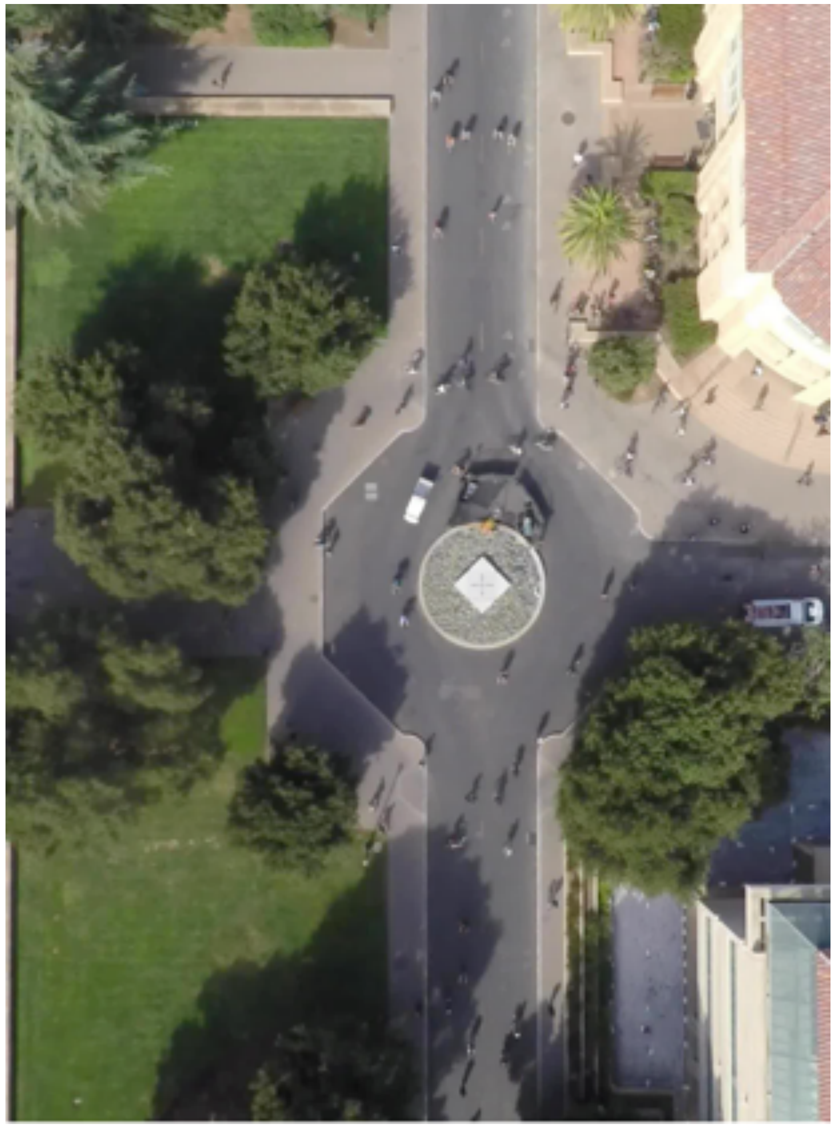
[A.Robicquet, A.Alahi, A.Sadeghian, B.Anenberg, J.Doherty, E.Wu, S.Savarese - arXiv 2016]

Goal: motion (*trajectory*) prediction

- Given a single picture and an observed agent, humans are able to predict the most likely future

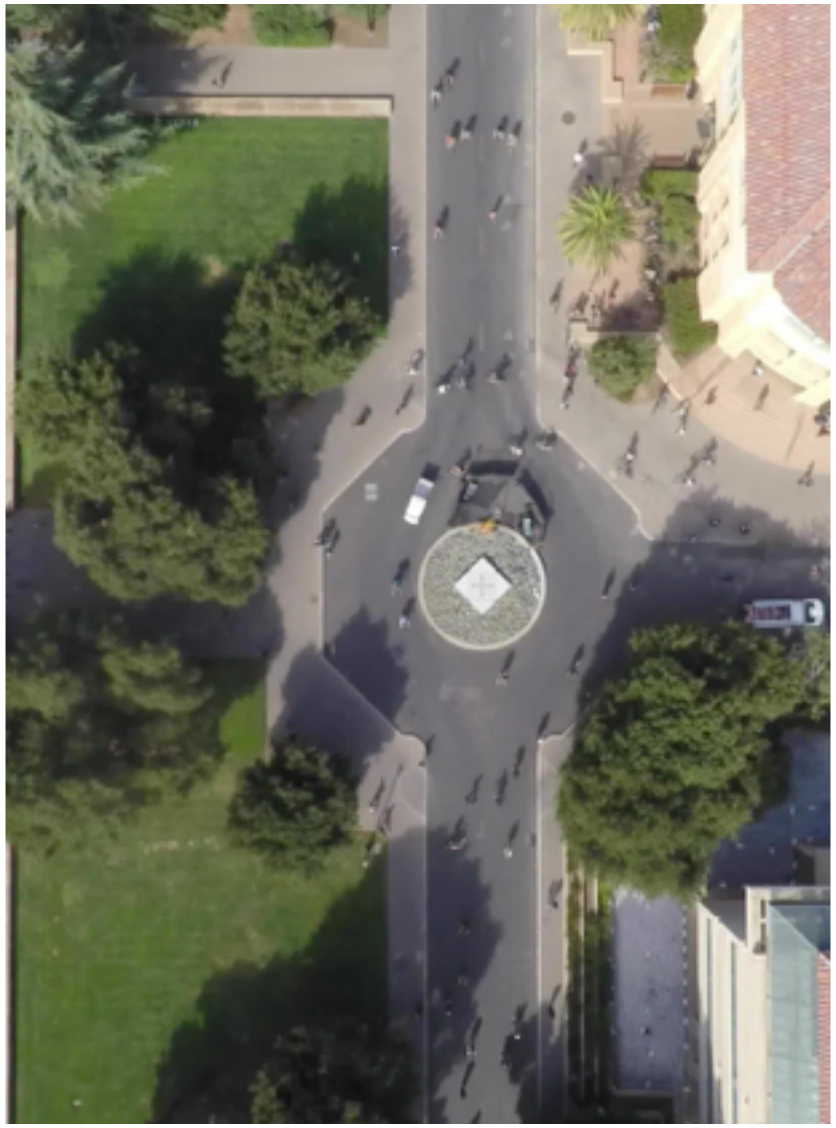
Motivation

- We believe this ability is mostly driven by two factors



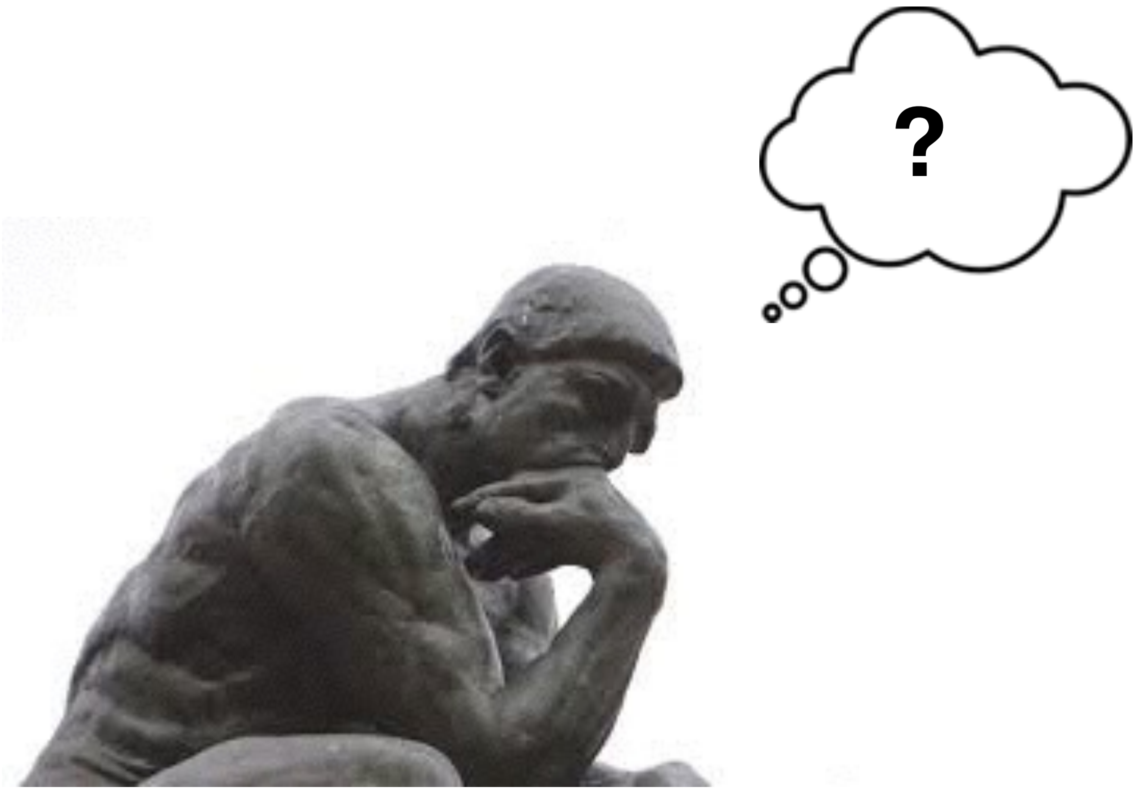
Motivation

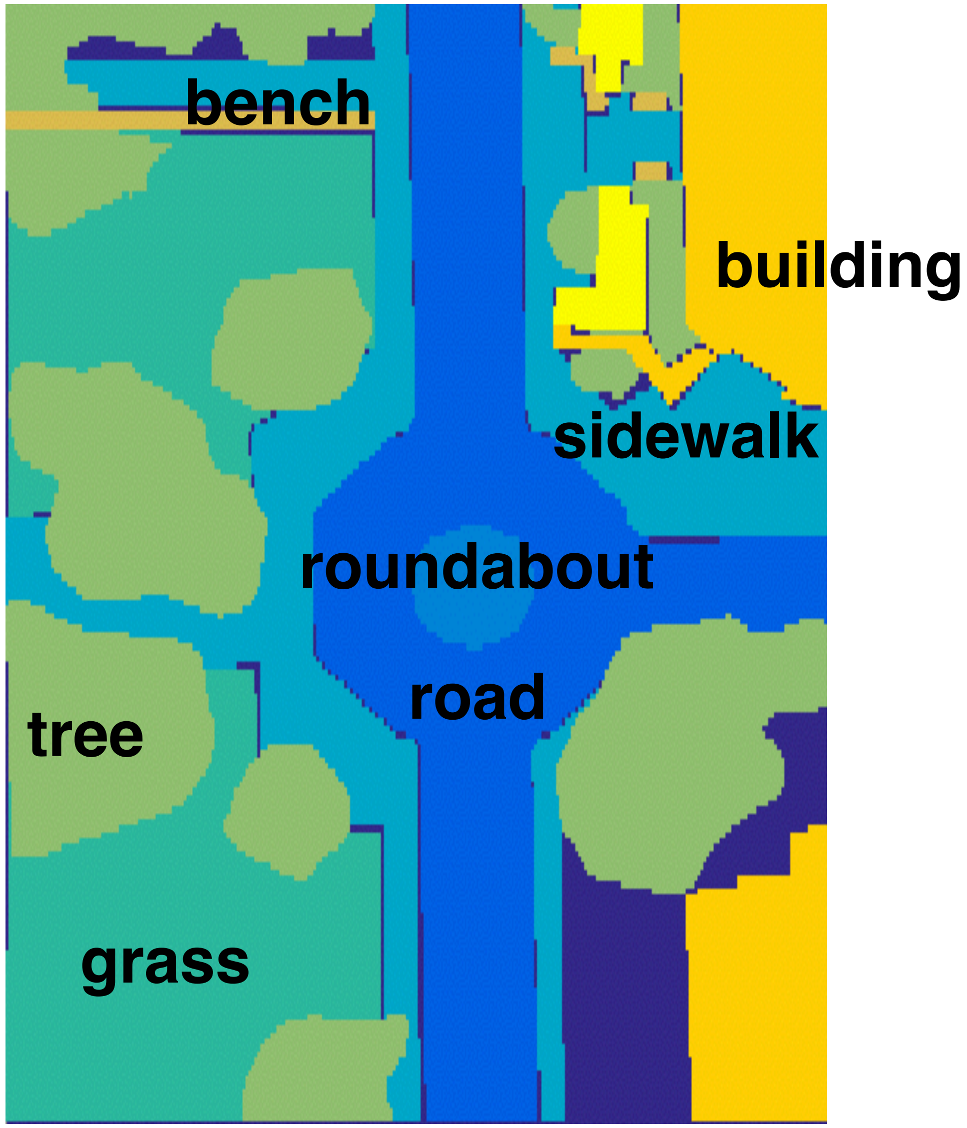
- (1) the *dynamics* of previously observed targets



Motivation

- (2) the *semantic of the scene*



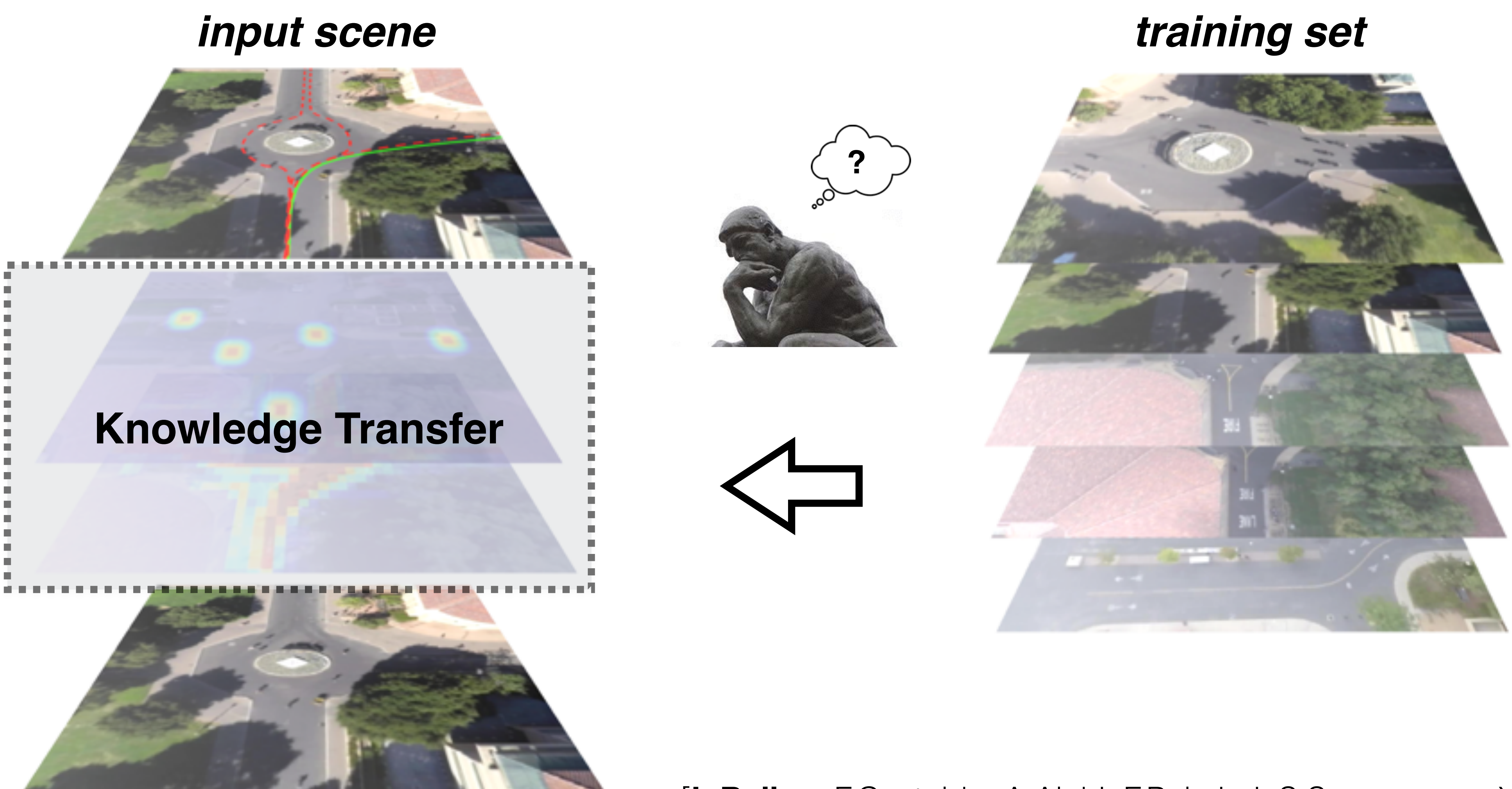


Challenges

- Our model should be able to exploit the interplay between scene semantics and agents
- Data collection is hard and expensive
- Q: how to scale to large dataset / new scenes?

Approach

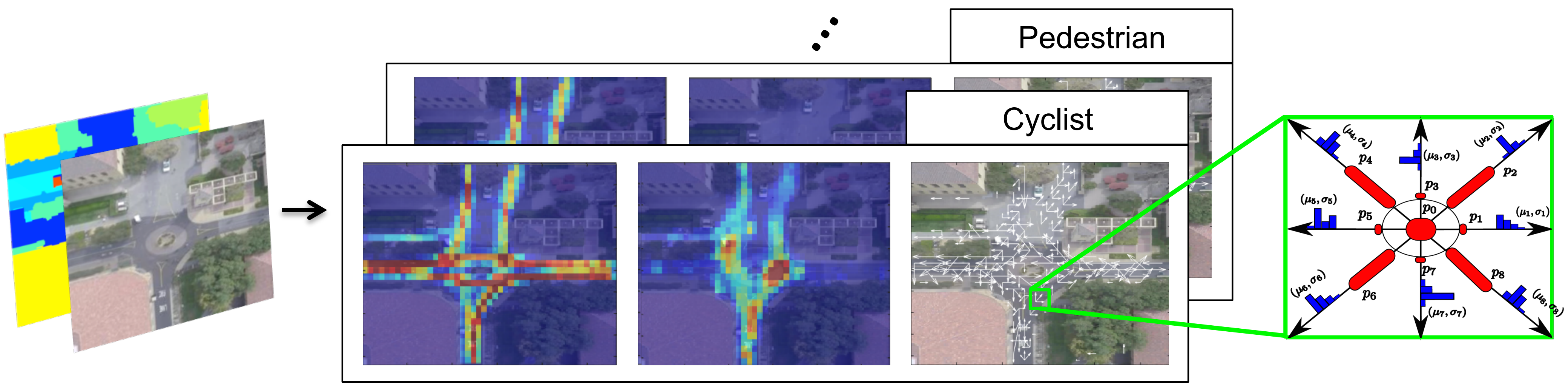
- This knowledge can be transferred to a new scene



[**L.Ballan**, F.Castaldo, A.Alahi, F.Palmieri, S.Savarese - arXiv 2016]

Approach

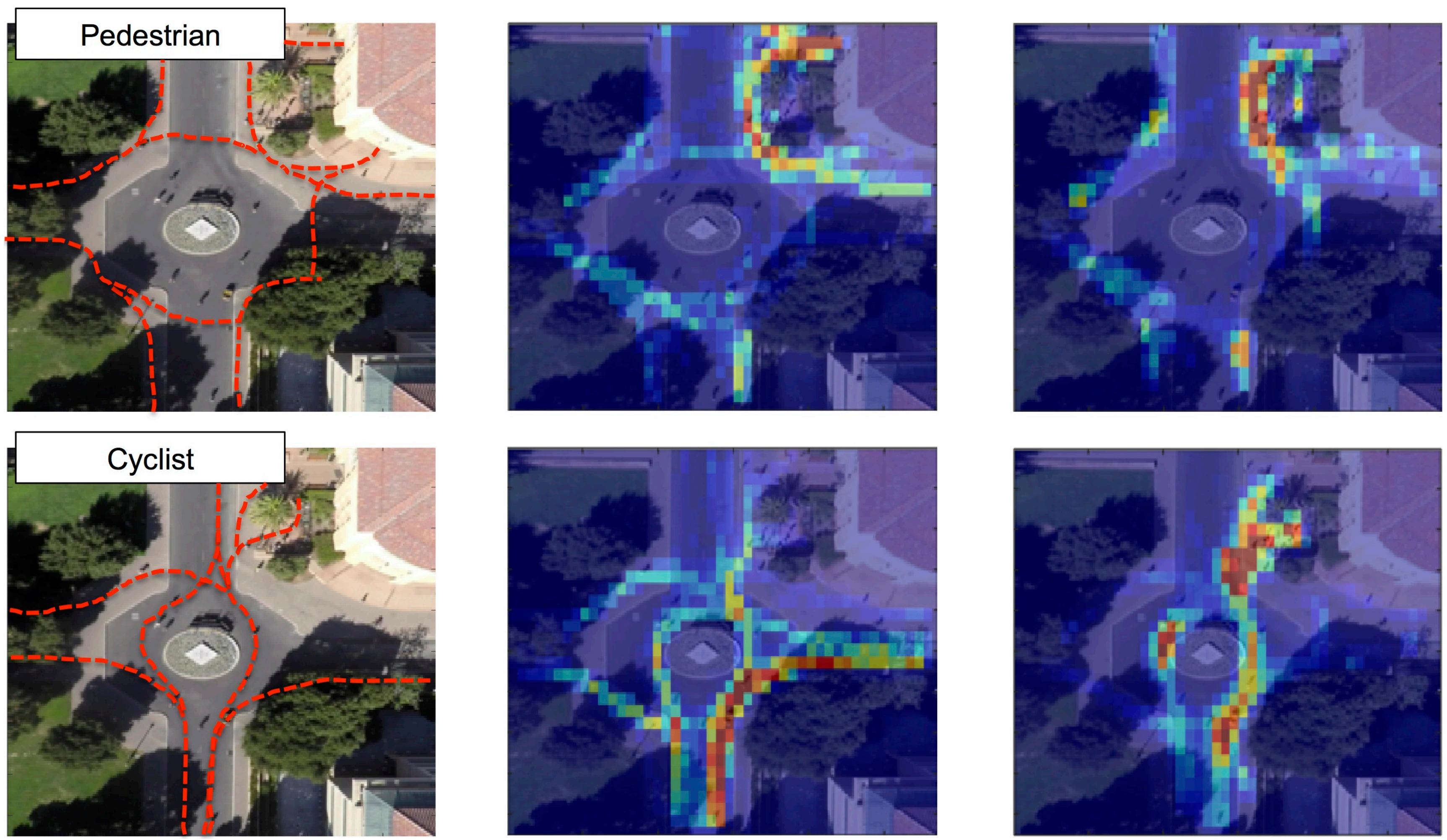
- Given an input scene we build a navigation map $\mathbf{M}$ which collects the navigation statistics
- For each patch in the map we collect:
    - Popularity score, Routing score, Histogram of Directions and Histogram of Speeds



(a) Input scene (b) "Navigation map" (c) *HoD* and *HoS*

Prediction model

- Navigation Map: two qualitative examples

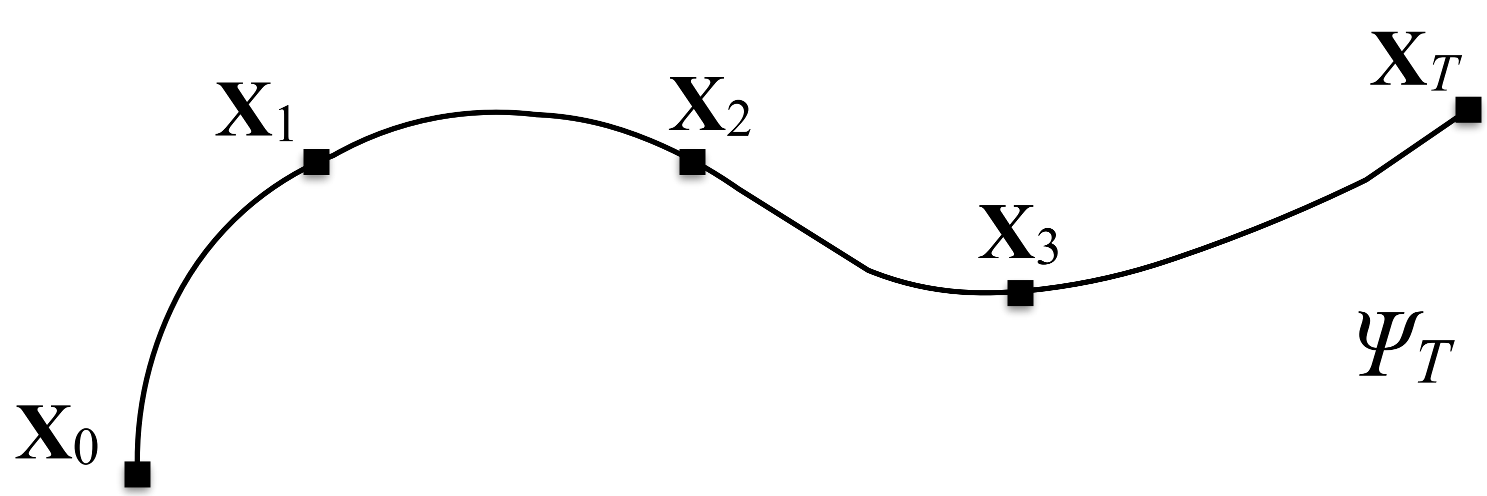


(a) (b) (c)

Column (a) visualizes the most common paths for both classes. Columns (b,c) show the corresponding popularity and routing maps.

Prediction model

- The target state variable is defined as $\mathbf{X}_k = (\mathbf{P}_k, \mathbf{V}_k)^T$
  - $\mathbf{P}_k = (X_k, Y_k)^T$ (position) and $\mathbf{V}_k = (\Omega_k, \Theta_k)^T$ (velocity)
- The target interacts with the map $\mathbf{M}$ by exploiting the navigation values for the patch he is occupying
- Given an initial condition $\mathbf{X}_0$, our goal is to generate a sequence of future states $\mathbf{X}_1, \ldots, \mathbf{X}_T$, i.e. a path Ψ_T



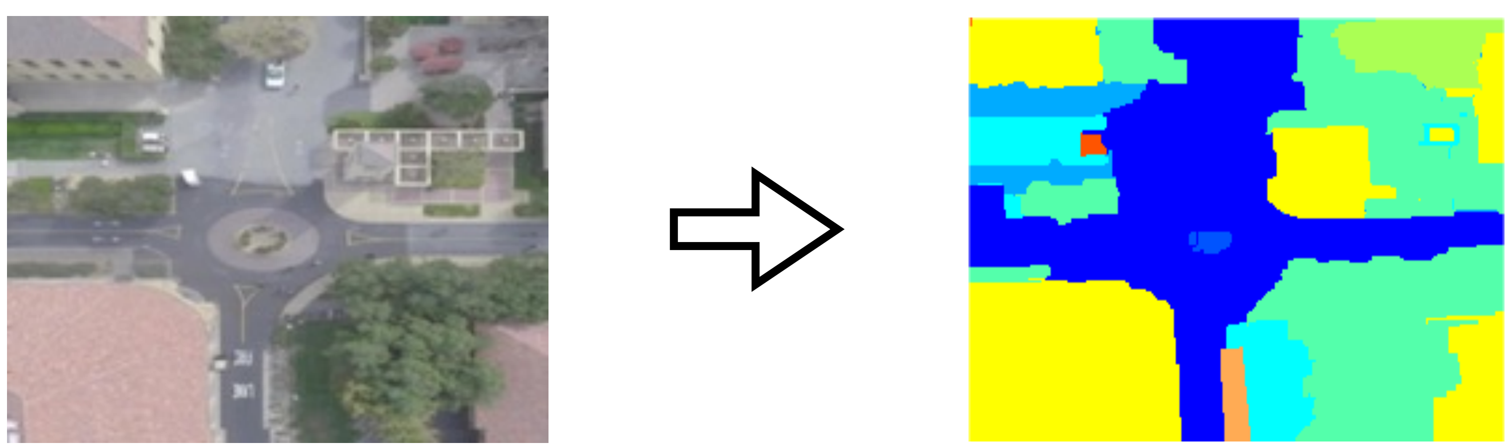
Prediction model

- The dynamic process describing the target motion is defined by:
  - $\mathbf{P}_{k+1} = \mathbf{P}_k + (\Omega_k \cos\Theta_k, \Omega_k \sin\Theta_k)' + \mathbf{w}_k$ (constant velocity)
  - $\mathbf{V}_{k+1} = \Phi(\mathbf{P}_k, \mathbf{V}_k; \mathbf{M})$
- The learned expected values in $\mathbf{M}$ allows our model to generate non-linear behaviors
- $\Phi(\cdot)$ is defined in probabilistic terms by means of a Dynamic Bayesian Network (DBN)

Knowledge transfer

“The elements of the scene define a semantic context, and they might determine similar behaviors in scenes characterized by a similar context”

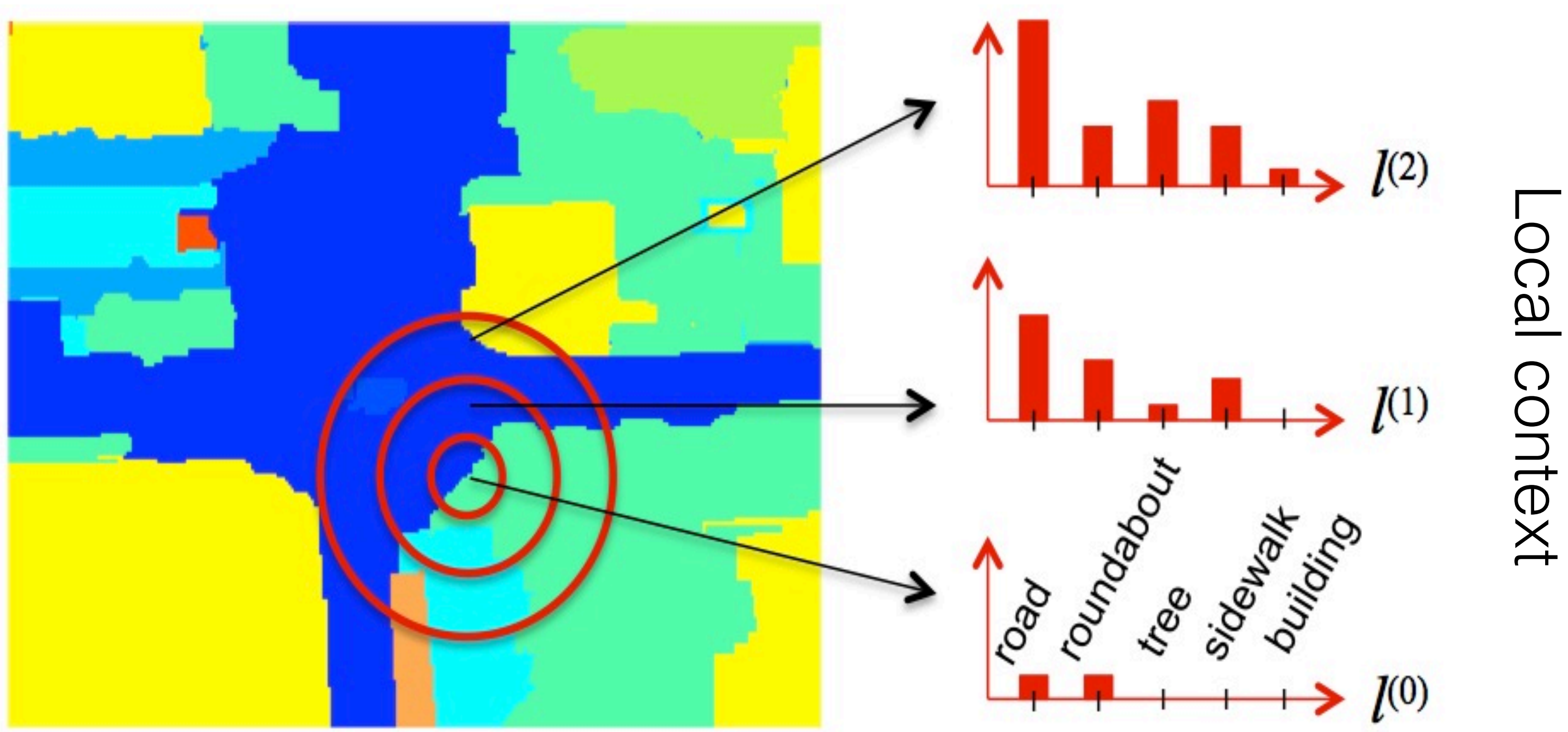
- Our data-driven approach uses scene similarity to transfer the functional properties to a new scene
- Scene parsing: we use a “non-parametric” algorithm (based on SIFT+LLC, GIST and MRF inference)



[J.Yang, B.Price, S.Cohen, M.Yang - CVPR 2014]

Knowledge transfer

- Context Descriptors: a weighted concatenation of the *global* and *local* semantic context components
    - *global context*: vector of distances between classes
    - *local context*: encodes the spatial configuration of nearby patches at multiple levels



Results: datasets

- **UCLA-courtyard**: 6 videos, 2 scenes, single-class (pedestrian), scene labeled with 8 semantic classes
- **Stanford-UAV**: 21 videos,15 scenes, multi-class (pedestrian and cyclist), scene labeled with 10 semantic classes
- **Evaluation metric:** *Modified Hausdorff Distance (MHD)* to measure the pixel distance between ground-truth trajectories and predicted paths

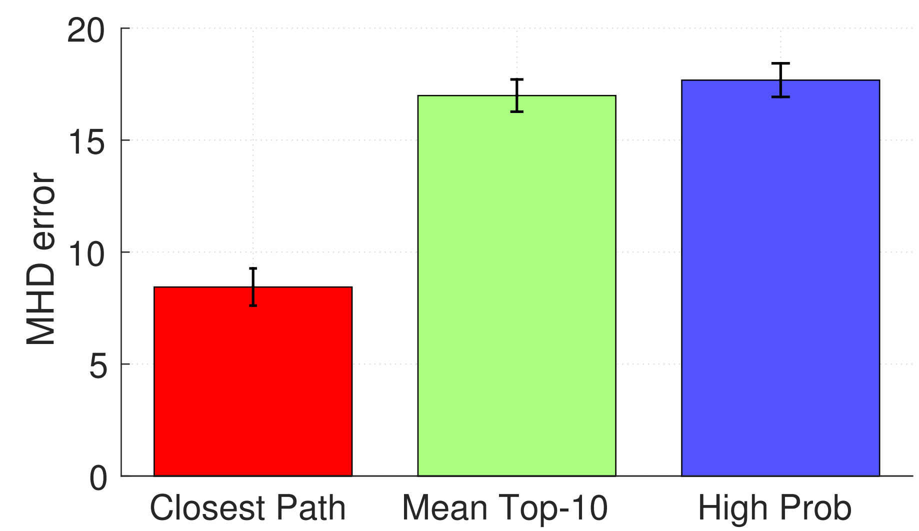
Results: path prediction

- Experiment 1: evaluates the ability of the proposed model to predict long-term trajectories

MHD error

	UCLA-courtyard	*Stanford-UAV*
LP	41.36±0.98	31.29±1.25
LP_{CA}	-	21.30±0.80
IOC [2]	14.47±0.77	14.02±1.13
SFM [14]	-	12.10±0.60
Ours	**10.32**±0.51	**8.44**±0.72

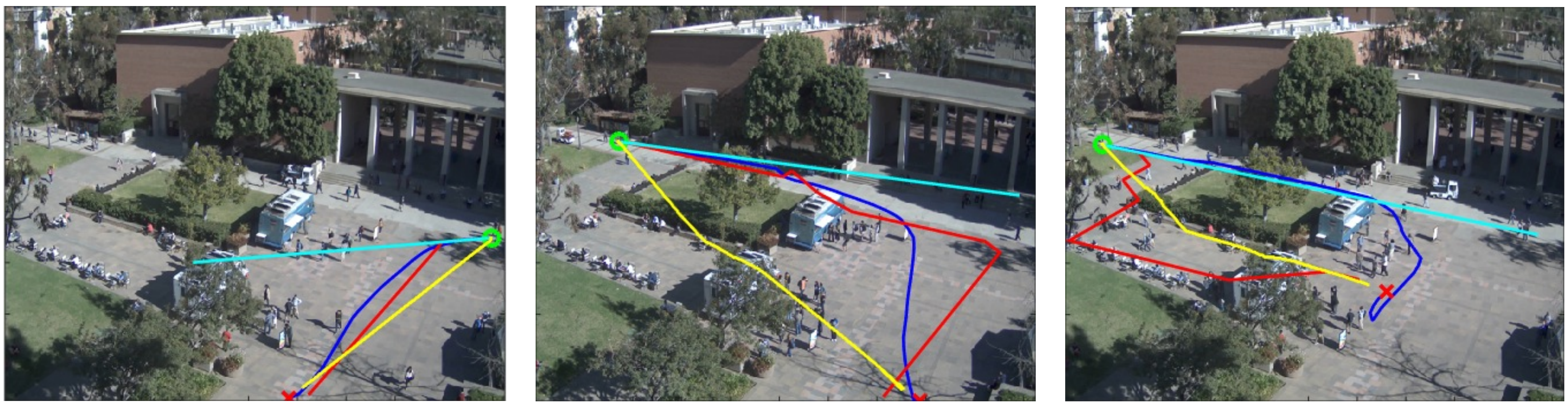
(a) MHD error for a given final destination



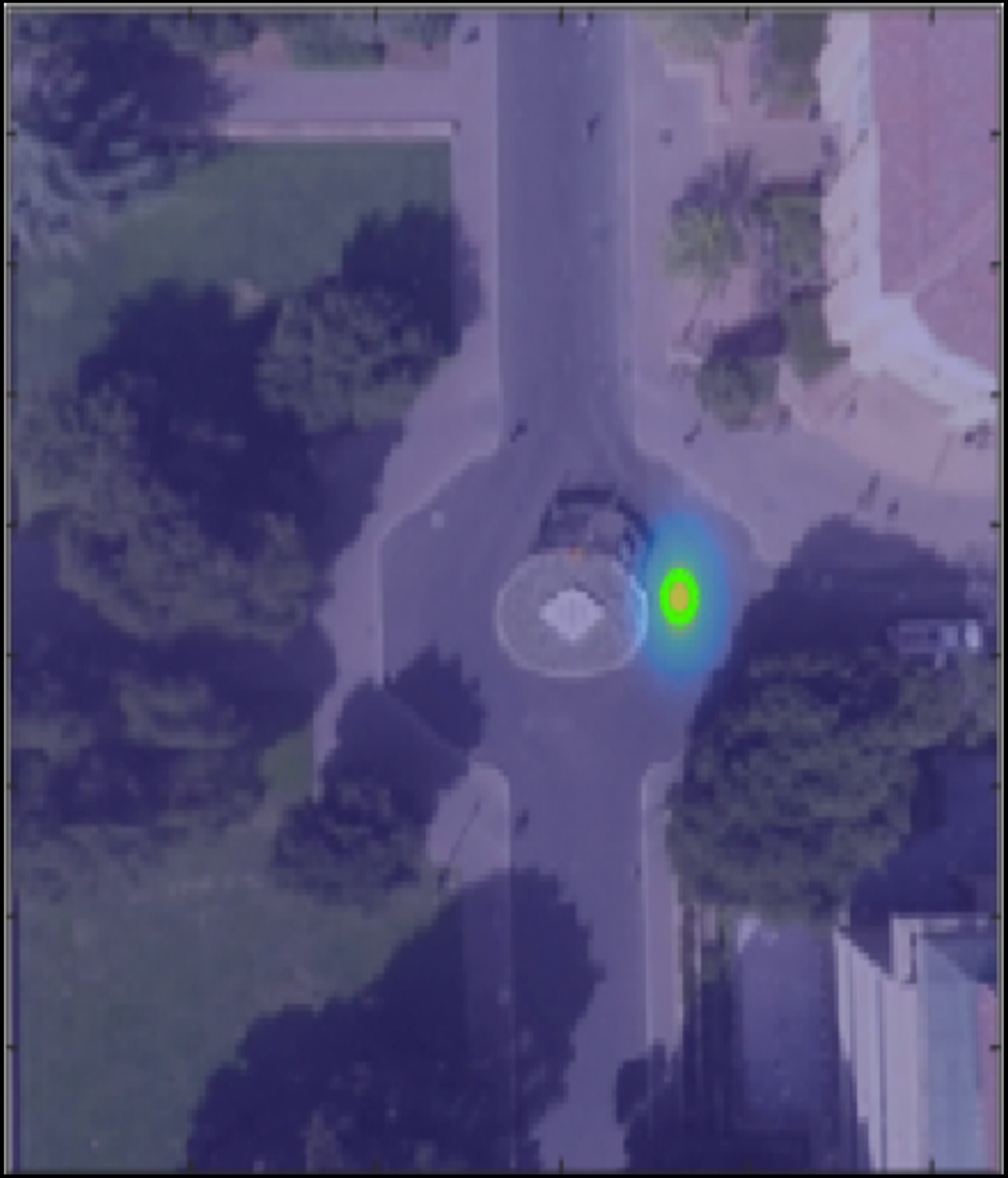
(b) Path generation strategies (ours)

Results: path prediction

- Qualitative examples on the UCLA-courtyard dataset
  (*blue* is ground-truth, *cyan* is LP, *yellow* is IOC, *red* is ours)




starting point
heatmap
path prediction



starting point

heatmap

path prediction

starting point

heatmap

path prediction

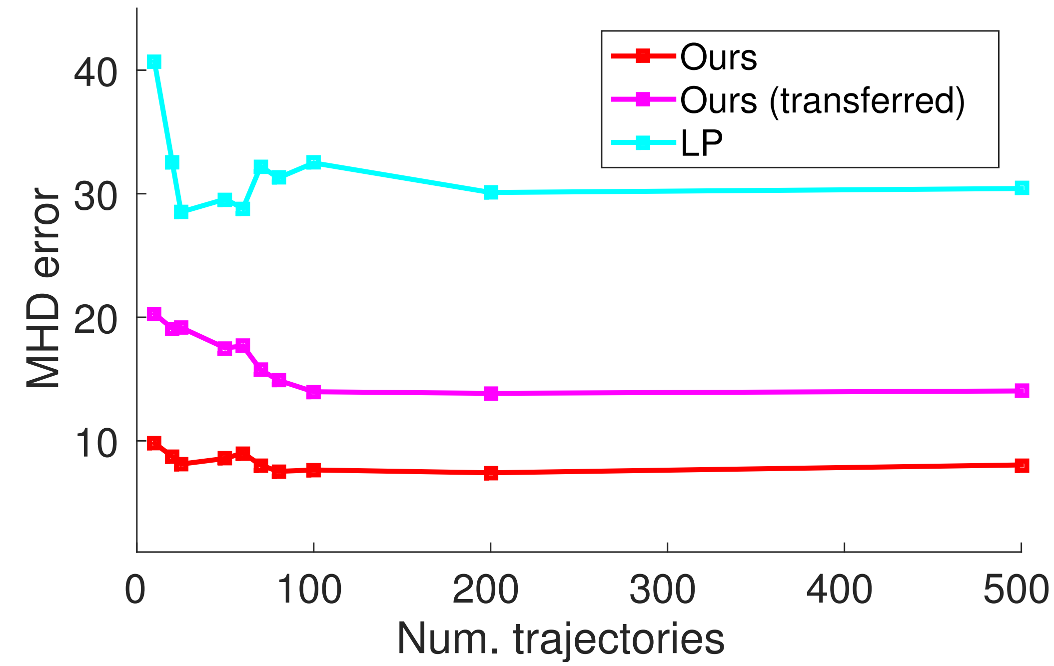
Results: knowledge transfer

- Experiment 2: evaluates the ability of our model to generalize and make predictions on novel scenes

MHD error

	Pedestrian	*Cyclist*	Overall
LP	34.07	26.15	31.29±1.25
IOC [2]	17.99	18.84	18.42±0.97
Ours	12.36	17.10	**14.29**±0.84

(a) Path prediction



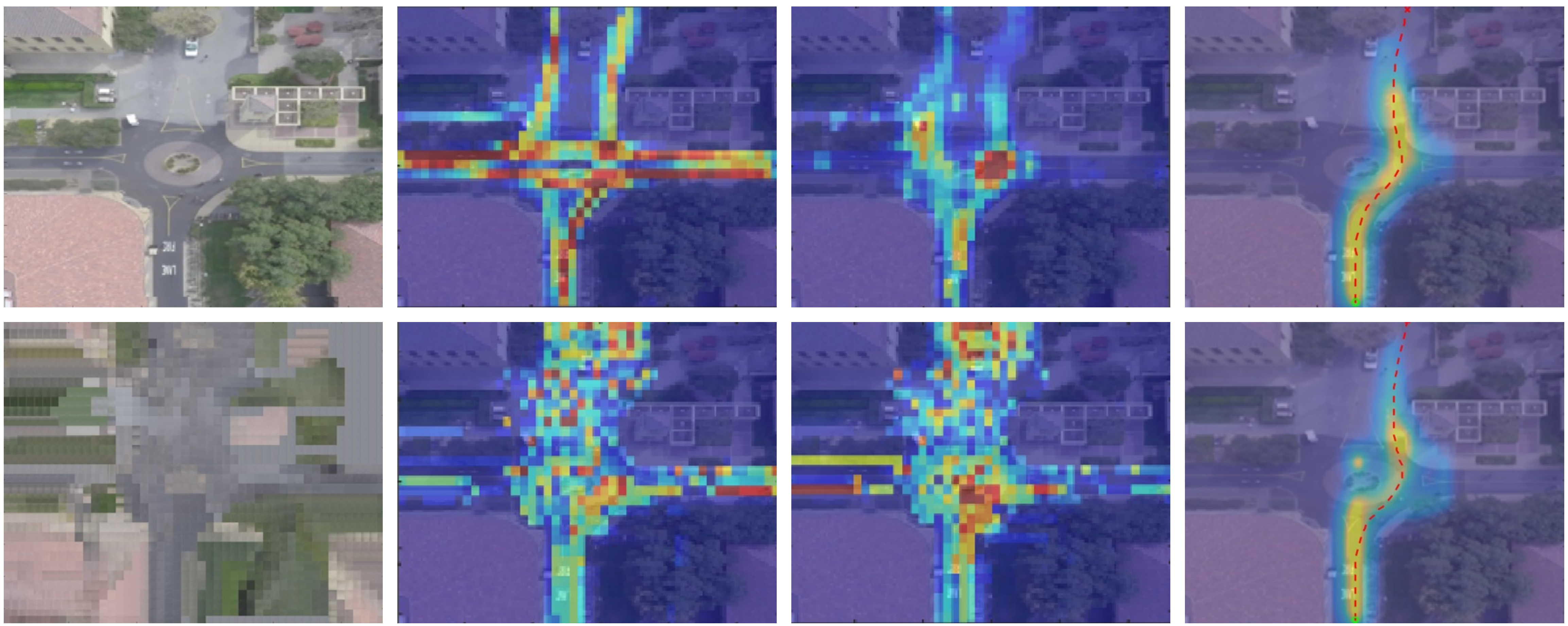
(b) Impact of training data

Results: knowledge transfer

- Qualitative examples: (1) path forecasting *vs.* (2) knowledge transfer

(1)

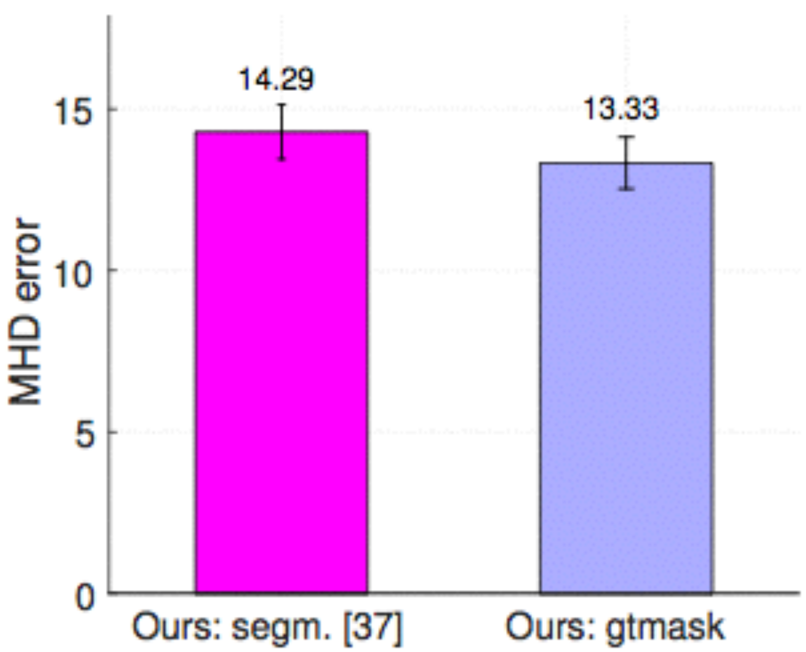
(2)



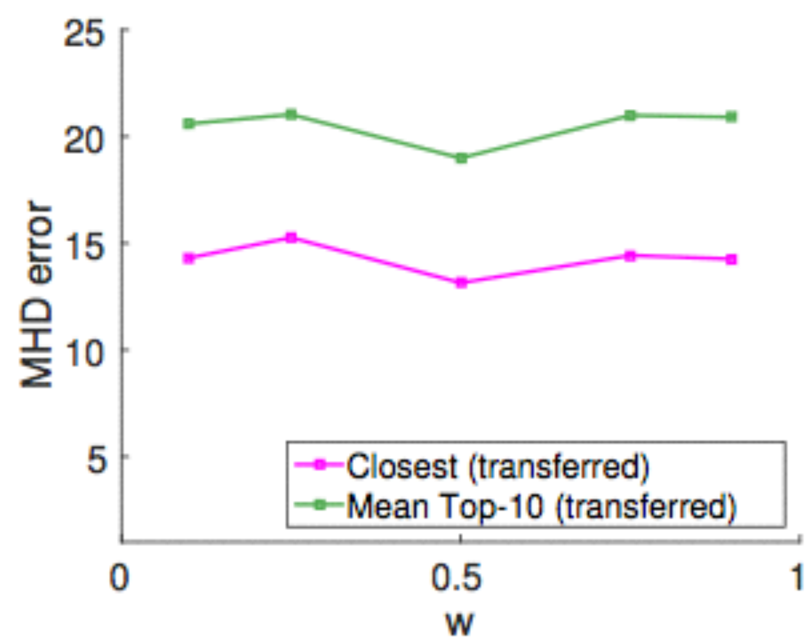
Results: impact of the parameters

- How the performance obtained with knowledge transfer is influenced by the different parameters?

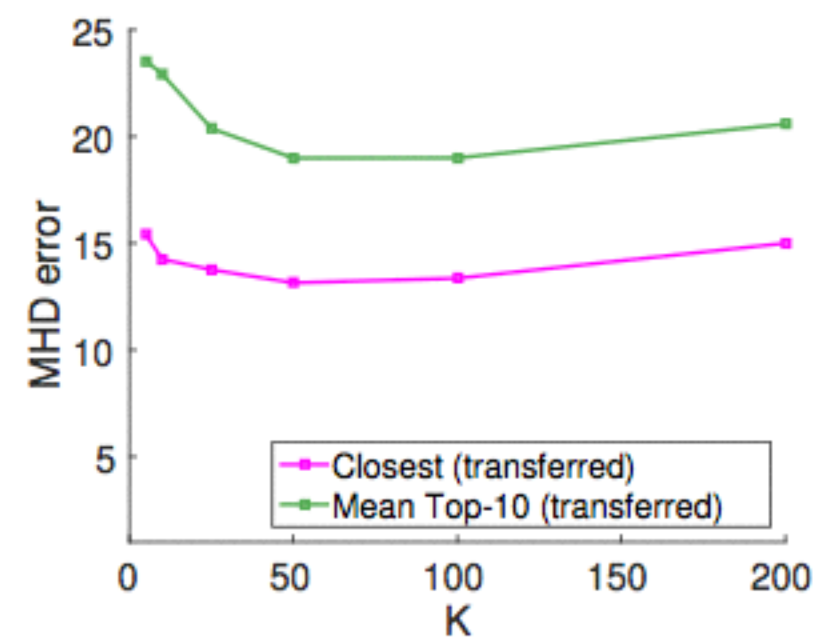
(a) image parsing



(b) context descriptors



(c) KNN / retrieval

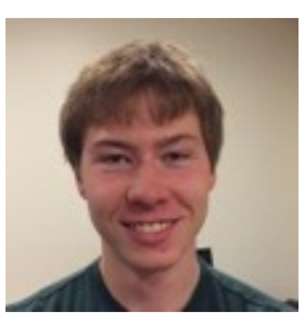


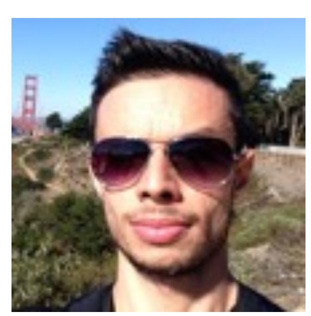
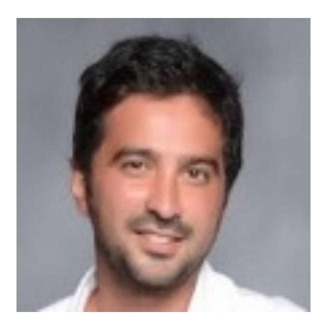
Contact Info

lballan@cs.stanford.eu

www.lambertoballan.net

Thanks!



Collaborators

References

[1] J. Johnson*, **L. Ballan***, L. Fei-Fei, “Love Thy Neighbors: Image Annotation by Exploiting Image Metadata”, ICCV 2015 (* equal contribution)

[2] **L. Ballan**, F. Castaldo, A. Alahi, F. Palmieri, S. Savarese, “Knowledge Transfer for Scene-specific Motion Prediction”, ECCV 2016

Collaborators