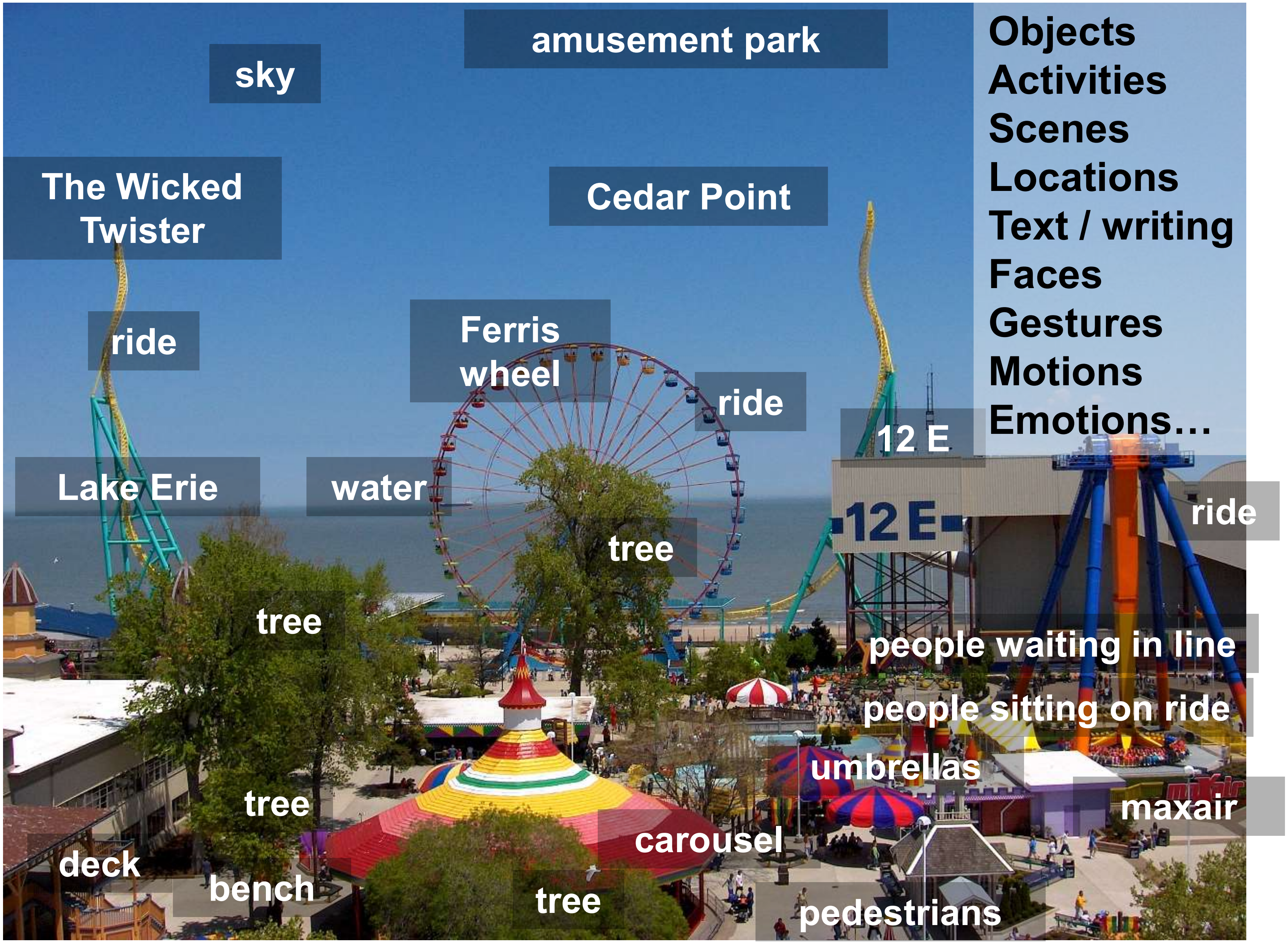
Weinberg Symposium on the Shared Frontiers of
Artificial Intelligence and Cognitive Science
University of Michigan, April 2018

Embodied Visual Learning and Recognition

Kristen Grauman
Department of Computer Science
University of Texas at Austin

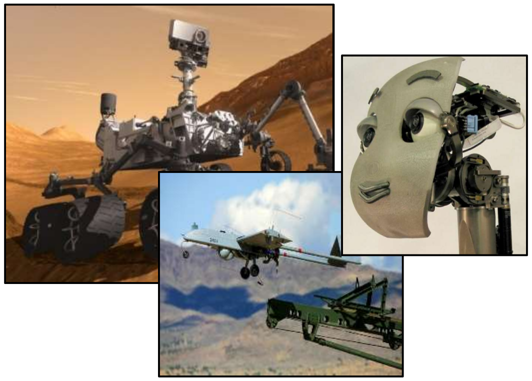
THE UNIVERSITY OF TEXAS AT AUSTIN

Visual recognition

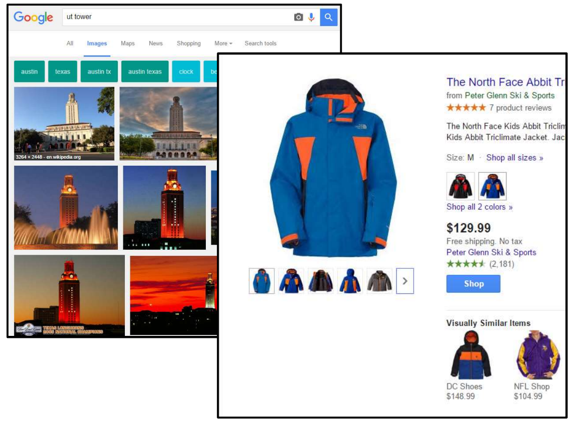


Objects
Activities
Scenes
Locations
Text / writing
Faces
Gestures
Motions
Emotions…

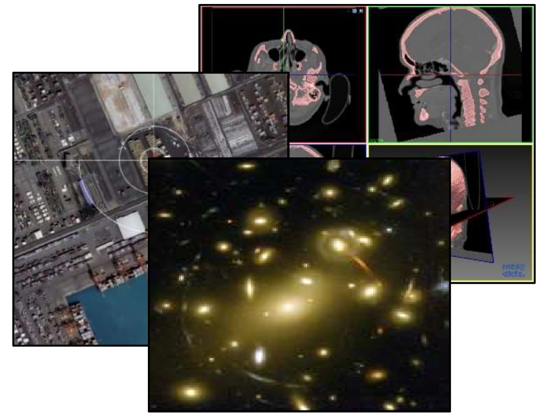
Visual recognition: applications



AI and autonomous robotics



Organizing visual content



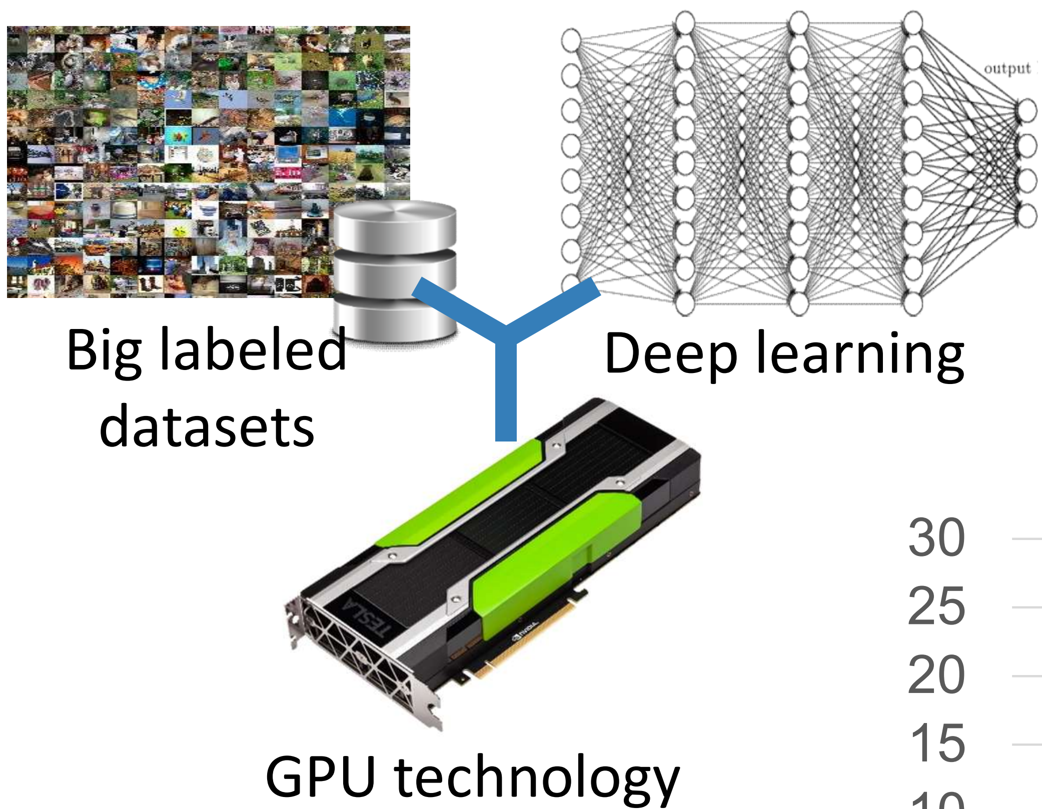
Science and medicine

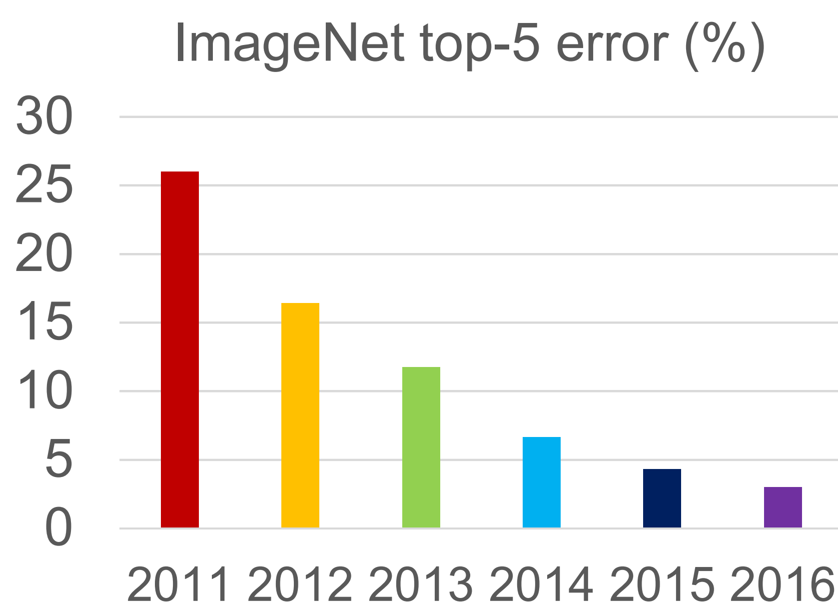
Gaming, HCI, Augmented Reality

Surveillance and security

Personal photo/video collections

Visual recognition: significant recent progress

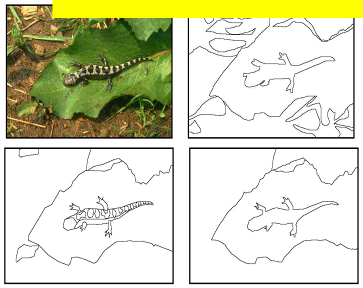




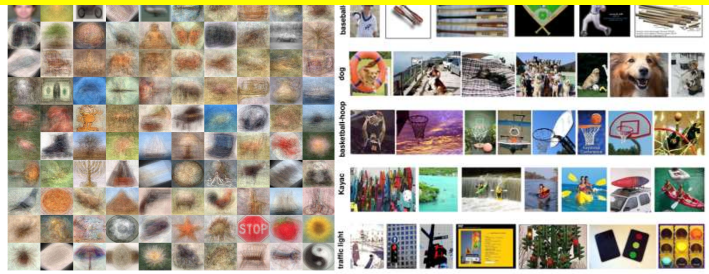
How do our systems learn about the visual world today?

Recognition benchmarks

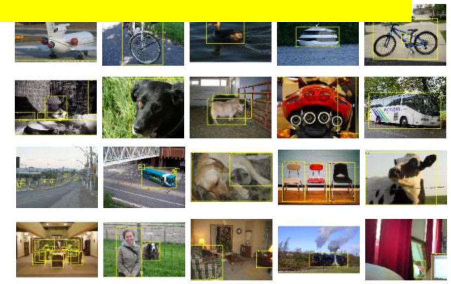
A “disembodied” well-curated moment in time



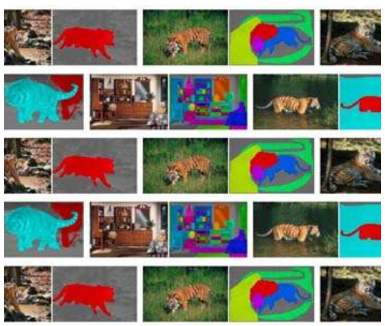
BSD (2001)



Caltech 101 (2004), Caltech 256 (2006)



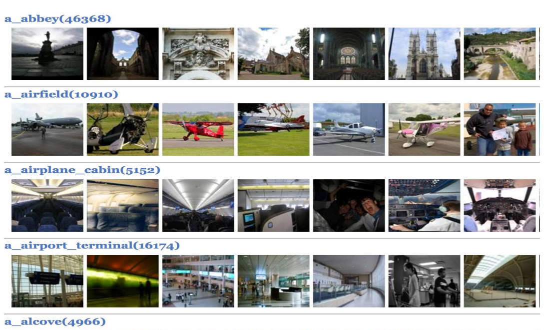
PASCAL (2007-12)



LabelMe (2007)

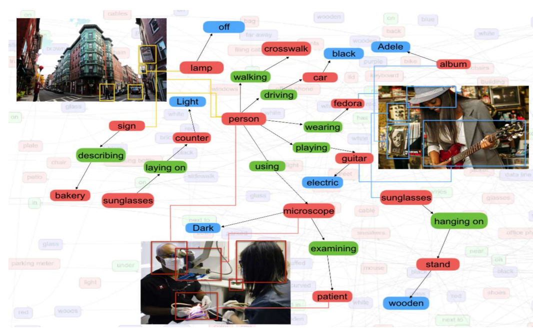
ImageNet (2009)

SUN (2010)



Places (2014)

MS COCO (2014)



Visual Genome (2016)

Egocentric perceptual experience

A tangle of relevant and irrelevant multi-sensory information

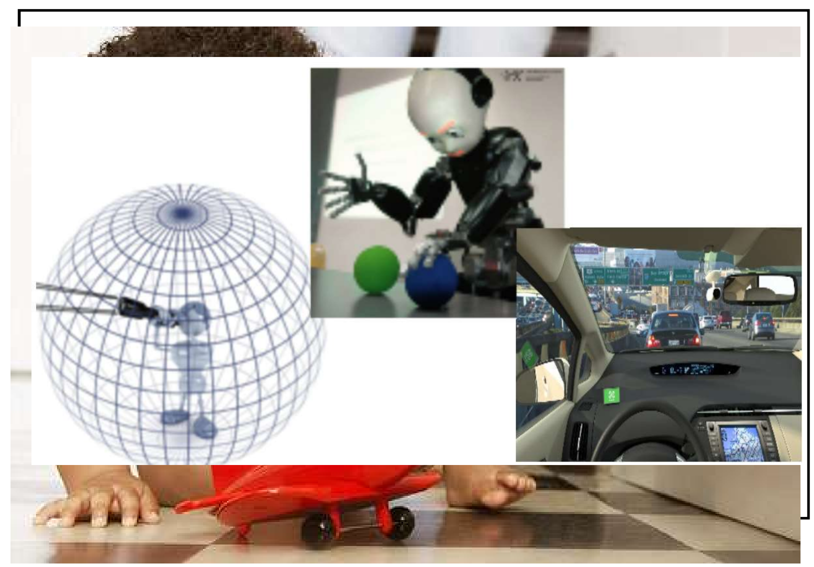
Big picture goal: Embodied visual learning

Status quo:

Learn from "disembodied" bag of labeled snapshots.

On the horizon:

Visual learning in the context of acting and moving in the world.

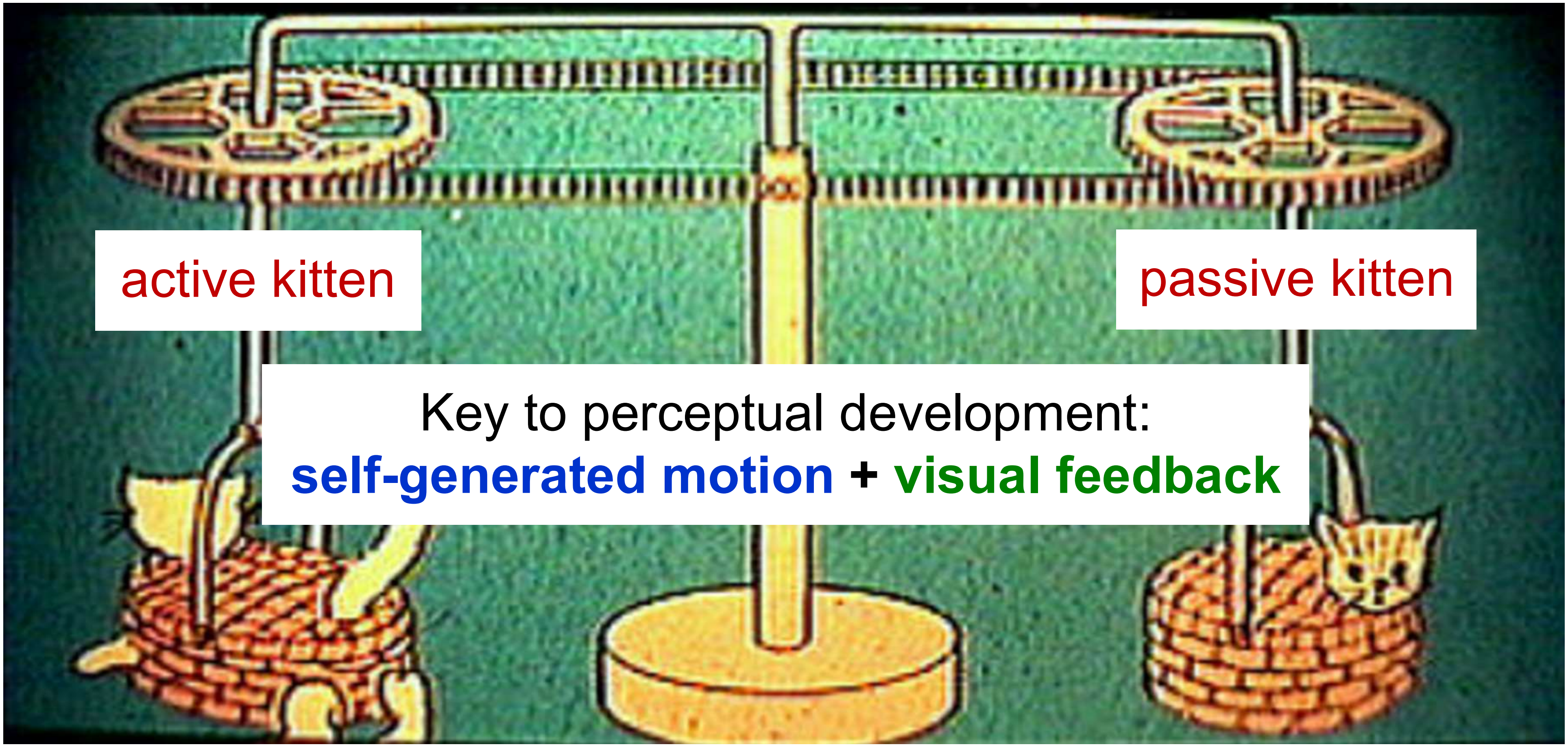


This talk

Towards embodied visual learning

1. Learning from unlabeled video and multiple sensory modalities
2. Learning policies for how to move for recognition and exploration

The kitten carousel experiment
[Held & Hein, 1963]



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

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

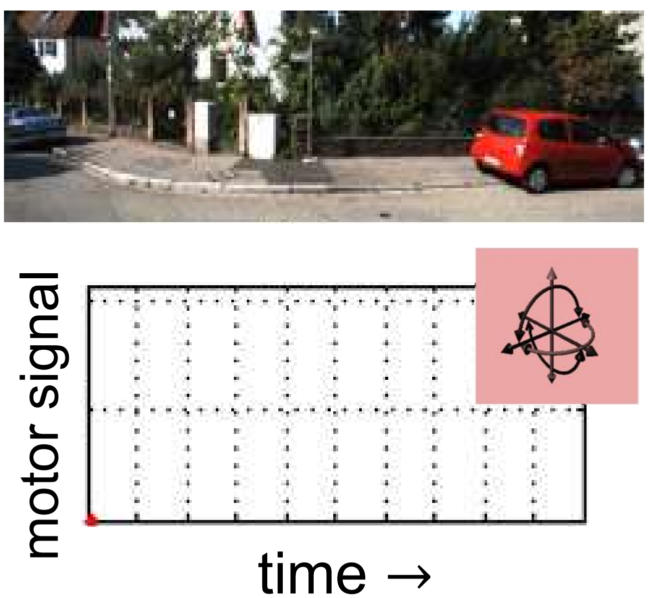
Ego-motion ↔ vision: view prediction

After moving:

Approach idea: Ego-motion equivariance

Training data

Unlabeled video +
motor signals



Learn

Equivariant embedding

organized by ego-motions

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

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

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

Results: Recognition

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

Geiger et al, IJRR '13

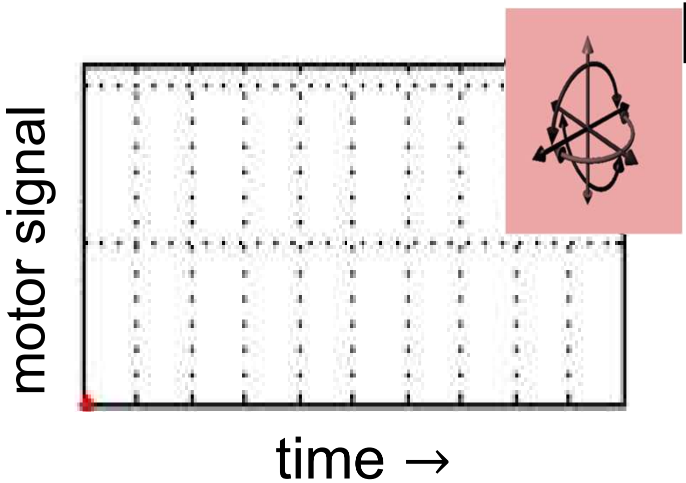
Exploit features for **static scene classification**
(SUN, 397 classes)

30% accuracy increase
when labeled data scarce

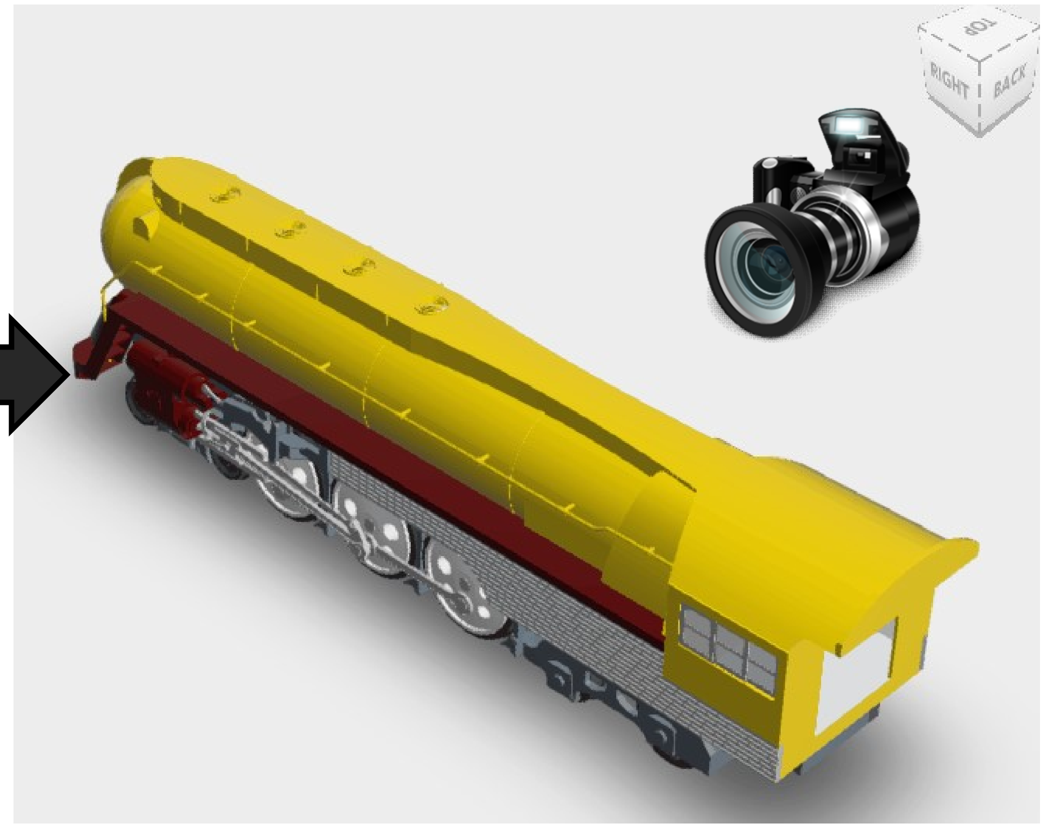
, CVPR '10

Passive → complete ego-motions

Pre-recorded video



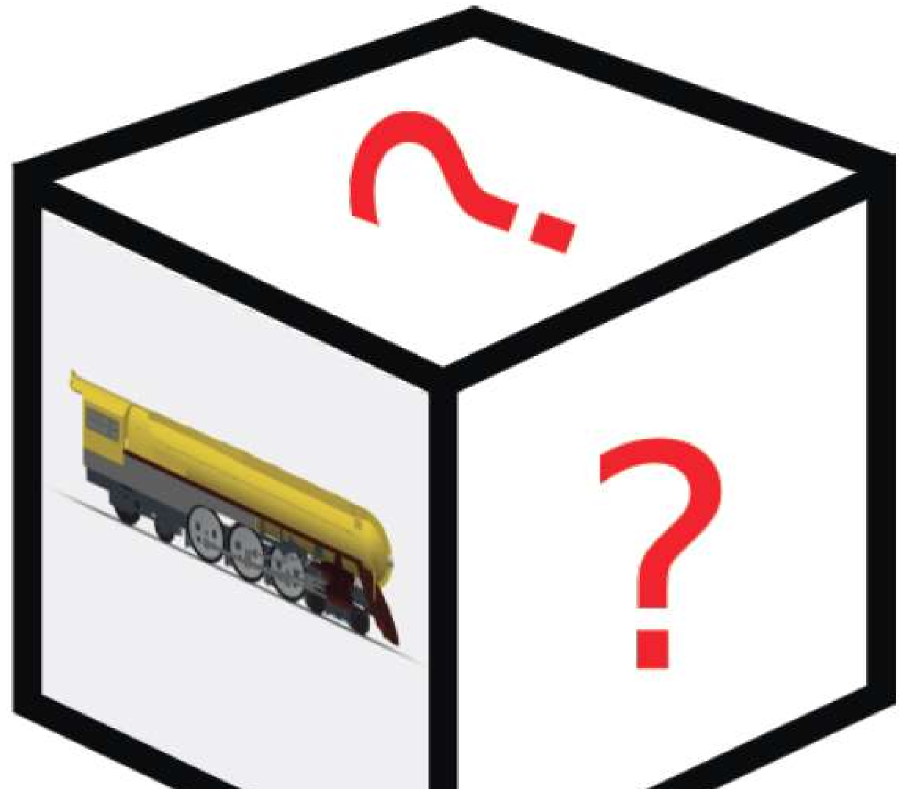
Comprehensive observation



One-shot reconstruction

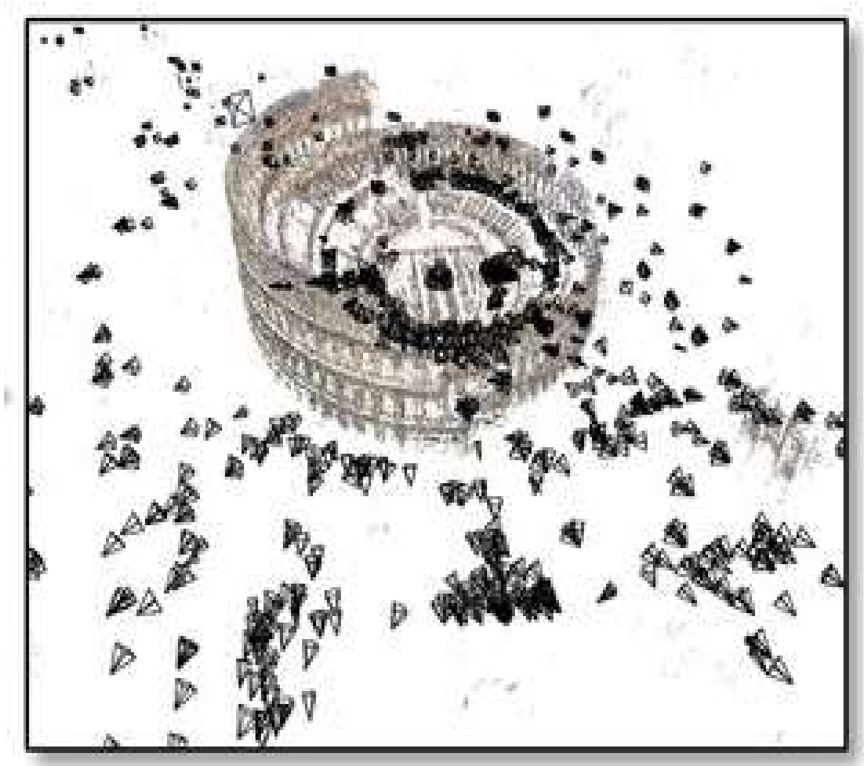
Viewgrid representation

Infer unseen views

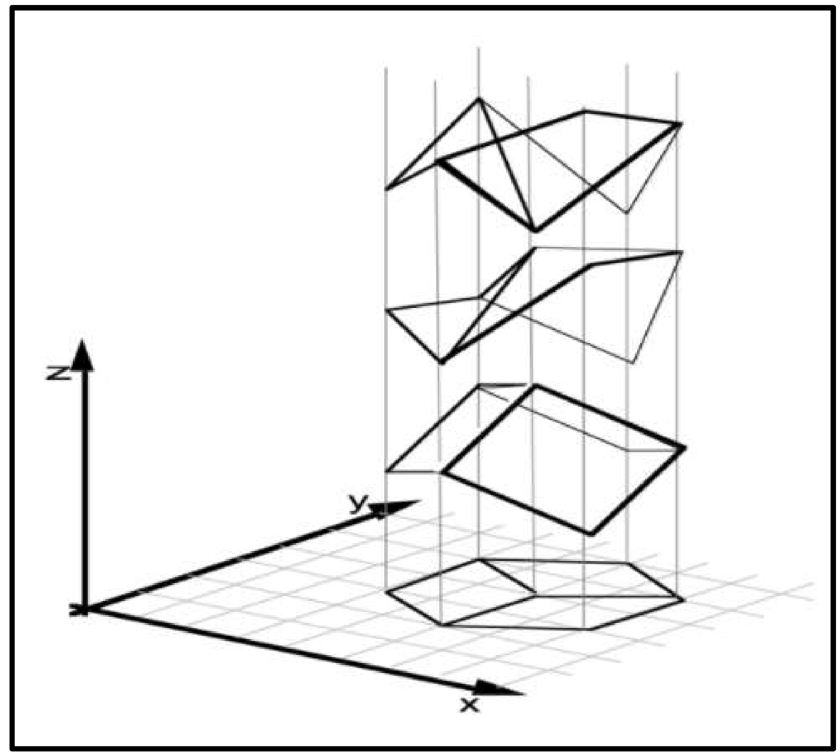


Key idea: One-shot reconstruction as a proxy task to learn semantic features.

One-shot reconstruction



[Snavely et al, CVPR '06]

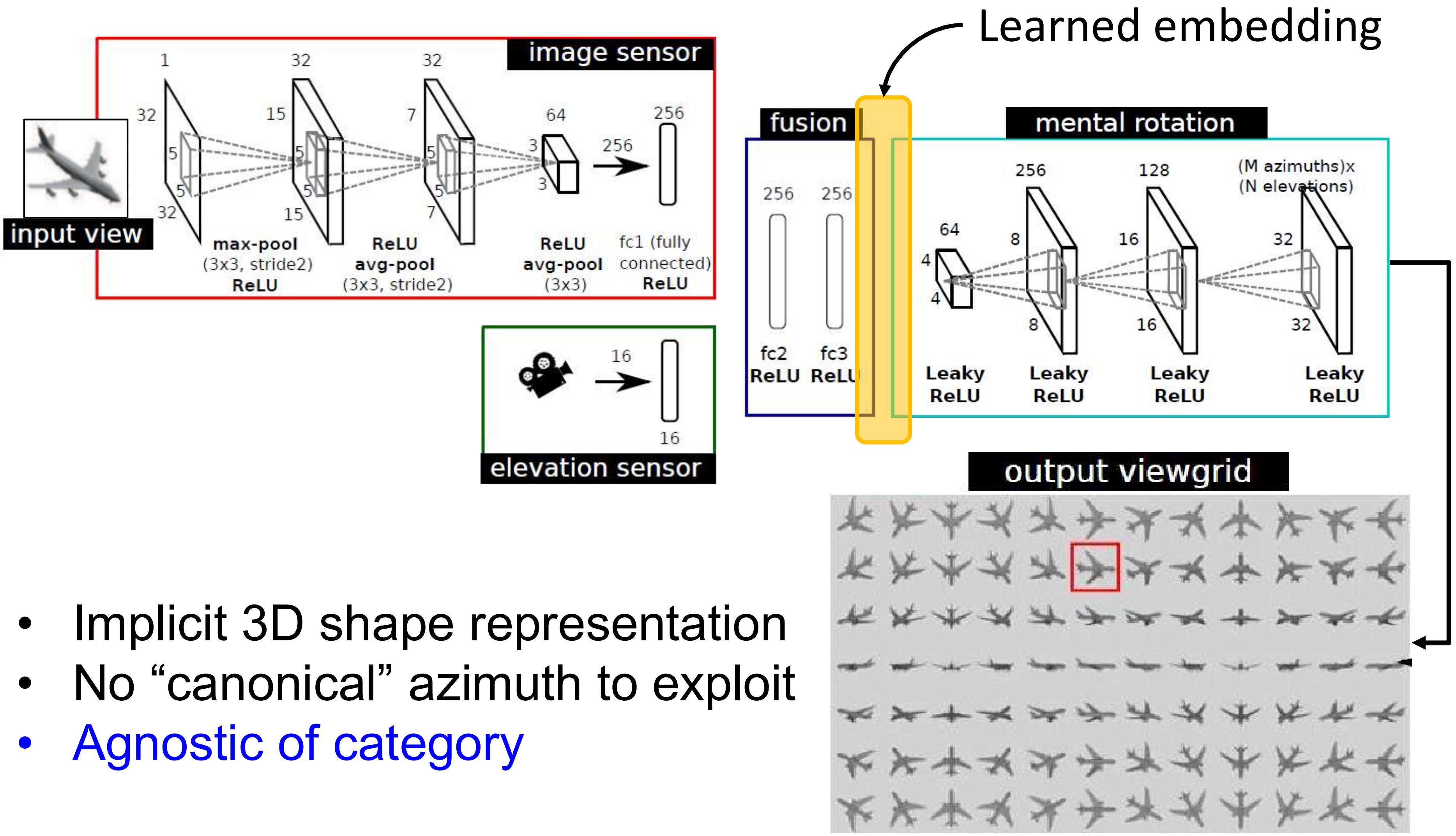


[Sinha et al, ICCV'93]

Shape from dense views
geometric problem

Shape from one view
semantic problem

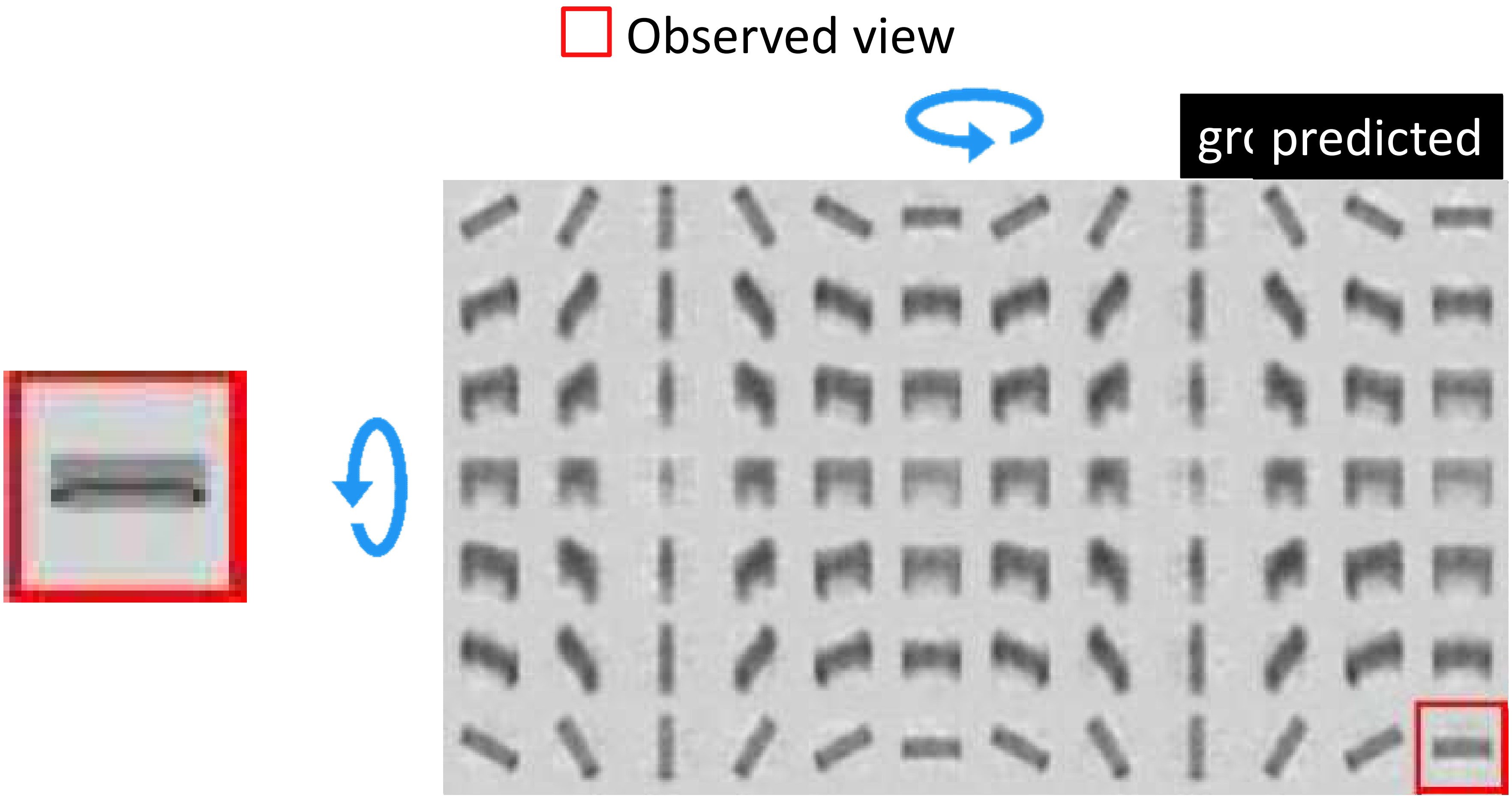
Approach: ShapeCodes



- Implicit 3D shape representation
- No "canonical" azimuth to exploit
- Agnostic of category

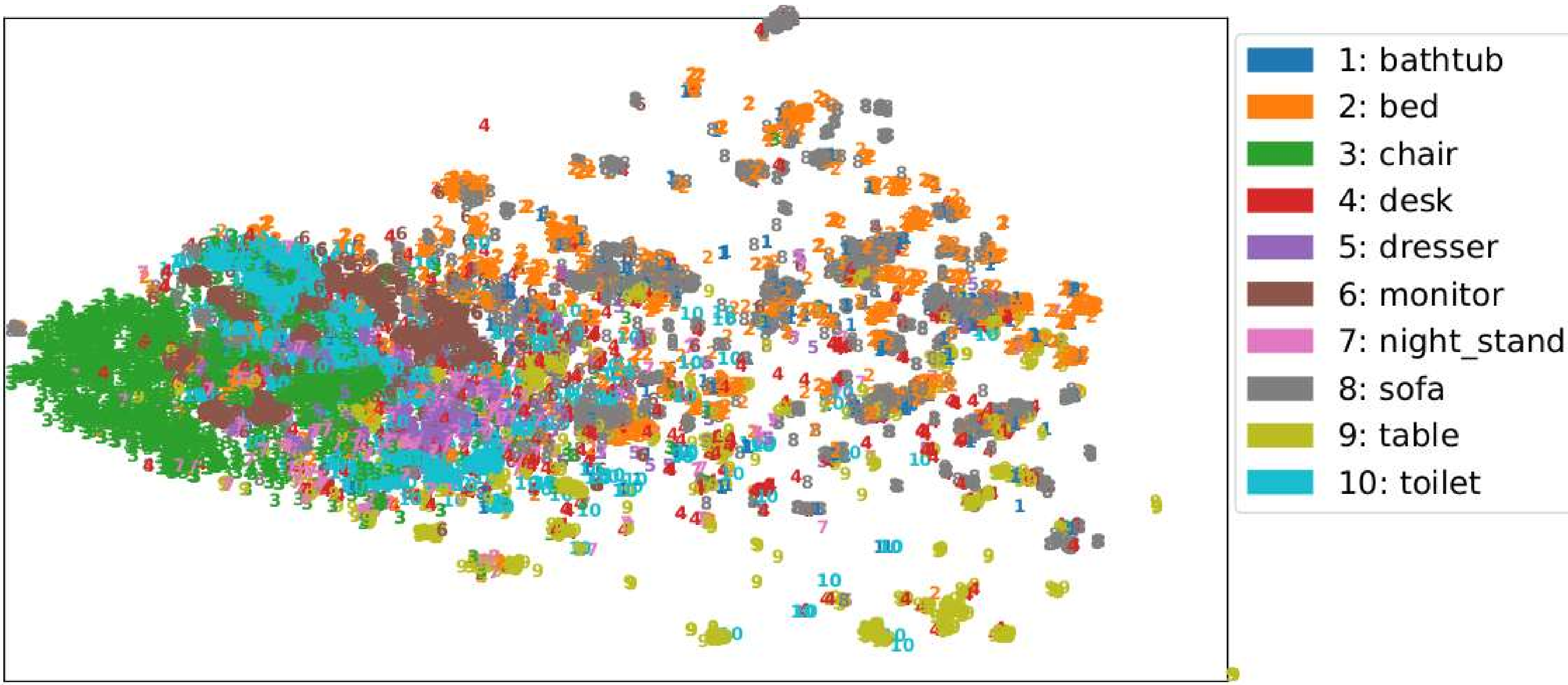
[Jayaraman & Grauman, arXiv 2017]

One-shot reconstruction example



[Jayaraman & Grauman, arXiv 2017]

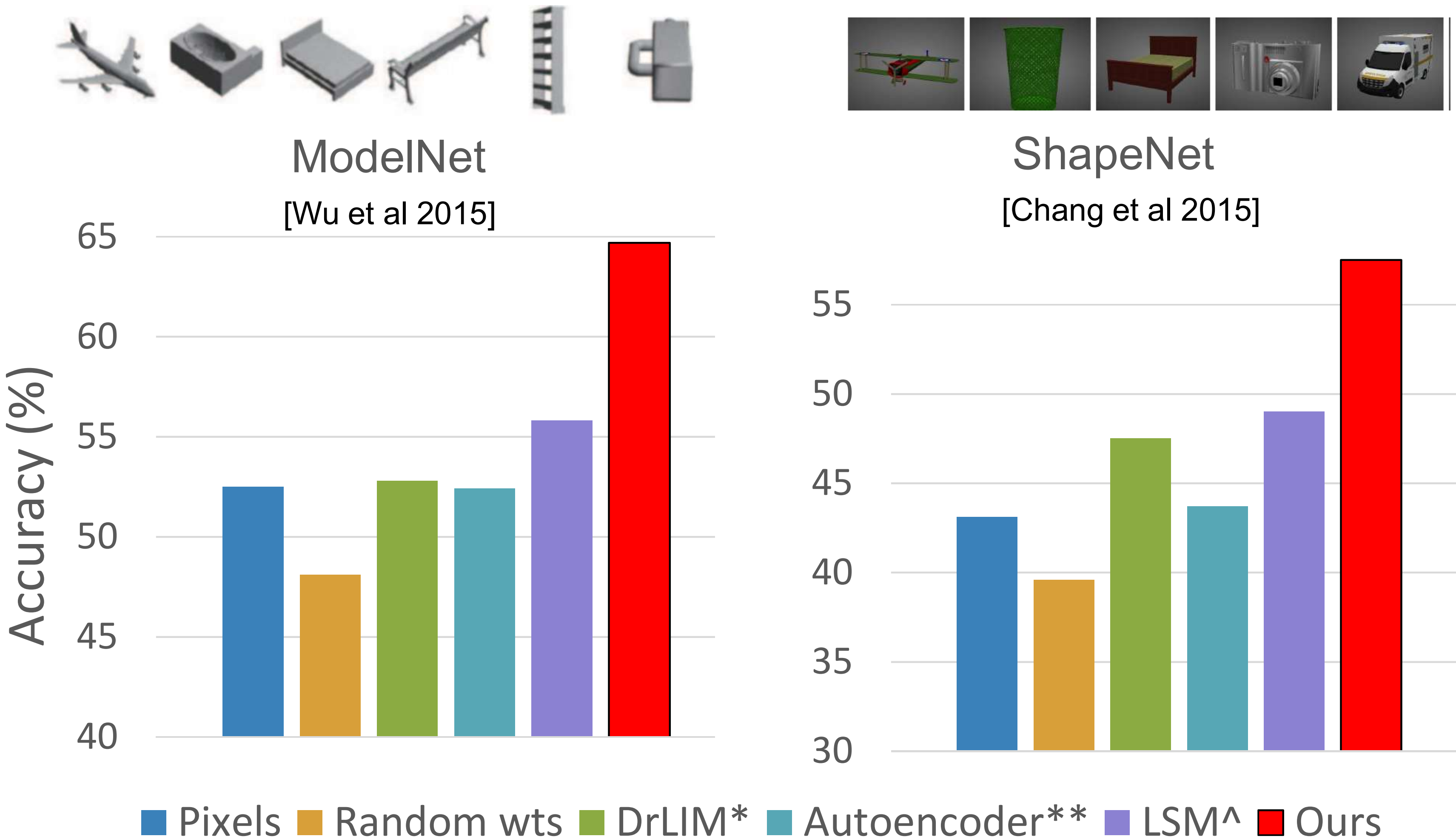
ShapeCodes capture semantics



t-SNE embedding for images of unseen object categories

[Jayaraman & Grauman, arXiv 2017]

ShapeCodes for recognition



*Hadsell et al, Dimensionality reduction by Learning an invariant mapping, CVPR 2005

** Masci et al, Stacked Convolutional Autoencoders for Hierarchical Feature Extraction, ICANN 2011

^Agrawal, Carreira, Malik, Learning to See by Moving, ICCV 2015

Ego-motion and implied body pose

Learn relationship between egocentric scene motion and 3D human body pose

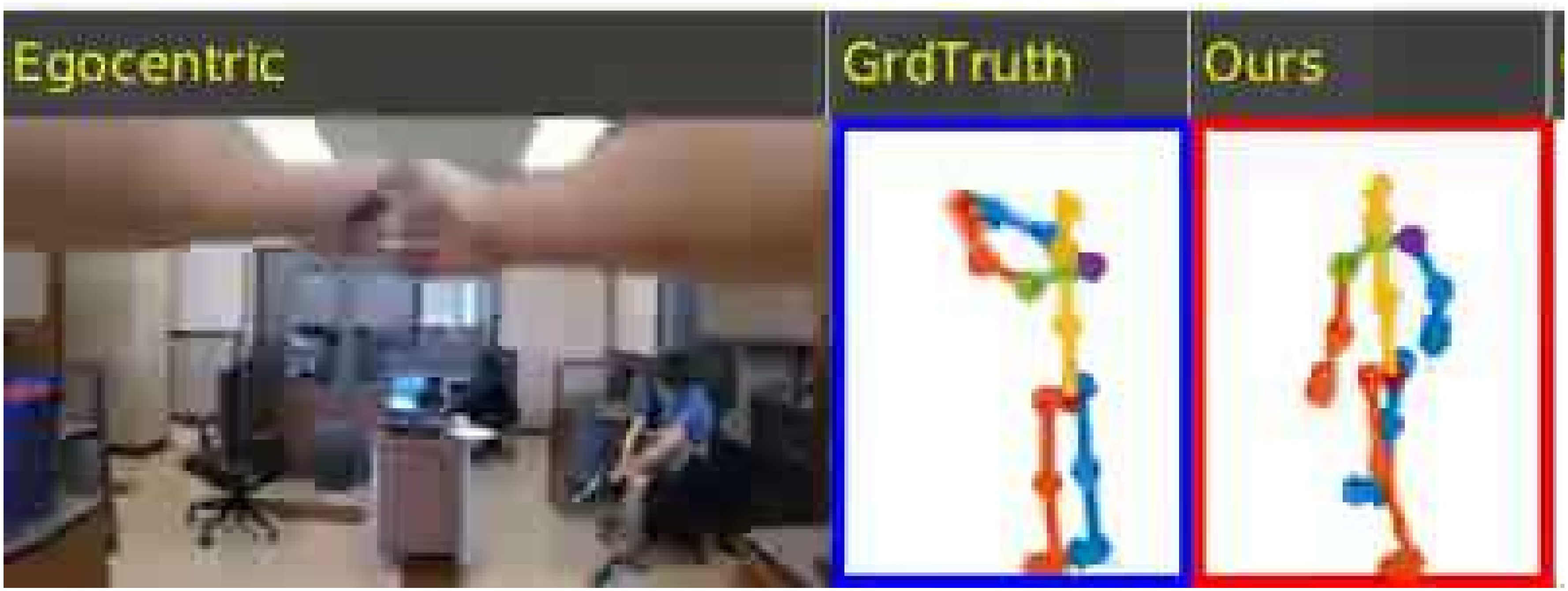
Input:
egocentric video

Output:
sequence of 3d joint positions

[Jiang & Grauman, CVPR 2017]

Ego-motion and implied body pose

Learn relationship between egocentric scene motion and 3D human body pose



Wearable camera video **Inferred pose of camera wearer**

`Videos: http://www.hao-jiang.net/egopose/index.html`

[Jiang & Grauman, CVPR 2017]

This talk

Towards embodied visual learning

1. Learning from unlabeled video and multiple sensory modalities
   a) Egomotion / motor signals
   b) Audio signals
2. Learning policies for how to move for recognition and exploration

Recall: Disembodied visual learning

...

...

Listening to learn

Listening to learn

Listening to learn

woof

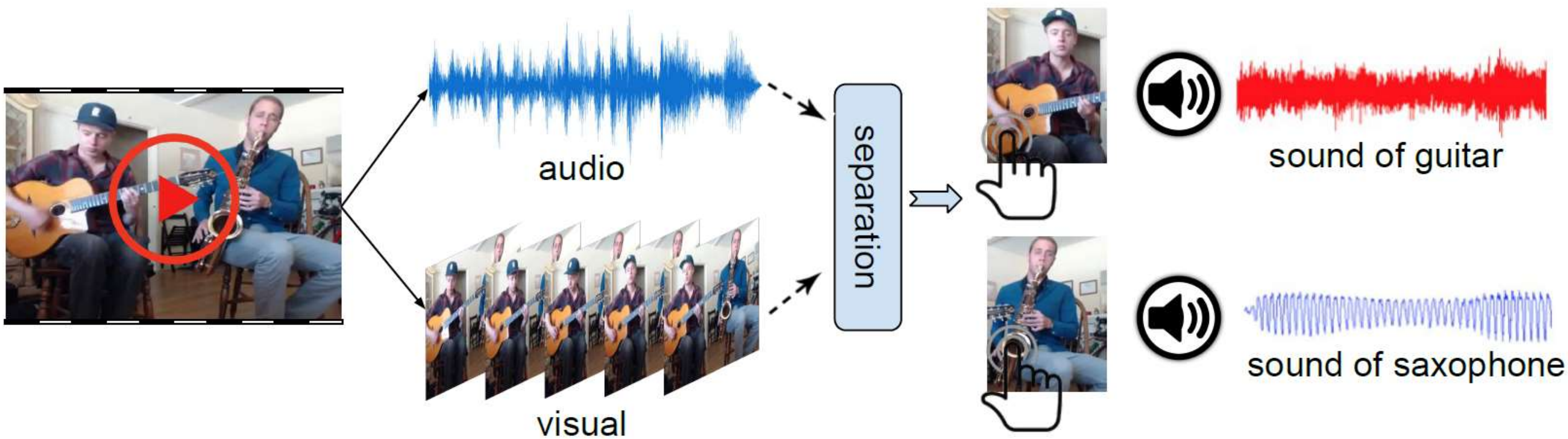
meow

ring

clatter

Goal: A repetoire of objects and their sounds

Visually-guided audio source separation



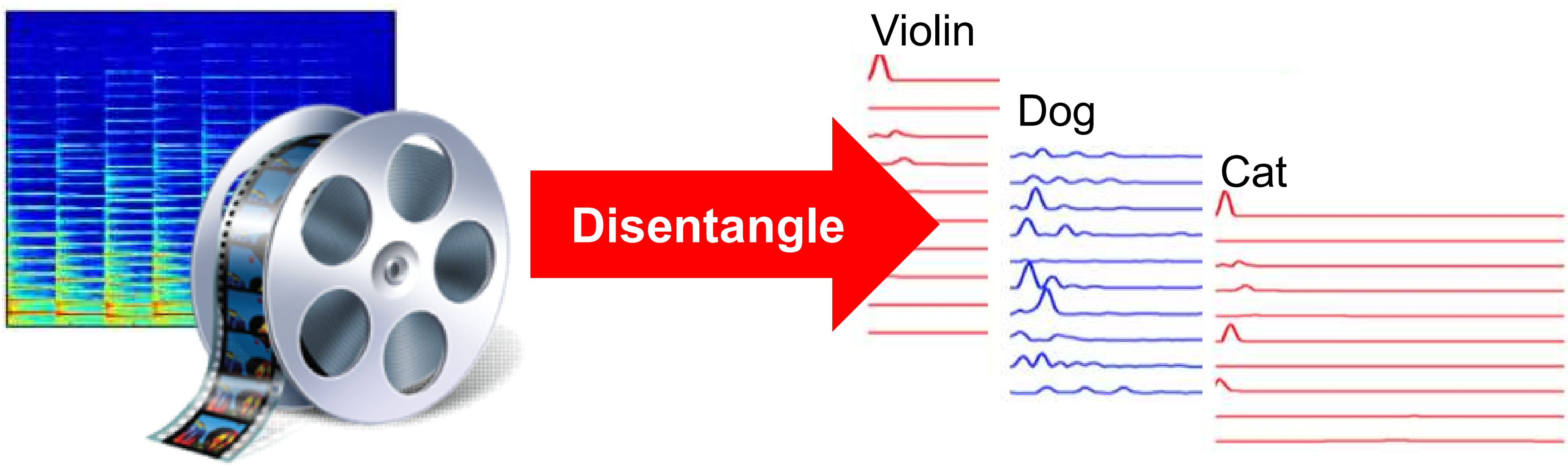
Traditional approach:

- Detect low-level correlations within a single video
- Learn from clean *single audio source* examples

[Darrell et al. 2000; Fisher et al. 2001; Rivet et al. 2007; Barzelay & Schechner 2007; Casanovas et al. 2010; Parekh et al. 2017; Pu et al. 2017; Li et al. 2017]

Learning to separate object sounds

Our idea: Leverage visual objects to learn from *unlabeled* video with *multiple* audio sources

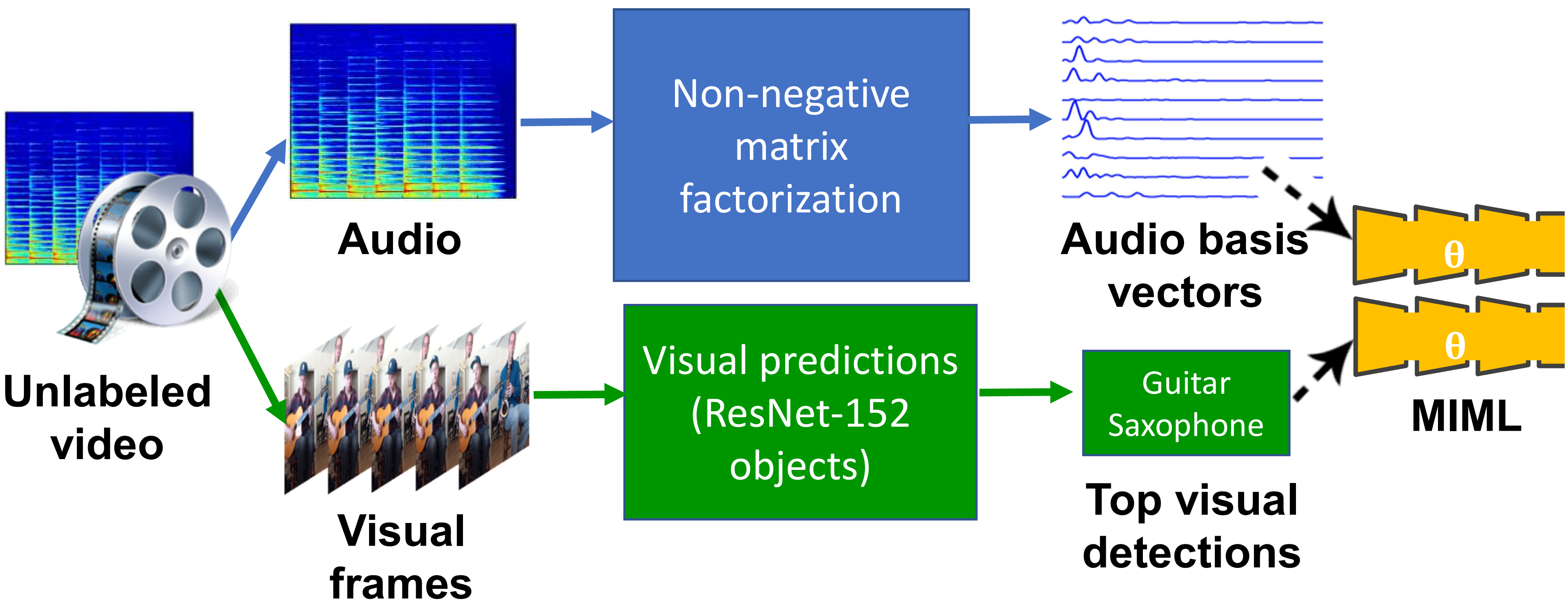


Unlabeled video **Object sound models**

[Gao, Feris, & Grauman, arXiv 2018]

Our approach: training

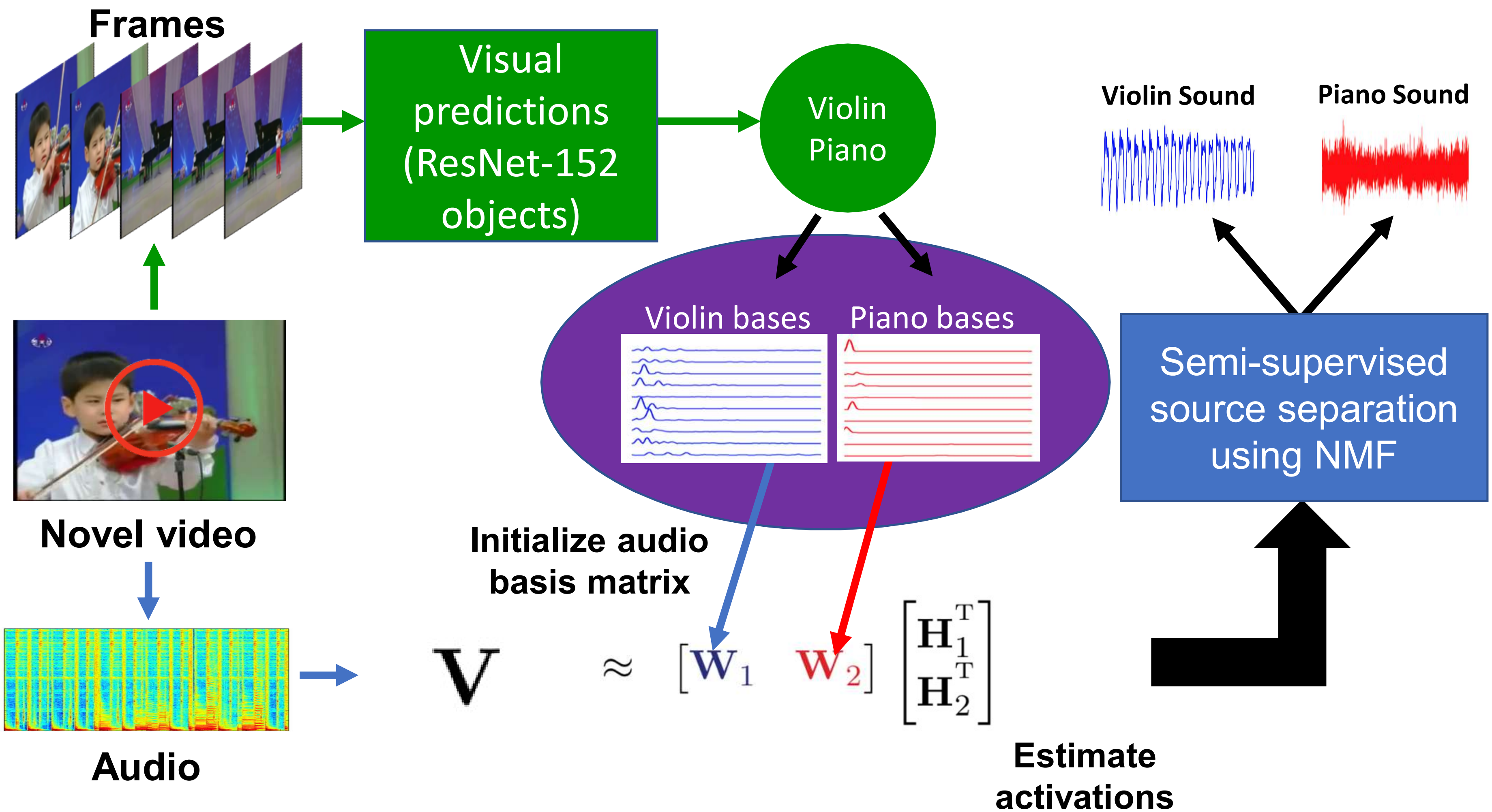
Deep multi-instance multi-label learning (MIML) to disentangle which visual objects make which sounds



Output: **Group of audio basis vectors per object class**

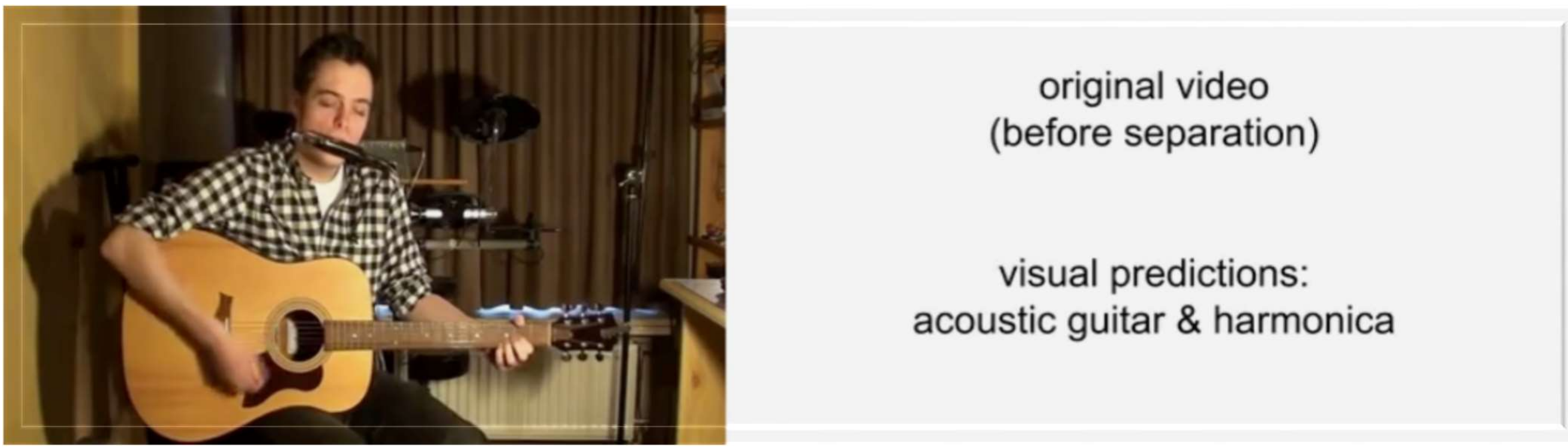
Our approach: inference

Given a novel video, use **discovered object sound models** to guide audio source separation.



Results

Train on 100,000 unlabeled video clips, then separate audio for novel video



Videos:
http://vision.cs.utexas.edu/projects/separating_object_sounds/

Baseline: M. Spiertz, Source-filter based clustering for monaural blind source separation. International Conference on Digital Audio Effects, 2009

[Gao, Feris, & Grauman, arXiv 2018]

Results

Train on 100,000 unlabeled video clips, then separate audio for novel video

Failure cases

Failure cases

Videos:
http://vision.cs.utexas.edu/projects/separating_object_sounds/

[Gao, Feris, & Grauman, arXiv 2018]

Results

	Instrument Pair	Animal Pair	Vehicle Pair	Cross-Domain Pair
Upper-Bound	2.05	0.35	0.60	2.79
K-means Clustering	-2.85	-3.76	-2.71	-3.32
MFCC Unsupervised [65]	0.47	-0.21	-0.05	1.49
Visual Exemplar	-2.41	-4.75	-2.21	-2.28
Unmatched Bases	-2.12	-2.46	-1.99	-1.93
Gaussian Bases	-8.74	-9.12	-7.39	-8.21
Ours	**1.83**	**0.23**	**0.49**	**2.53**

Visually-aided audio source separation (SDR)

	Wooden Horse	Violin Yanni	Guitar Solo	Average
Sparse CCA (Kidron et al. [43])	4.36	5.30	5.71	5.12
JIVE (Lock et al. [50])	4.54	4.43	2.64	3.87
Audio-Visual (Pu et al. [56])	8.82	5.90	**14.1**	9.61
Ours	**12.3**	**7.88**	11.4	**10.5**

Visually-aided audio denoising (NSDR)

Train on 100K unlabeled video clips from AudioSet [Gemmeke et al. 2017]

This talk

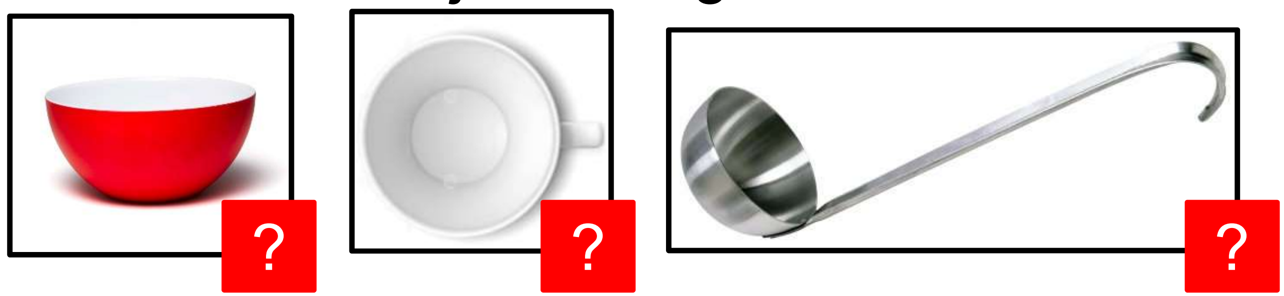
Towards embodied visual learning

1. Learning from unlabeled video and multiple sensory modalities
2. Learning policies for how to move for recognition and exploration

Current recognition benchmarks

Passive, disembodied snapshots at *test* time, too

Object recognition



Scene recognition

Moving to recognize

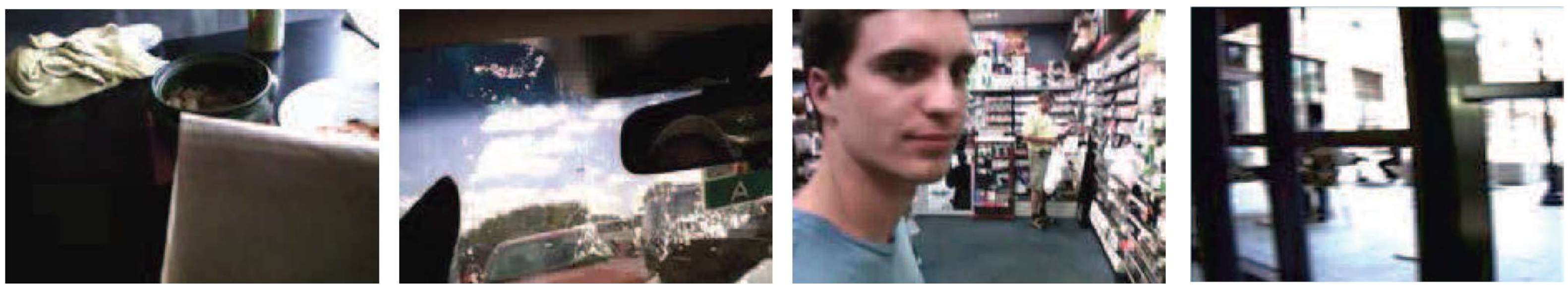


Time to revisit active recognition in challenging settings!

Bajcsy 1985, Aloimonos 1988, Ballard 1991, Wilkes 1992, Dickinson 1997, Schiele & Crowley 1998, Tsotsos 2001, Denzler 2002, Soatto 2009, Ramanathan 2011, Borotschnig 2011, …

Moving to recognize

Difficulty: unconstrained visual input

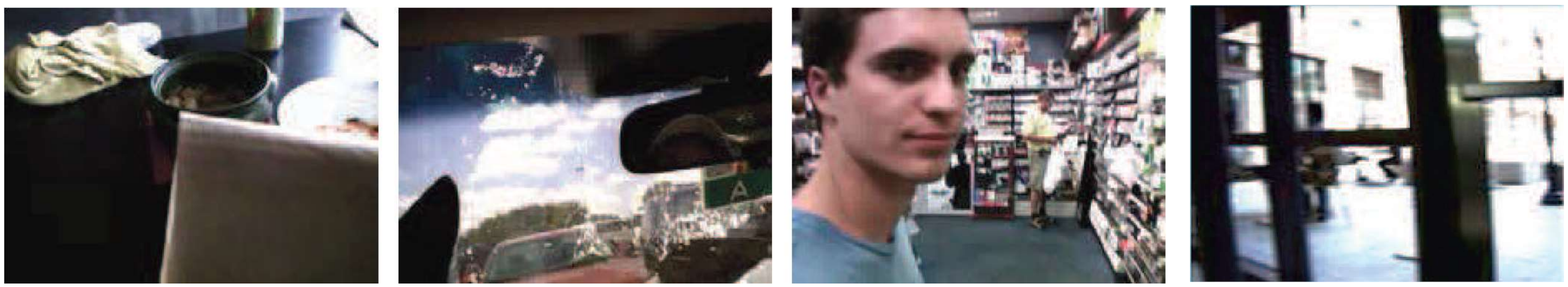


vs.

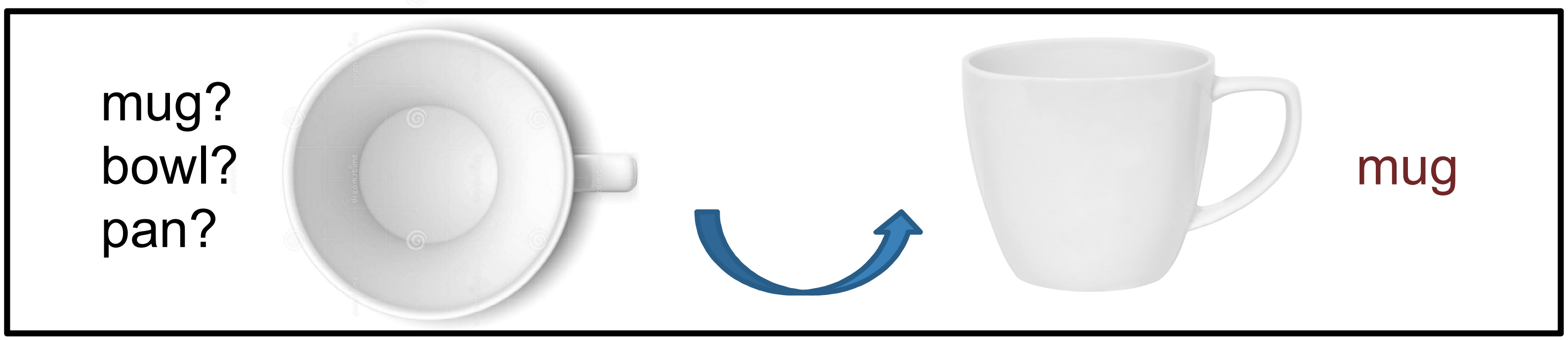
ImageNet Web images

Moving to recognize

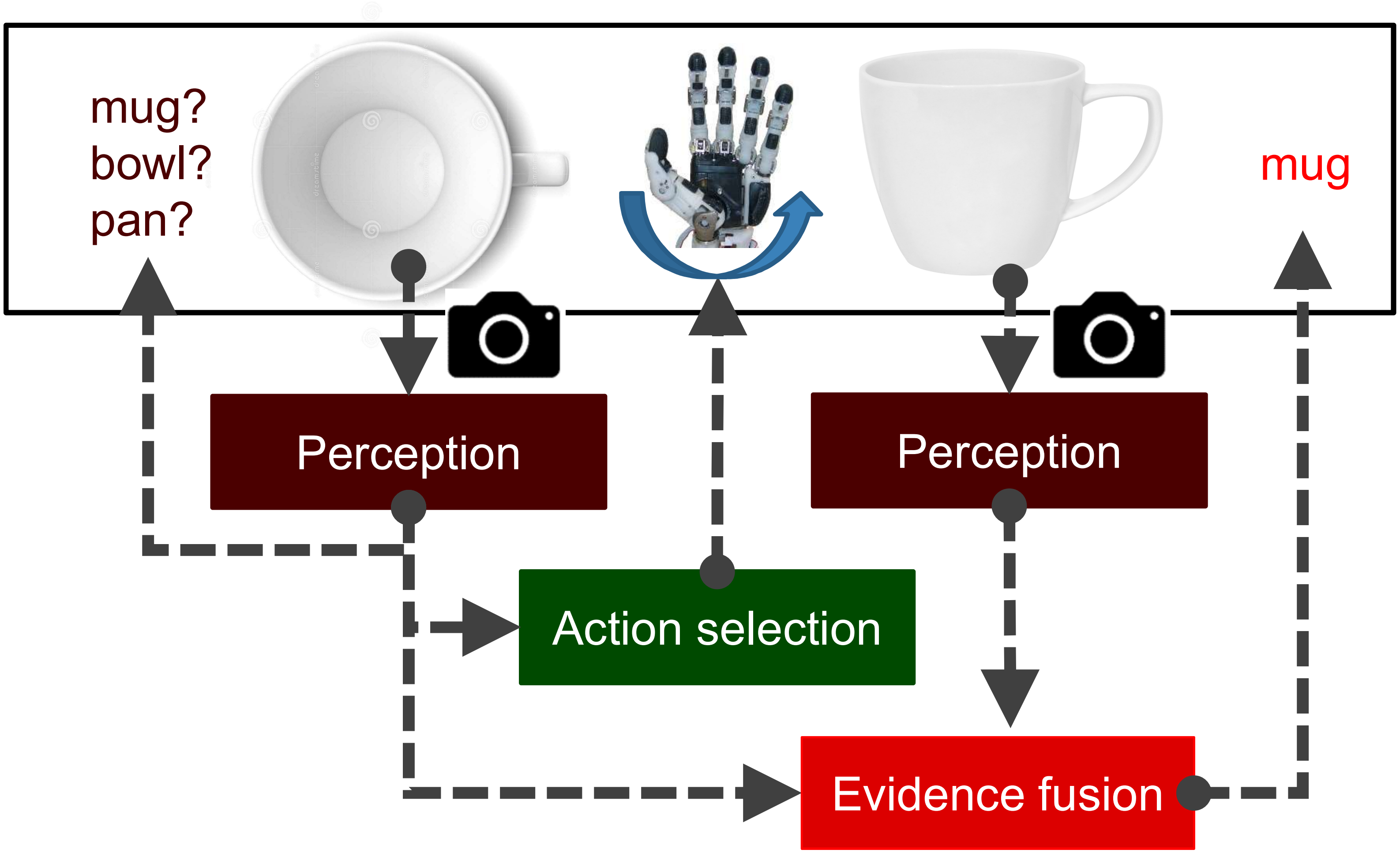
Difficulty: unconstrained visual input



Opportunity: ability to move to *change* input



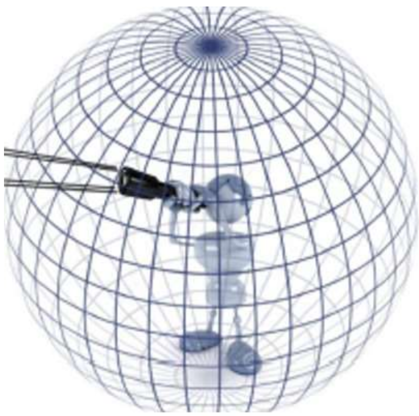
End-to-end active recognition



Jayaraman and Grauman, ECCV 2016

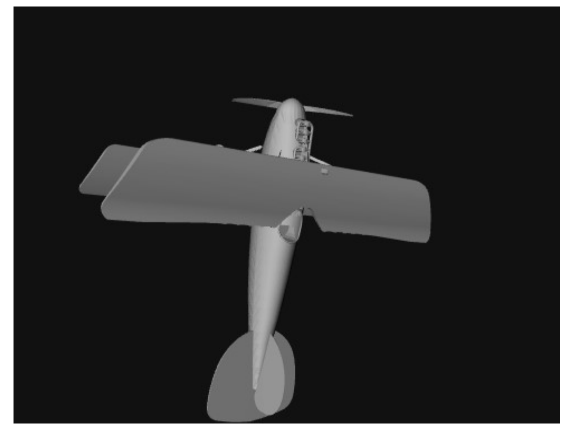
End-to-end active recognition

Look around scene



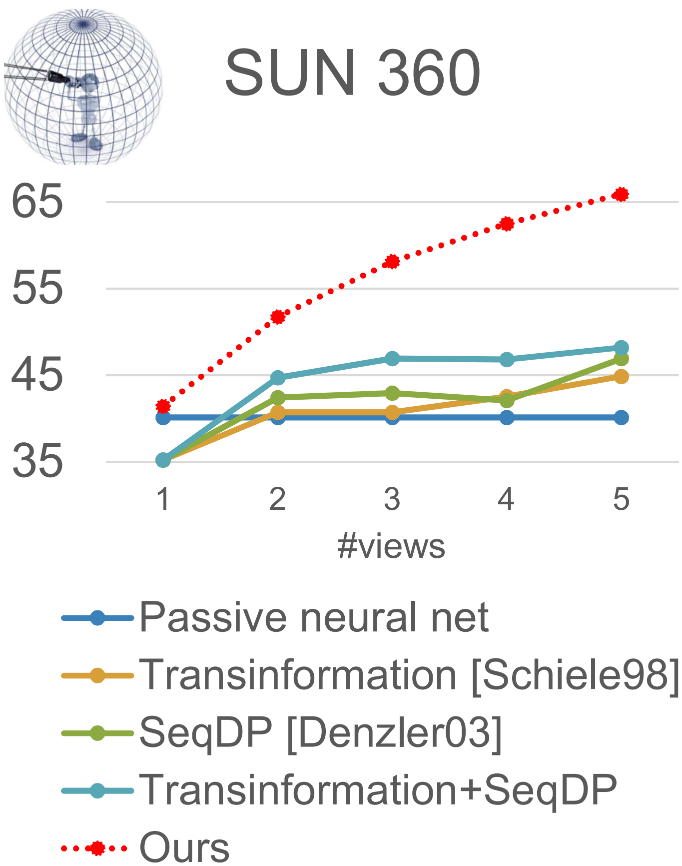
Manipulate object

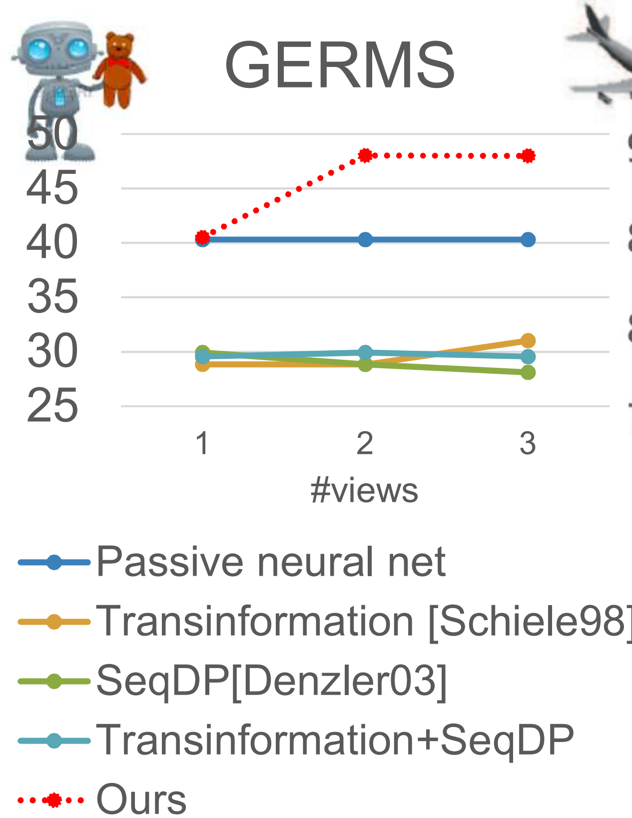
Move around an object

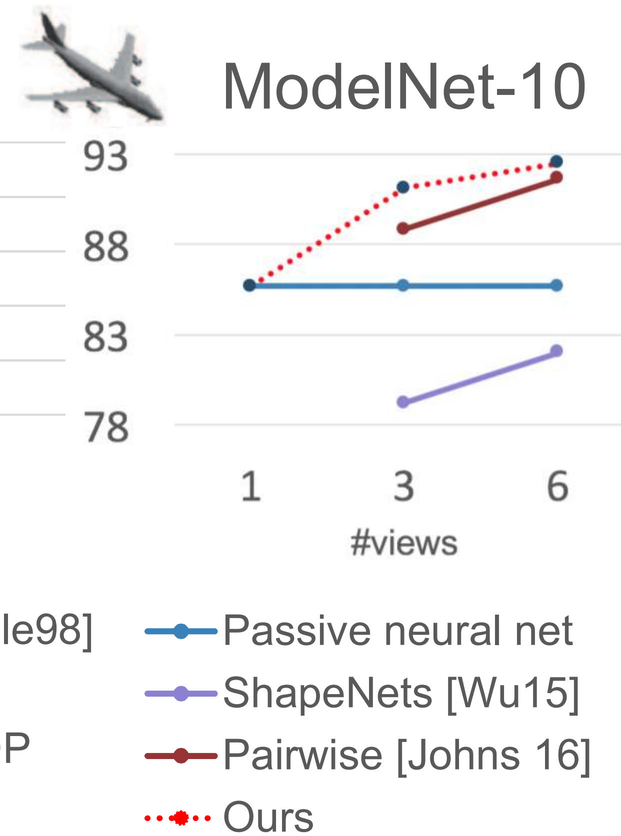


[Jayaraman and Grauman, ECCV 2016]

End-to-end active recognition







Agents that learn to look around intelligently can recognize things faster.

[Jayaraman and Grauman, ECCV 2016]

End-to-end active recognition: example

Plaza courtyard

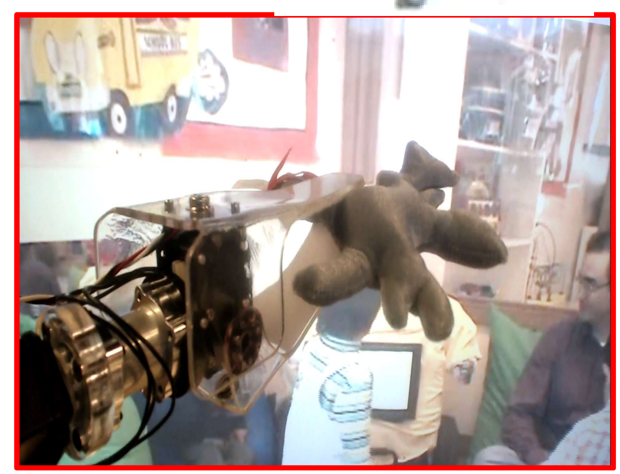
[Jayaraman and Grauman, ECCV 2016]

End-to-end active recognition: example

Predicted label:

T=1

T=2



T=3

GERMS dataset: Malmir et al. BMVC 2015

[Jayaraman and Grauman, ECCV 2016]

Goal: Learn to "look around"

vs.

recognition

reconnaissance

search and rescue

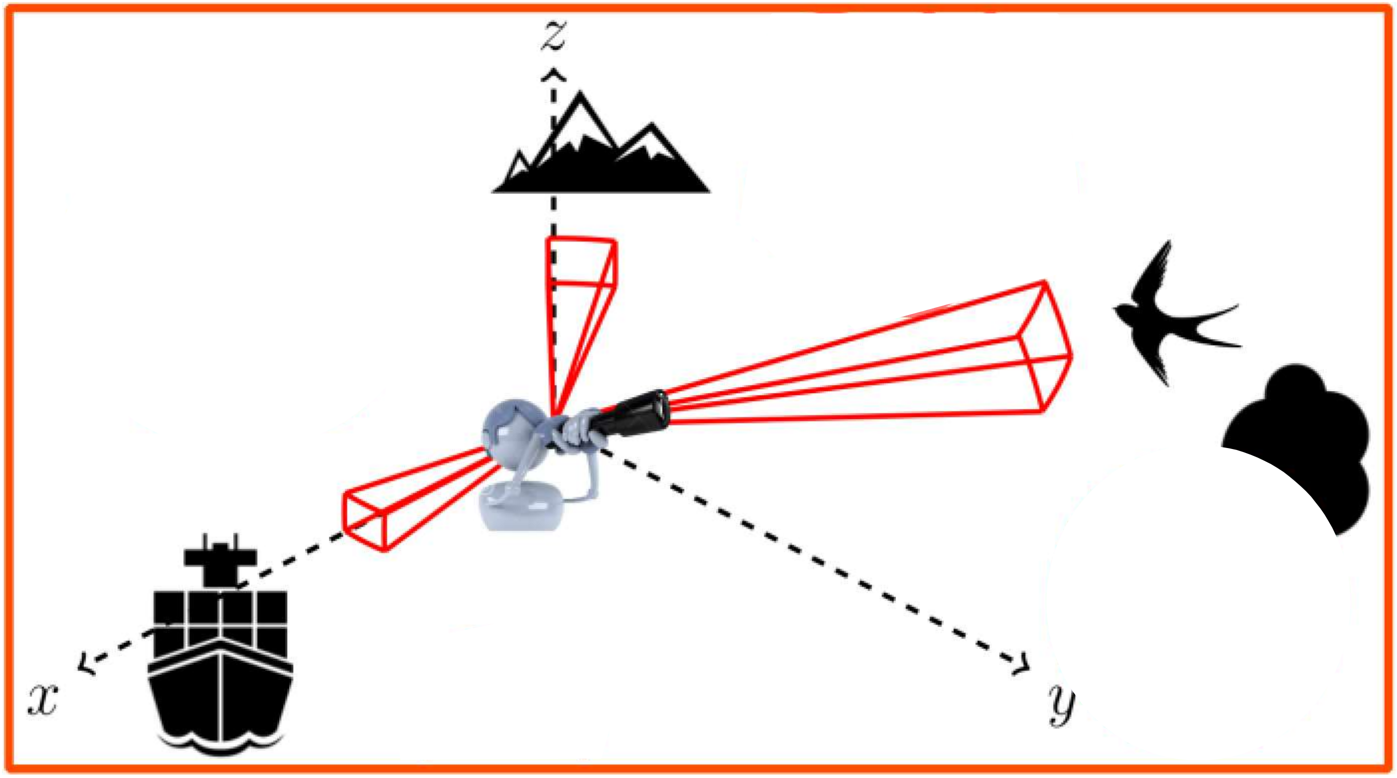
task predefined

task unfolds dynamically

Can we learn look-around policies for visual agents that are curiosity-driven, exploratory, and generic?

Key idea: Active observation completion

Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment

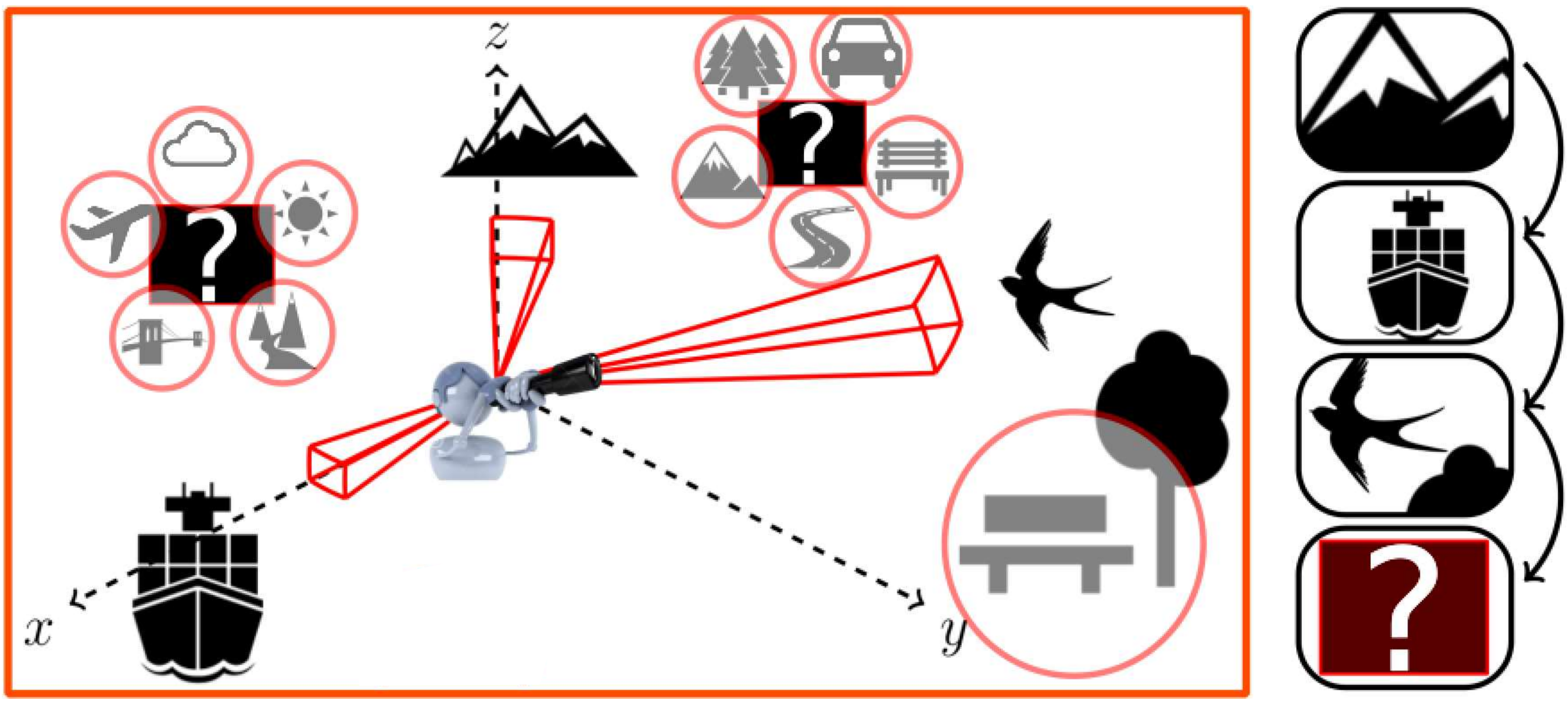


Agent must choose where to look *before* looking there.

Jayaraman and Grauman, CVPR 2018

Key idea: Active observation completion

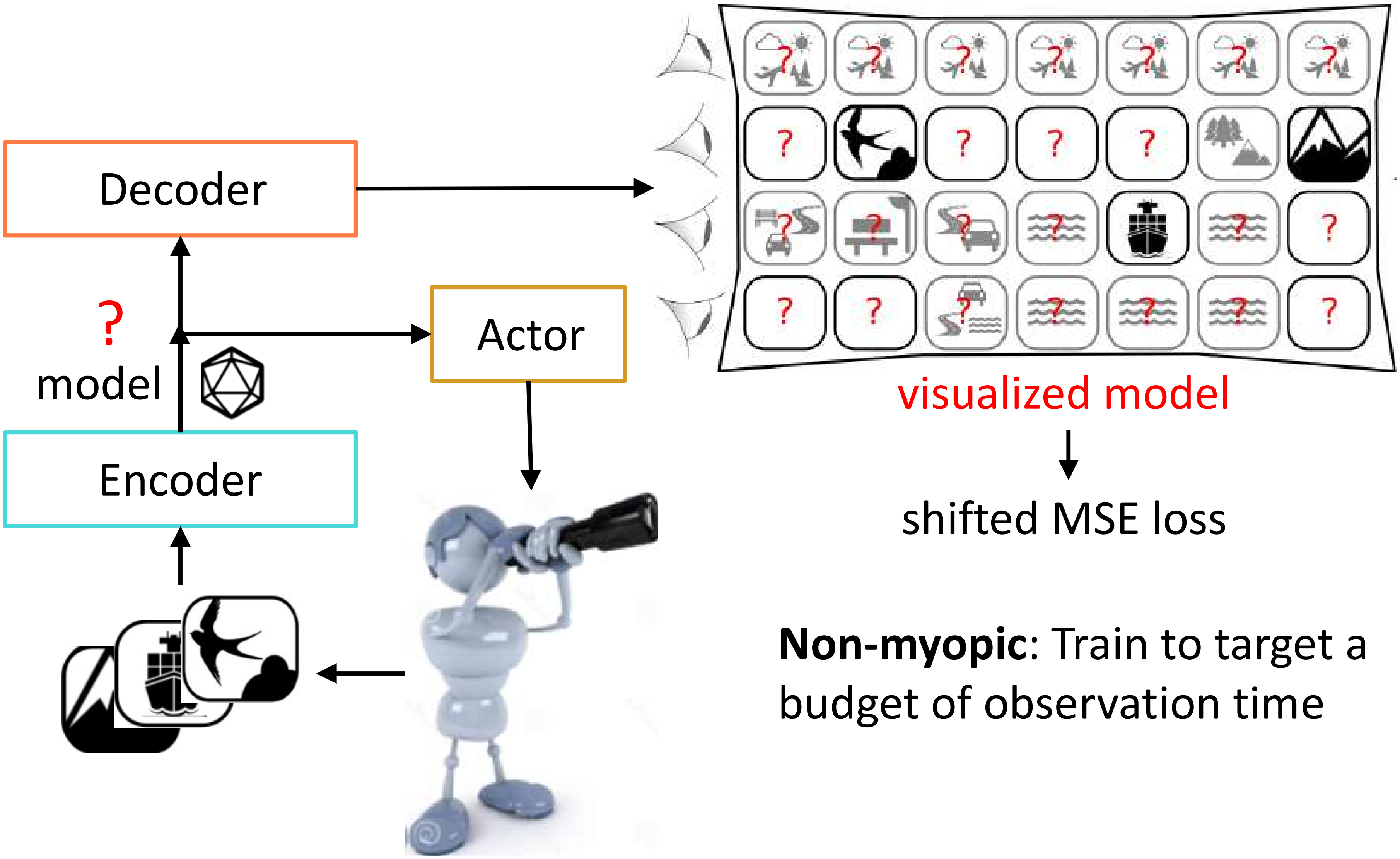
Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment



Agent must choose where to look *before* looking there.

Jayaraman and Grauman, CVPR 2018

Approach: Active observation completion

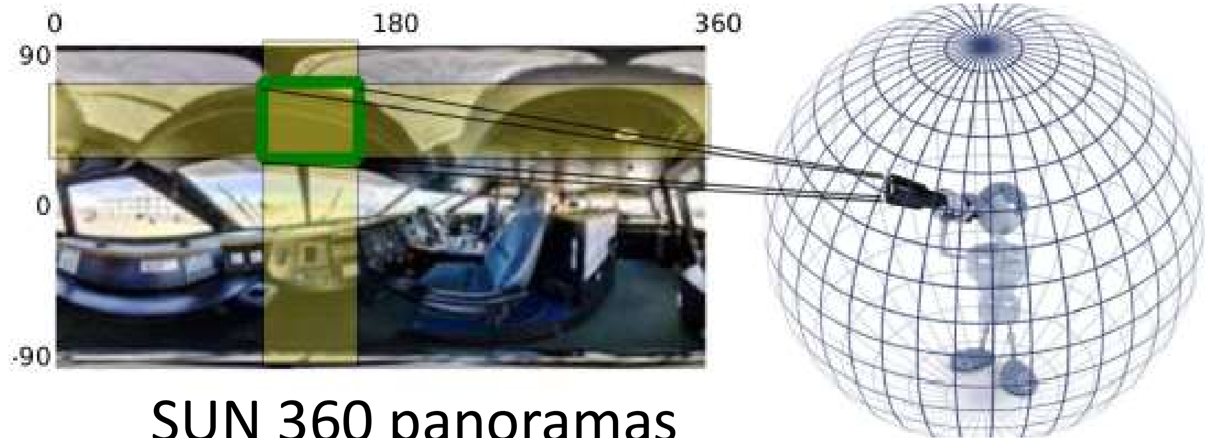


Non-myopic: Train to target a budget of observation time

Jayaraman and Grauman, CVPR 2018

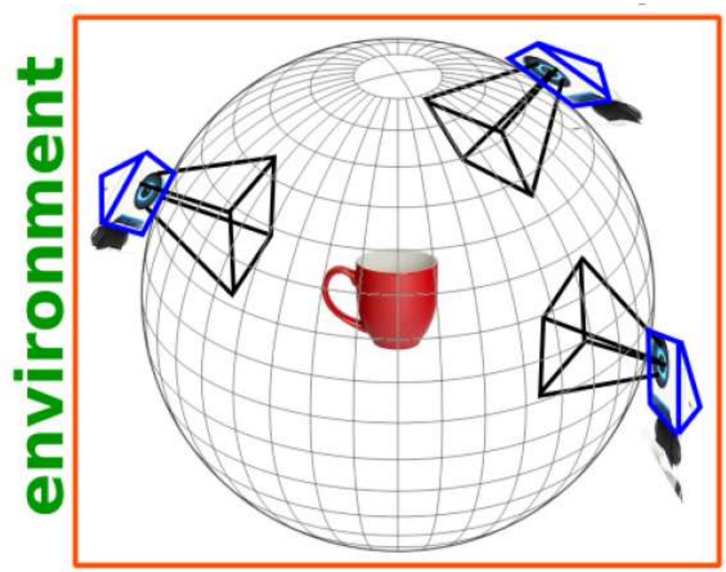
Two scenarios

Where to
look next?

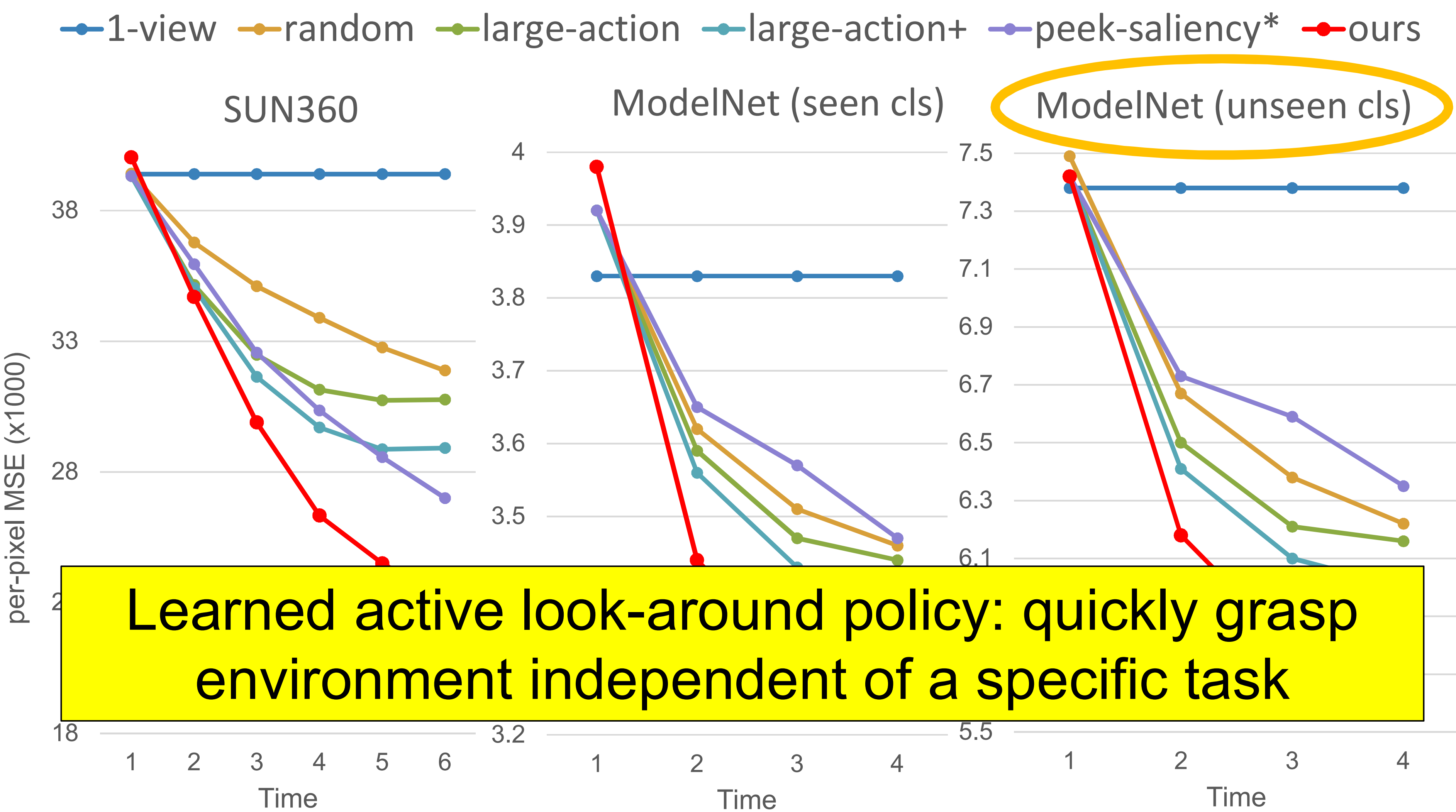


SUN 360 panoramas
[Xiao 2012]

How to
manipulate?



Active "look around" results

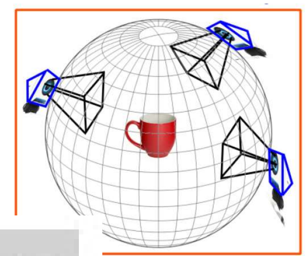


Learned active look-around policy: quickly grasp environment independent of a specific task

*Harel et al, Graph based Visual Saliency, NIPS'07

Jayaraman and Grauman, CVPR 2018

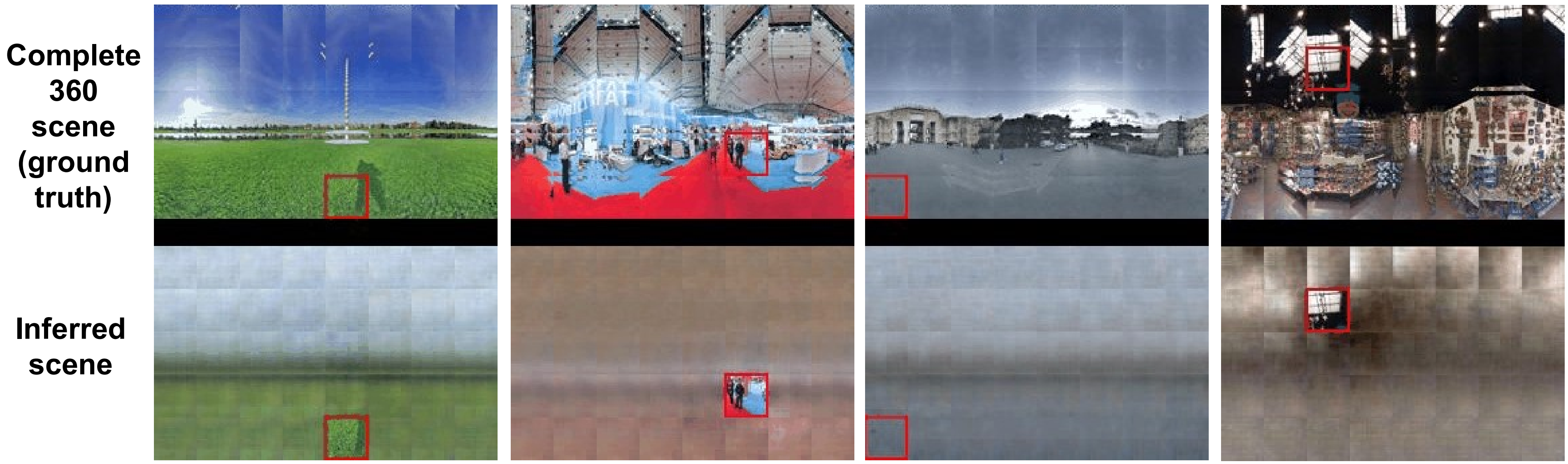
Active "look around" visualization



Agent's mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

Active "look around" visualization

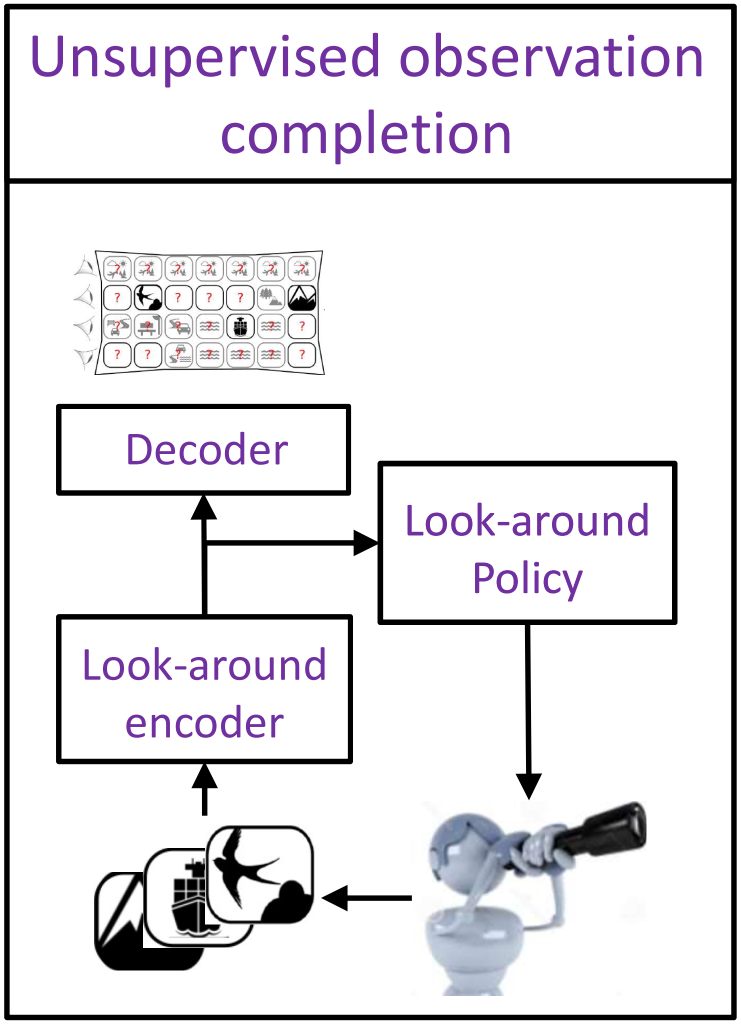


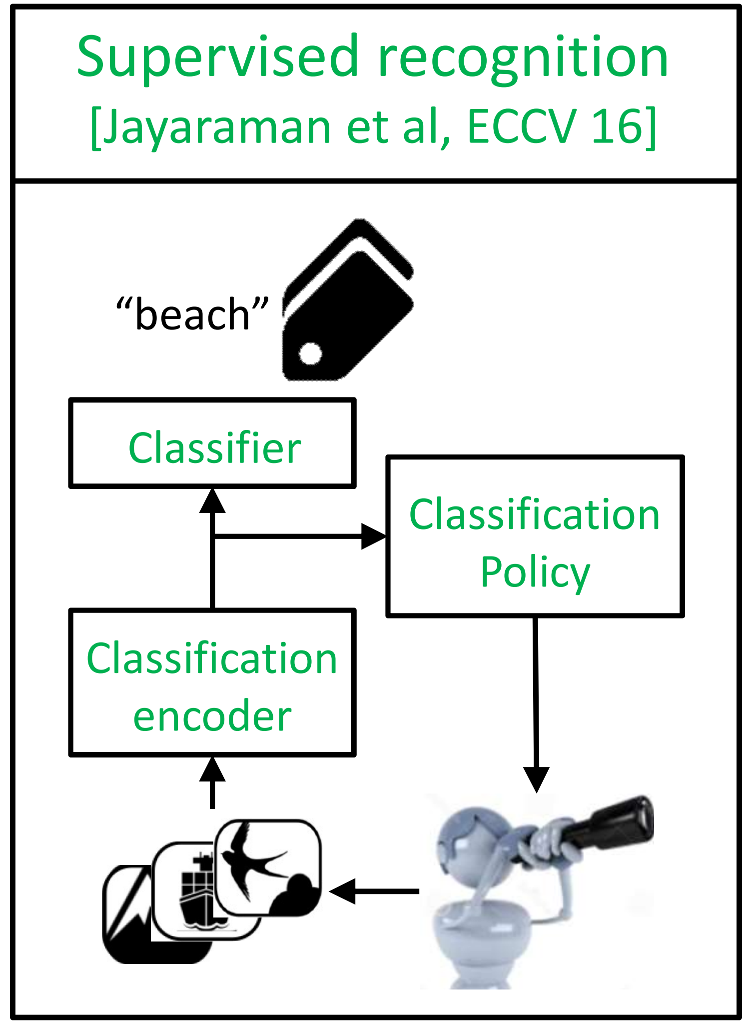
= observed views

Agent's mental model for 360 scene evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

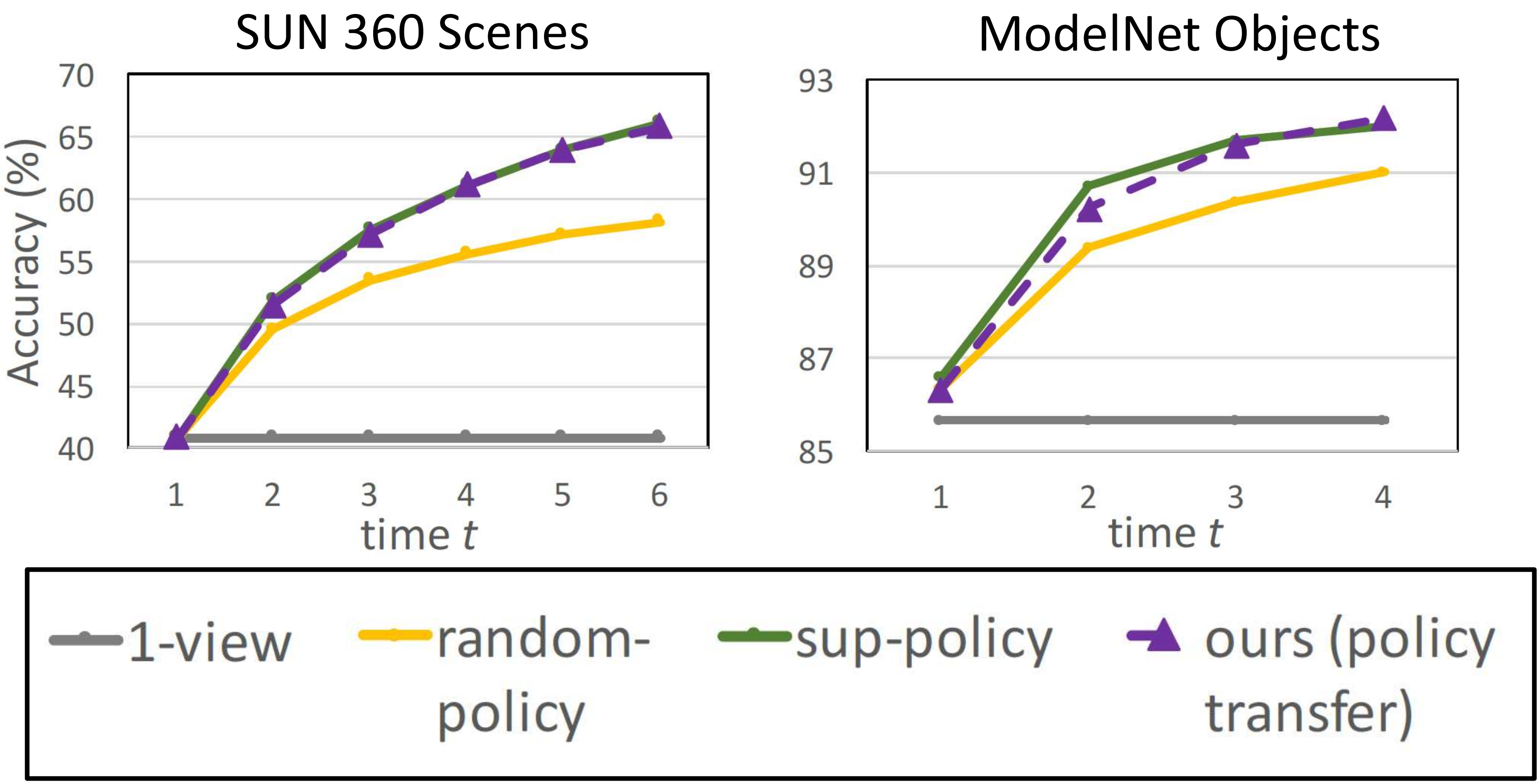
Motion policy transfer





Plug observation completion policy in for new task

Motion policy transfer



Unsupervised exploratory policy approaches supervised task-specific policy accuracy!

Summary

- Visual learning benefits from
  - context of action and motion in the world
  - continuous unsupervised observations
- New ideas:
  - Embodied feature learning via visual and motor signals
  - Learning to separate object sound models from unlabeled video
  - Active policies for view selection and camera control

Dinesh Jayaraman

Ruohan Gao

Papers

- **Learning to Separate Object Sounds by Watching Unlabeled Video**. R. Gao, R. Feris, and K. Grauman. arXiv:1804.01665, April 2018. videos
- **Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks**. D. Jayaraman and K. Grauman. CVPR 2018.
- **Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video**. H. Jiang and K. Grauman. CVPR 2017.
- **Learning Image Representations Tied to Egomotion from Unlabeled Video**. D. Jayaraman and K. Grauman. International Journal of Computer Vision (IJCV), Special Issue for Best Papers of ICCV 2015, Mar 2017.
- **Look-Ahead Before You Leap: End-to-End Active Recognition by Forecasting the Effect of Motion**. D. Jayaraman and K. Grauman. ECCV 2016.
- **Unsupervised learning through one-shot image-based shape reconstruction**, D. Jayaraman, R. Gao, K. Grauman. arXiv 2017

http://www.cs.utexas.edu/~grauman/research/pubs.html