CVPR 2016

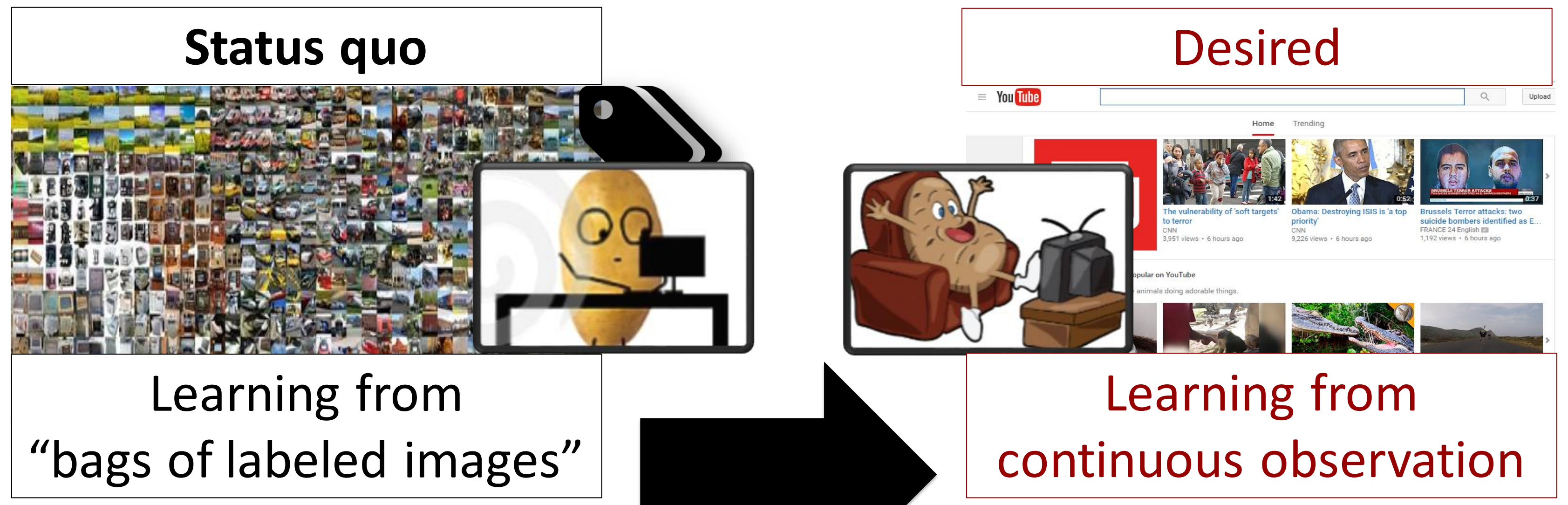
Slow and steady feature analysis: higher order temporal coherence in video

Dinesh Jayaraman and Kristen Grauman
UT Austin

THE UNIVERSITY OF TEXAS AT AUSTIN

Problem

Learning *unsupervised* generic visual features from unlabeled video



Status quo

Learning from "bags of labeled images"

- Expensive
- Limited data
- Task-specific

Desired

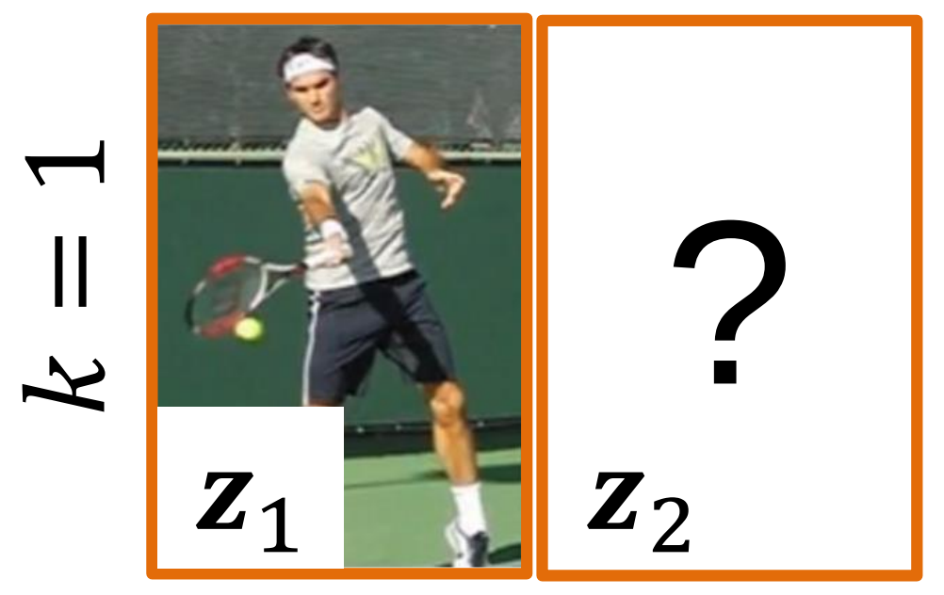
Learning from continuous observation

- Free
- Unlimited data
- Generic

Learning from natural world temporal dynamics

Prior work

Slowness: "concepts only change slowly over time", i.e., given $k = 1$ video frame, next frame is close in a semantic space $\boldsymbol{z}$:

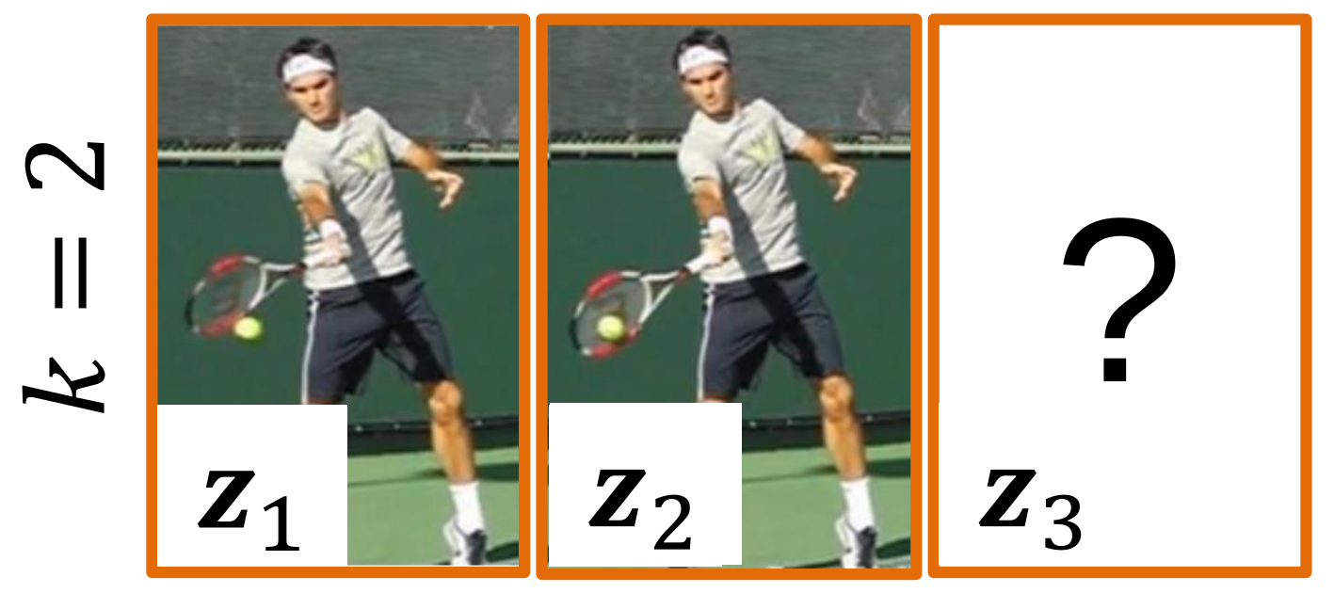


Predict: $\widehat{z_2} \approx z_1$

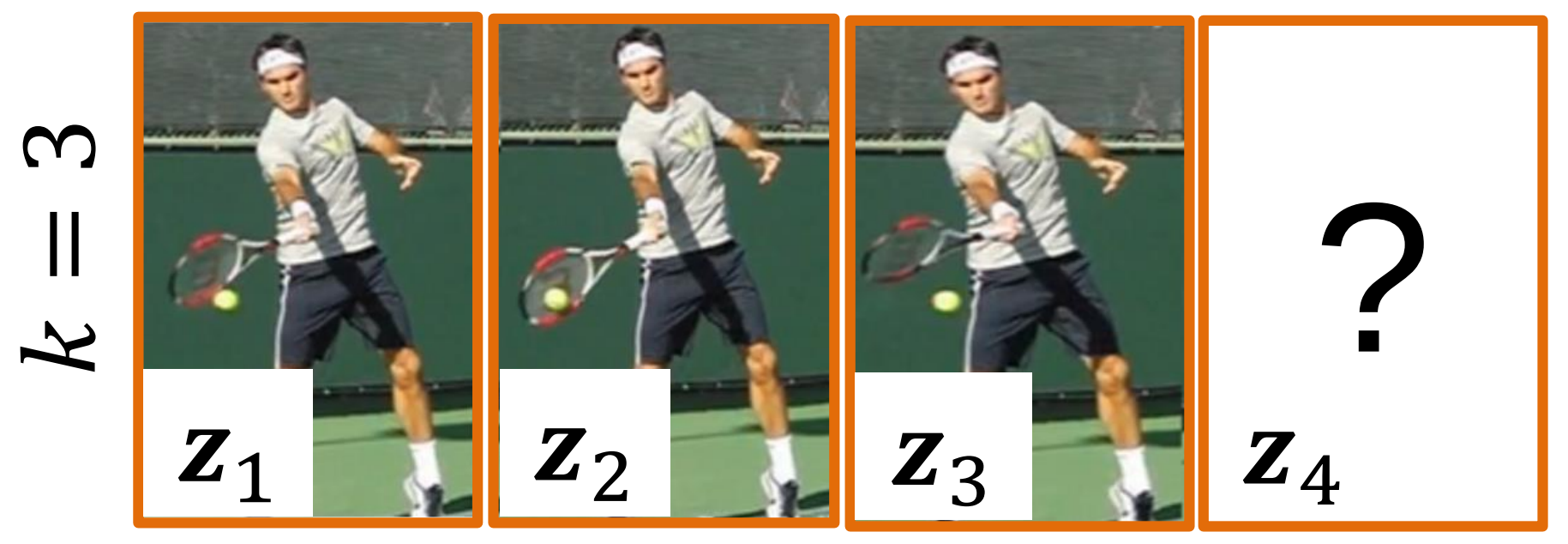
[Wiskott 2002], [Hadsell 2006], [Mobahi 2009], [Bergstra 2009], [Goroshin 2013], [Wang 2015] …

Our idea

Steadiness: "concepts evolve '*steadily*' over time", i.e., given k frames, perform k^{th} order extrapolation to guess next frame:



Predict: $\widehat{z_3} \approx z_2 + (z_2 - z_1)$



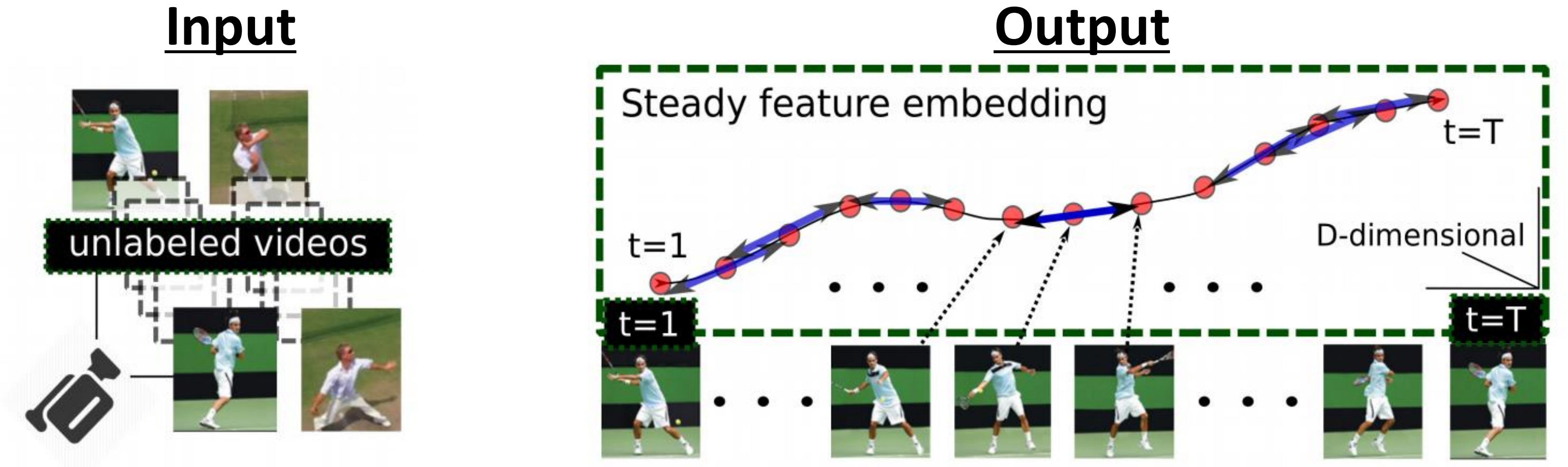
Predict: $\widehat{z_4} \approx z_3 + (z_3 - z_2) + ((z_3 - z_2) - (z_2 - z_1))$

⋮

Idea: train a feature mapping $\boldsymbol{z}(.)$ to be a "steady" semantic space i.e. k^{th} order extrapolation must predict future frames well:

- Induces desirable feature properties such as invariance ($k = 1$), equivariance ($k = 2$) and so on.

Target representation

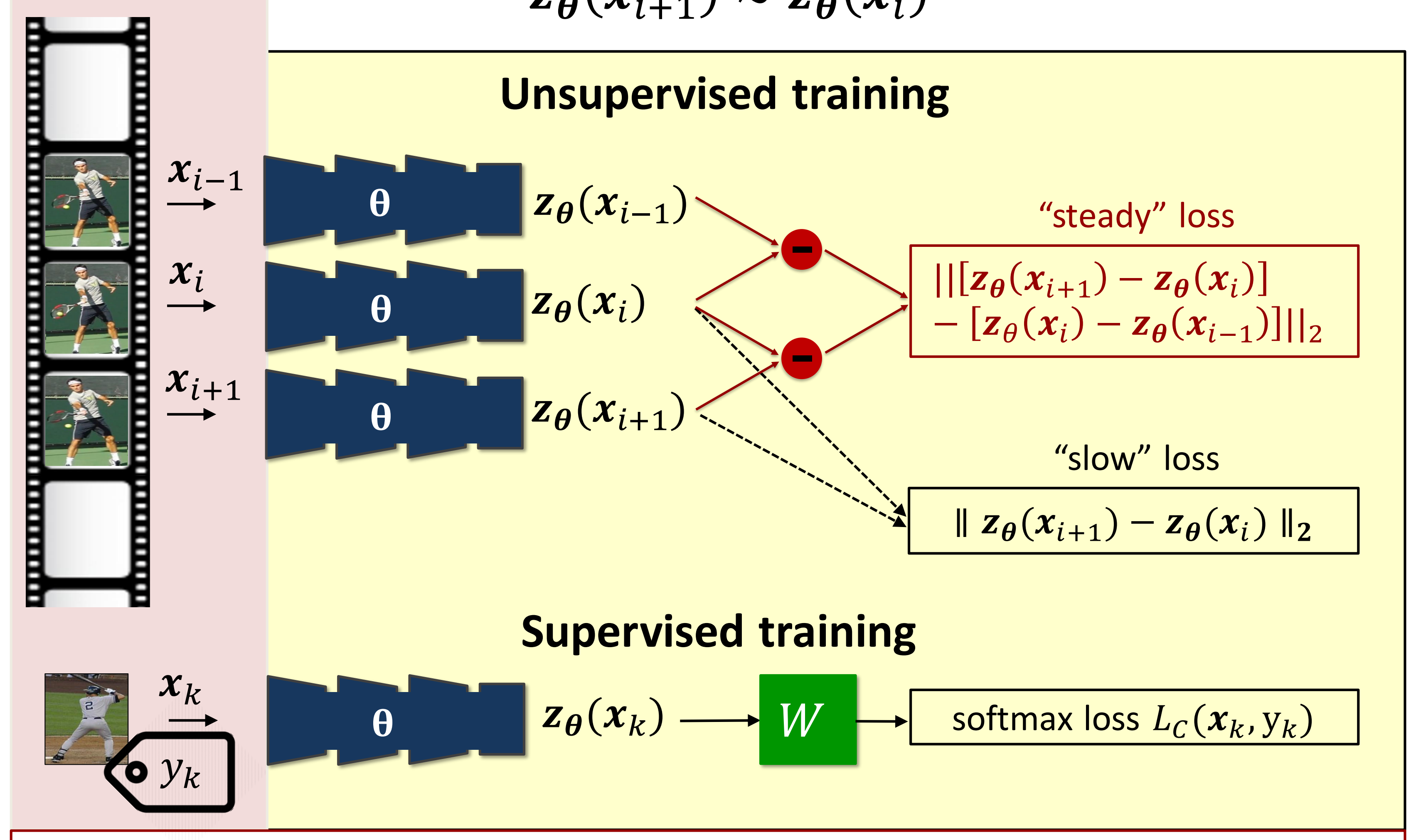


"K^{th} order steady" features $\boldsymbol{z} \Rightarrow \frac{dz}{dt} \approx \boldsymbol{0}, \frac{d^2z}{dt^2} \approx \boldsymbol{0}, \ldots, \frac{d^Kz}{dt^K} \approx \boldsymbol{0}$

Steady representation learning

Desired: Feature space with second-order temporal coherence:

Given: $z_\theta(x_{i+1}) - z_\theta(x_i) \approx z_\theta(x_i) - z_\theta(x_{i-1})$, and
$z_\theta(x_{i+1}) \approx z_\theta(x_i)$



Training within a Siamese triplet neural network architecture

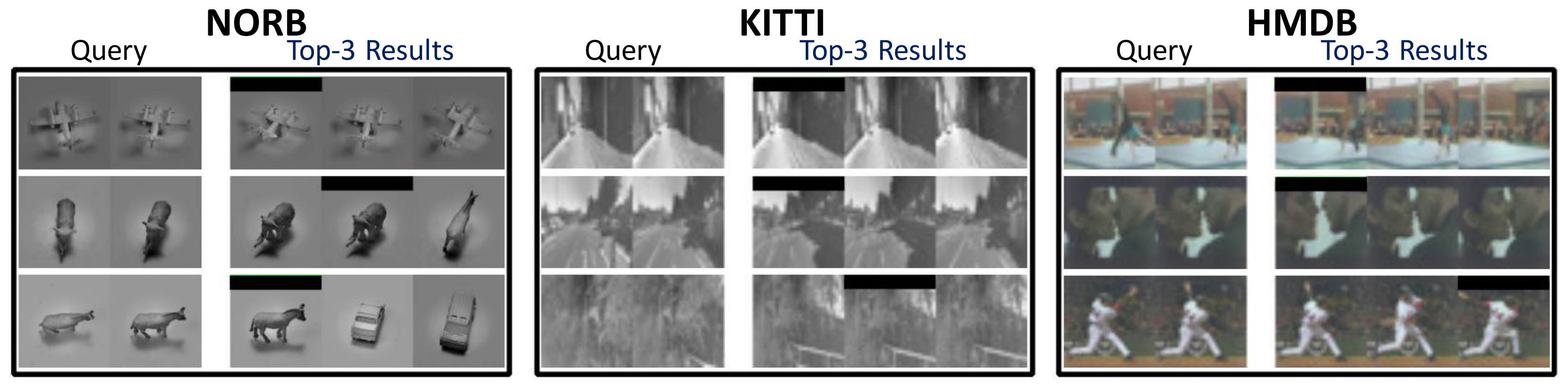
Datasets

unlabeled video

target task

Task	Img/frame dims	#Classes	Recog. Task	#Train	#Test	Unsup. Input Type	#Pairs (1:3)	#Triplets (1:1)
NORB→NORB	96×96×1	25	object	150	8100	pose-reg. images	50,000	75,000
KITTI→SUN	32×32×1	397	scene	2382	7940	car-mounted video	100,000	100,000
HMDB→PASCAL-10	32×32×3	10	action	50	2000	web video	100,000	100,000

Sequence completion



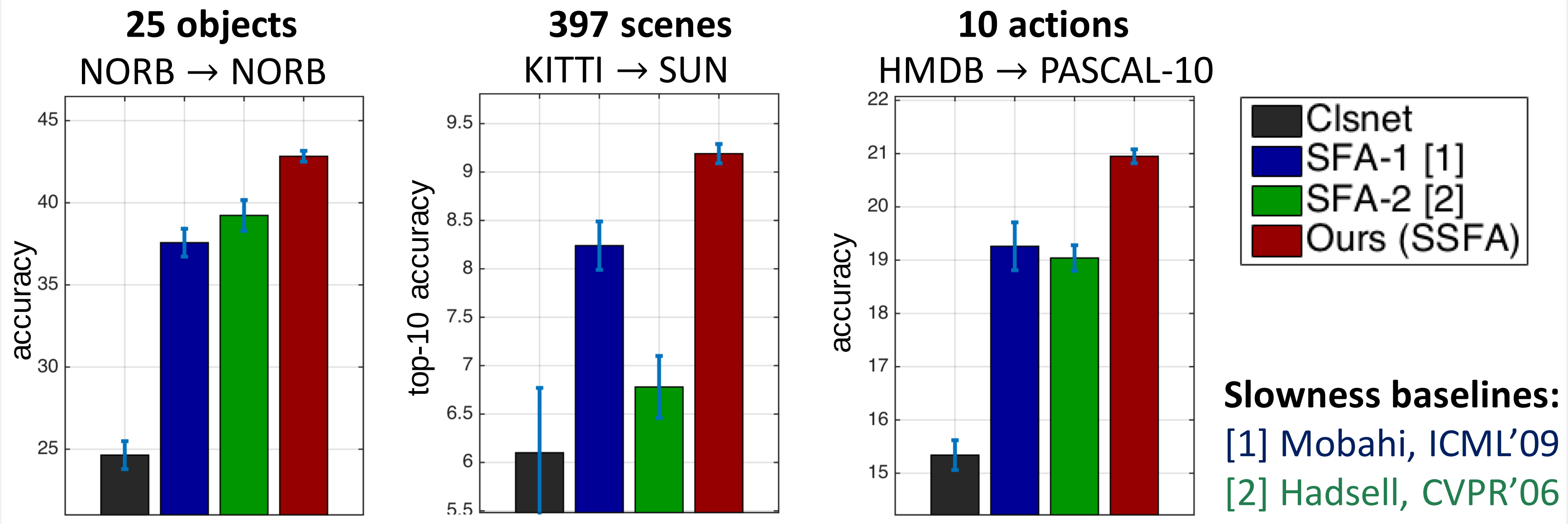
Problem: Given two frames, find third in sequence, as: $z_3 = z_2 + (z_2 - z_1)$

Score: $\eta = E\left[\frac{rank}{\#candidates}\right] \times 100$

Datasets	NORB	KITTI	HMDB
SFA-1 [1]	0.95	31.04	2.70
SFA-2 [2]	0.91	8.39	2.27
SSFA (Ours)	**0.53**	**7.79**	**1.78**

Qualitative and quantitative feature "steadiness" verification

Regularized category recognition from few samples

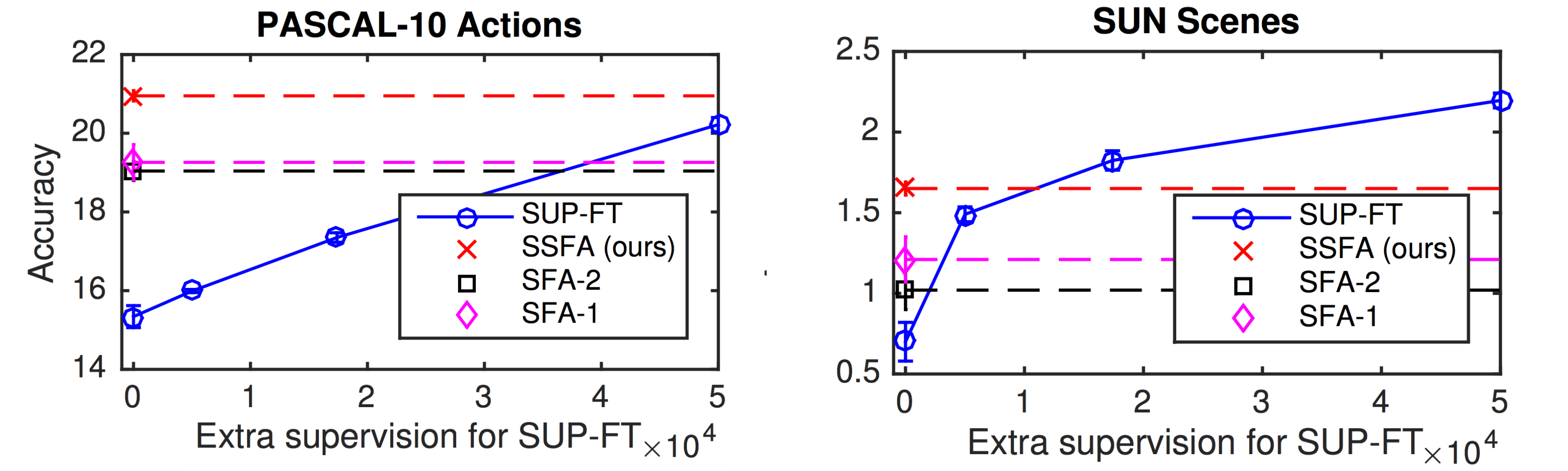


Slowness baselines:
[1] Mobahi, ICML'09
[2] Hadsell, CVPR'06

Strong and consistent accuracy gains for *higher-order* temporal coherence vs. *slow* feature learning methods

Unsupervised vs. supervised pretraining and finetuning

Pretraining on unlabeled video vs. labeled CIFAR-100 images:



Our unsupervised features can even surpass supervised pretraining with up to 50,000 additional class labels for an auxiliary task!