See, Hear, Move: Towards Embodied Visual Perception

Kristen Grauman

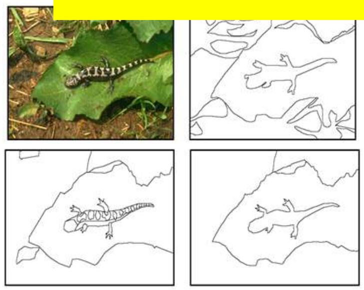
Facebook AI Research

University of Texas at Austin

How do recognition systems typically learn today?

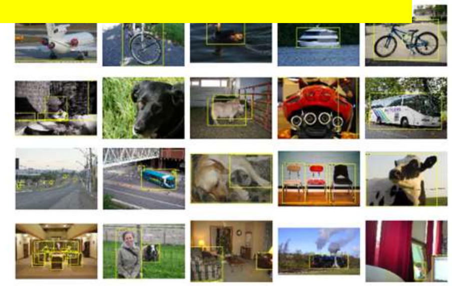
Web photos + recognition

A "disembodied" well-curated moment in time

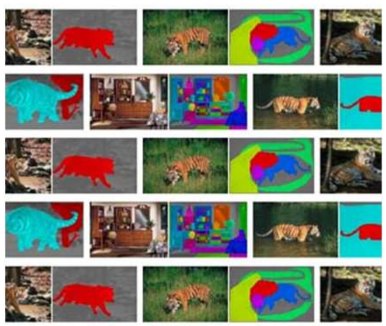


BSD (2001)

Caltech 101 (2004), Caltech 256 (2006)



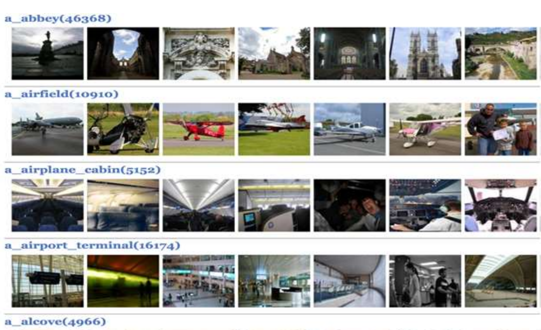
PASCAL (2007-12)



LabelMe (2007)

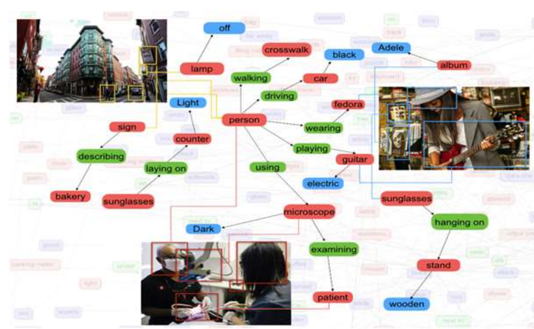
ImageNet (2009)

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 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.

Kristen Grauman

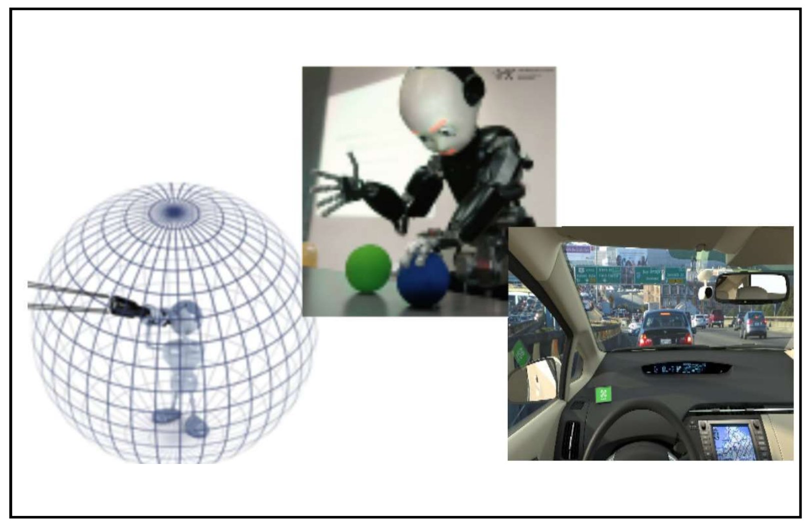
Big picture goal: Embodied visual learning

Status quo:

Learn from "disembodied" bag of labeled snapshots.

On the horizon:

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

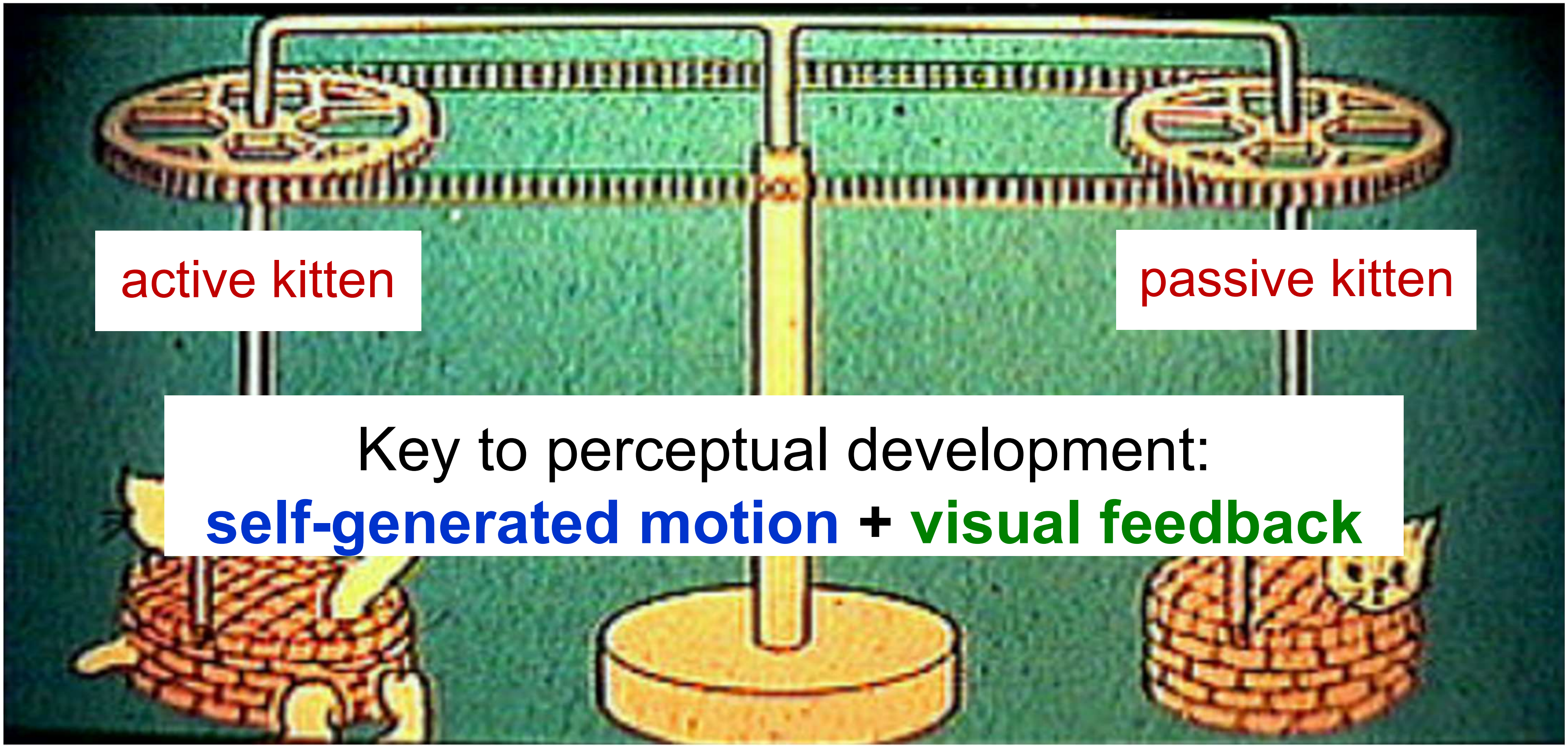


Kristen Grauman

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: Egomotion ↔ vision

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

Ego-motion motor signals

+

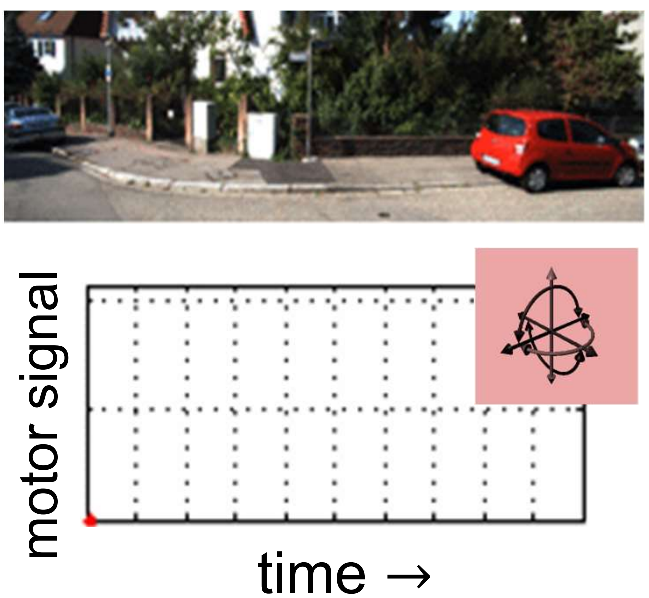
Unlabeled video

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

Approach: Egomotion equivariance

Training data

Unlabeled video +
motor signals



Learn

Equivariant embedding

organized by egomotions

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

Pairs of frames related by similar egomotion should be related by same feature transformation

[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

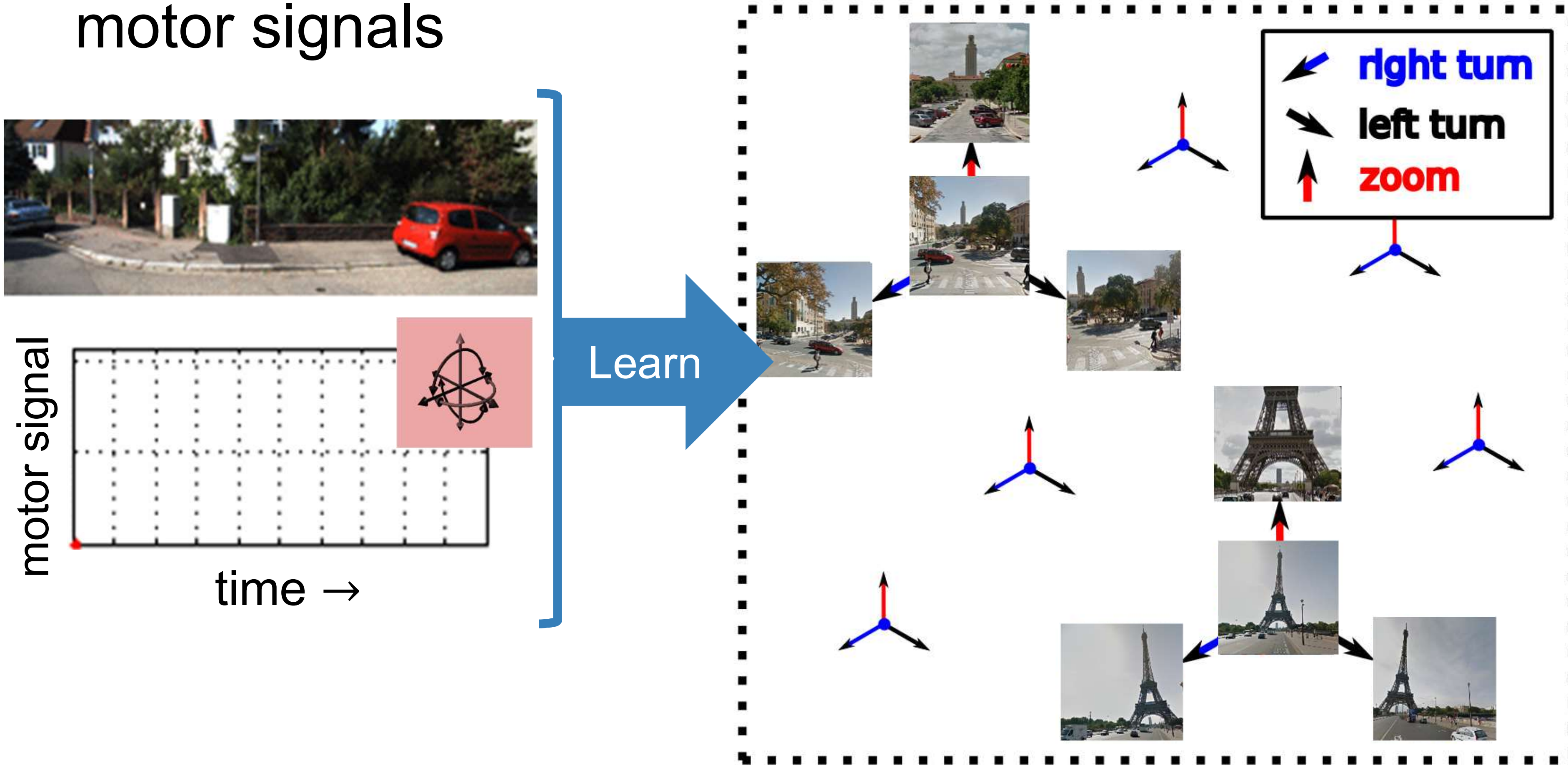
Approach: Egomotion equivariance

Training data

Unlabeled video + motor signals

Equivariant embedding

organized by egomotions



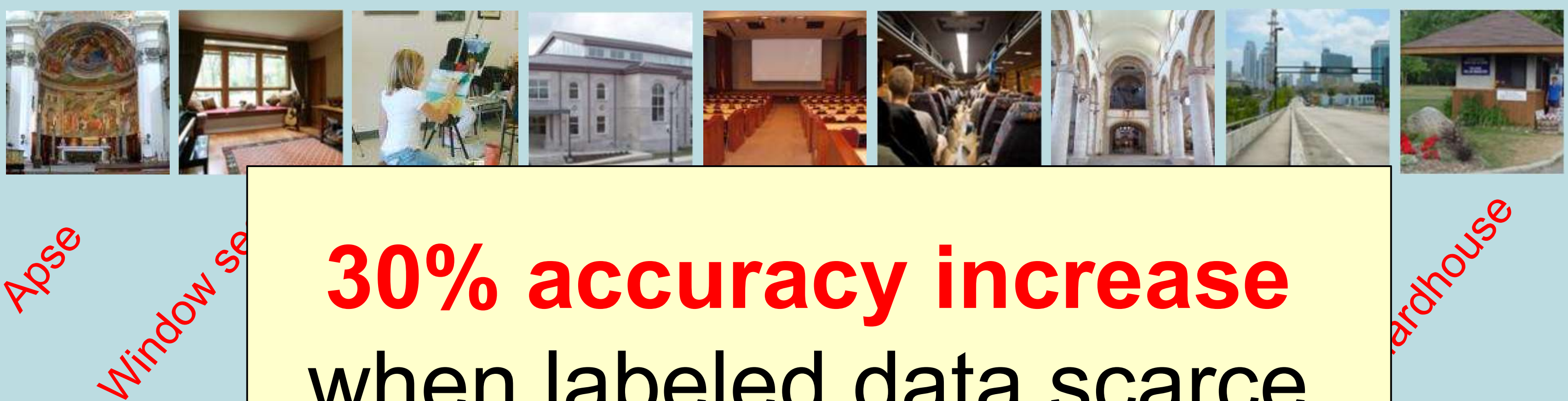
[Jayaraman & Grauman, ICCV 2015, IJCV 2017]

Impact on recognition

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

Geiger et al, IJRR '13

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

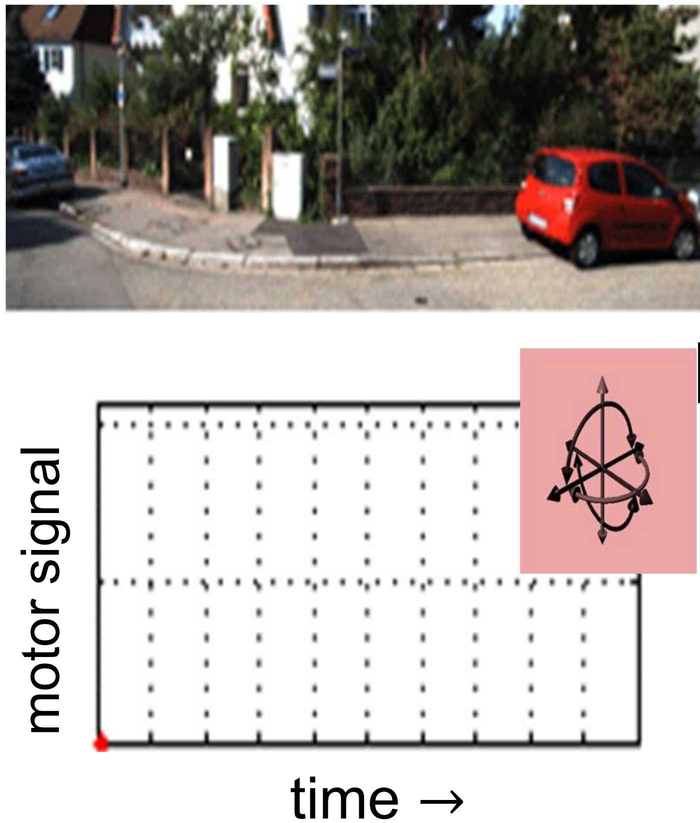


30% accuracy increase
when labeled data scarce

, CVPR '10

Passive → complete egomotions

Pre-recorded video

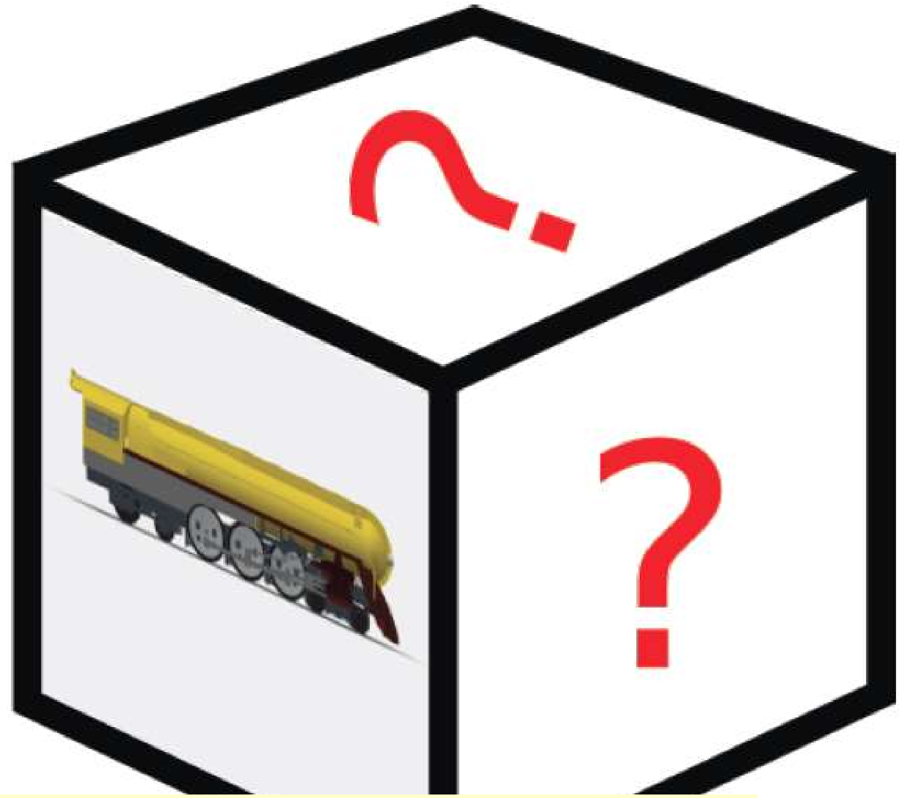


Moving around to inspect

One-shot reconstruction

Infer unseen views

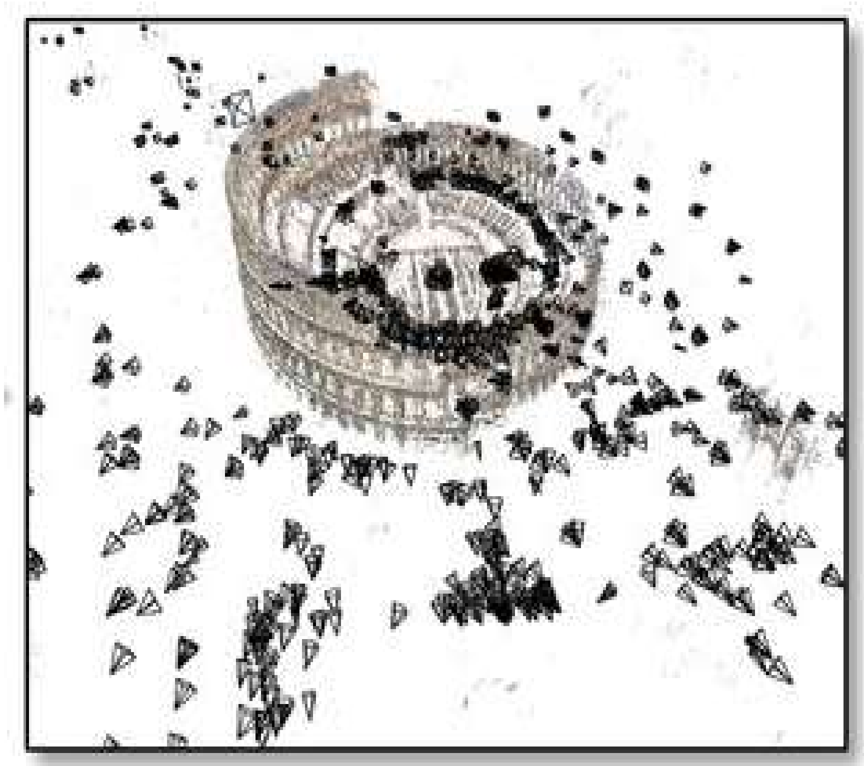
Viewgrid representation



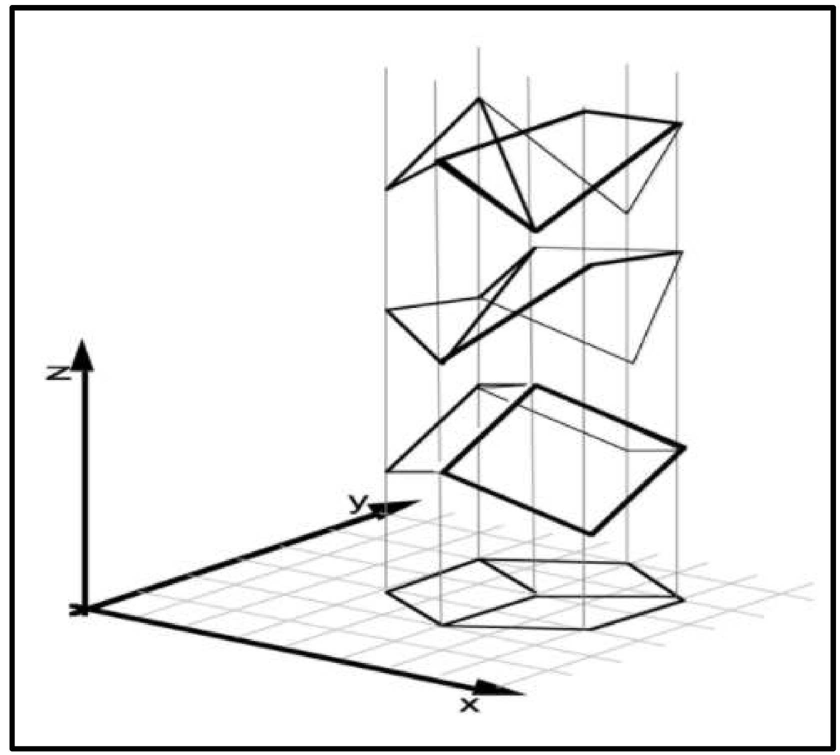
Key idea: One-shot reconstruction as a proxy task to learn semantic shape features.

[Jayaraman et al., ECCV 2018]

One-shot reconstruction



[Snavely et al, CVPR '06]



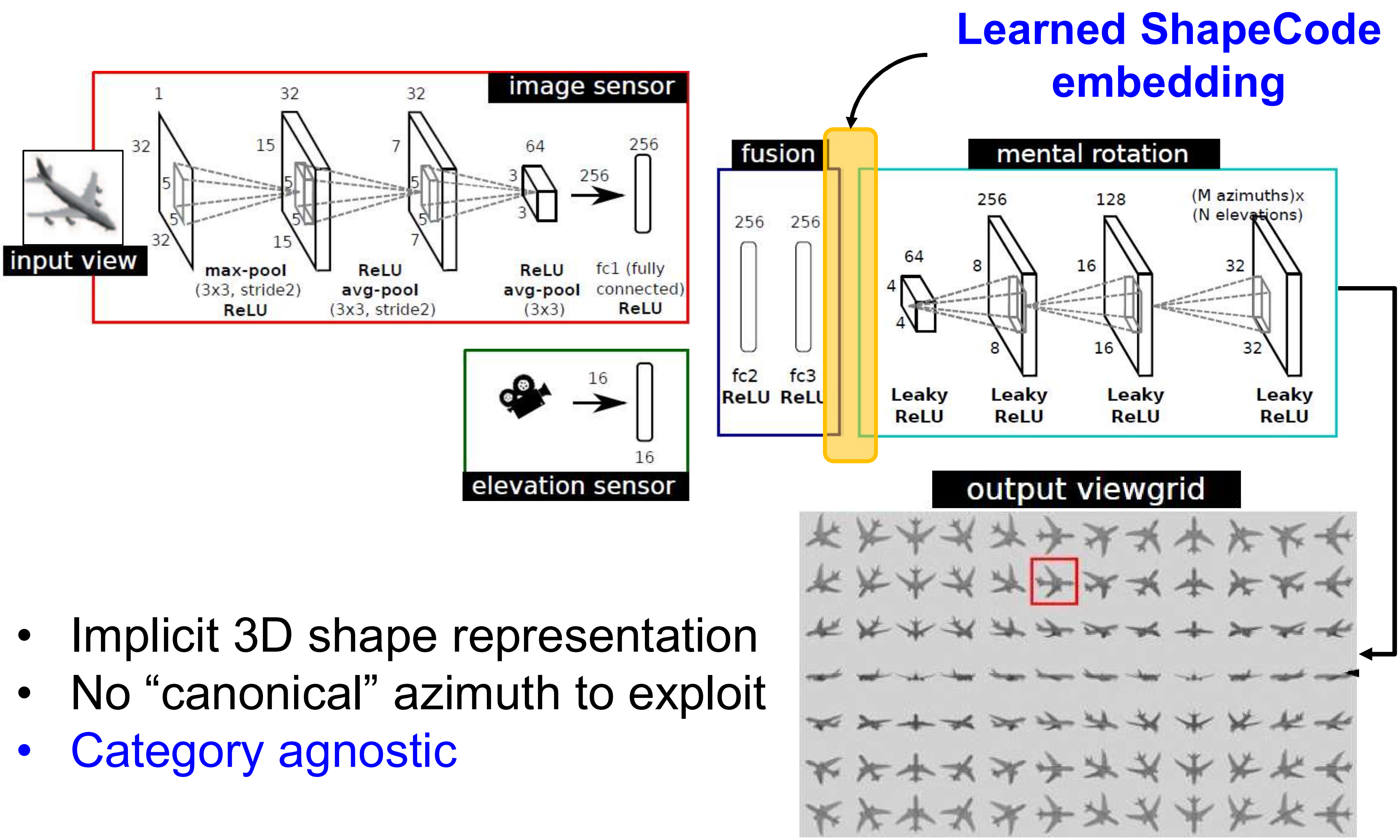
[Sinha et al, ICCV'93]

Shape from many views
geometric problem

Shape from one view
semantic problem

[Jayaraman et al., ECCV 2018]

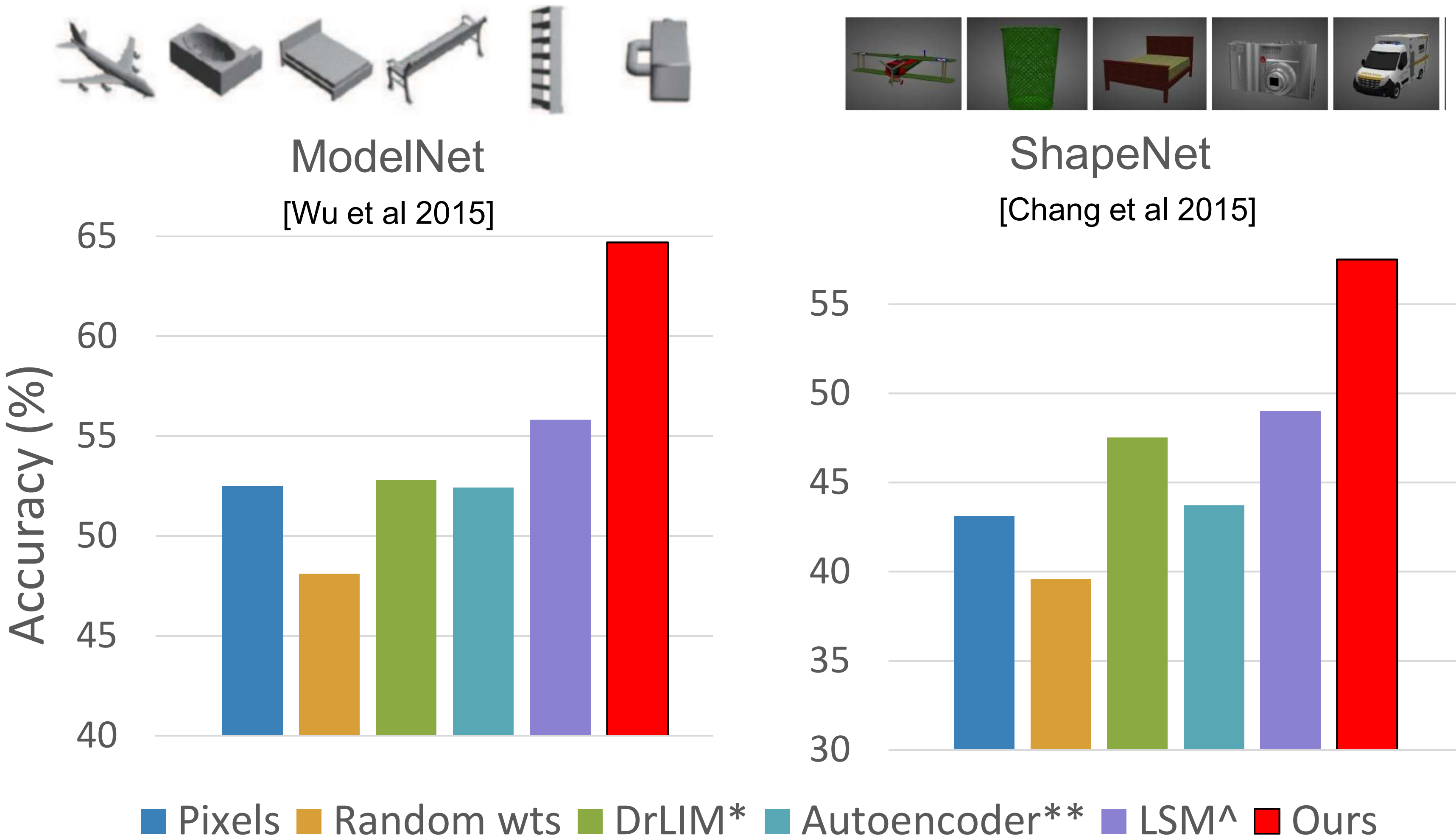
Approach: ShapeCodes



- Implicit 3D shape representation
- No "canonical" azimuth to exploit
- Category agnostic

[Jayaraman et al., ECCV 2018]

ShapeCodes for recognition



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

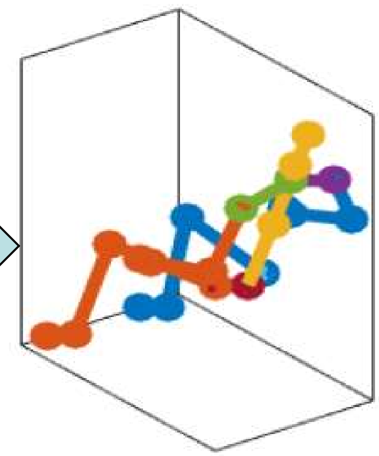
** Masci et al, Stacked Convolutional Autoencoders for Hierarchical Feature Extraction, ICANN 2011

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

Egomotion 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]

Egomotion and implied body pose

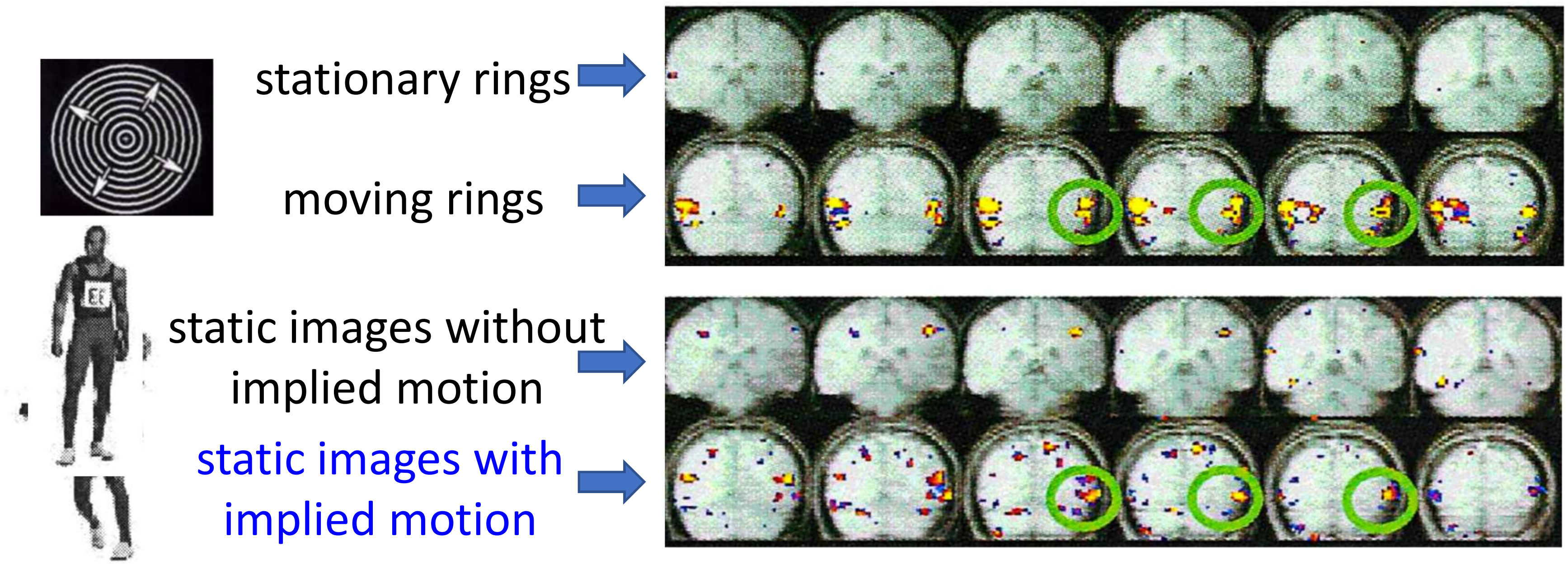
Learn relationship between egocentric scene motion and 3D human body pose

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

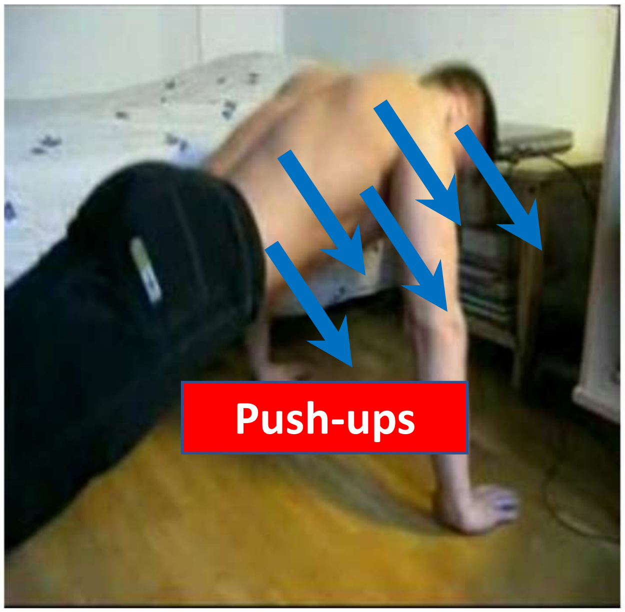
[Jiang & Grauman, CVPR 2017]

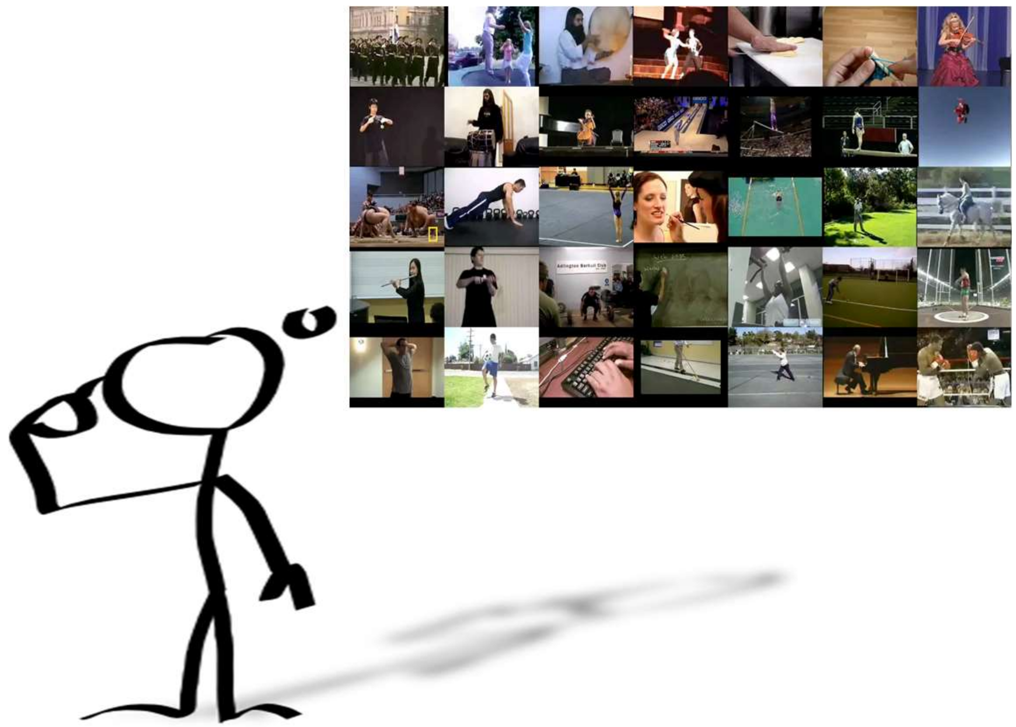
Implied motion in static images
[Kourtzi & Kanwisher, 2000]

Activation in medial temporal / medial superior temporal (MT/MST) cortex by static images with implied motion



Im2Flow:
Infer next motion in a static image





Unlabeled video as rich source of motion experience

[Gao & Grauman, CVPR 2018]

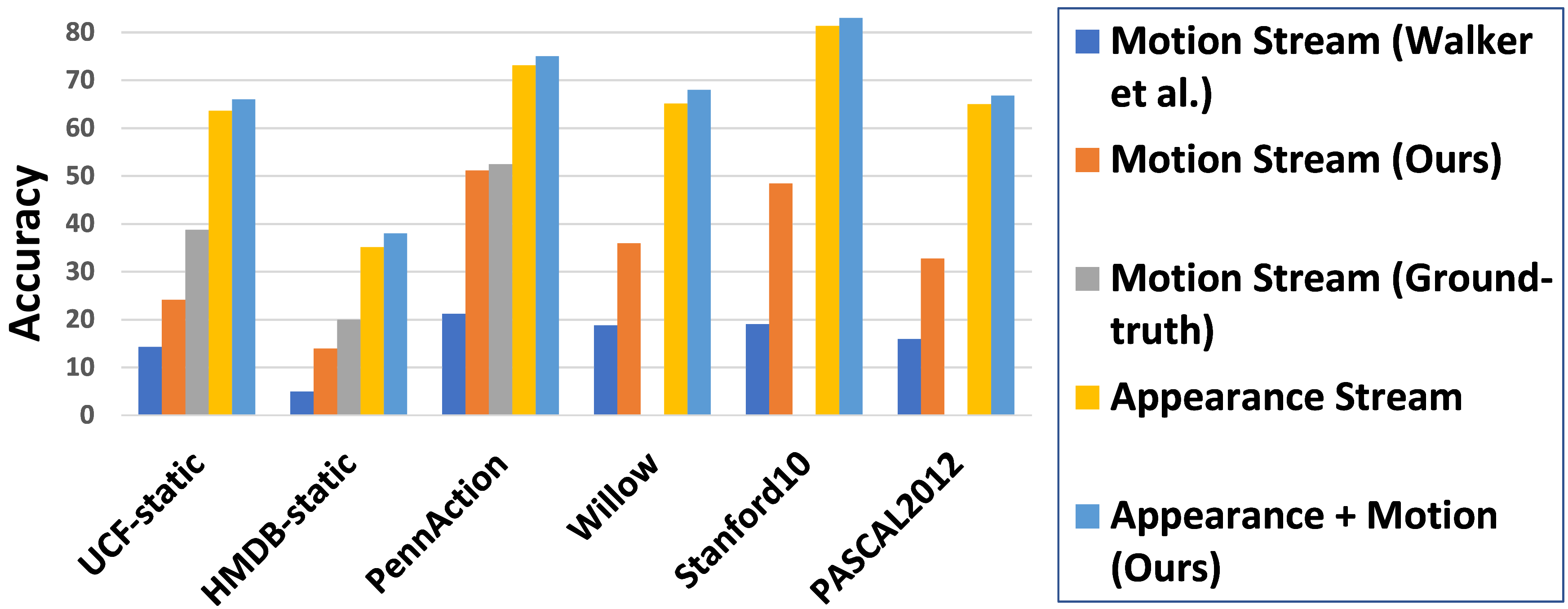
Im2Flow for "motion potential"

Identify static images that are most suggestive of motion or coming events

[Gao & Grauman, CVPR 2018]

Im2Flow for action recognition in photos

Two-stream network with RGB and inferred flow



- Inferred motion from Im2Flow framework boosts recognition
- Up to 6% relative gain vs. appearance stream alone

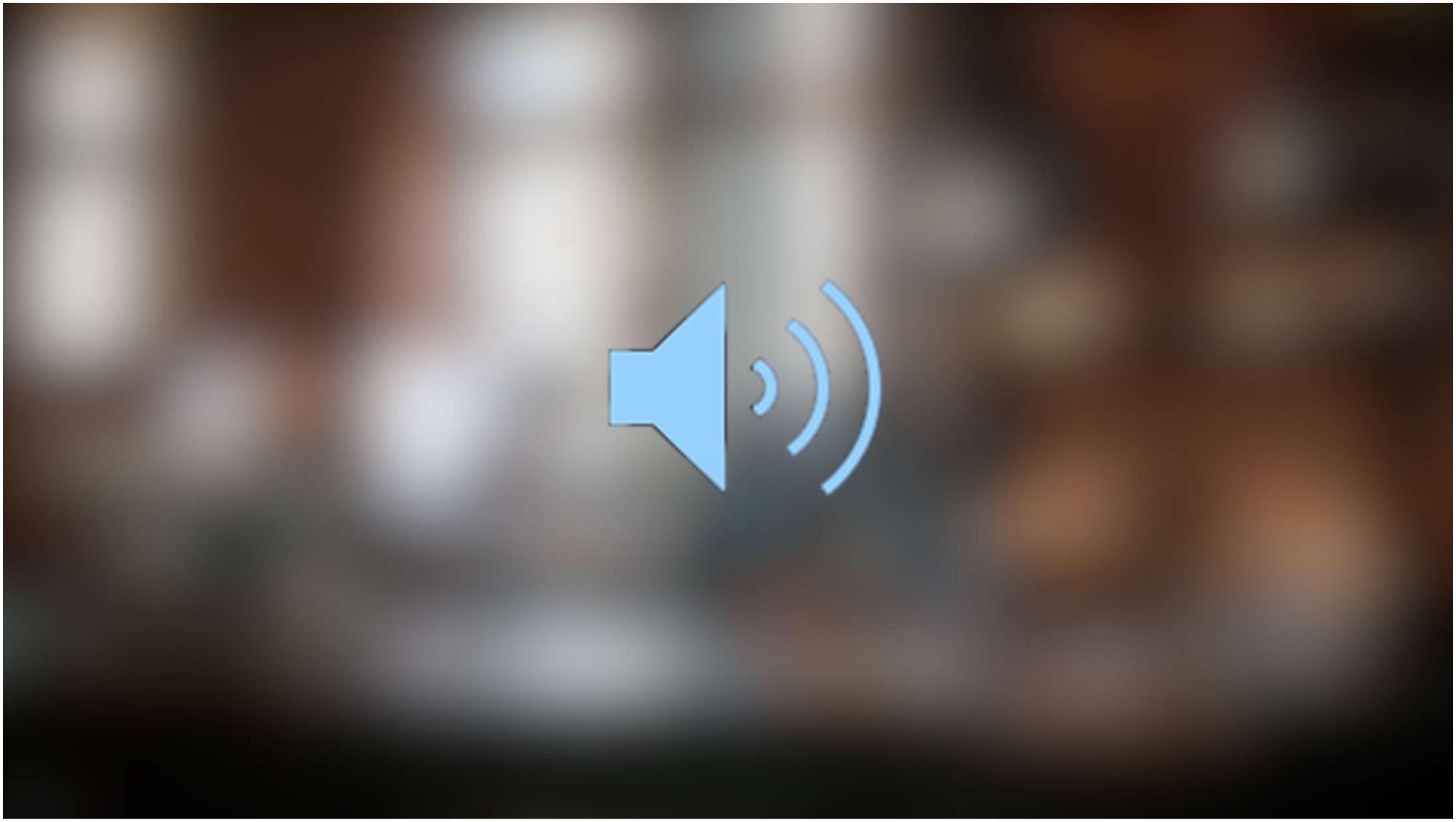
[Gao & Grauman, CVPR 2018]

Recall: Disembodied visual learning

...

...

Listening to learn



Listening to learn

Listening to learn



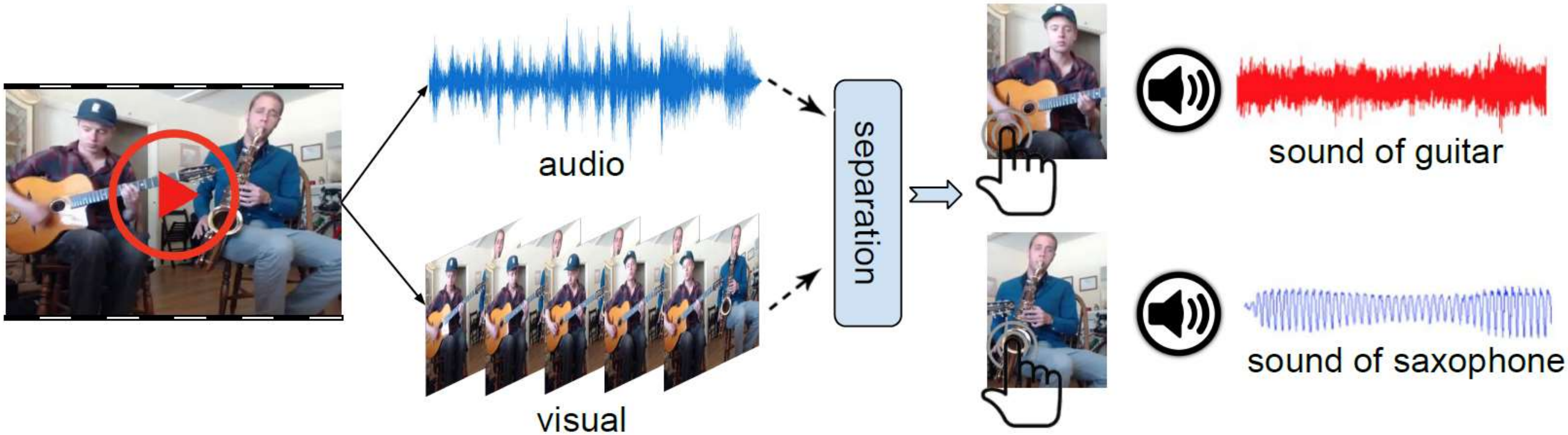

woof **meow** **ring** **clatter**

Goal: a repertoire of objects and their sounds

Challenge: a single audio channel mixes sounds of multiple objects

Kristen Grauman

Visually-guided audio source separation



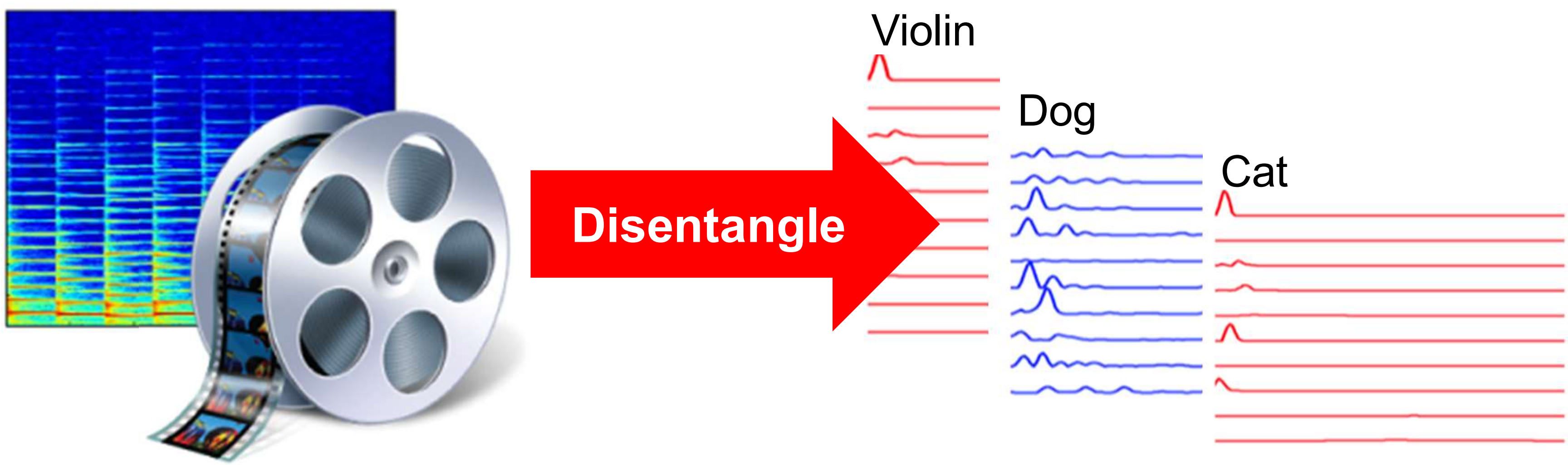
Traditional approach:

- 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



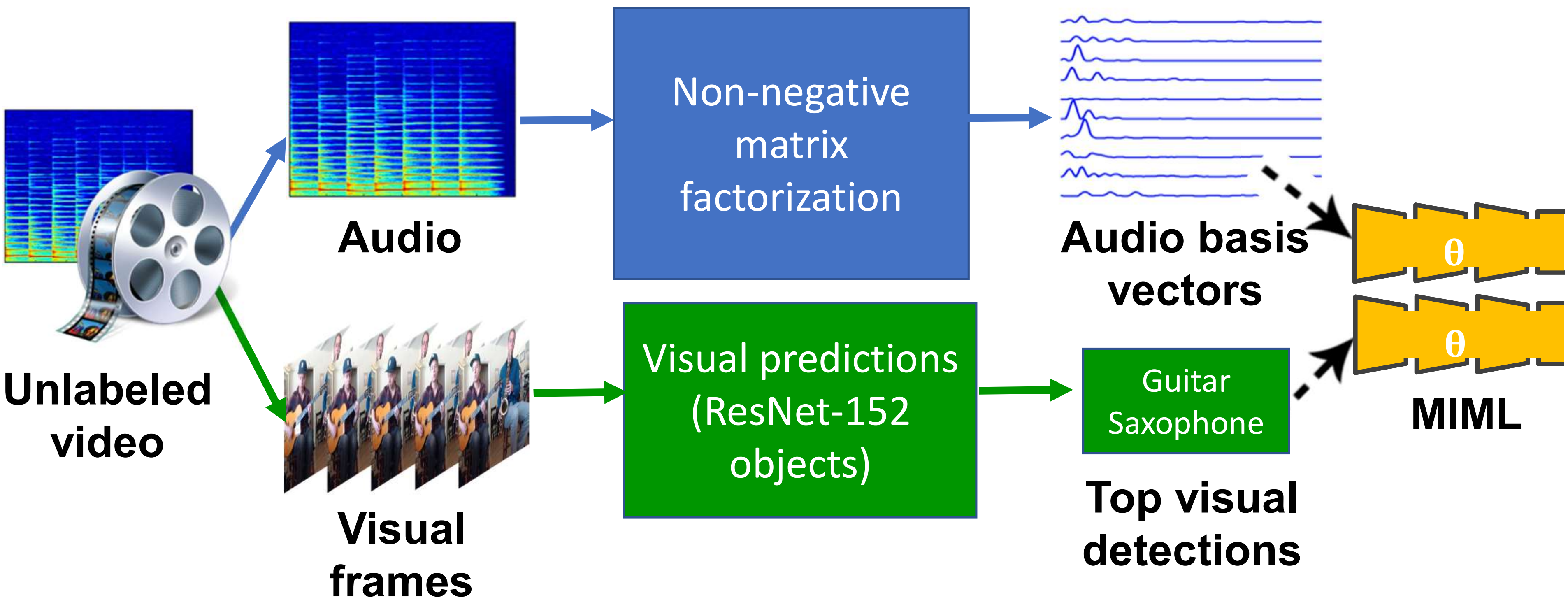
Unlabeled video **Object sound models**

Kristen Grauman

[Gao, Feris, & Grauman, ECCV 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**

[Gao, Feris, & Grauman, ECCV 2018]

Our approach: learning

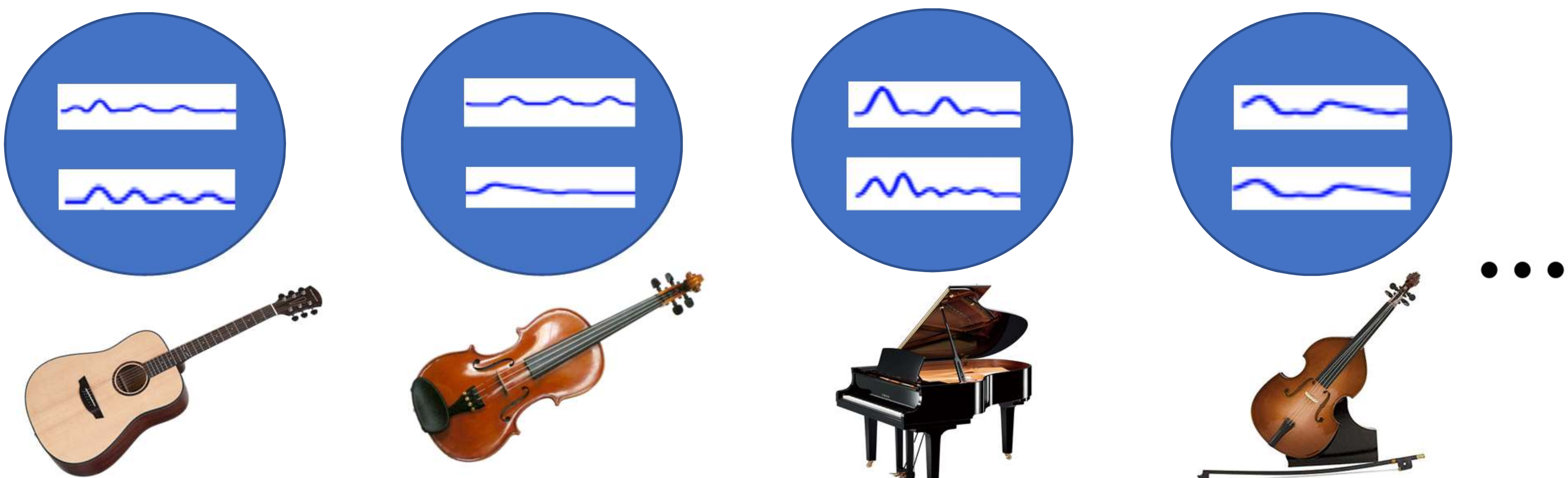
MIML detangles sounds via visually detected objects

Guitar + Violin

Guitar + Piano

Cello + Piano

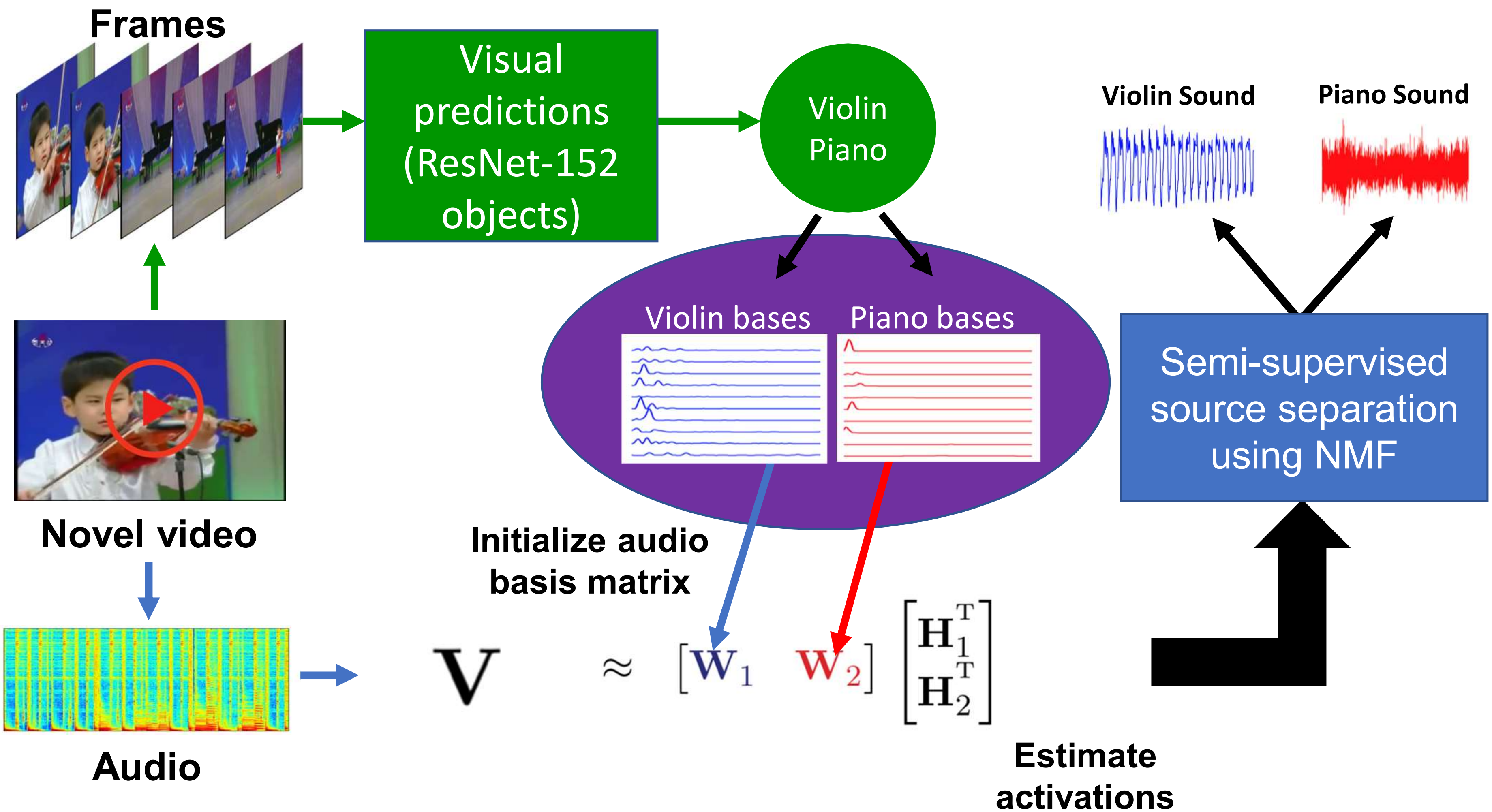
Audio Bases



[Gao, Feris, & Grauman, ECCV 2018]

Our approach: inference

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



Results

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, ECCV 2018]

Results

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, ECCV 2018]

Results

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, ECCV 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

Towards embodied visual learning

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

Active perception

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, …

Kristen Grauman

End-to-end active recognition

Predicted label:

T=1 T=2 T=3

[Jayaraman and Grauman, ECCV 2016, PAMI 2018]

Goal: Learn to "look around"

recognition

reconnaissance

search and rescue

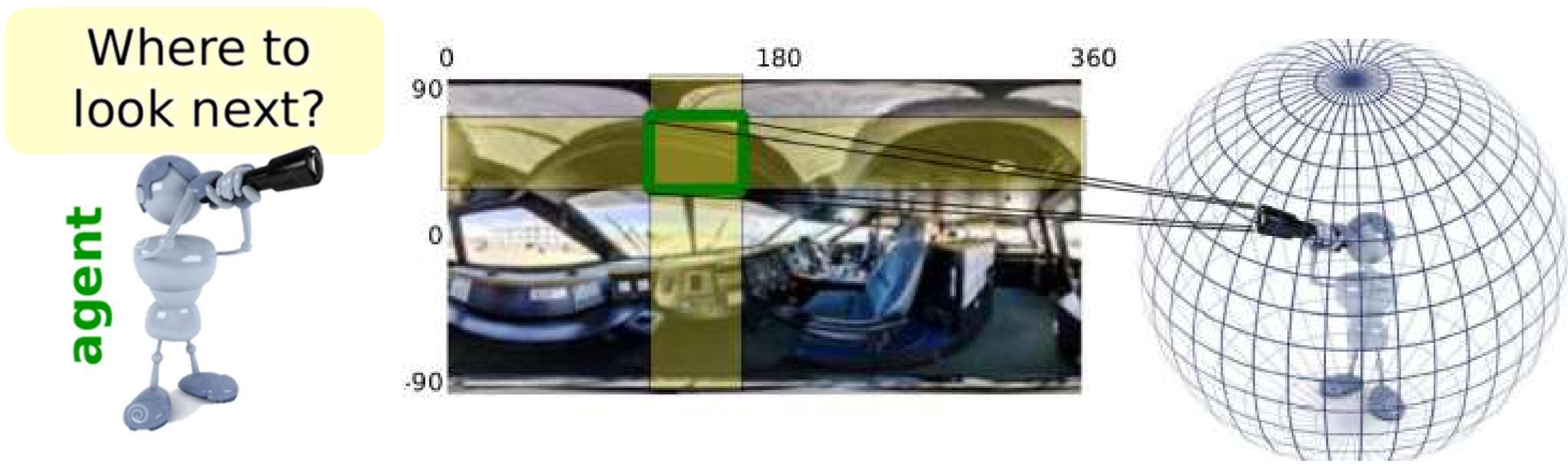
task predefined

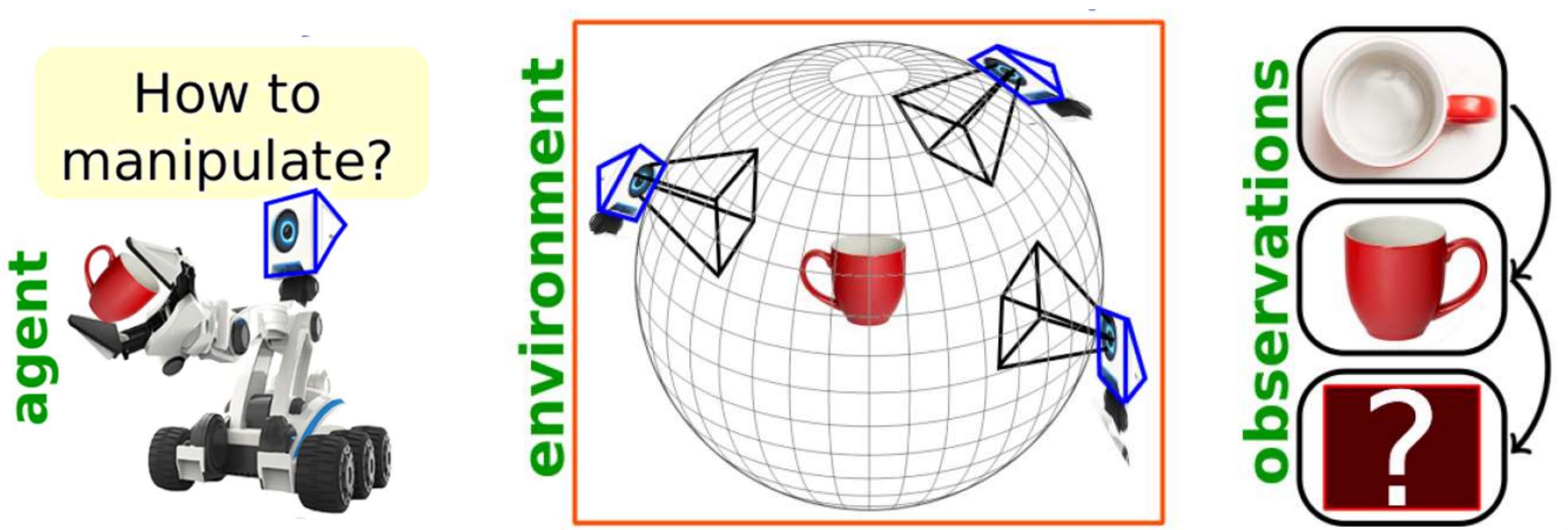
task unfolds dynamically

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

Kristen Grauman

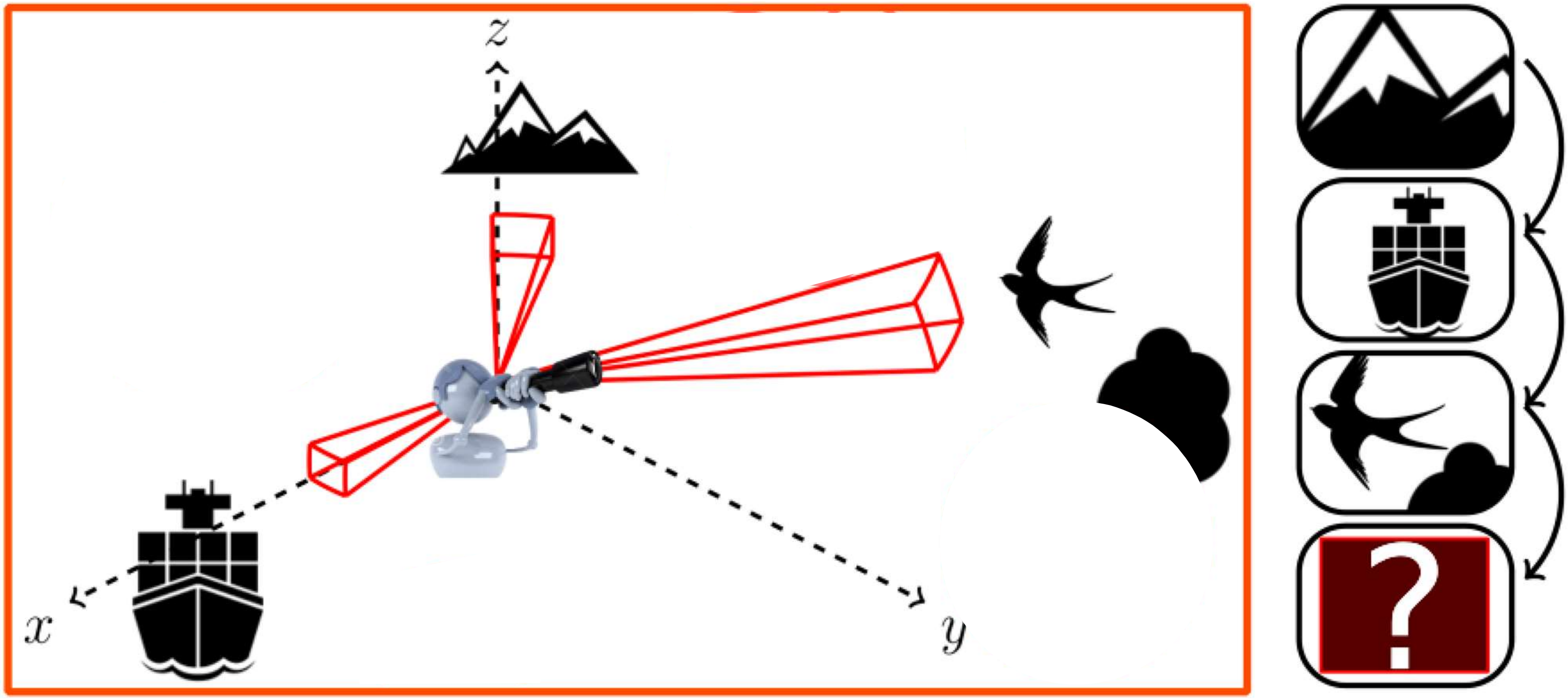
Two scenarios





Key idea: Active observation completion

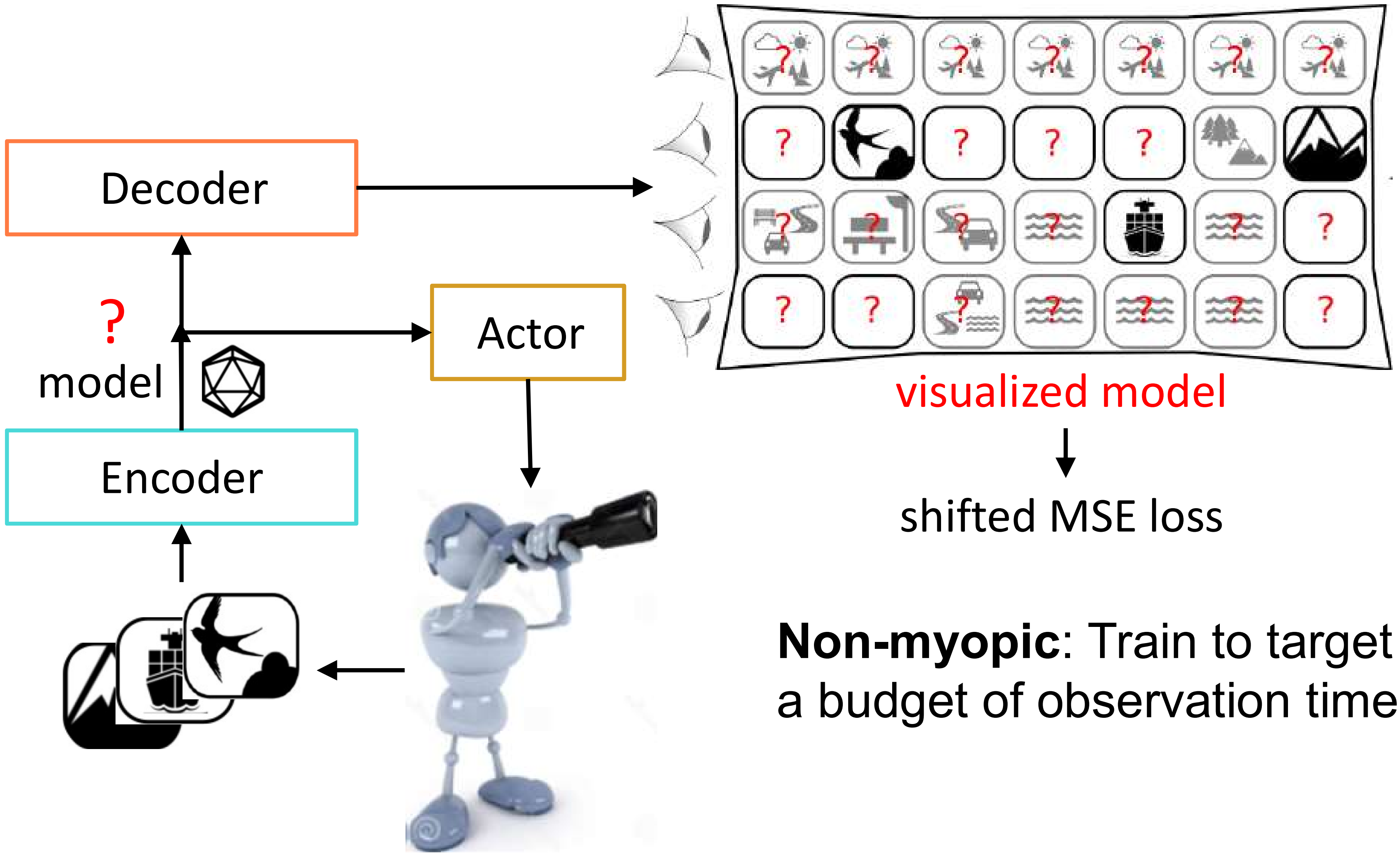
Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment



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

Jayaraman and Grauman, CVPR 2018

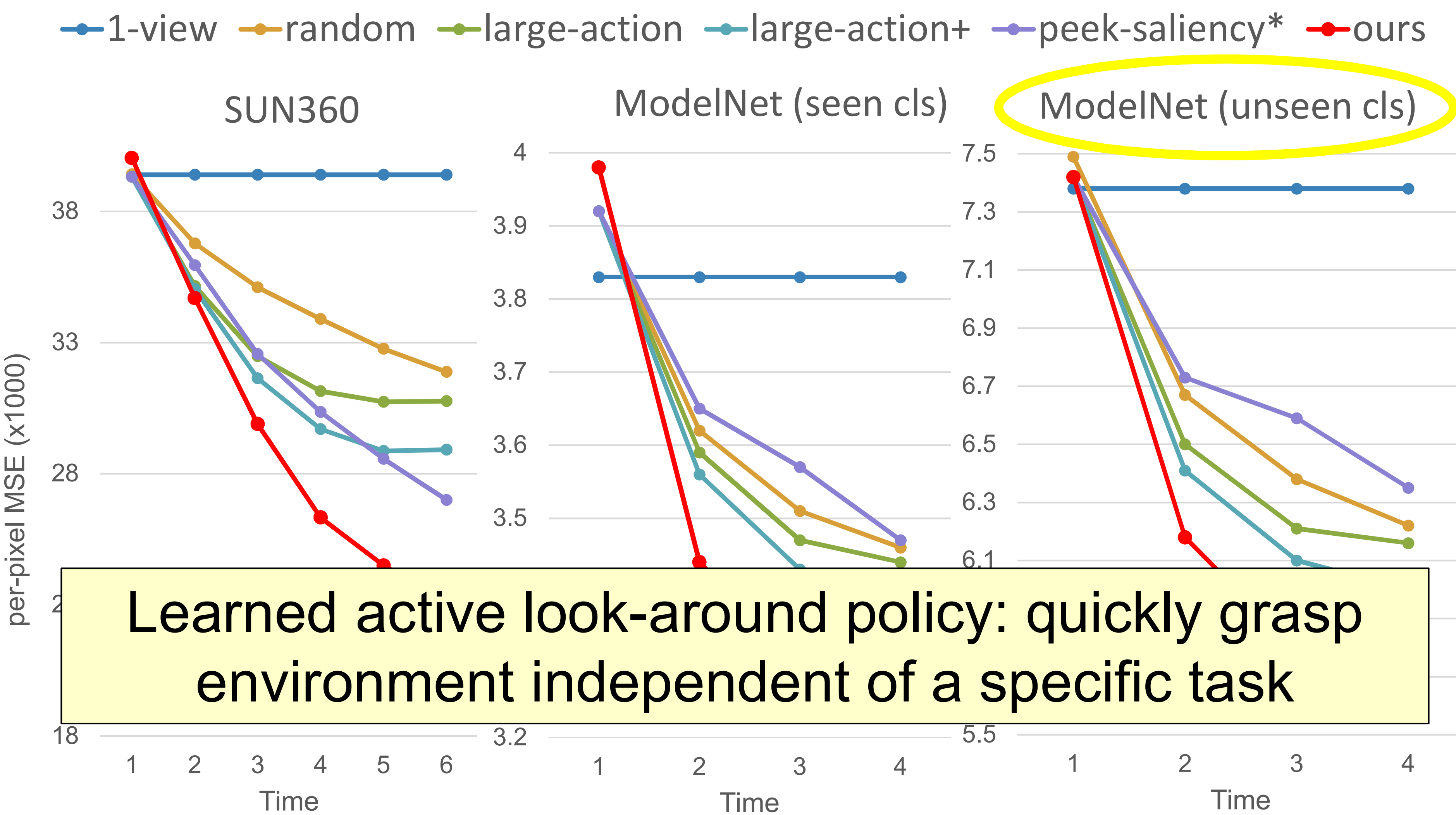
Approach: Active observation completion



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

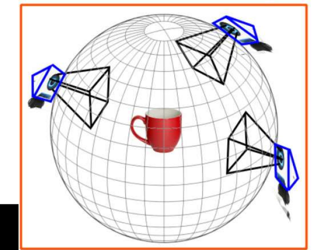
Jayaraman and Grauman, CVPR 2018

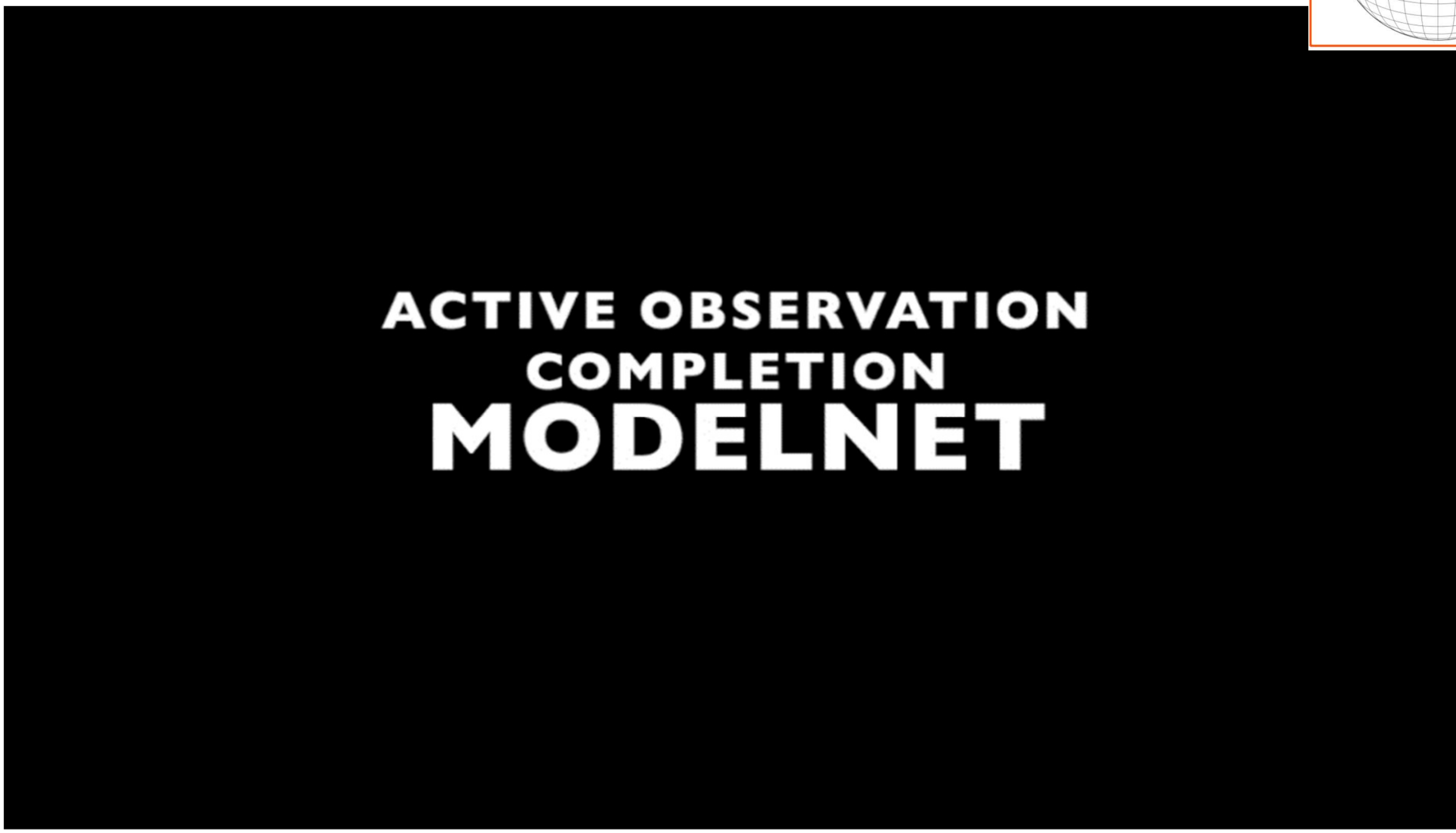
Active “look around” results



*Saliency -- 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

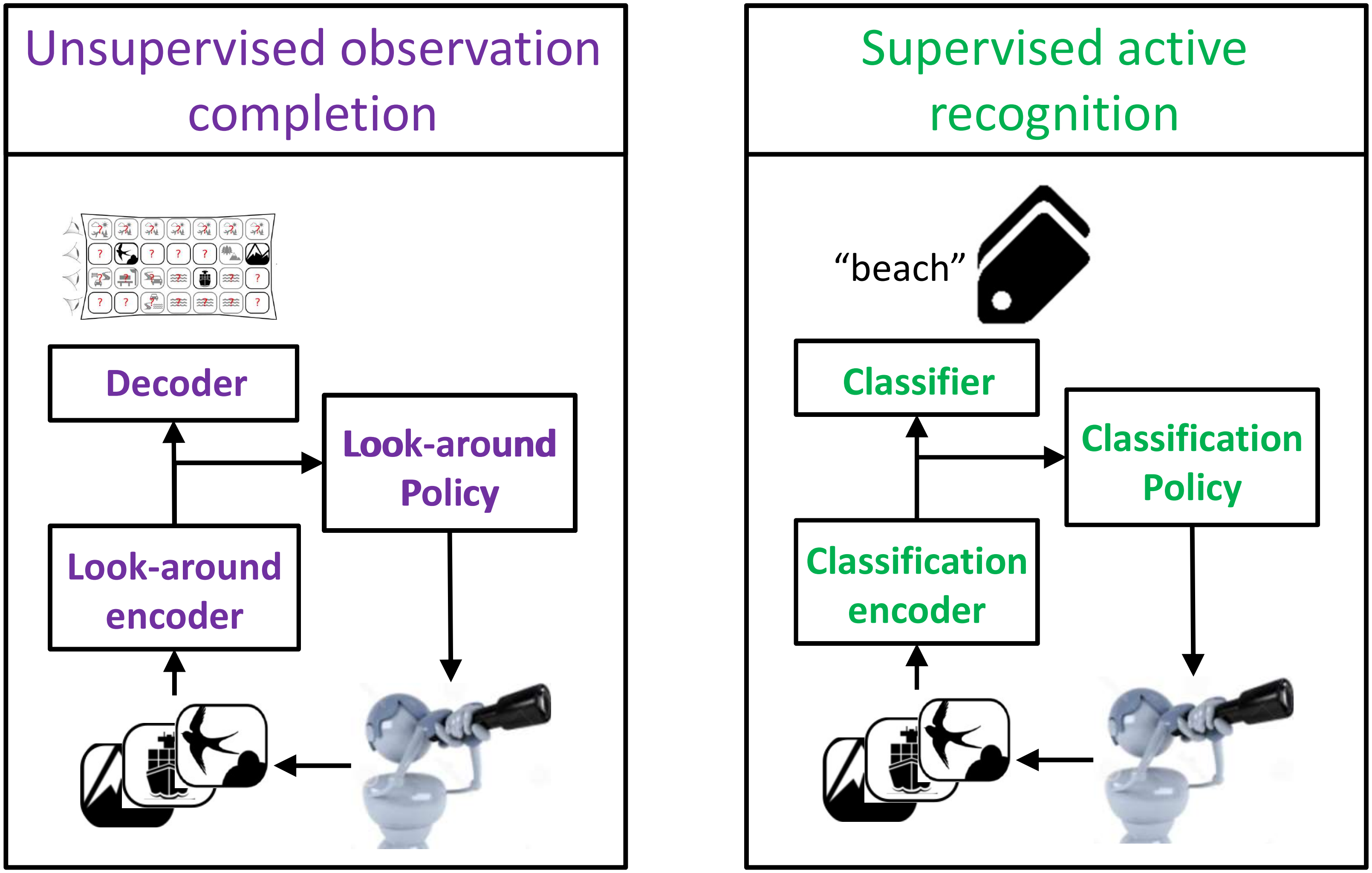
Jayaraman and Grauman, CVPR 2018; Ramakrishnan & Grauman, ECCV 2018

Active "look around" visualization

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

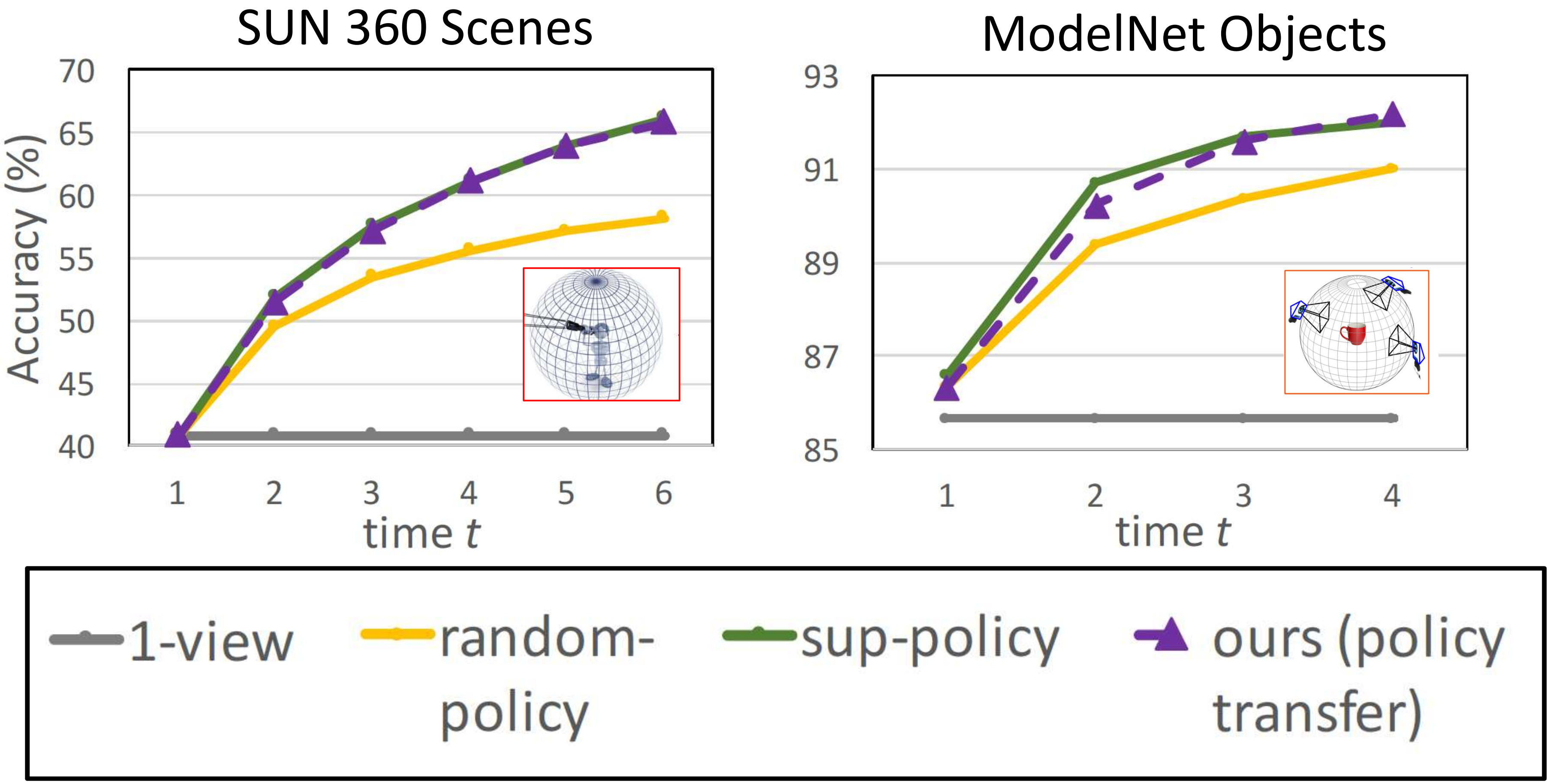
Jayaraman and Grauman, CVPR 2018; Ramakrishnan & Grauman, ECCV 2018

Egomotion policy transfer



Plug observation completion policy in for new task

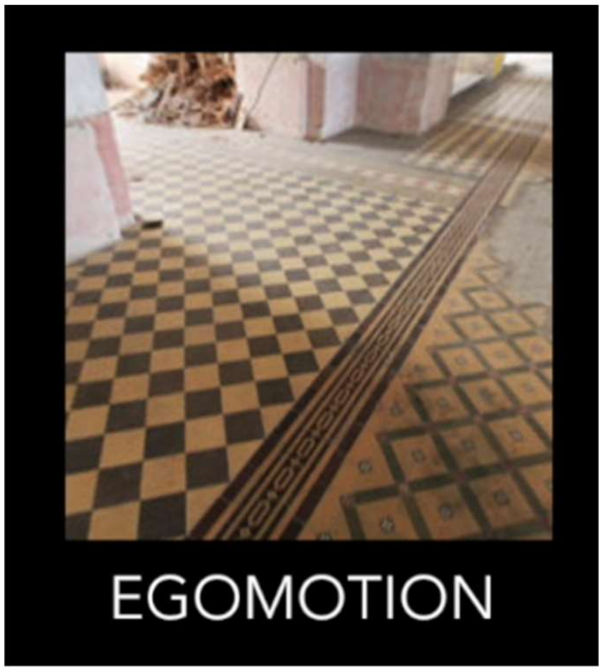
Egomotion policy transfer



Unsupervised exploratory policy approaches supervised task-specific policy accuracy!

Jayaraman and Grauman, CVPR 2018

Challenge: Motion policy learning with partial observability

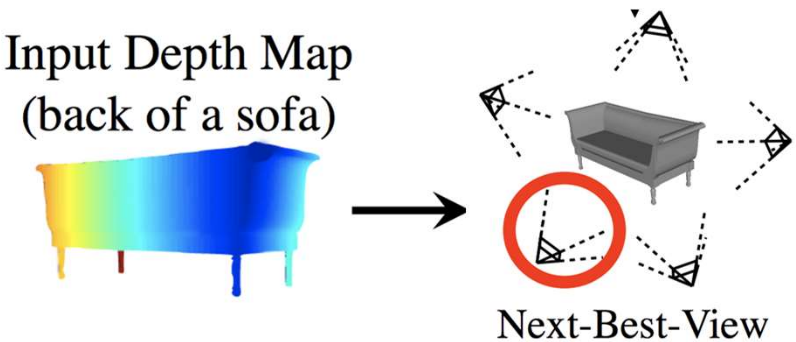


Exploration with limited observability impedes policy learning

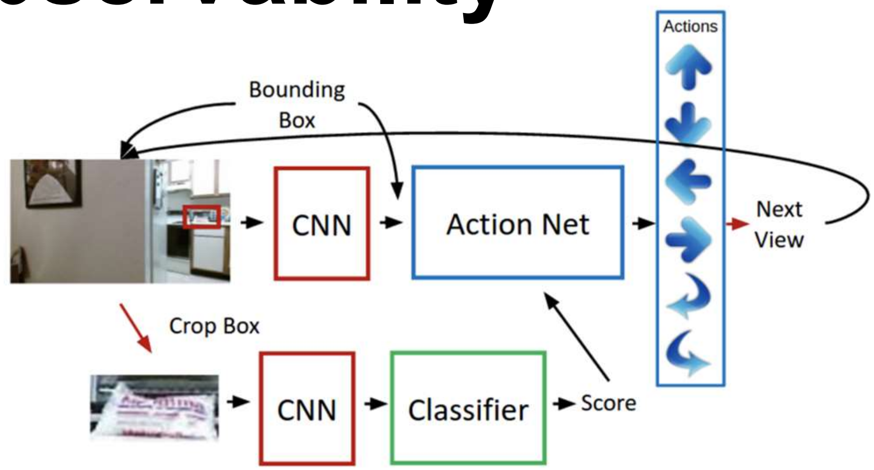
Yet during training, full state may be available

Kristen Grauman

Challenge: Motion policy learning with partial observability

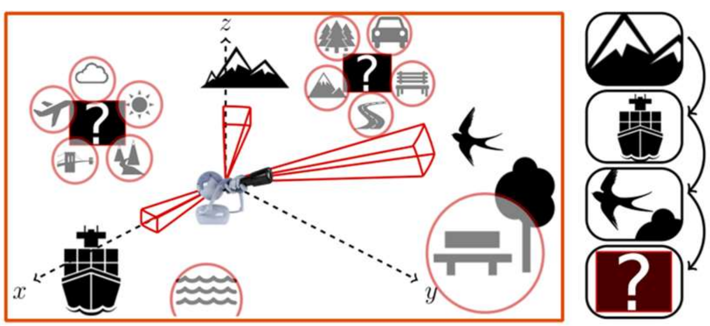


Wu et al., 2015



Ammirato et al., 2017

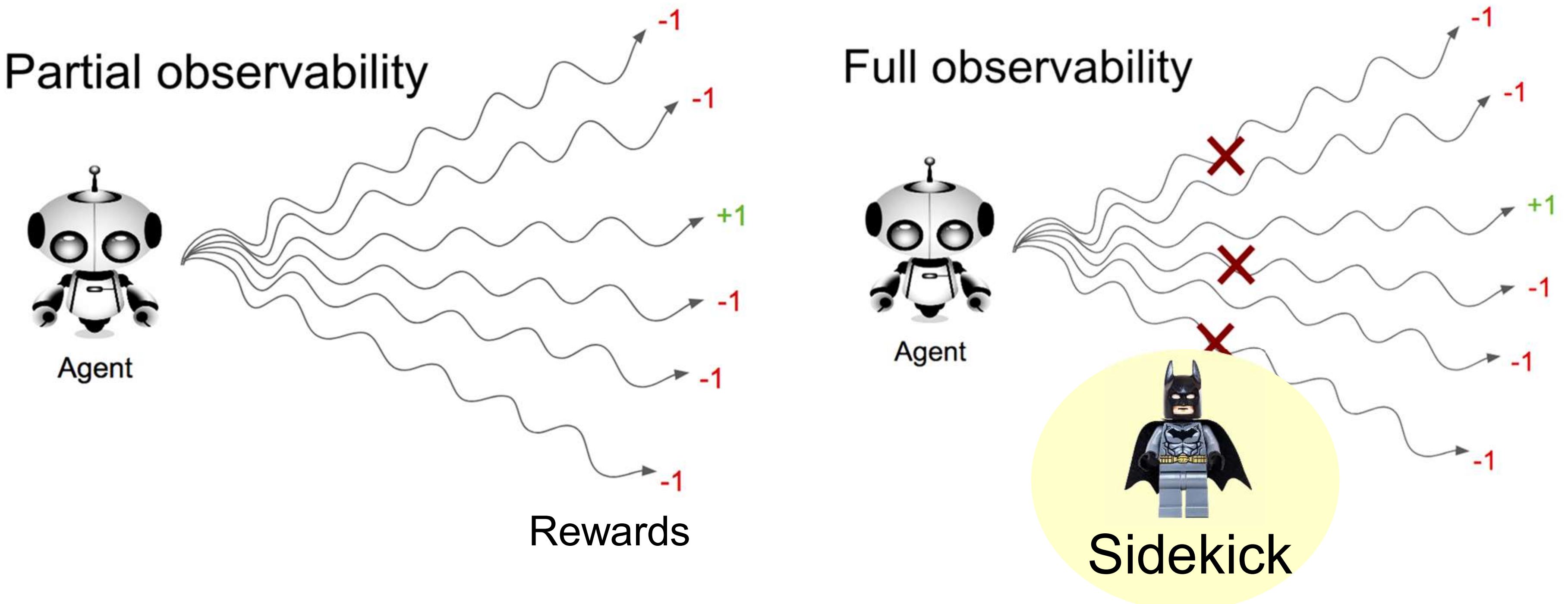
Status quo: ignore full observability available at training time



Jayaraman and Grauman., 2018

Jayaraman and Grauman., 2016

Idea: Sidekick policy learning



Sidekick agent with full observability guides policy towards valuable states during training

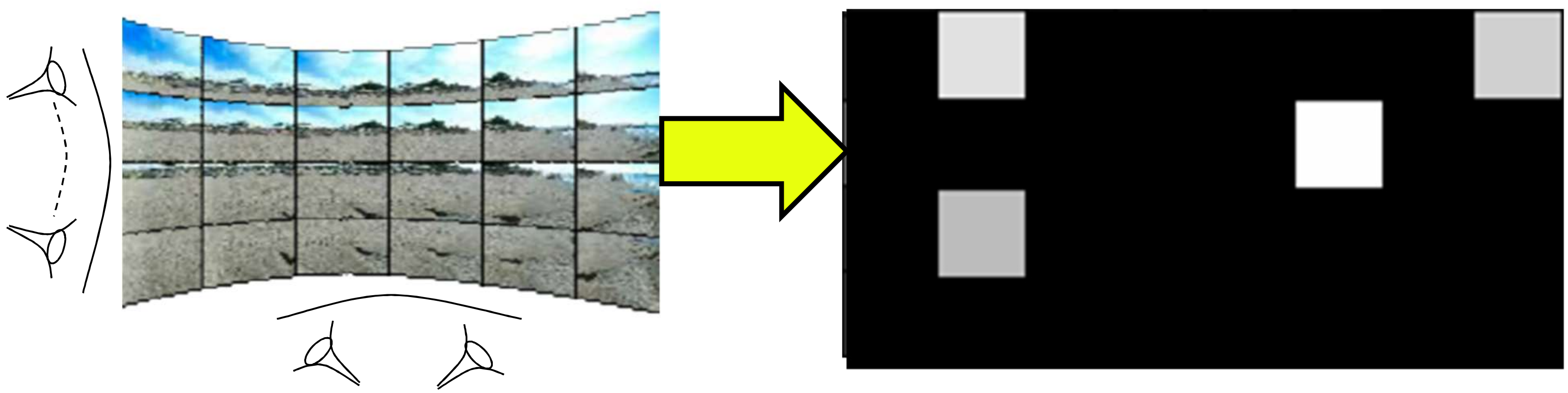
Ramakrishnan & Grauman, ECCV 2018

1) Reward-based sidekick

Preview and transfer knowledge of environment

360 environment - *X*

Identify informative views



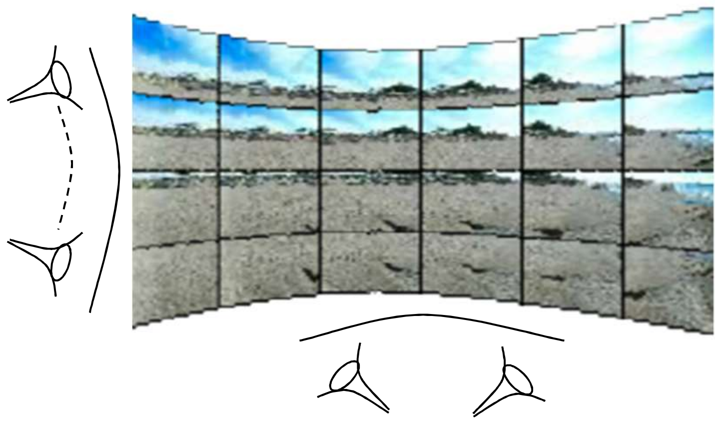
Shape reward function

Ramakrishnan & Grauman, ECCV 2018

2) Demonstration-based sidekick

Generate information-gathering trajectories to initially supervise policy learning

360 environment - X



Selected views

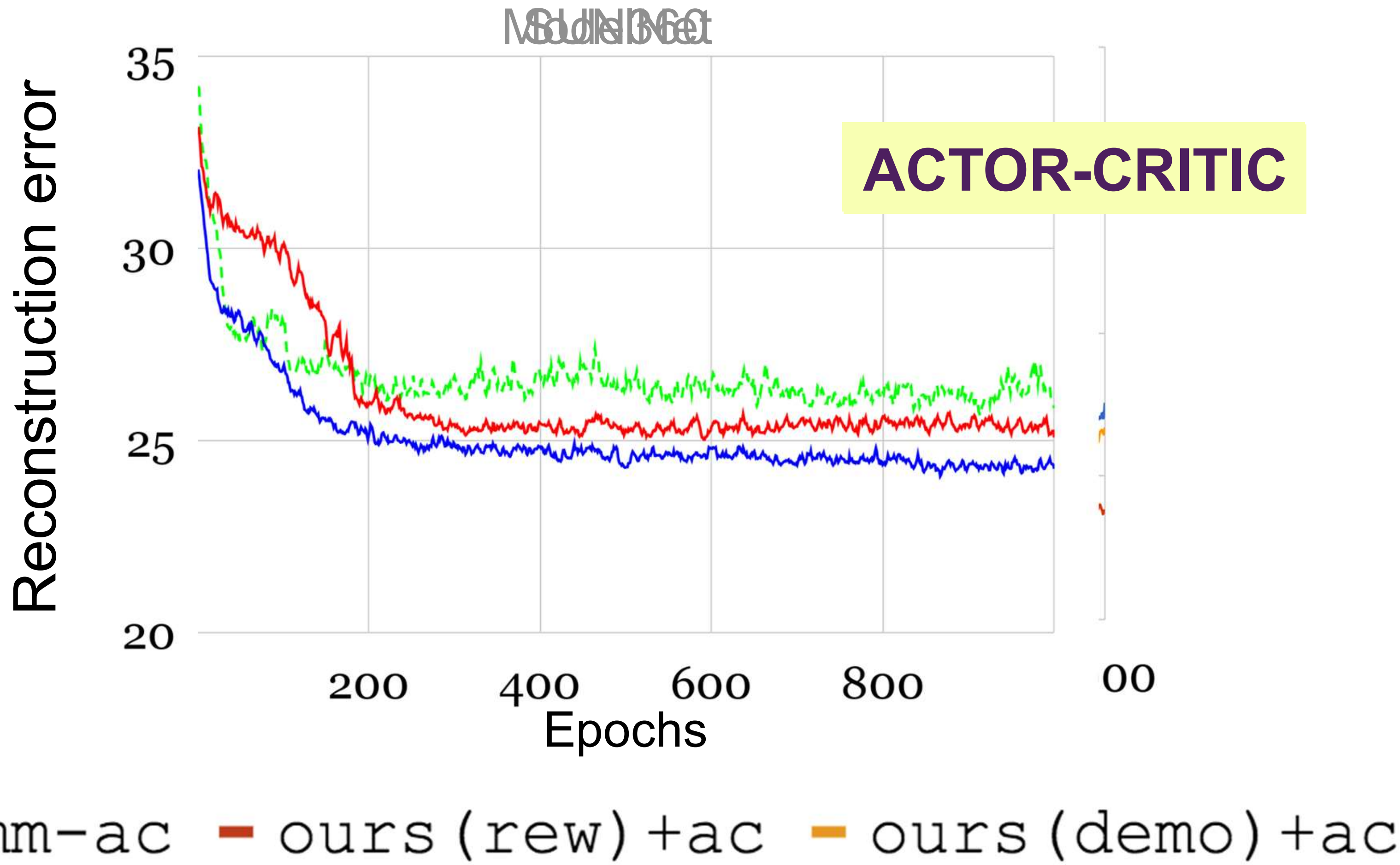
Current view

Cumulative information

Ramakrishnan & Grauman, ECCV 2018

Sidekick results

Accelerate training and obtain better policies



ltla: Jayaraman & Grauman, Learning to look around, CVPR 2018
asymm-ac: Pinto et al. Asymmetric actor-critic, RSS 2018

Summary

- Visual learning benefits from
  - context of action and multiple senses
  - continuous unsupervised observations
- Key ideas:
  - Embodied feature learning via multi-sensory signals
  - Active policies for view selection and camera control


Ruohan Gao


Santhosh Ramakrishnan


Dinesh Jayaraman


Rogerio Feris

Papers/code/videos

- **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]
- **ShapeCodes: Self-Supervised Feature Learning by Lifting Views to Viewgrids**. D. Jayaraman, R. Gao, and K. Grauman. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sept 2018. [pdf]
- **Sidekick Policy Learning for Active Visual Exploration**. S. Ramakrishnan and K. Grauman. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sept 2018. [pdf] [supp] [videos/code]
- **End-to-end Policy Learning for Active Visual Categoriza**tion. D. Jayaraman and K. Grauman. To appear, Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2018. [pdf]
- **Im2Flow: Motion Hallucination from Static Images for Action Recognition**. R. Gao, B. Xiong, and K. Grauman. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, June 2018. (Oral) [pdf] [code] [project page]
- **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]