Aligning Videos In Space and Time

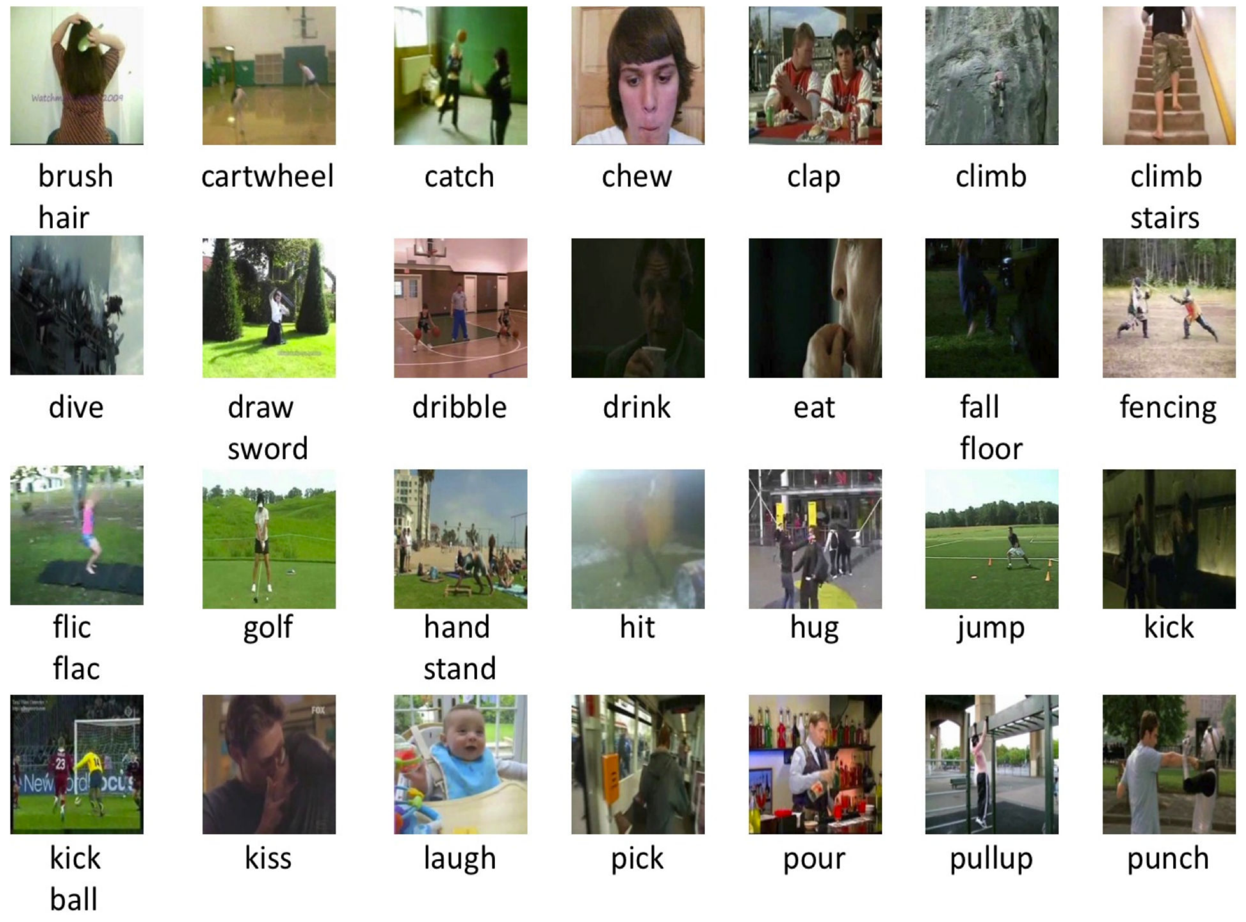
Senthil Purushwalkam[1], Tian Ye[1], Saurabh Gupta[3], Abhinav Gupta[1,2]

[1]Carnegie Mellon University
[2]Facebook AI Research
[3]University of Illinois Urbana-Champaign (UIUC)

European Conference on Computer Vision (ECCV), 2020

Long Presentation

Video Understanding in Computer Vision

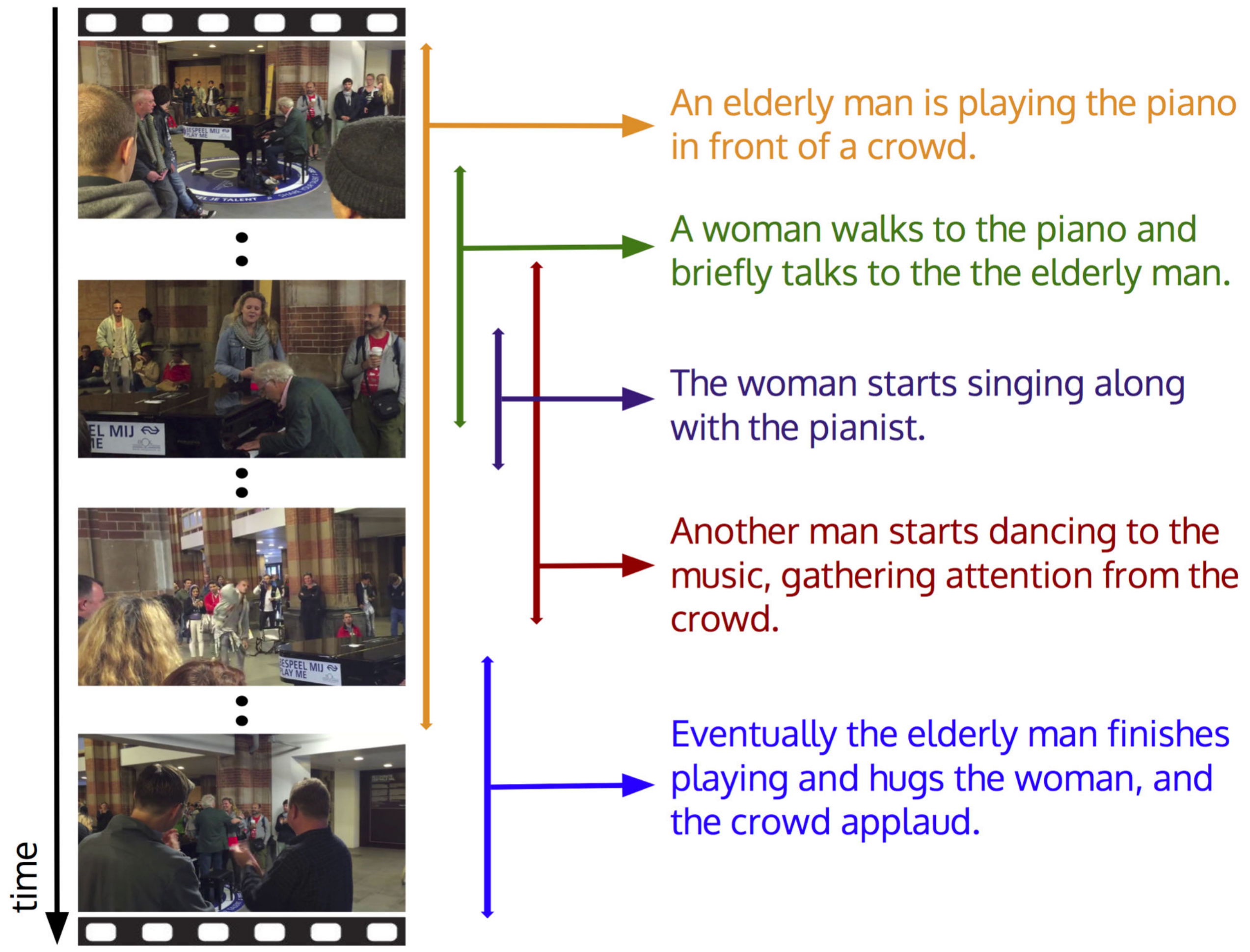


Action Recognition

Classification of Videos into Predefined Action Categories

Action Detection

Localizing Predefined Actions Temporally in Videos



Video Captioning

Generating Textual Descriptions for Videos

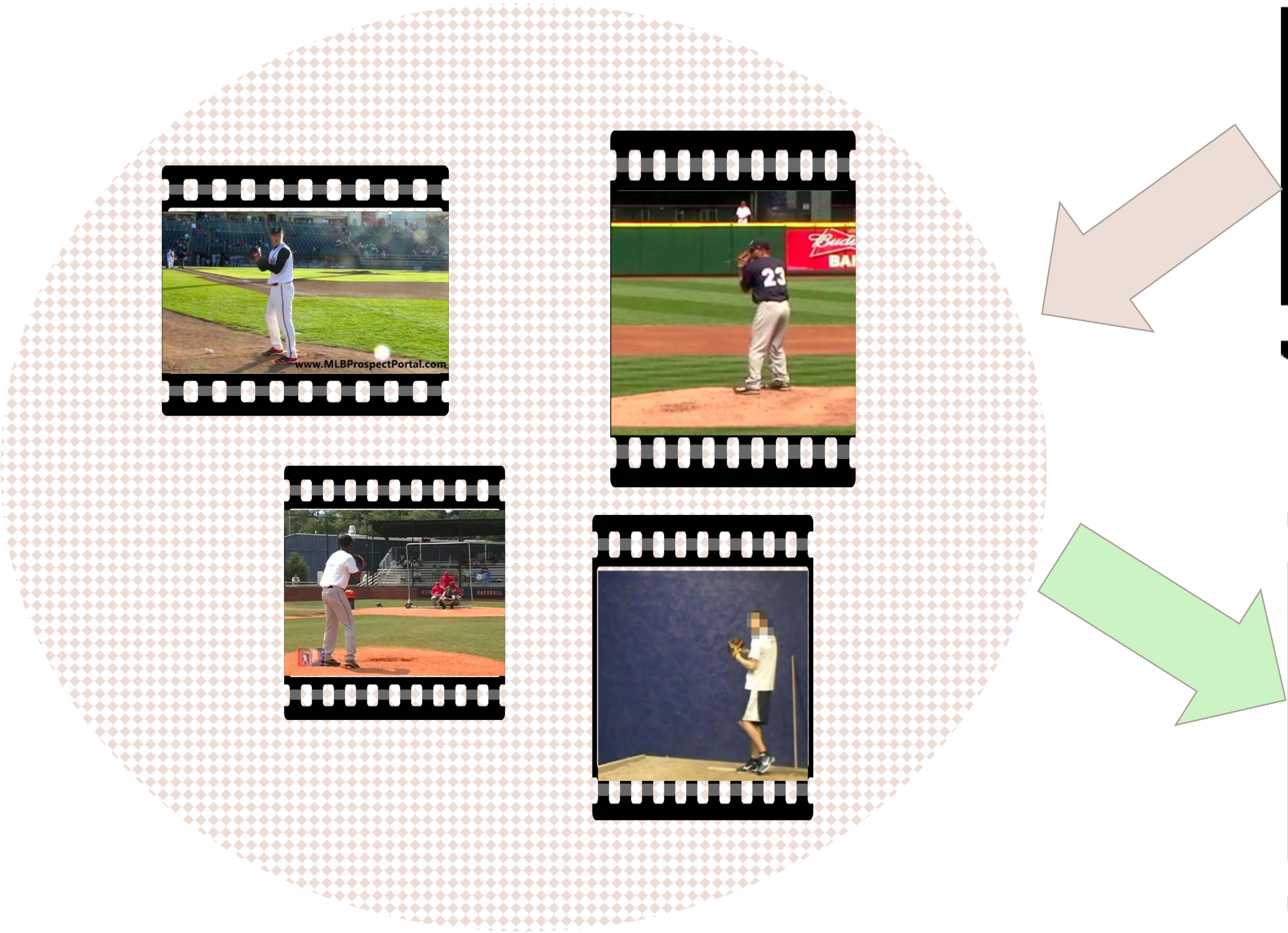
Coarse Understanding of Videos

Data Collection is Not Scalable

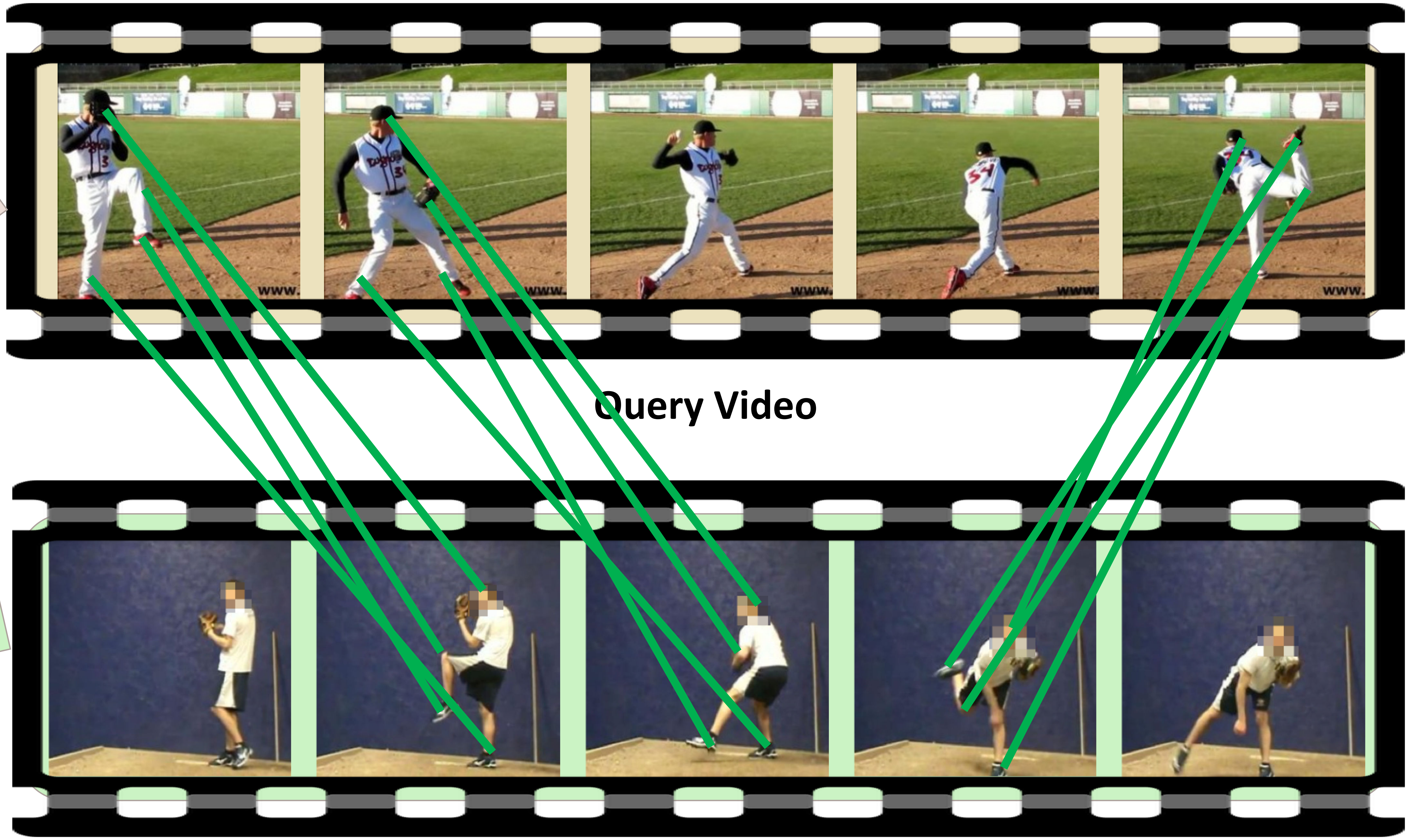
Figures taken from Kuehne et al. ICCV 2011; Krishna et al. ICCV 2017; Xiong et al. arXiv 2017.

Video Understanding via Association

Ask not "what is this?", ask "what is this like".
-Moshe Bar



Baseball Bowling



Query Video

Retrieved Video

Video Understanding via Association

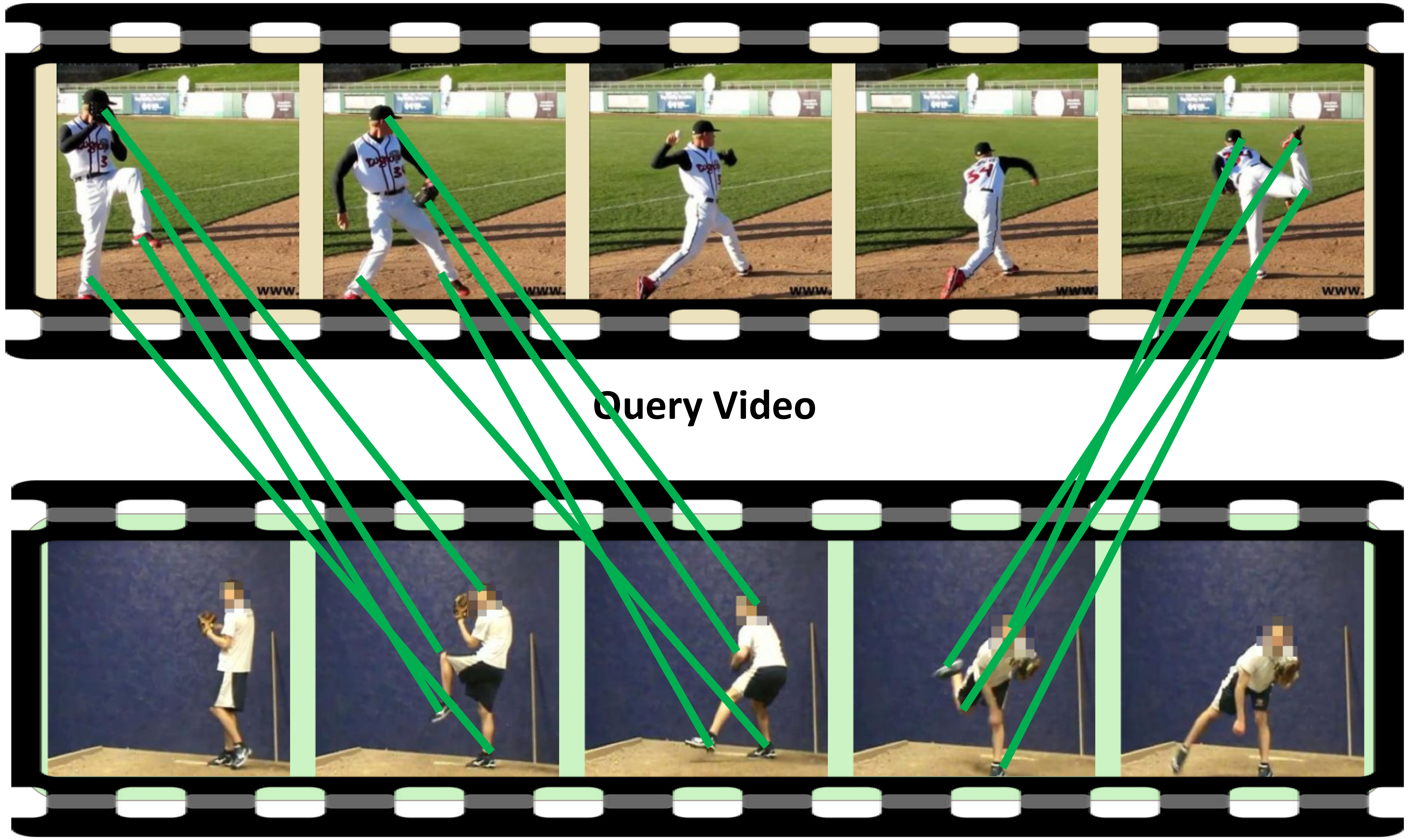
Ask not "what is this?", ask "what is this like".
-Moshe Bar

What does this achieve?

Describing the states object states in one video in terms of known reference videos

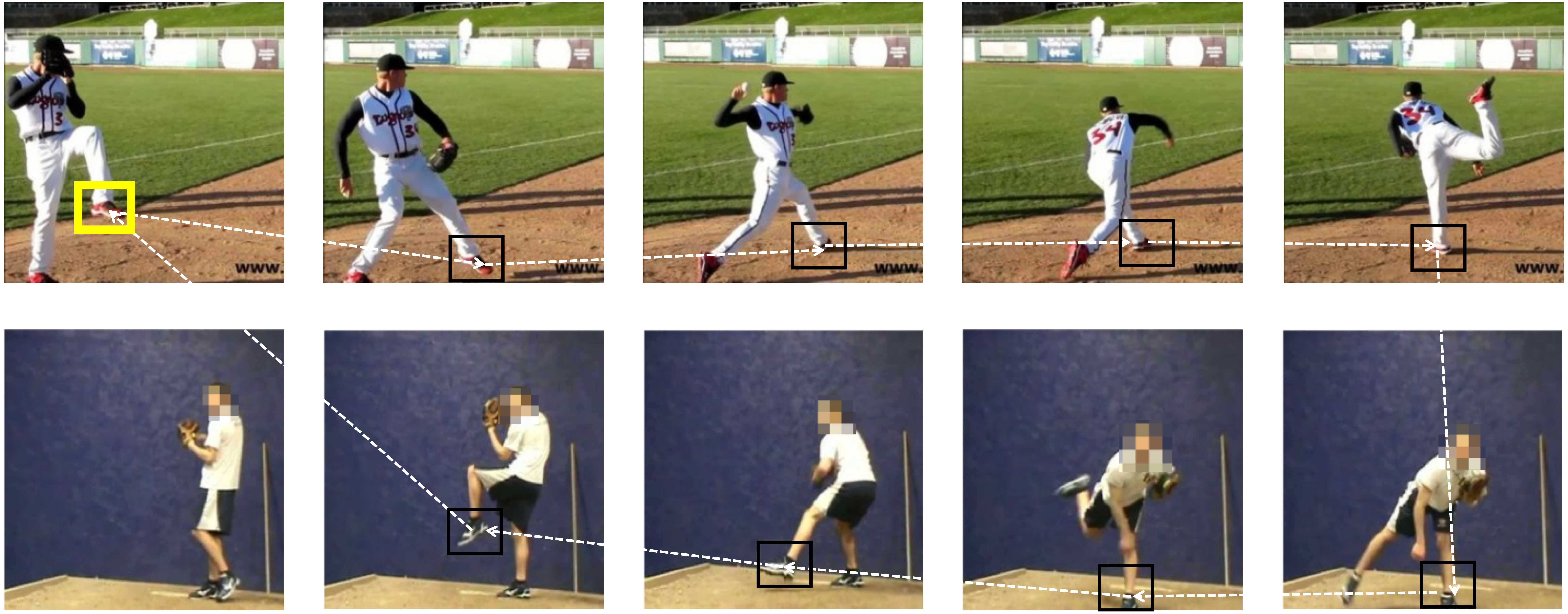
Any knowledge about the reference video can be transferred to the query video

Data collection for this is still infeasible!



Retrieved Video

Spatio-Temporal Associations Through Cycle Consistency



What is a cycle?

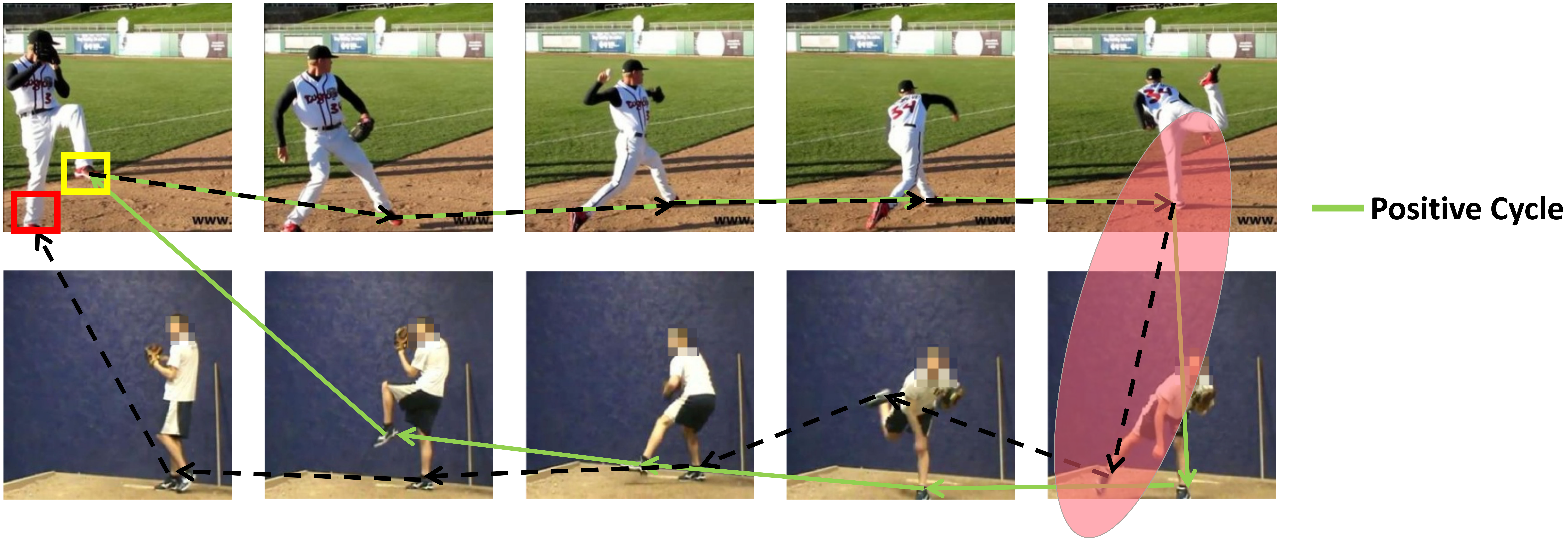
Match forward in time

Match to another video

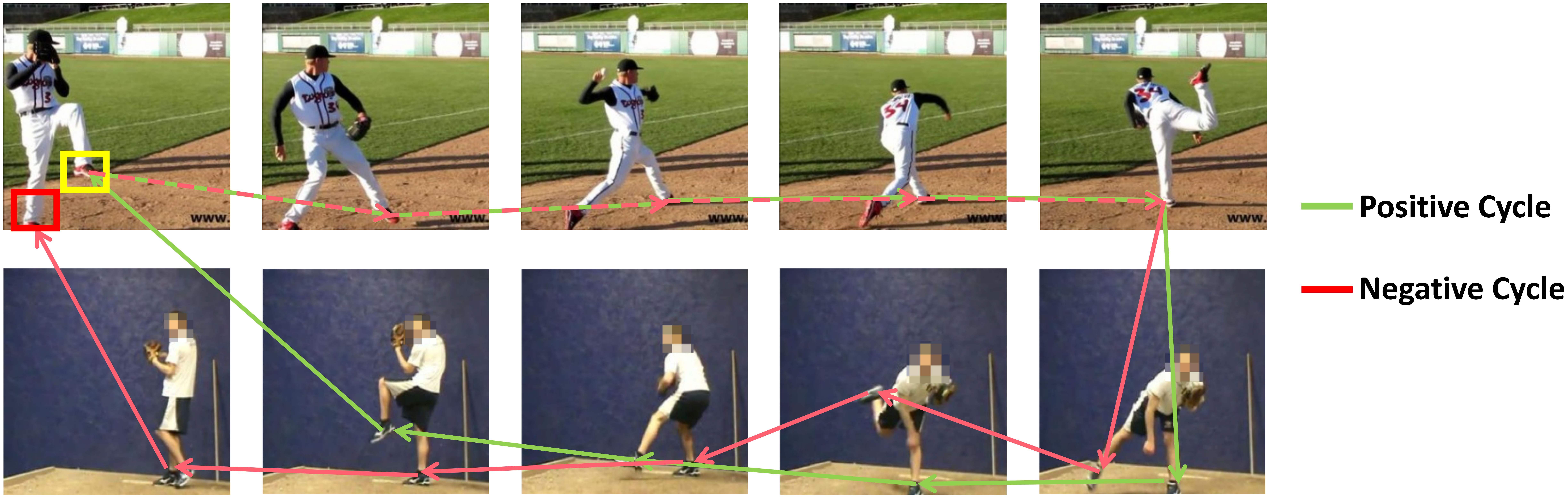
Match backward in time

Match back to first video

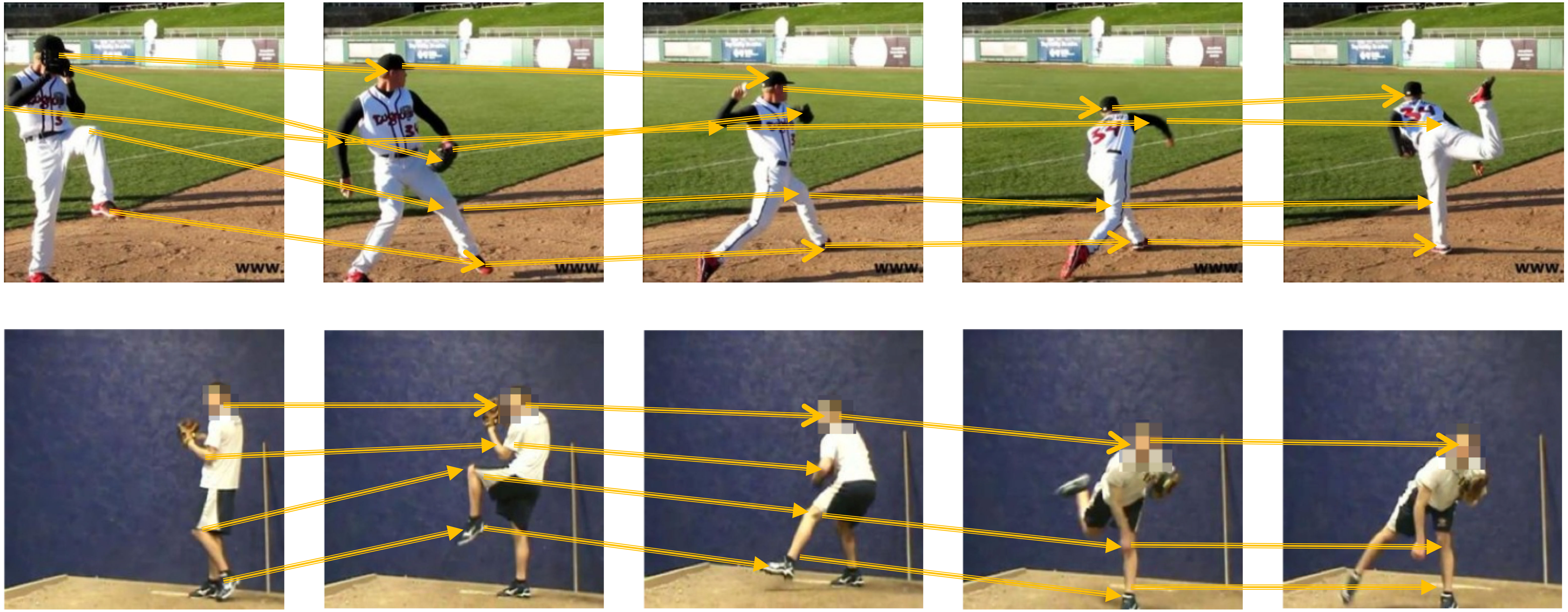
Spatio-Temporal Associations Through Cycle Consistency



Spatio-Temporal Associations Through Cycle Consistency



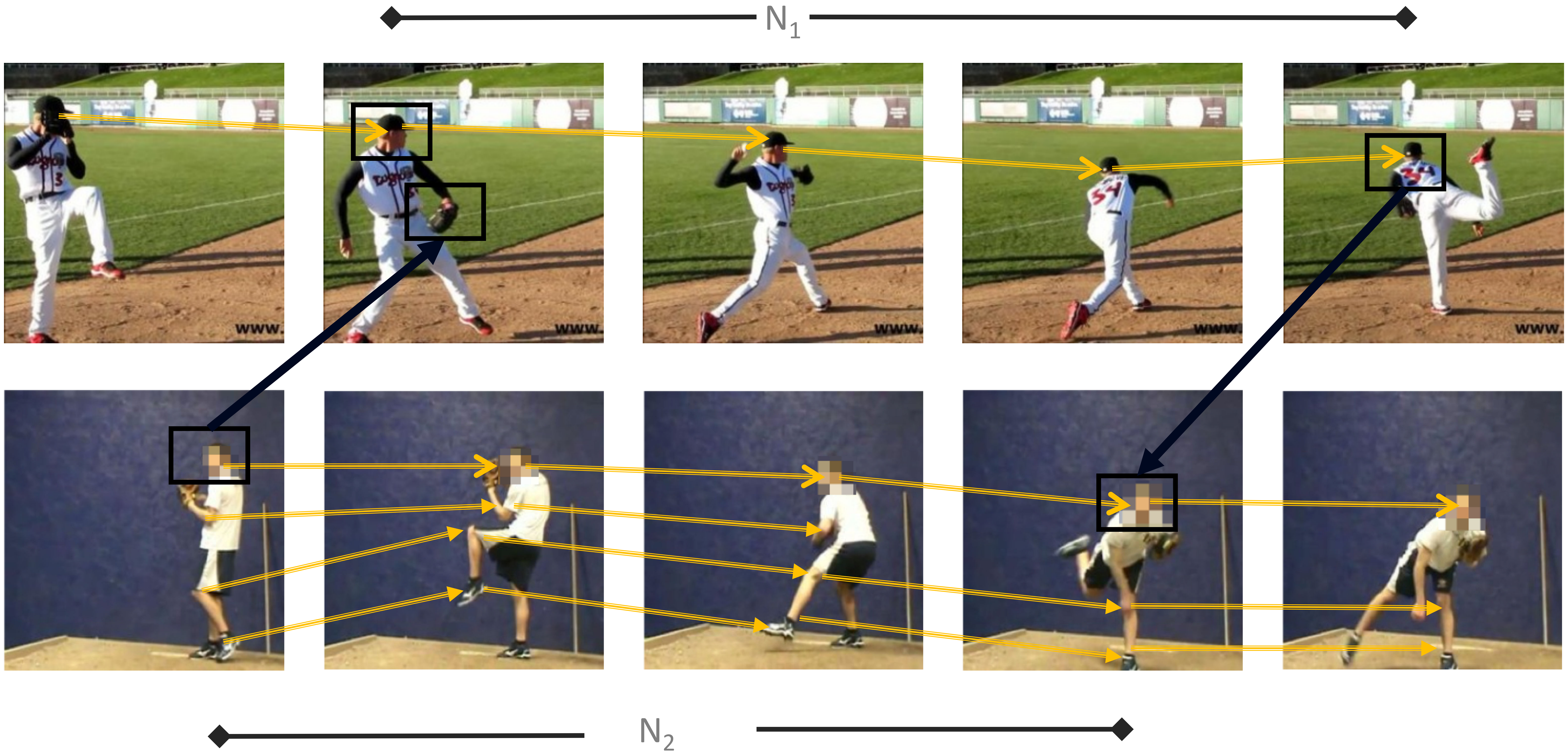
Spatio-Temporal Associations Through Cycle Consistency



Precompute tracks using an unsupervised tracker

Learning Correspondence From the Cycle-Consistency of Time
Xiaolong Wang, Allan Jabri, Alexei A. Efros; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2566-2576

Spatio-Temporal Associations Through Cycle Consistency



What is a cycle?

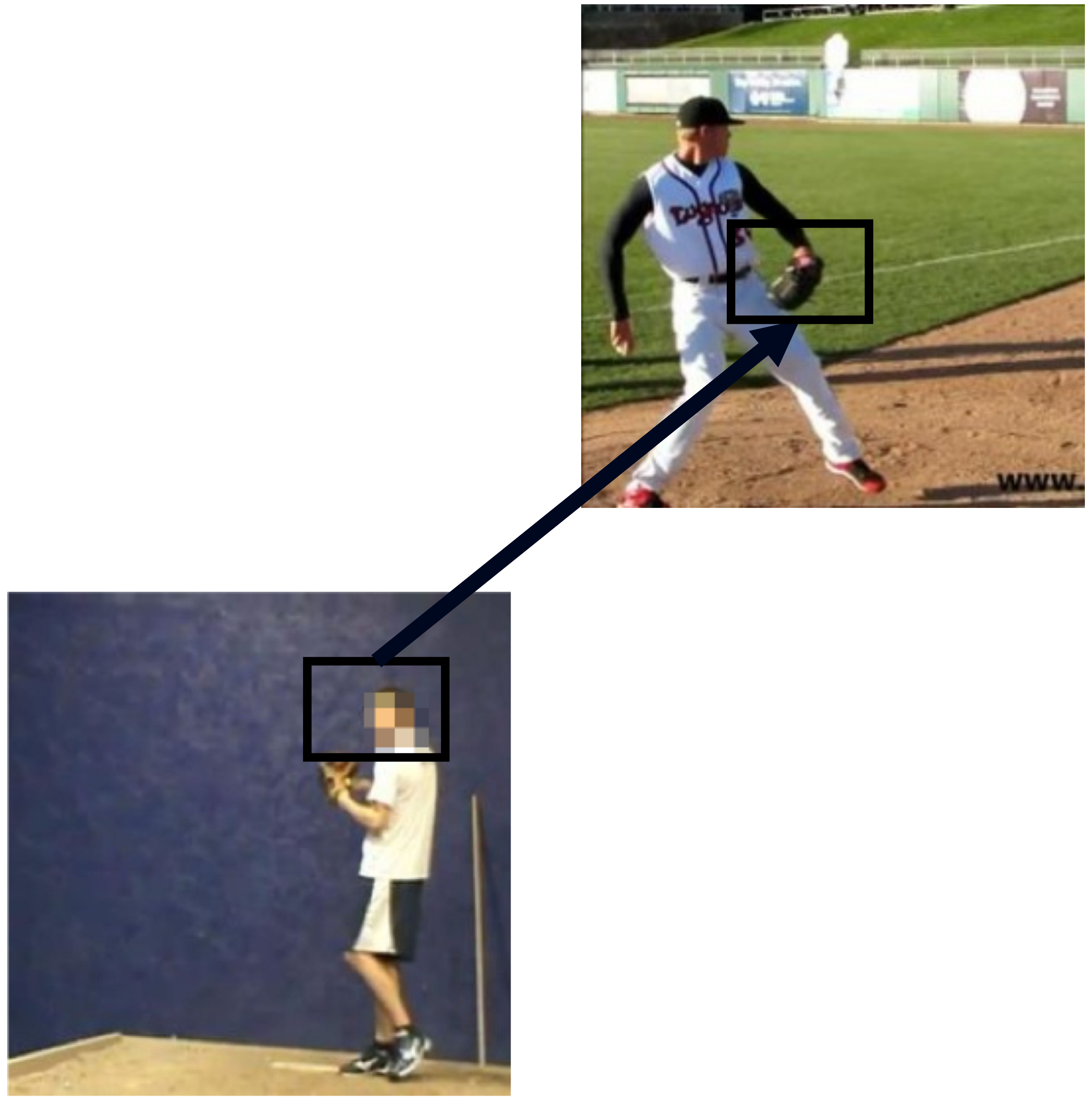
Follow track forward in time

Match to another video

Follow track backward in time

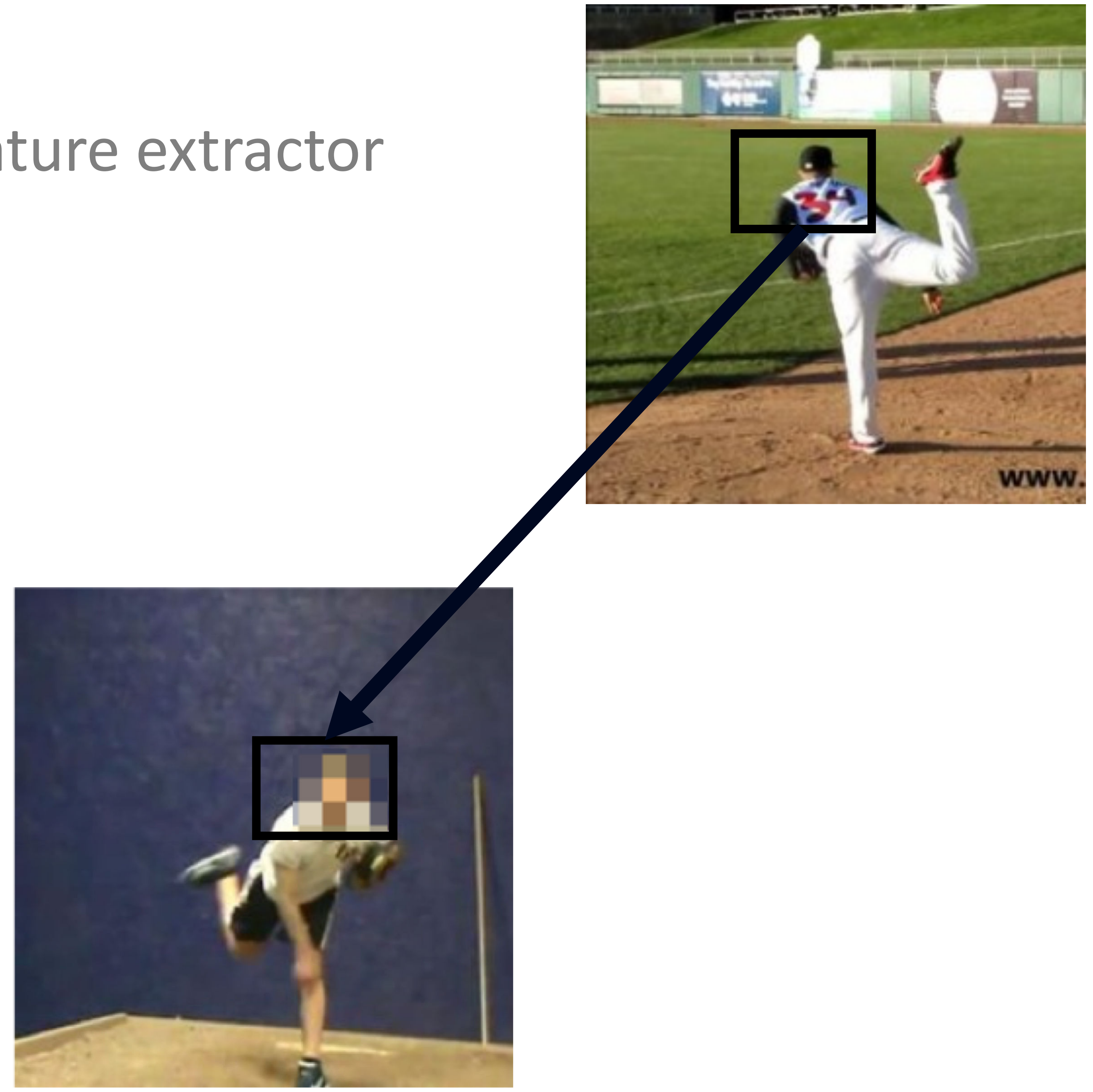
Match back to first video

Spatio-Temporal Associations Through Cycle Consistency



Depends on the feature extractor

$$f_\theta$$



What is a cycle?

Follow track forward N_1 frames

Match to another video

Follow track backward N_2 frames

Match back to first video

Spatio-Temporal Associations Through Cycle Consistency

Depends on the feature extractor

$$f_\theta$$

Score of a cycle

$$S_{21} \quad + \quad S_{12}$$

What is a cycle?

Follow track forward N_1 frames

Match to another video

Follow track backward N_2 frames

Match back to first video

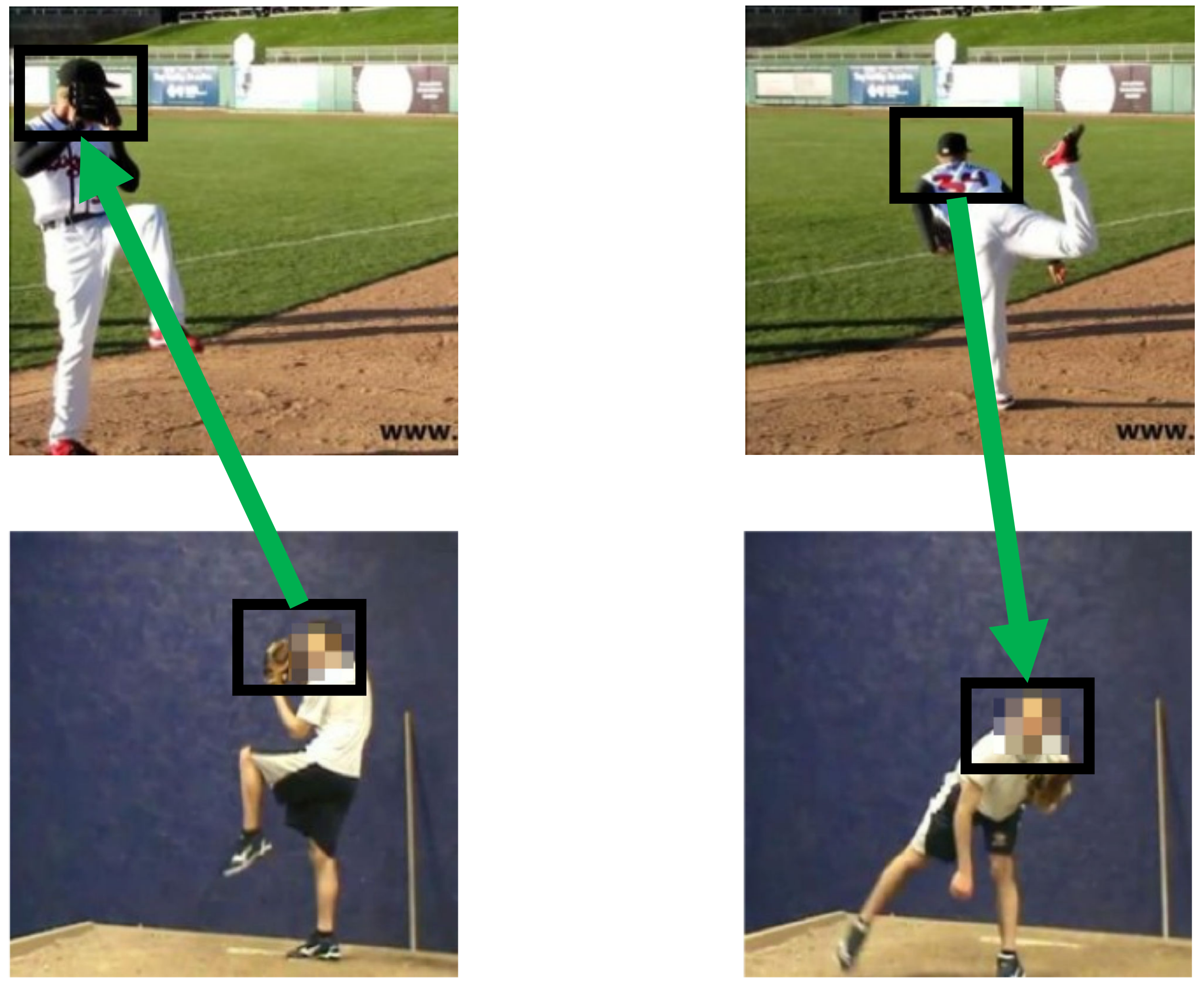
Spatio-Temporal Associations Through Cycle Consistency

Objective for training f_θ:

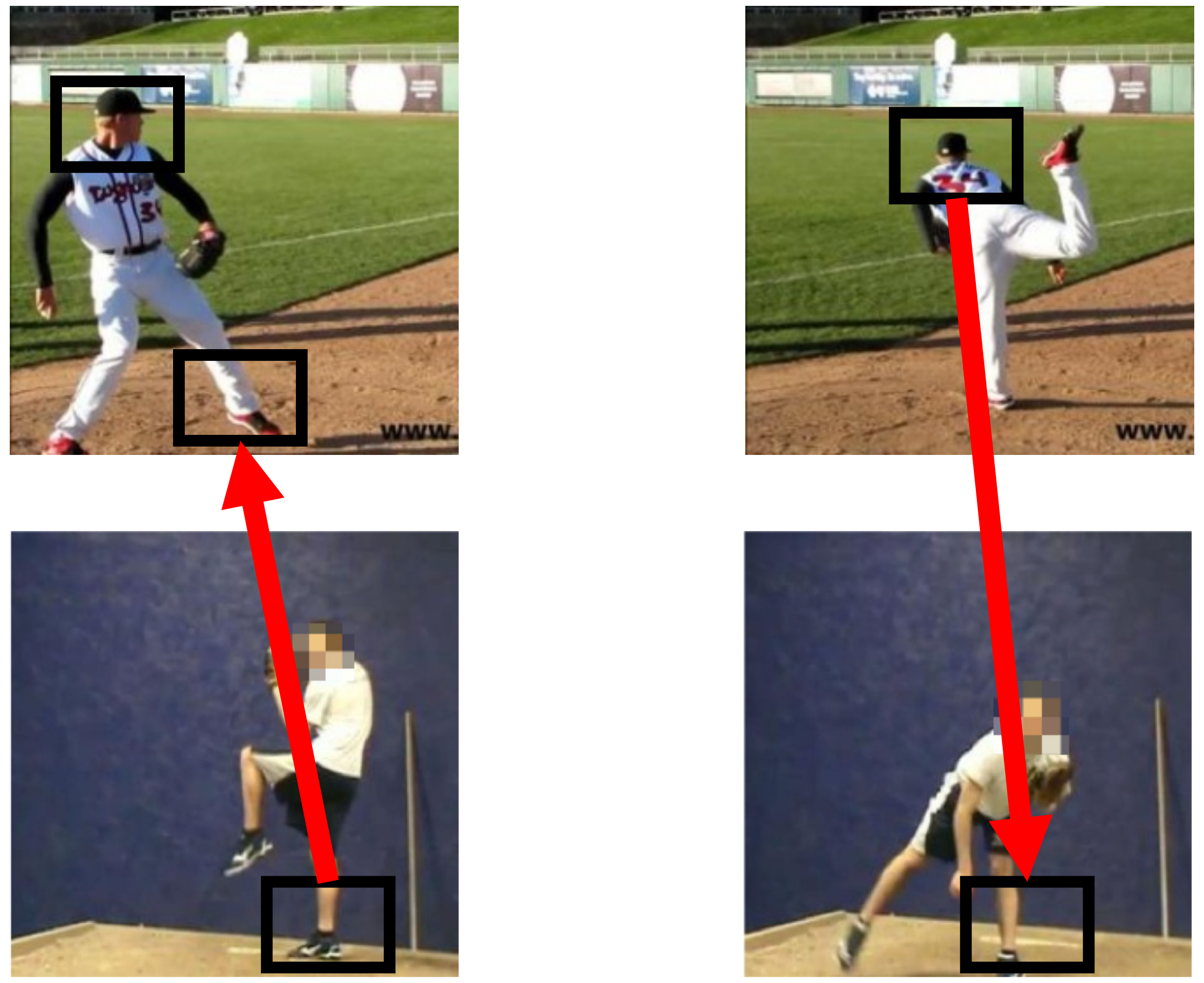
$$L = \max(0, S^- - S^+ + \delta)$$

For a given starting patch

S^+ = **Score of highest scoring Positive Cycle**



S^- = **Score of highest scoring Negative Cycle**



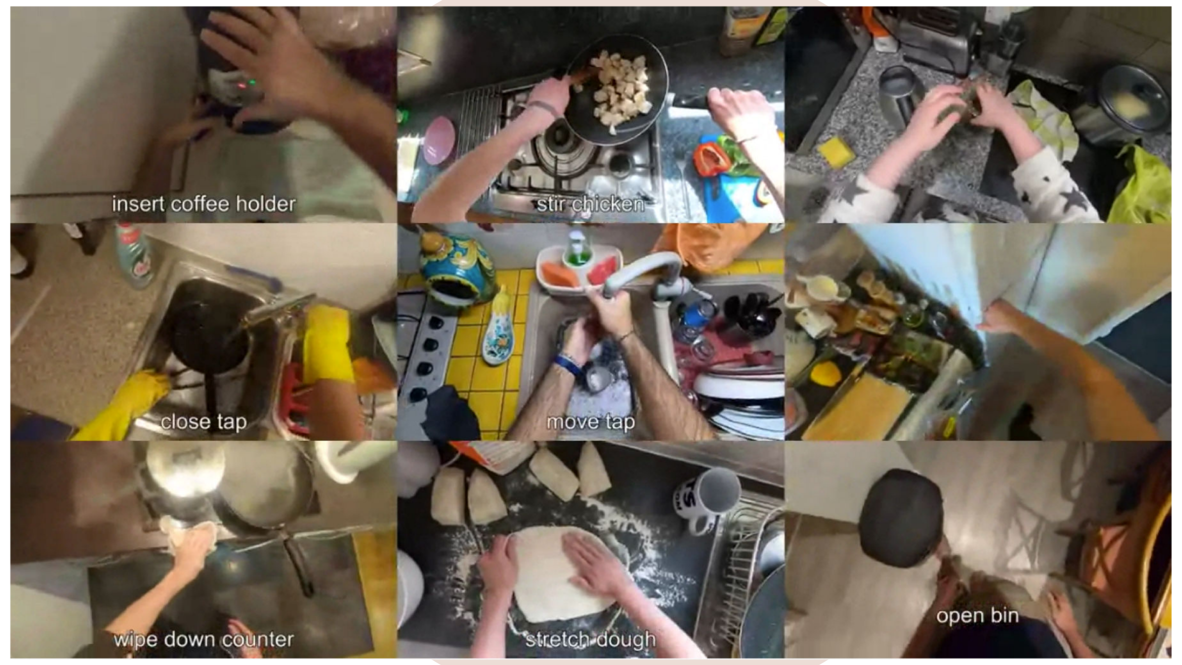
Training Datasets

Penn Action Dataset[1]

Videos depicting 15 different actions with human joint annotations

Pouring Dataset[2]

Videos depicting pouring from one container into another



Epic Kitchens Dataset[3]

First-person videos depicting activities in kitchens

1. Weiyu Zhang, Menglong Zhu and Konstantinos Derpanis, "From Actemes to Action: A Strongly-supervised Representation for Detailed Action Understanding" International Conference on Computer Vision (ICCV). Dec 2013.
2. Sermanet, Pierre, Kelvin Xu, and Sergey Levine. "Unsupervised perceptual rewards for imitation learning." *arXiv preprint arXiv:1612.06699* (2016).
3. Damen, Dima, et al. "Scaling egocentric vision: The epic-kitchens dataset." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

Qualitative Evaluation: Patch Nearest Neighbor

A representation that can **encode patch appearance** while **accounting for object states**

Qualitative Results: Spatio-Temporal Alignment

Learned representation can effectively ***spatio-temporally align videos***

Aligning Patches
Choose tracks that form high scoring cycles

Aligning Frames
Frames with high cumulative patch alignment scores

Quantitative Results: Spatio-Temporal Alignment

Temporal Alignment Error
Mean difference in joint angles between aligned frames

Spatial Alignment Accuracy
Accuracy of aligning keypoint patches
(within some neighborhood)

Initialization Method	Temporal Alignment Err	Spatial Alignment Acc
ImageNet	0.509	0.153
Mask-RCNN [1]	0.504	0.202
Unsupervised Tracker [2]	0.501	0.060
Kinetics Action Classification Model	0.492	0.150
Penn Action Classification Model	0.521	0.157
Our features	**0.448**	**0.284**

Summary

A spatio-temporal alignment formulation for
dense video understanding via association to known videos

A method to
learn representations using cycle-consistency

Demonstrate that the learned
representation encodes object appearance and object states

Demonstrate that the proposed approach can be
effectively used to spatio-temporally align videos

Thank you for listening!
Checkout our project paper for relevant links:
http://www.cs.cmu.edu/~spurushw/publication/alignvideos/