Anticipating the Unseen and Unheard for Embodied Perception

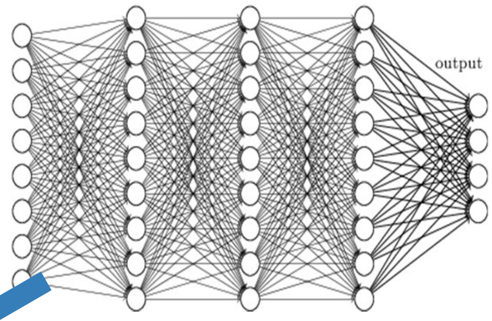
Kristen Grauman

University of Texas at Austin

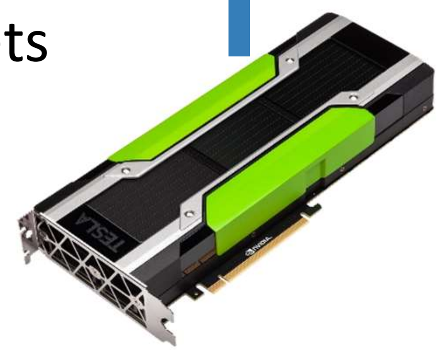
Facebook AI Research

Visual recognition: significant recent progress

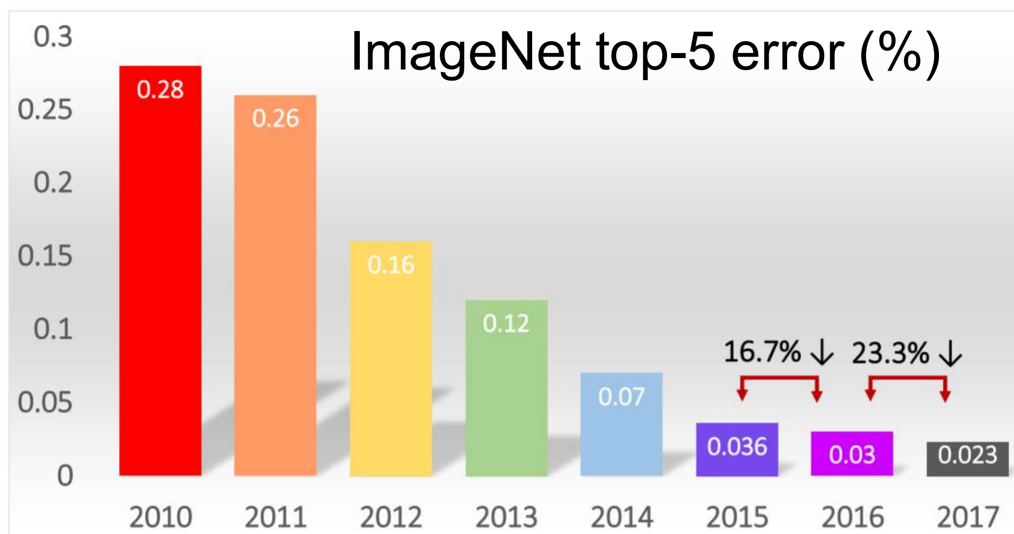
Big labeled datasets



Deep learning



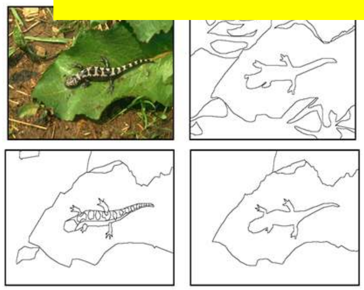
GPU technology



Kristen Grauman

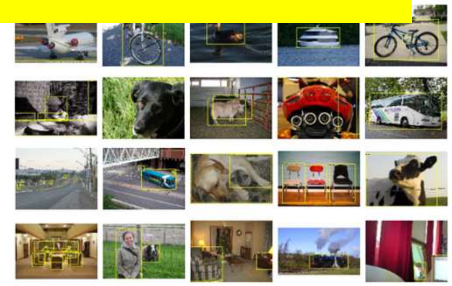
The Web photo perceptual experience

A “disembodied” well-curated moment in time

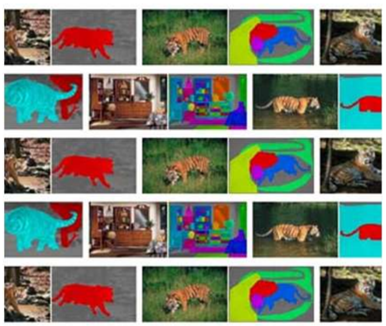


BSD (2001)

Caltech 101 (2004), Caltech 256 (2006)



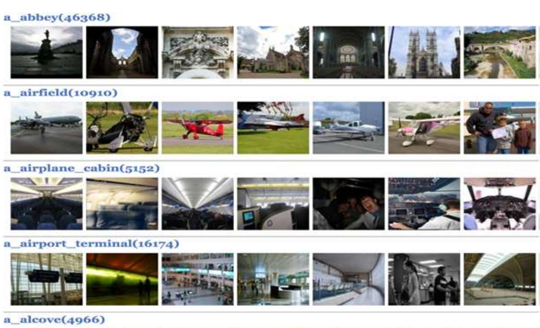
PASCAL (2007-12)



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

Kristen Grauman

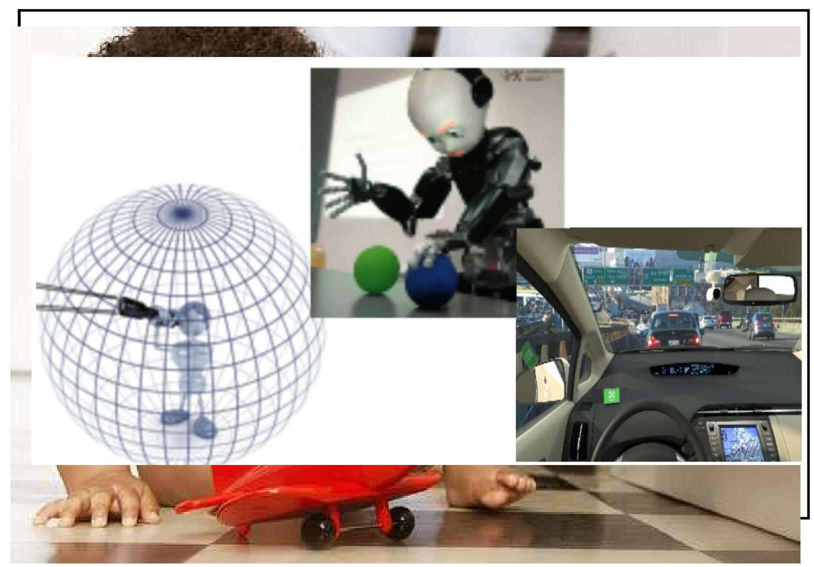
Big picture goal: Embodied perception

Status quo:

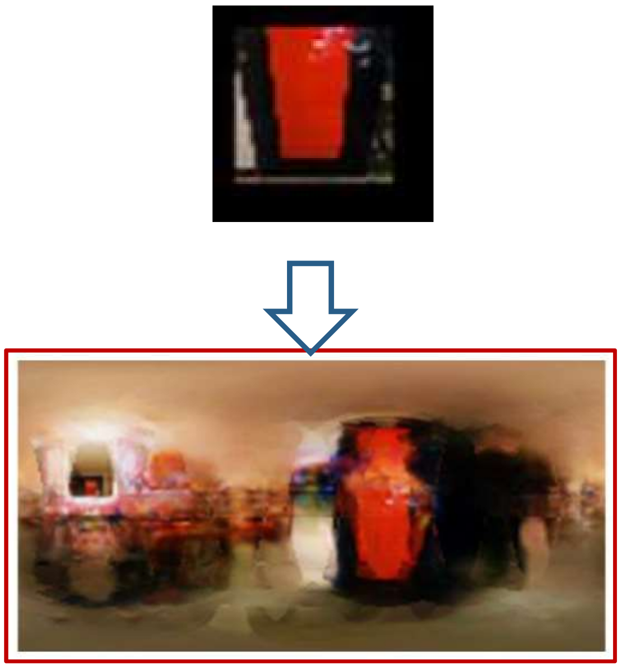
Learning and inference with "disembodied" snapshots.

On the horizon:

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

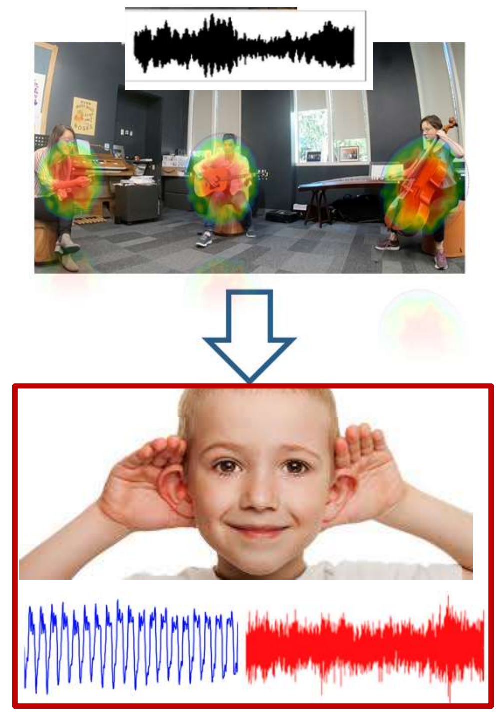


Anticipating the unseen and unheard



Look-around policies

Affordance learning

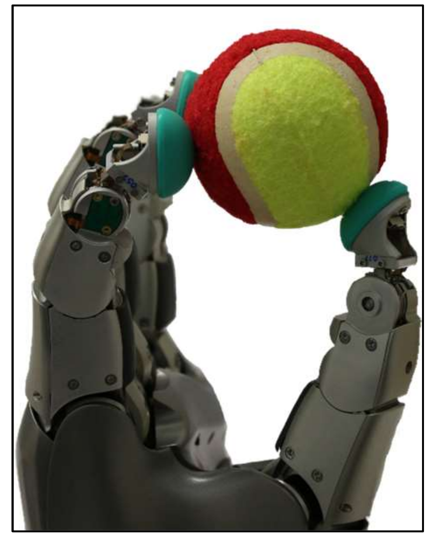


Audio-visual learning

Towards embodied perception

Active perception

From learning *representations* to learning *policies*

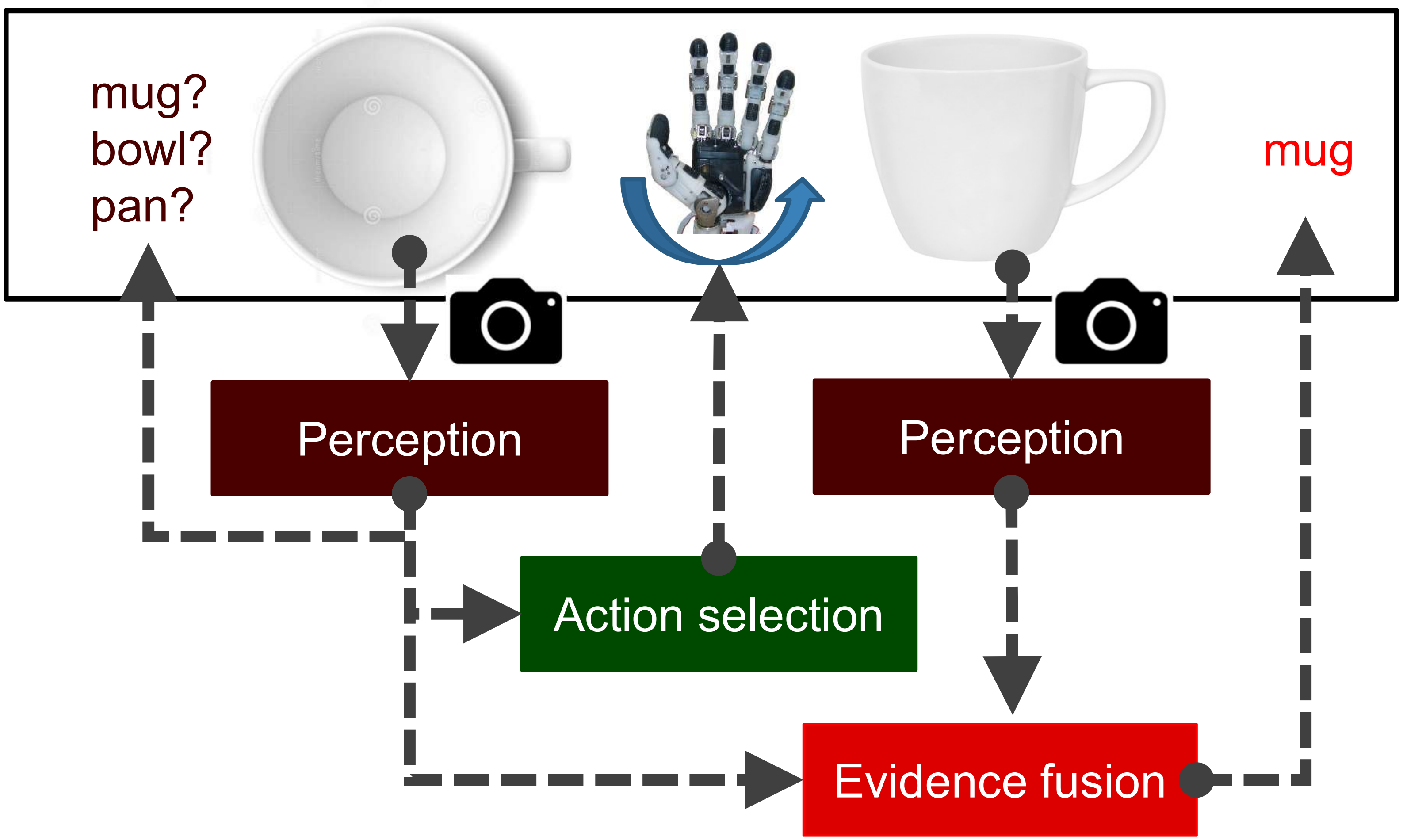


Bajcsy 1985, Aloimonos 1988, Ballard 1991, Wilkes 1992, Dickinson 1997, Schiele & Crowley 1998, Tsotsos 2001, Denzler 2002, Soatto 2009, Ramanathan 2011, Borotschnig 2011, …

Kristen Grauman

End-to-end active recognition

Main idea: Deep reinforcement learning approach that anticipates visual changes as a function of egomotion



Kristen Grauman

Jayaraman and Grauman, ECCV 2016, PAMI 2018

End-to-end active recognition

T=1 T=2 T=3

[Jayaraman and Grauman, ECCV 2016, PAMI 2018]

Kristen Grauman

Goal: Learn to "look around"

recognition

vs.

reconnaissance

search and rescue

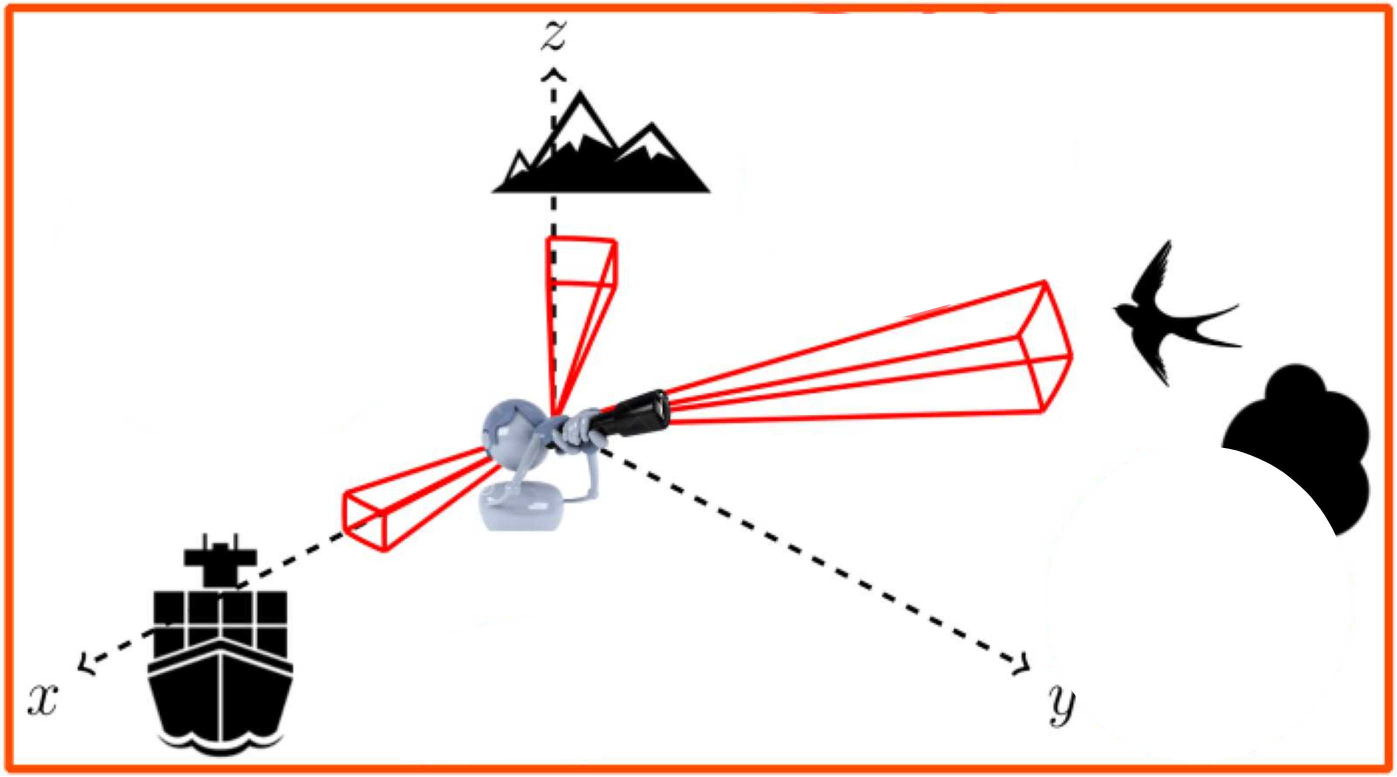
task predefined

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



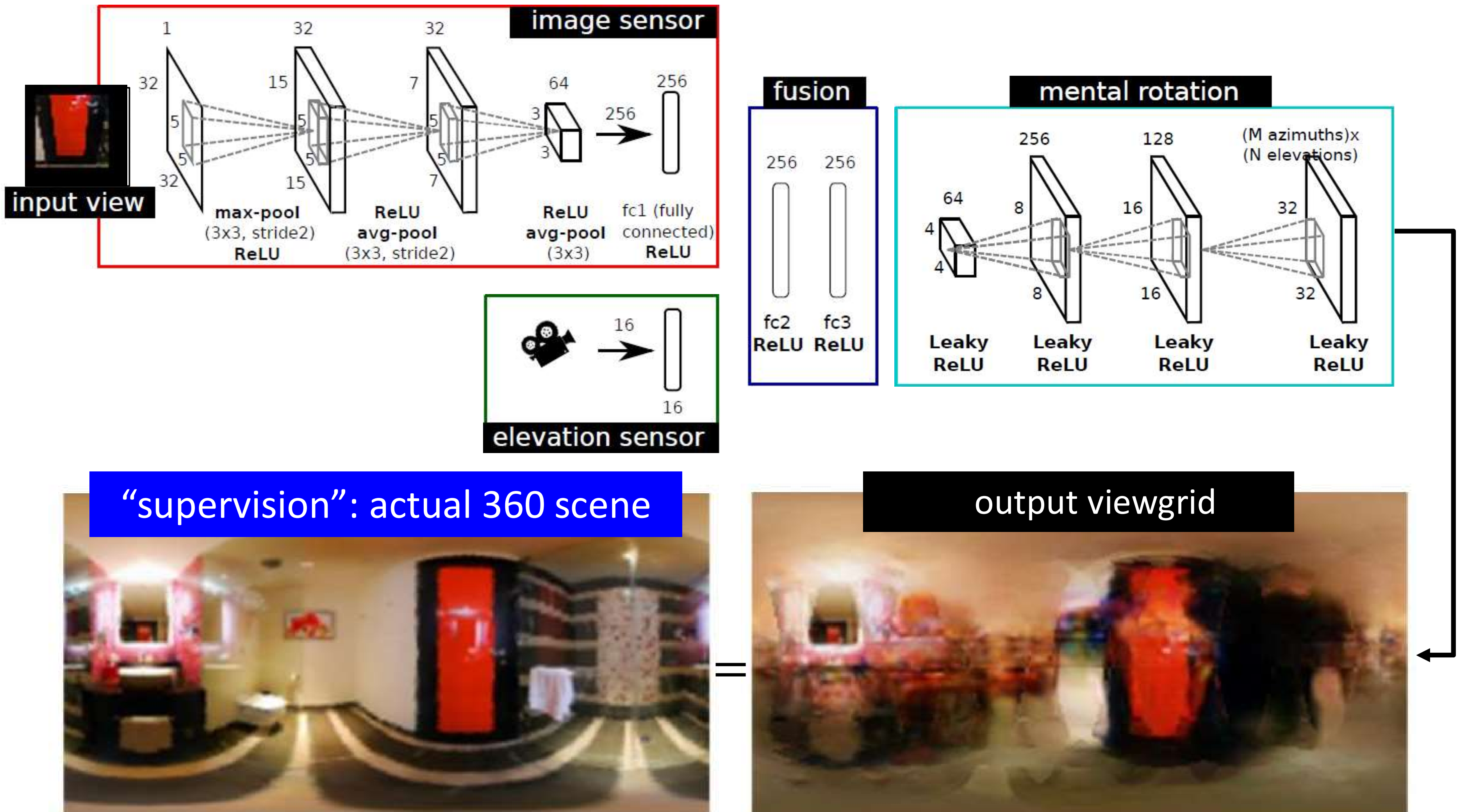
Agent must choose where to look *before* looking there.

Kristen Grauman

Jayaraman and Grauman, CVPR 2018

Completing unseen views

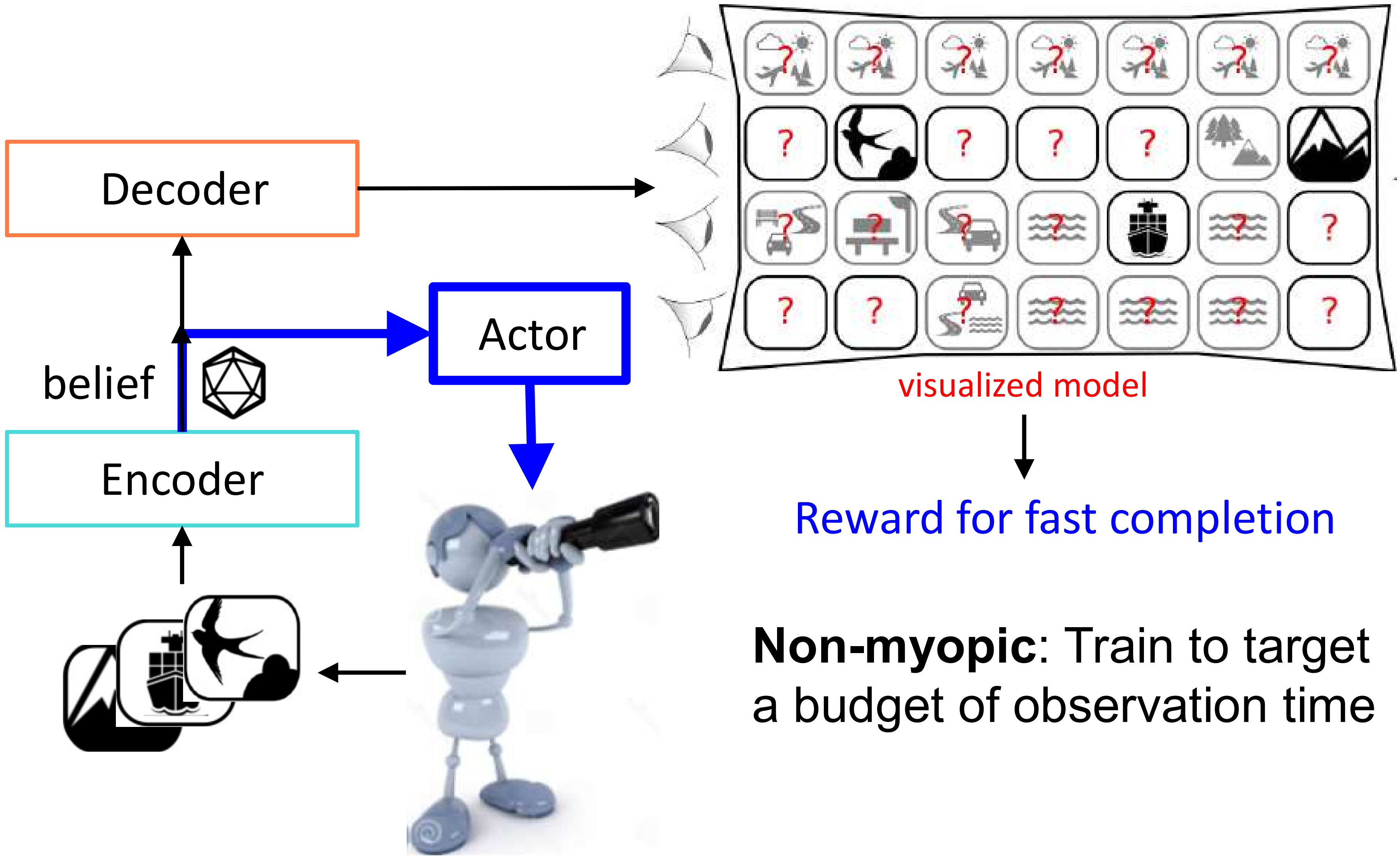
Encoder-decoder model to infer unseen viewpoints



Kristen Grauman

Jayaraman and Grauman, CVPR 2018; Ramakrishnan & Grauman, ECCV 2018

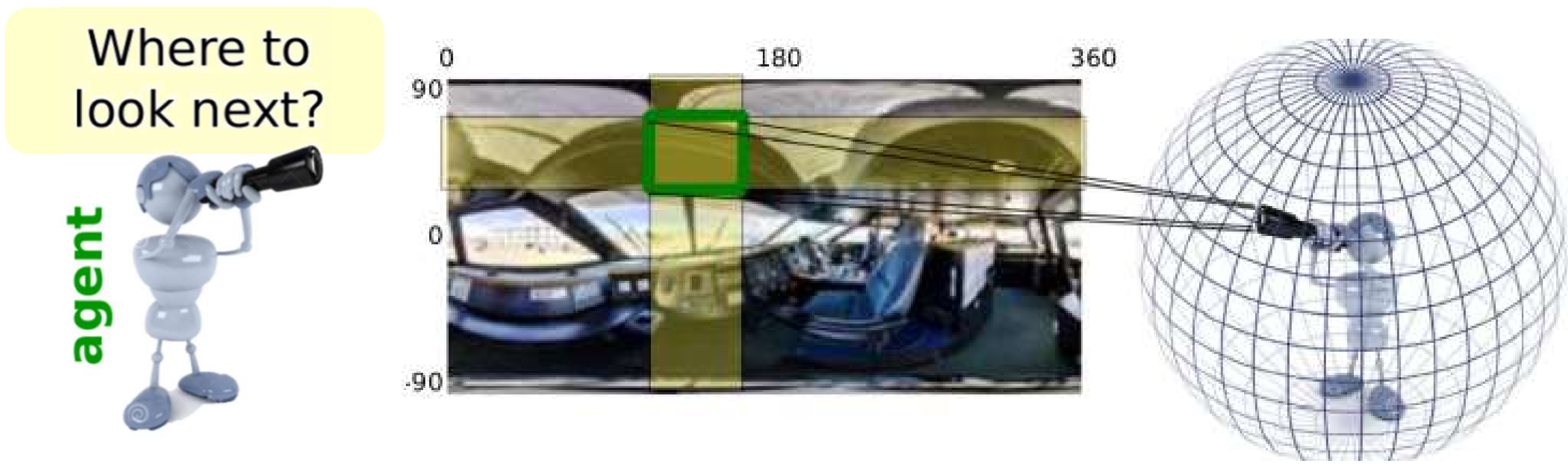
Actively selecting observations

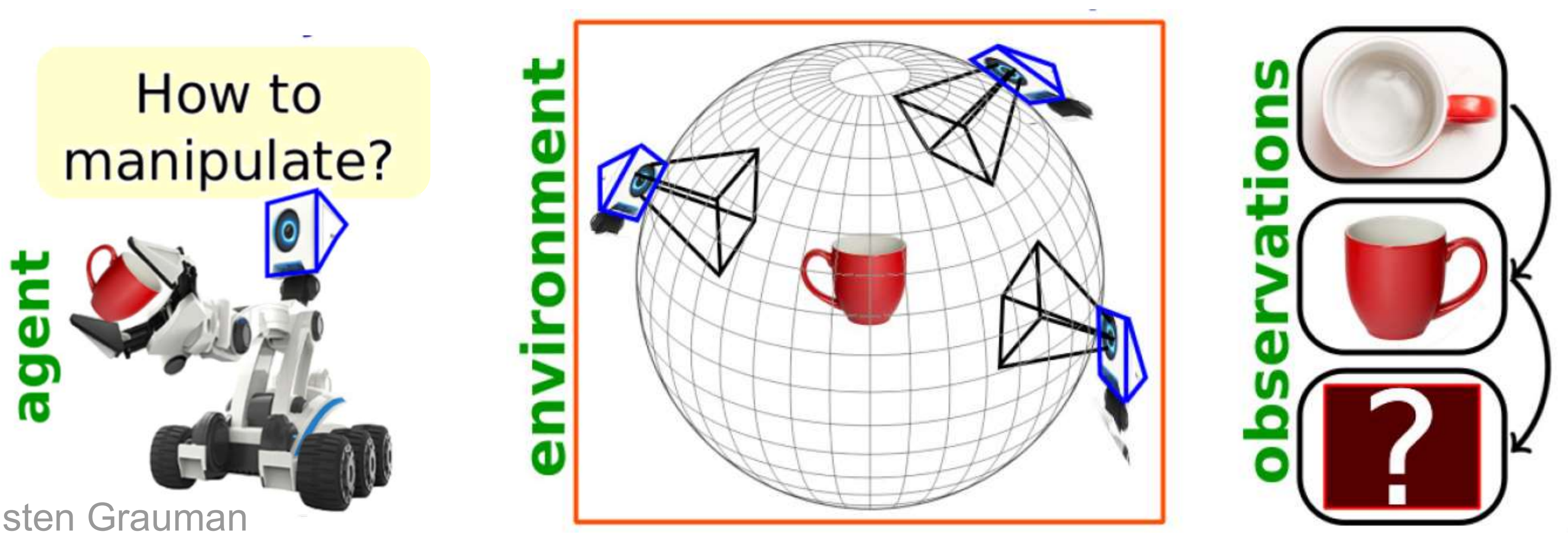


Non-myopic: Train to target a budget of observation time

Jayaraman and Grauman, CVPR 2018; Ramakrishnan & Grauman, ECCV 2018

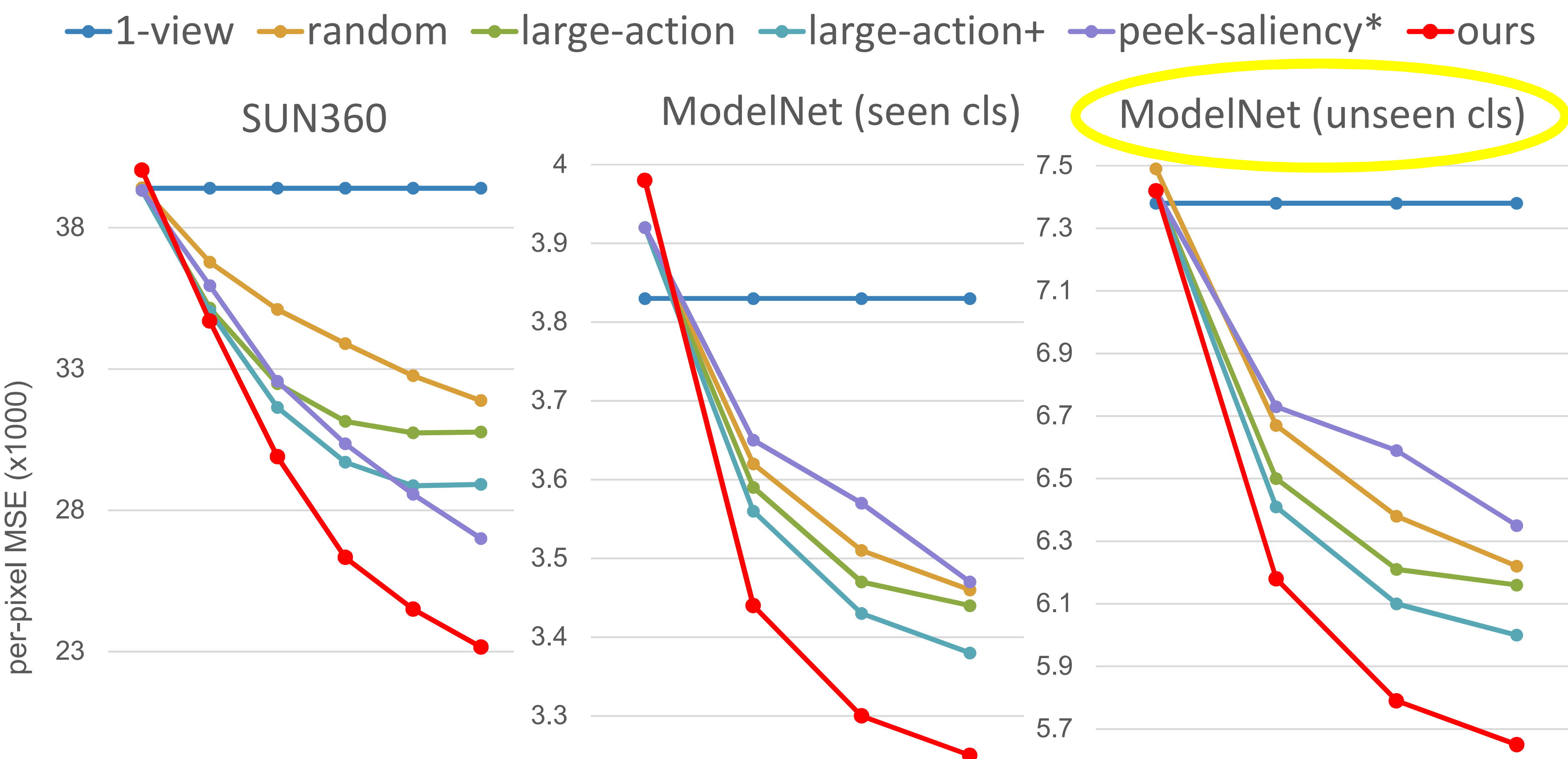
Two scenarios





Kristen Grauman

Active “look around” results

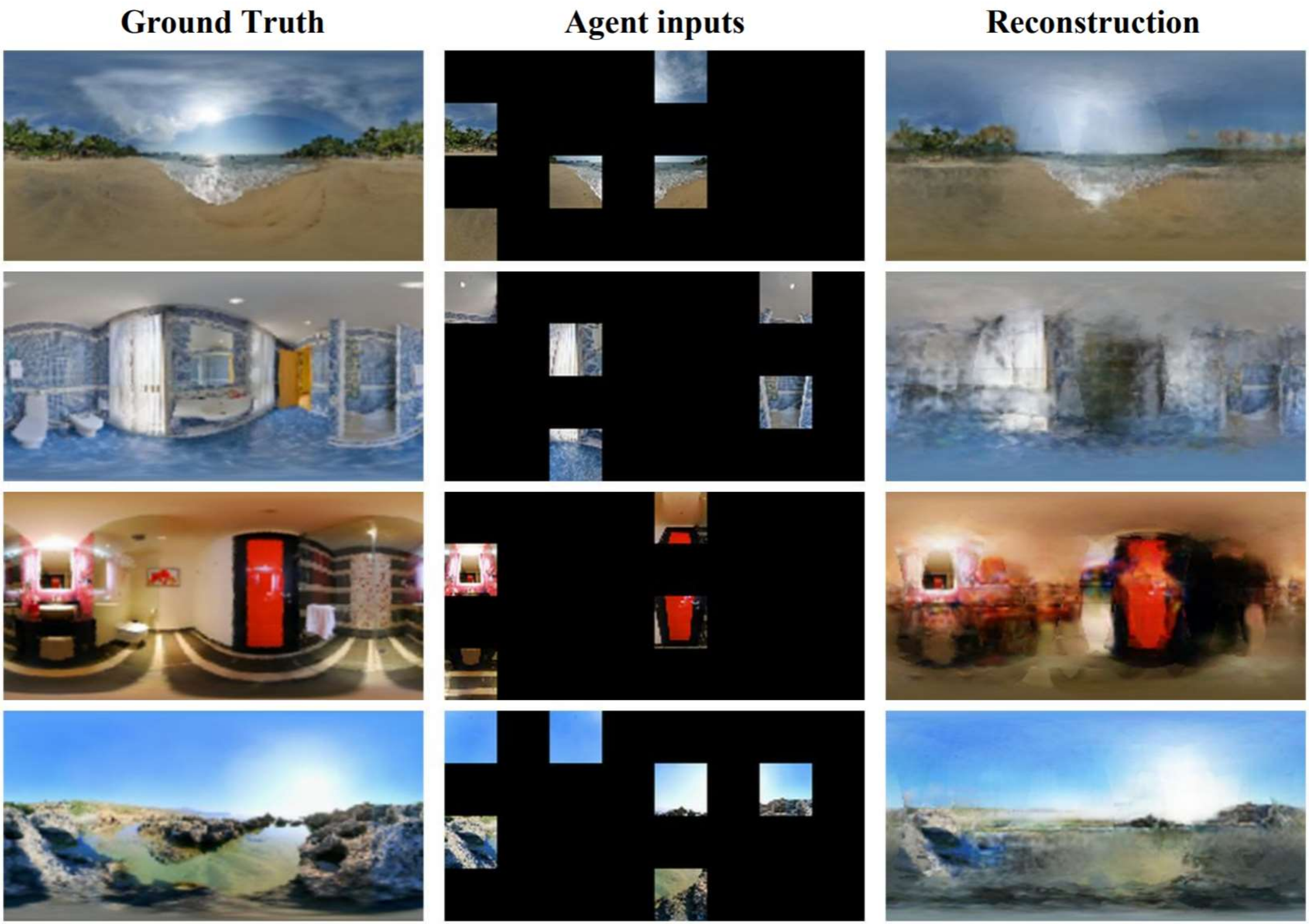


Learned active look-around policy: quickly grasp environment independent of a specific task

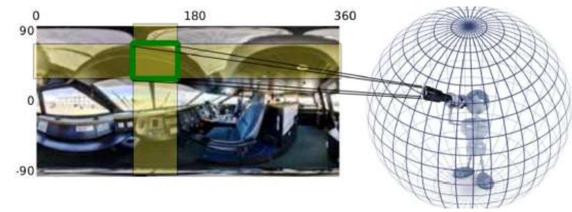
*Saliency -- Harel et al, Graph based Visual Saliency, NIPS’07

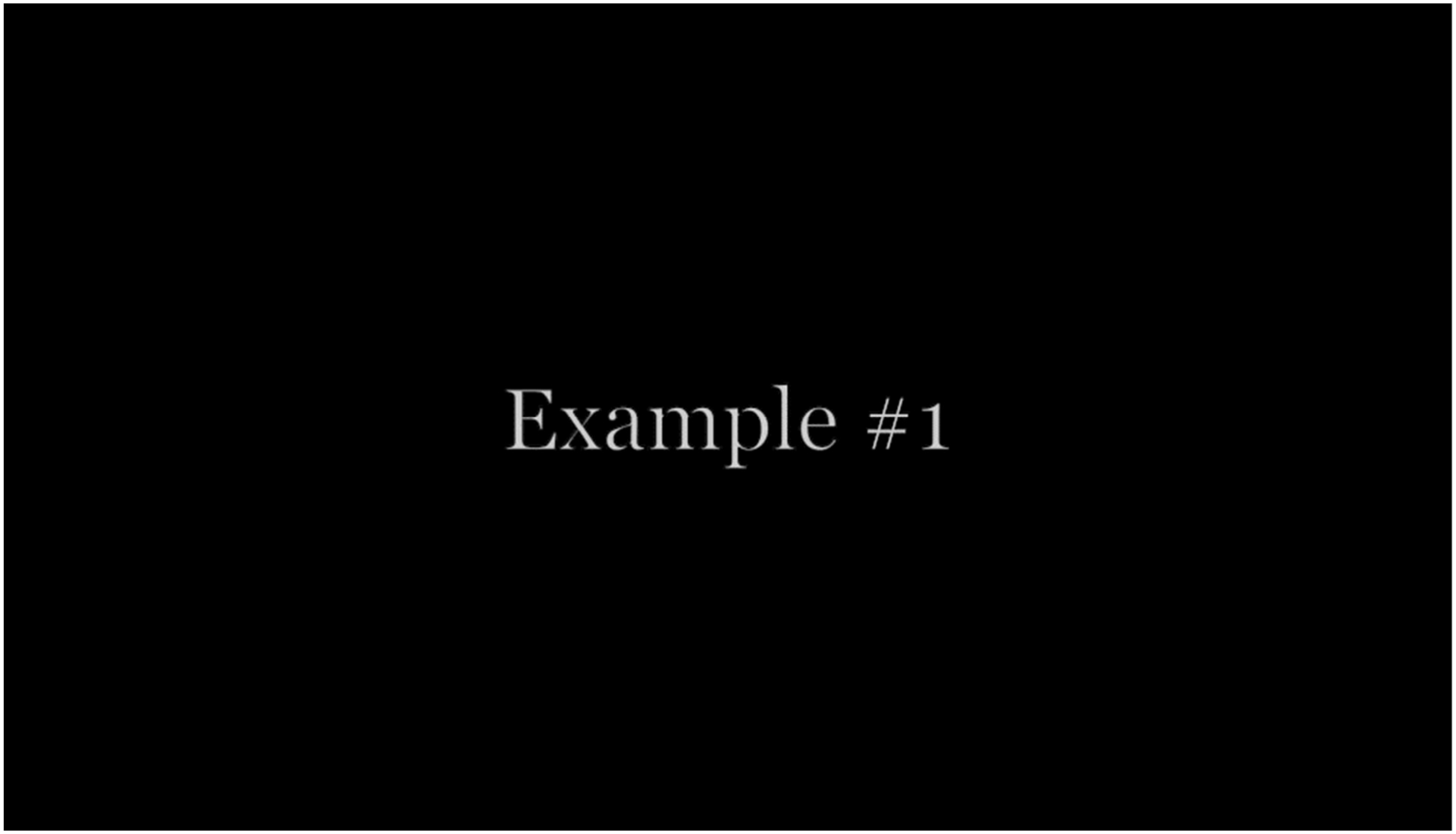
Jayaraman and Grauman, CVPR 2018

Active “look around” results



Active “look around”

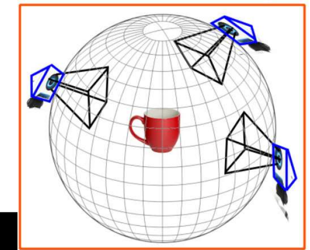




Agent’s mental model for 360 scene evolves with actively accumulated glimpses

Jayaraman and Grauman, CVPR 2018; Ramakrishnan & Grauman, ECCV 2018

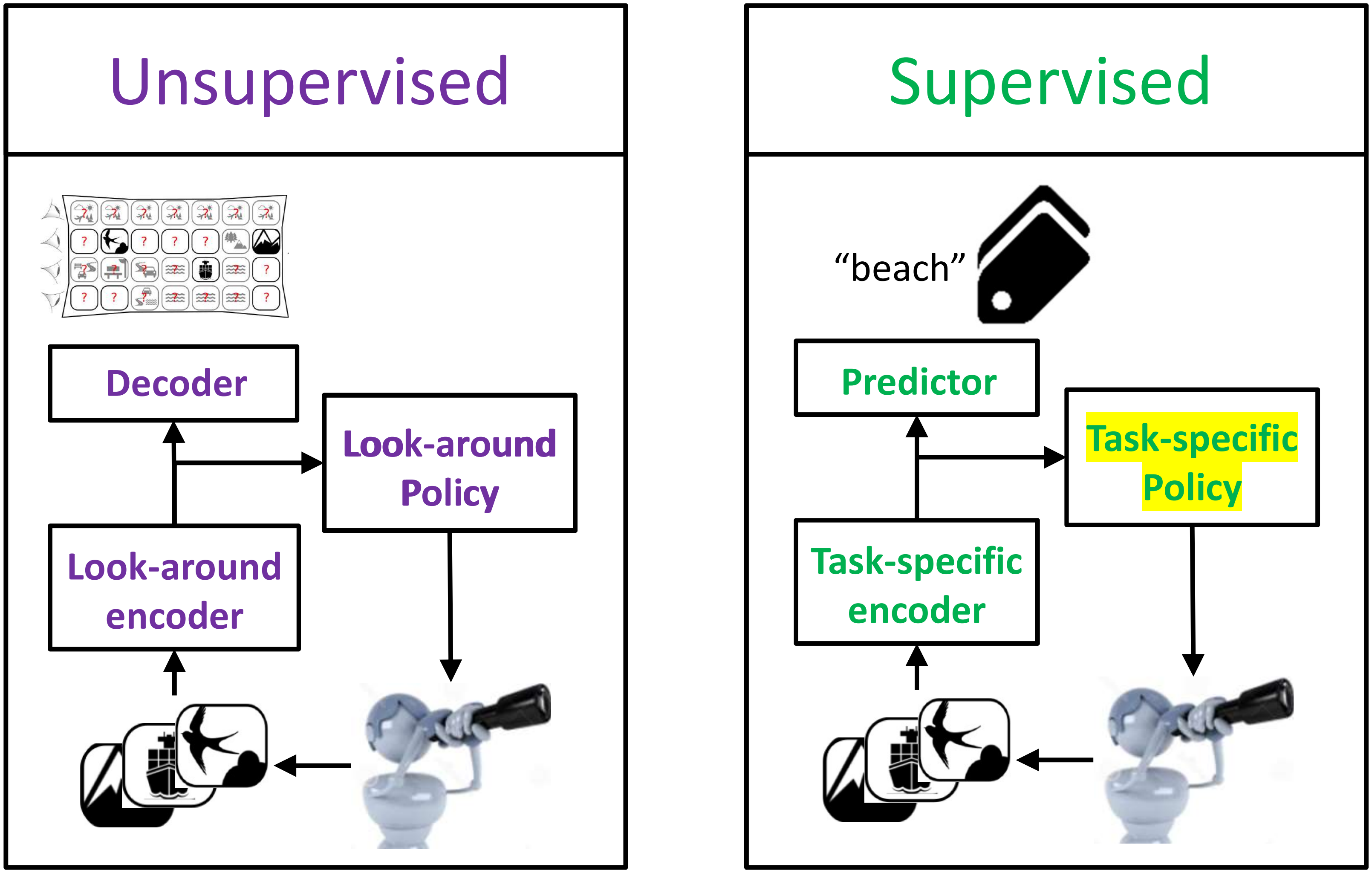
Active “look around”



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

Jayaraman and Grauman, CVPR 2018; Ramakrishnan & Grauman, ECCV 2018

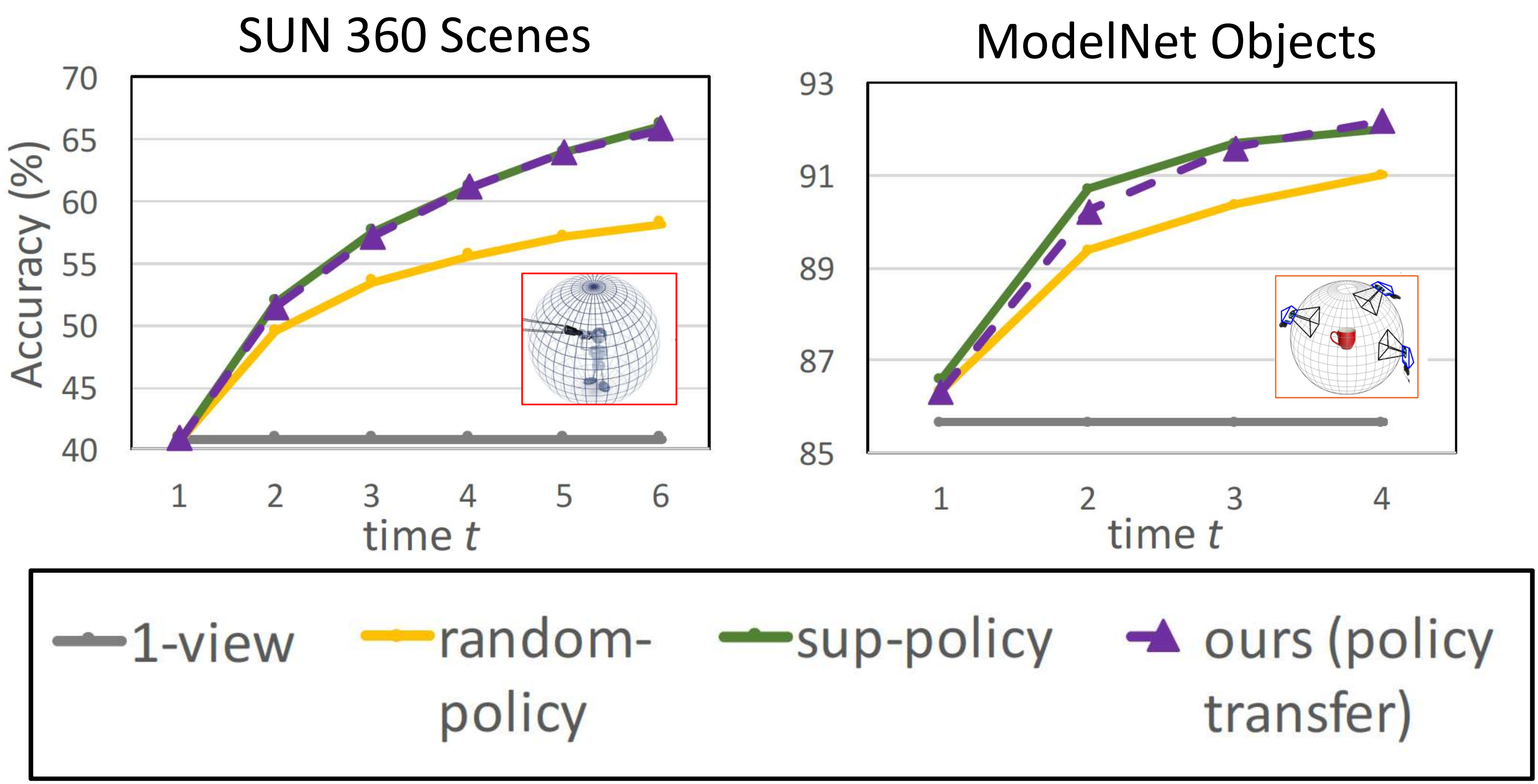
Look-around policy transfer



Plug observation completion policy in for new task

Kristen Grauman

Look-around policy transfer



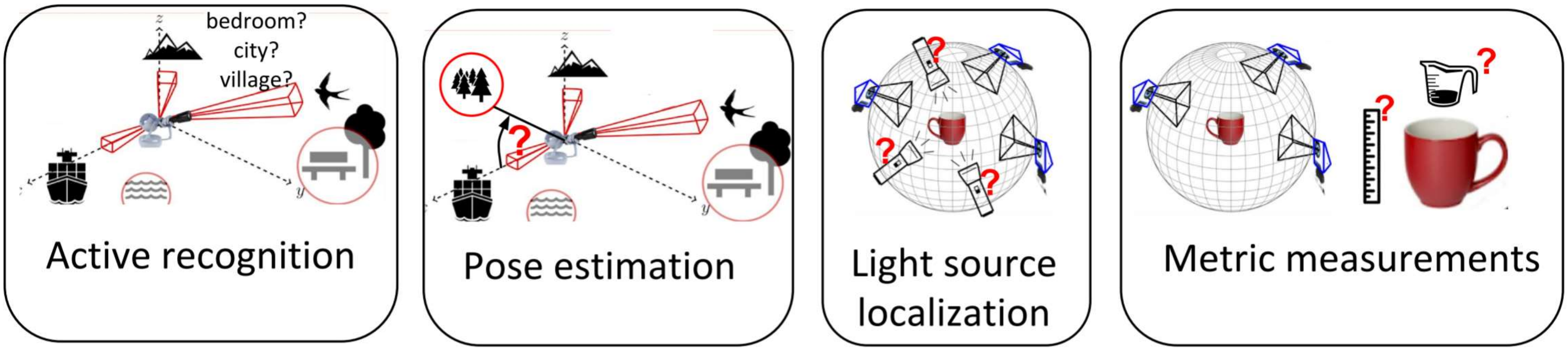
Plug task

Unsupervised exploratory policy approaches supervised task-specific policy accuracy!

Kristen Grauman

Jayaraman and Grauman, CVPR 2018

Look-around policy transfer



Multiple perception tasks

Task / Method	SUN360 Active recogn. Accuracy↑	SUN360 Pose estimation. AE azim.↓	SUN360 Pose estimation. AE elev.↓	ModelNet Active recogn. Accuracy↑	ModelNet Light source loc. Accuracy ↑	ModelNet Surface area RMSE×100 ↓
`one-view`	51.94	75.74	30.32	83.60	58.74	21.22
`rnd-actions`	62.90	66.18	19.53	88.46	72.97	19.04
`large-action`	63.73	67.57	19.94	89.05	75.14	18.38
`peek-saliency`	64.20	65.46	19.76	88.74	71.19	18.85
`supervised`	68.21	51.36	9.81	88.58	**86.30**	18.43
`lookaround`	68.89	50.00	9.94	89.00	83.29	18.82
`lookaround+spl`	**69.32**	**47.13**	**9.36**	**89.38**	83.08	**18.14**

Kristen Grauman

Ramakrishnan et al. 2019

Look-around policy transfer

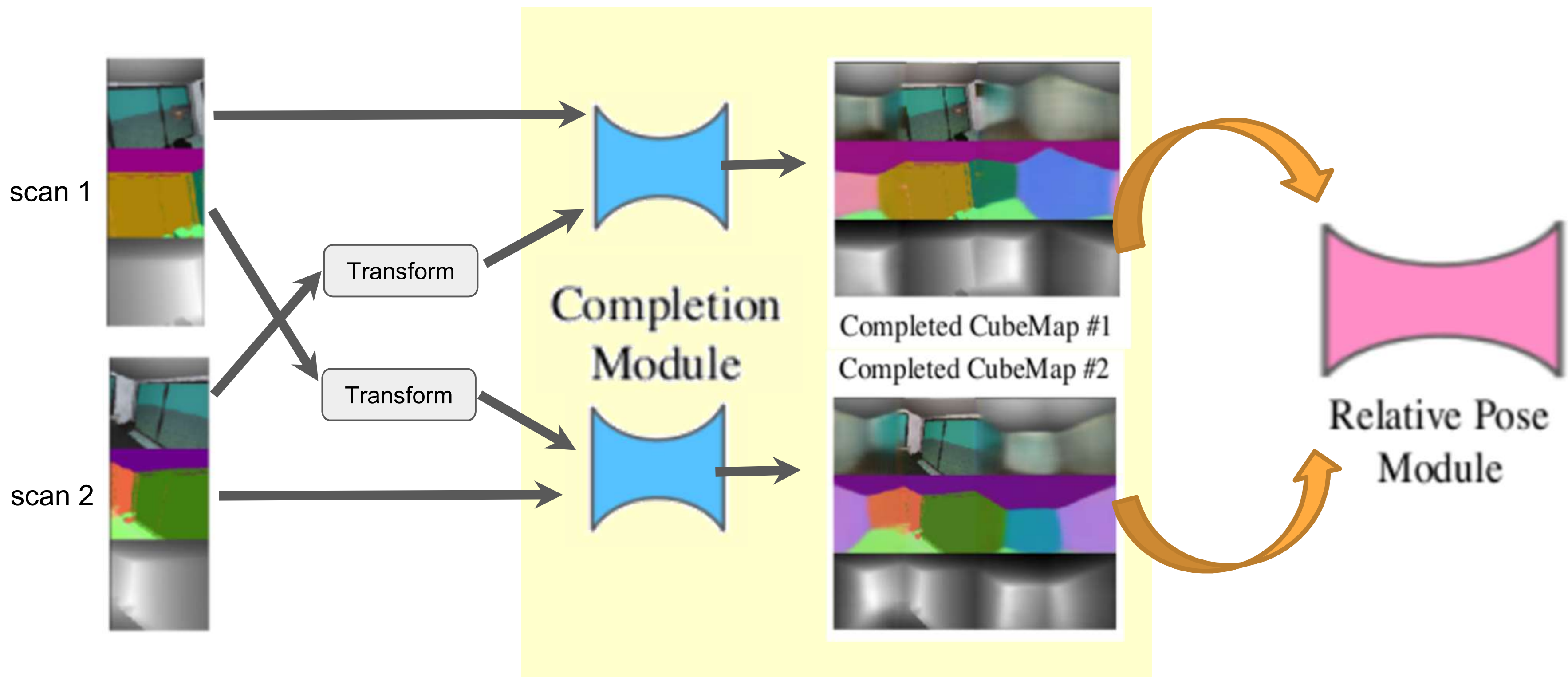
Agent navigates 3d environment leveraging active exploration

Kristen Grauman

Extreme relative pose from RGB-D scans

Input: Pair of RGB-D scans with little or *no* overlap

Output: Rigid transformation (R,t) that separates them

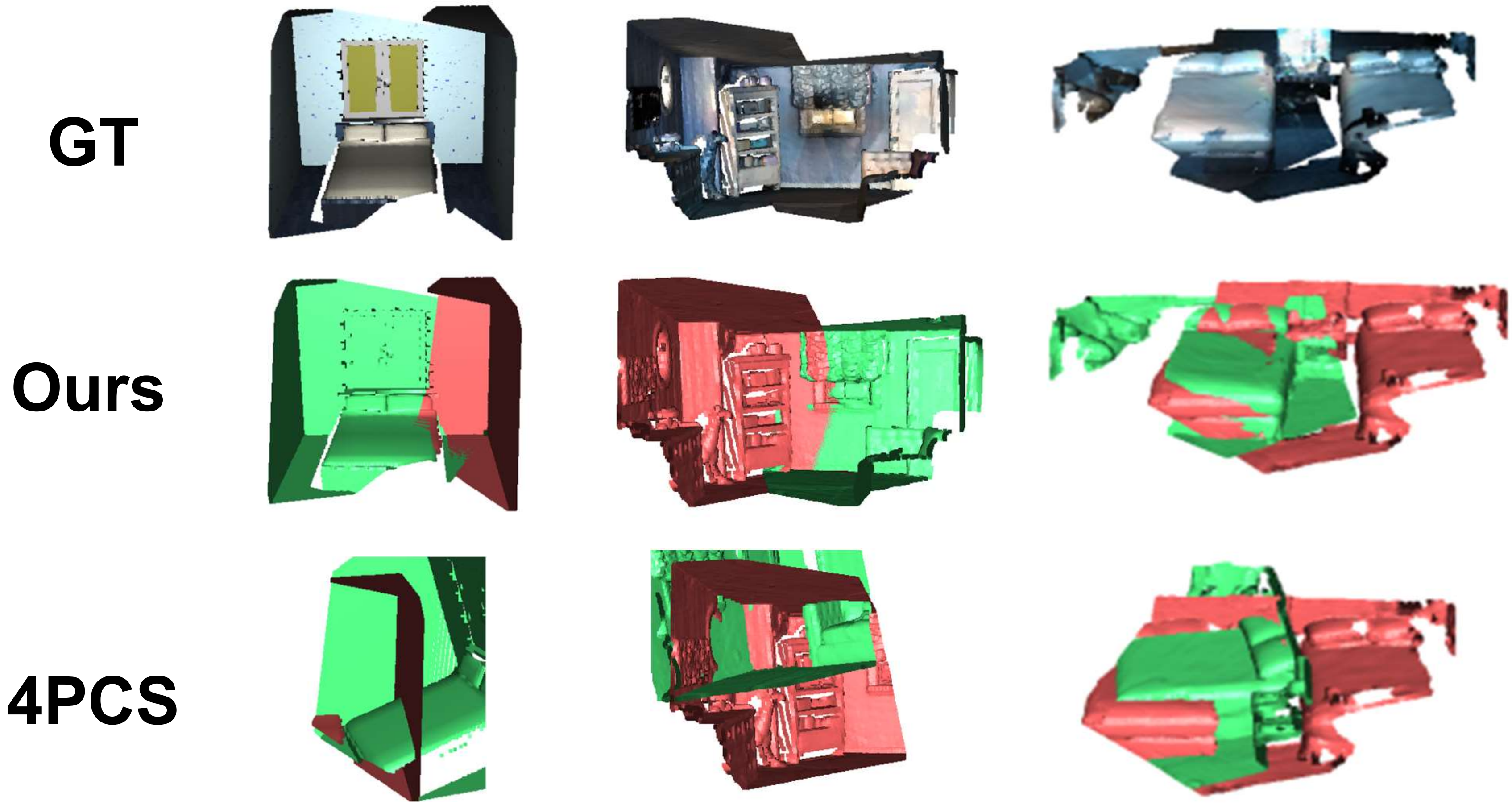


Approach: Alternate between completion and matching

Kristen Grauman

Yang et al. CVPR 2019

Extreme relative pose from RGB-D scans



Outperform existing methods on SUNCG / Matterport / ScanNet, particularly for small overlap case (10% to 50%)

Yang et al. CVPR 2019

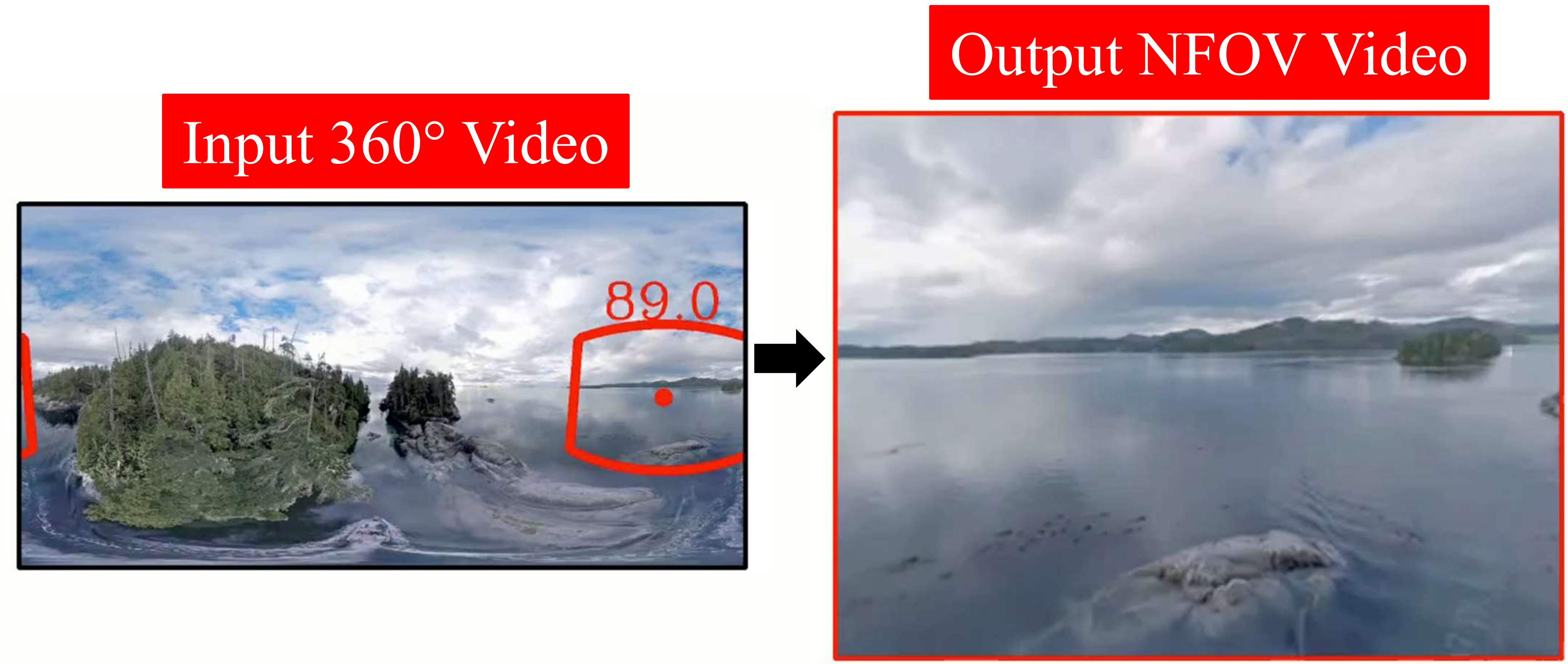
360° video: a “look around” problem for people

Control by mouse

Where to look when?

Kristen Grauman

AutoCam



Automatically select FOV and viewing direction

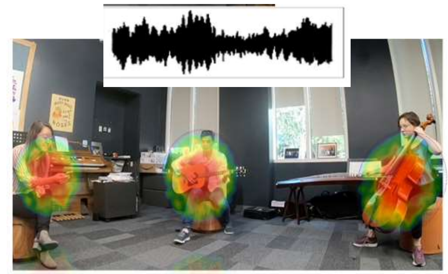
Kristen Grauman

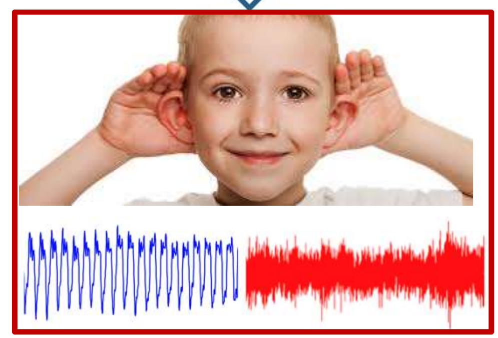
[Su & Grauman, ACCV 2016, CVPR 2017]

Anticipating the unseen and unheard

Look-around policies

Affordance learning





Audio-visual learning

Towards embodied perception

Kristen Grauman

Object interaction

Embodied perception system

Object manipulation

Turn on

Increase height

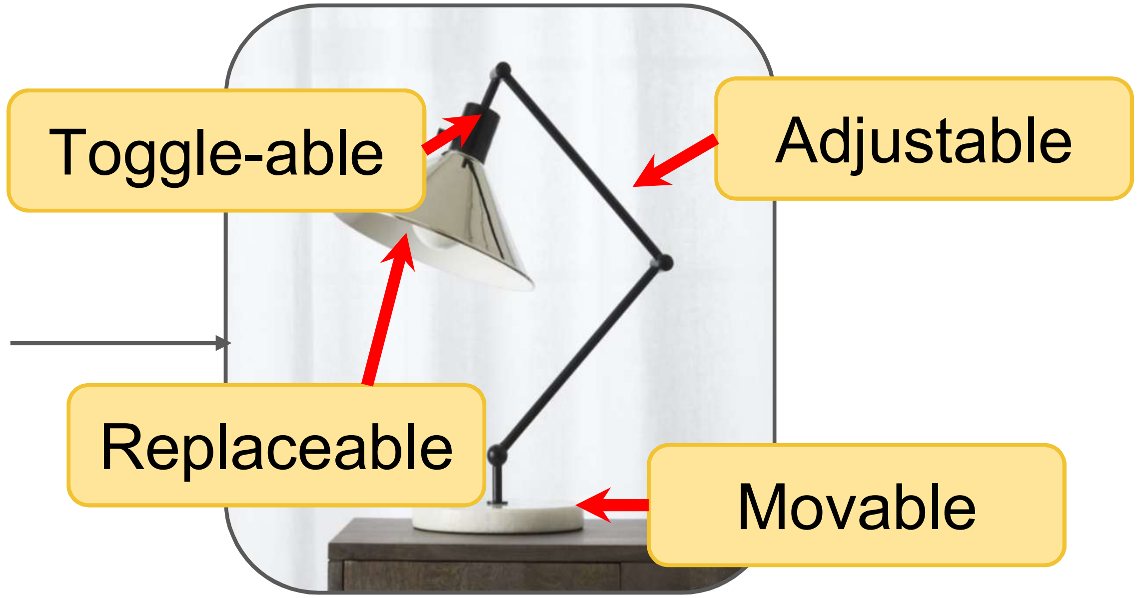
Move lamp

Replace lightbulb

Kristen Grauman

What actions does an object *afford*?

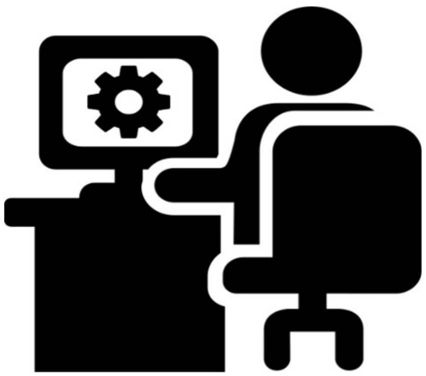
Embodied
perception system



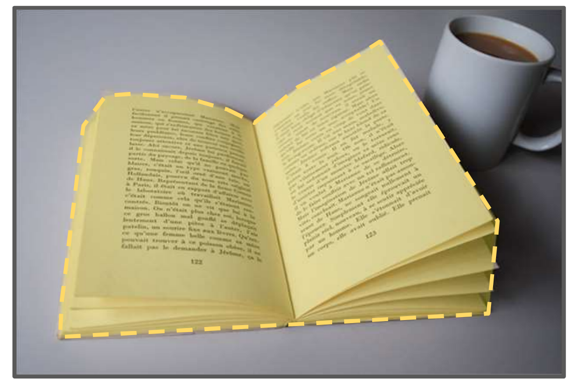
Object
manipulation

Kristen Grauman

Current approaches: affordance as semantic segmentation



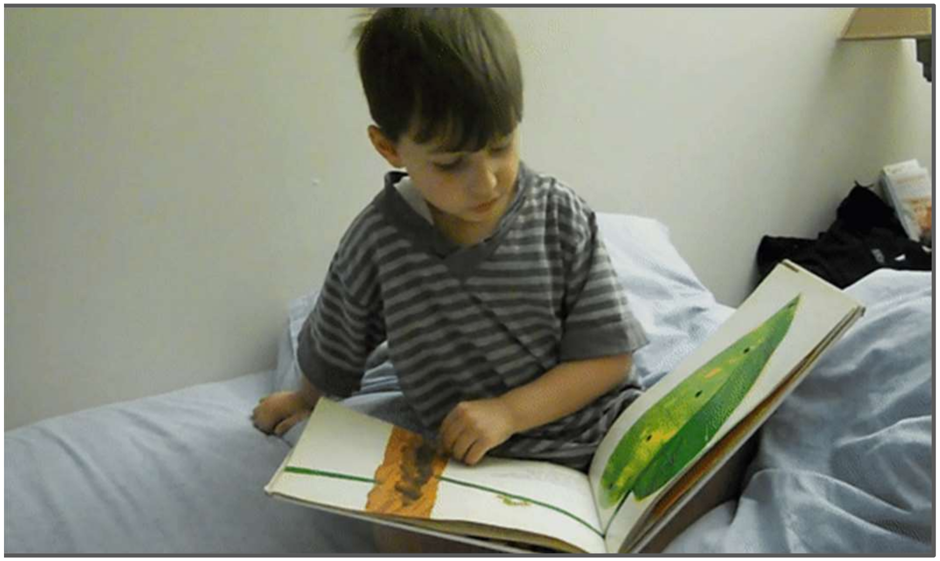
Label “holdable” regions

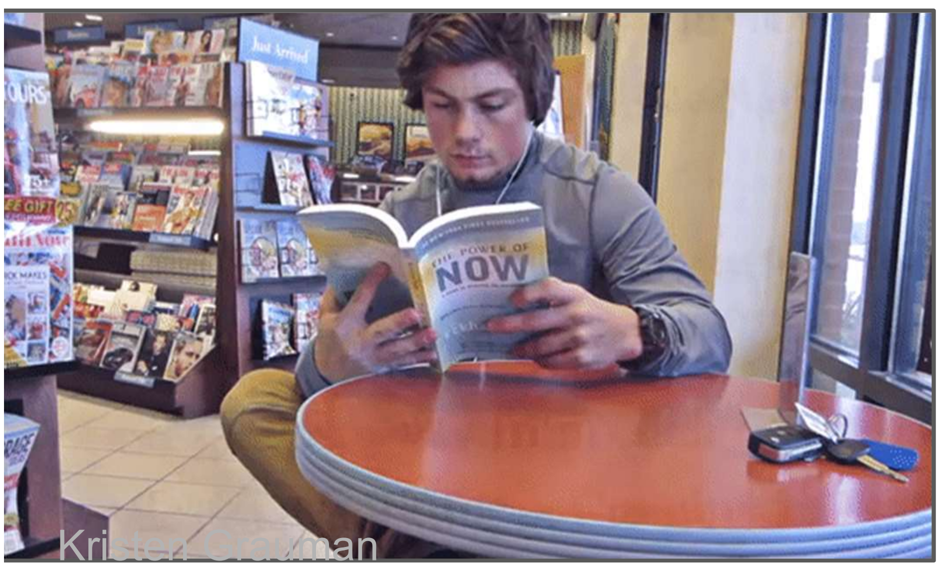


Captures annotators’ expectations of what is important

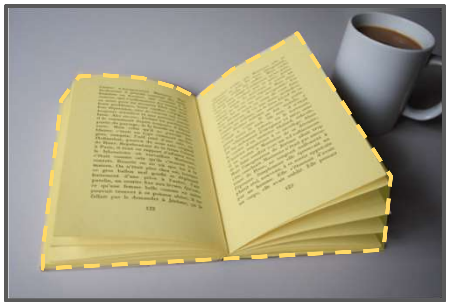
Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …

…but real human behavior is complex



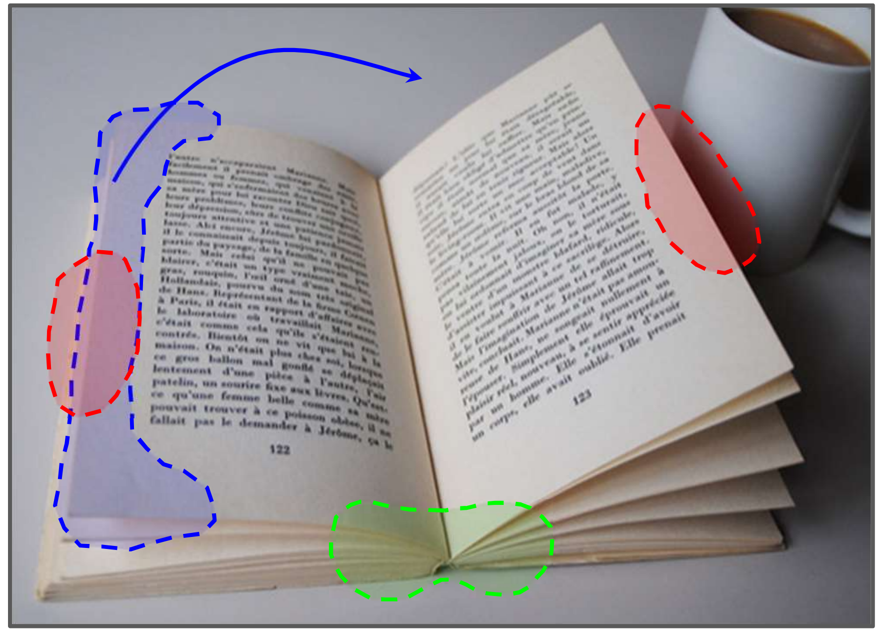


How to learn object affordances?



Manually curated affordances

V S.



Real human interactions?

Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …

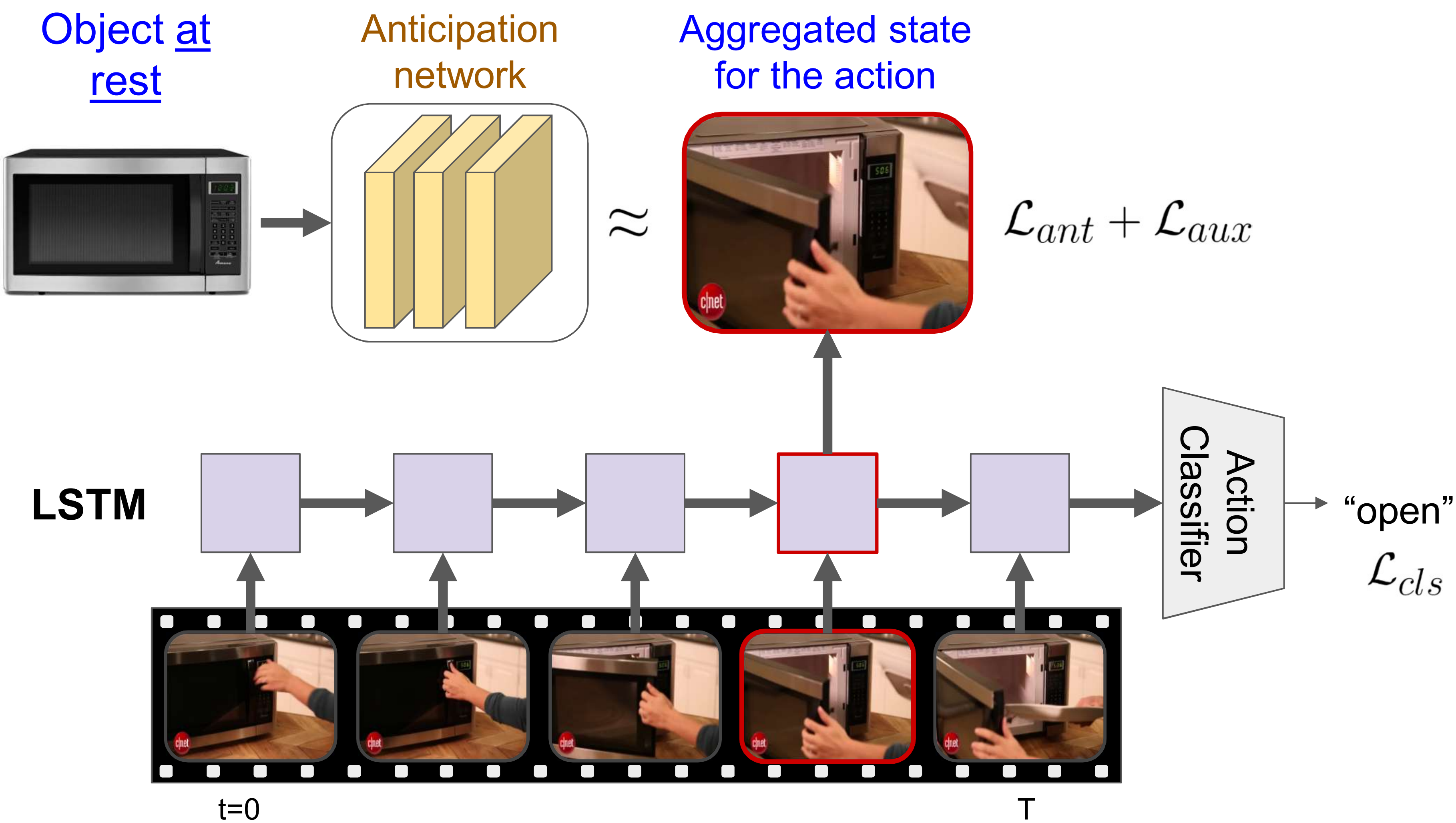
Kristen Grauman

Our idea: Learn directly by watching people (video)

Kristen Grauman

[Nagarajan et al. 2019]

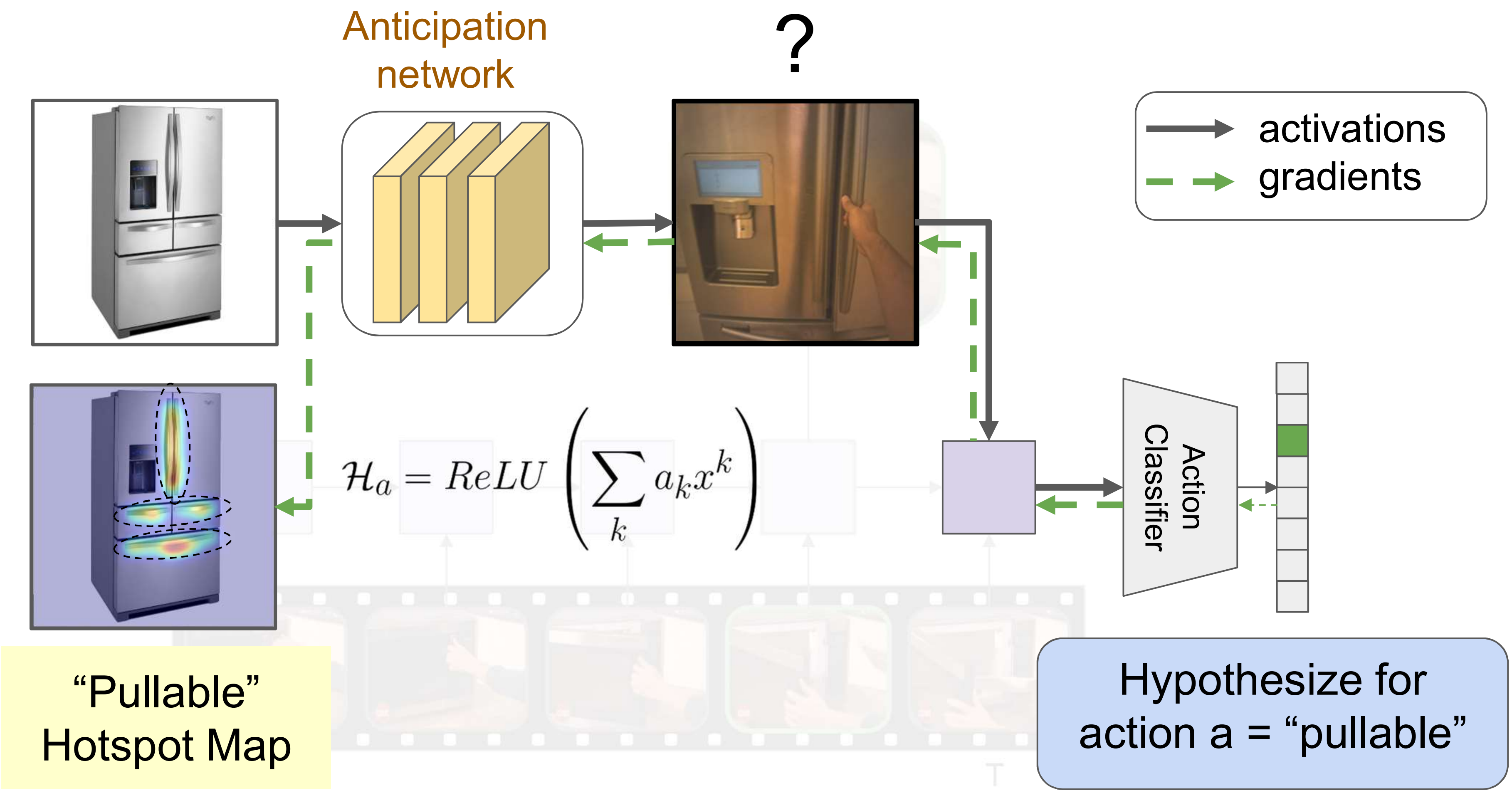
Learning affordances from video



Kristen Grauman

[Nagarajan et al. 2019]

Extracting interaction hotspot maps



Activation mapping to identify responsible spatial regions

Kristen Grauman

[Nagarajan et al. 2019]

Wait, is this just action recognition?

Action recognition + Grad-CAM

Ours

No: Hotspot anticipation model maps
object at rest to *potential for interaction*

Kristen Grauman

Evaluating interaction hotspots

OPRA
(Fang et al., CVPR 18)

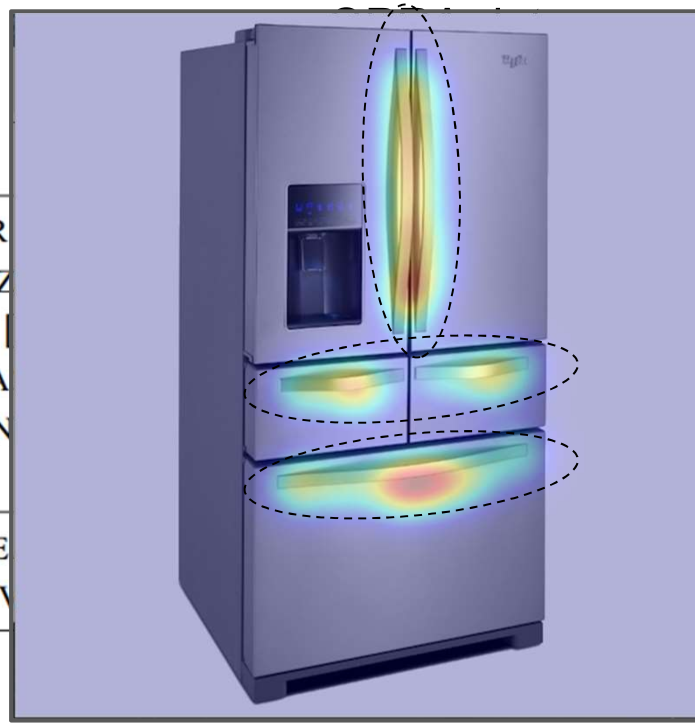
EPIC Kitchens
(Damen et al., ECCV 18)

MS COCO
(Lin et al., ECCV 14)

Train on video datasets, generate heatmaps on novel images---
even from unseen categories

Results: interaction hotspots

Given static image of object at rest, infer affordance regions



Weakly Supervised

Strongly Supervised

	EPIC data		
	KLD ↓	SIM ↑	AUC-J ↑
CENTER	9.20	0.24	0.66
EGOGAZ	2.13	0.29	0.64
MLNET [	3.70	0.32	0.75
DEEPGA	1.34	0.40	0.76
SALGAN	1.48	0.40	**0.78**
OURS	**1.25**	**0.41**	**0.78**
IMG2HE	1.56	0.35	0.77
DEMO2V	–	–	–

Up to 24% increase vs. weakly supervised methods

Kristen Grauman

[Nagarajan et al. 2019]

Results: interaction hotspots

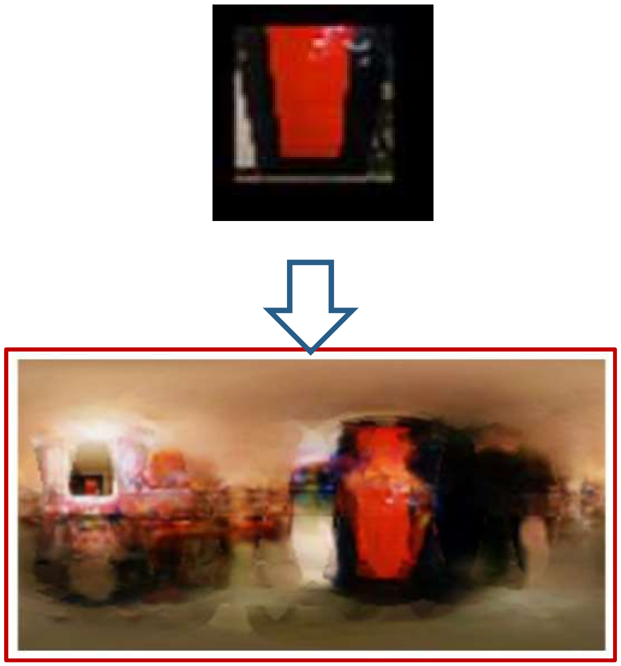
Kristen Grauman

Results: hotspots for recognition

	COCO			
N →	5	25	100	3300 (all)
VANILLA	44.3 ± 0.3	56.6 ± 0.2	**65.6 ± 0.4**	**75.2 ± 0.1**
AUTOENCODER	39.4 ± 0.4	51.2 ± 0.2	59.1 ± 0.2	72.8 ± 0.3
OURS	**46.8 ± 0.3**	**57.9 ± 0.1**	63.2 ± 0.2	73.9 ± 0.3

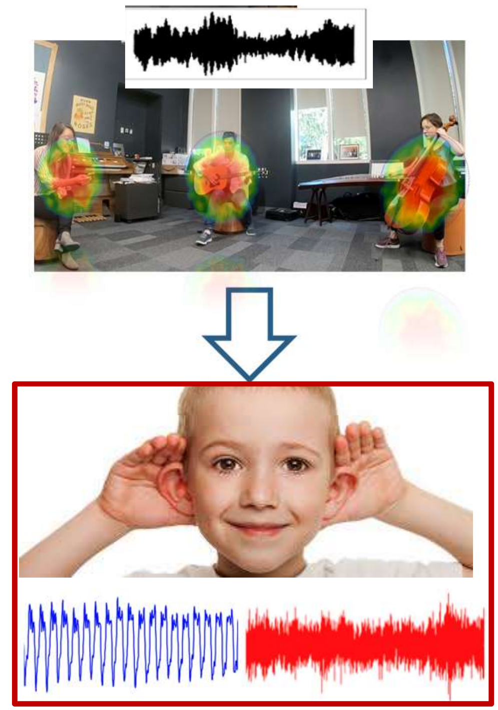
Better low-shot object recognition
by anticipating object function

Anticipating the unseen and unheard



Look-around policies

Affordance learning



Audio-visual learning

Towards embodied perception

Kristen Grauman

Listening to learn

woof **meow** **ring** **clatter**

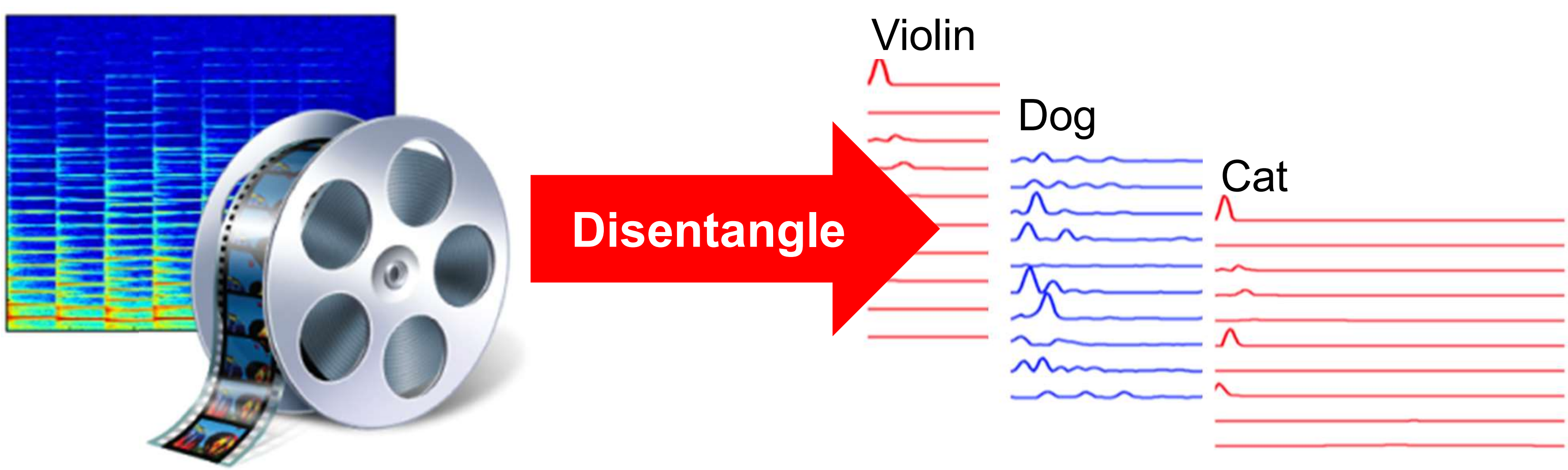
Goal: a repertoire of objects and their sounds

Challenge: a single audio channel mixes sounds of multiple objects

Kristen Grauman

Learning to separate object sounds

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



Unlabeled video

Object sound models

Apply to **separate** simultaneous sounds in novel videos

Kristen Grauman

[Gao, Feris, & Grauman, ECCV 2018]

Results: audio-visual source separation

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

original video
(before separation)

visual predictions:
violin & acoustic guitar

Dataset: AudioSet [Gemmeke et al. 2017]

Kristen Grauman

[Gao et al. ECCV 2018]

Results: audio-visual source separation

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

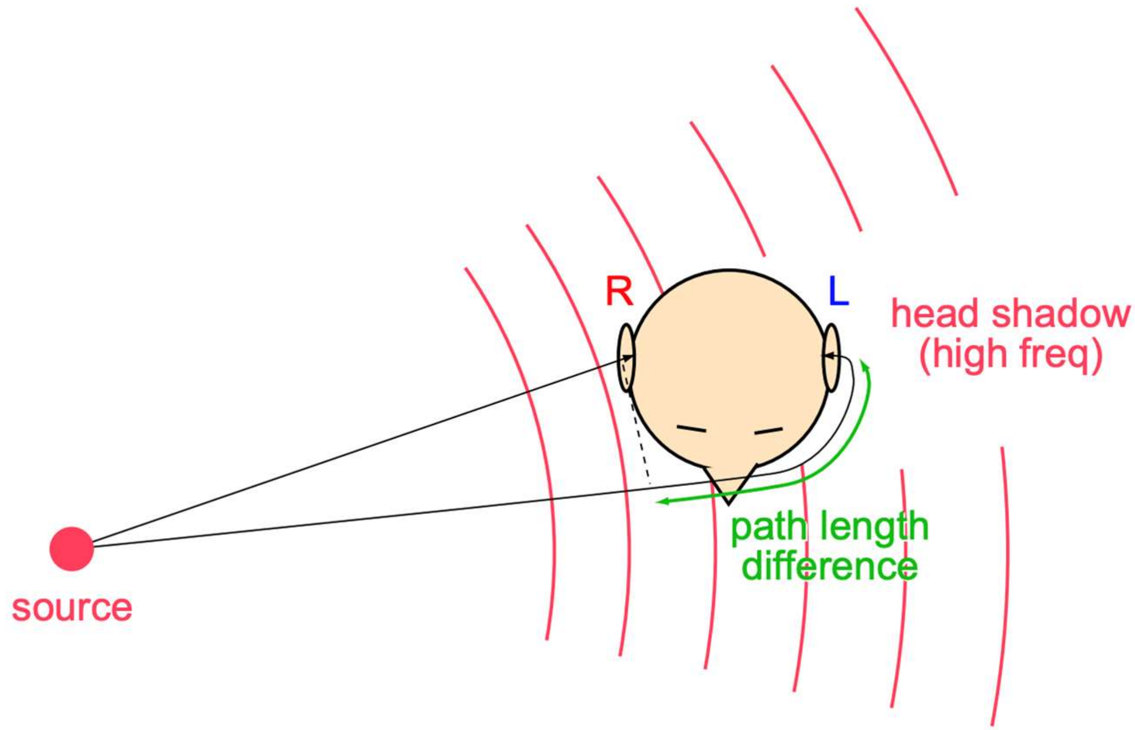
original video
(before separation)

visual predictions:
dog & violin

[Gao et al. ECCV 2018]

Spatial effects in audio

Spatial effects absent in monaural audio



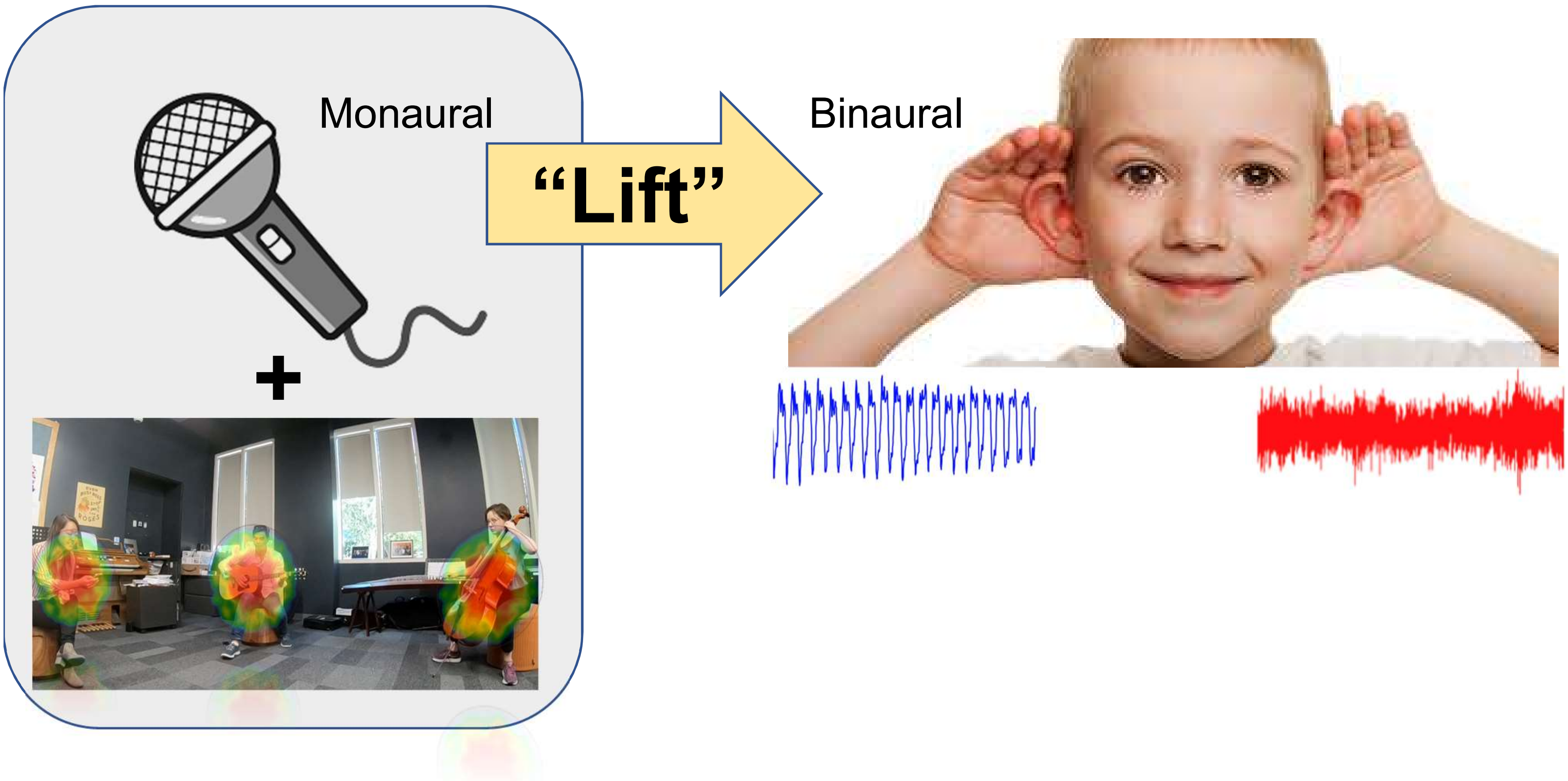
Cues for spatial hearing:

- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)

Image Credit: Michael Mandel

Our idea: 2.5D visual sound

"Lift" mono audio to spatial audio via visual cues

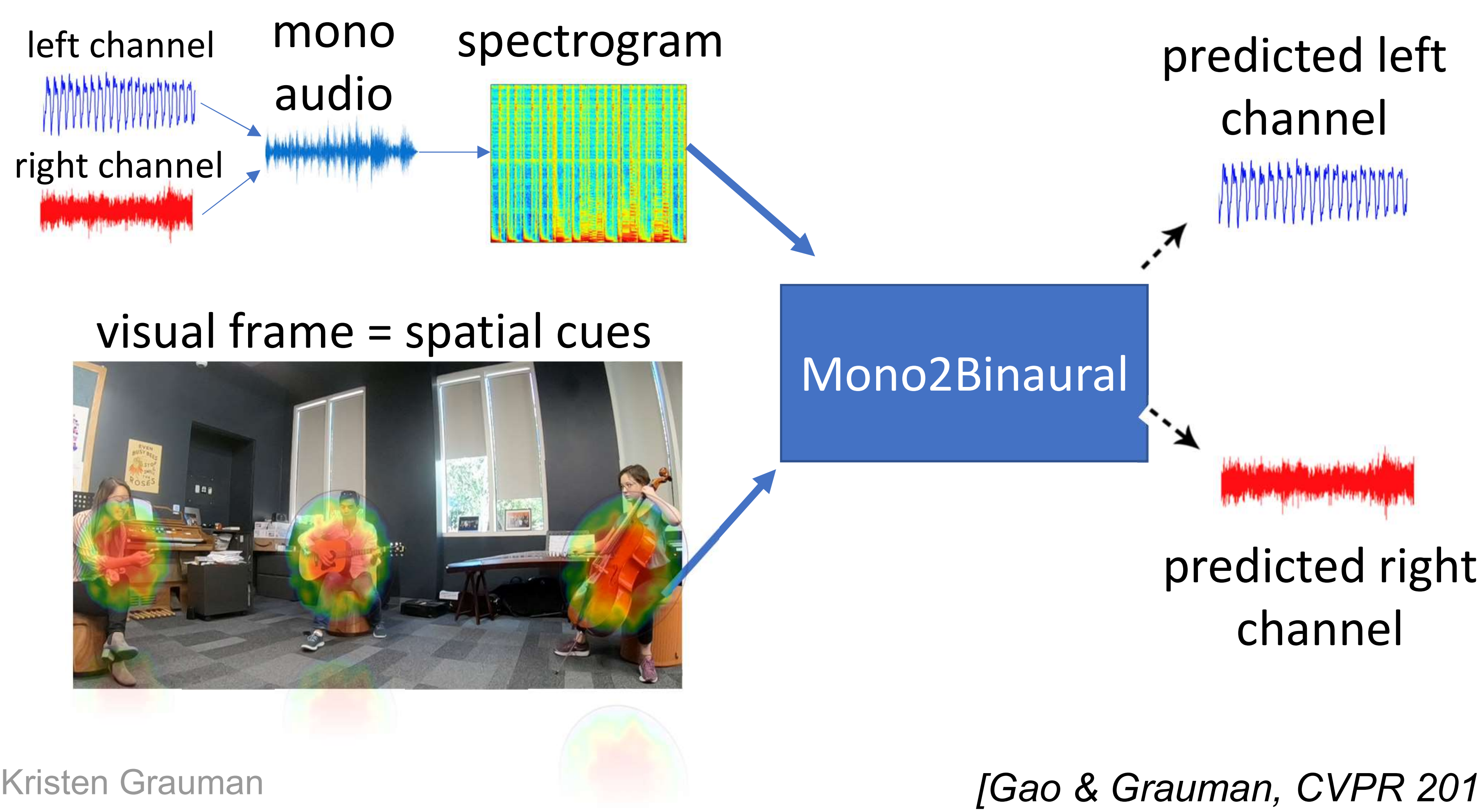


Kristen Grauman

[Gao & Grauman, CVPR 2019]

Our idea: 2.5D visual sound

"Lift" mono audio to spatial audio via visual cues



Kristen Grauman

[Gao & Grauman, CVPR 2019]

New: FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Binaural microphone rig linked to camera and monoaural mic

Capture ~5 hours video and binaural sound in music room

[Gao & Grauman, CVPR 2019]

Datasets

FAIR-Play Binaural

- 10 musical instruments, e.g., cello, guitar, harp, ukulele, trumpet, etc.
- ~5 hours of performances

Ambisonics Datasets

[Morgado *et al.* NIPS 2018]

- Streets, random YouTube
- ~1000 360° video clips
- Converted to binaural audio using decoder

Kristen Grauman

Results: 2.5D visual sound

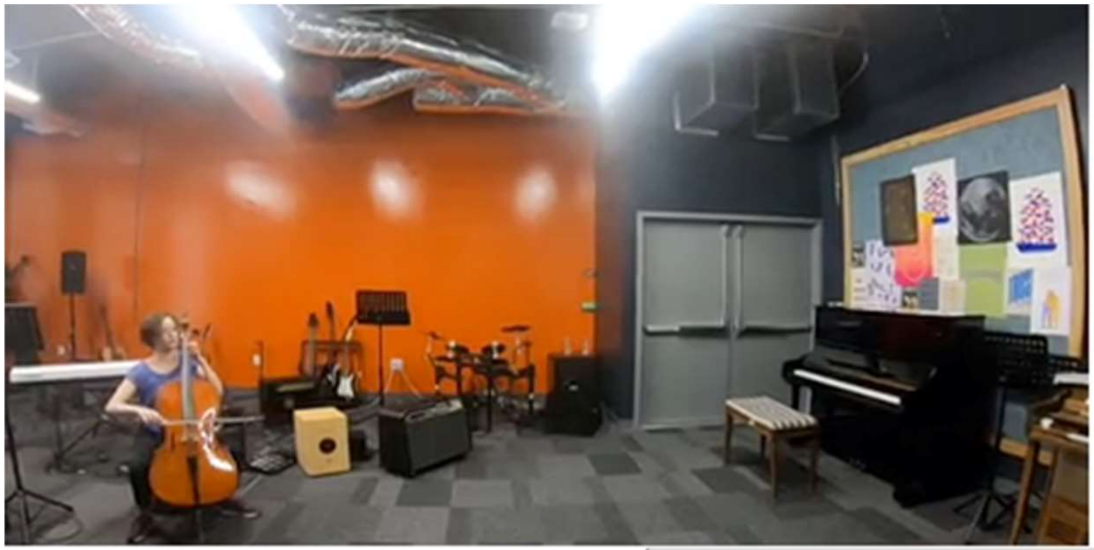
Left channel Monaural input Right channel

vision.cs.utexas.edu/projects/2.5D_visual_sound/

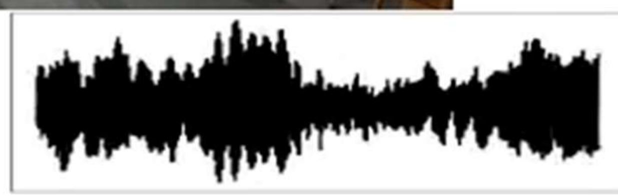
Kristen Grauman

[Gao & Grauman, CVPR 2019]

Results: 2.5D visual sound



input monaural audio



Left channel　　Monaural input　　Right channel

vision.cs.utexas.edu/projects/2.5D_visual_sound/

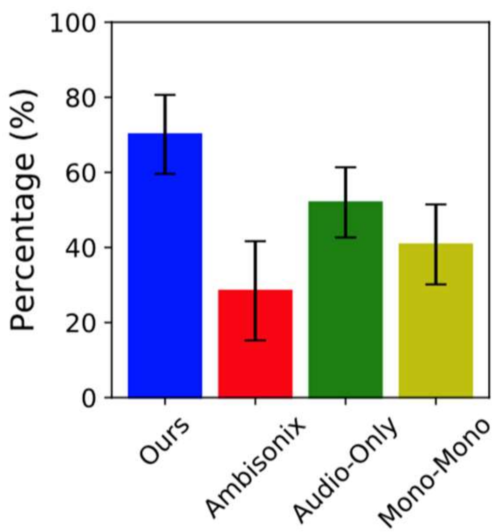
Kristen Grauman

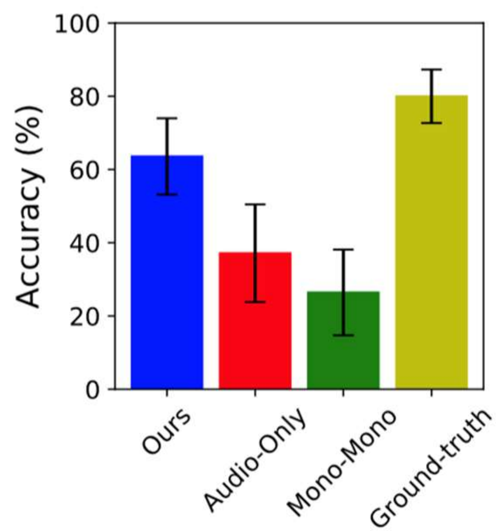
[Gao & Grauman, CVPR 2019]

Results: 2.5D visual sound

	BINAURAL-MUSIC-ROOM		REC-STREET		YT-CLEAN		YT-MUSIC	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Ambisonics [29]	-	-	0.744	0.126	1.435	0.155	1.885	0.183
Audio-Only	1.022	0.143	0.590	0.114	1.065	0.131	1.553	0.167
Flipped-Visual	1.136	0.148	0.658	0.123	1.095	0.132	1.590	0.165
Mono-Mono	1.141	0.152	0.774	0.136	1.369	0.153	1.853	0.184
MONO2BINAURAL (Ours)	**0.875**	**0.133**	**0.565**	**0.109**	**1.027**	**0.130**	**1.451**	**0.156**

Binaural audio generation error, all four datasets





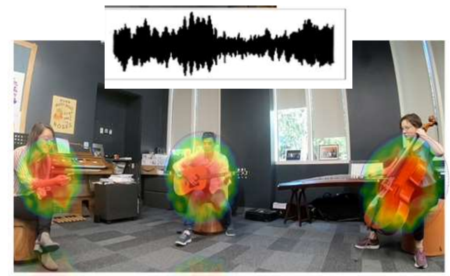
User studies:
perceived realism

Binaural audio offers "embodied" 3D sensation.
…and improves sound source separation!

Ambisonics: *Morgado et al. NIPS 2018*

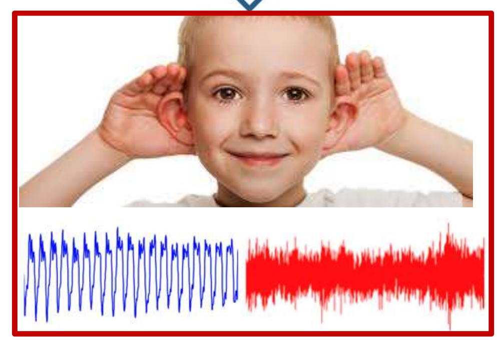
[Gao & Grauman, CVPR 2019]

Anticipating the unseen and unheard



Look-around policies

Affordance learning



Audio-visual learning

Towards embodied perception

Kristen Grauman

Summary

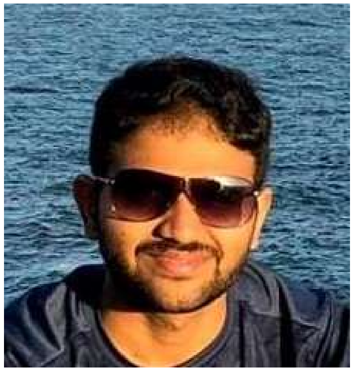
Towards embodied perception

- self-supervised learning via anticipation
- learning to autonomously direct the camera
- multi-sensory observations (audio, motion, visual)
- object interaction from video


Ruohan Gao


Tushar Nagarajan


Dinesh Jayaraman


Santhosh Ramakrishnan


Christoph Feichtenhofer