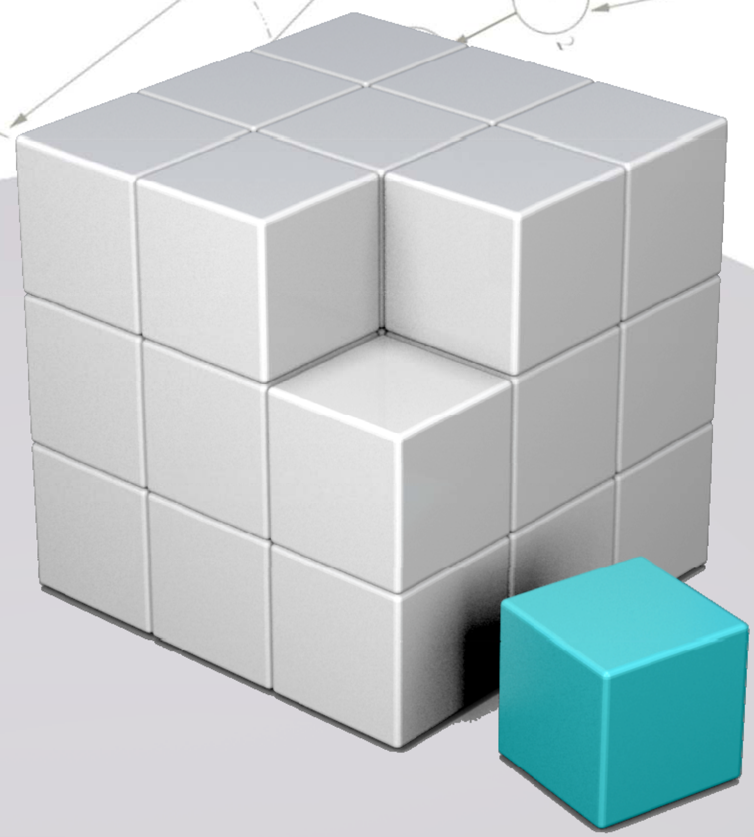
University of Texas at Austin
Visual Recognition



Presentation (paper review)

Learning Image Representations Tied to Ego-motion

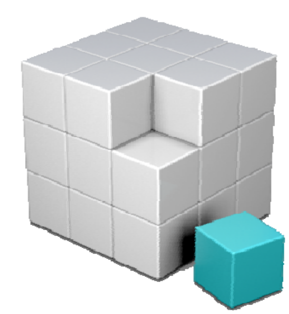
Jayaraman and Grauman. ICCV 2015

Hilgad Montelo

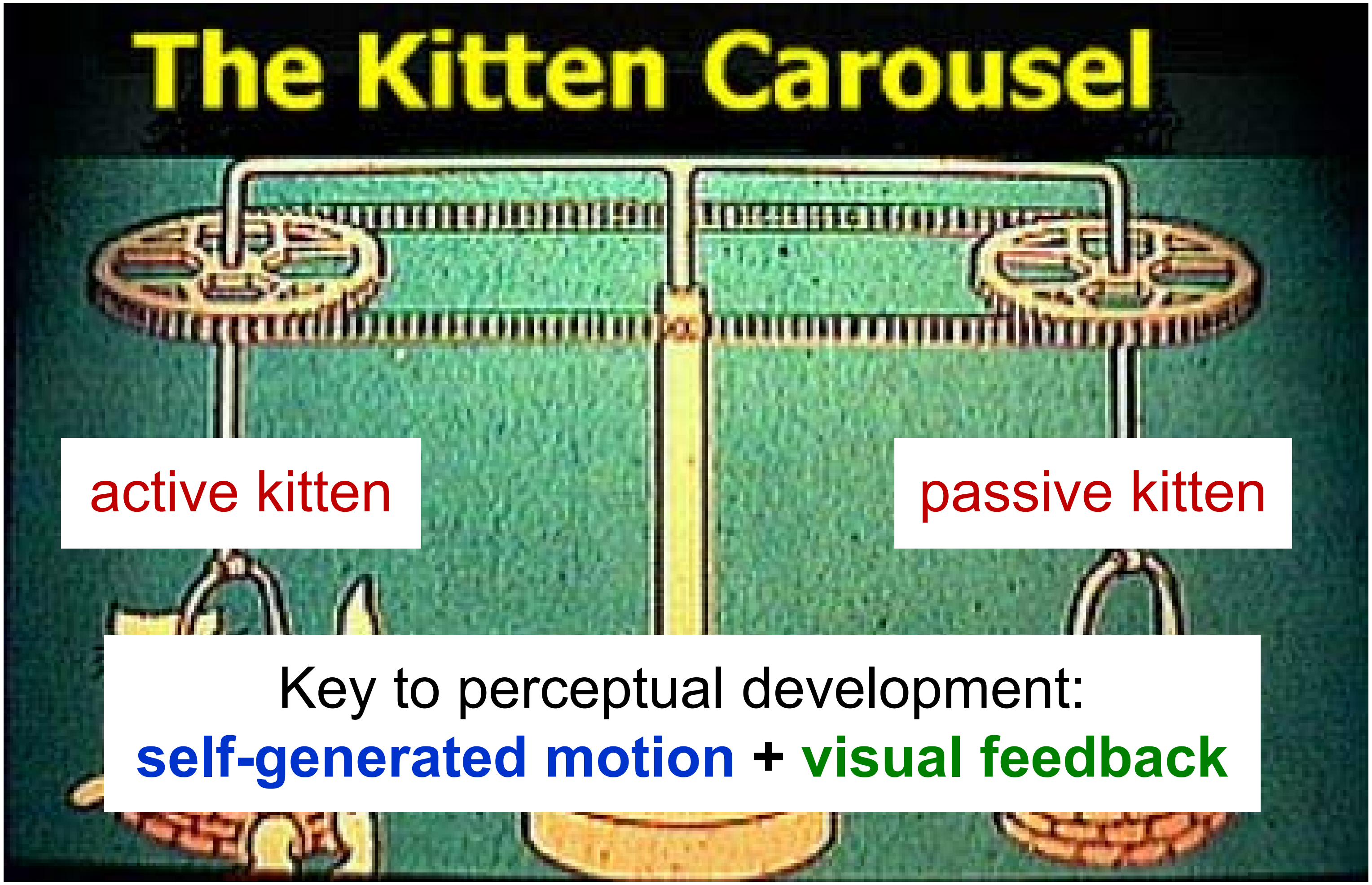
2016 March

Outline

- The "Kitten Carousel" Experiment
- Problem
- Objective
- Main Idea
- Related Work
- Approach
- Experiments and Results
- Conclusions



The "Kitten Carousel" Experiment
(Held & Hein, 1963)



[Slide credit: Dinesh Jayaraman]

Problem

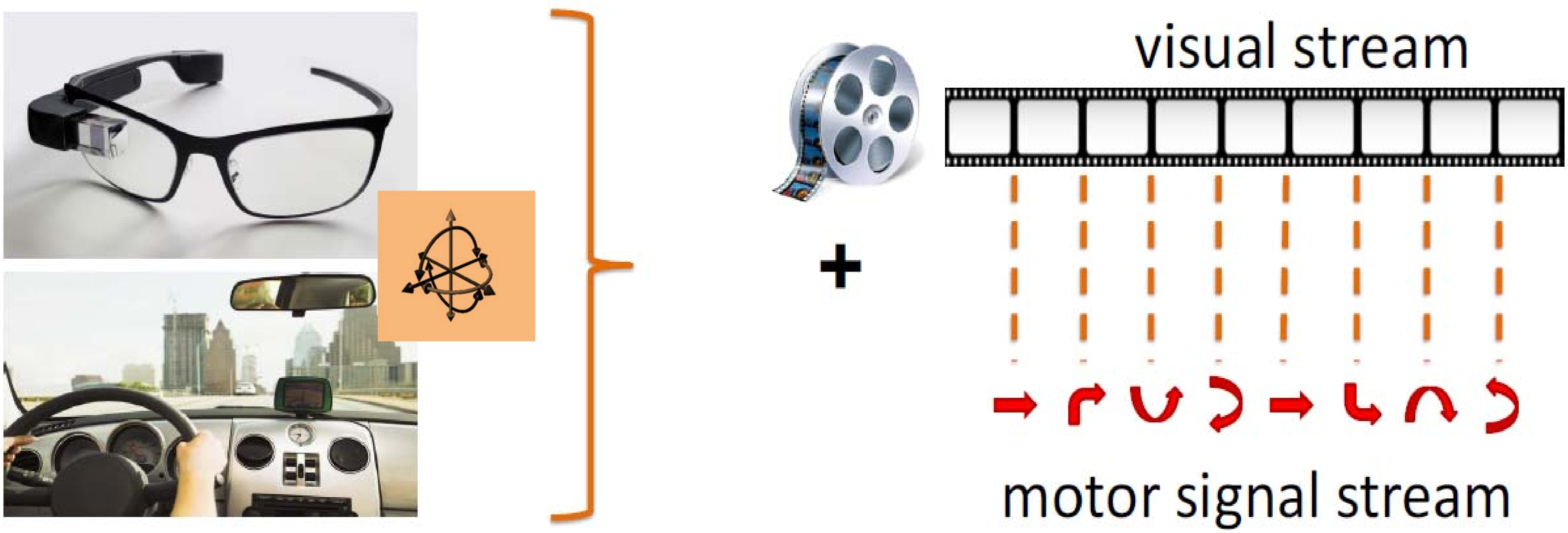
- Today's visual recognition algorithms learn from "disembodied" bag of labeled snapshots.

Objective

- Provide visual recognition algorithm that learns in the context of **acting** and **moving** in the world.

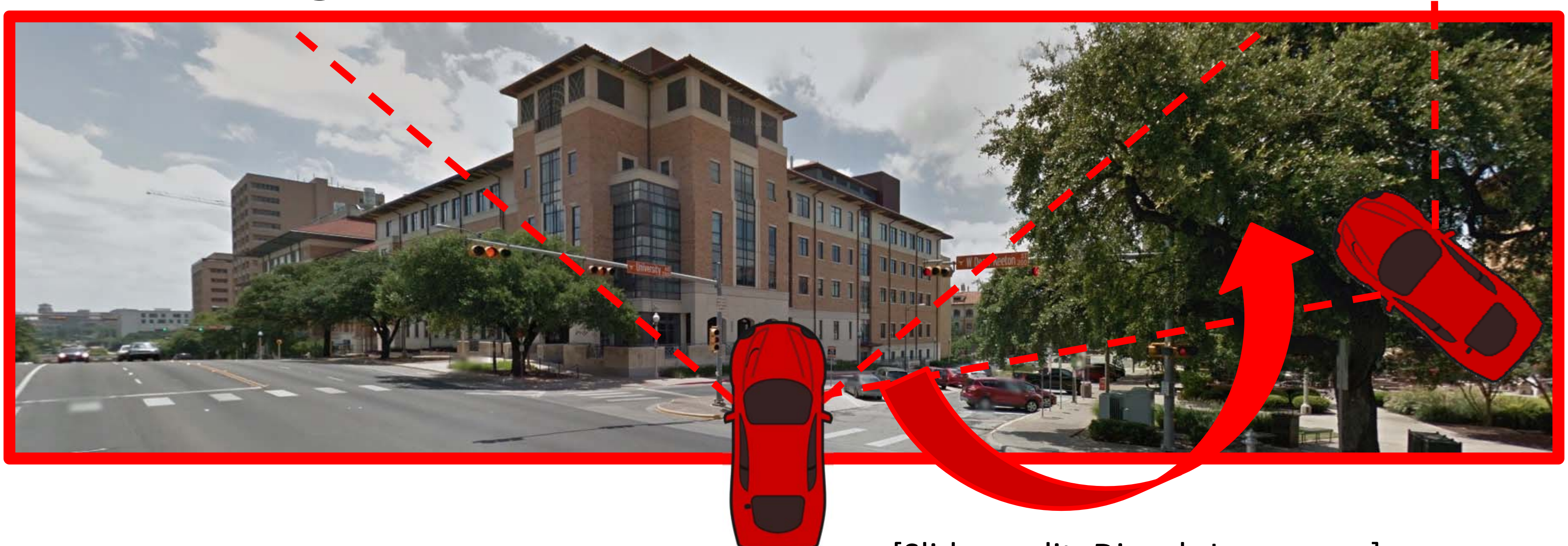
Main Idea

- Associate Ego-Motion and vision by teaching computer vision system the connection:
  - “how I move” ↔ “how my visual surroundings change”



Ego-motion ↔ vision: view prediction

After moving:



[Slide credit: Dinesh Jayaraman]

Ego-motion ↔ vision for recognition

- Learning this connection requires:

  - Depth, 3D geometry
  - Semantics
  - Context

  Also key to recognition!

- Can be learned without manual labels!

Approach: unsupervised feature learning using egocentric video + motor signals

Related Works

Integrating vision and motion

Agrawal, Carreira, Malik, “Learning to see by moving”, ICCV 2015
Watter, Springenberg, Boedecker, Riedmiller, “Embed to control...”, NIPS 2015
Levine, Finn, Darrell, Abbeel, “… visuomotor policies”, arXiv 2015
Konda, Memisevic, “Learning visual odometry ...”, VISAPP 2015

Visual prediction

Doersch, Gupta, Efros, “… context prediction”, ICCV 2015
Oh, Guo, Lee, Lewis, Singh, “Action-conditional video …”, NIPS 2015
Kulkarni, Whitney, Kohli, Tenenbaum, “… inverse graphics ...”, NIPS 2015
Vondrick, Pirsiavash, Torralba, “Anticipating the future ...”, arXiv 2015

Video for unsupervised image features

Wang, Gupta, “Unsupervised learning of visual …”, ICCV 2015
Goroshin, Bruna, Tompson, Eigen, LeCun, “Unsupervised ...”, ICCV 2015

Approach

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"

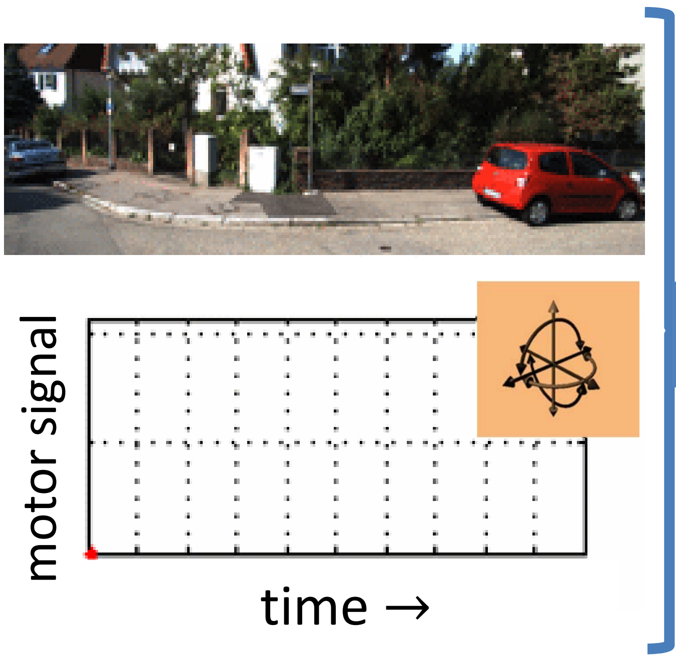
$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

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

Approach

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

Source: “Learning image representations equivariant to ego-motion”, Jayaraman and Grauman, ICCV 2015

Approach

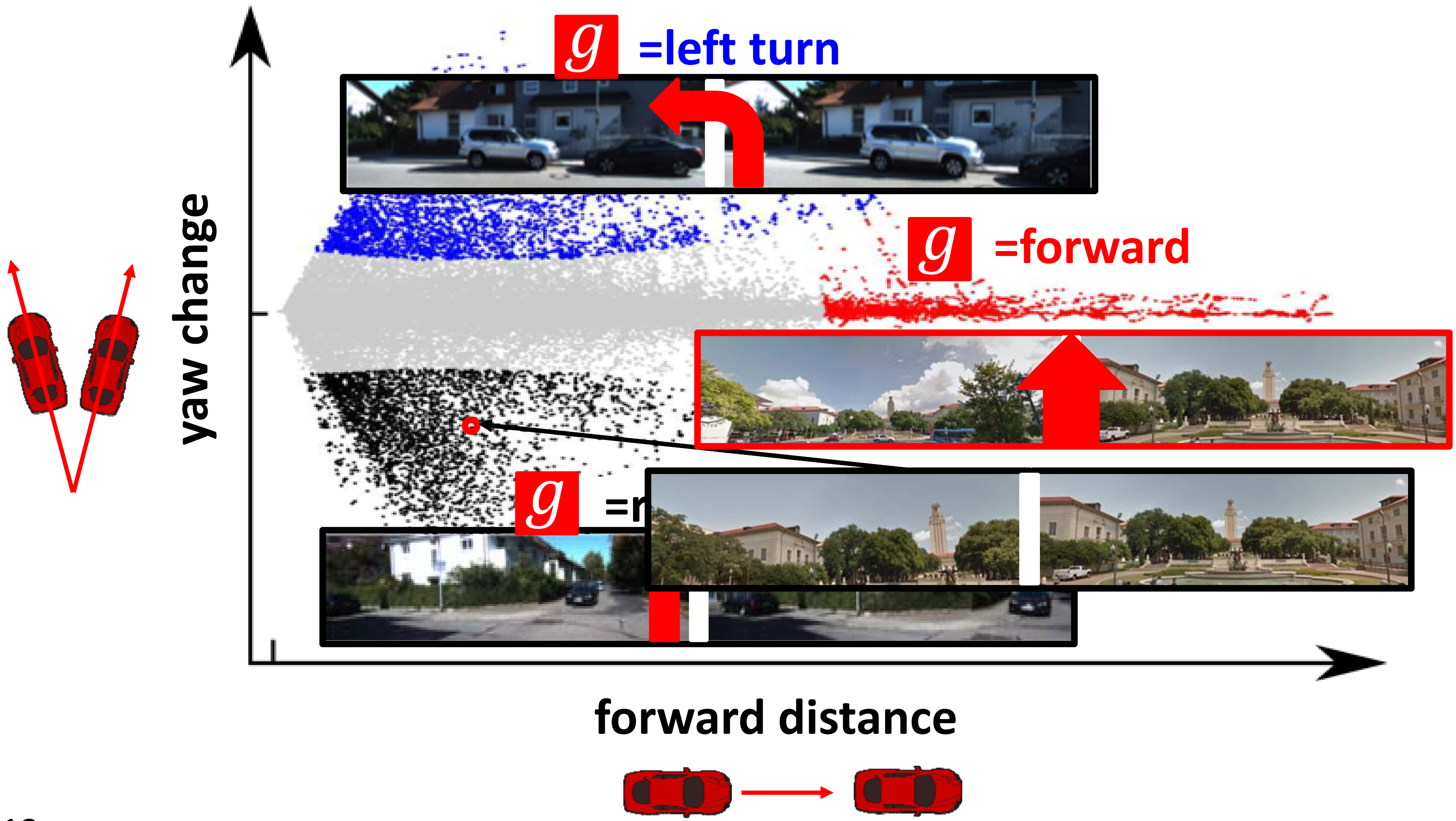
1. Extract training frame pairs from video

2. Learn ego-motion-equivariant image features

3. Train on target recognition task in parallel

Training frame pair mining

Discovery of ego-motion clusters

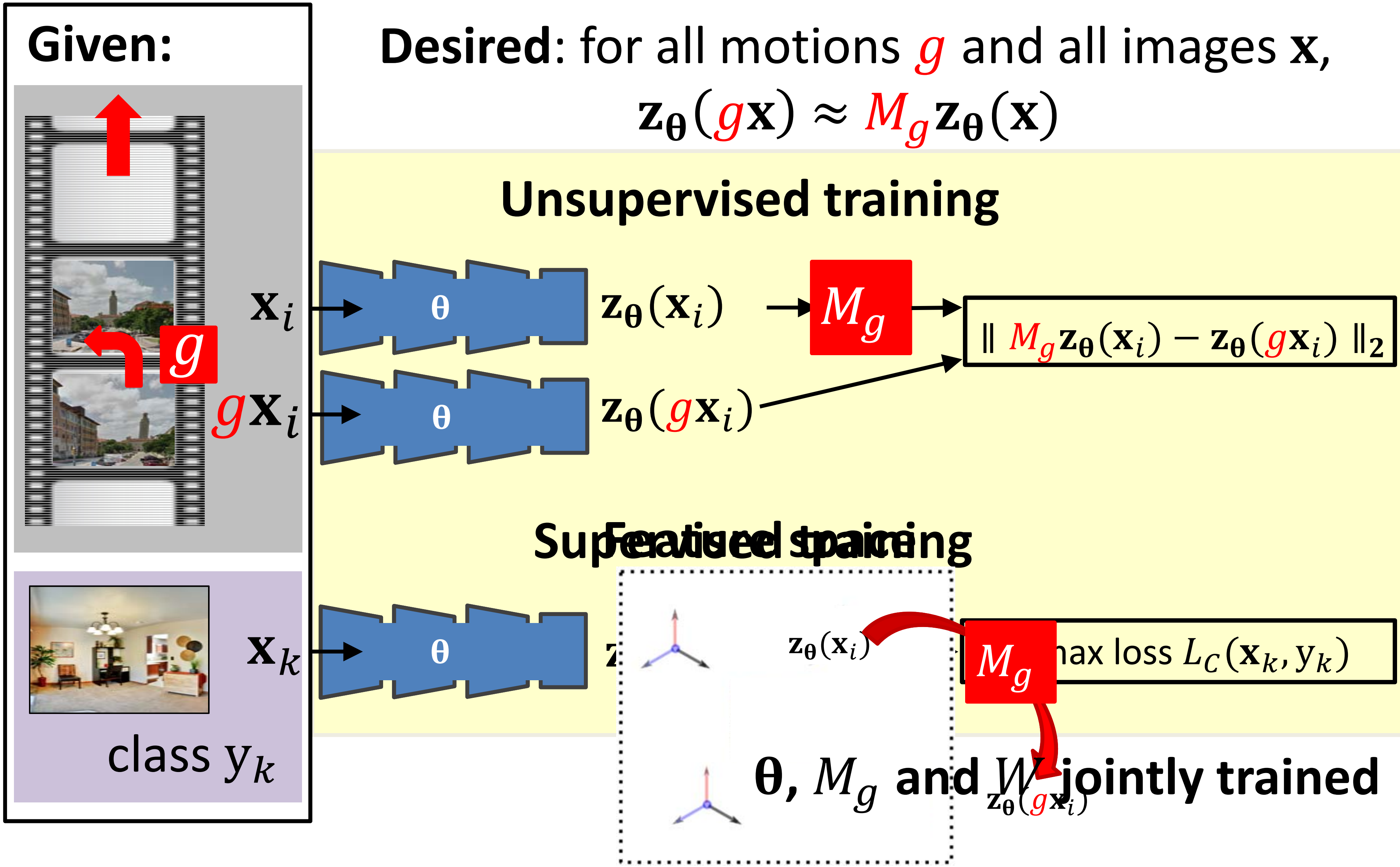


[Slide credit: Dinesh Jayaraman]

Ego-motion equivariant feature learning

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

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



[Slide credit: Dinesh Jayaraman]

Experiments

- Validation using 3 public datasets: **NORB**, **KITTI**, **SUN**.
- Comparison with different methods: CLSNET, TEMPORAL, DRLIM.

Results: Recognition

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

Geiger et al, IJRR ’13

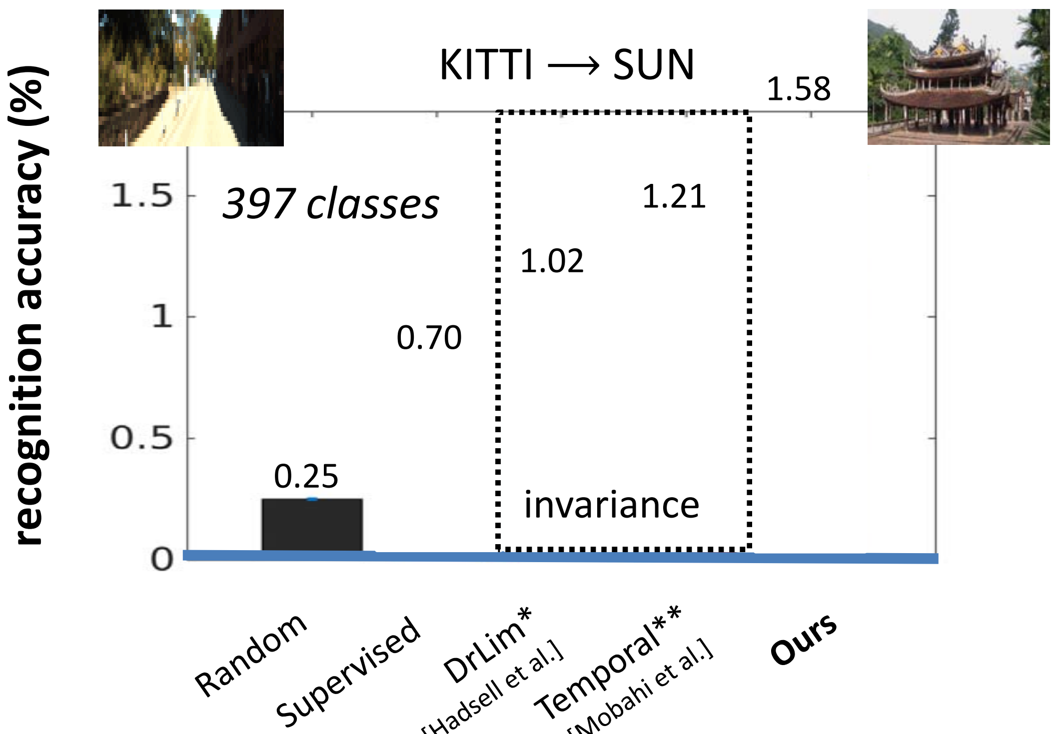
Exploit features for **static scene classification**
(SUN, 397 classes)

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

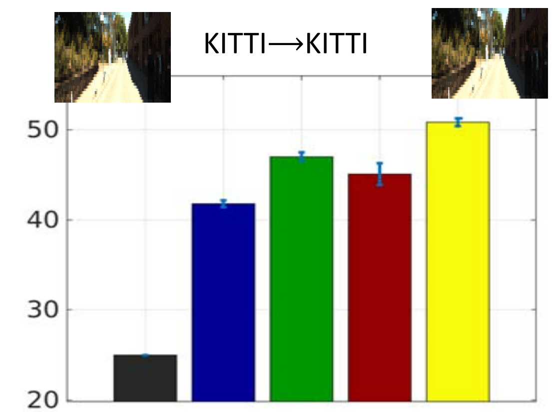
Xiao et al, CVPR ’10

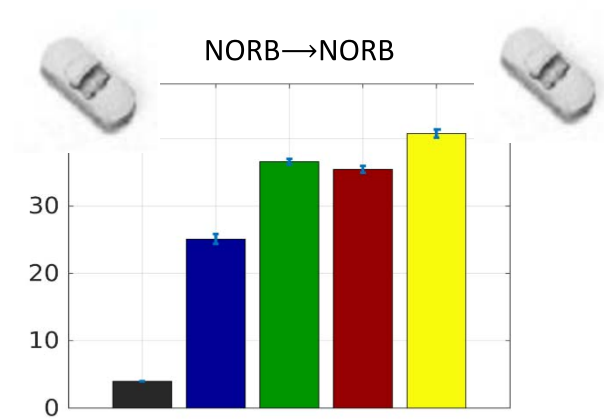
Results: Recognition

Do ego-motion equivariant features improve recognition?



6 labeled training examples per class





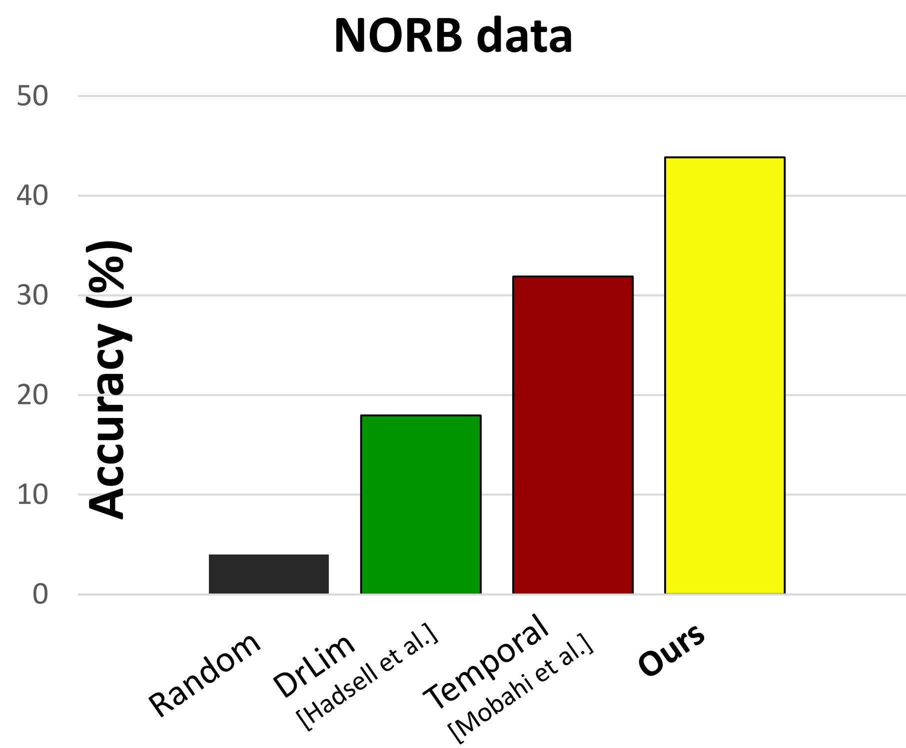
Up to 30% accuracy increase over state of the art!

*Hadsell et al., Dimensionality Reduction by Learning an Invariant Ma

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

Results: Active recognition

- Leverage proposed equivariant embedding to select next best view for object recognition



[Slide credit: Dinesh Jayaraman]

Conclusion and Future Work

- The paper provided a new *embodied* visual feature learning paradigm.
- The Ego-motion equivariance boosts performance across multiple challenging recognition tasks.

Questions

- Why KITTI training and not some other domain based training?
- Why does incorporating DRLIM improve EQUIV? Still Temporal coherence properties left to be learned?
- Is it meaningful to compare EQUIV or EQUIV + DRLIM with the other cases with respect to equivariance error?


Thank You