Time is Precious: Self-Supervised Learning Beyond Images

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO

2024

Part (1): Learning Image Encoders from Videos - A Review

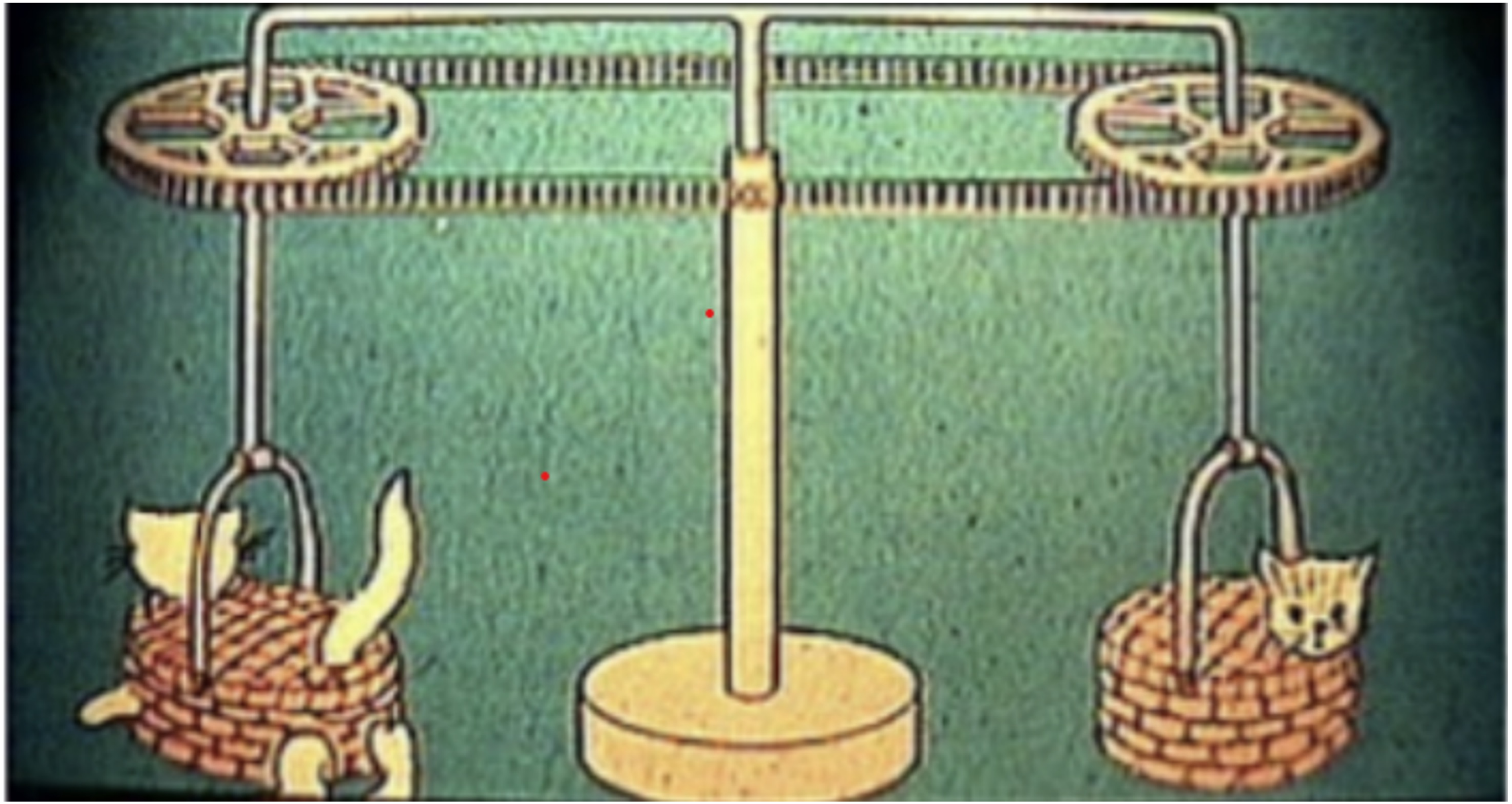
Learning Image Encoders from Videos

Integrating Vision and Motion

Visual Prediction

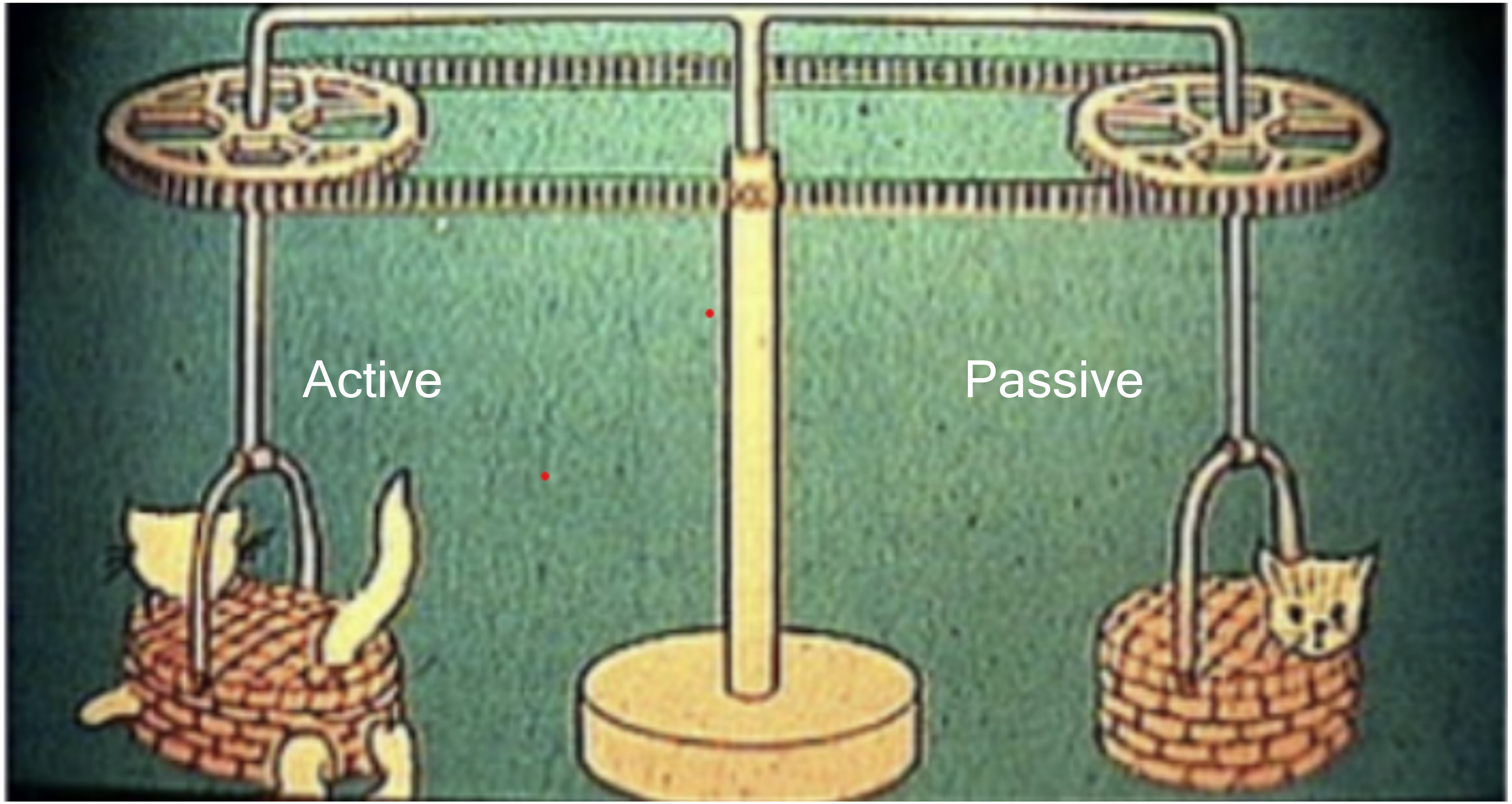
Videos for unsupervised image features

The Kitten Carousel Experiment



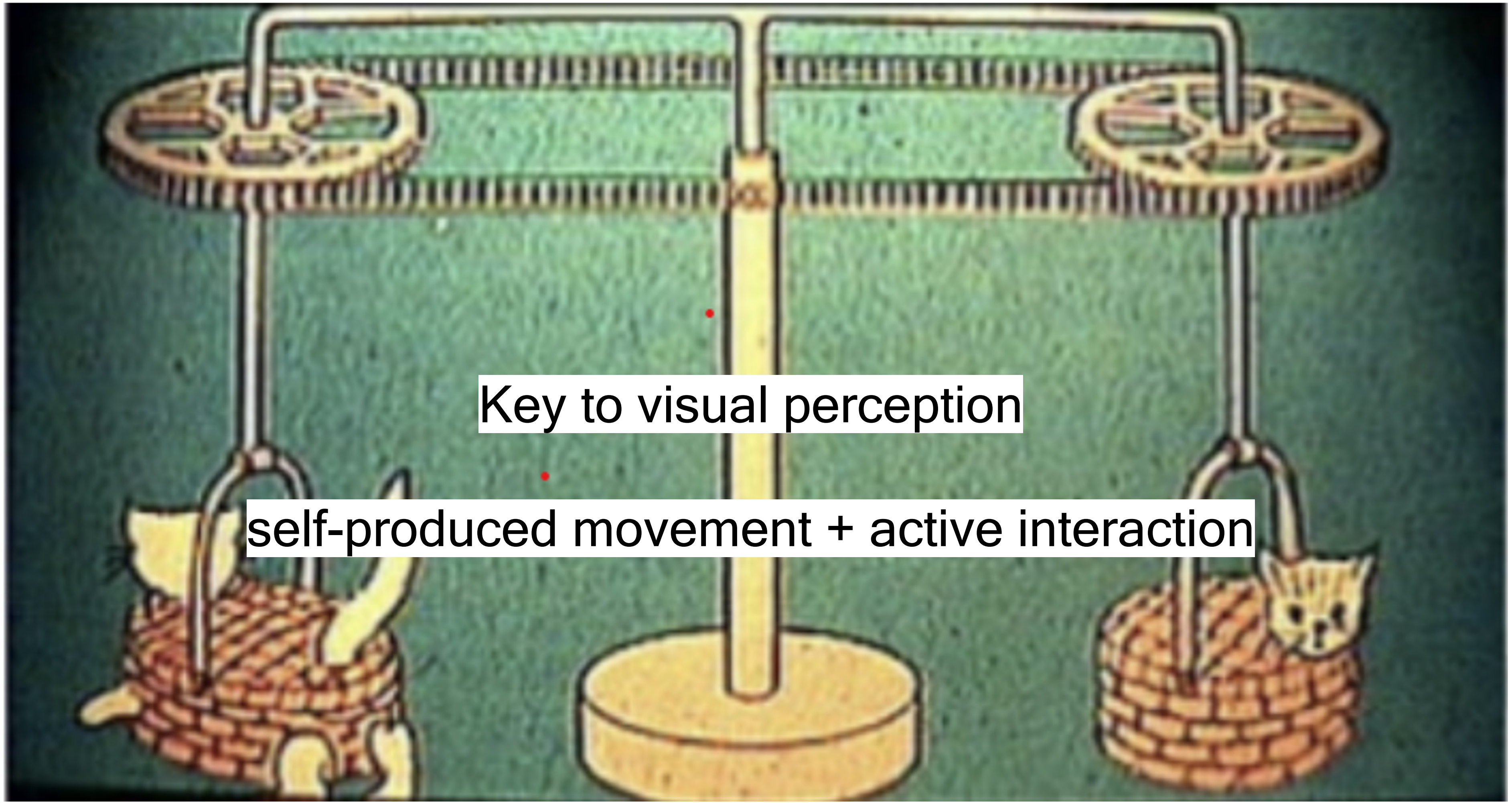
Held & Hein, Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 1963
Slide inspired from Kristen Grauman, Egomotion and Visual Learning, 2016

The Kitten Carousel Experiment



Held & Hein, Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 1963
Slide inspired from Kristen Grauman, Egomotion and Visual Learning, 2016

The Kitten Carousel Experiment



Held & Hein, Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 1963
Slide inspired from Kristen Grauman, Egomotion and Visual Learning, 2016

Egocentric Videos ↔ Vision

Goal: Teach computer vision systems:

"How I move" ↔ "how my visual surroundings change"

Egocentric Videos ↔ Vision

Goal: Teach computer vision systems:

"How I move" ↔ "how my visual surroundings change"

Autonomous vehicle

+

Unlabeled videos

Egocentric Videos ↔ Vision

Goal: Teach computer vision systems:

"How I move" ↔ "how my visual surroundings change"

Mobile camera

+

Unlabeled videos

Egocentric Videos ↔ Vision

Goal: Teach computer vision systems:

"How I move" ↔ "how my visual surroundings change"

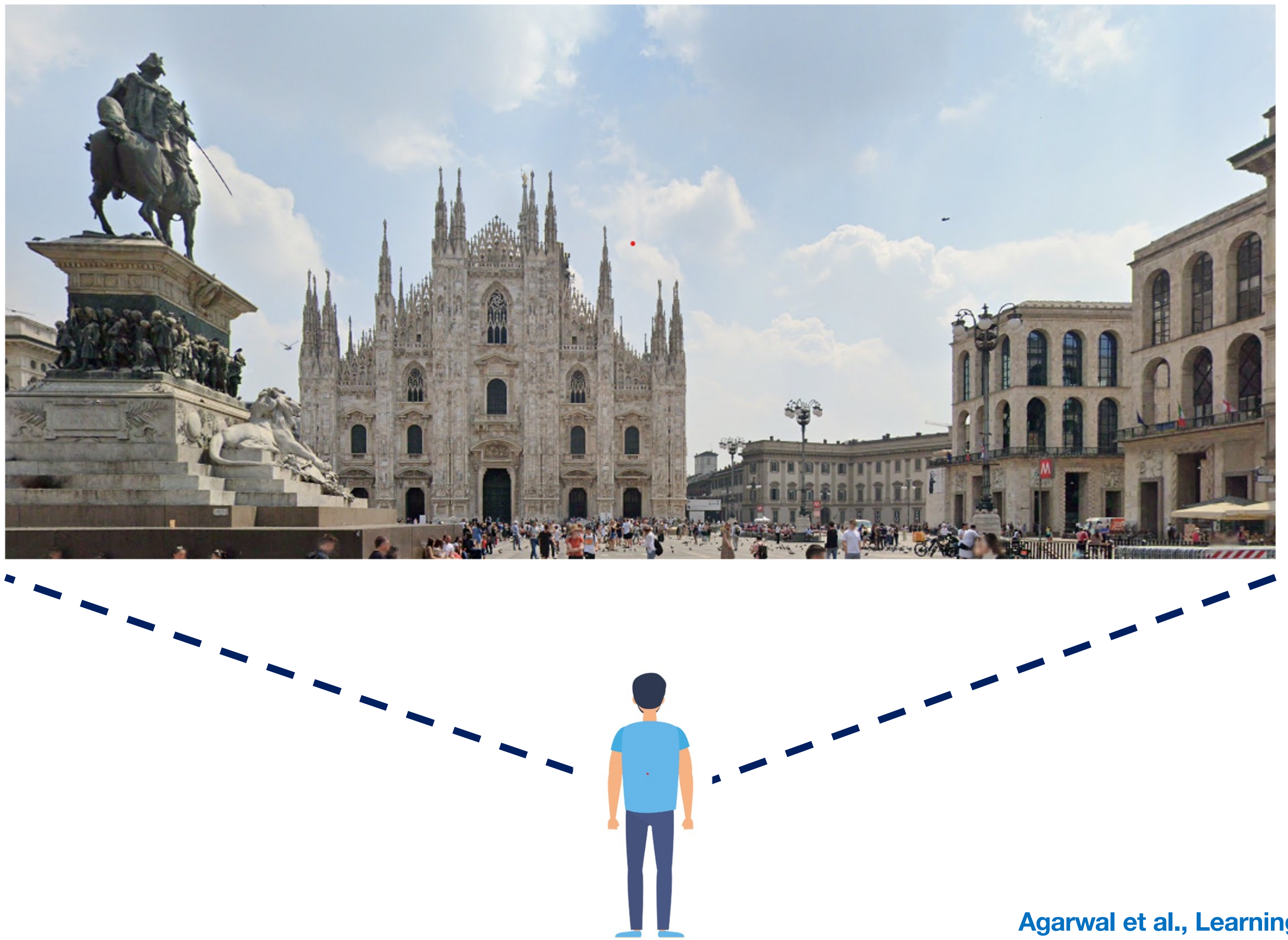
Head-mount camera

+

Unlabeled videos

Damen et al., Scaling Egocentric Vision: The EPIC-KITCHENS Dataset, ECCV 2018

EgoMotion ↔ Vision: Predicting camera transformations

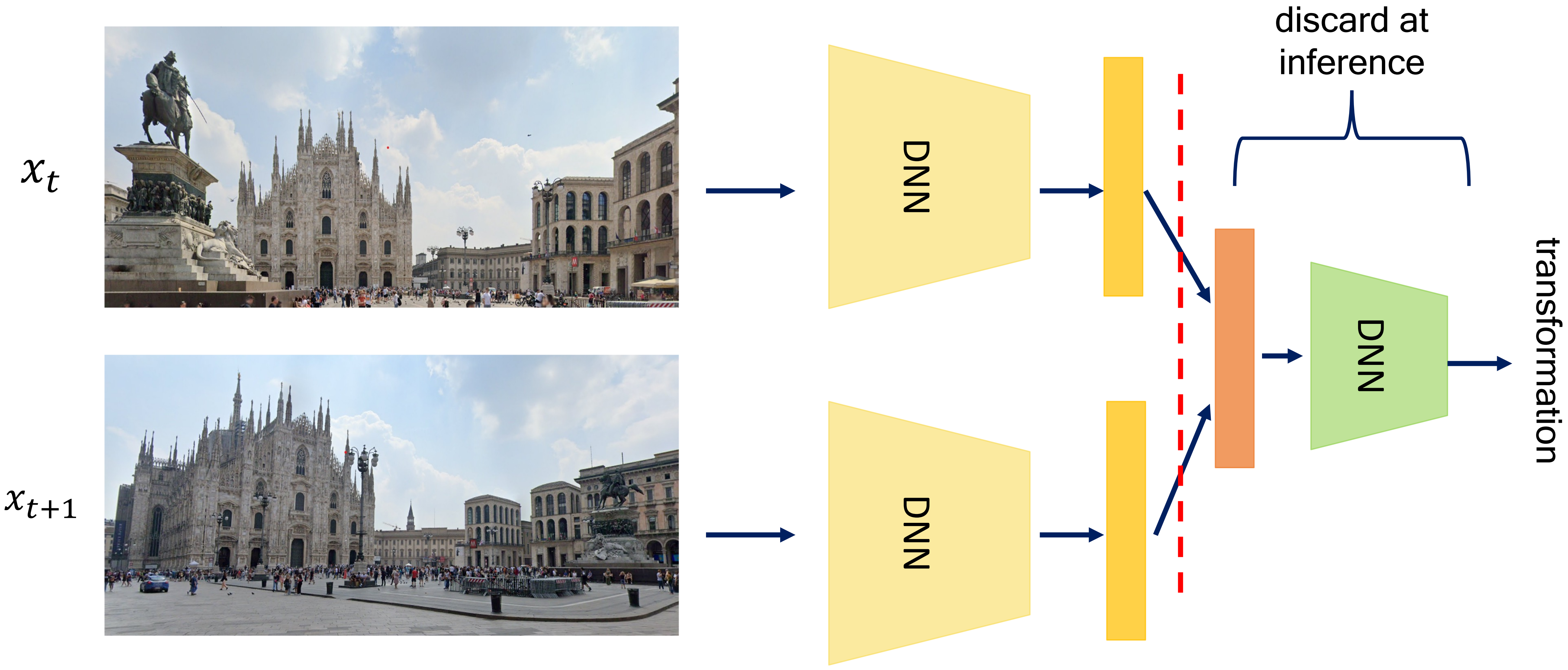


Agarwal et al., Learning to see by moving, ICCV 2015

EgoMotion ↔ Vision: Predicting camera transformations

Agarwal et al., Learning to see by moving, ICCV 2015

EgoMotion ↔ Vision: Predicting camera transformations



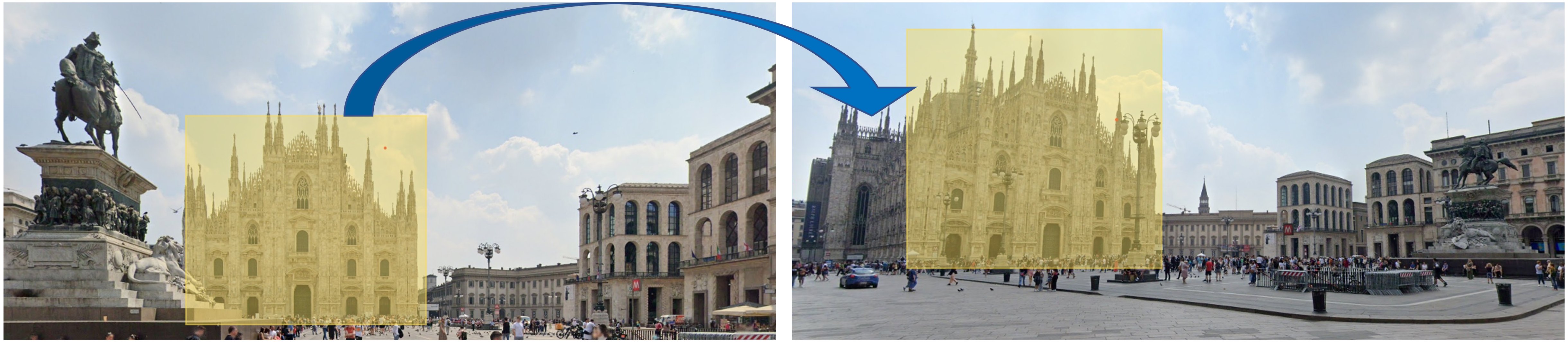
Agarwal et al., Learning to see by moving, ICCV 2015

EgoMotion ↔ Vision: Egomotion Equivariance

Objective: Pairs of frames related by same ego-motion should be related by same feature transformation.

Jayaraman & Grauman, Learning image representations tied to ego-motion, ICCV 2015

EgoMotion ↔ Vision: Egomotion Equivariance



time

Jayaraman & Grauman, Learning image representations tied to ego-motion, ICCV 2015

EgoMotion ↔ Vision: Egomotion Equivariance

time

Jayaraman & Grauman, Learning image representations tied to ego-motion, ICCV 2015

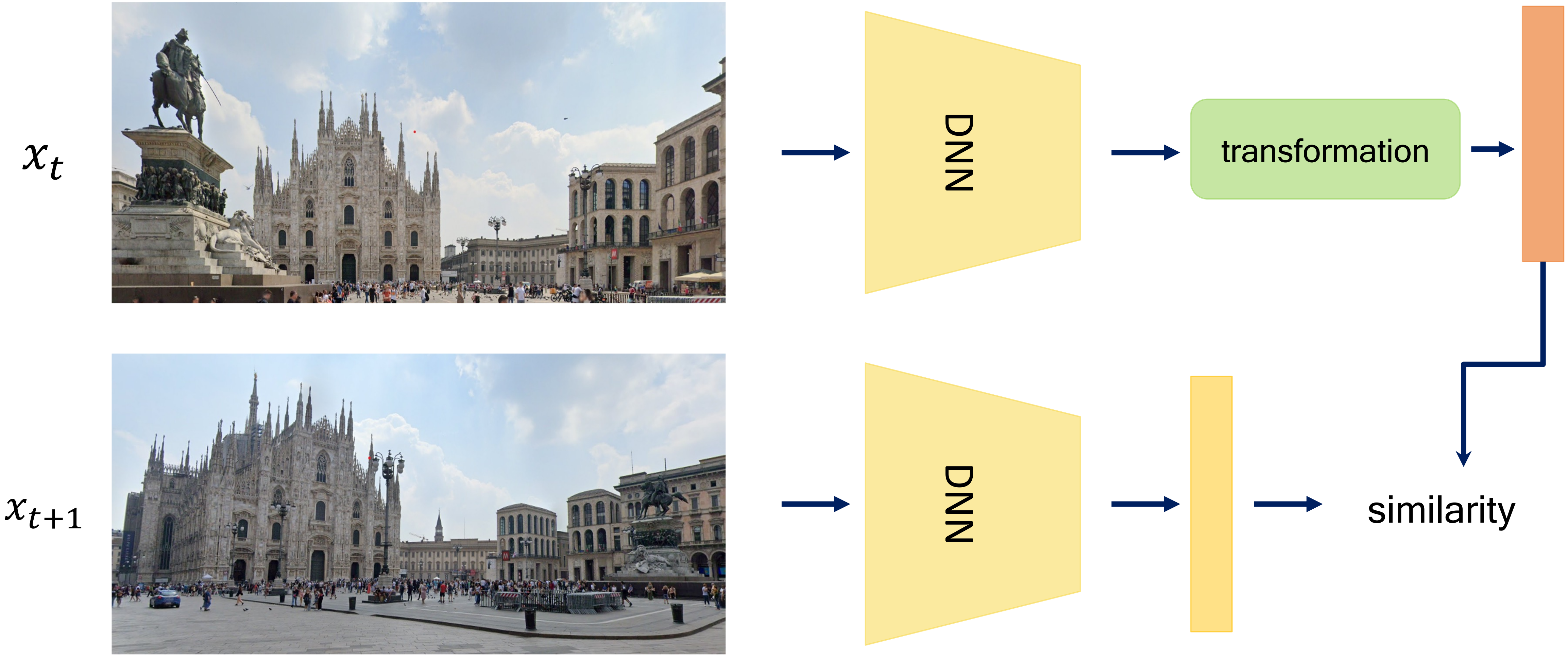
EgoMotion $\leftrightarrow$ Vision: Egomotion Equivariance

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

Key to recognition

EgoMotion $\leftrightarrow$ Vision: Egomotion Equivariance



Jayaraman & Grauman, Learning image representations tied to ego-motion, ICCV 2015

Time to bring back active recognition, in a challenging setting

EgoMotion ↔ Vision: Learning How to Move

Perception

Perception

Jayaraman & Grauman, Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion., ECCV 2016

EgoMotion ↔ Vision: Learning How to Move

Perception

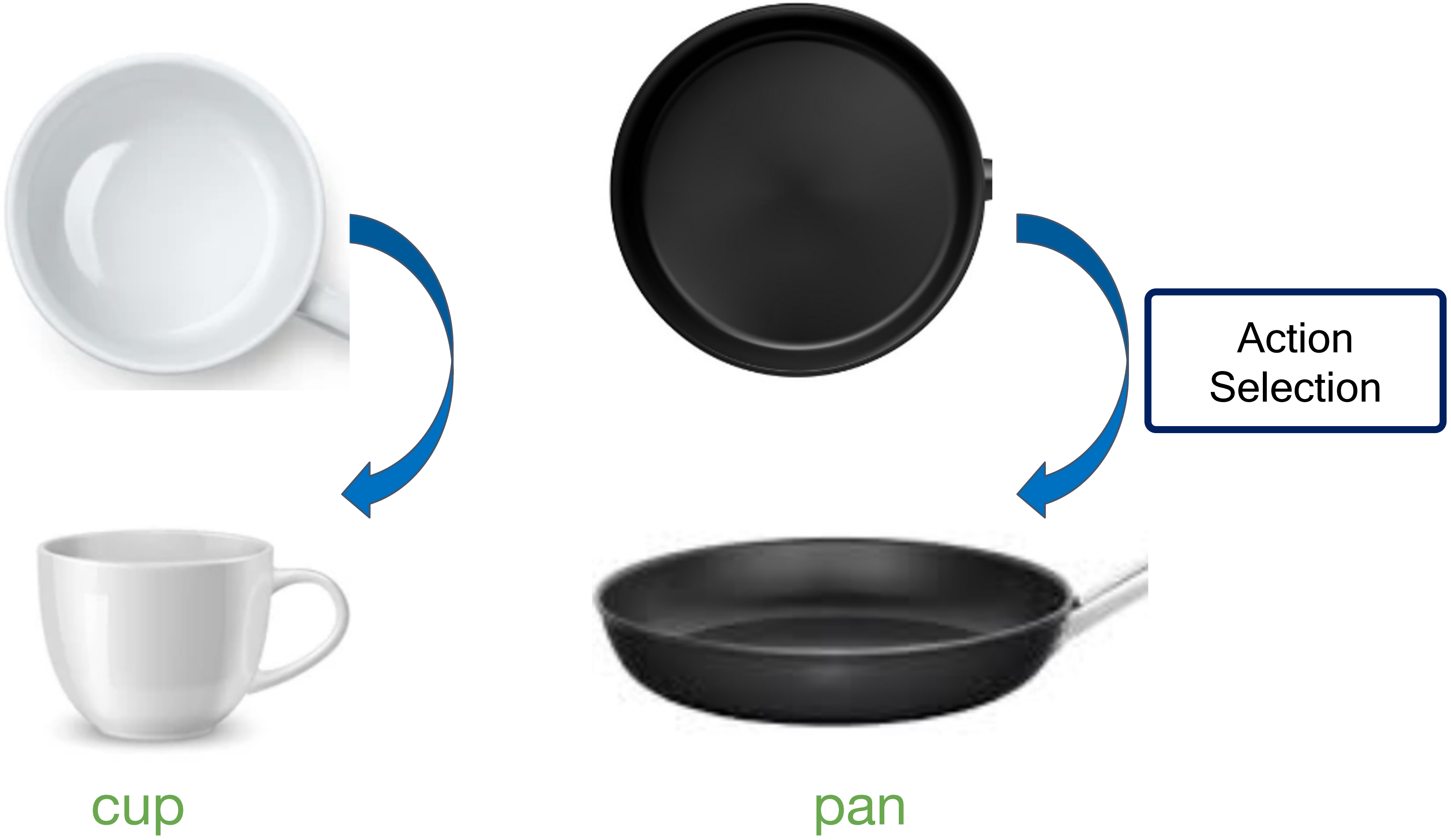
cup/bowl/pan?

Perception

cup/bowl/pan?

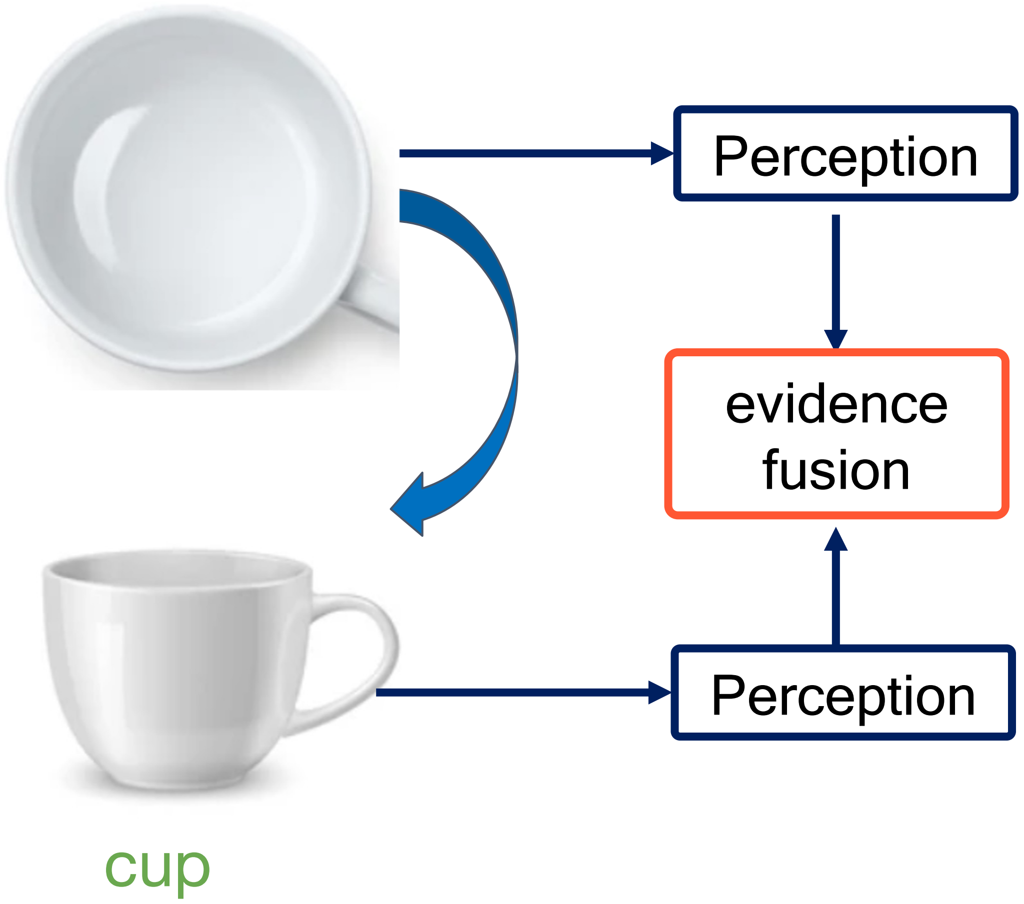
Jayaraman & Grauman, Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion., ECCV 2016

EgoMotion ↔ Vision: Learning How to Move



Jayaraman & Grauman, Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion., ECCV 2016

EgoMotion ↔ Vision: Learning How to Move



Egomotion equivariance to select **next best view**

Jayaraman & Grauman, Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion., ECCV 2016

EgoMotion $\leftrightarrow$ Vision: Recap

Visual learning benefits from:

- Context of action and motion in the wild
- Continuous self-acquired feedback

Ego-motion equivariance boots performance across multiple challenging recognition tasks.

Learning Image Encoders from Videos

Integrating Vision and Motion

Visual Prediction

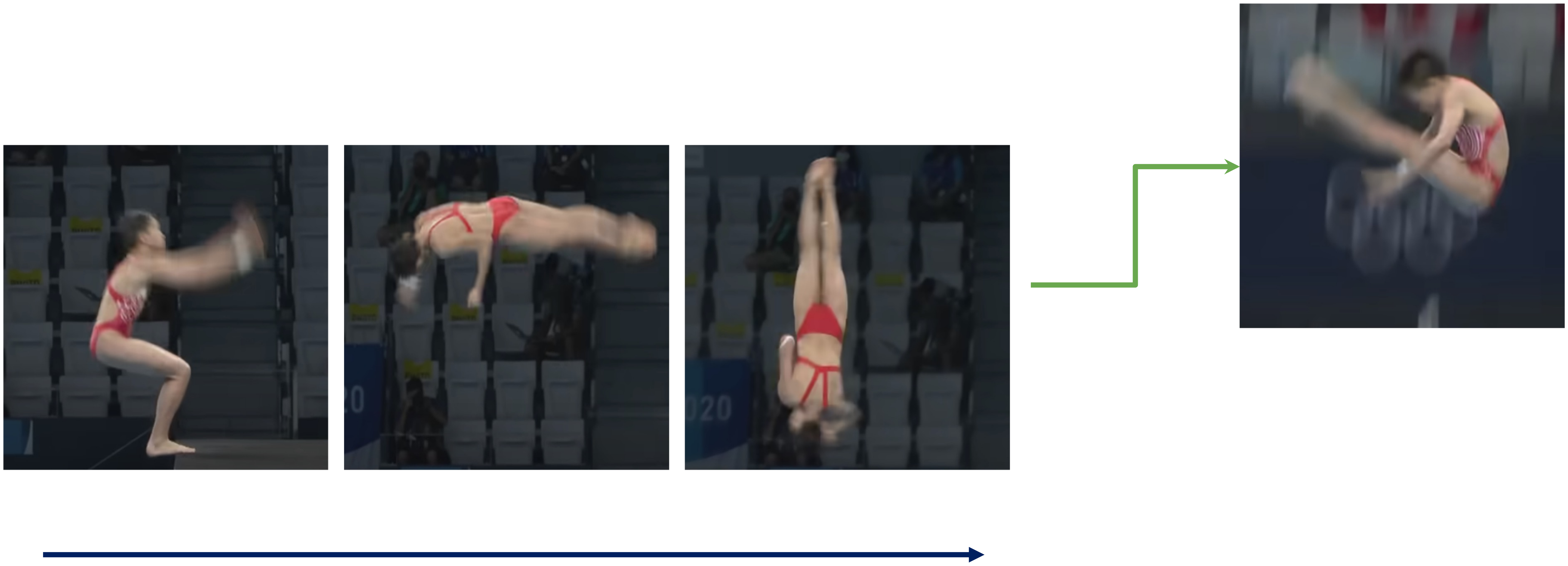
Videos for unsupervised image features

Visual Prediction: Anticipating future

What happens next?

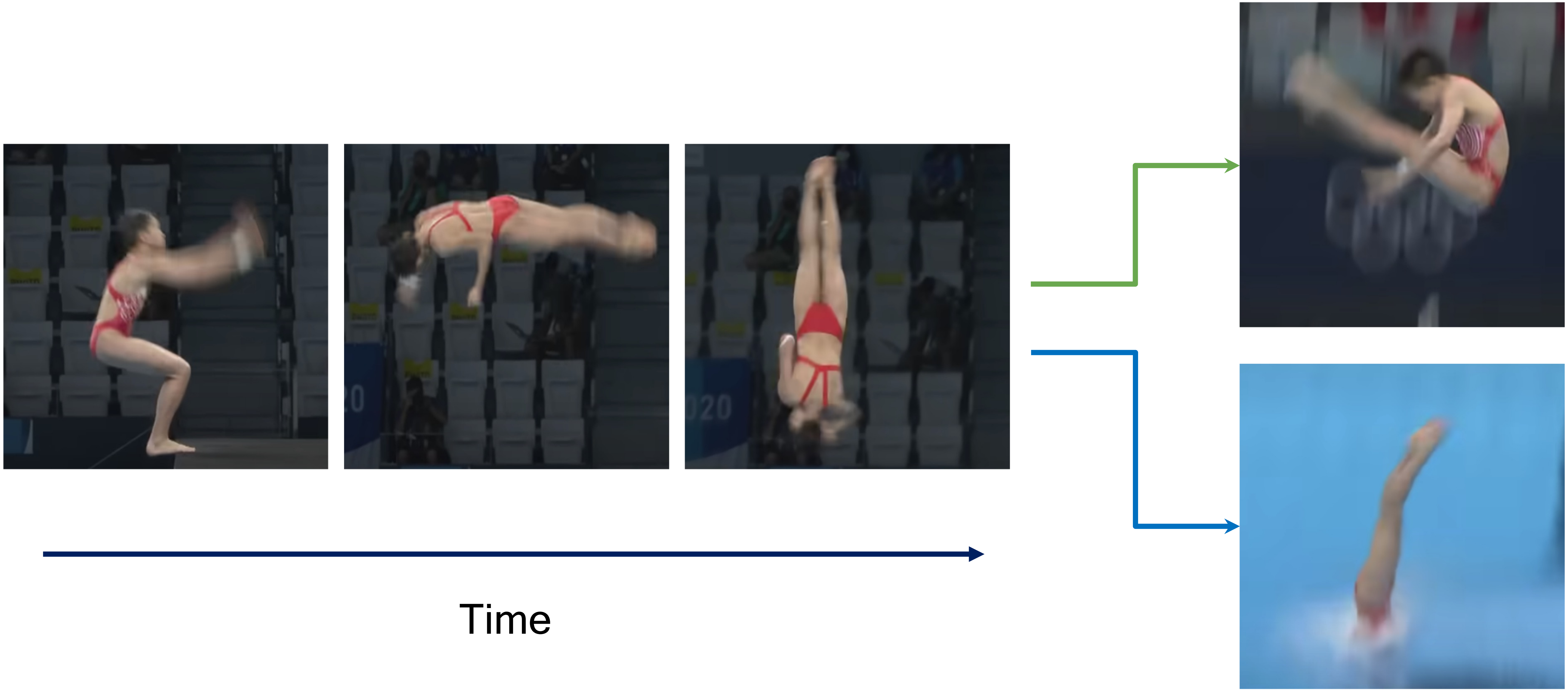
Time

Visual Prediction: Anticipating future

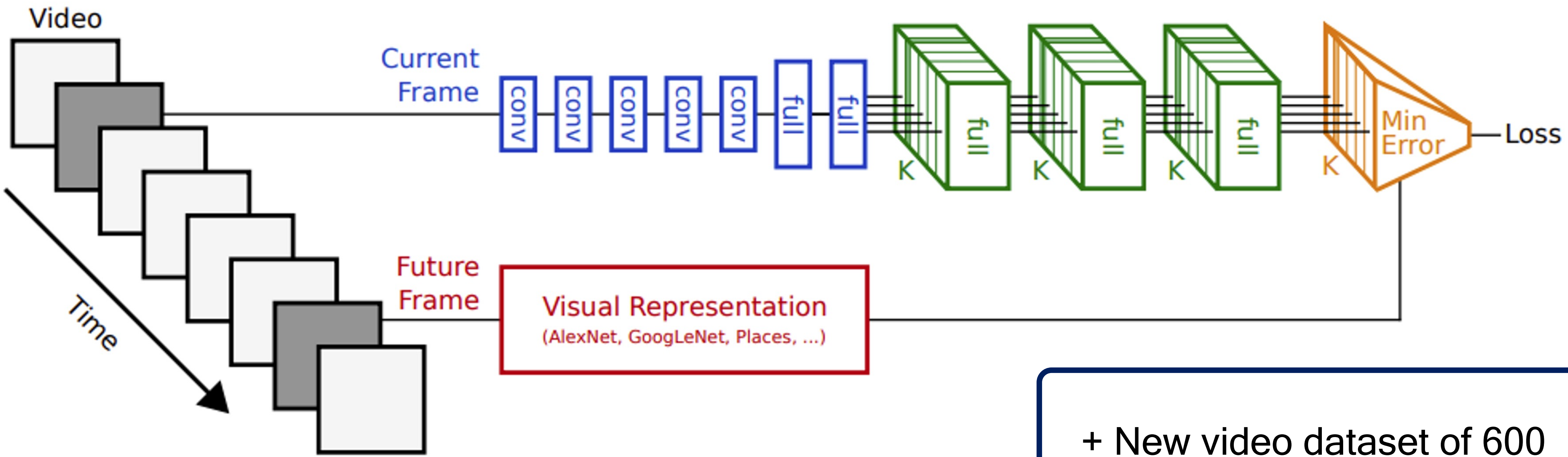


Time

Visual Prediction: Anticipating future



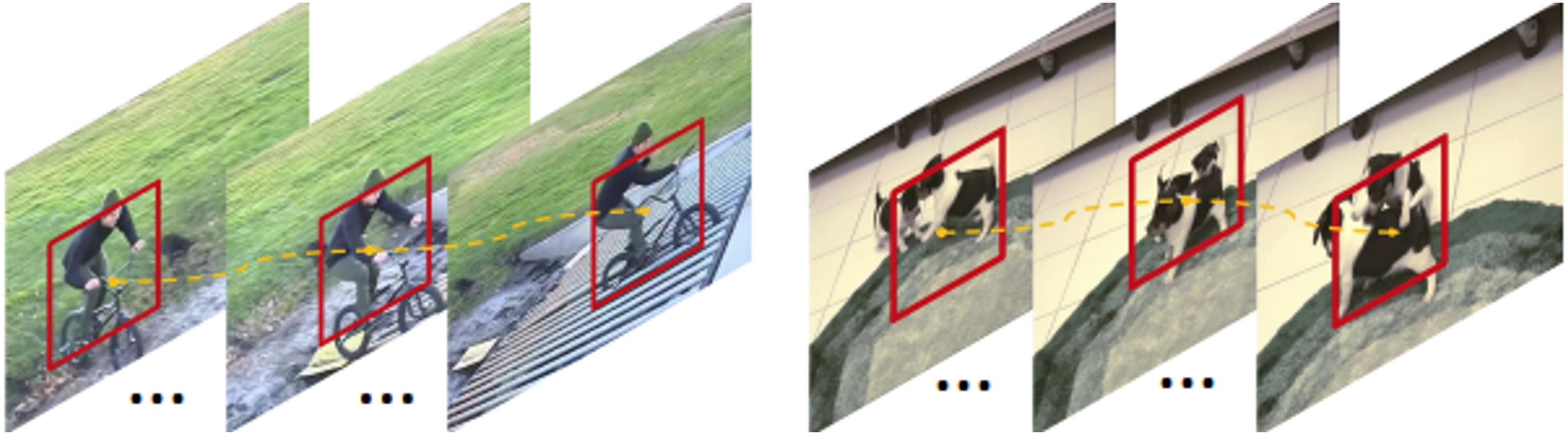
Visual Prediction: Anticipating future



+ New video dataset of 600 hours Youtube videos

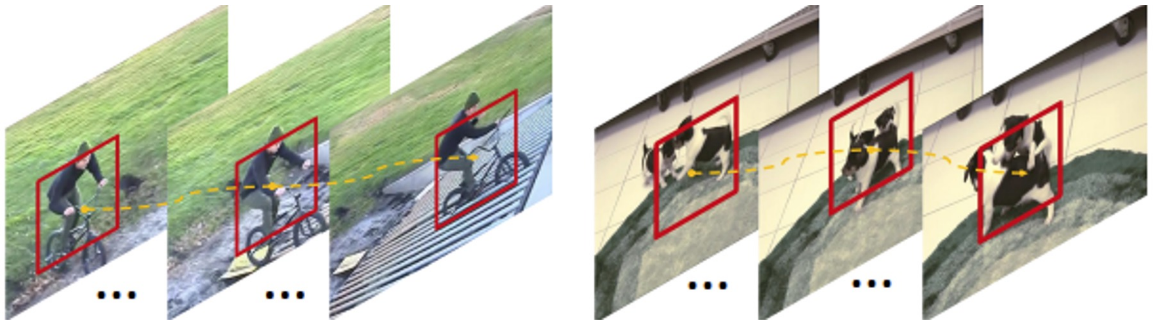
Vondrick *et al.*, Anticipating Visual Representations from Unlabeled Video, CVPR 2016

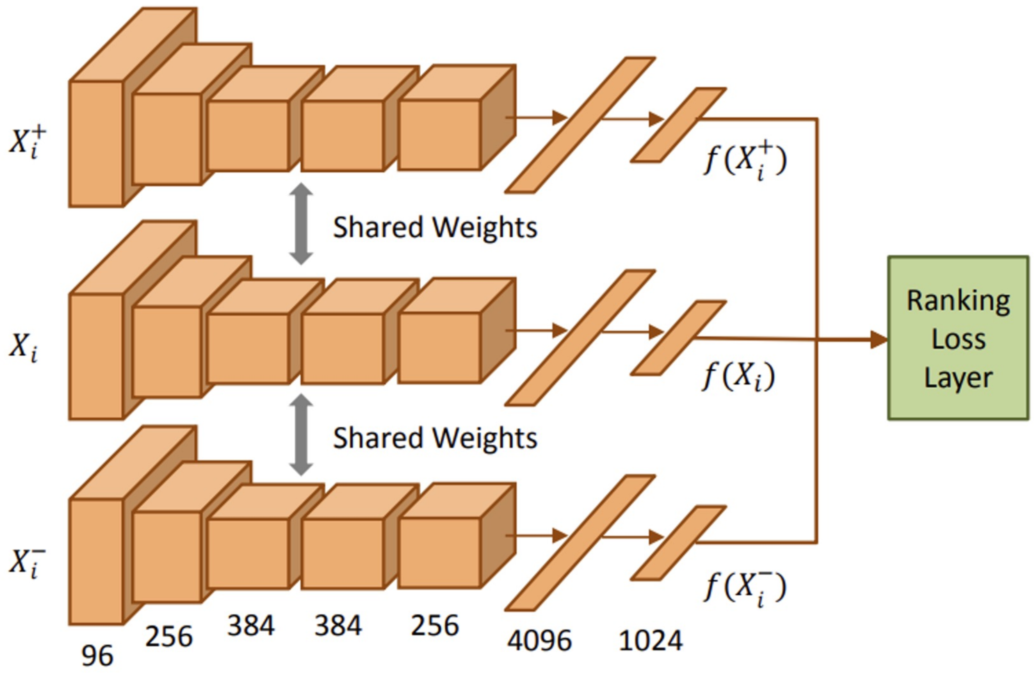
Visual Prediction: Learning to track



Wang & Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015

Visual Prediction: Learning to track





Anchor: object from 1st frame
Positive: same object from last frame
Negative: Random crop from any frame

Wang & Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015

Visual Prediction: Recap

- Learning embeddings by tracking patches of similar objects acts as a strong semantic supervision.

- Predicting visual representations enables scalable, generalizable anticipation models for improved forecasting systems.

- Opens avenues for integrating model-based reasoning and intuitive physics in self-supervised learning frameworks.

Learning Image Encoders from Videos

Integrating Vision and Motion

Visual Prediction

Videos for unsupervised image features

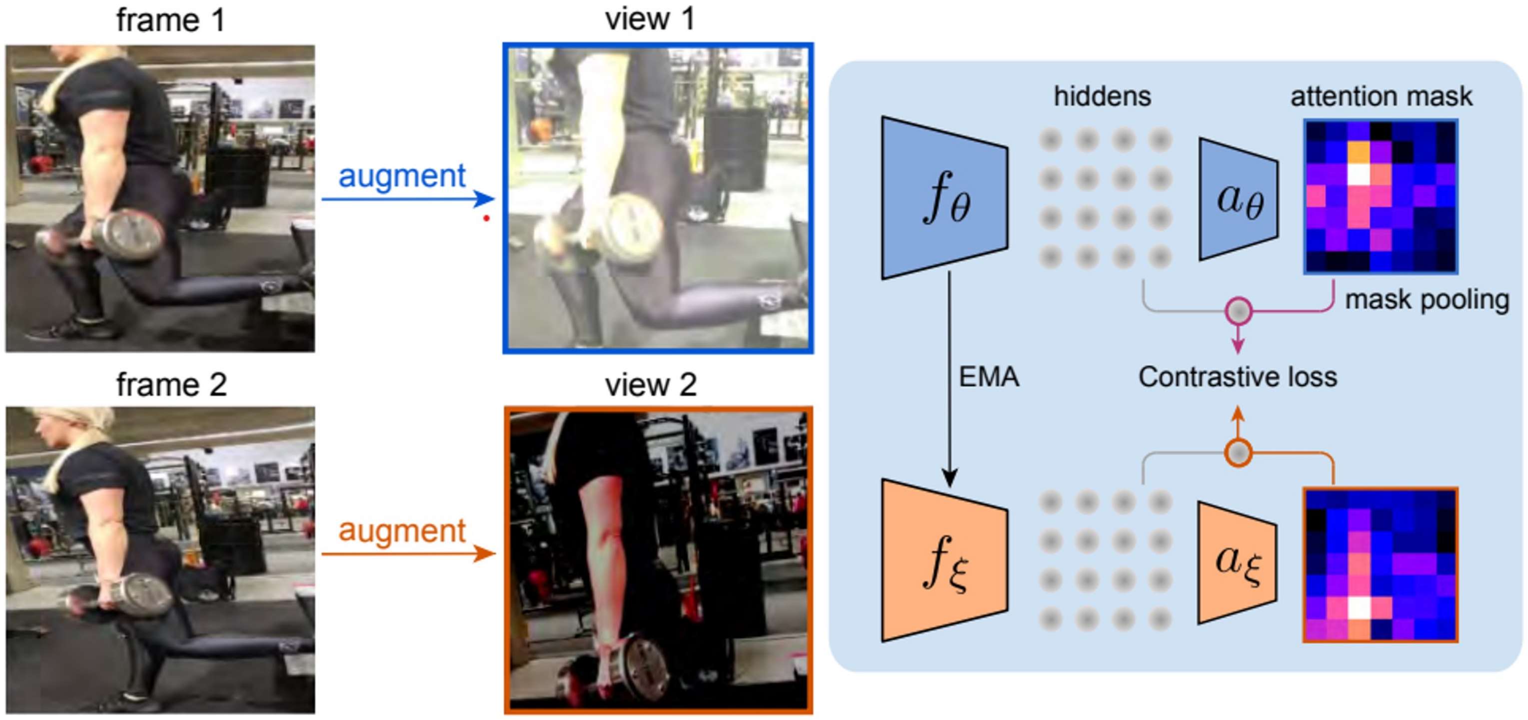
Videos for unsupervised image features

- Current SSL methods focus on invariant representations (scale, color, translation) through synthetic augmentations.

- Natural videos provide rich signals (pose, viewpoint, motion) that are crucial for learning intuitive physics and reasoning.

- Distills natural video transformations into image-based representations.

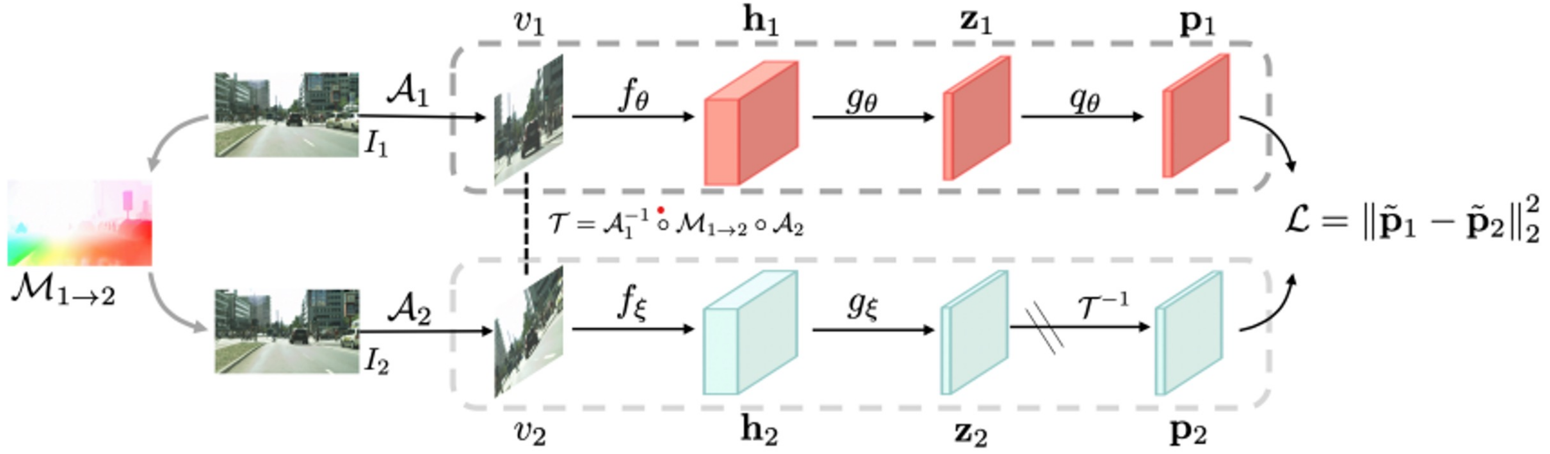
Videos for unsupervised image features: Temporal Equivariance



Enforces temporal equivariance of masked features across temporally displaced views.

Parthasarathy et al., Self-supervised video pretraining yields human-aligned visual representations, NeurIPS 2023

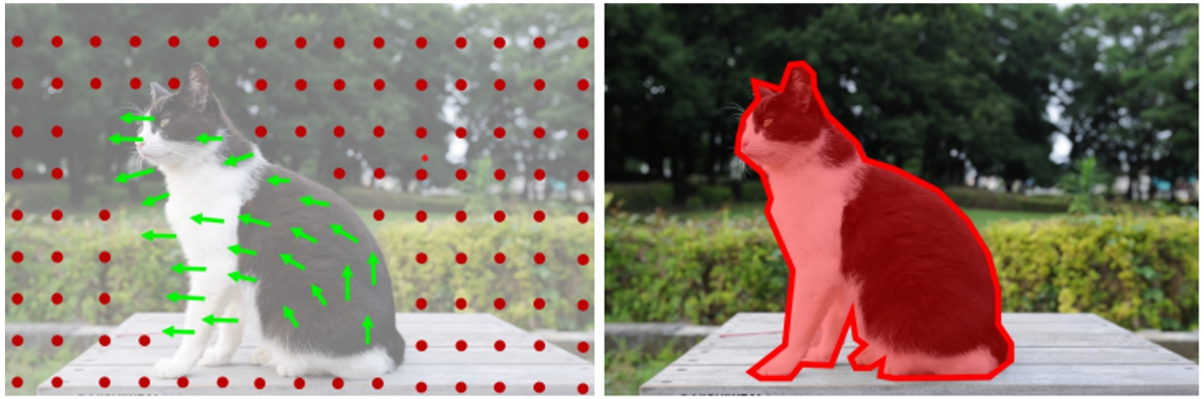
Videos for unsupervised image features: Flow Equivariance



Enforces flow equivariance: applies flow transformation to the features of the current frame to predict features of another frame.

Xiong et al., Self-Supervised Representation Learning from Flow Equivariance, ICCV 2021

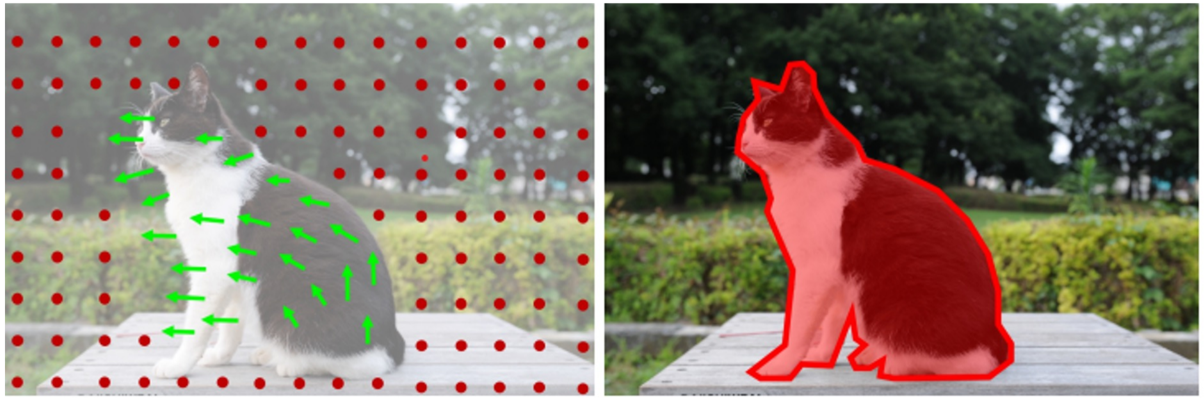
Videos for unsupervised image features: Motion

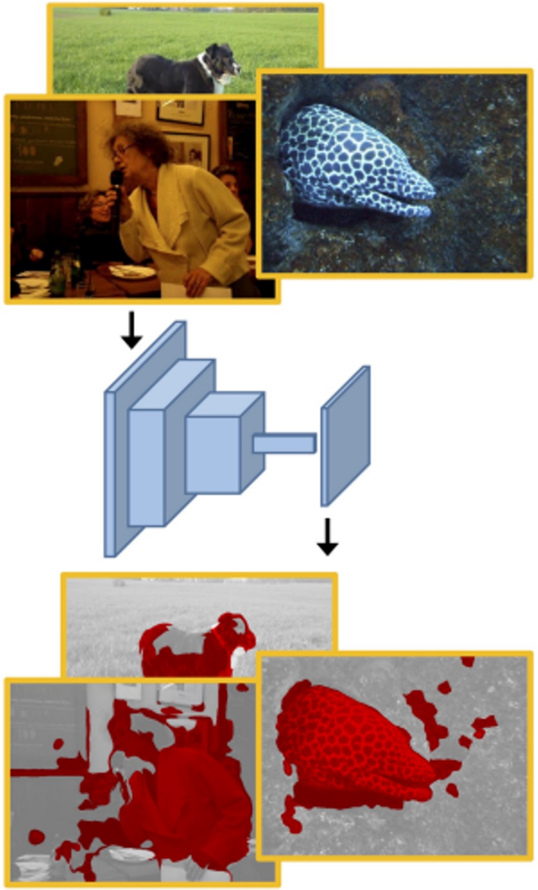


Use motion information to segment moving object and enforce ConvNet to predict the segmentation masks

Pathak et al., Learning Features by Watching Objects Move, CVPR 2017

Videos for unsupervised image features: Motion





Use motion information to segment moving object and enforce ConvNet to predict the segmentation masks

Pathak et al., Learning Features by Watching Objects Move, CVPR 2017

This field has a rich history, And now it's time to get back to it

2002 Wiskott and Sejnowski, Slow Feature Analysis: Unsupervised Learning of Invariances

2014 Pintea *et al.*, Deja Vu: Motion Prediction in Static Images

2015 Agarwal et al., Learning to see by moving
2015 Jayaraman & Grauman, Learning Image Representations Equivariant to Ego-Motion
2015 Wang & Gupta, Unsupervised Learning of Visual Representations using Videos
2015 Oh, Guo, Lee, Lewis, Singh, Action-Conditional Video Prediction using Deep Networks in Atari Games
2015 Kulkarni, Whitney, Kohli, Tenenbaum, Deep Convolutional Inverse Graphics Network
2015 Misra et al., Watch and Learn: Semi-Supervised Learning of Object Detectors from Videos
2015 Doersch, Gupta, Efros, Unsupervised Visual Representation Learning by Context Prediction
2015 Goroshin, Bruna, Tompson, Eigen, LeCun, Unsupervised Learning of Spatiotemporally Coherent Metrics

2016 Gao et al., Object-Centric Representation Learning From Unlabeled Videos
2016 Jayaraman & Grauman, Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion
2016 Jayaraman & Grauman, Slow and steady feature analysis: higher order temporal coherence in video
2016 Vondrick *et al.*, Anticipating Visual Representations from Unlabeled Video
2016 Bertasius et al., First Person Action-Object Detection with Egonet
2016 Leal-Taixe et al., Learning By Tracking- Siamese CNN for Robust Target Association
2016 Misra et al., Shuffle and Learn- Unsupervised Learning Using Temporal Order Verification
2016 Fragkiadaki et al., Learning Visual Predictive Models of Physics for Playing Billiards

2017 Chakraborty & Namboodiri, Learning to Estimate Pose by Watching Videos
2017 Croitoru et al., Unsupervised learning from video to detect foreground objects in single images
2017 Pathak et al., Learning Features by Watching Objects Move
2017 Wang et al., Transitive Invariance for Self-supervised Visual Representation Learning

2018 Jayaraman & Grauman, Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks
2018 Pot et al., Self-supervisory Signals for Object Discovery and Detection
2018 Xia et al., Gibson Env: Real-World Perception for Embodied Agents
2018 Mahendran et al., Cross Pixel Optical Flow Similarity for Self-Supervised Learning
2018 Wei et al., Learning and Using the Arrow of Time
2018 Redondo-Cabrera & López-Sastre, Unsupervised learning from videos using temporal coherency deep networks