Learning image representations from unlabeled video

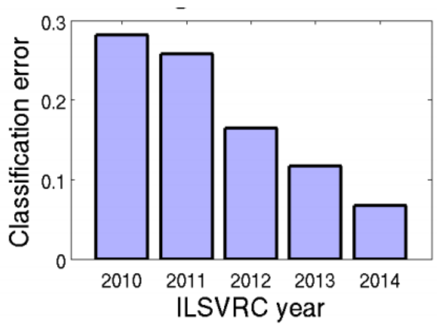
Kristen Grauman
Department of Computer Science
The University of Texas at Austin

Work with Dinesh Jayaraman

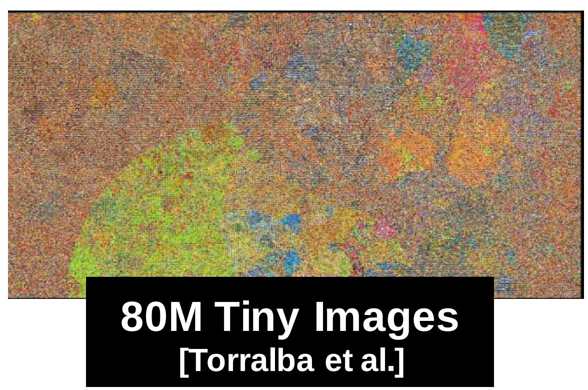
THE UNIVERSITY OF TEXAS AT AUSTIN

Learning visual categories

- Recent major strides in category recognition



- Facilitated by large labeled datasets



[Papageorgiou & Poggio 1998, Viola & Jones 2001, Dalal & Triggs 2005, Grauman & Darrell 2005, Lazebnik et al. 2006, Felzenszwalb et al. 2008, Krizhevsky et al. 2012, Russakovsky IJCV 2015…]

Big picture goal: Embodied vision

Status quo:

Learn from "disembodied" bag of labeled snapshots.

Our goal:

Learn in the context of acting and moving in the world.

Beyond "bags of labeled images"?

Visual development in nature is based on:

- *continuous observation*
- *multi-sensory feedback*
- *motion and action*

… in an environment.

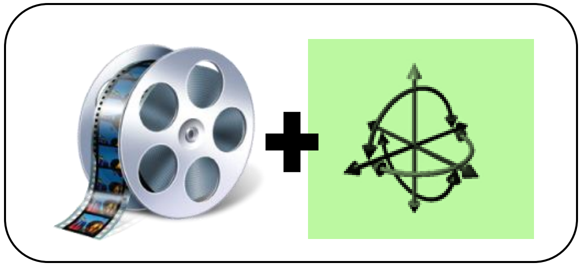
Inexpensive, and unrestricted in scope

Evidence from: psychology, evolutionary biology, cognitive science.

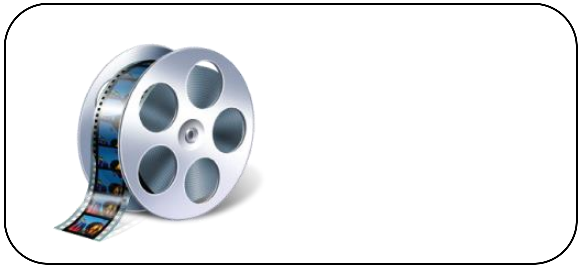
[Held et al, 1964][Moravec et al, 1984][Wilson et al, 2002]

Talk overview

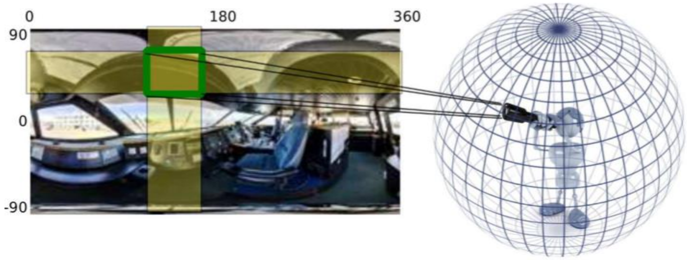
1. Learning representations tied to ego-motion



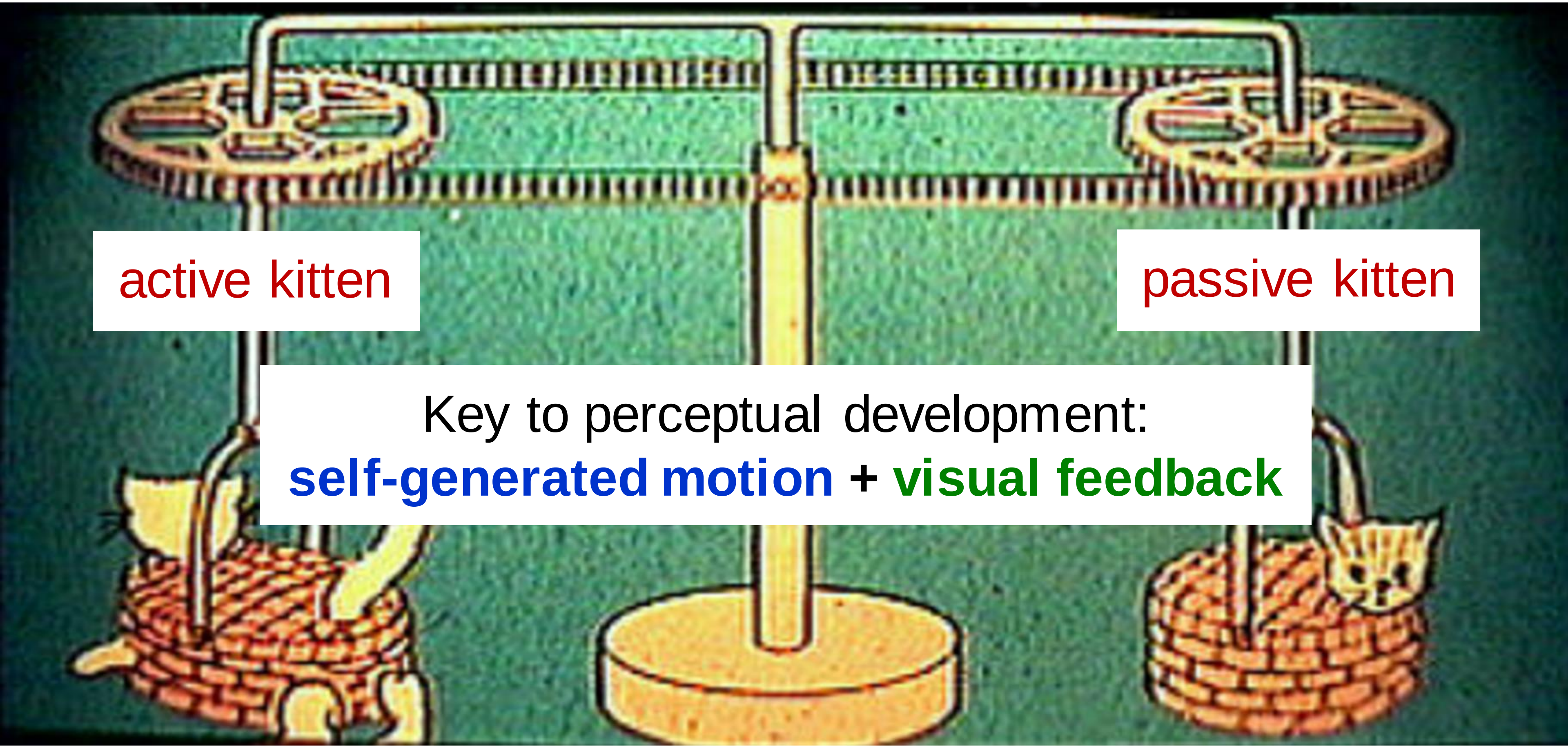
2. Learning representations from unlabeled video



3. Learning how to move and where to look



The kitten carousel experiment
[Held & Hein, 1963]



Key to perceptual development:
self-generated motion + visual feedback

Our idea: Ego-motion ↔ vision

Goal: Teach computer vision system the connection:
"how I move" ↔ "how my visual surroundings change"

Ego-motion motor signals

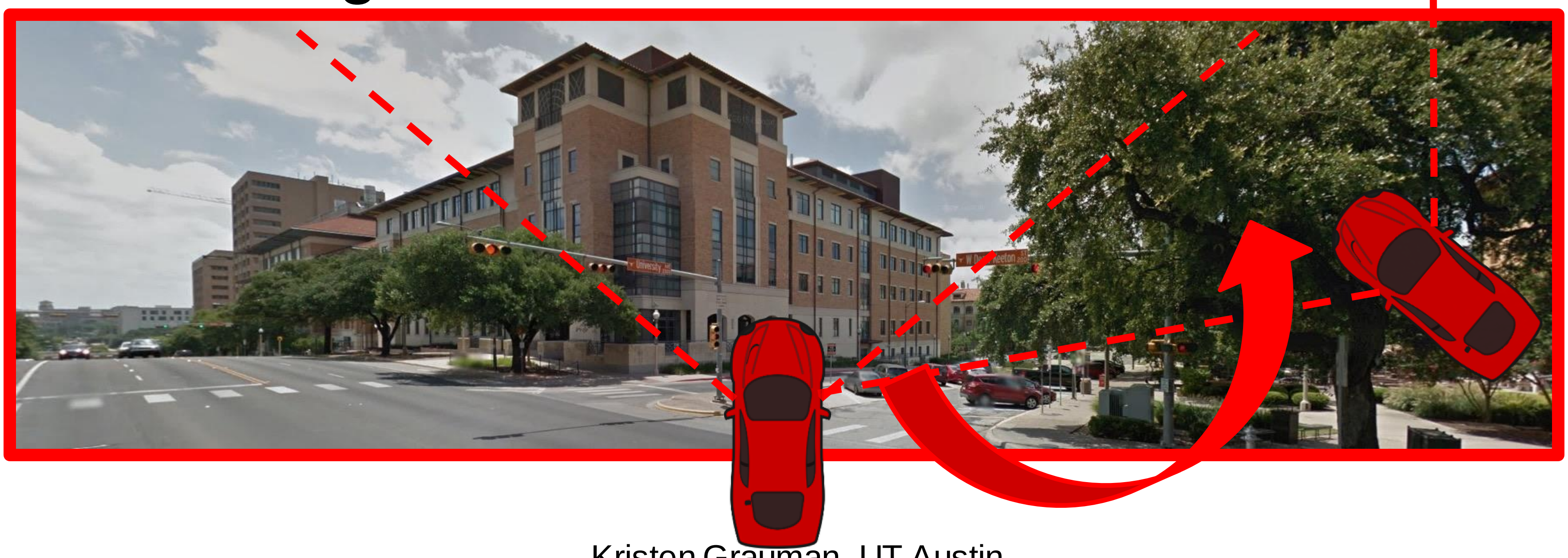
+

Unlabeled video

Kristen Grauman, UT Austin

Ego-motion ↔ vision: view prediction

After moving:



Ego-motion ↔ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

Also key to recognition!

Can be learned without manual labels!

Our approach: unsupervised feature learning using egocentric video + motor signals

Approach idea: Ego-motion equivariance

Invariant features: unresponsive to some classes of transformations

$$\mathbf{z}(g\mathbf{x}) \approx \mathbf{z}(\mathbf{x})$$

Simard et al, Tech Report, '98
Wiskott et al, Neural Comp '02
Hadsell et al, CVPR '06
Mobahi et al, ICML '09
Zou et al, NIPS '12
Sohn et al, ICML '12
Cadieu et al, Neural Comp '12
Goroshin et al, ICCV '15
Lies et al, PLoS computation biology '14
…

Approach idea: Ego-motion equivariance

Invariant features: unresponsive to some classes of transformations

$$\mathbf{z}(g\mathbf{x}) \approx \mathbf{z}(\mathbf{x})$$

Equivariant features: *predictably* responsive to some classes of transformations, through simple mappings (e.g., linear)

"equivariance map"

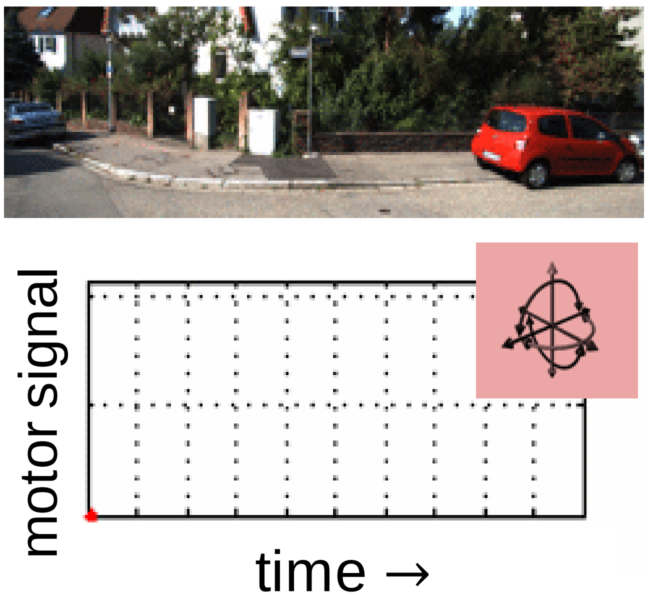
$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

Invariance *discards* information;
equivariance *organizes* it.

Approach idea: Ego-motion equivariance

Training data

Unlabeled video +
motor signals



Learn

Equivariant embedding

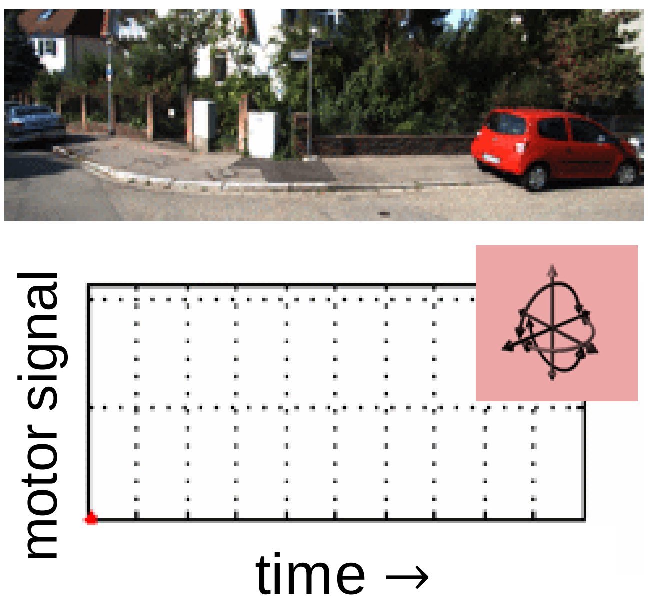
organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Approach idea: Ego-motion equivariance

Training data

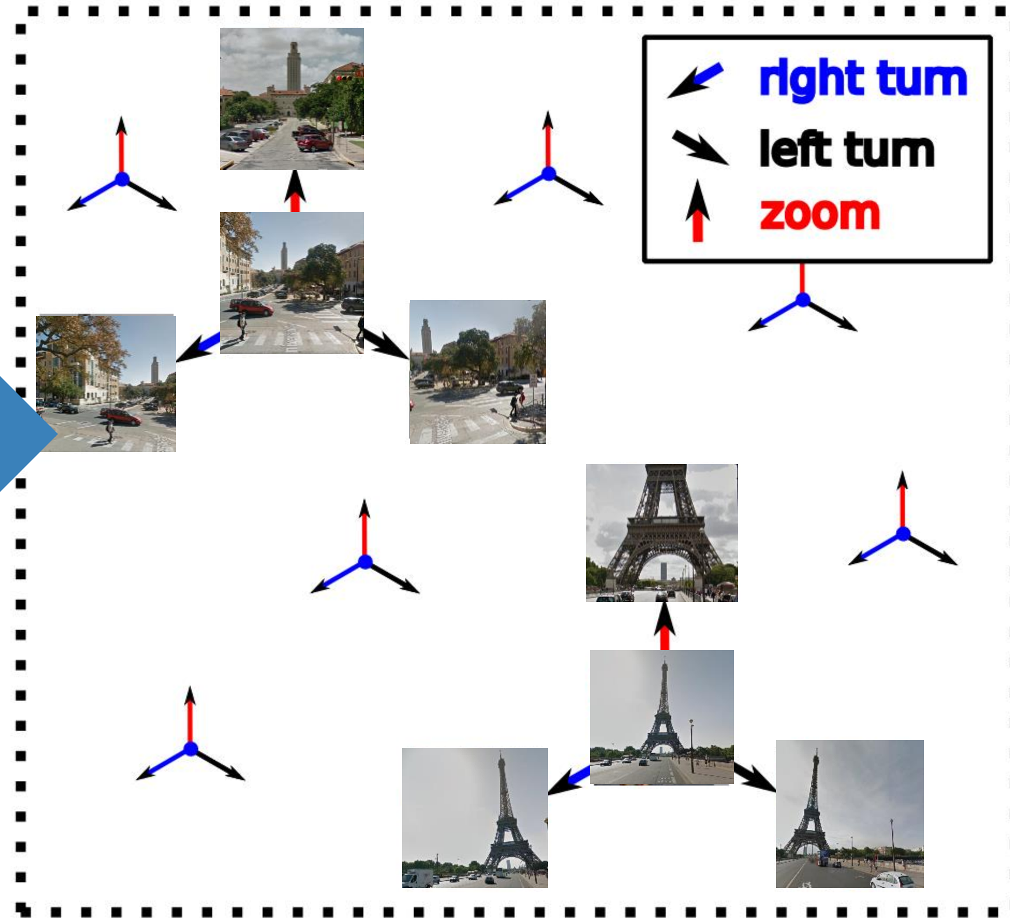
Unlabeled video + motor signals



Learn

Equivariant embedding

organized by ego-motions



Approach overview

Our approach: unsupervised feature learning using egocentric video + motor signals

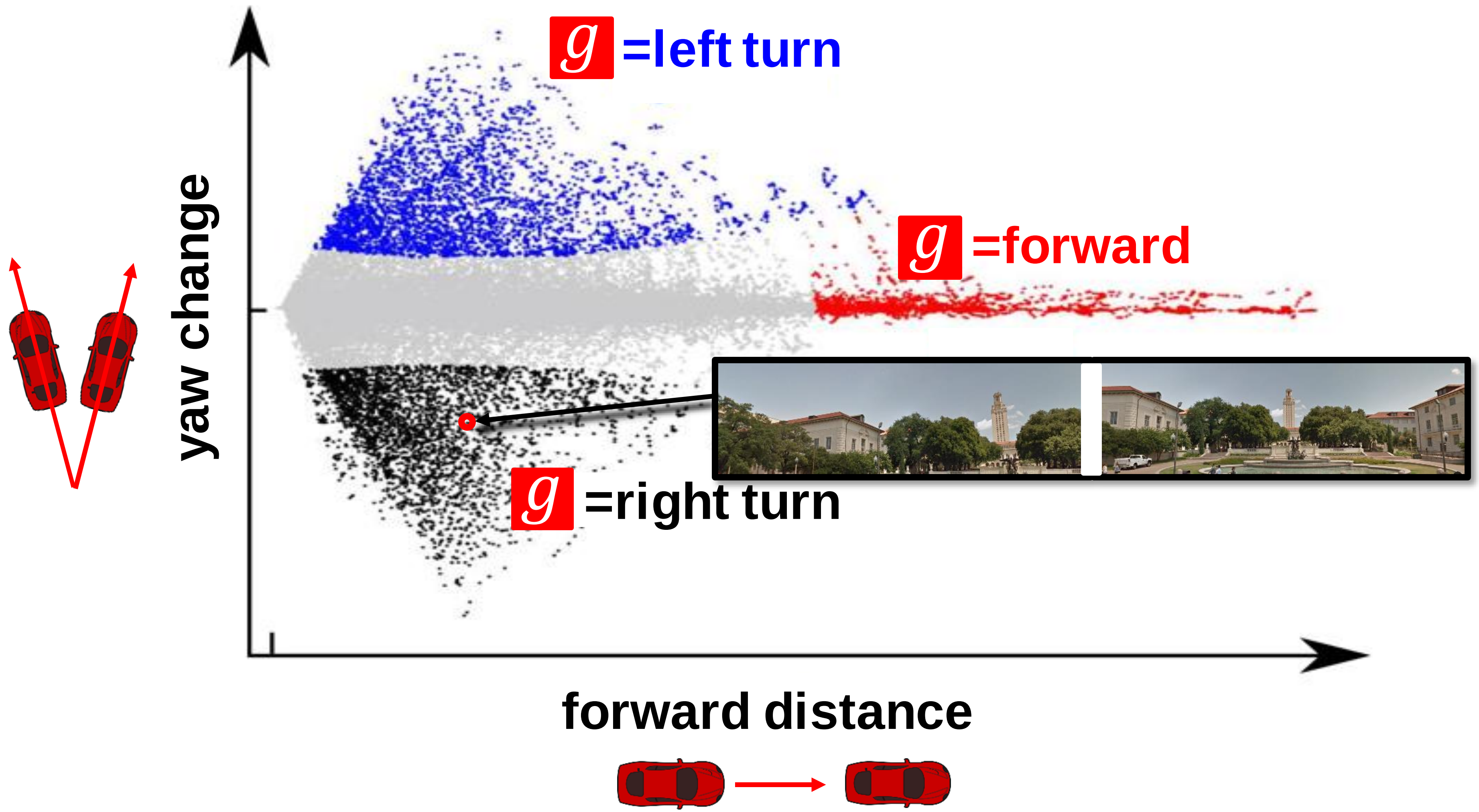
1. Extract training frame pairs from video

2. Learn ego-motion-equivariant image features

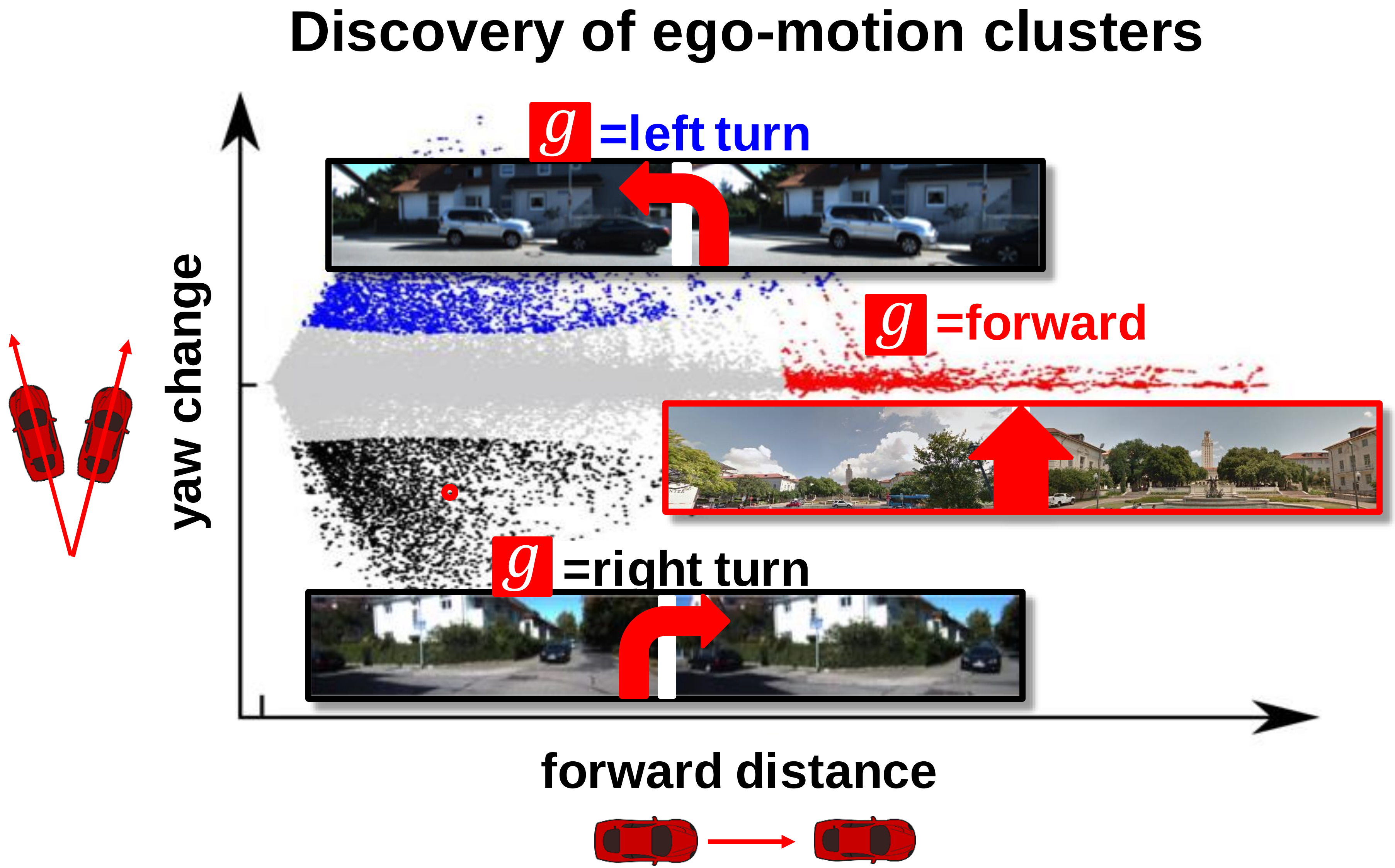
3. Train on target recognition task in parallel

Training frame pair mining

Discovery of ego-motion clusters



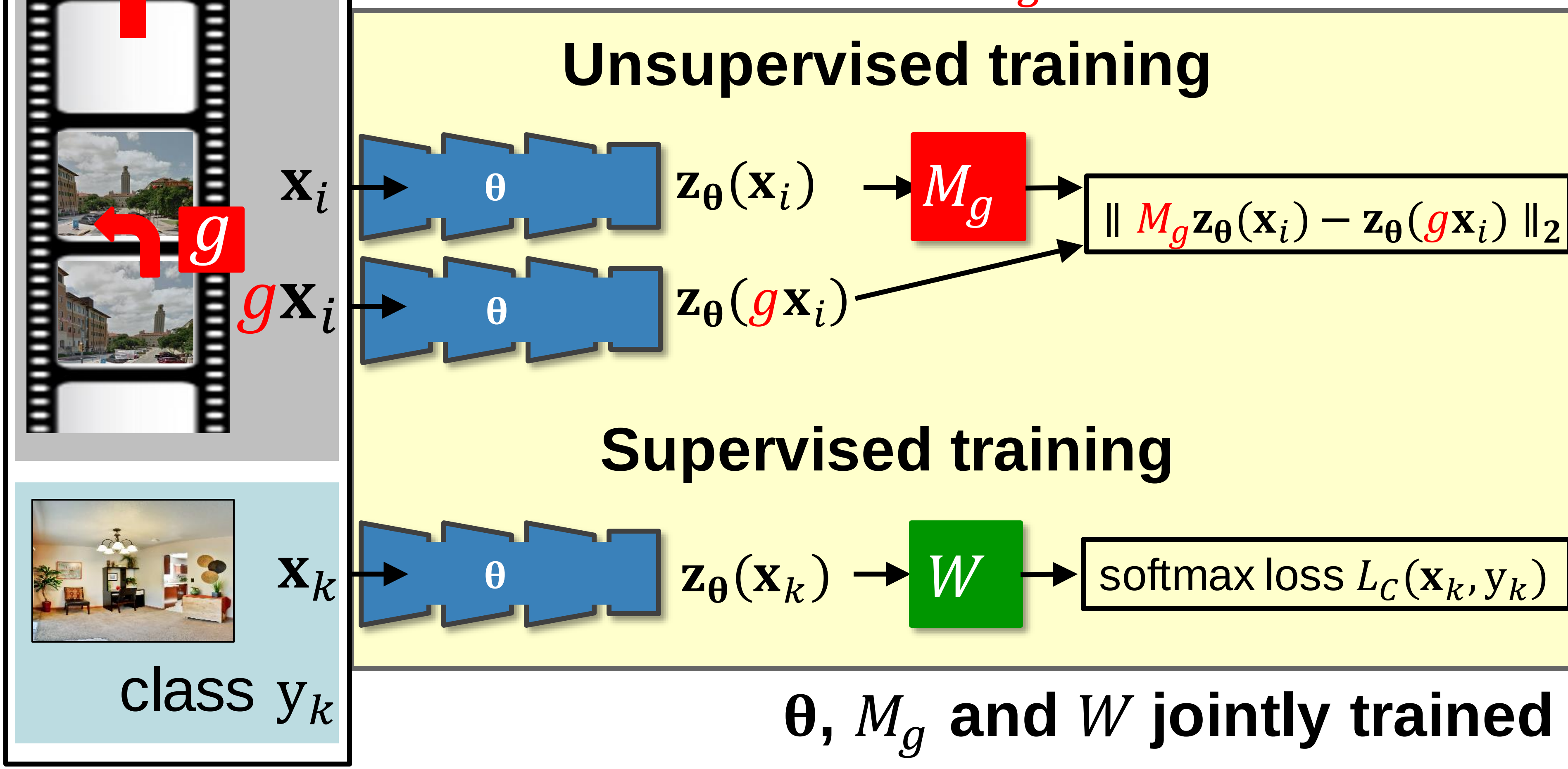
Training frame pair mining



Ego-motion equivariant feature learning

Desired: for all motions g and all images $\mathbf{x}$,

$$\mathbf{z}_{\boldsymbol{\theta}}(g\mathbf{x}) \approx M_g \mathbf{z}_{\boldsymbol{\theta}}(\mathbf{x})$$

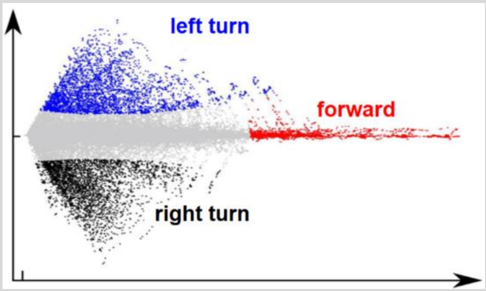


$\boldsymbol{\theta}$, M_g and W jointly trained

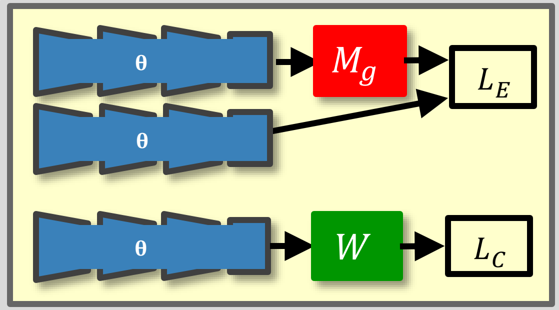
Method recap

APPROACH

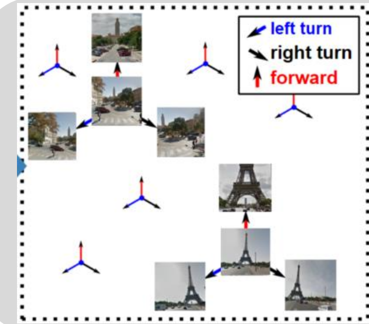
Ego-motion training pairs



Neural network training



Equivariant embedding



RESULTS

Scene and object recognition

Football field?
Pagoda?
Airport?
Cathedral?
Army base?

Next-best view selection

cup

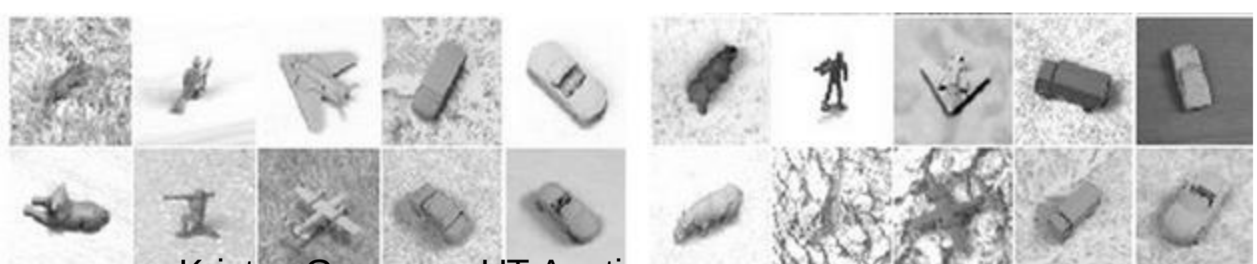
frying pan

Datasets

KITTI video
[Geiger et al. 2012]
Car platform
Egomotions: yaw and forward distance

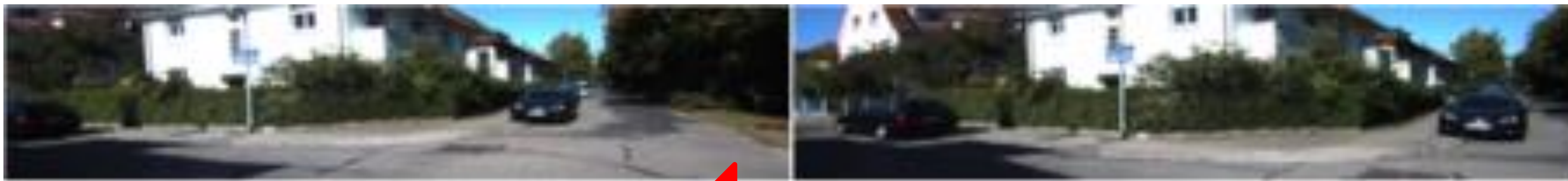
SUN images
[Xiao et al. 2010]
Large-scale scene classification task with 397 categories (static images)

NORB images
[LeCun et al. 2004]
Toy recognition
Egomotions: elevation and azimuth



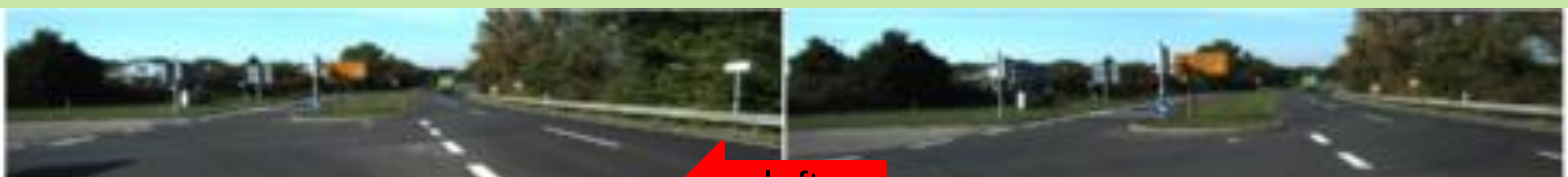
Results: Equivariance check

Visualizing how well equivariance is preserved



left

Query pair



left

Neighbor pair (our features)

zoom

Neighbor pair (pixel space)

Results: Equivariance check

How well is equivariance preserved?

Recognition loss only

Temporal coherence

Ours

Methods↓	atomic	composite
random	1.0000	1.0000
CLSNET	0.9239	0.9145
TEMPORAL [19]	0.7587	0.8119
DRLIM [7]	0.6404	0.7263
EQUIV	**0.6082**	**0.6982**
EQUIV+DRLIM	**0.5814**	**0.6492**

Normalized error:

$$\rho_g = E\left[\|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - M_g^{'}\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2 / \|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2\right]$$

Temporal coherence: Hadsell et al. CVPR 2006, Mohabi et al. ICML 2009

Results: Recognition

Learn from **unlabeled car video** (KITTI)

Geiger et al, IJRR '13

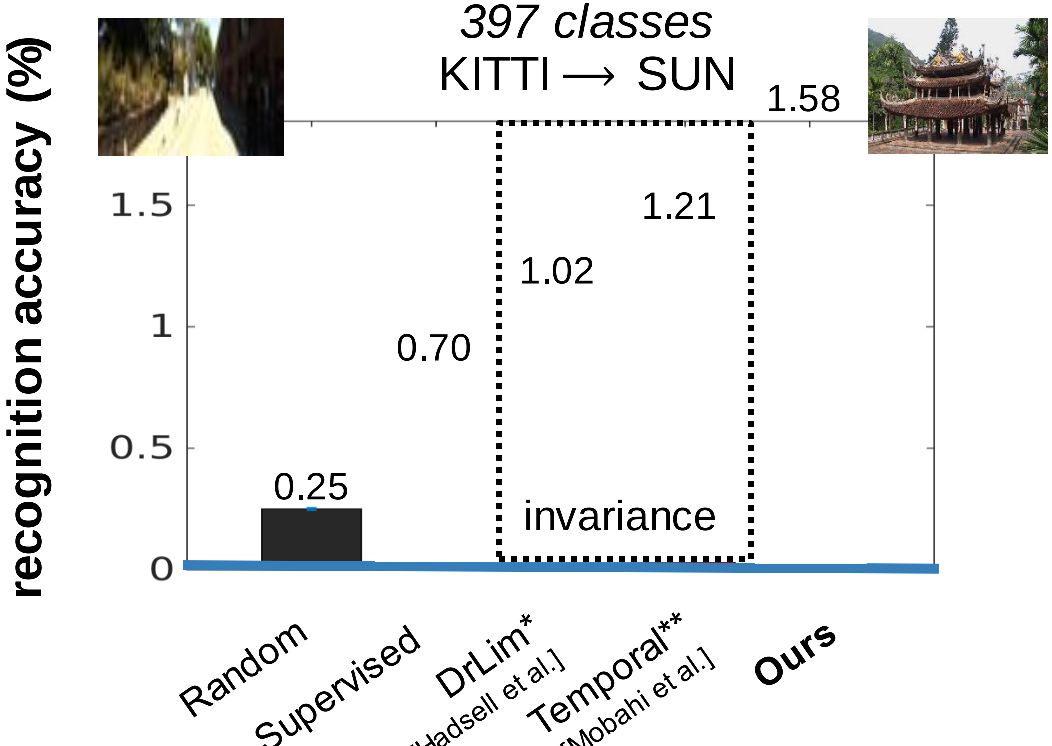
Exploit features for **static scene classification**
(SUN, 397 classes)

Apse
Window seat
Art school
Library
Auditorium
Bus interior
Cathedral
Freeway
Guardhouse

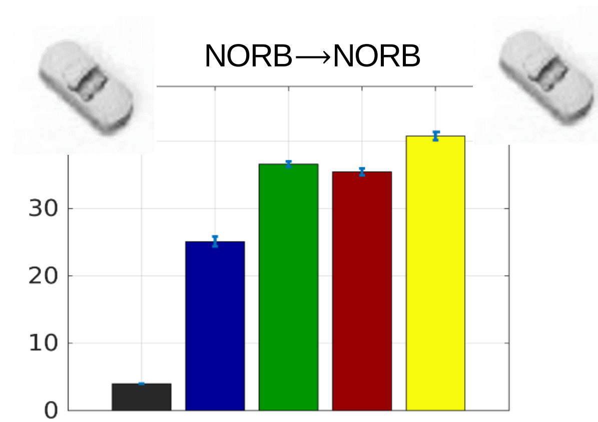
Xiao et al, CVPR '10

Results: Recognition

Do ego-motion equivariant features improve recognition?



6 labeled training examples per class

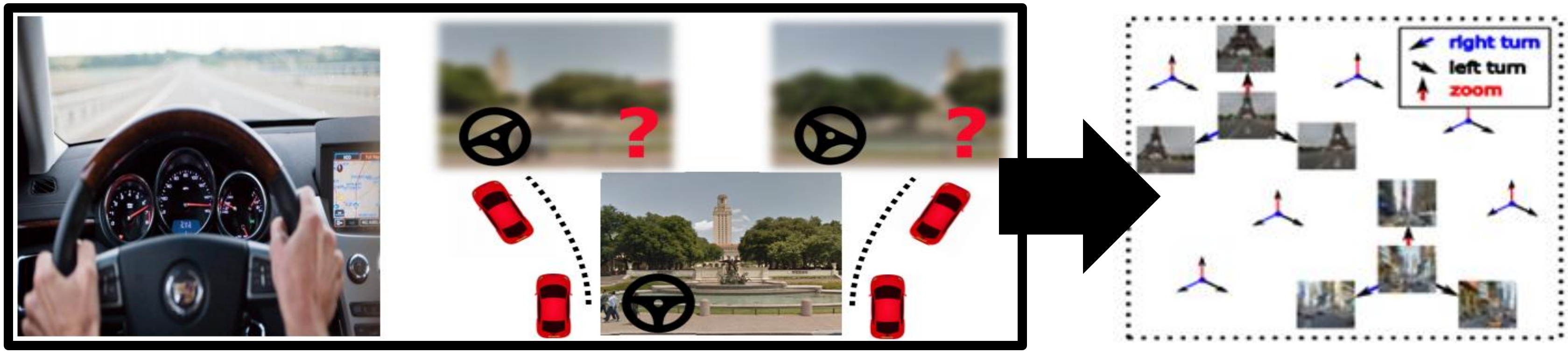


Up to 30% accuracy increase over state of the art!

*Hadsell et al., Dimensionality Reduction by Learning an Invaria

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML'09

Recap so far

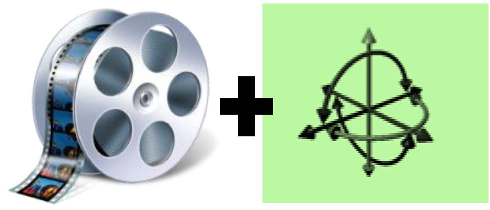


http://vision.cs.utexas.edu/projects/egoequiv/

- New *embodied* visual feature learning paradigm
- Ego-motion equivariance boosts performance across multiple challenging recognition tasks
- Future work: volition at training time too

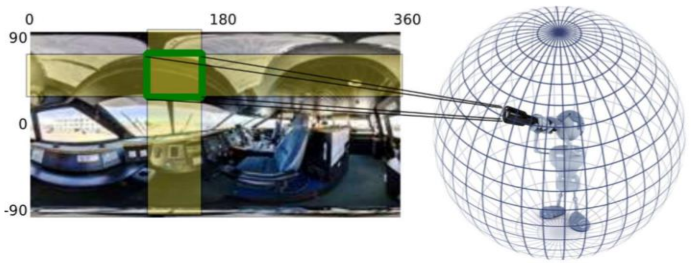
Talk overview

1. Learning representations tied to ego-motion

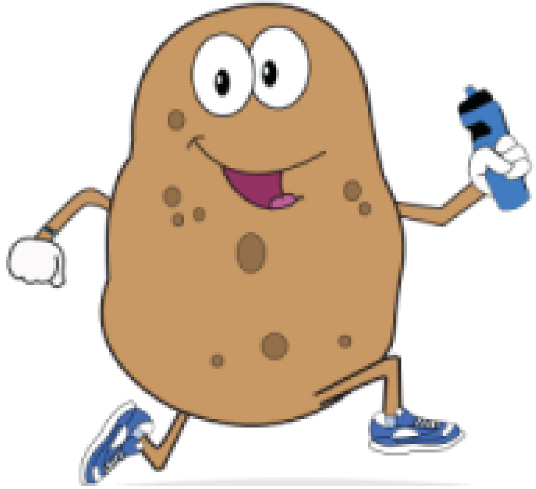


2. Learning representations from unlabeled video

3. Learning how to move and where to look



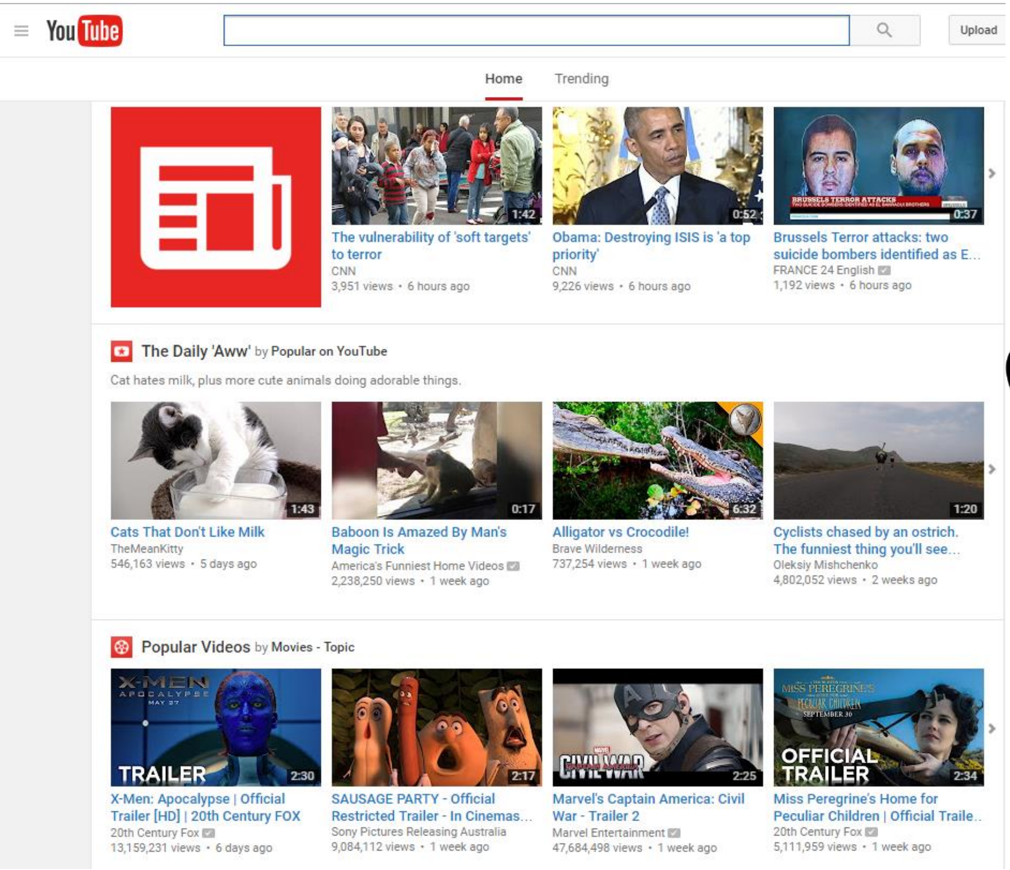
Learning from arbitrary unlabeled video?



Unlabeled video + ego-motion

Unlabeled video

Learning from arbitrary unlabeled video?

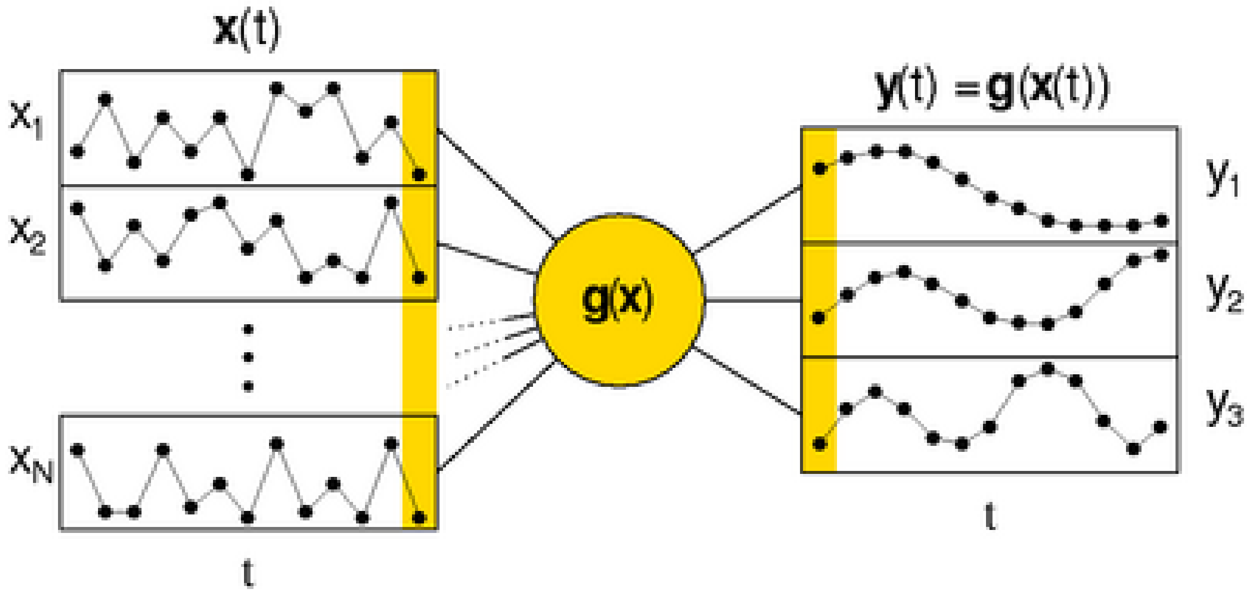


Unlabeled video

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

Find functions **g(x)** that map



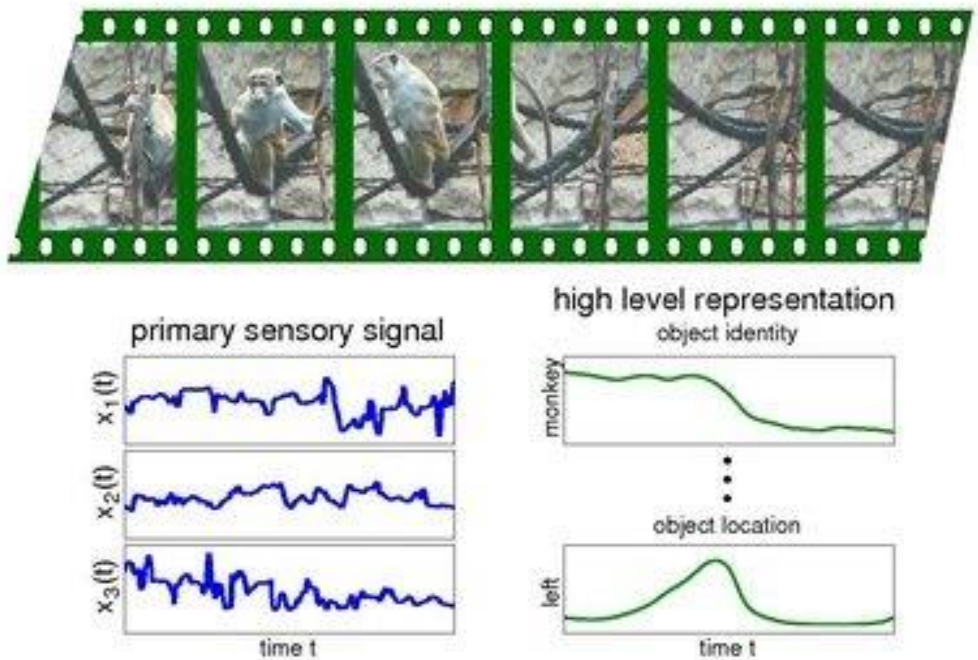
quickly varying input signal **x(**t**)** → slowly varying features **y(**t**)**

Figure: Laurenz Wiskott, http://www.scholarpedia.org/article/File:SlowFeatureAnalysis-OptimizationProblem.png

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]

Find functions **g(x)** that map



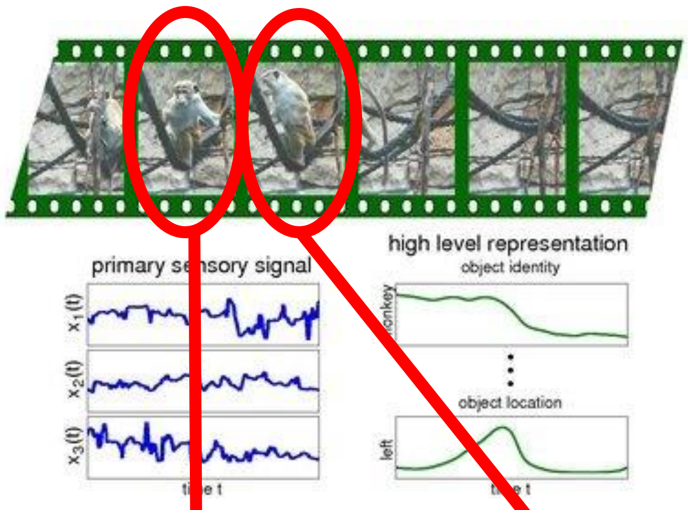
quickly varying input signal **x(**t**)** → slowly varying features **y(**t**)**

Figure: Laurenz Wiskott, http://www.scholarpedia.org/article/File:SlowFeatureAnalysis-OptimizationProblem.png

Kristen Grauman, UT Austin

Background: Slow feature analysis

[Wiskott & Sejnowski, 2002]



$$\mathbf{z}(\boldsymbol{a}) \approx \mathbf{z}(\boldsymbol{b})$$

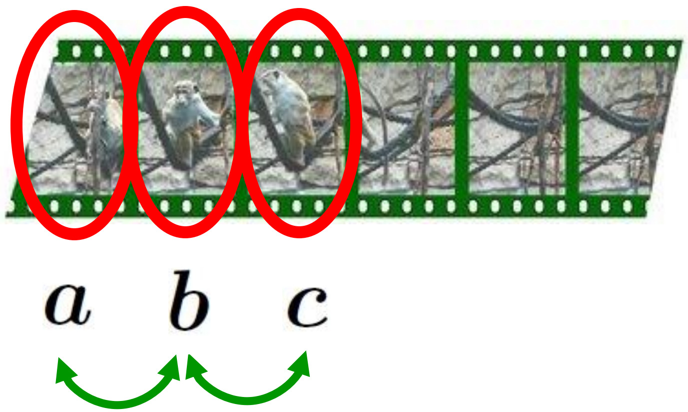
in learned embedding

- Existing work exploits "slowness" as temporal coherence in video → learn invariant representation

  [Hadsell et al. 2006; Mobahi et al. 2009; Bergstra & Bengio 2009; Goroshin et al. 2013; Wang & Gupta 2015,…]

- Fails to capture *how* visual content changes over time

Our idea: Steady feature analysis



- Higher order temporal coherence in video → learn equivariant representation

Second order slowness operates on frame triplets:

$$\mathbf{z}(\boldsymbol{b}) - \mathbf{z}(\boldsymbol{a}) \approx \mathbf{z}(\boldsymbol{c}) - \mathbf{z}(\boldsymbol{b})$$

in learned embedding

[Jayaraman & Grauman, CVPR 2016]

Approach: Steady feature analysis

Learn classifier *W* and representation *θ* jointly,

$$(\boldsymbol{\theta}^*, W^*) = \underset{\boldsymbol{\theta}, W}{\arg\min}\, L_s(\boldsymbol{\theta}, W, \mathcal{S}) + \lambda L_u(\boldsymbol{\theta}, \mathcal{U})$$

with unsupervised regularization loss:

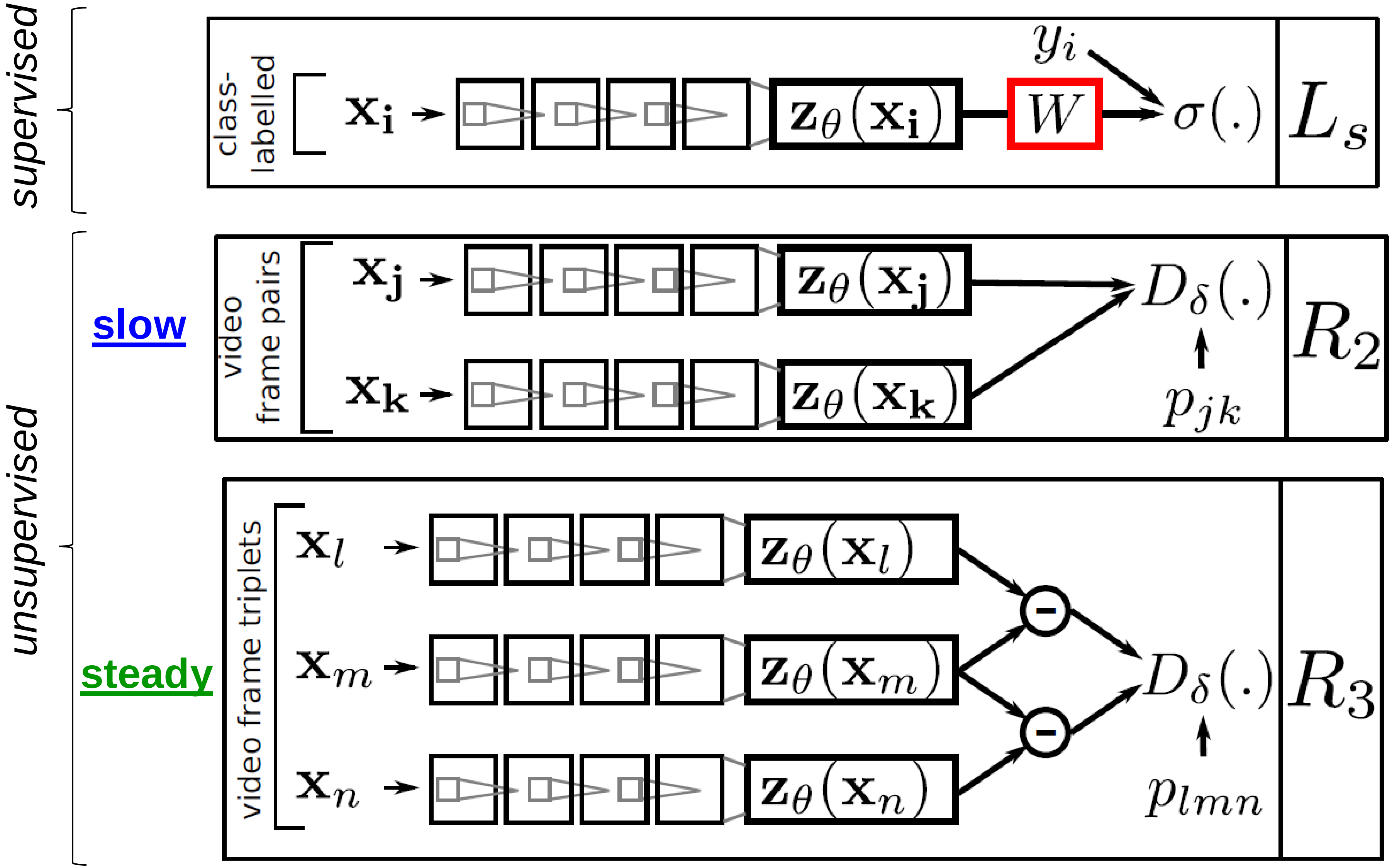
$$L_u(\boldsymbol{\theta}, \mathcal{U}) = \underbrace{R_2(\boldsymbol{\theta}, \mathcal{U})}_{\text{slow}} + \lambda' \underbrace{R_3(\boldsymbol{\theta}, \mathcal{U})}_{\text{steady}}$$

slow $\sum_{(j,k)\in\mathcal{U}_2} D_\delta(\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_k), p_{jk})$

steady $\sum_{(l,m,n)\in\mathcal{U}_3} D_\delta(\mathbf{z}_{\boldsymbol{\theta}l} - \mathbf{z}_{\boldsymbol{\theta}m},\ \mathbf{z}_{\boldsymbol{\theta}m} - \mathbf{z}_{\boldsymbol{\theta}n},\ p_{lmn})$

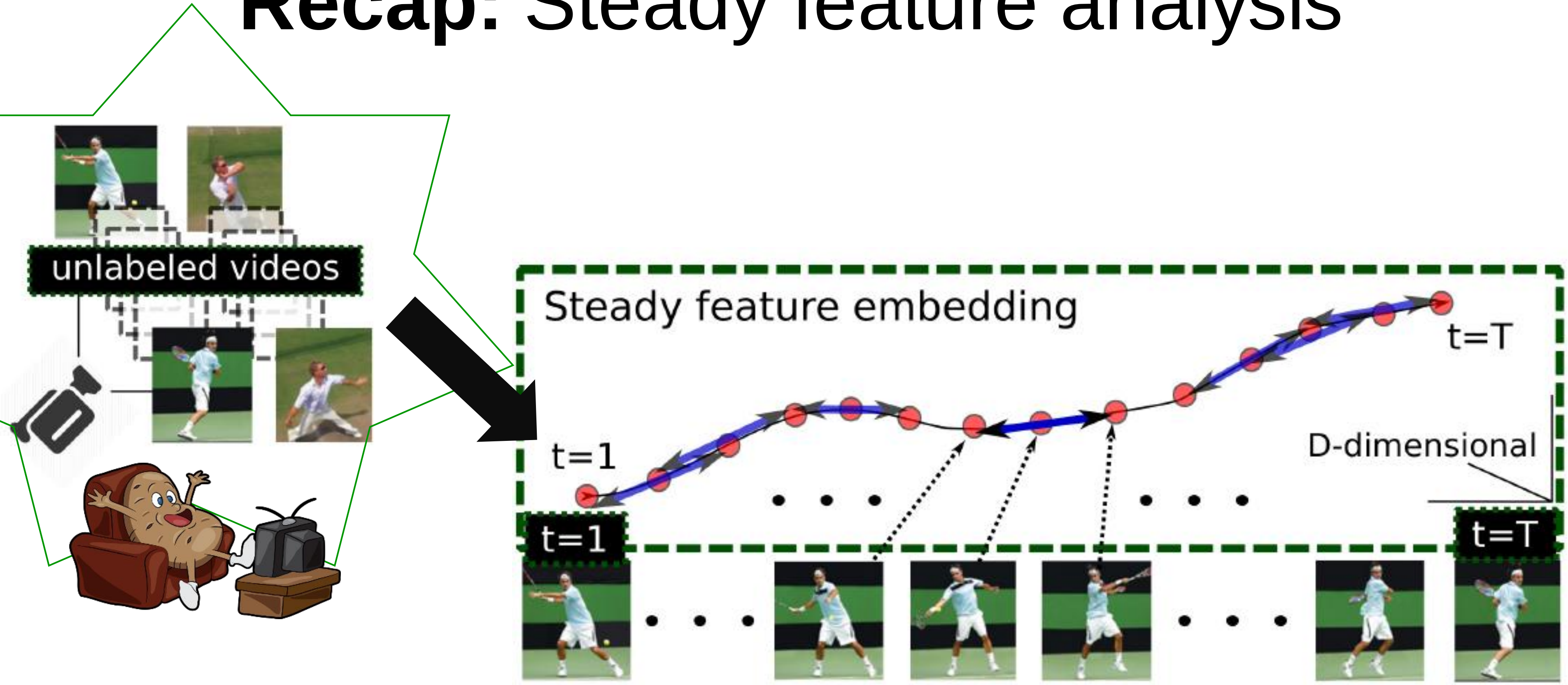
Contrastive loss that also exploits "negative" tuples

Approach: Steady feature analysis



[Jayaraman & Grauman, CVPR 2016]

Recap: Steady feature analysis



Equivariance ≈ “steadily” varying frame features!

$$d^2\mathbf{z}_{\theta}(\mathbf{x}t)/dt^2 \approx \mathbf{0}$$

[Jayaraman & Grauman, CVPR 2016]

Datasets

Unlabeled video

Human Motion Database (HMDB)

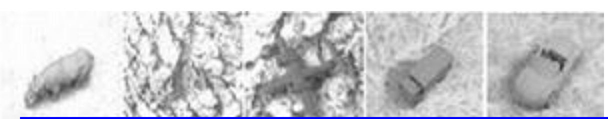
KITTI Video

NORB

Target task (few labels)

PASCAL 10 Actions

SUN 397 Scenes



NORB 25 Objects

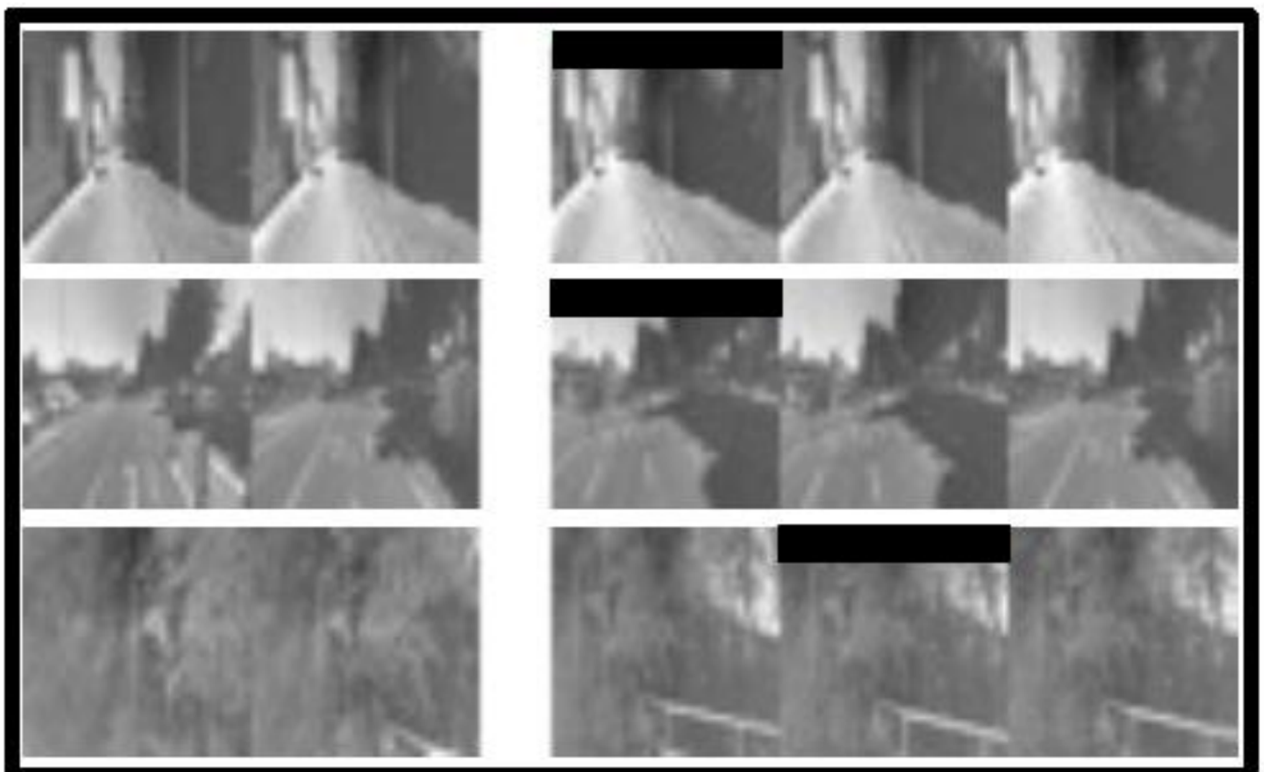
32 x 32 images or 96 x 96 images

Results: Sequence completion

Given sequential pair, infer next frame (embedding)

$$\tilde{\mathbf{z}}_{\boldsymbol{\theta}}(\boldsymbol{x}_3) = 2\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_2) - \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_1)$$

$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ Our top 3 estimates for $\boldsymbol{x}_3$



KITTI datase

Results: Sequence completion

Given sequential pair, infer next frame (embedding)

	Datasets→	NORB	KITTI	HMDB
slow	SFA-1 [30] *	0.95	31.04	2.70
slow	SFA-2 [14] **	0.91	8.39	2.27
slow & steady	SSFA (ours)	**0.53**	**7.79**	**1.78**

Percentile rank of correct completion (lower is better)

*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR'06

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML'09

Results: Recognition

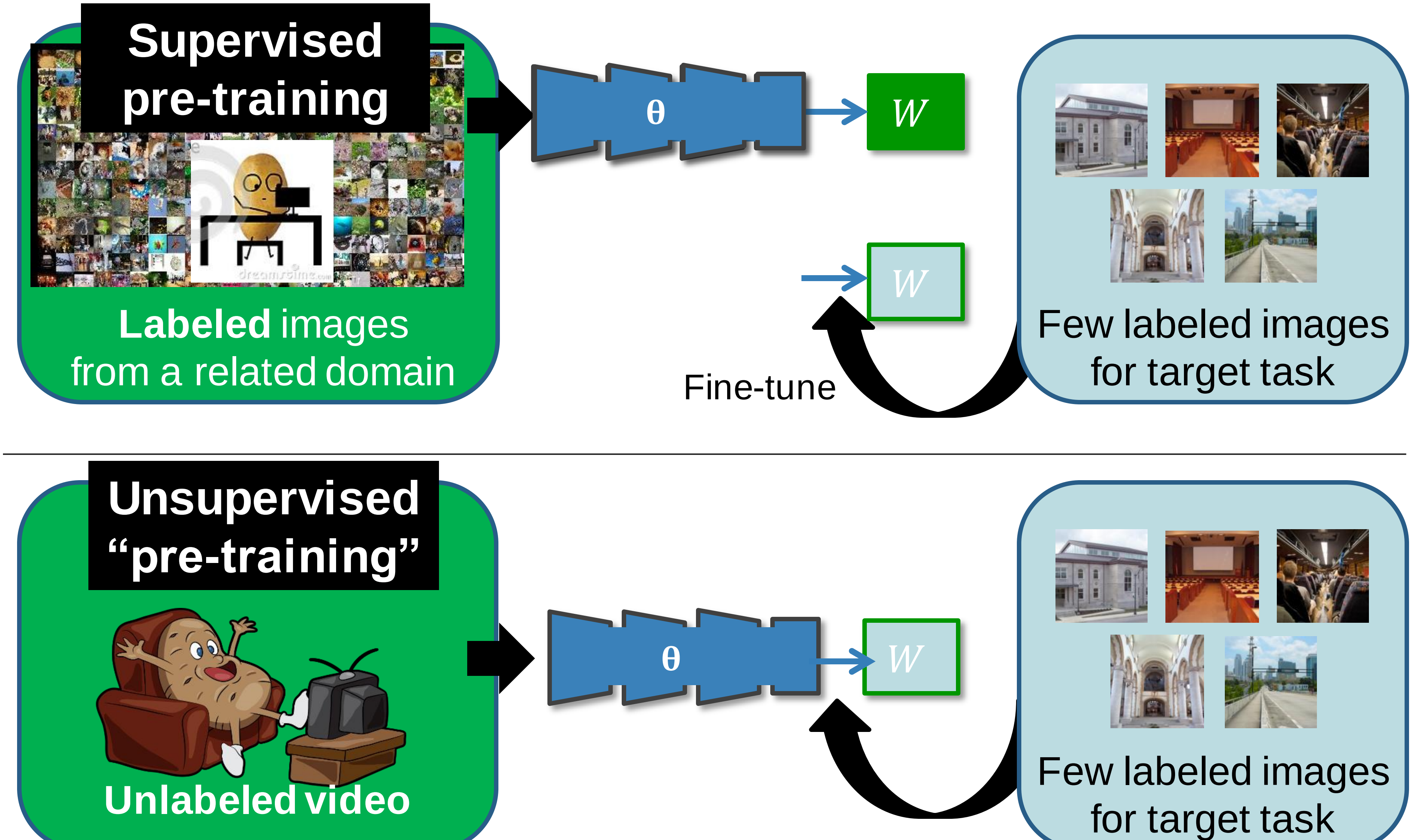
Task type→	Objects	Scenes		Actions
Datasets→	NORB→NORB	KITTI→SUN		HMDB→PASCAL-10
Methods↓	[25 cls]	[397 cls]	[397 cls, top-10]	[10 cls]
random	4.00	0.25	2.52	10.00
UNREG	24.64±0.85	0.70±0.12	6.10±0.67	15.34±0.28
SFA-1 [30]*	37.57±0.85	1.21±0.14	8.24±0.25	19.26±0.45
SFA-2 [14]**	39.23±0.94	1.02±0.12	6.78±0.32	19.04±0.24
SSFA (ours)	**42.83±0.33**	**1.65±0.04**	**9.19±0.10**	**20.95±0.13**

Multi-class recognition accuracy

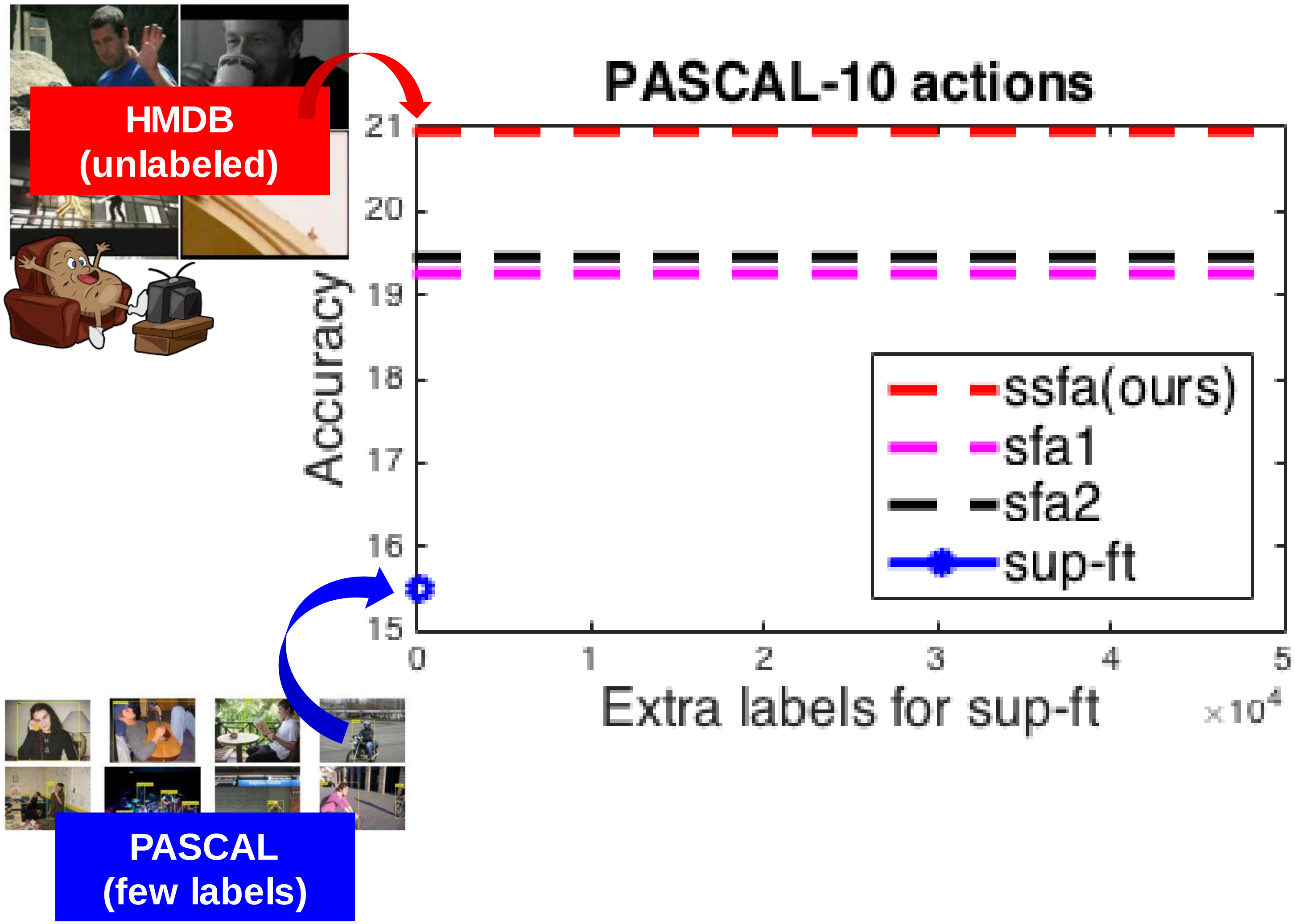
*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR'06

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML'09

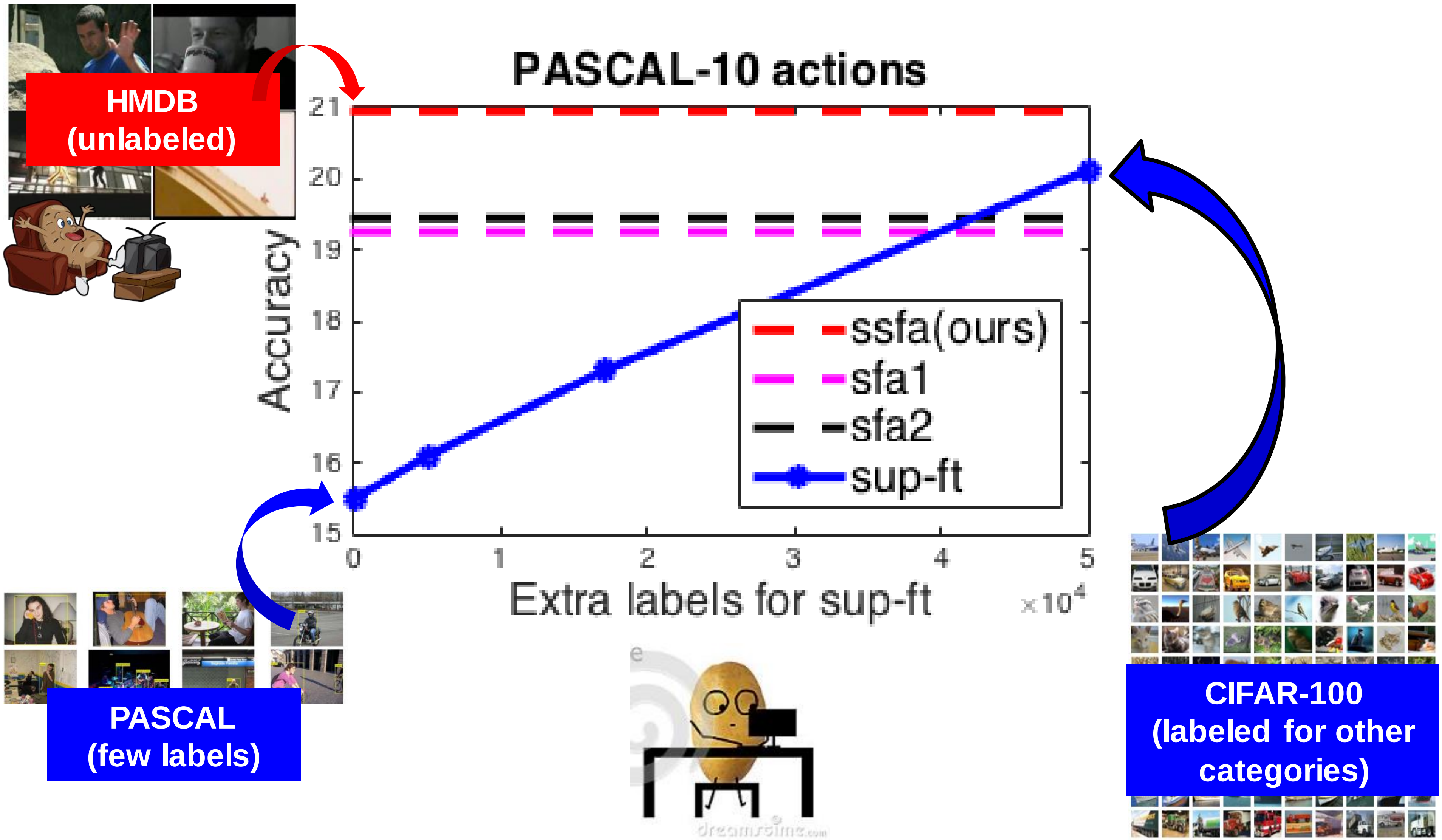
Pre-training a representation



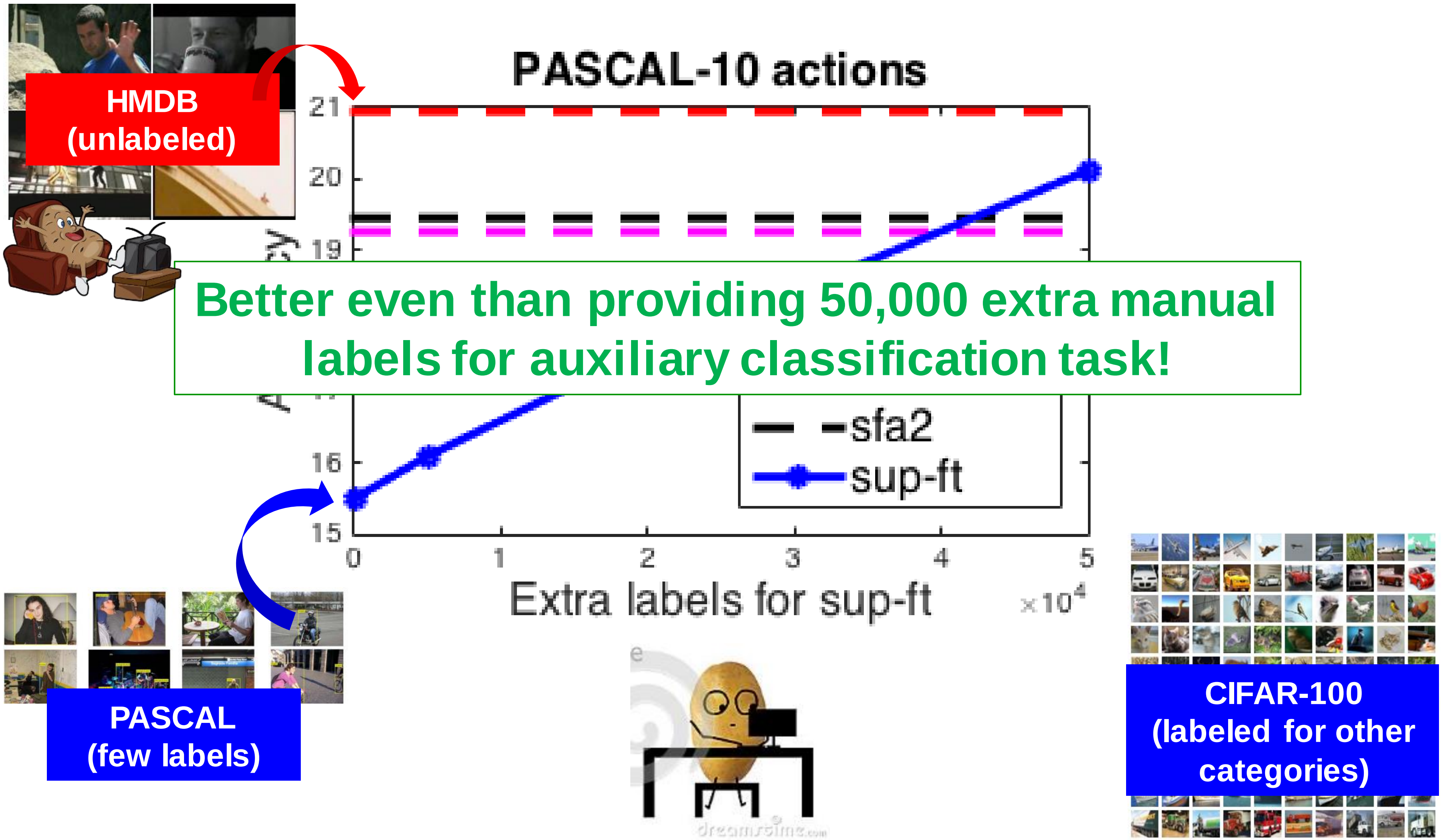
Results: Can we learn *more* from unlabeled video than “related” labeled images?



Results: Can we learn *more* from unlabeled video than “related” labeled images?



Results: Can we learn *more* from unlabeled video than “related” labeled images?

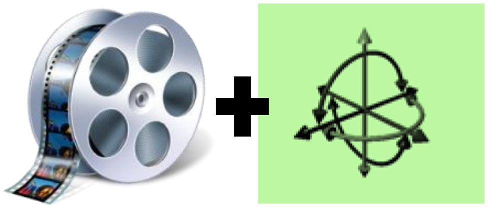


Better even than providing 50,000 extra manual labels for auxiliary classification task!

Kristen Grauman, UT Austin

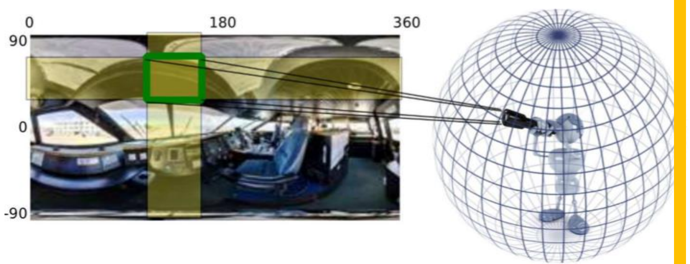
Talk overview

1. Learning representations tied to ego-motion



2. Learning representations from unlabeled video

3. Learning how to move and where to look



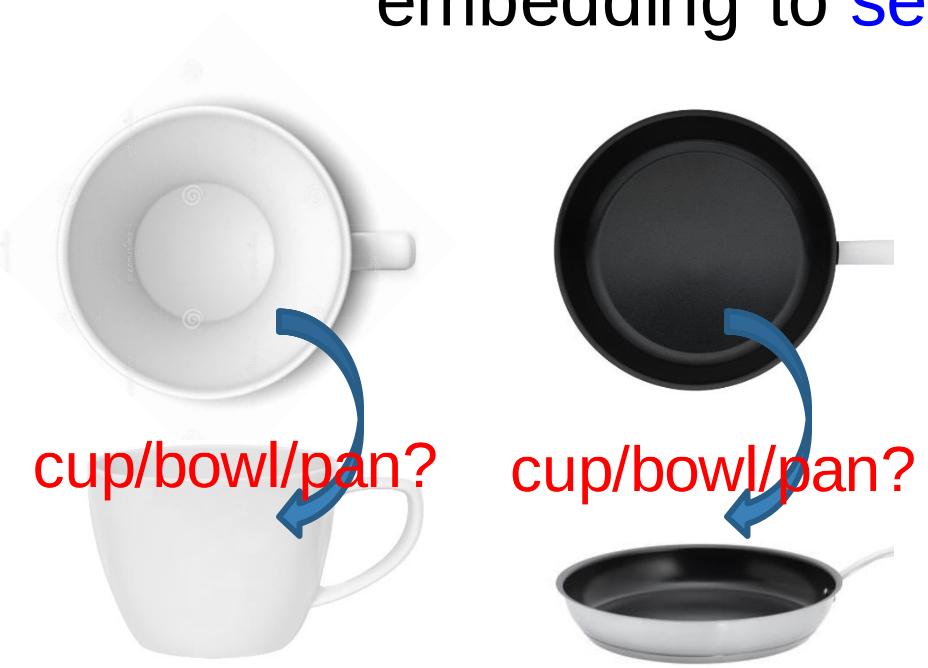
Learning how to move for recognition

Time to revisit active recognition in challenging settings!

[Bajcsy 1985, Schiele & Crowley 1998, Dickinson et al. 1997, Tsotsos et al. 2001, Soatto 2009,…]

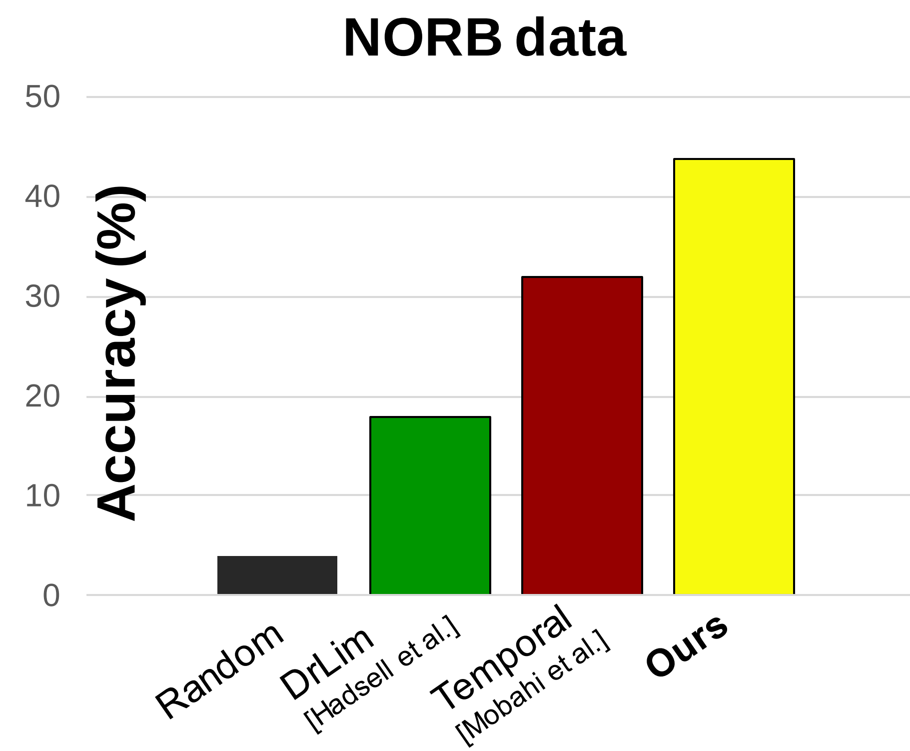
Learning how to move for recognition

Leverage proposed ego-motion equivariant embedding to select next best view



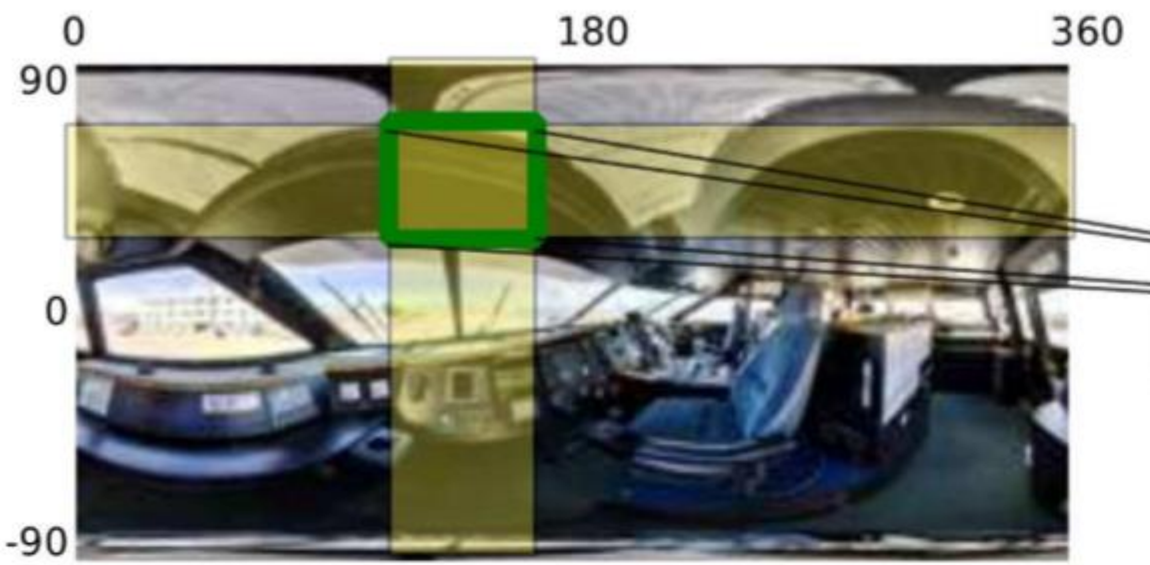
cup

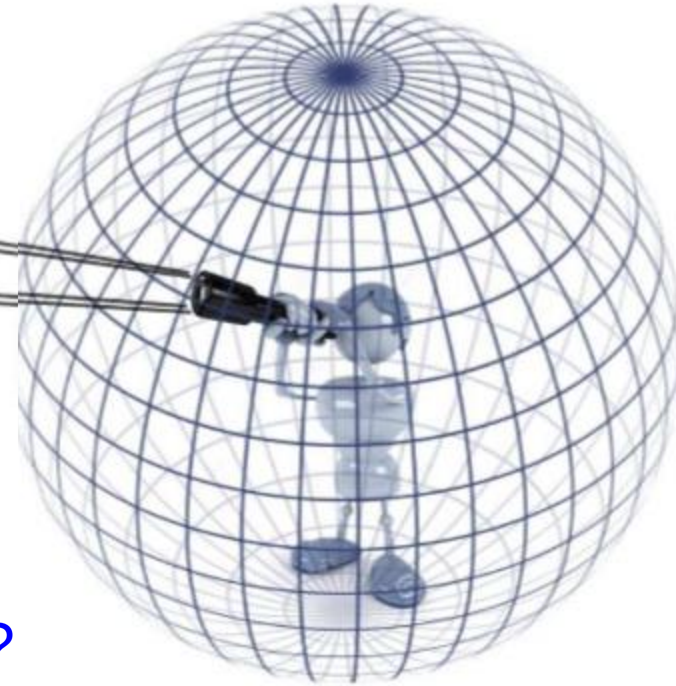
frying pan



[Jayarman & Grauman, ICCV 2015]

Learning how to move for recognition





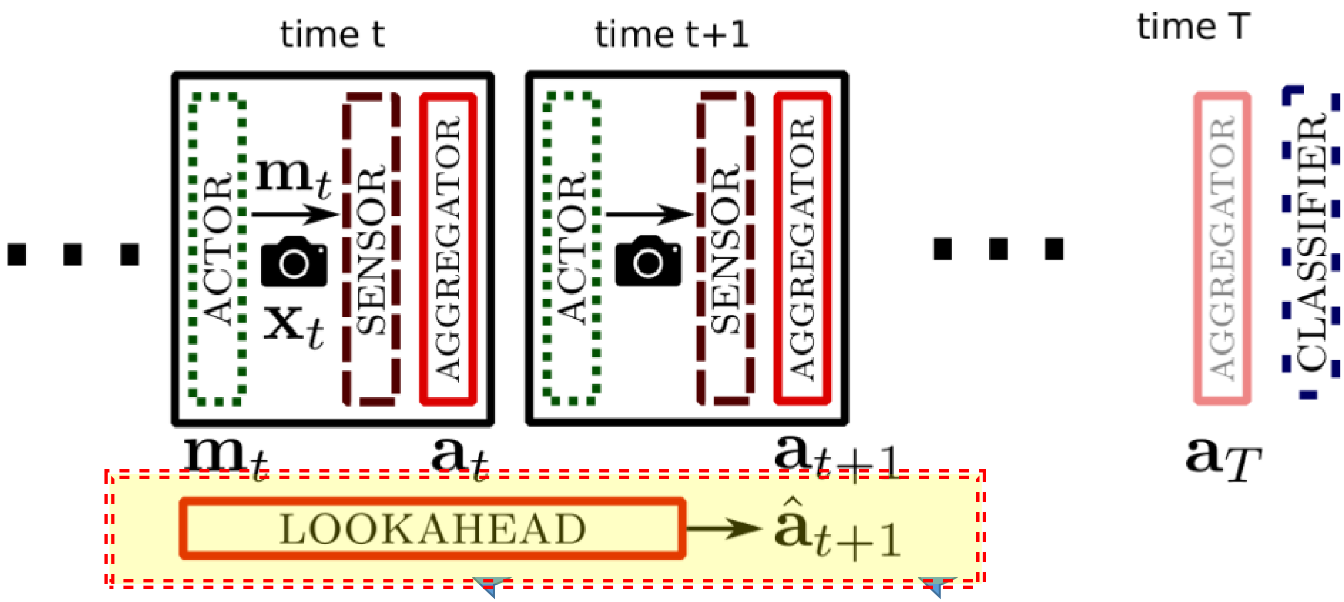
Best sequence of glimpses in 3D scene?

Requires:
- Action selection
- Per-view processing
- Evidence aggregation
- Look-ahead prediction
- Final class belief prediction

Learn all end-to-end

Jayaraman and Grauman, UT TR AI15-06

Active visual recognition



Requires several separate functionalities:

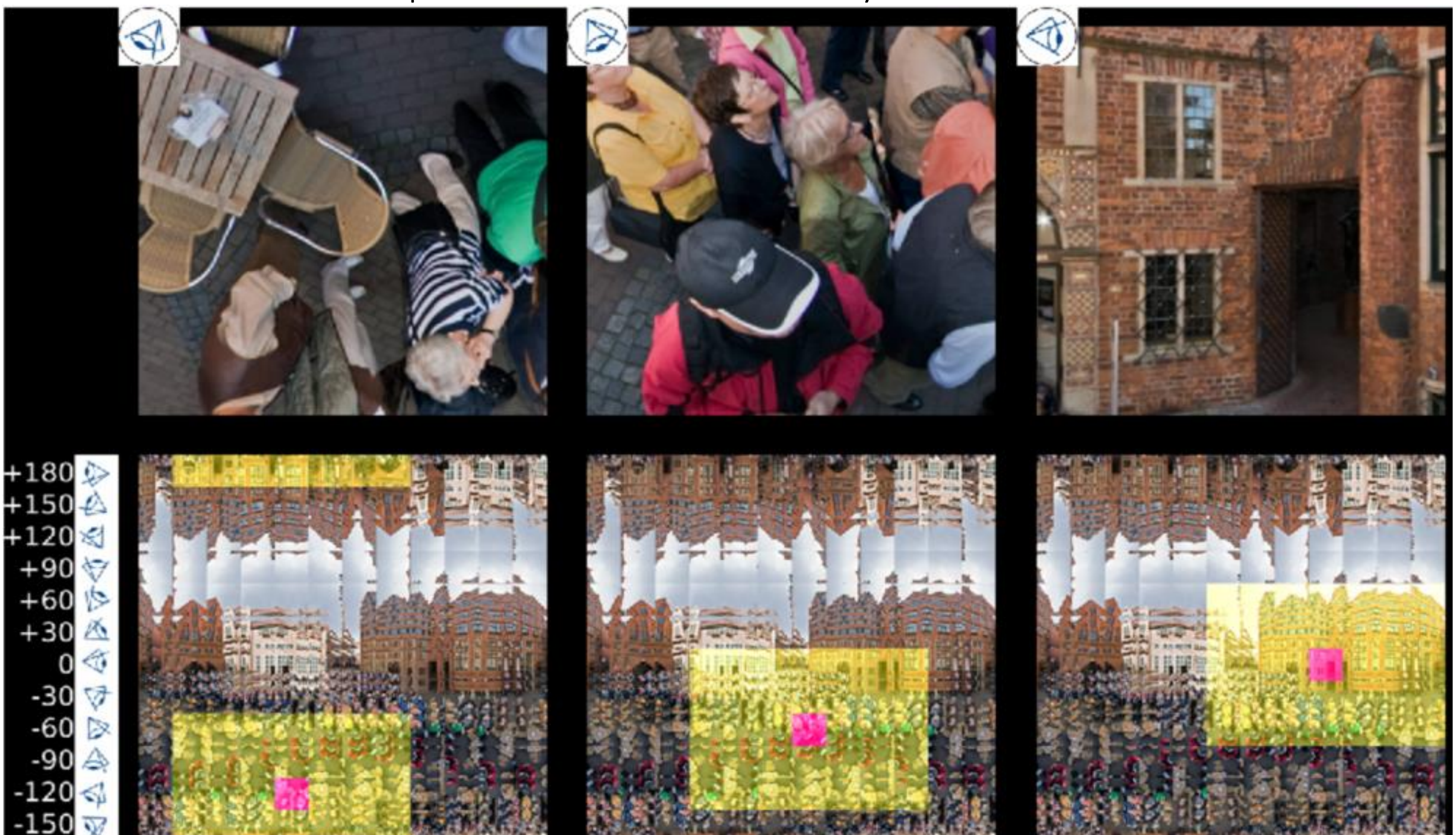
- Action selection
- Per-view processing
- Across-view evidence aggregation
- Next-view prediction
- Final class belief prediction

Learn all end-to-end

Kristen Grauman, UT Austin

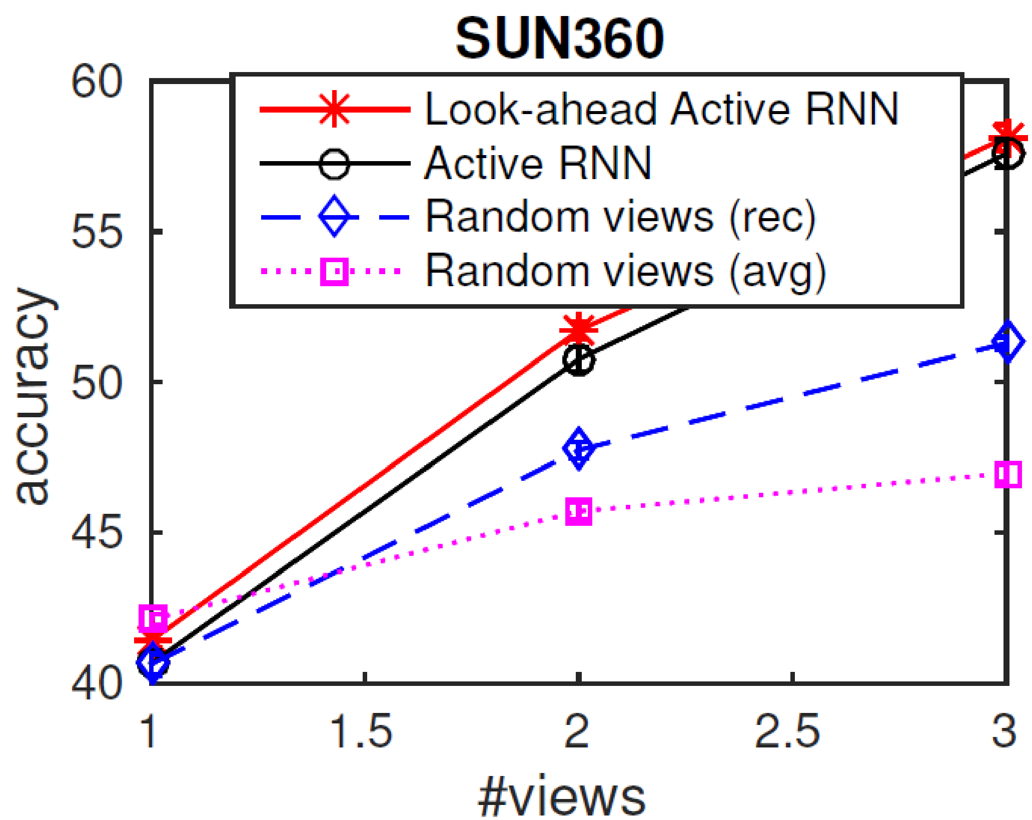
Active recognition: example results

P("Plaza courtyard"): Top 3 guesses:	(6.28) Restaurant Train interior Shop	(11.95) Theater Restaurant Plaza courtyard	(68.38) Plaza courtyard Street Theater



Jayaraman and Grauman, UT TR AI15-06

Active recognition: Results



Active selection + look-ahead → better scene categorization from sequence of glimpses in 360 panorama

Summary

- Visual learning requires
  - context of action and motion in the world
  - with continuous self-acquired feedback

- New ideas:
  - "Embodied" feature learning using both visual and motor signals
  - Feature learning from unlabeled video via higher order temporal coherence
  - Steps towards active view selection in 360 scenes

References

- Learning Image Representations Tied to Ego-Motion. D. Jayaraman and K. Grauman. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec 2015. (Oral)

- Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video. D. Jayaraman and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 2016. (Spotlight)

- Look Ahead Before You Leap: End-to-End Active Recognition by Forecasting the Effect of Motion. D. Jayaraman and K. Grauman. UT Tech Report A115-06, Dec 2015.