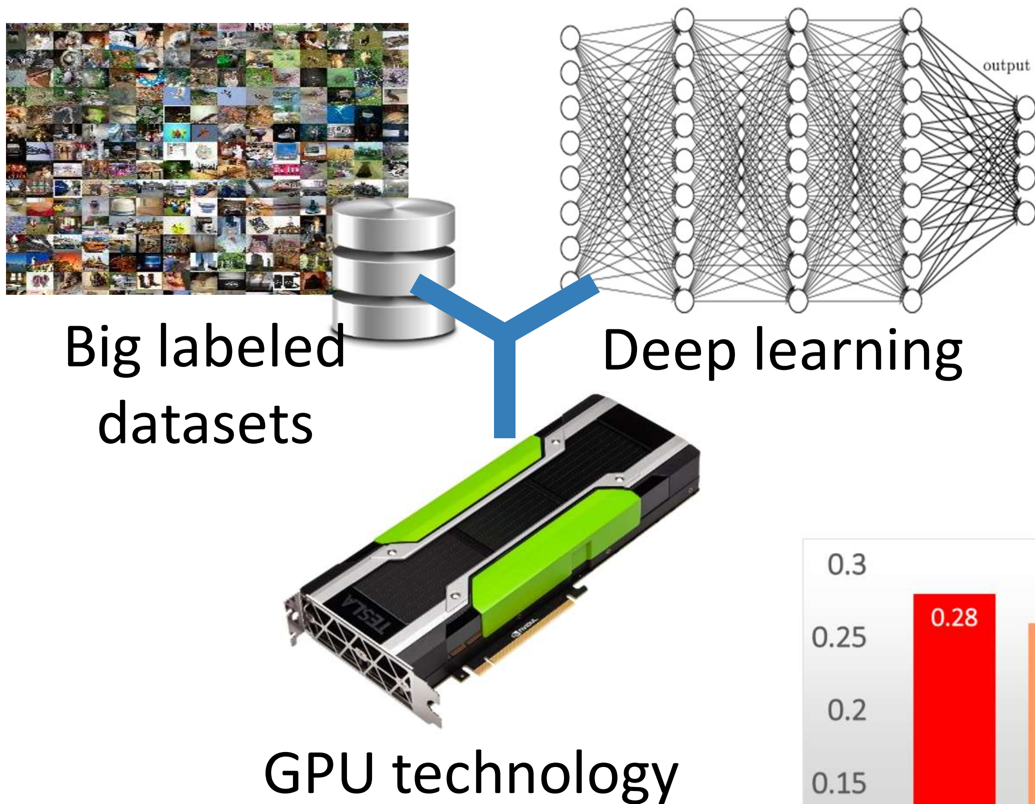
Learning Where to Look and Listen: Egocentric and 360 Computer Vision

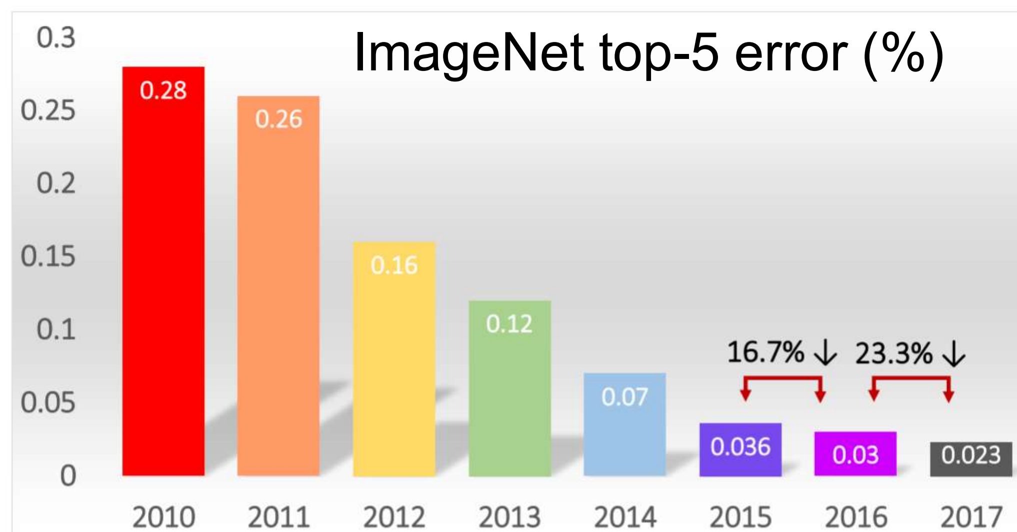
Kristen Grauman

Facebook AI Research

University of Texas at Austin

Visual recognition: significant recent progress





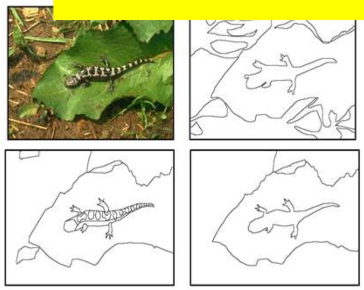
How do vision systems learn today?

...

...

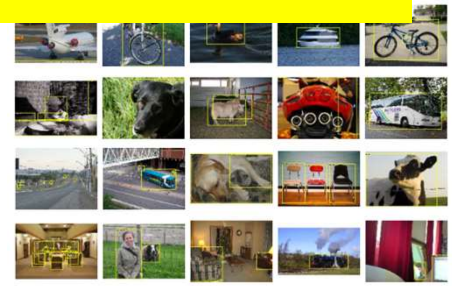
Web photos + vision

A “disembodied” well-curated moment in time

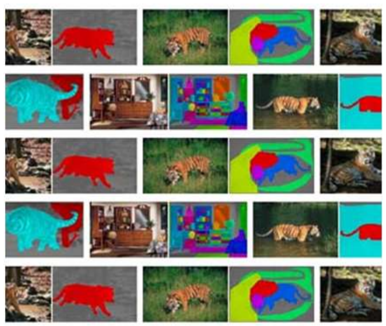


BSD (2001)

Caltech 101 (2004), Caltech 256 (2006)



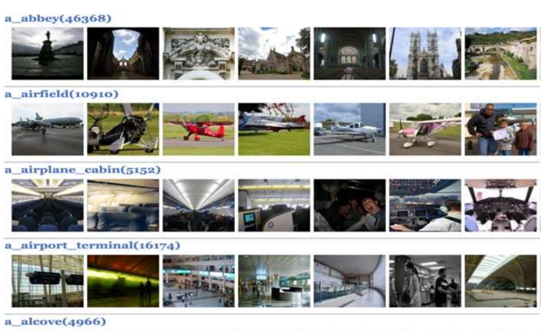
PASCAL (2007-12)



LabelMe (2007)

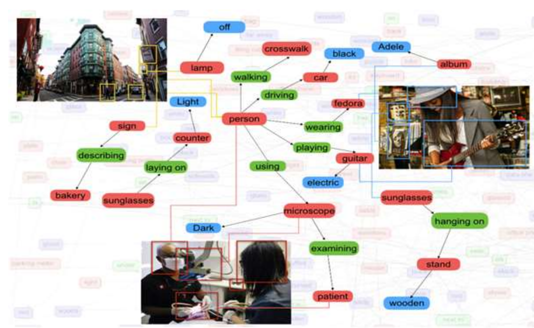
ImageNet (2009)

SUN (2010)



Places (2014)

MS COCO (2014)



Visual Genome (2016)

Egocentric perceptual experience

A tangle of relevant and irrelevant multi-sensory information

Egocentric perceptual experience

A tangle of relevant and irrelevant multi-sensory information

First-person video

360 video

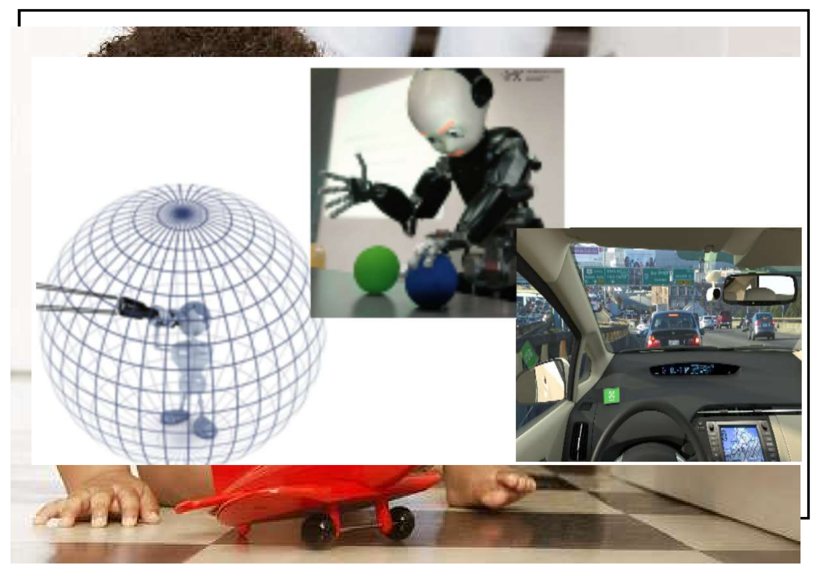
Big picture goal: Embodied visual learning

Status quo:

Learn from "disembodied" bag of labeled snapshots.

On the horizon:

Visual learning in the context of action, motion, and multi-sensory observations.



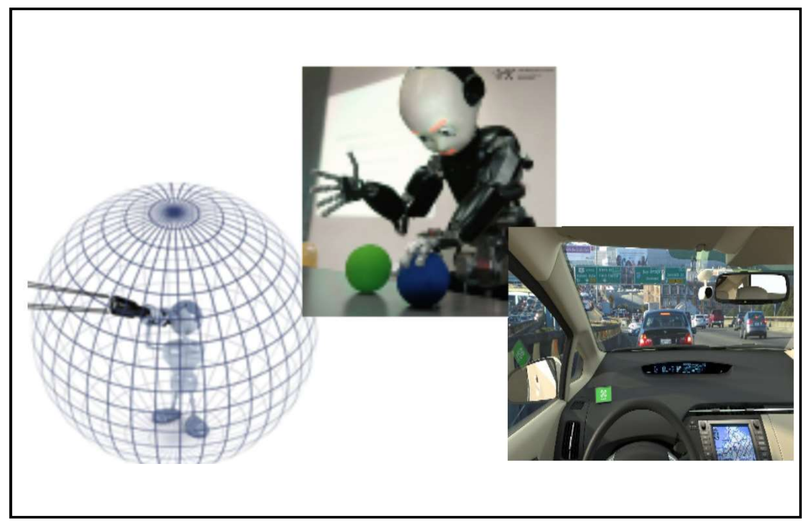
Big picture goal: Embodied visual learning

Status quo:

Learn from "disembodied" bag of labeled snapshots.

On the horizon:

Visual learning in the context of action, motion, and multi-sensory observations.

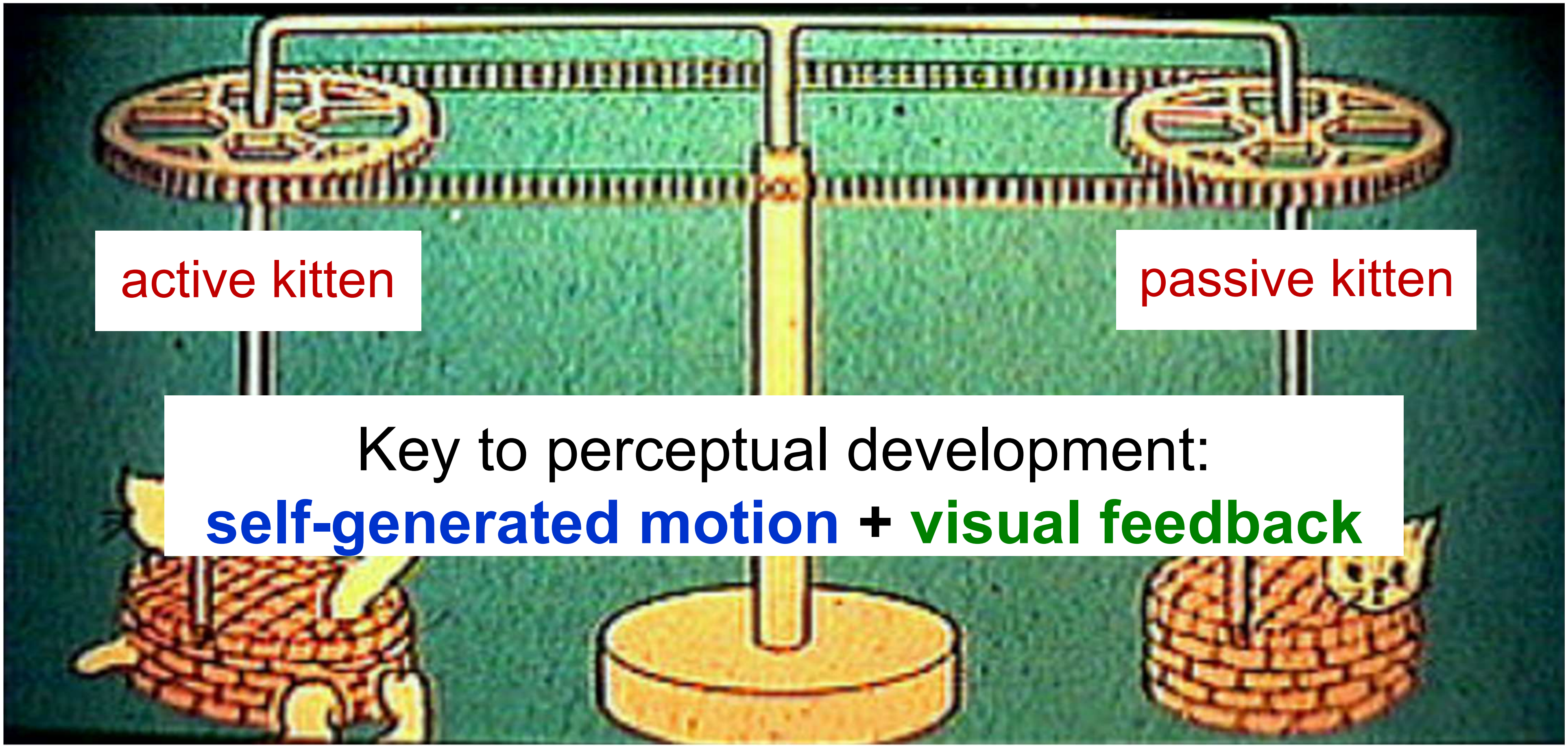


This talk

Learning where to look and listen

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

+

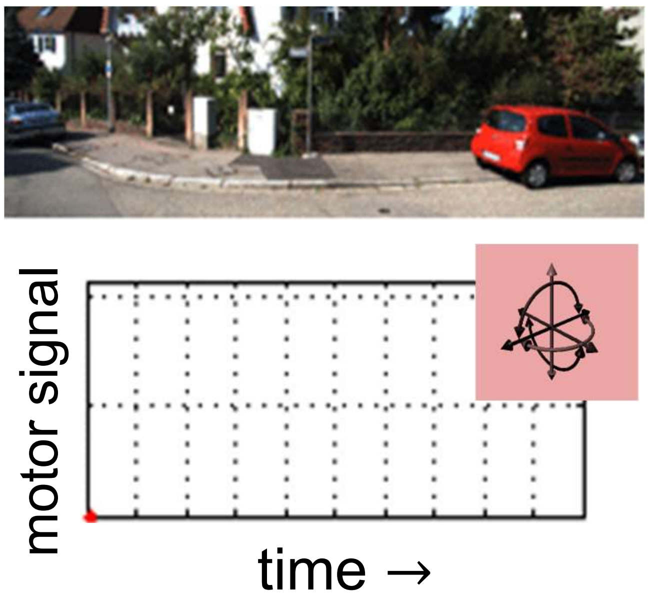
Unlabeled video

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

Approach: Ego-motion equivariance

Training data

Unlabeled video +
motor signals



Learn

Equivariant embedding

organized by ego-motions

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

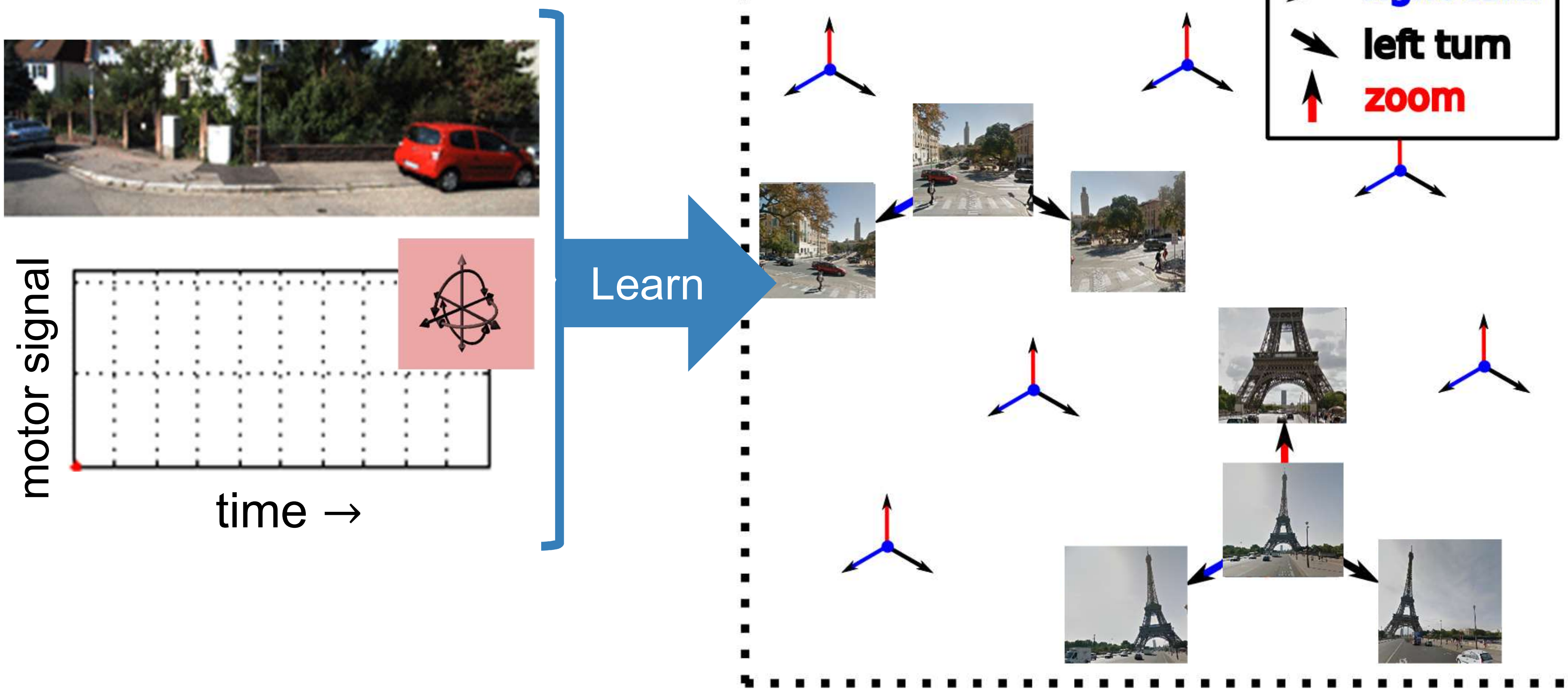
Pairs of frames related by similar ego-motion should be related by same feature transformation

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

Approach: Ego-motion equivariance

Training data
Unlabeled video +
motor signals

Equivariant embedding
organized by ego-motions



[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

Example result: Recognition

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

Geiger et al, IJRR ’13

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

Apse

Window se

ardhouse

30% accuracy increase
when labeled data scarce

, CVPR ’10

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**

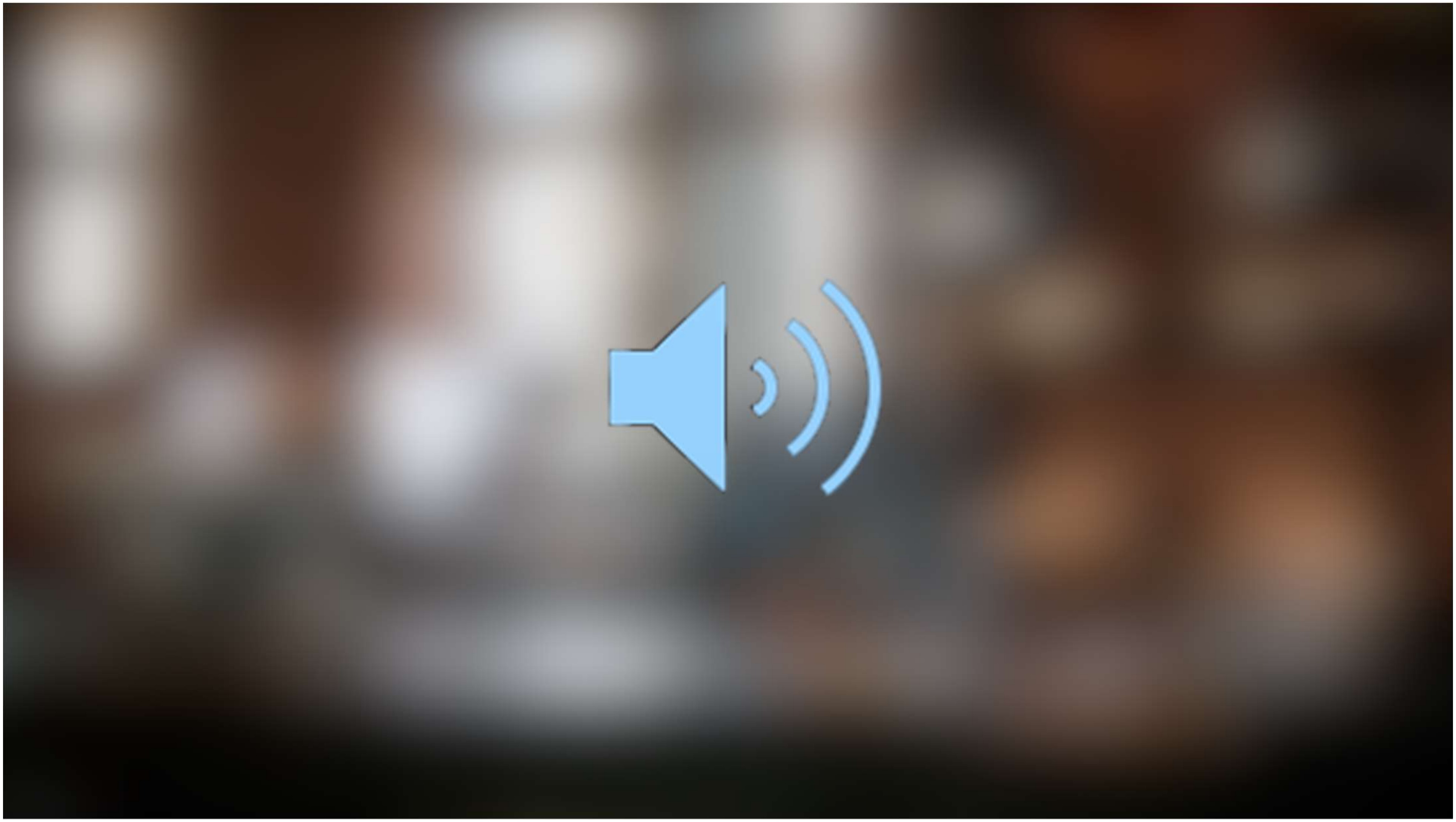
[Jiang & Grauman, CVPR 2017]

This talk

Learning where to look and listen

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

Listening to learn



Listening to learn

Listening to learn

woof

meow

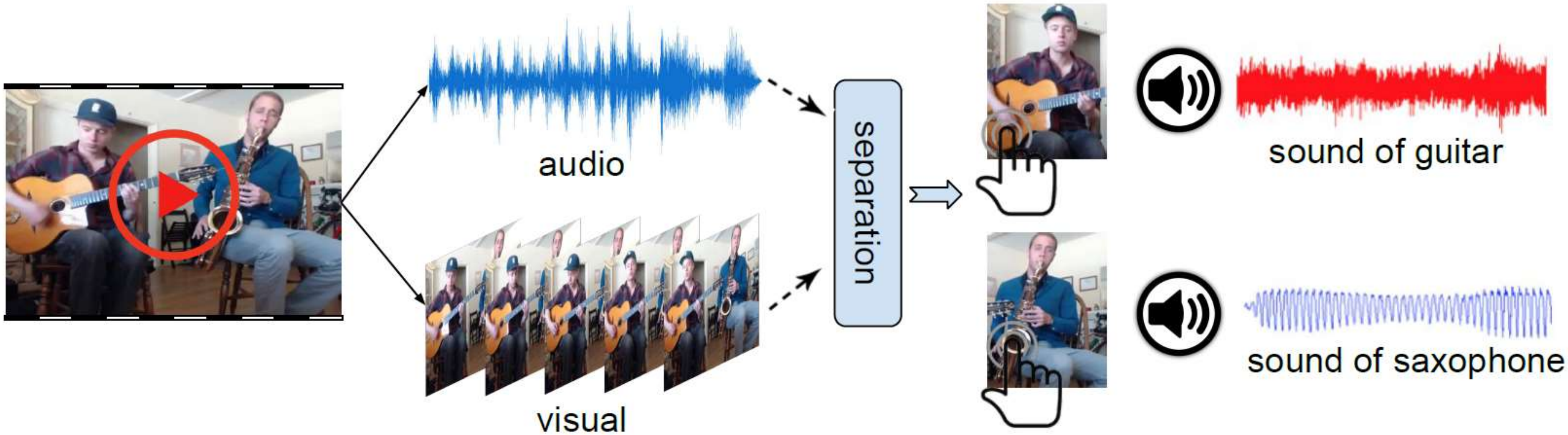
ring

clatter

Goal: a repertoire of objects and their sounds

Challenge: a single audio channel mixes sounds of multiple objects

Visually-guided audio source separation



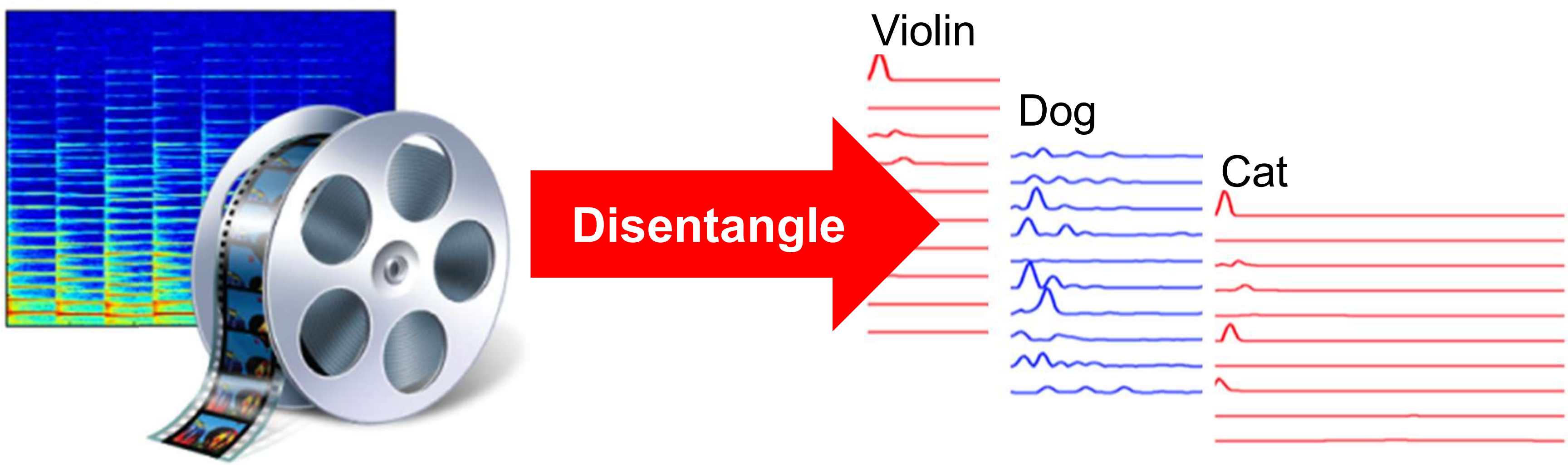
Traditional approach:

- 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

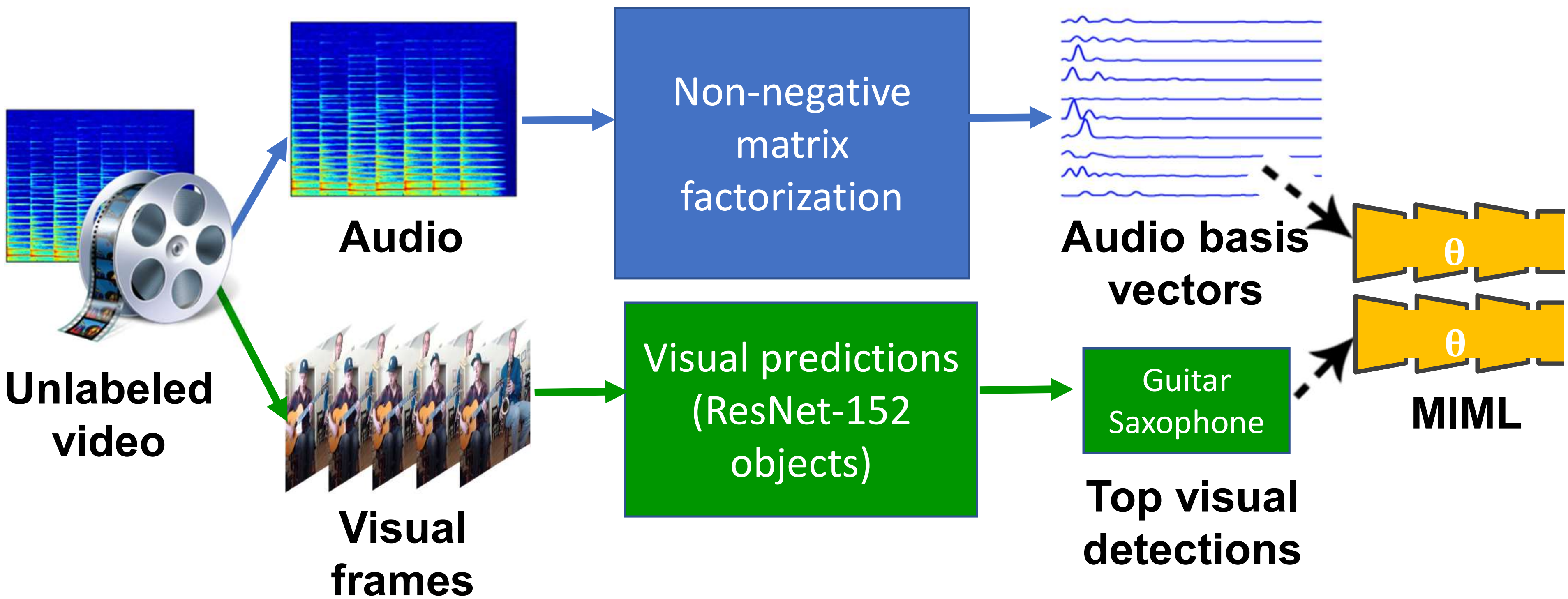


Unlabeled video **Object sound models**

[Gao, Feris, & Grauman, arXiv 2018]

Our approach: learning

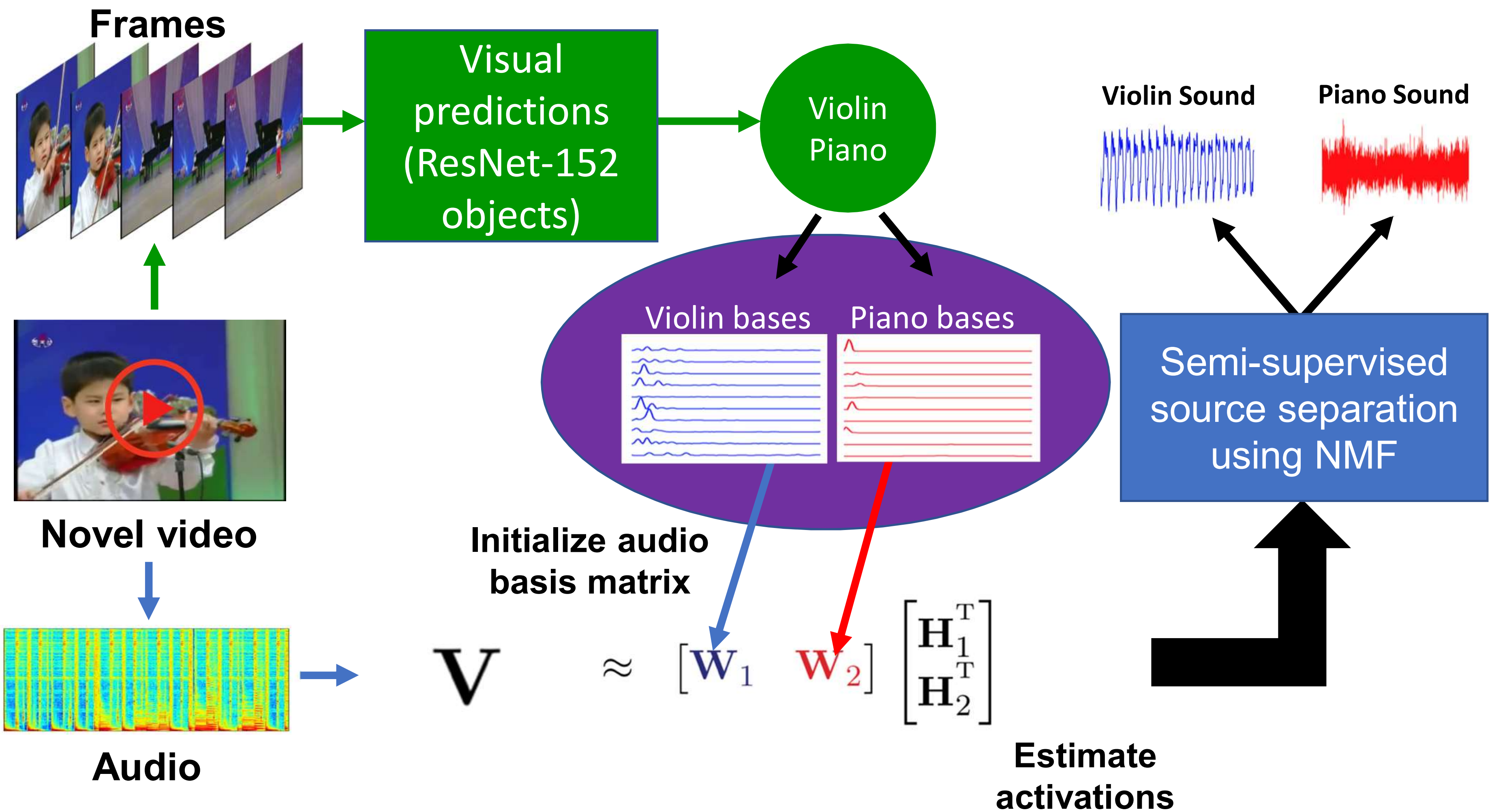
Deep multi-instance multi-label learning (MIML) to disentangle which visual objects make which sounds



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: learning to separate sounds

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

original video
(before separation)

visual predictions:
acoustic guitar & harmonica

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: learning to separate sounds

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

original video
(before separation)

visual predictions:
dog & violin

[Gao, Feris, & Grauman, arXiv 2018]

Results: learning to separate sounds

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

Failure case

original video
(before separation)

visual predictions:
accordion & acoustic guitar

Failure cases

[Gao, Feris, & Grauman, arXiv 2018]

Results: Separating object sounds

	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)

Lock et al. Annals Stats 2013; Spiertz et al. ICDAE 2009; Kidron et al. CVPR 2006; Pu et al. ICASSP 2017

This talk

Learning where to look and listen

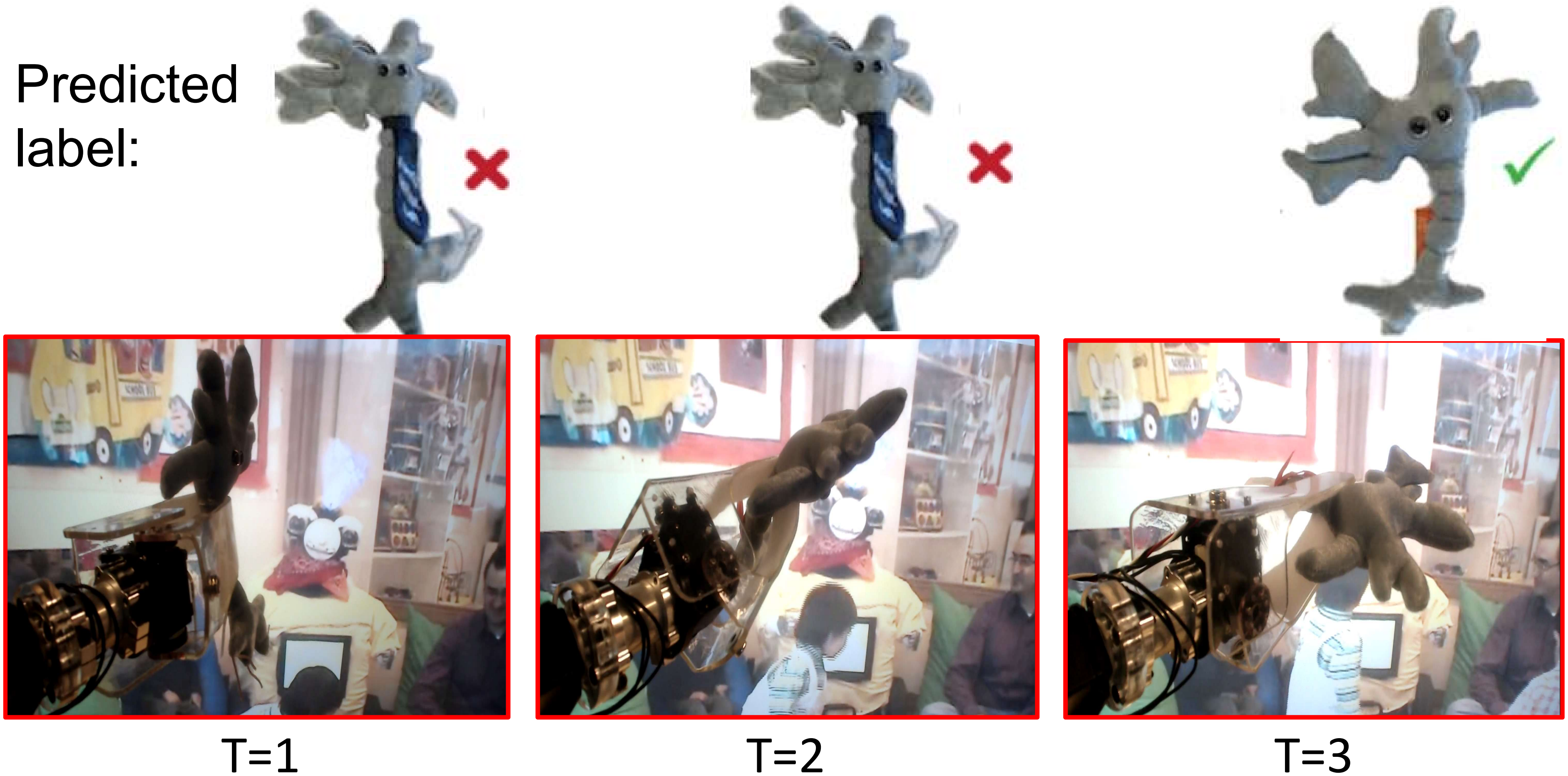
1. Learning from unlabeled video and multiple sensory modalities
2. Learning policies for how to move for recognition and exploration
   a) Active perception
   b) 360 video

Agents that move intelligently to see

Time to revisit active perception 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, …

End-to-end active recognition



[Jayaraman and Grauman, ECCV 2016, PAMI 2018]

Goal: Learn to "look around"

vs.

recognition

reconnaissance

search and rescue

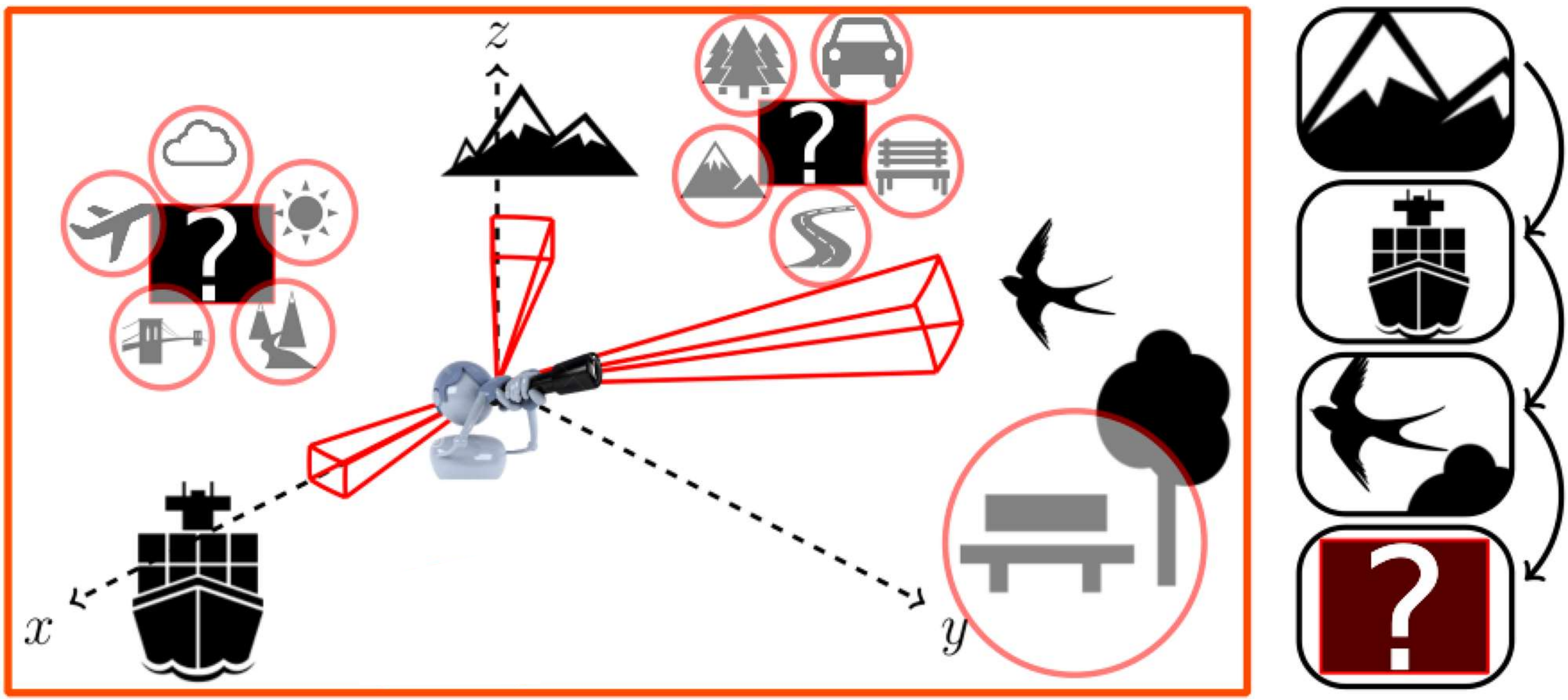
task predefined

task unfolds dynamically

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

Key idea: Active observation completion

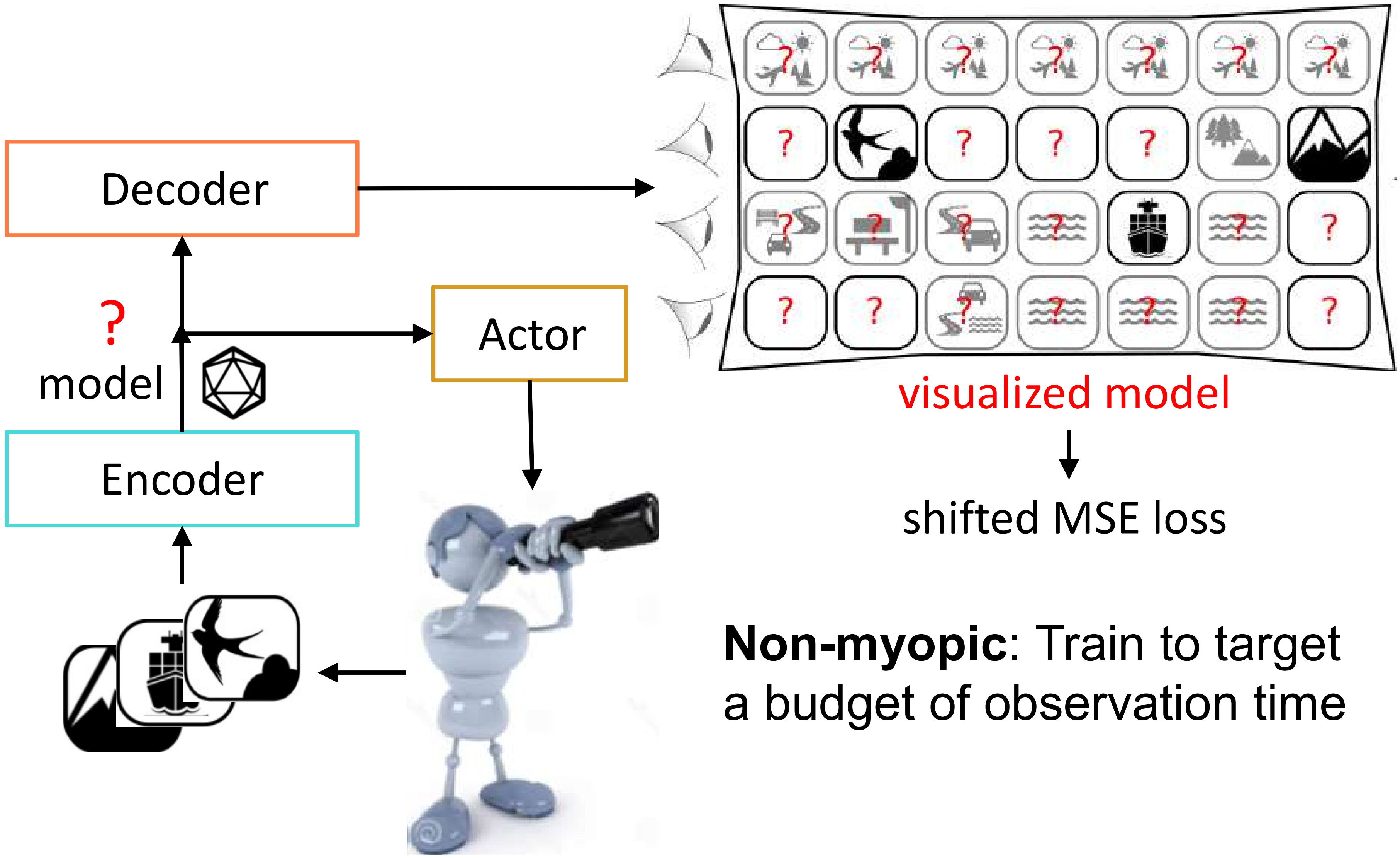
Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment



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

Jayaraman and Grauman, CVPR 2018

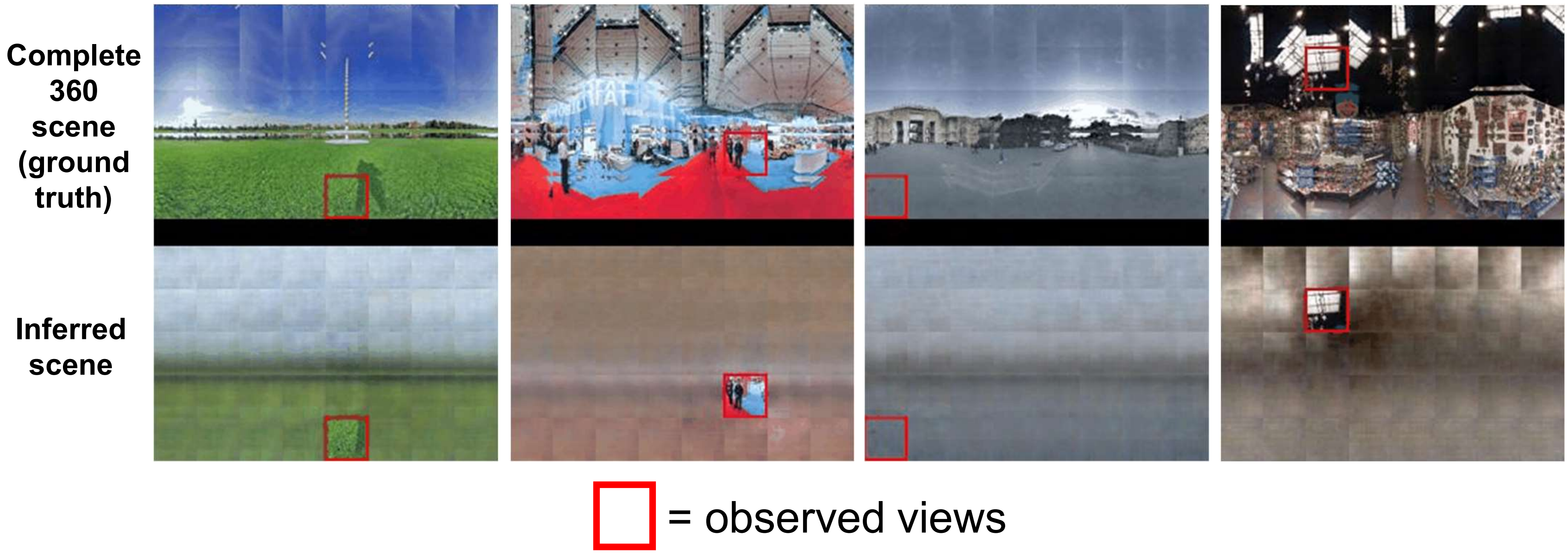
Approach: Active observation completion



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

Jayaraman and Grauman, CVPR 2018

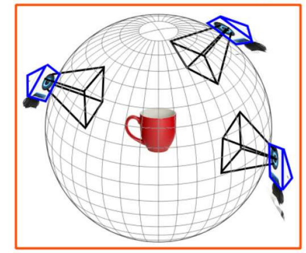
Active "look around" visualization



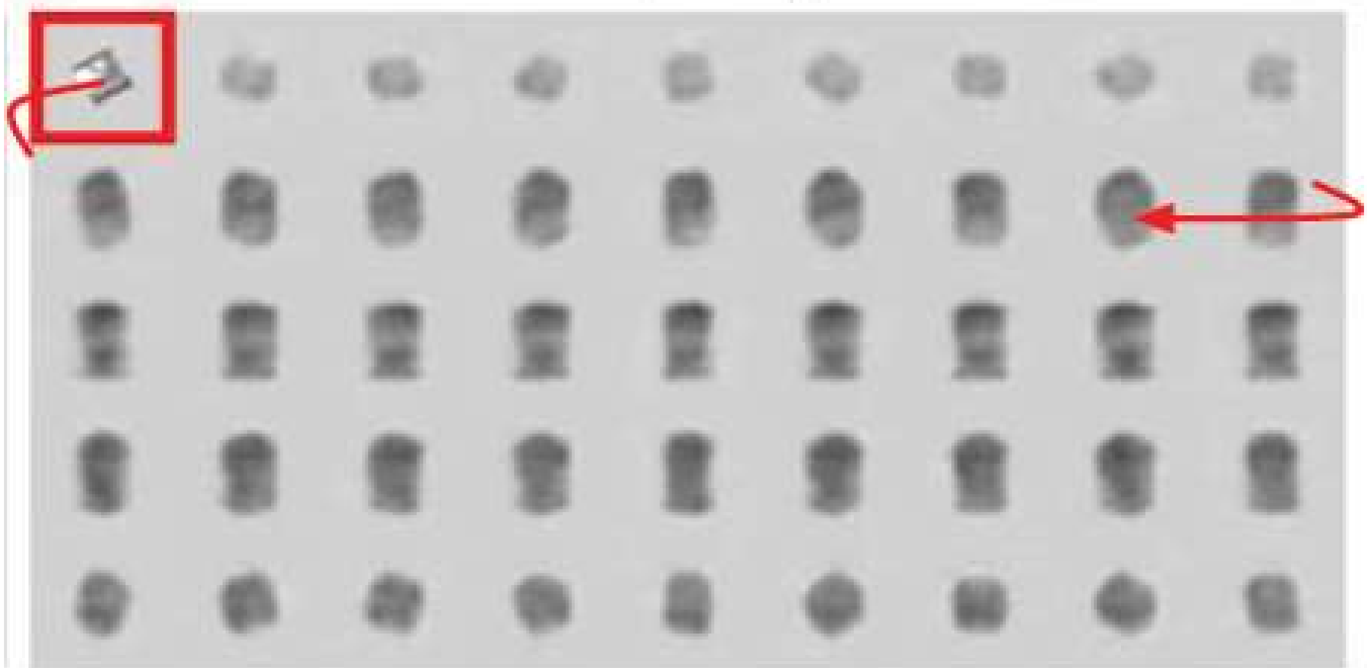
Agent's mental model for 360 scene evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

Active “look around” visualization



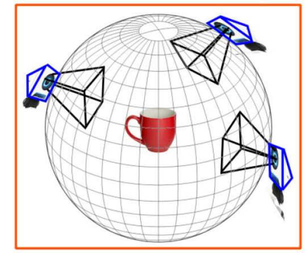
$t = 1$

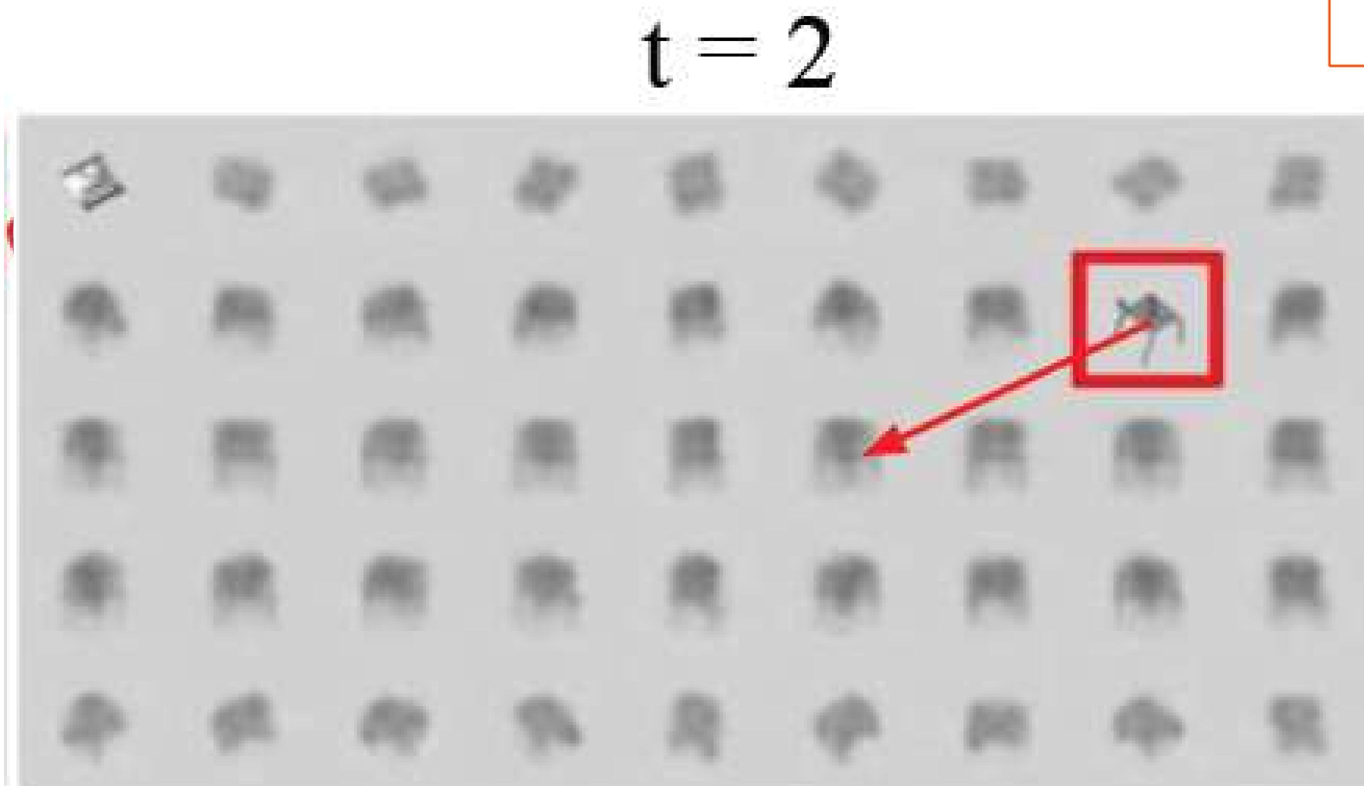


Agent’s mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

Active "look around" visualization

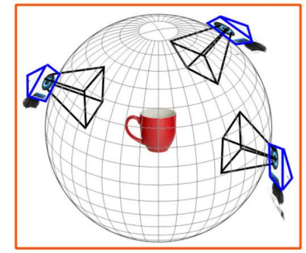




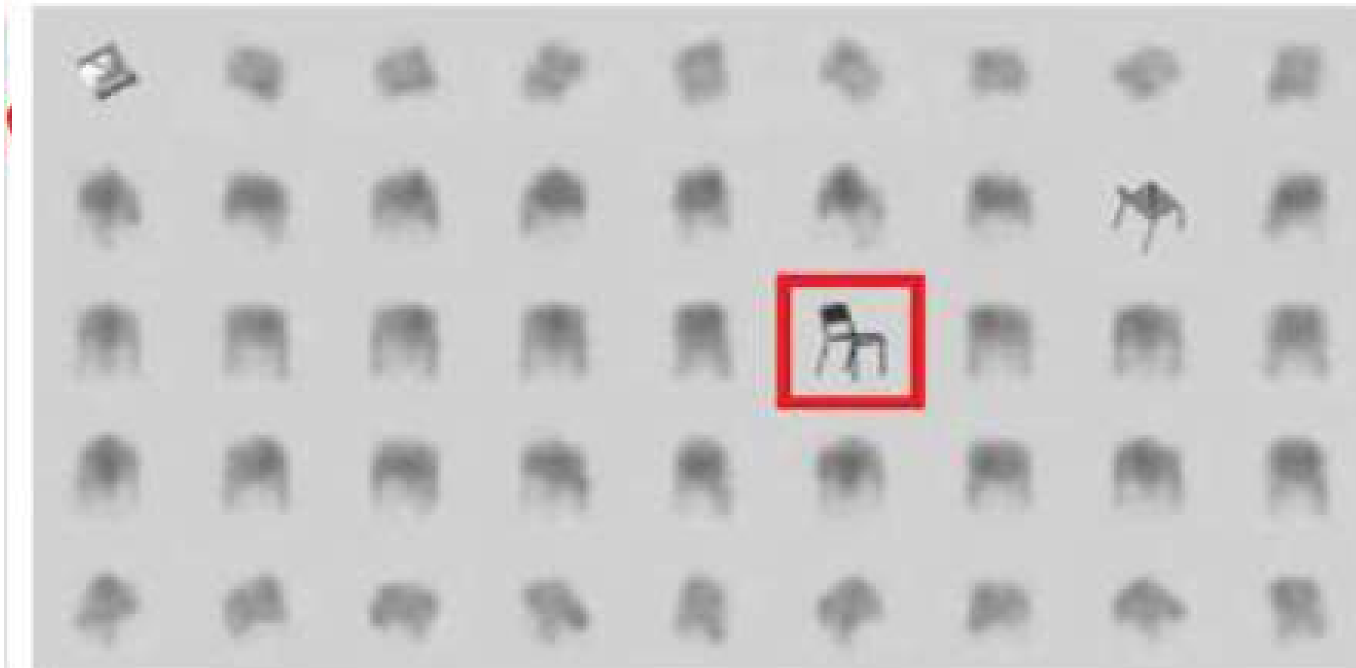
Agent's mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

Active "look around" visualization



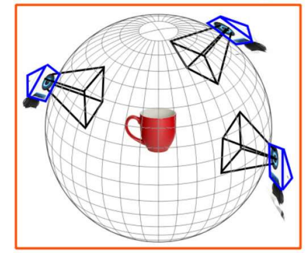
$t = 3$

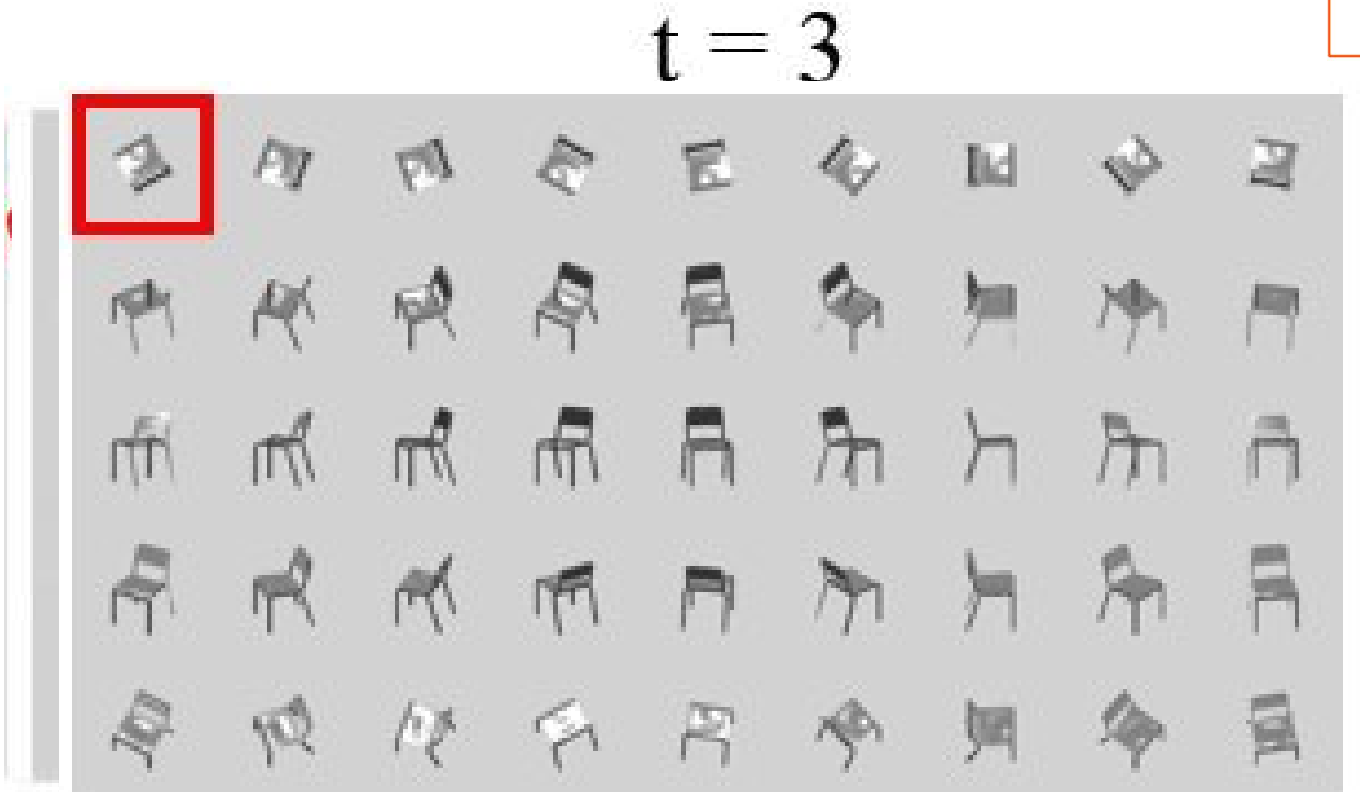


Agent's mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

Active "look around" visualization

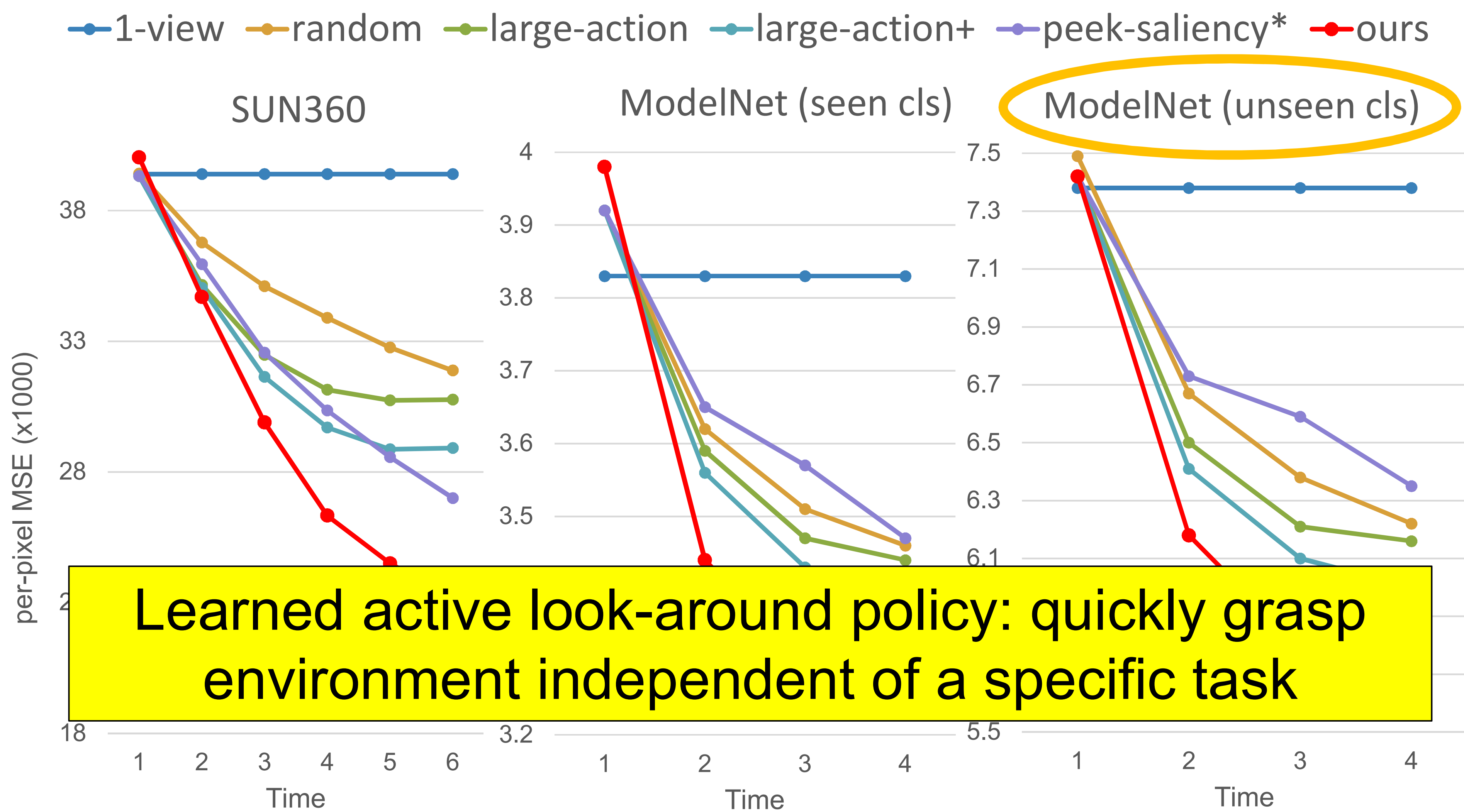




Agent's mental model for 3D object evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018

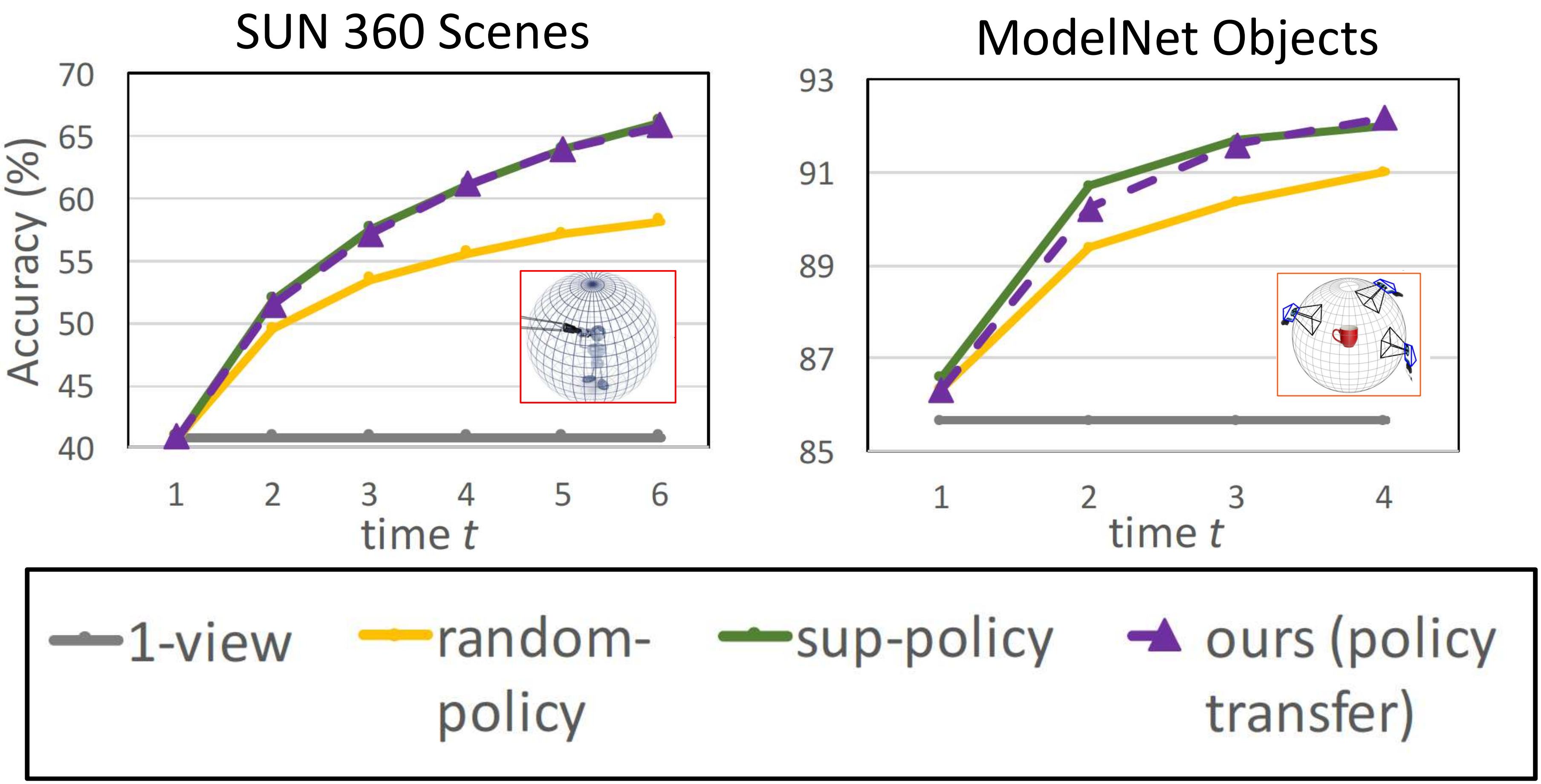
Active "look around" results



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

Jayaraman and Grauman, CVPR 2018

Egomotion policy transfer



Unsupervised exploratory policy approaches supervised task-specific policy accuracy!

Jayaraman and Grauman, CVPR 2018

This talk

Learning where to look and listen

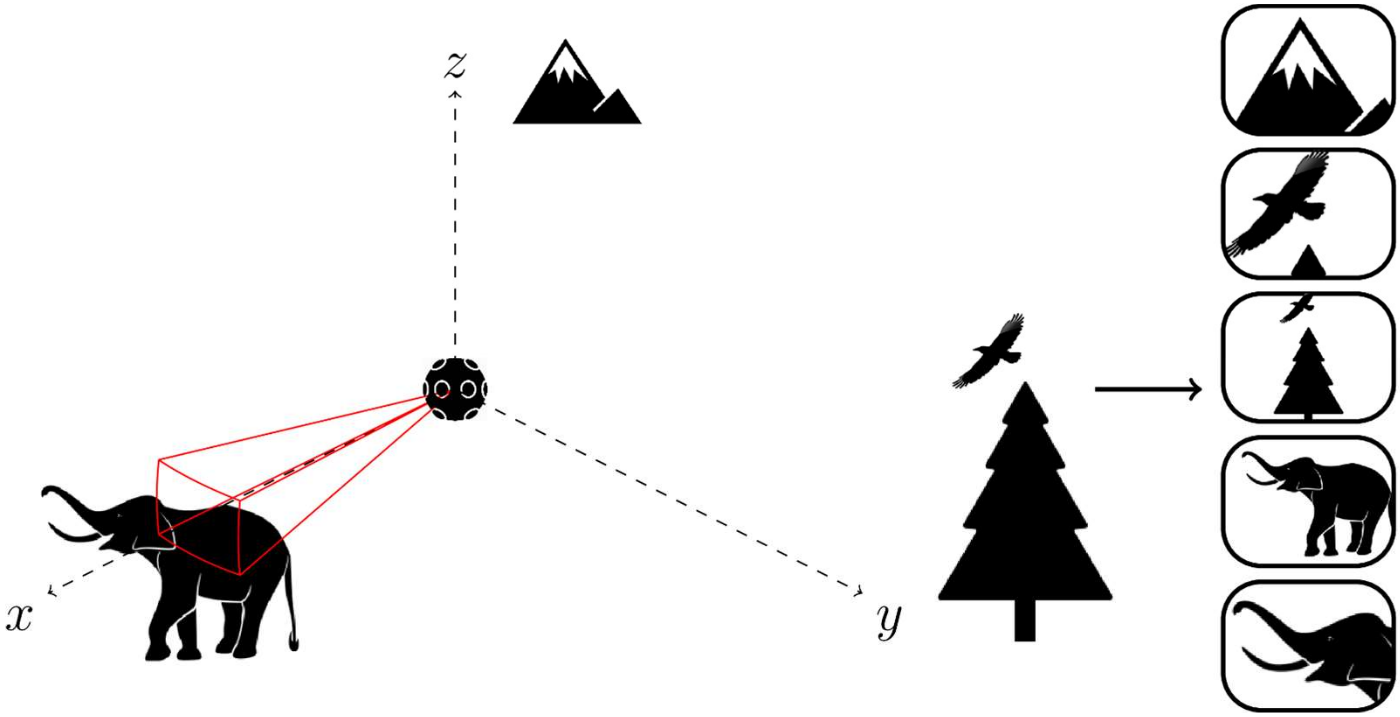
1. Learning from unlabeled video and multiple sensory modalities
2. Learning policies for how to move for recognition and exploration
   a) Active perception
   b) 360 video

Challenge of viewing 360° videos

Control by mouse

Where to look when?

Pano2Vid: automatic videography



Definition

Input: 360° video

Output: "natural-looking" normal FOV video

Task: control virtual camera direction and FOV

[Su et al. ACCV 2016, CVPR 2017]

Our approach – AutoCam

Learn videography tendencies from unlabeled Web videos

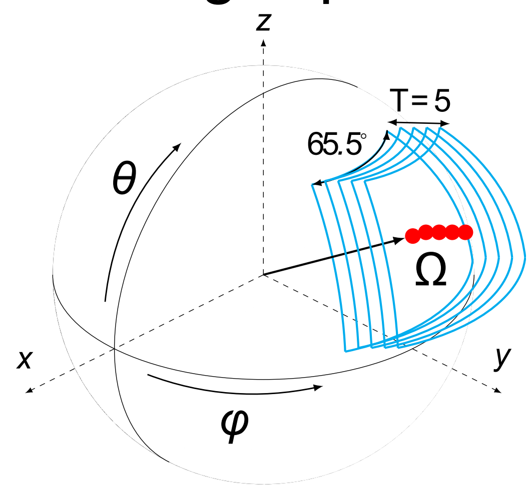
- Diverse capture-worthy content
- Proper composition

Human-captured NFOV videos (“HumanCam”)

Unlabeled video

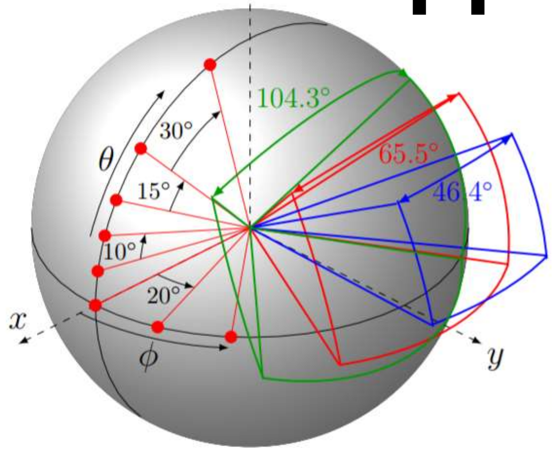
How close?

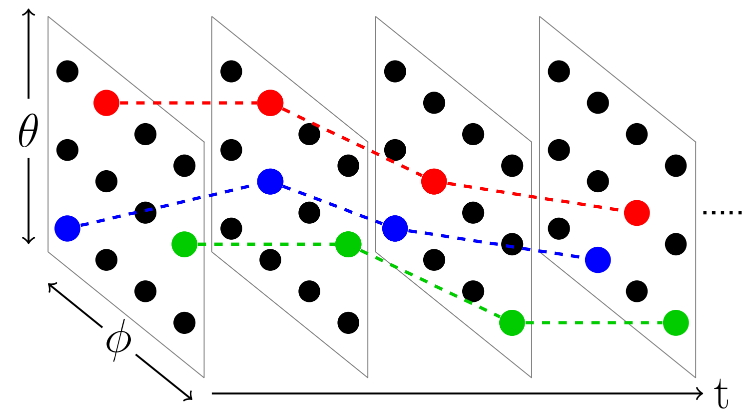
ST-glimpses



[Su et al. ACCV 2016, CVPR 2017]

Our approach – AutoCam

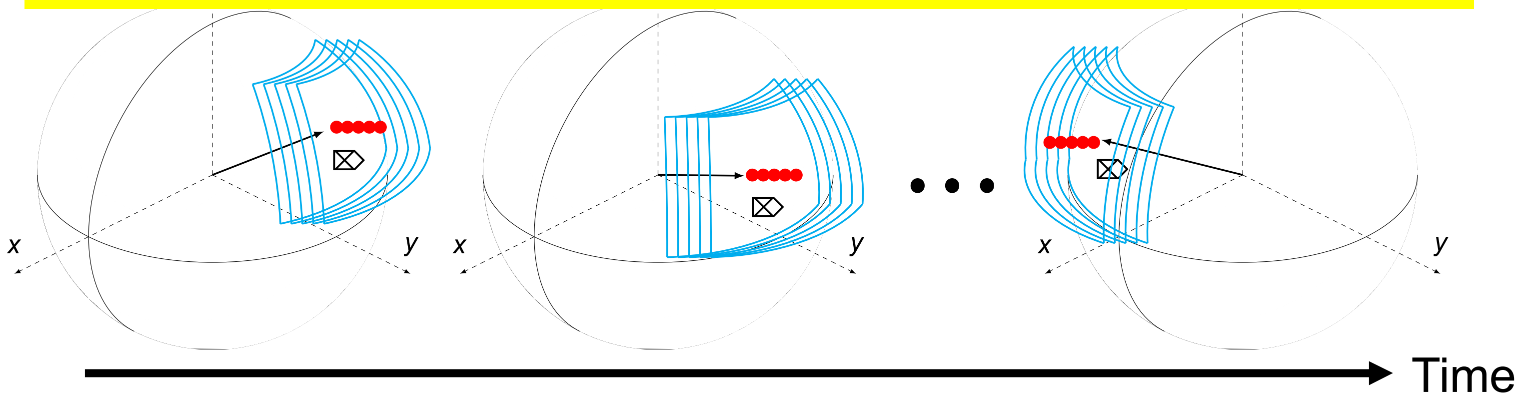




Densely sample and score glimpses

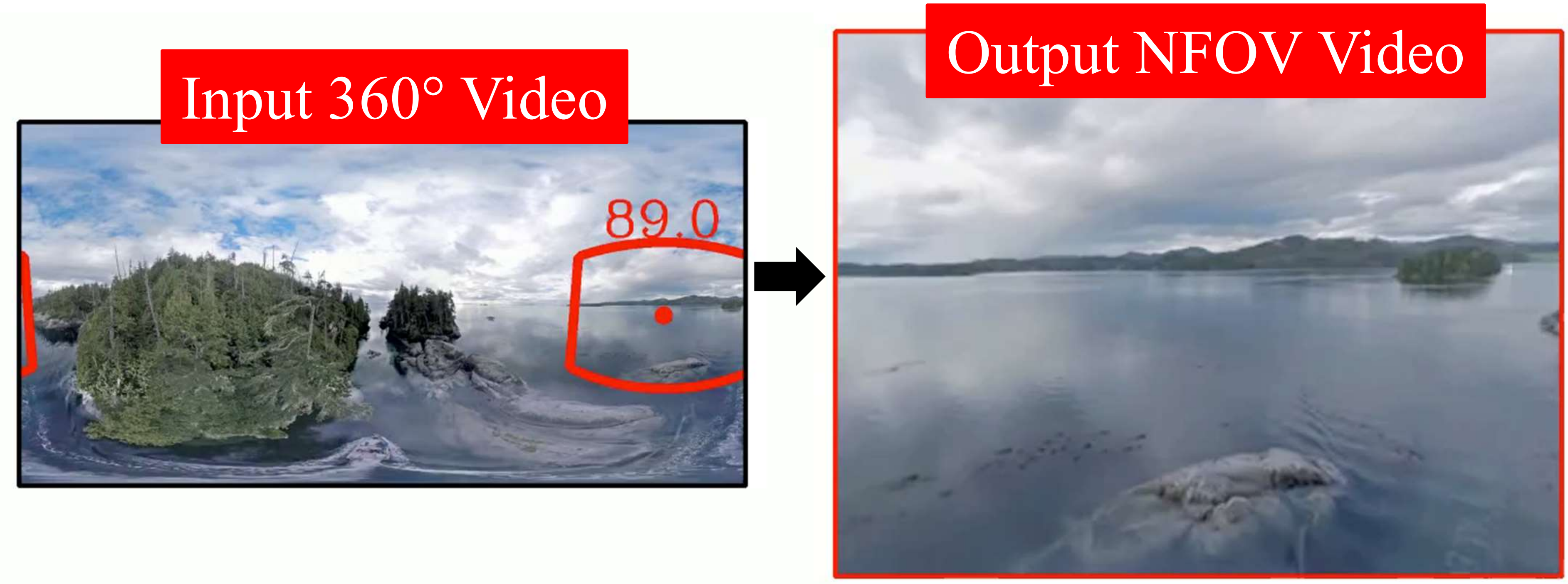
Pose selection as shortest path(s) problem

Optimize for *multiple diverse* hypotheses



Output smooth view path maximizing capture-worthiness

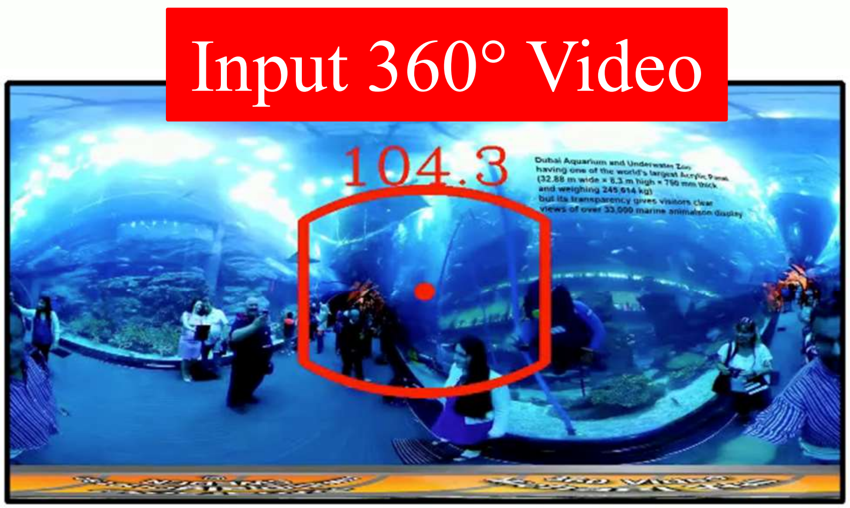
AutoCam results



Automatically select FOV and viewing direction

[Su & Grauman, CVPR 2017]

AutoCam results



Automatically select FOV and viewing direction

[Su & Grauman, CVPR 2017]

AutoCam results:
Multiple diverse hypotheses

Input Video & Cam. Trajectory

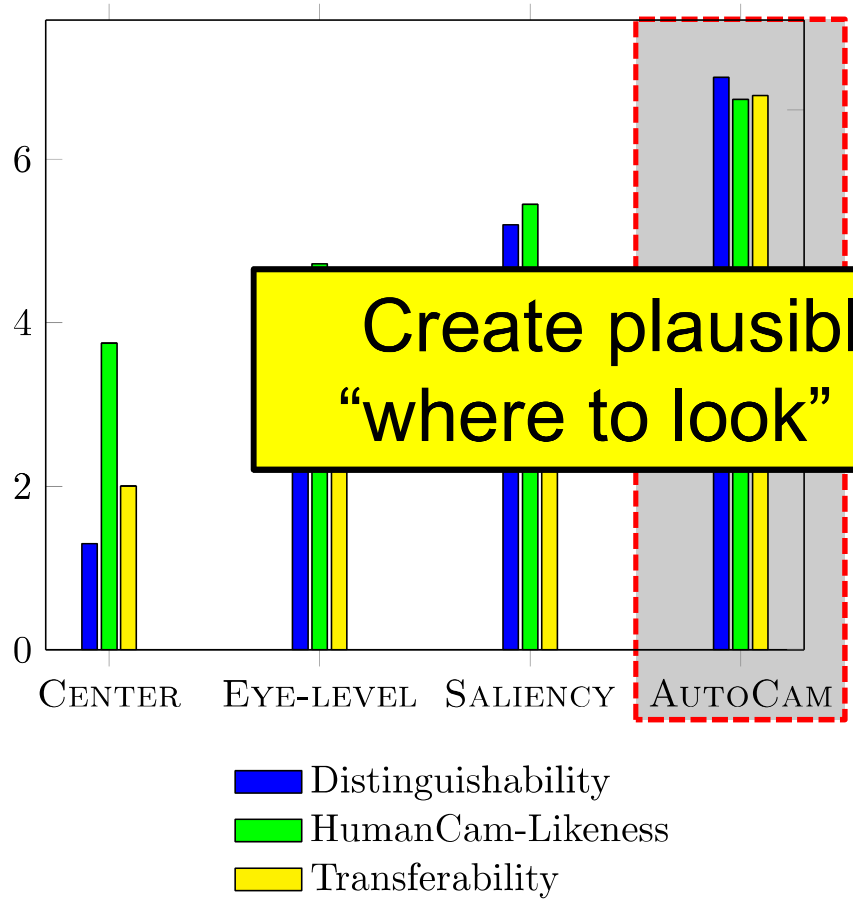
Output Videos

Hypothesis 1

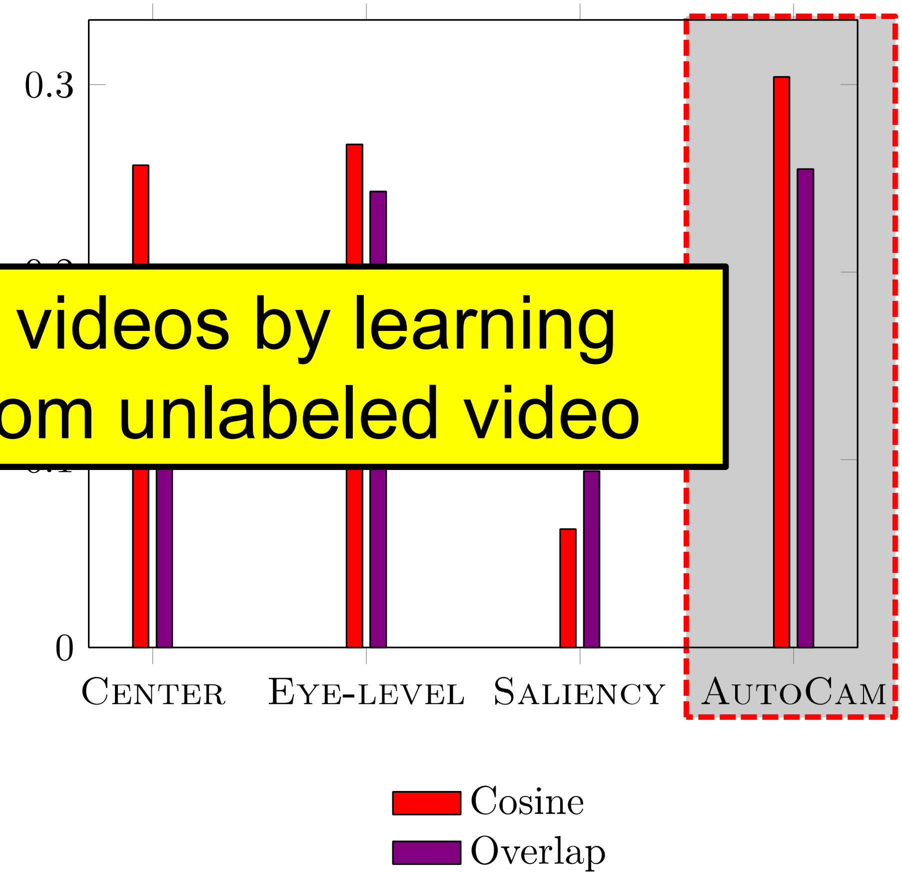
Hypothesis 2

AutoCam results

Similarity to user-uploaded standard web videos



Similarity to human-selected camera trajectories

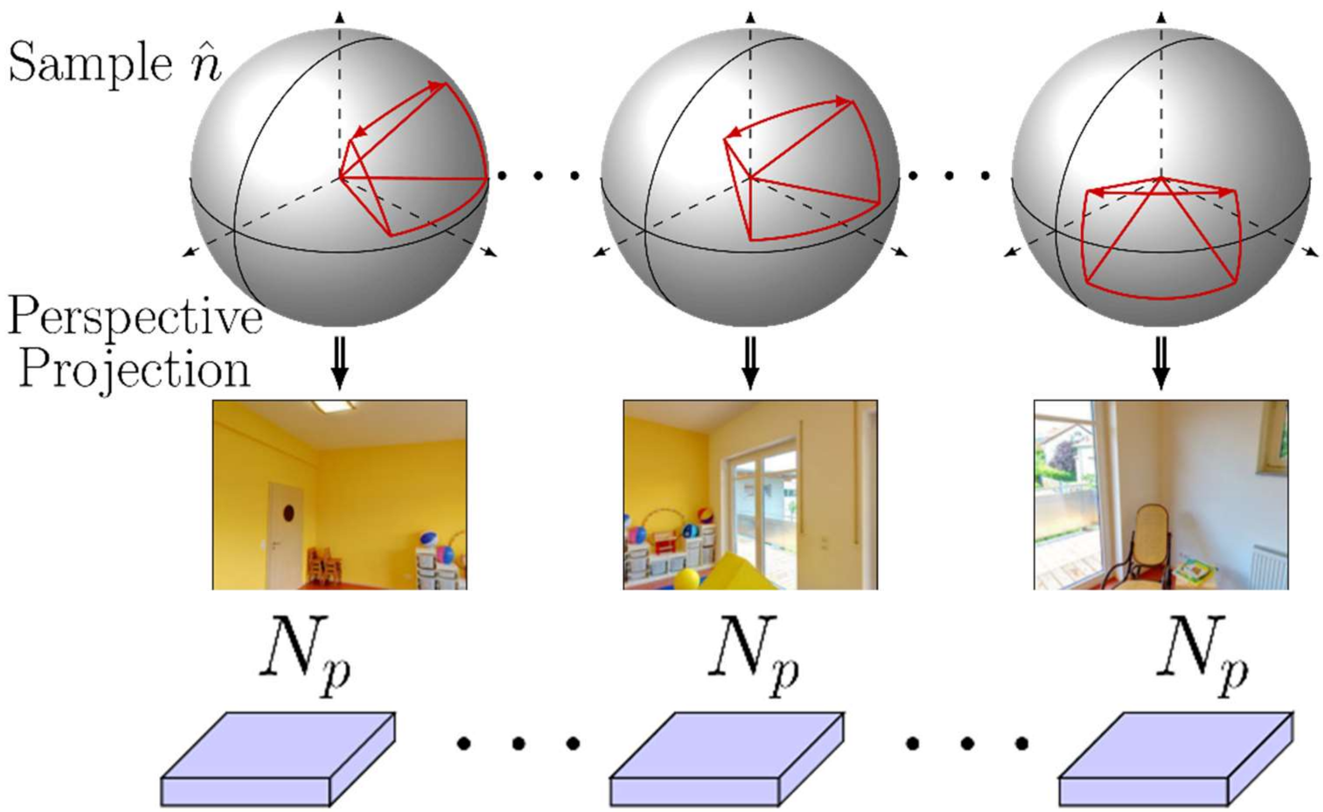


Create plausible videos by learning "where to look" from unlabeled video

[Su et al. ACCV 2016, CVPR 2017]

Applying CNNs to 360 imagery

Existing strategy 1: Reproject



Accurate but slow

Applying CNNs to 360 imagery

Existing strategy 2: Equirect

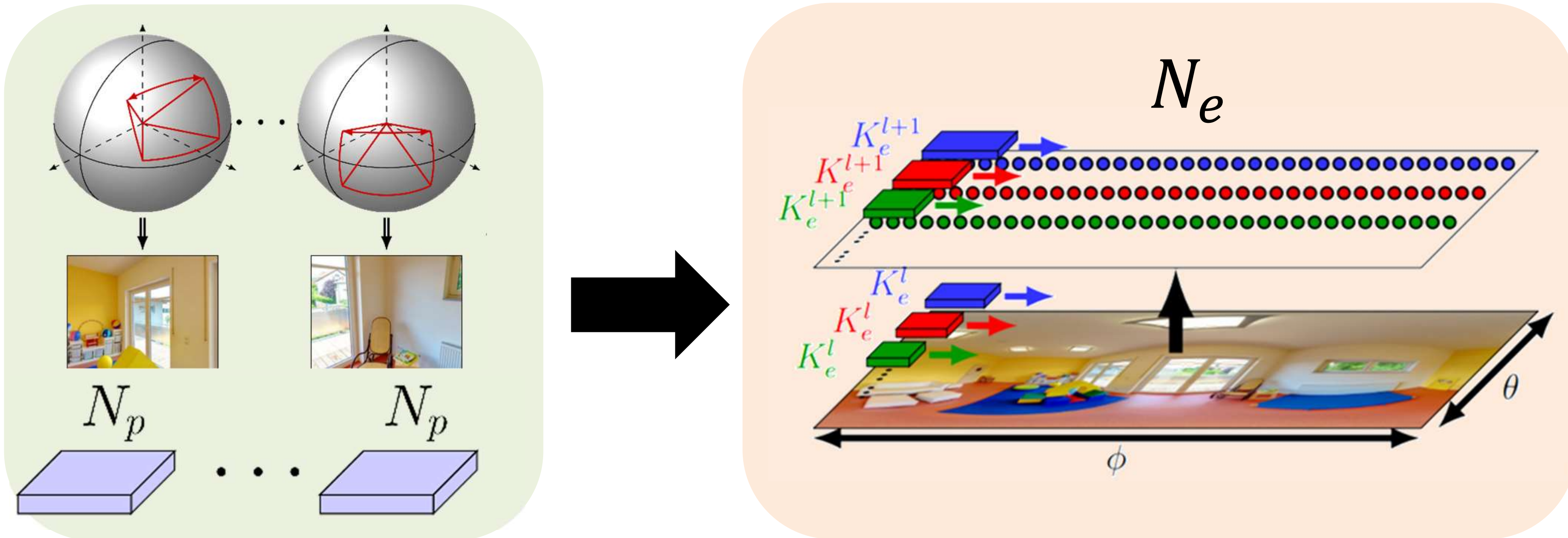
equirectangular projection of spherical 360 image

vs.

standard FOV "flat" image

Fast but inaccurate

Our idea: Learning spherical convolution



$$\min_{N_e} |N_e(I_e)[x,y] - N_p(I_s)[\theta,\phi]|^2$$

- Fast and accurate
- Enable off-the-shelf “flat” CNNs for 360

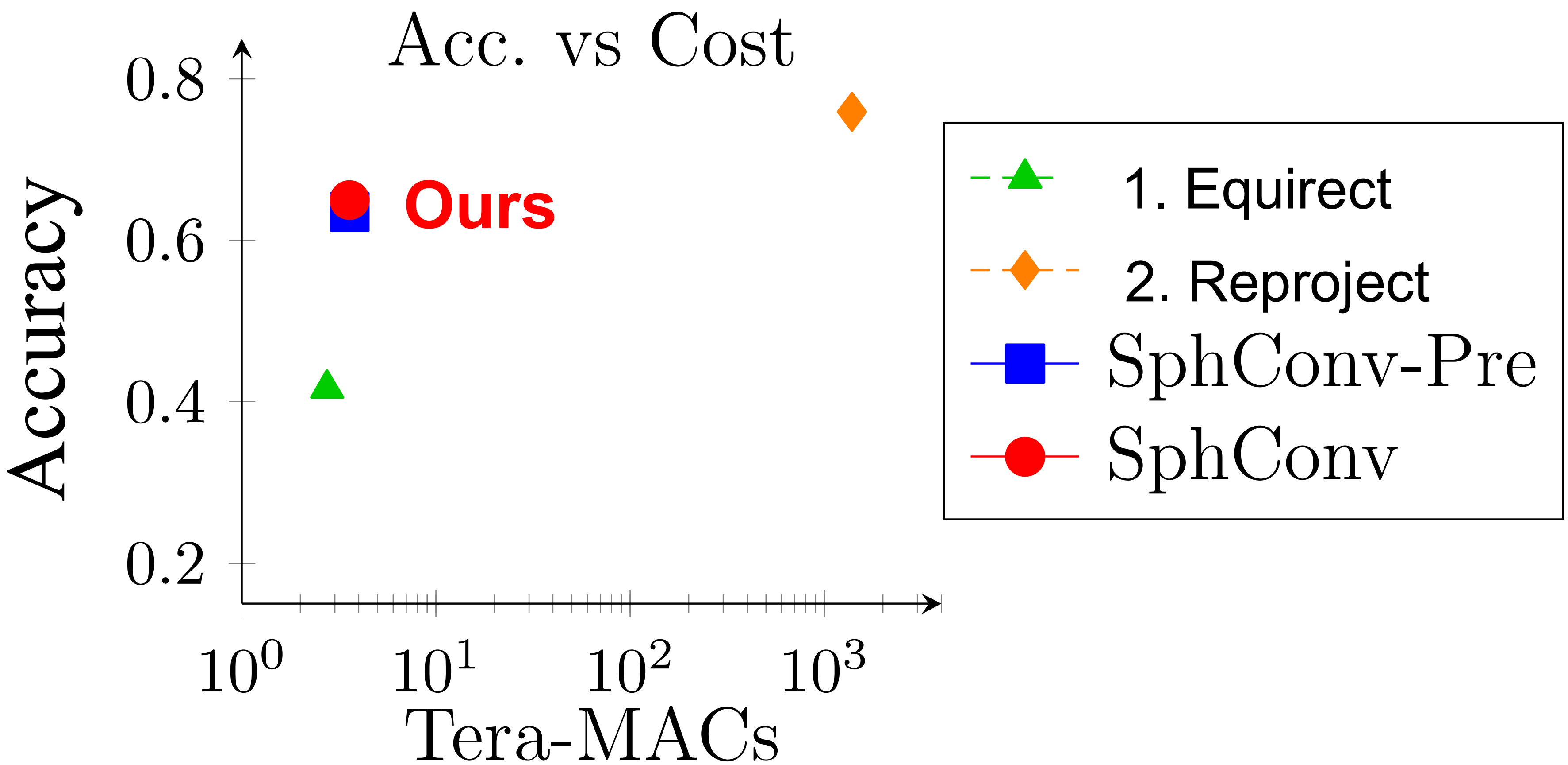
[Su & Grauman, NIPS 2017]

Spherical convolution for object detection

Spherical convolution + Faster RCNN *[Ren et al. 2016]*

[Su & Grauman, NIPS 2017]

Results: Spherical convolution

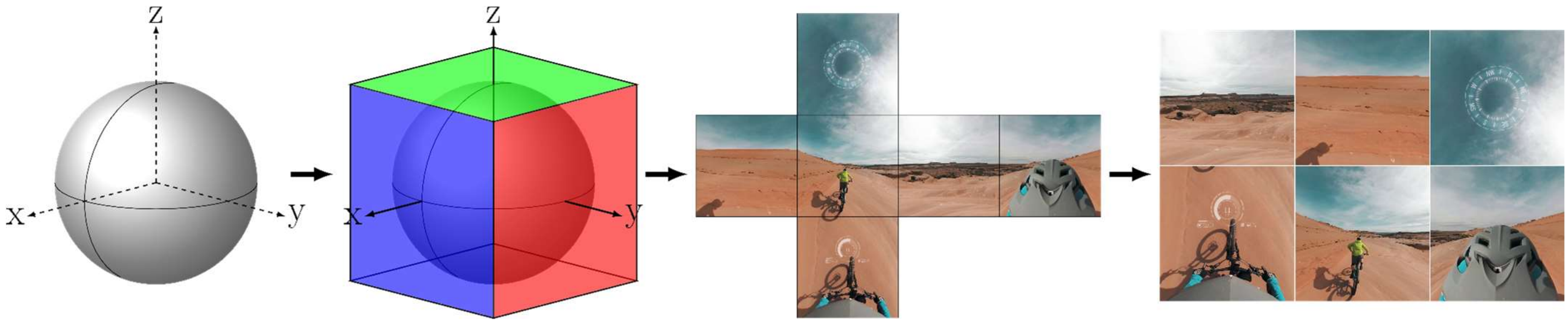


Fast and (quite) accurate

[Su & Grauman, NIPS 2017]

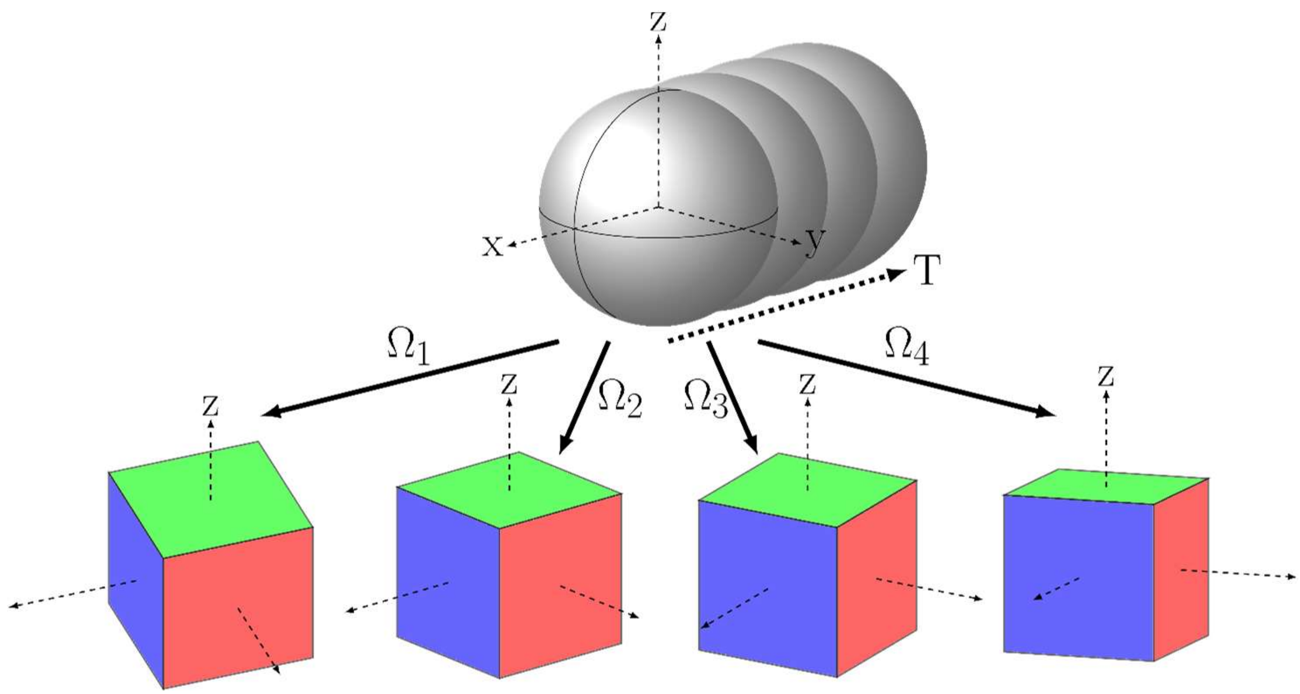
How to compress a 360 video?

Cubemap projection



From spherical to 6 perspective images

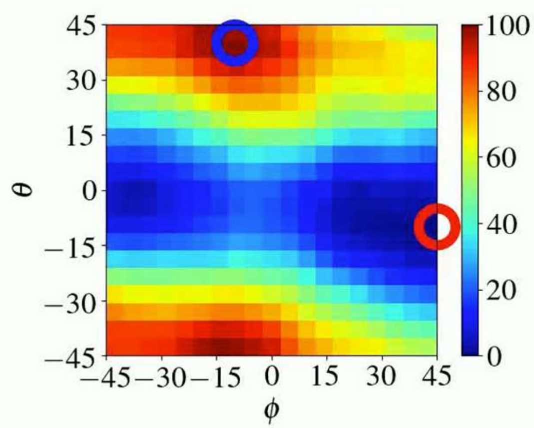
Problem: 360 video isomers



- Video content is invariant to projection axis
- However, the encoded bit-streams are not

[Su & Grauman, CVPR 2018]

Problem: 360 video isomers



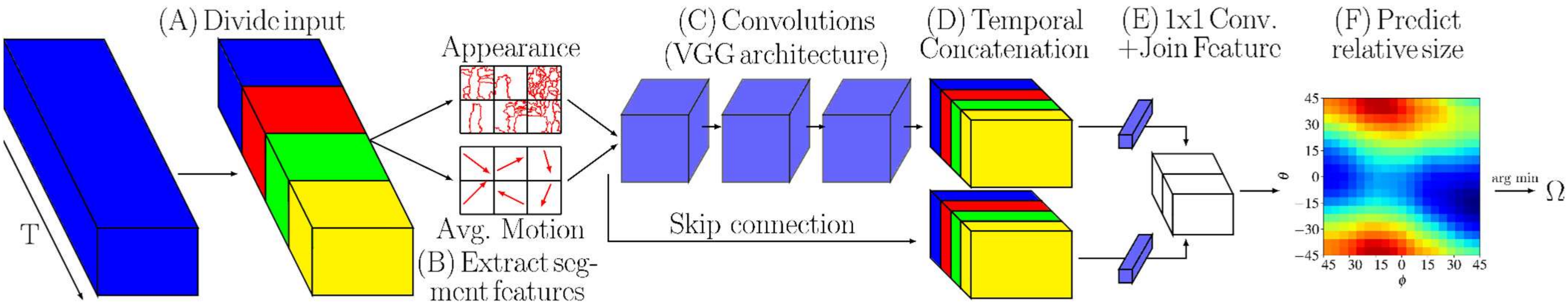
Video size vs.
cube rotation angle

- Video content is invariant to projection axis
- However, the encoded bit-streams are not

[Su & Grauman, CVPR 2018]

Our idea: Compressible 360 isomers

Given video, predict most compressible isomer (angle)



	H264	HEVC	VP9
RANDOM	50.75	51.62	51.20
CENTER	74.35	63.34	72.92
OURS	82.10	79.10	81.55

% size reduction achieved

[Su & Grauman, CVPR 2018]

Summary

Ruohan Gao

Yu-Chuan Su

Dinesh Jayaraman

- Visual learning benefits from
  - context of action and multiple senses
  - continuous unsupervised observations

- Key ideas:
  - Learning from egomotion and sound with unlabeled video
  - Look-around motion policies to quickly explore new environments
  - Spherical convolution and compression

Papers/code/videos

Embodied vision and multi-modal:

- **Learning to Separate Object Sounds by Watching Unlabeled Video**. R. Gao, R. Feris, and K. Grauman. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sept 2018. (Oral) [pdf] [videos]
- **End-to-end Policy Learning for Active Visual Categorization**. D. Jayaraman and K. Grauman. To appear, Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2018. [pdf]
- **Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks**. D. Jayaraman and K. Grauman. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, June 2018. [pdf] [animations]
- **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. [pdf] [preprint] [project page, pretrained models]

360 images/video:

- **Learning Spherical Convolution for Fast Features from 360° Imagery**. Y-C. Su and K. Grauman. In Advances in Neural Information Processing (NIPS), Long Beach, CA, Dec 2017. [pdf]
- **Learning Compressible 360 Video Isomers**. Y-C. Su and K. Grauman. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, June 2018. [pdf]
- **Making 360 Video Watchable in 2D: Learning Videography for Click Free Viewing.** Y-C. Su and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, July 2017. (Spotlight)
- Code and models: `http://www.cs.utexas.edu/~grauman/research/pubs.html`