Video Understanding

Asim Kadav

Researcher, NEC Labs America

Topics in Deep Learning: Methods and Biomedical Applications
Spring 2020

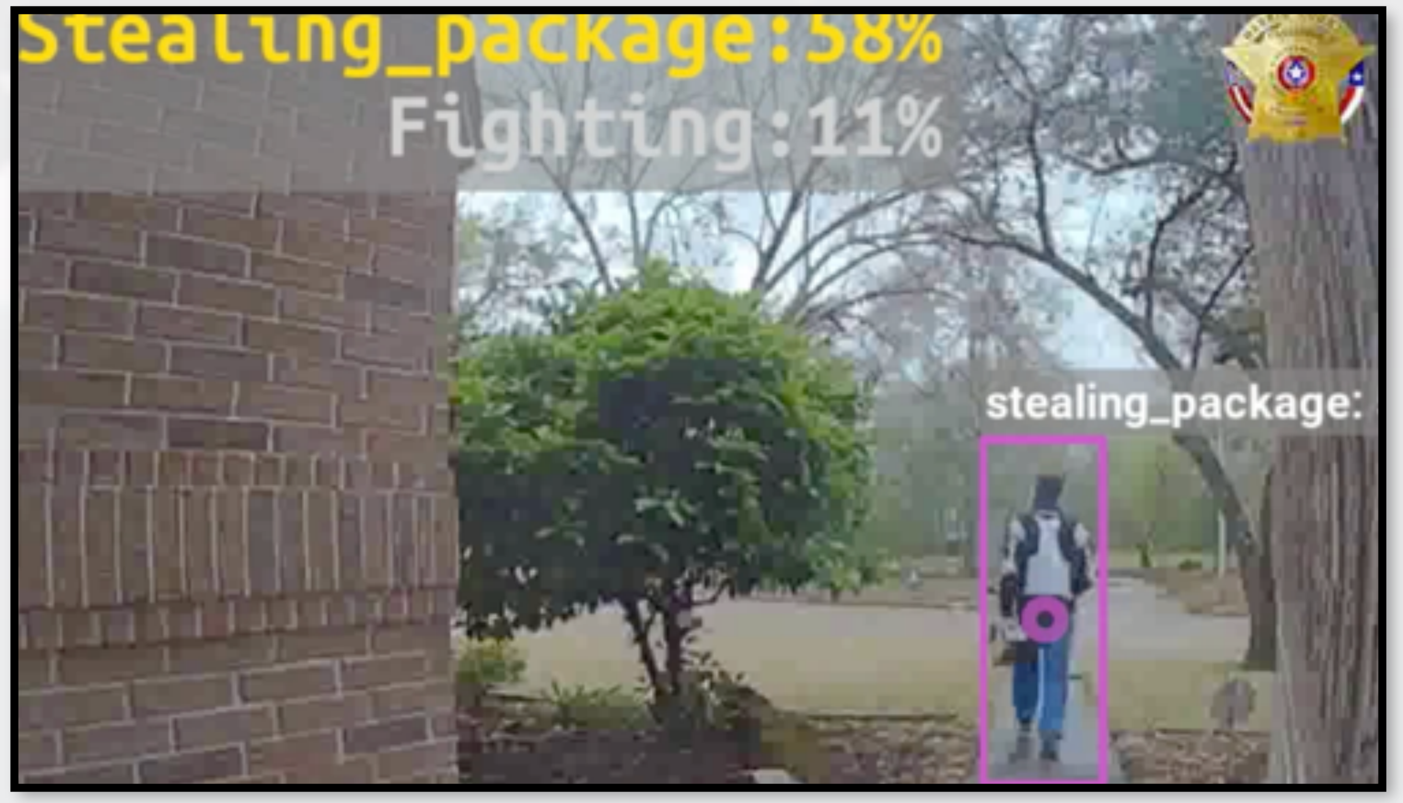
MOTIVATION: WIDESPREAD VIDEO-BASED APPLICATIONS

Video organization

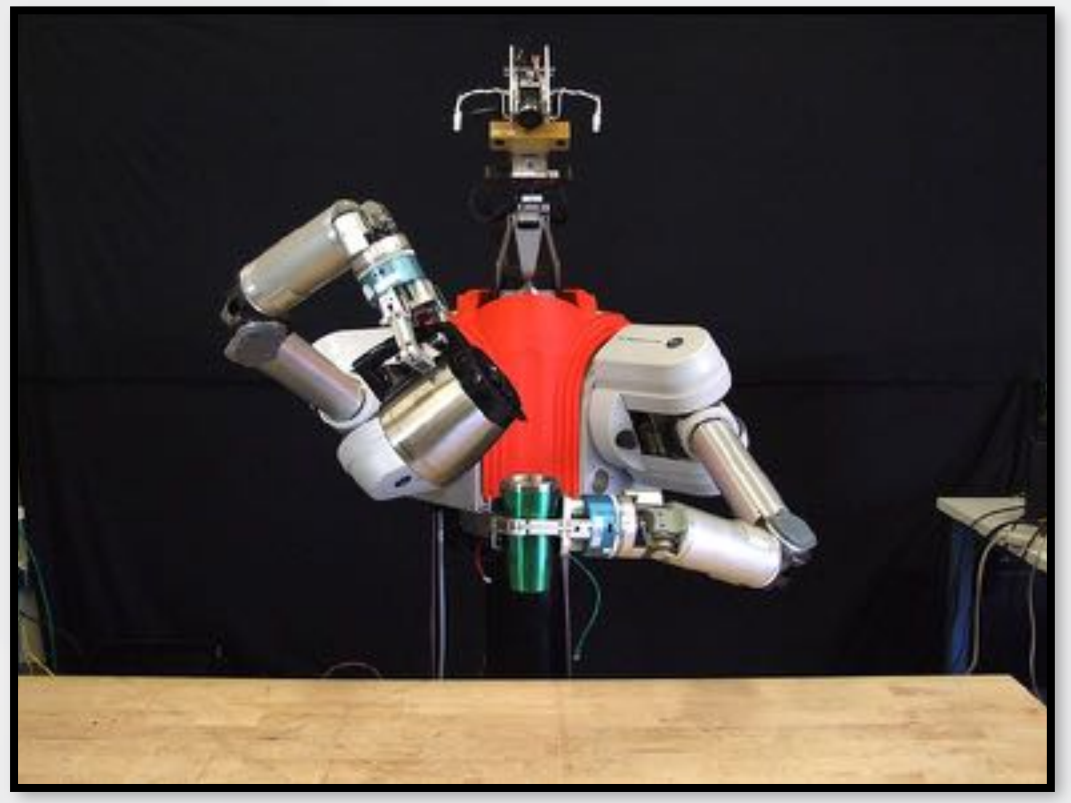
Retail intelligence

AR/VR

Public safety



Home safety



Robotic manipulation

SOME VIDEO UNDERSTANDING TASKS AND DATASETS

Segmentation

YouTube-VOS

Skeleton

COCO
PoseTrack’17

Detection/ Tracking

VOT’18

Classification

Youtube 8M
Kinetics
Charades
Something something

Localization

AVA
Charades

Language

ActivityNet Captions
Charades-STA
TALL

Key challenges:

1) Understand the higher-level spatio-temporal concept in the overall video snippet
2) Understand temporal motion of objects to reduce inaccuracies due to occlusion, background clutter, lighting conditions as objects

WHAT IS A VIDEO?

3D video tensor

Sequence of frames representing temporal motion

Each second represented by multiple frames (FPS)

Each color pixel represented by 3 channels (R, G, B)

Fine-grained spatial relationship

Short and long temporal relationships

Example: Video from kinetics dataset

10 seconds x 720p (1280x720)

Raw space necessary:

3 (Channels)x 8 bits per channel x 1280 x 720 x 10 seconds * 15 FPS = 395 MB

After compression

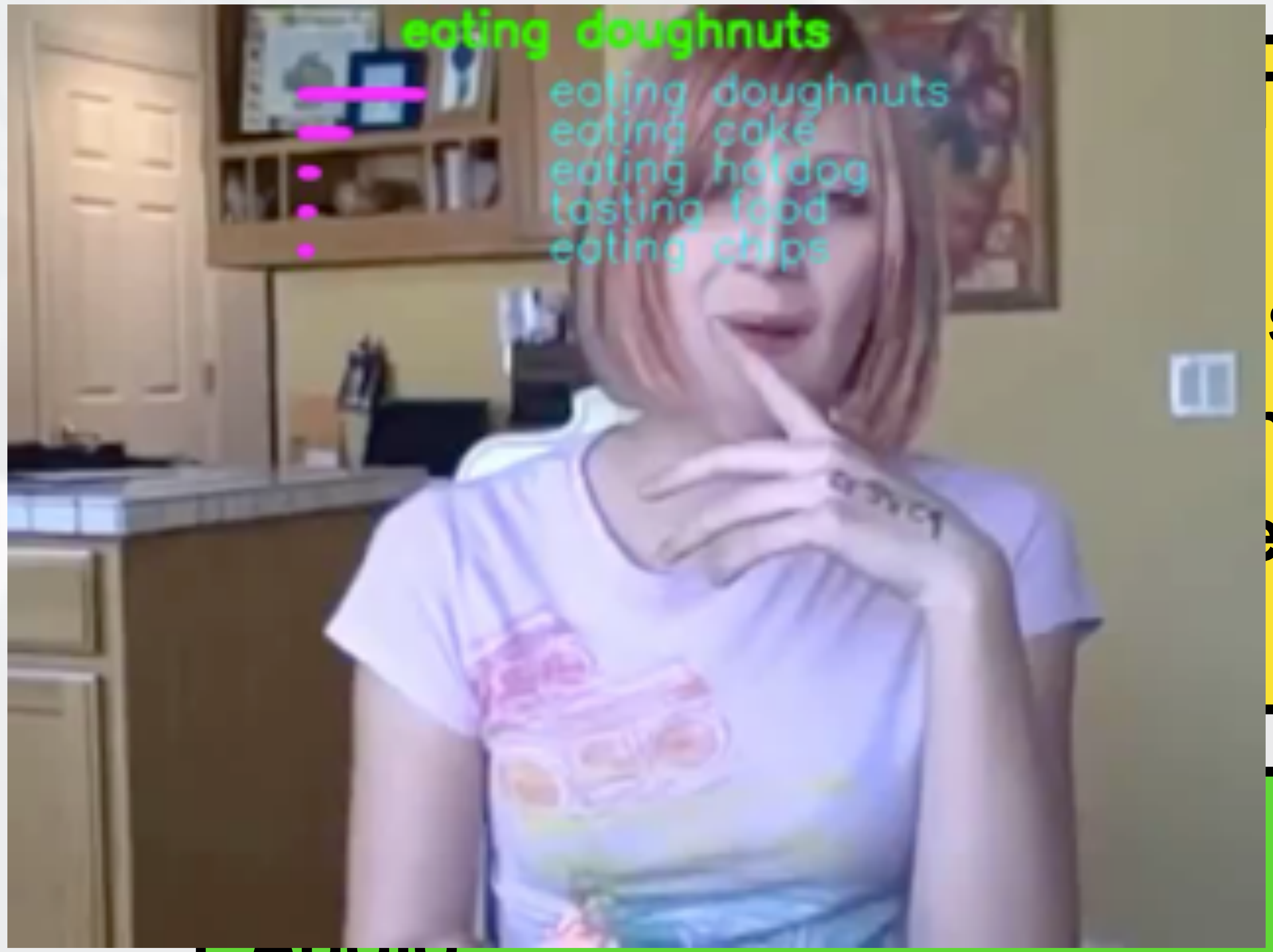
~ 5 MB (H.264)

- High encoding cost but supported by most modern processors
- But ML algorithms operate on raw frames (~395 MB every 10s)

ACTION RECOGNITION

- Action recognition (video classification) is the most well studied video understanding task
- Most interesting videos (and complex motion) are based around human actions

What spatio-temporal features does the model need to learn?



Temporal

s in motion
n tracking
etry

Audio
Action length
Specific combination of all of the above

PRE-DEEP LEARNING APPROACHES — SVM BASED

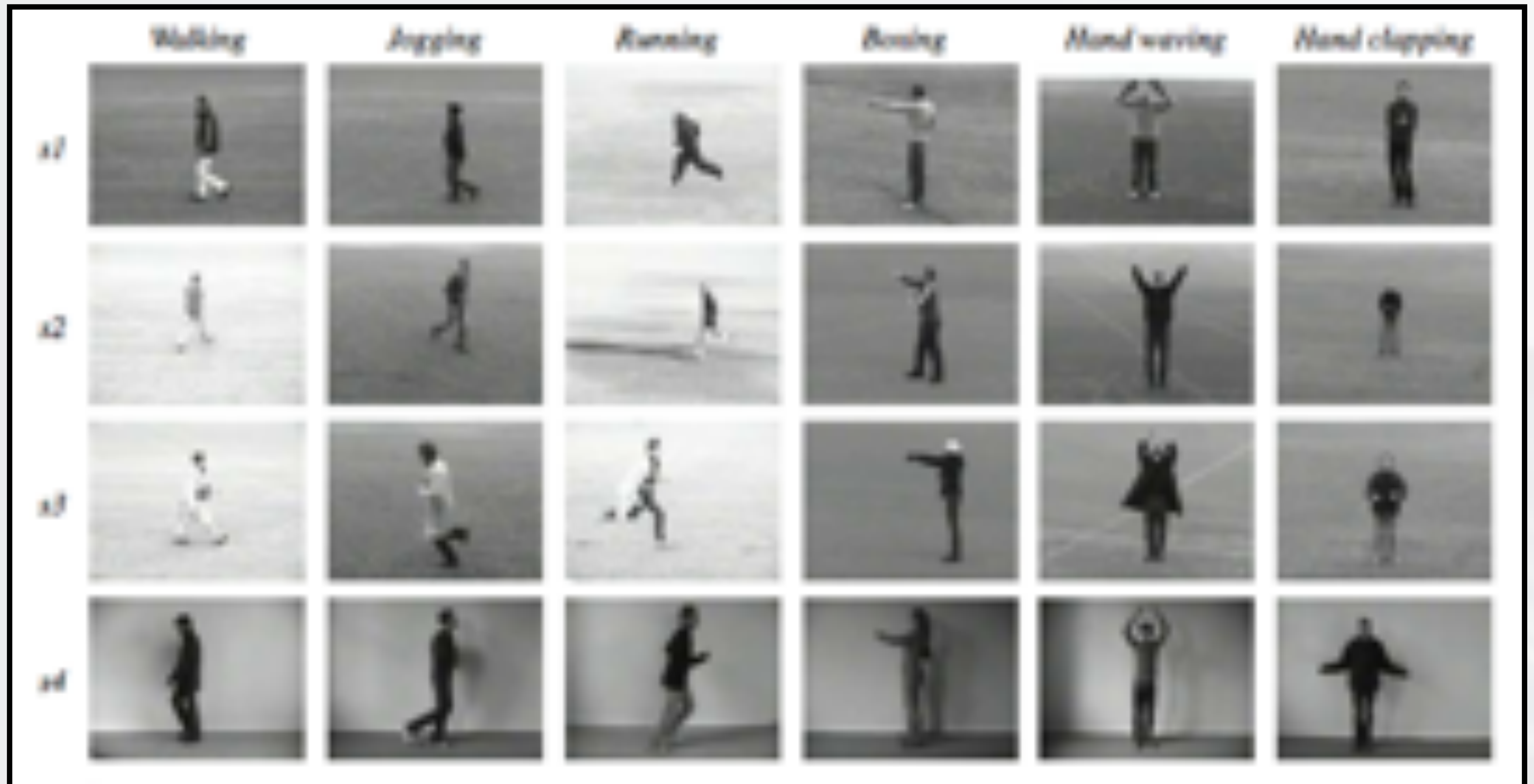
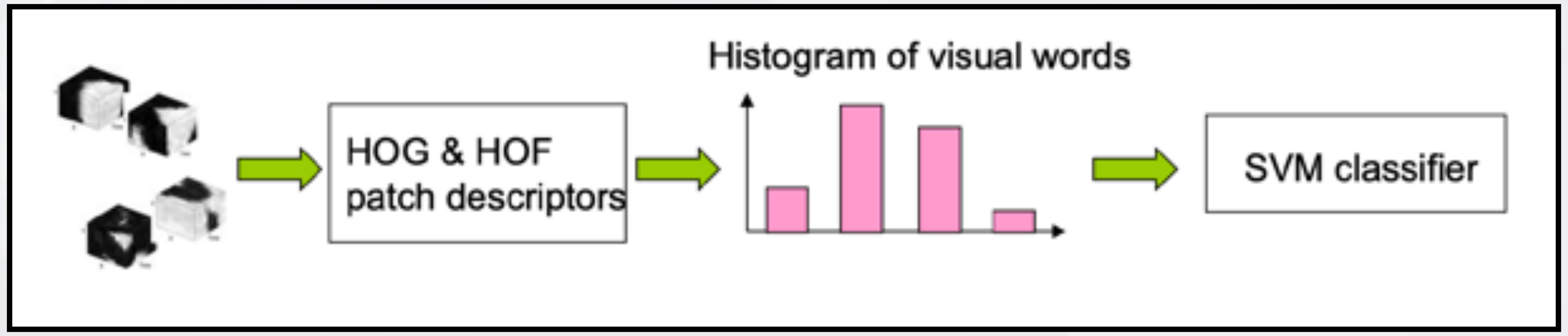


Figure 2. Action database (available on request): examples of sequences corresponding to different types of actions and scenarios.



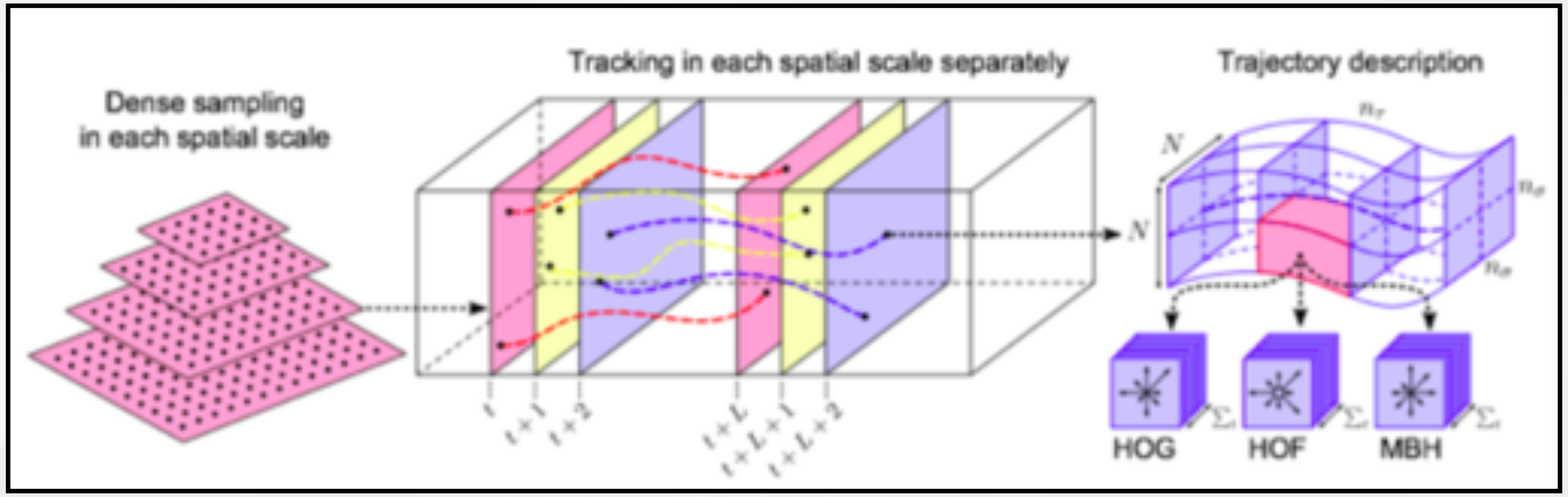
Recognizing Human Actions: A Local SVM Approach
Schuldt et. al. (2004)

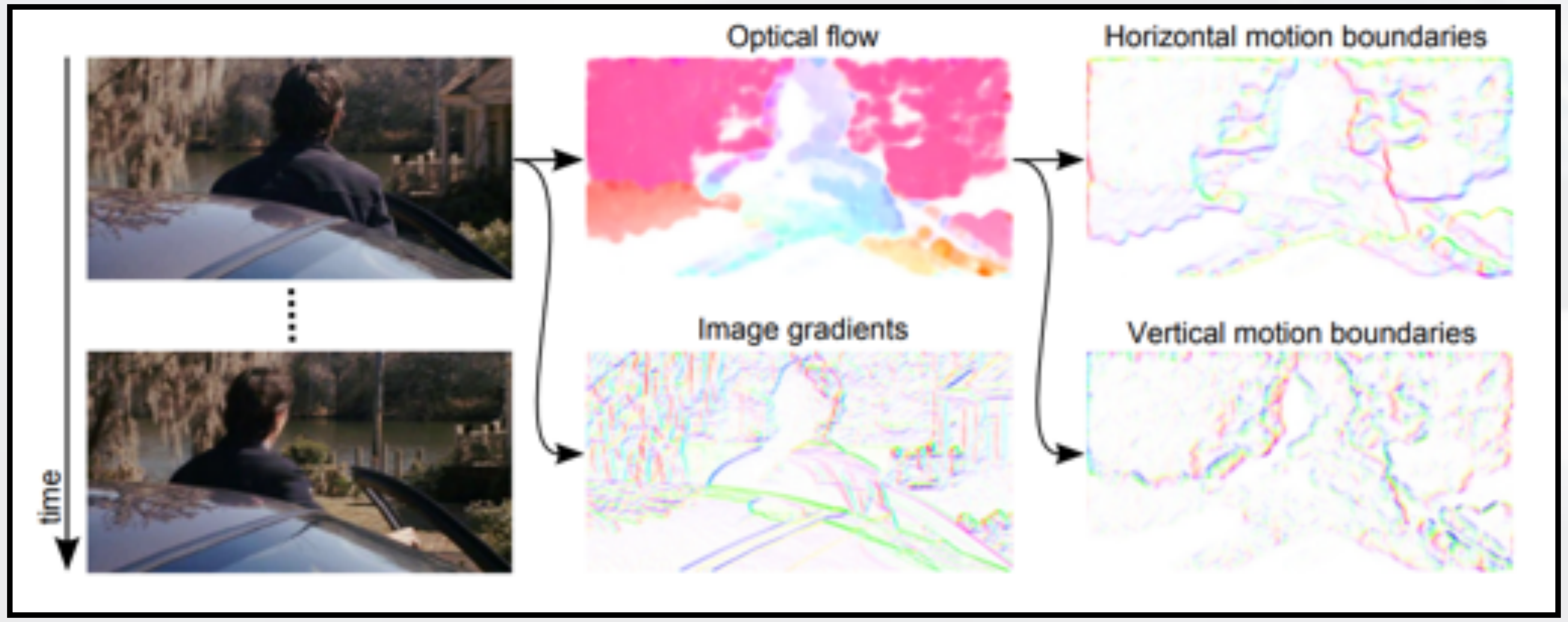
OPTICAL FLOW

Optical flow computes a motion field that gives:

1. Motion field of overall scene
2. Object tracking
3. Visual odometry

PRE-DEEP LEARNING APPROACHES - DENSE TRAJECTORIES





	Hollywood2	UCFSports
Trajectory	47.8%	75.4%
HOG	41.2%	84.3%
HOF	50.3%	76.8%
MBH	55.1%	84.2%
Combined	58.2%	88.0%

Dense trajectories and motion boundary descriptors for action recognition,
International Journal of Computer Vision, H Wang et. al. 2013
Action recognition with improved trajectories, Wang et. al. ICCV 2013

EX. VIDEO CLASSIFICATION TASK (UCF-11)

b_shooting

v_spiking

swinging

dog walking

tennis swing

cycling

diving

soccer juggling

r_riding

golf swing

t_jumping

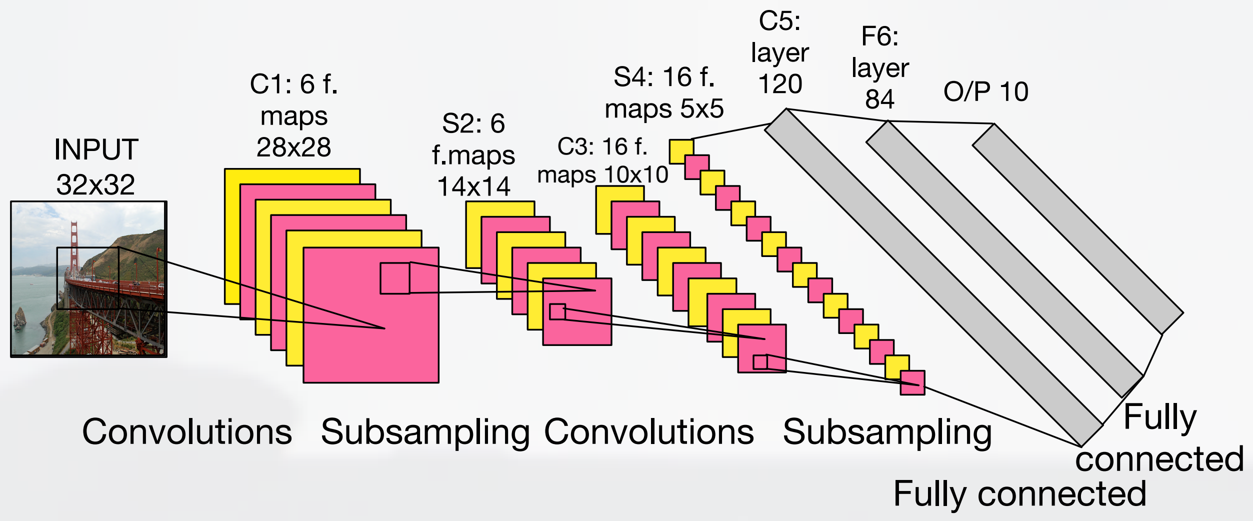
Detect human actions in video classification instead of objects in image classification

USING DEEP VISUAL FEATURES FROM 2D CNNs

```
for f in frames:
```

`224x224x3x1`

Video frames



CNN

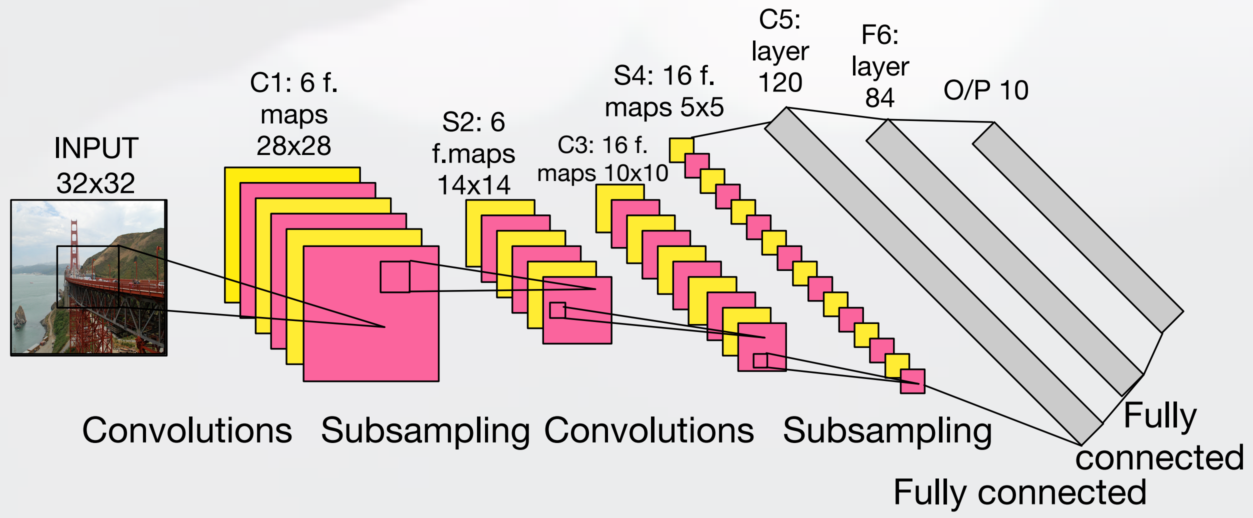
`101x1`

Tennis forehand

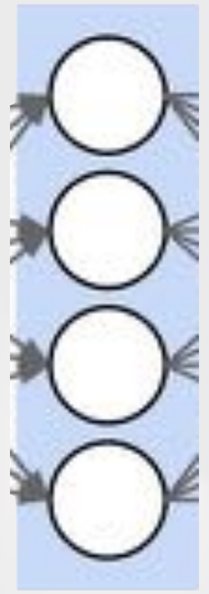
67% on UCF-101 dataset

`224x224x3x15`

Video frames



CNN



Pool/Concat intermediate features for all frames

`101x1`

Tennis forehand

EX. VIDEO CLASSIFICATION TASK (UCF-11)

b_shooting

v_spiking

swinging

dog walking

tennis swing

cycling

diving

soccer juggling

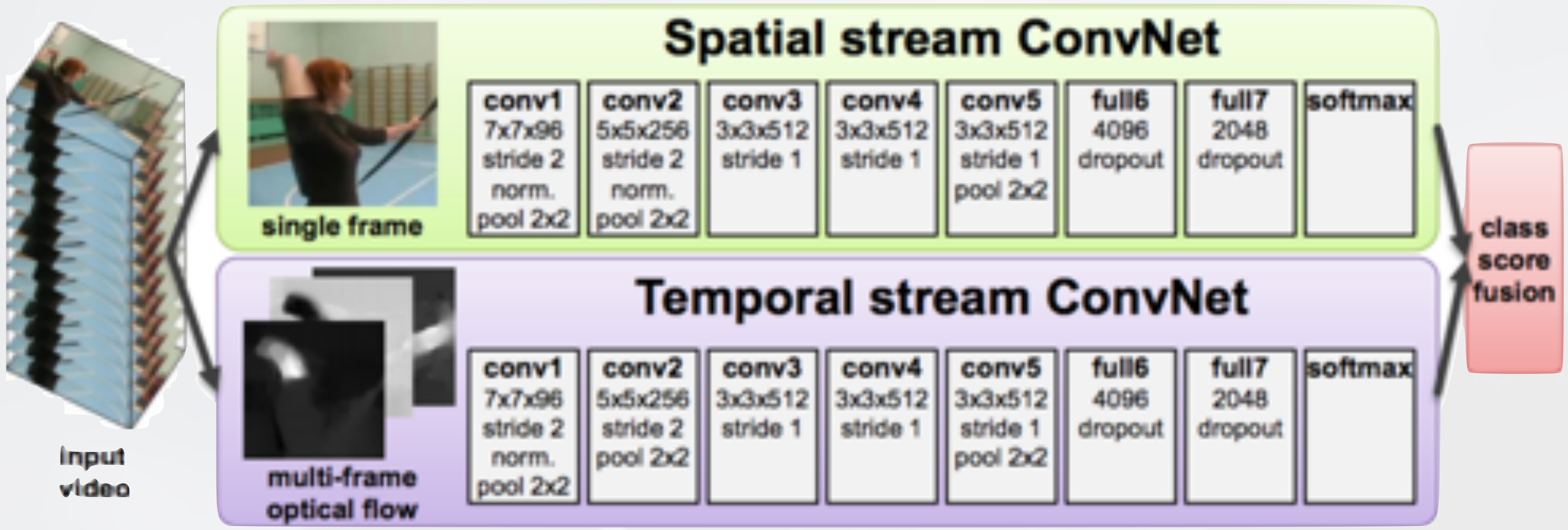
r_riding

golf swing

t_jumping

What is the problem here if we just use RGB features? Or even use RGB+flow features?

TWO STREAM NETWORKS: FUSING RGB AND FLOW SCORES



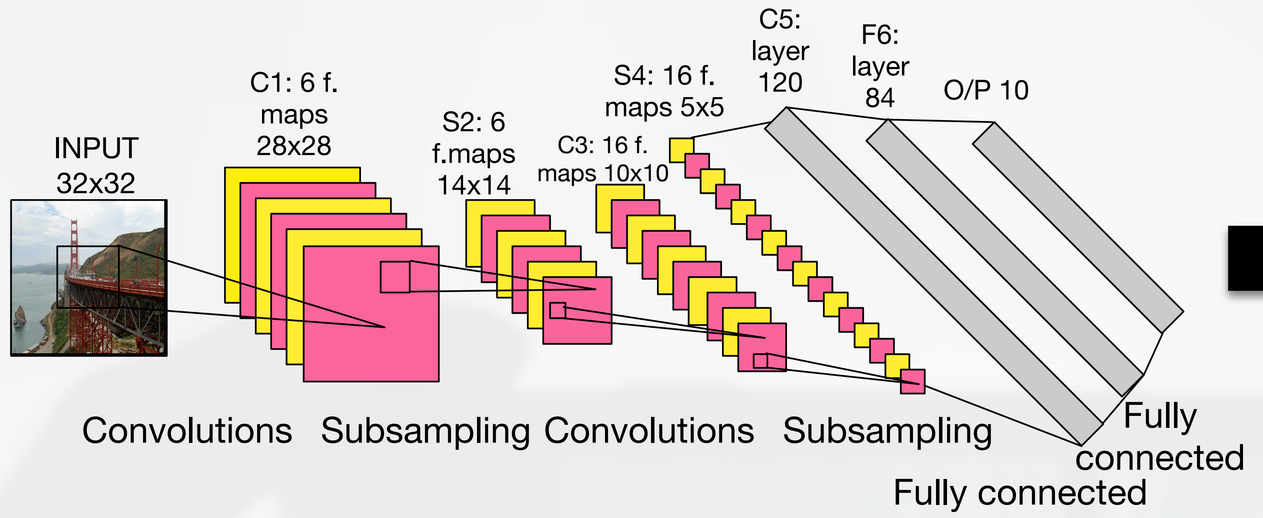
Two Stream Networks for Action Recognition in Videos. Simoyan et. al. NIPS 2014

Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

USING 2D CNN FEATURES WITH LSTM

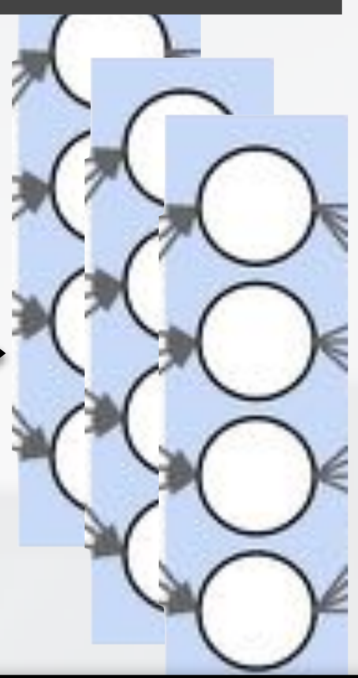
224x224x3x15

Video frames



CNN

15x2048



Intermediate Features

Sequence model

101x1

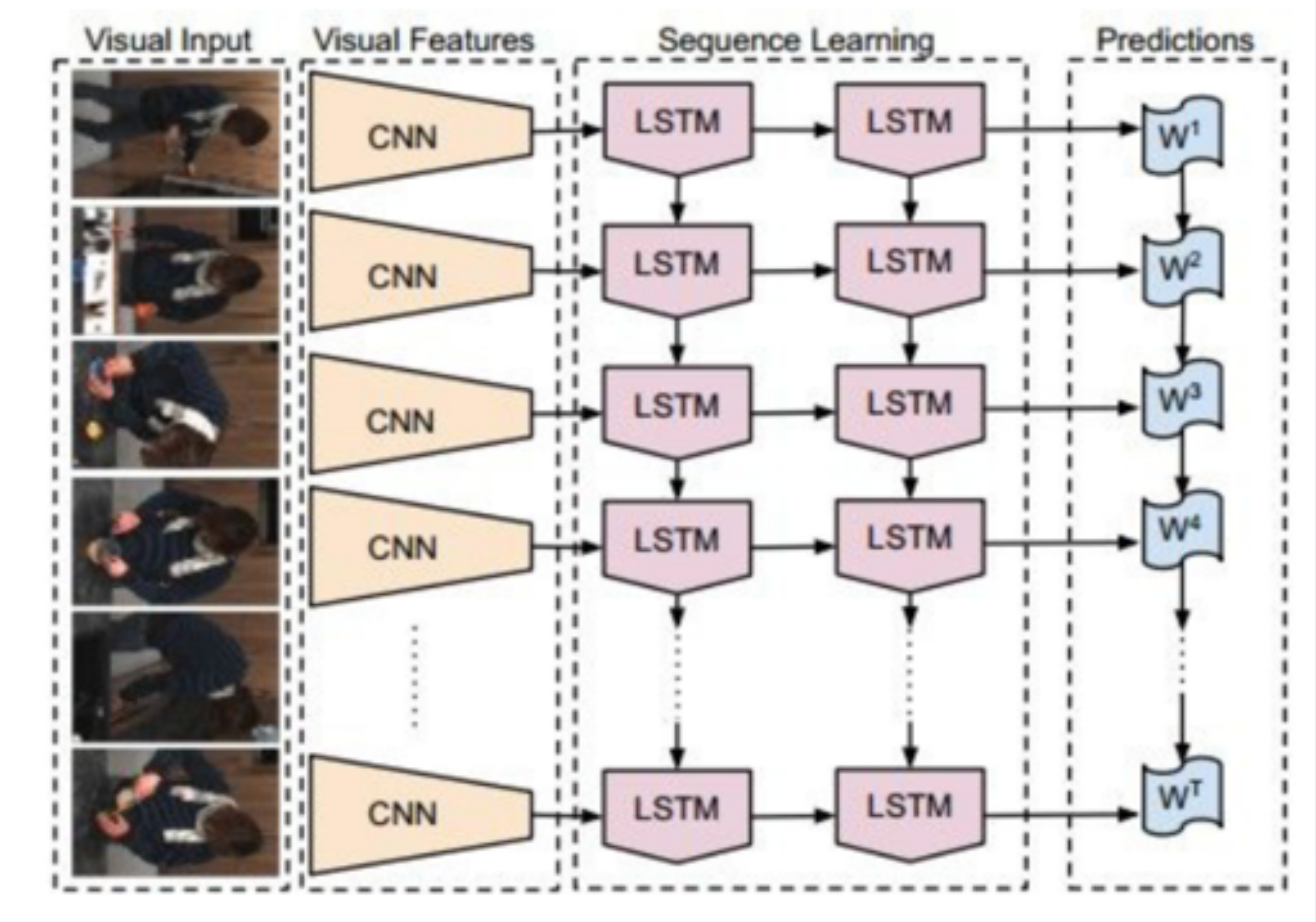
Tennis forehand

5

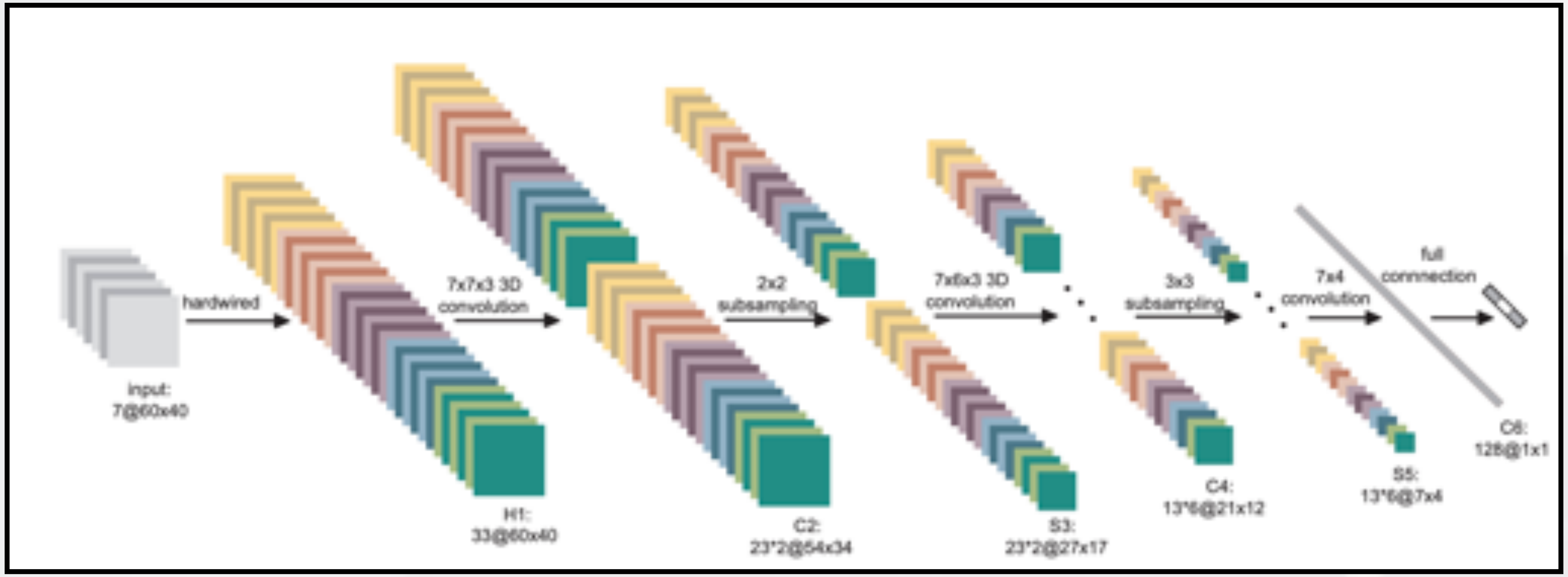
Model	Single Input Type		Weighted Average	
	RGB	Flow	1/2, 1/2	1/3, 2/3
Single frame	67.37	74.37	75.46	78.94
LRCN-fc6	**68.20**	**77.28**	80.90	**82.34**

TABLE 1
Activity recognition: Comparing single frame models to LRCN networks for activity recognition on the UCF101 [25] dataset, with RGB and flow inputs. Average values across all three splits are shown. LRCN consistently and strongly outperforms a model based on predictions from the underlying convolutional network architecture alone.

Long-term recurrent CNNs for Visual Recognition and Description, Donahue et. al., CVPR 2015



3D CONVOLUTION NETWORKS



3D Convolution Neural Networks for Human Action Recognition, Ji et. al. ICML 2010

VGG

Learning station-temporal features with 3D convolutional networks. Tran et. al., 2015

- Convolution in time and space domain (e.g. 5x5xT filters)
- Huge increase in parameters (e.g. UCF-101 2D -> 3D, 5M -> 33M params), C3D is 39.5 GFlop (as compared to resnext 8GFlop)
- Slowly learns time and space relationships through depth of the network
- 2D -> pooling/concat instead bring the temporal information all at once

USING CHANNELS FOR BETTER VISUAL FEATURES

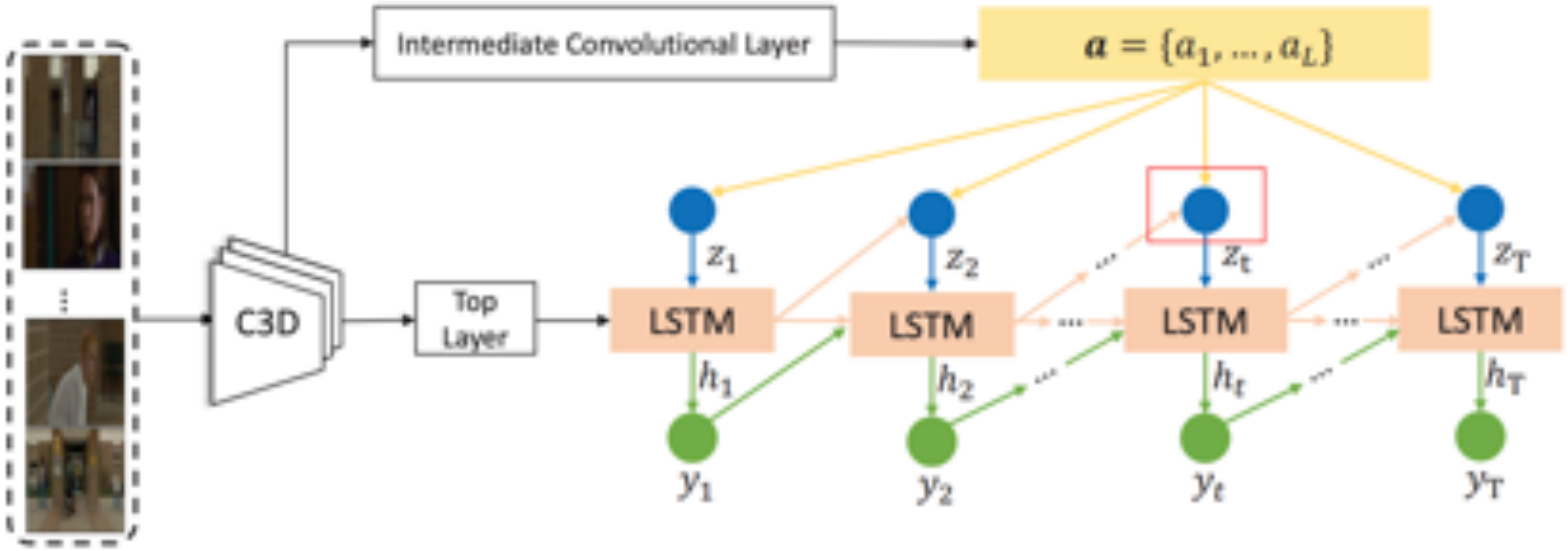


Figure 1: Illustration of our proposed caption-generation model. The model leverages a fully-connected map from the top layer as well as convolutional maps from different mid-level layers of a pretrained 3D convolutional neural network (C3D). The context vector z_t is generated from the previous hidden unit h_{t-1} and the convolutional maps $\{a_1, \ldots, a_L\}$ (the red frame), which is detailed in Figure 2.

- Uses 3-D convolution (C3D) features from attended intermediate layers with LSTM
- Used to solve the video captioning task, but the intermediate features can be used for any video understanding task

Adaptive Feature Abstraction for Translating Video to Text. Pu, Martin Renqiang Min et. al., AAAI 2018

ACTION RECOGNITION DATASETS

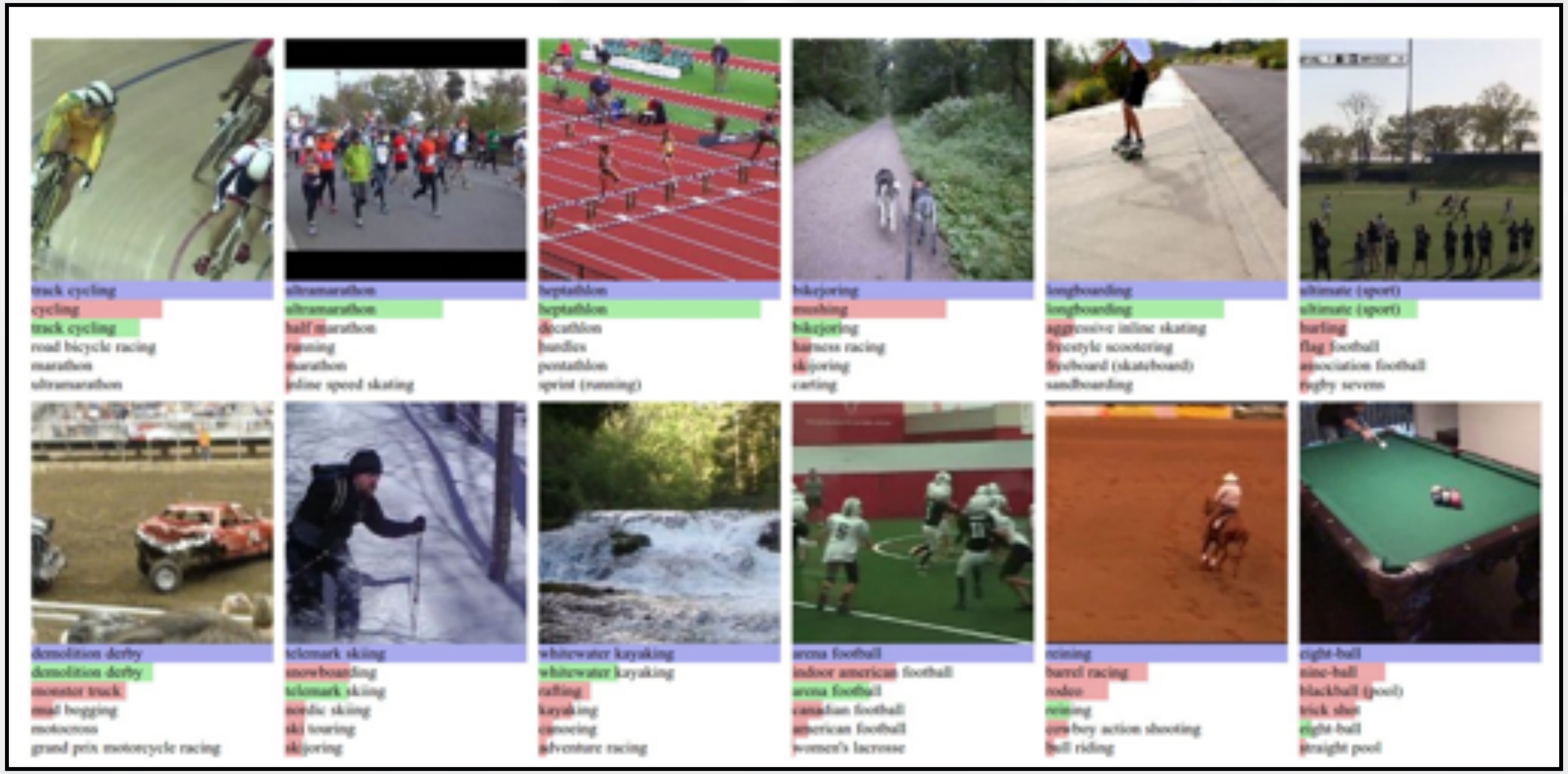
EARLY DATASETS: UCF-101 & HMDB-51

- UCF 101: 101 classes, 7 sec videos, 13K videos
- Large and commonly used dataset until 2017
- Youtube videos: variety of camera angles (first person, ego-centric, TV), illuminations, background, pose etc.

HMDB-51 and J-HMDB (21)

- HMDB 51: 51 classes, 4sec videos, 5K videos from movies
- J-HMDB dataset (21 classes from HMDB relying w/ joint information)

SPORTS 1M



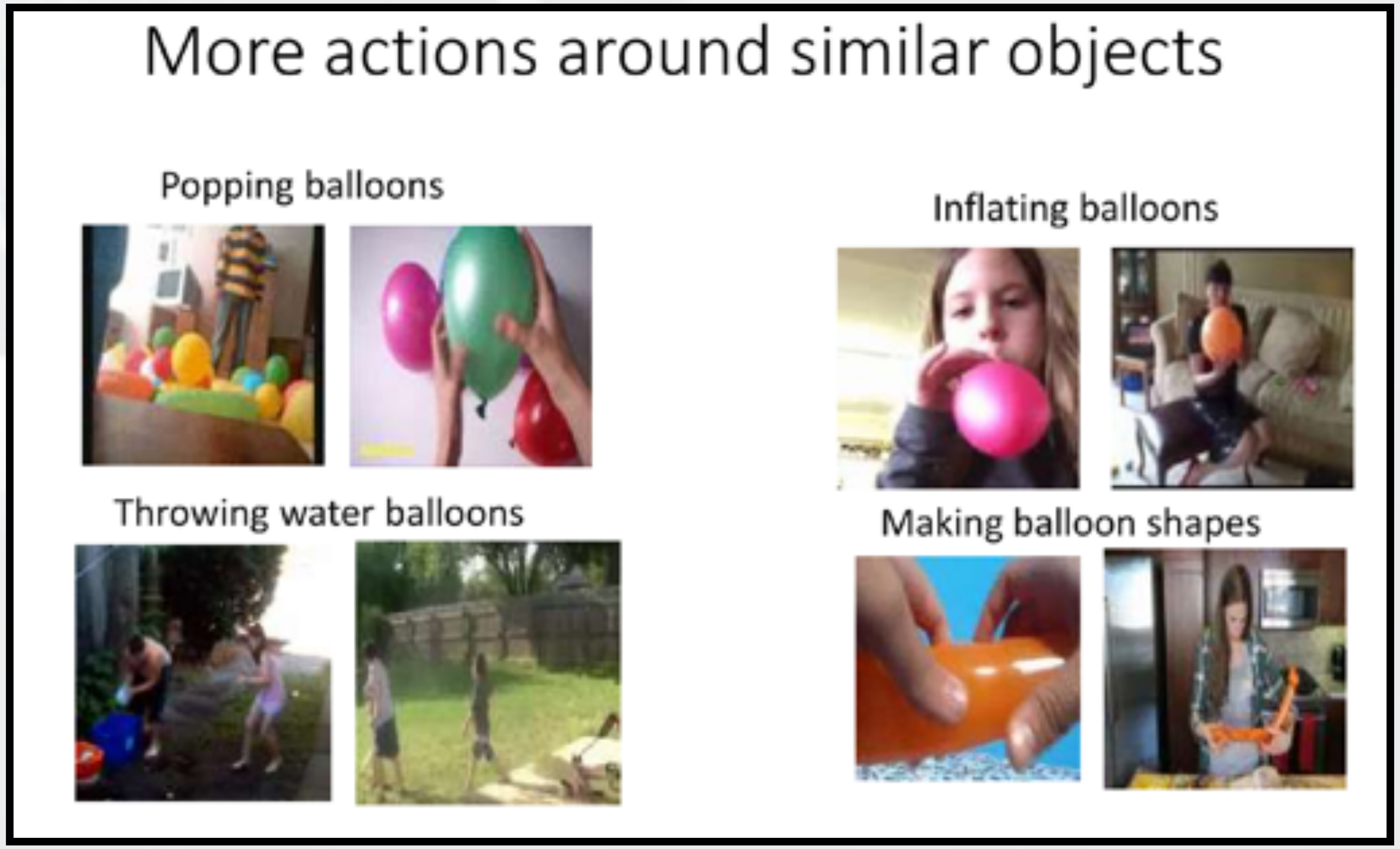
- YouTube videos: 1M
- 487 classes
- Fine-grained sports classes
- Pre-training on sports 1M and fine-tuning on UCF-101 generally improves performance

KINETICS

	Year	Actions	Clips per class	Total
Kinetics-400	2017	400	400-1000	300k
Kinetics-600	2018	600	600-1000	500k

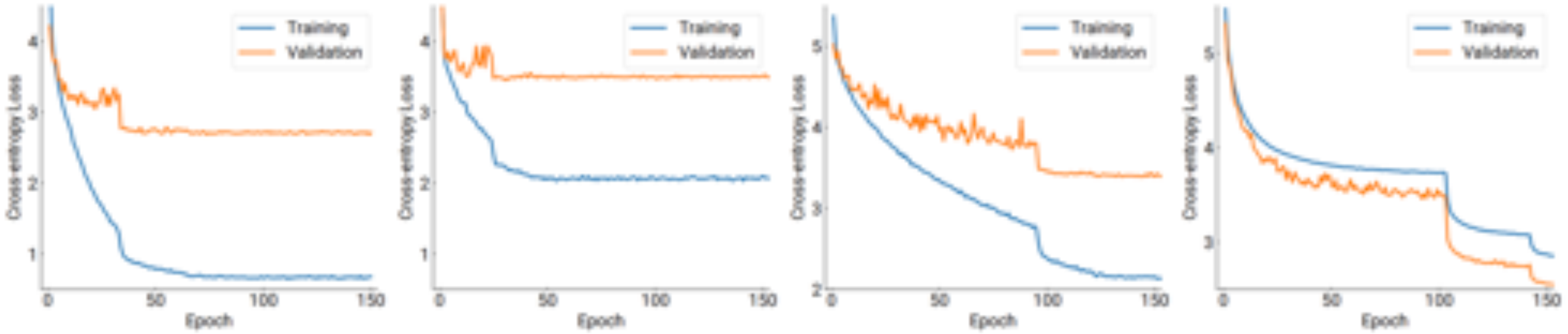
- 10s clips
- Every clip is from a different YouTube video
- For each action, huge variety in people, viewpoint, execution

- **Person Actions (Singular):** e.g. waving, blinking, running, jumping
- **Person-Person Actions:** e.g. hugging, kissing, shaking hands
- **Person-Object Actions:** e.g. opening door, mowing lawn, washing dishes



The Kinetics Human Action Video Dataset. Kay et. al. arXiv, 2017

KINETICS DATASET



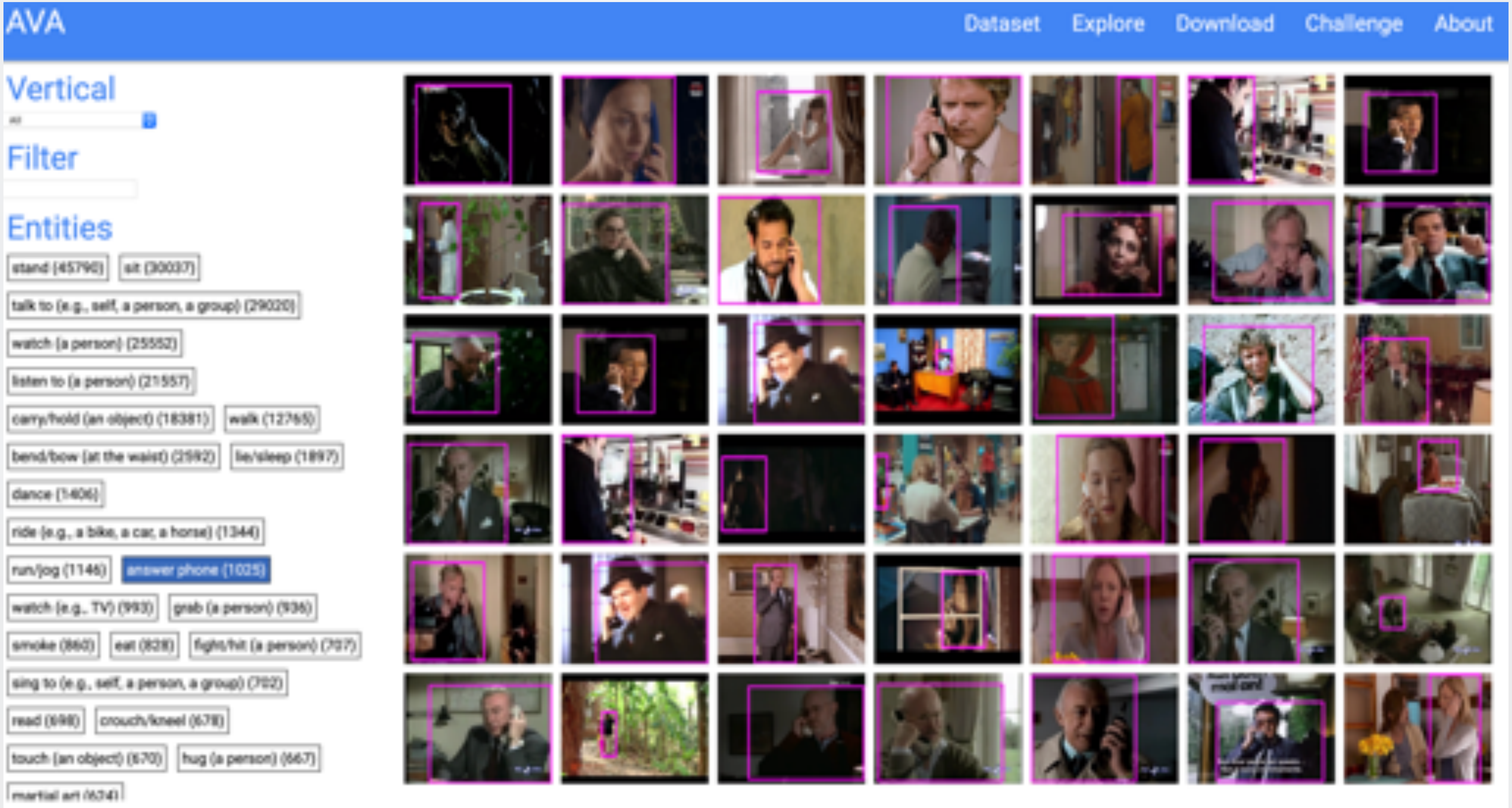
(a) UCF-101 (split 1). (b) HMDB-51 (split 1). (c) ActivityNet. (d) Kinetics.

Figure 4: ResNet-18 training and validation losses. The validation losses on UCF-101, HMDB-51, and ActivityNet quickly converged to high values and were clearly higher than their corresponding training losses. The validation losses on Kinetics were slightly higher than the corresponding training losses, significantly different than those on the other datasets.

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? Hara et. al., CVPR 2018

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500
ActivityNet-200 [3]	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245

AVA: ATOMIC VISUAL ACTIONS



SOMETHING SOMETHING DATASET

20BN-SOMETHING-SOMETHING-DATASET	
Total number of videos	220,847
Training Set	168,913
Validation Set	24,777
Test Set (w/o labels)	27,157
Labels	174
Putting something on a surface	4,081
Moving something up	3,750
Covering something with something	3,530
Pushing something from left to right	3,442
Moving something down	3,242
Pushing something from right to left	3,195
Uncovering something	3,004
Taking one of many similar things on the table	2,969
Turning something upside down	2,943
Tearing something into two pieces	2,849
Putting something into something	2,783

TOWARDS BETTER DATASETS FOR FINE-GRAINED UNDERSTANDING

Video classification datasets often suffer from visual bias (scene, objects) and difficulties in learning temporal relationships (long and very short temporal relationships)

E.g. 1) Eating watermelon involves a watermelon and with lack of other watermelon actions in the dataset, model infers "eating watermelon" when it sees a visually similar object to a watermelon

2) Short actions like 'slapping' are very short, as compared to median length of other actions

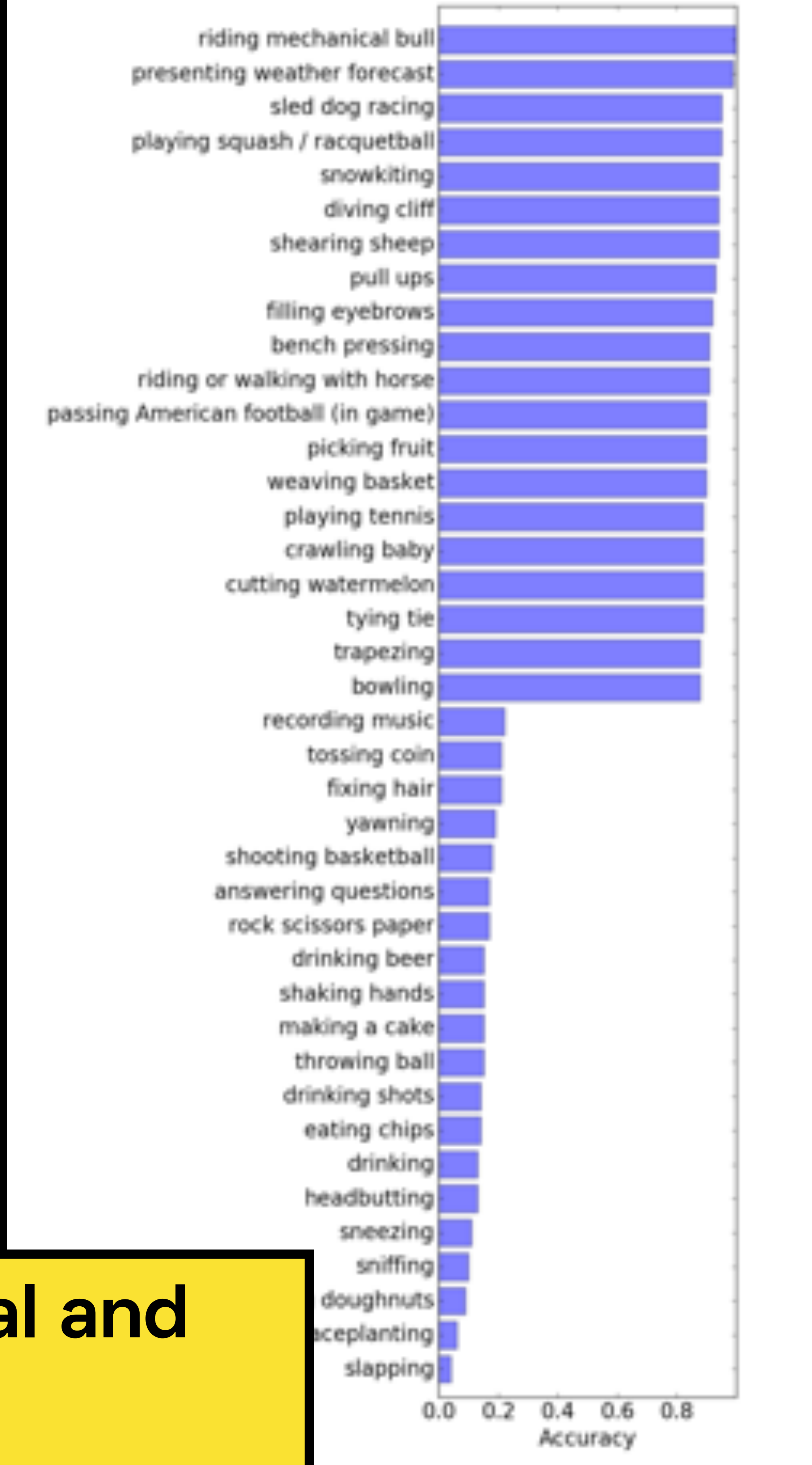


Figure 4: List of 20 easiest and 20 hardest Kinetics classes sorted by class accuracies obtained using the two-stream model.

How do we design a dataset to include spatial and temporal understanding?

CATER DATASET

- Synthetic video dataset built over CLEVR (Johnson et. al, 2017)

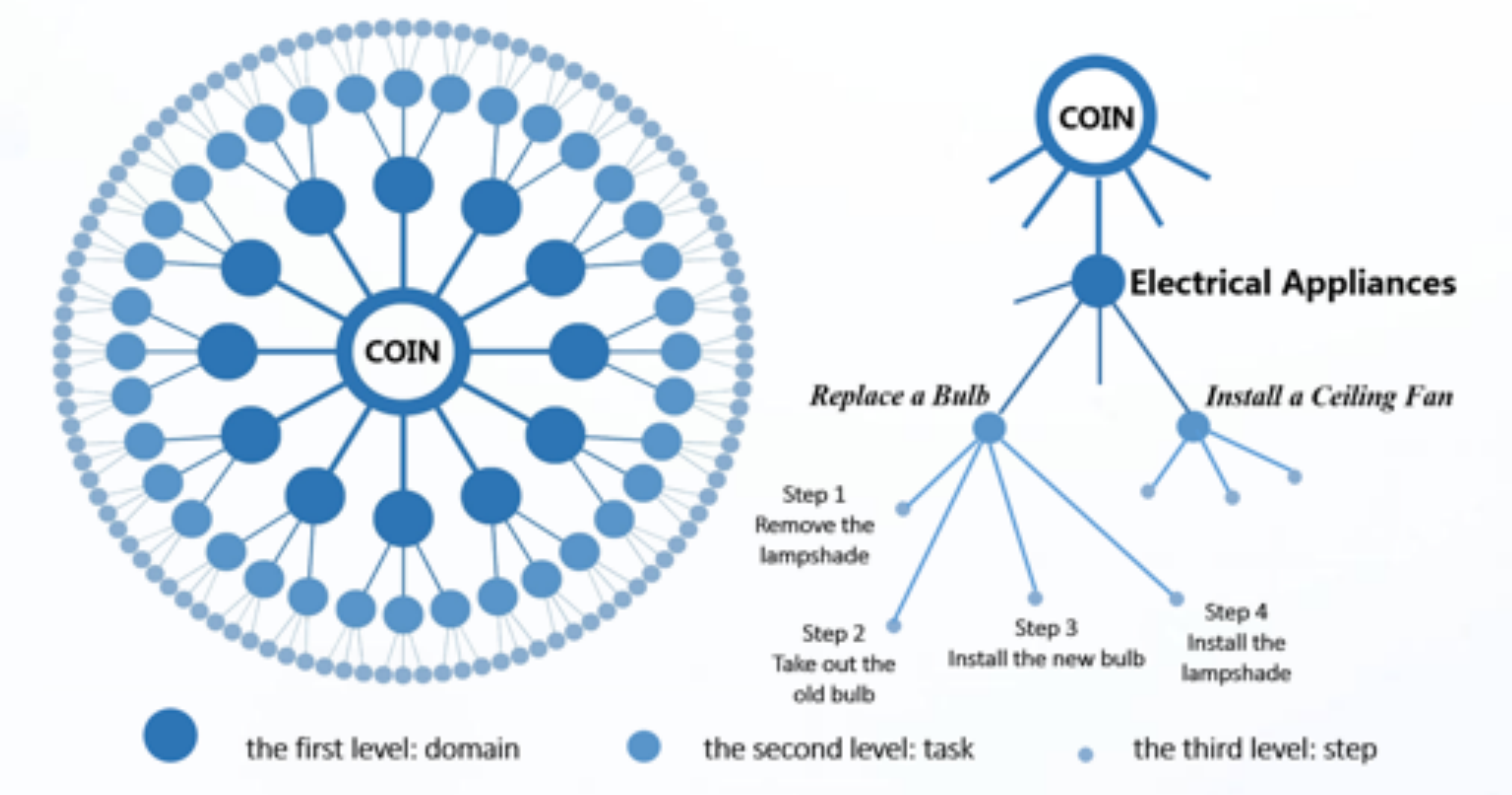
CATER: A diagnostic dataset for compositional actions & temporal reasoning. Giridhar et. al., ICLR 2020

Atomic action recognition (13 classes)

Compositional action recognition (301 classes)

Localization (36 classes, 6x6 grid)

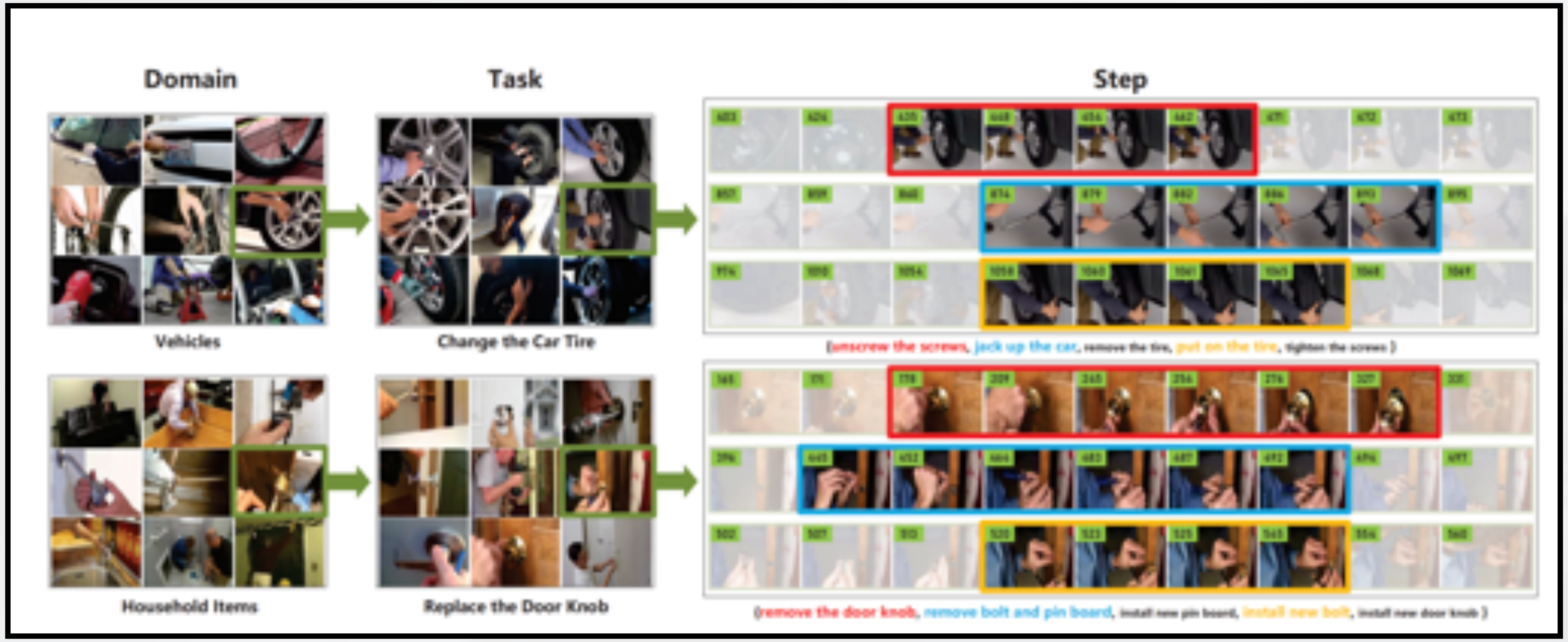
COIN DATASET



- 11, 827 videos, 180 tasks in 12 domains
- Domains: nursing & caring, vehicles, leisure & performance, gadgets, electric appliances, household items, science & craft, plants & fruits, snacks & drinks dishes, sports, and housework
- Tasks include: replace a bulb, install a ceiling fan (domain: electric appliance
- Steps ”remove the lampshade”, ”take out the old bulb”, ”install the new bulb” and ”install the lampshade” are associated with the tasks ”replace a bulb”.

COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, Tang et. al. arXiv 2019

COIN DATASET



- 11, 827 videos, 180 tasks in 12 domains
- Domains: nursing & caring, vehicles, leisure & performance, gadgets, electric appliances, household items, science & craft, plants & fruits, snacks & drinks dishes, sports, and housework
- Tasks include: replace a bulb, install a ceiling fan (domain: electric appliance
- Steps ”remove the lampshade”, ”take out the old bulb”, ”install the new bulb” and ”install the lampshade” are associated with the tasks ”replace a bulb”.

Table 1. Comparisons of existing instructional video datasets.

Dataset	Duration	Samples	Segments	Type of Task	Video Source	Hierarchical	Classes	Year
MPII [35]	9h,48m	44	5,609	cooking activities	self-collected	✗	-	2012
YouCook [14]	2h,20m	88	-	cooking activities	YouTube	✗	-	2013
50Salads [40]	5h,20m	50	966	cooking activities	self-collected	✗	-	2013
Breakfast [28]	77h	1,989	8,456	cooking activities	self-collected	✗	10	2014
"5 tasks" [10]	5h	150	-	comprehensive tasks	YouTube	✗	5	2016
Ikea-FA [41]	3h,50m	101	1,911	assembling furniture	self-collected	✗	-	2017
YouCook2 [52]	176h	2,000	13,829	cooking activities	YouTube	✗	89	2018
EPIC-KITCHENS [13]	55h	432	39,596	cooking activities	self-collected	✗	-	2018
COIN (Ours)	**476h,38m**	**11,827**	**46,354**	**comprehensive tasks**	**YouTube**	✓	**180**	

COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, Tang et. al. arXiv 2019

PROGRESSION OF THE FIELD THROUGH DATASETS

- Datasets have generally grown larger (demands of the deep learning models being used)
- Models have grown diverse from simple movies scenes to various view points, camera, lighting (mostly from youtube)
- Addition of fine-grained spatial actions (Kinetics, Sports-1M); also attempt to capture complex temporal relationships (Sth-sth, CATER)
- Many language grounded tasks beyond captioning

Dataset	Size	Len	Task	#cls	TO	STR	LTR	CSB
UCF101 (Soomro et al., 2012)	13K	7s	cls	101	✗	✗	✗	✗
HMDB51 (Kuehne et al., 2011)	5K	4s	cls	51	✗	✗	✗	✗
Kinetics (Kay et al., 2017)	300K	10s	cls	400	✗	✓	✗	✗
AVA (Gu et al., 2018)	430	15m	det	80	✗	✓	✗	✗
VLOGs (Fouhey et al., 2018)	114K	10s	cls	30	✗	✓	✗	✗
DAHLIA (Vaquette et al., 2017)	51	39m	det	7	✓	✓	✓	✗
TACoS (Regneri et al., 2013)	127	6m	align	-	✓	✓	✓	✗
DiDeMo (Anne Hendricks et al., 2017)	10K	30s	align	-	✓	✓	✓	✗
Charades (Sigurdsson et al., 2016)	10K	30s	det	157	✓	✓	✗	✗
Something Something (Goyal et al., 2017)	108K	4s	cls	174	✓	✓	✗	✓
Diving48 (Li et al., 2018)	18K	5s	cls	48	✓	✓	✗	✓
Cooking (Rohrbach et al., 2012a)	44	3-41m	cls	218	✓	✓	✗	✓
IKEA (Toyer et al., 2017)	101	2-4m	gen	-	✓	✓	✓	✓
Composite (Rohrbach et al., 2012b)	212	1-23m	cls	44	✓	✓	✓	✓
TFGIF-QA (Jang et al., 2017)	72K	3s	qa	-	✓	✓	✗	✗
MovieQA (Tapaswi et al., 2016)	400	200s	qa	-	✓	✓	✓	✗
Robot Pushing (Finn et al., 2016)	57K	1s	gen	-	✓	✓	✗	✓
SVQA (Song et al., 2018)	12K	4s	qa	-	✓	✓	✗	✓
Moving MNIST (Srivastava et al., 2015)	-	2s	gen	-	✓	✓	✗	✓
Flash MNIST (Long et al., 2018)	100K	2s	cls	1024	✗	✓	✗	✓
CATER (ours)	5.5K	10s	cls	36-301	✓	✓	✓	✓

Table from CATER, Giridhar et. al., ICLR 2020

LEARNING INTERACTIONS BETWEEN SPATIO-TEMPORAL SCENE OBJECTS

Motivation: Human actions involve complex interactions between the scene objects

Skiing

≠

Snowboarding

How do we learn complex interactions between scene elements?

Idea: Train an object detector to extract regions and learn their interaction across spatial and temporal domain

Skiing

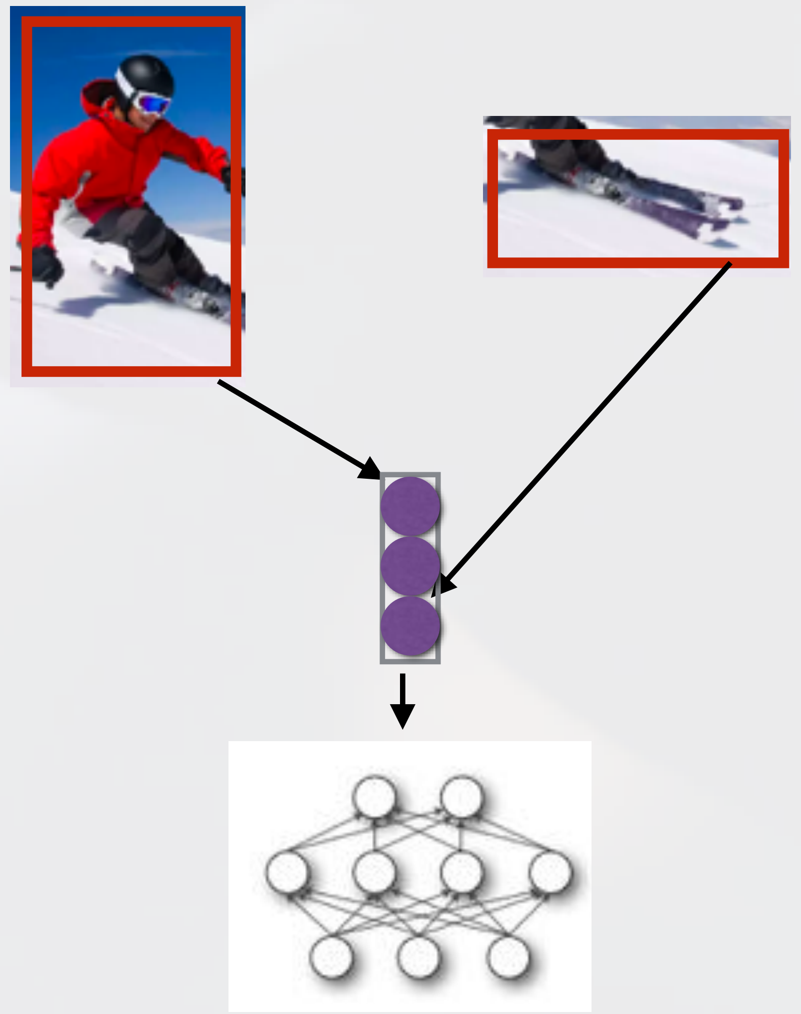
LEARNING PAIR-WISE INTERACTIONS CAN BE EXPENSIVE

- For each pair of objects, mean pool the vectors and train an MLP
  - Computation grows quickly with the number of projects
  - Adding more objects in a single vector drops accuracy quickly
  - Cannot learn higher-order interactions
  - Need to do this across frames

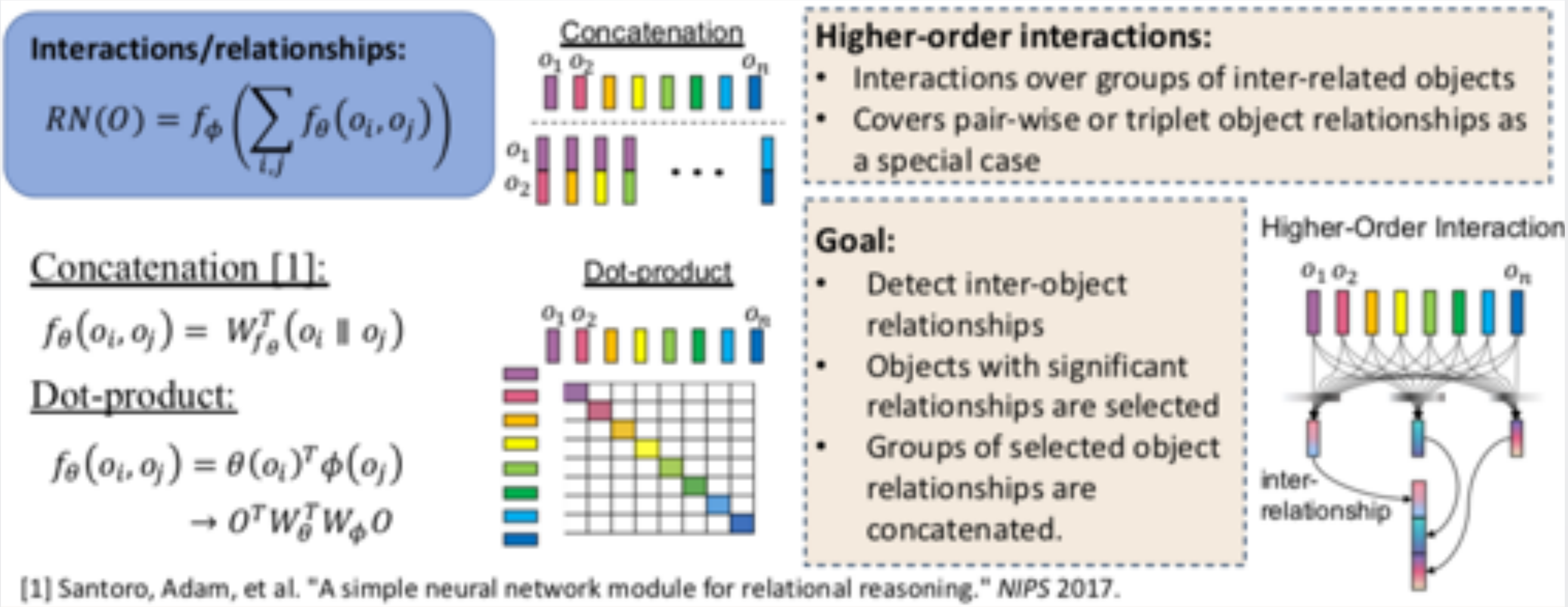
For each pair of detected ROIs

Skiing

=

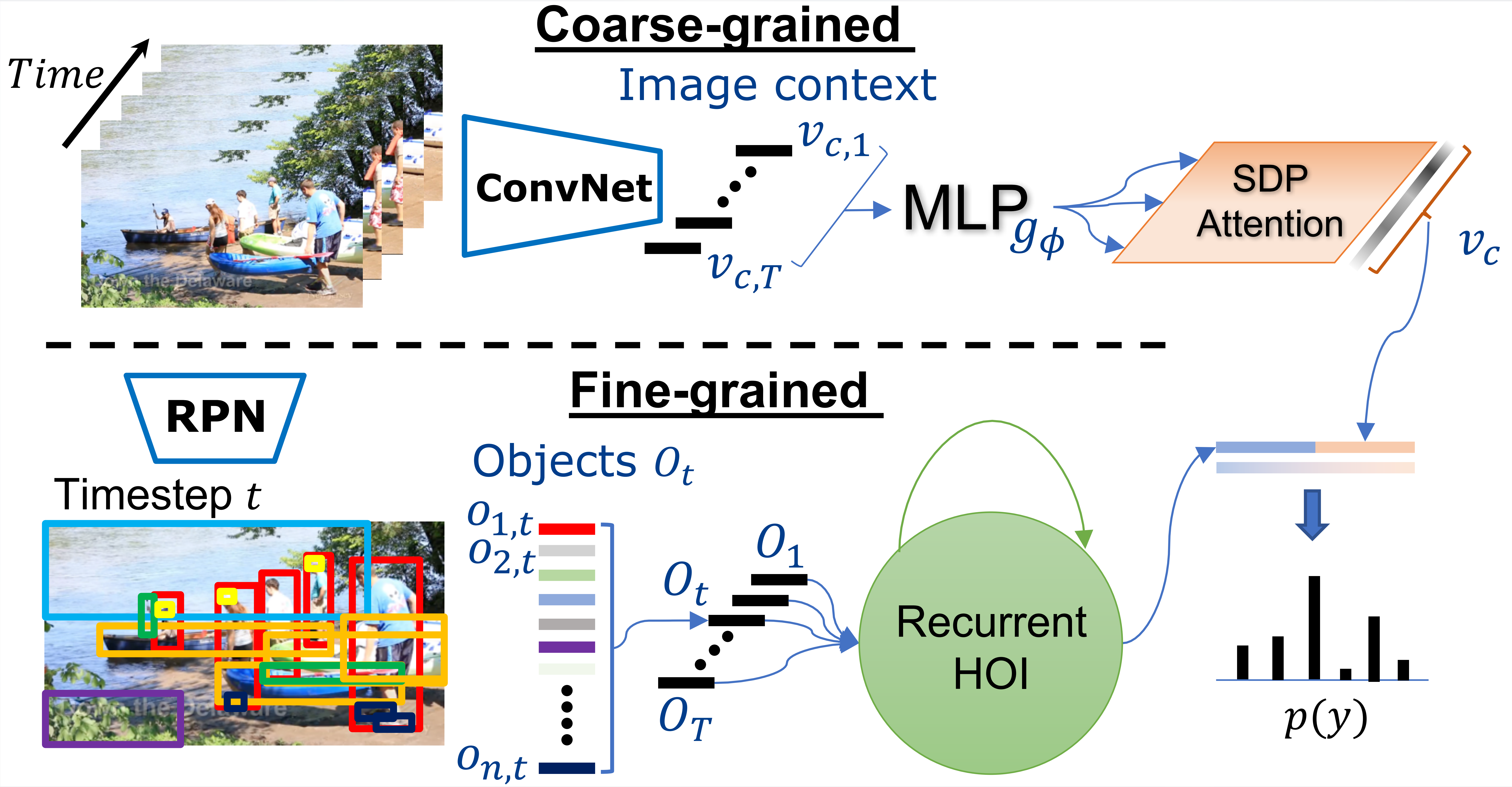


USING SELF-ATTENTION TO SELECT SALIENT OBJECTS



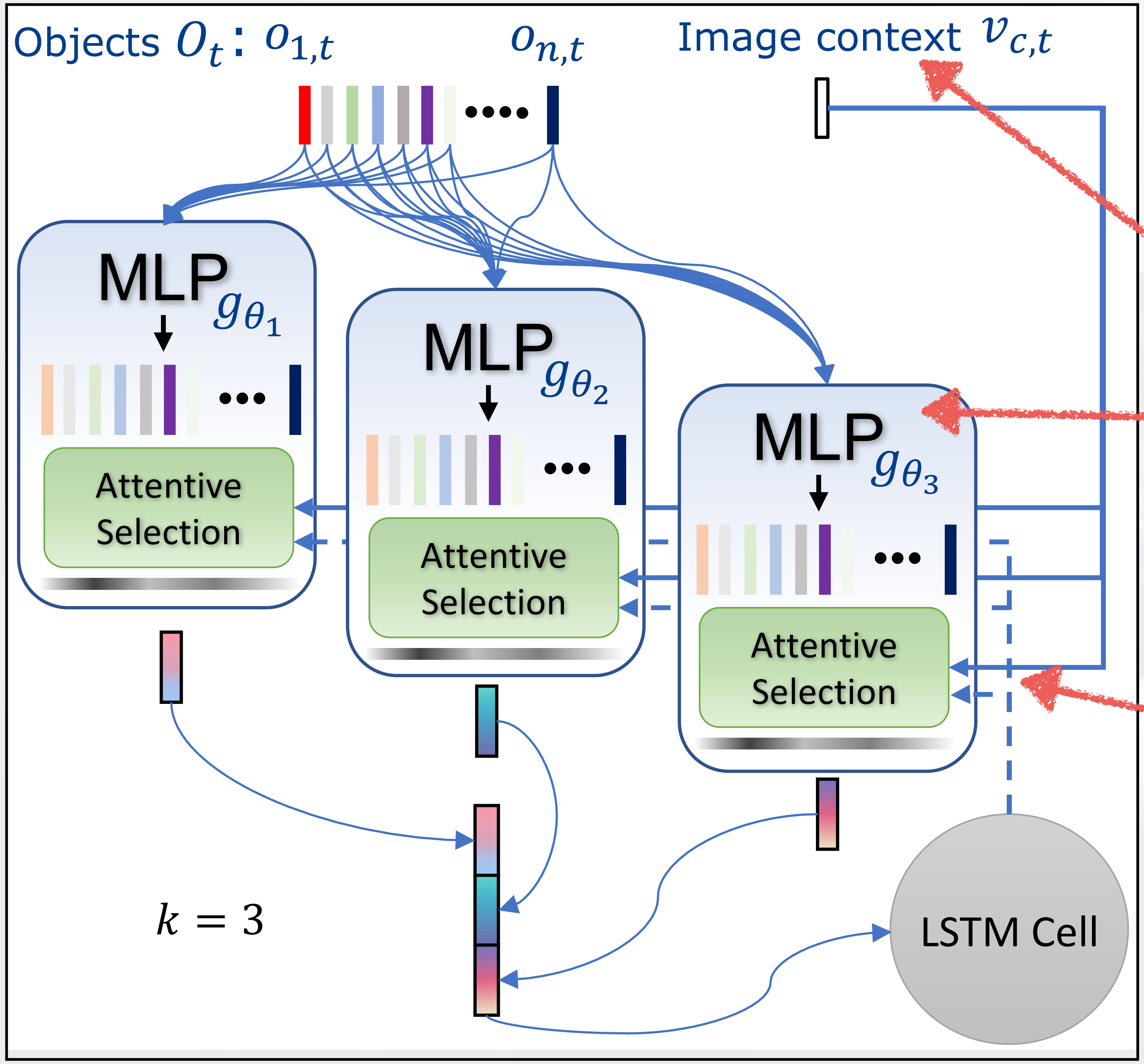
Attend and Interact: Higher-Order Object Interactions
for Video Understanding. Ma et. al. CVPR 2018.

SINET LEARNS INTERACTION BETWEEN SCENE ELEMENTS (ROIs)



SINET obtains a global video representation via the Scale Dot-Product Attention and a fine-grained representation (over objects) via recurrent higher-order interaction (HOI) module. The latter selects groups of objects with inter-relationships via an attention mechanism, and encodes the attended object features with LSTM. The coarse and fine-grained representations are concatenated for final prediction.

LEARNING HIGHER-ORDER INTERACTIONS WITH SELF-ATTENTION



Goal: Learn higher-order interactions between arbitrary (learnt) subgroups of objects

- Introduce learnable parameters via MLP to address domain shift problem
- Attentive selection with image context to co-attend with overall context
- This is combined with all previous interactions to generate a probability distribution over all objects using self-attention

Attend and Interact: Higher-Order Object Interactions for Video Understanding. Ma et. al. CVPR 2018.

QUALITATIVE RESULTS OF SINET

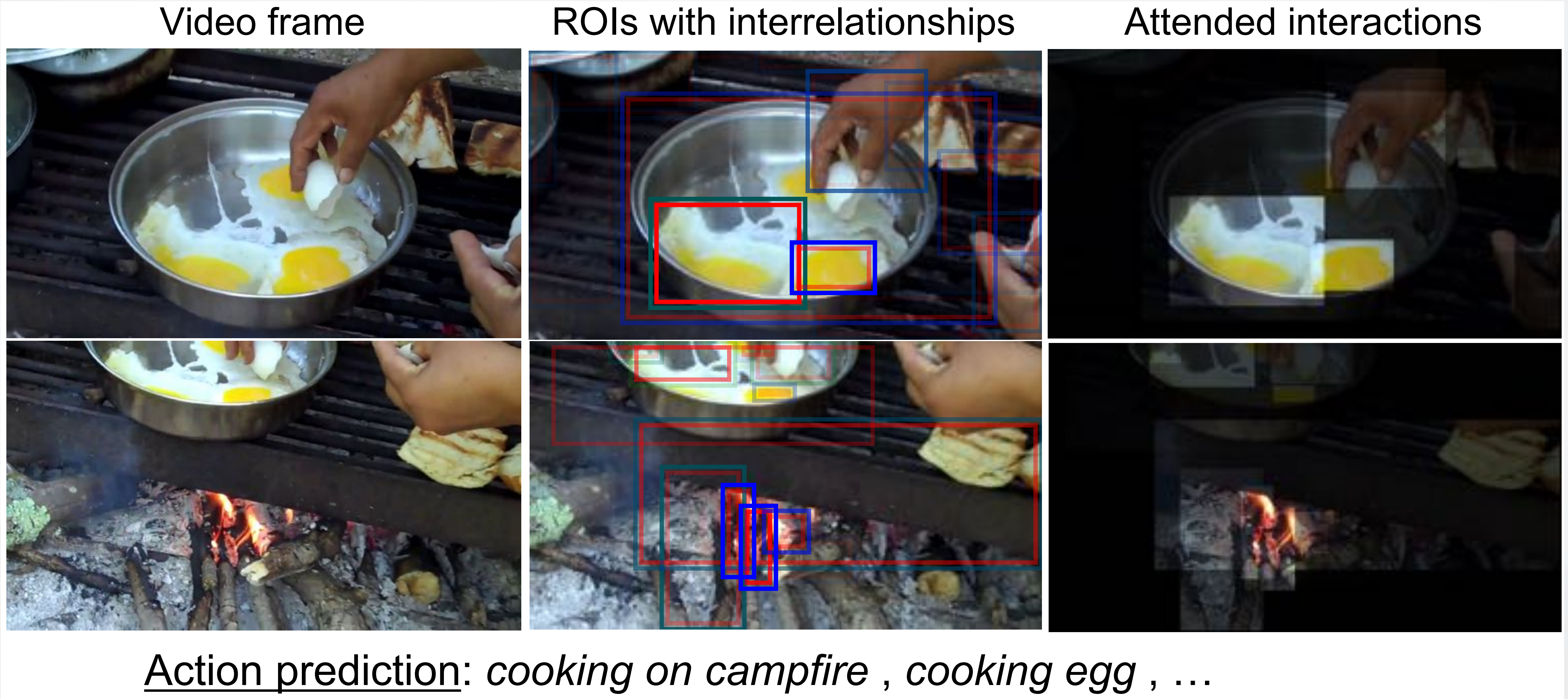
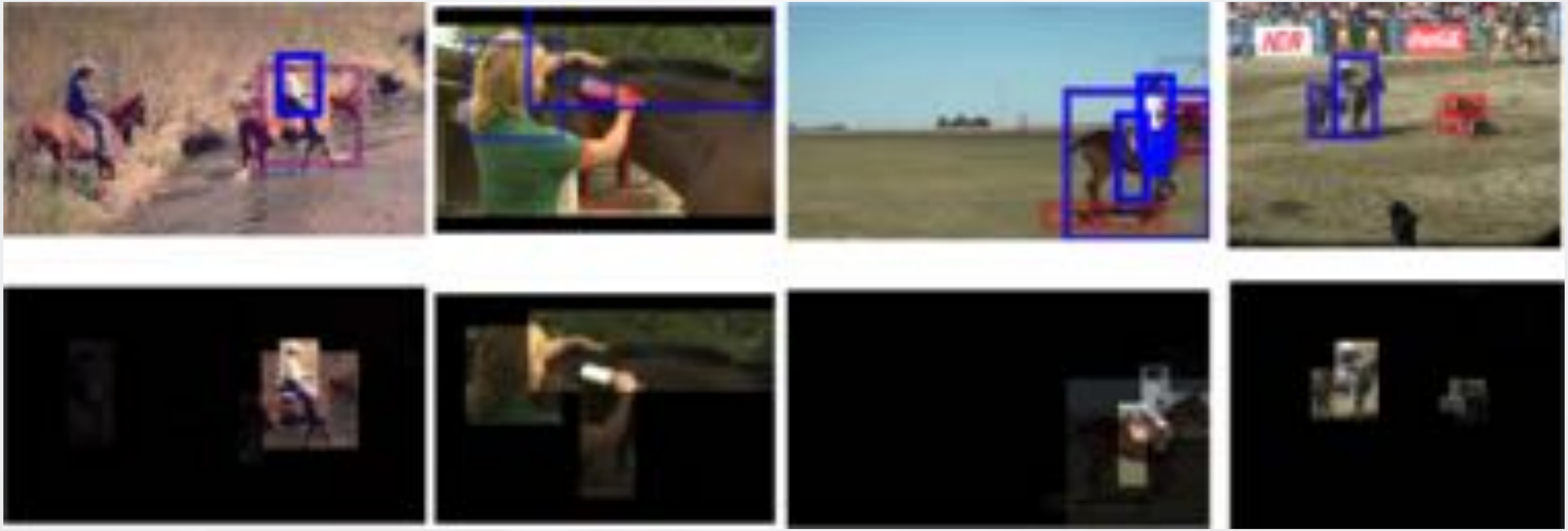


Figure 1. *Higher-order object interactions* are progressively detected based on selected lower-order interrelationships. ROIs with the same color (weighted r, g, b) indicating there exist inter-object relationships, e.g. eggs in the same bowl, hand breaks egg, and bowl on top of campfire (interaction within the same color). Groups of inter-relationships then jointly model higher-order object interaction of the scene (interaction between different colors). *Bottom*: ROIs are highlighted with their attention weights for higher-order interactions. The model further reasons the interactions through time and predicts *cooking on campfire* and *cooking egg*. Images are generated from SINet (best viewed in color).

QUALITATIVE RESULTS CONTINUED..



Riding

Brushing

Play Polo

Tie Up

Example: All images show a horse and a person. But actions are very different and difficult to distinguish for other methods

KINETICS EFFICIENCY OF SINET

Table 2. Comparison of pairwise (or triplet) object interaction with the proposed higher-order object interaction with dot-product attentive selection method on Kinetics. The maximum number of objects is set to be 15. FLOP is calculated per video. For details on calculating FLOP, please refer to Sec. 7.5.

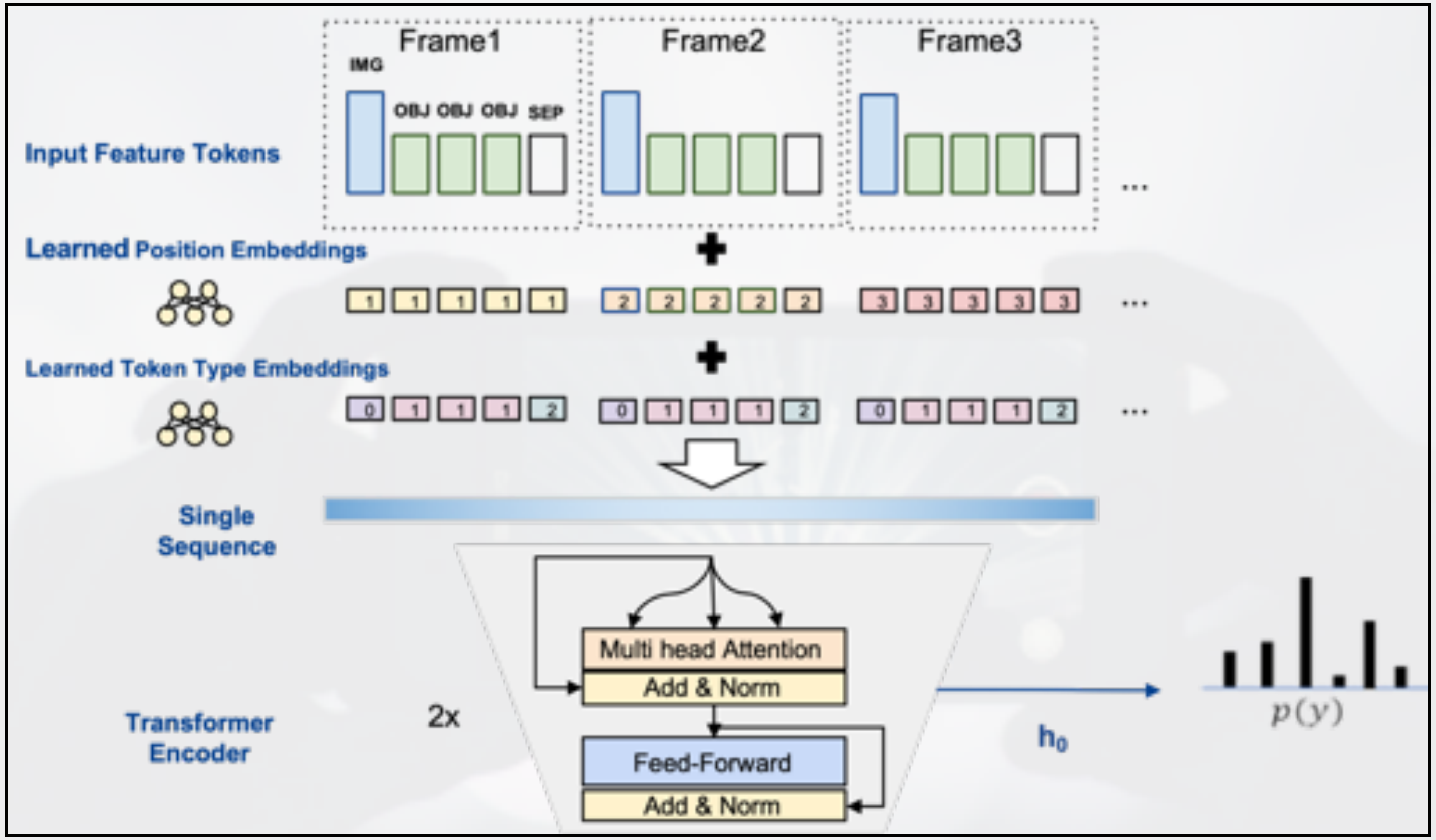
Method	Top-1	Top-5	FLOP (e^9)
Obj (mean-pooling)	73.1	90.8	1.9
Obj pairs (mean-pooling)	73.4	90.8	18.3
Obj triplet (mean-pooling)	72.9	90.7	77.0
SINet ($K = 1$)	73.9	91.3	2.7
SINet ($K = 2$)	74.2	91.5	5.3
SINet ($K = 3$)	**74.2**	**91.7**	8.0

Overall efficiency

I3D: 216 GFlops

SINET: 69 (53+8+8) GFlops

USING TRANSFORMERS INSTEAD OF SINET-HOI



- Object features -> Position encodings -> Type encodings (object and frame) -> Transformer (higher-order learning)
- Large increase in memory requirements
- Performs comparably to SINET (0.1 % lower)

MAKING ACTION RECOGNITION PRACTICAL

- Subsample videos : Sample videos at 1-5 FPS
- Reduce computation along the temporal dimension: Most modern benchmarks heavily rely on spatial information
- Use parallel operations : Transformers and convolution blocks can execute in parallel but the former have huge memory costs
- On-device processing : Suppress dead-frames; Use motion or audio to trigger processing
- Limit to RGB modality : Most Activity-Net contests are won using combination of optical-flow, audio, and skeletal modalities but recently RGB-only approaches have been competitive

FUTURE

- Compositional Methods: Understand videos in a compositional, spatio-temporal format.

COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, Tang et. al. arXiv 2019

Compositionality in Computer Vision

June 15th, Held in conjunction with CVPR 2020 in Seattle, US

- Using keypoints beyond pose: We have lot of experience modeling key point modality. Can we use these tools to solve other problems beyond pose estimation?

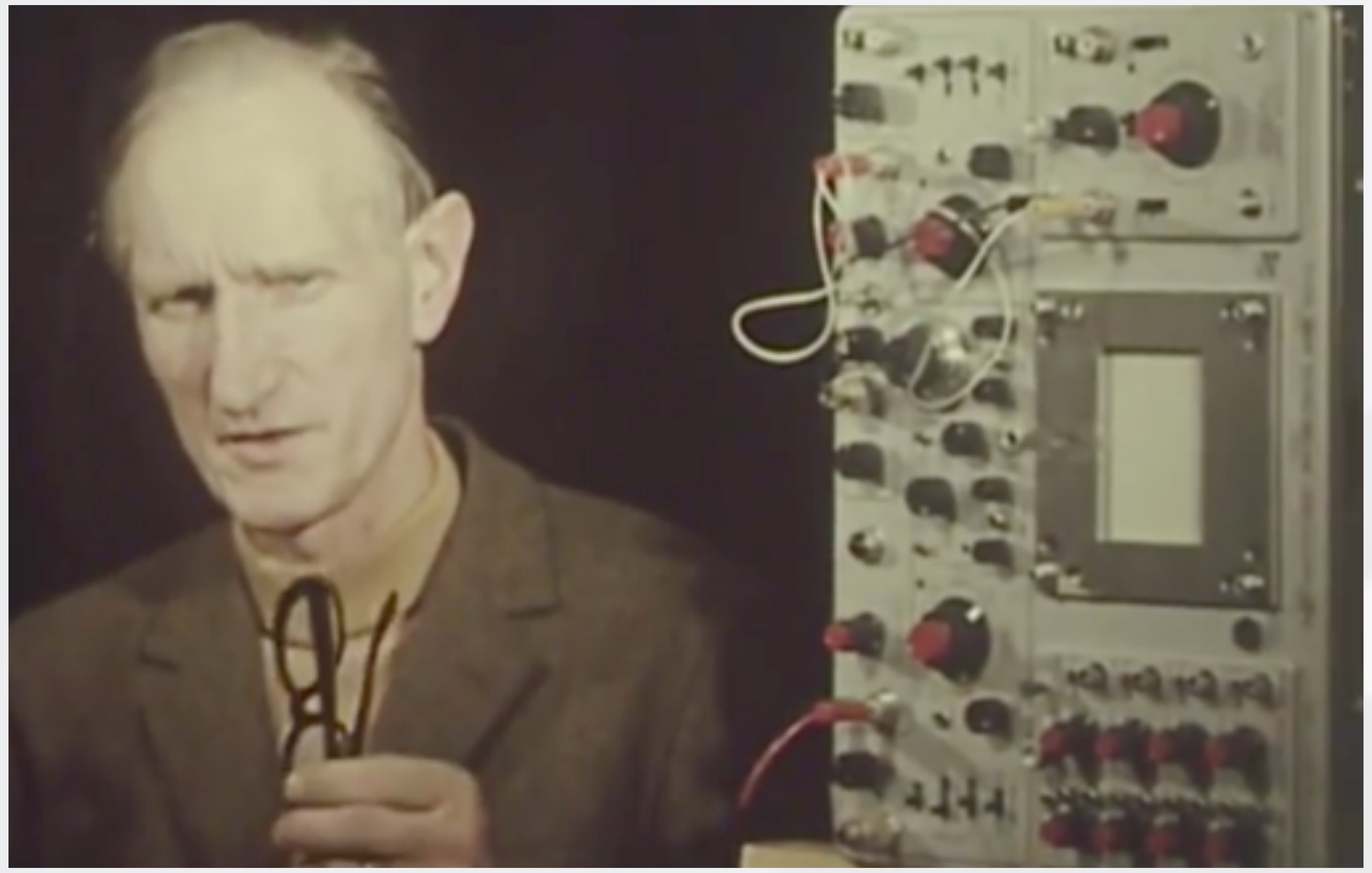
CornerNet: Detecting Objects as Paired Keypoints. Law et. al. ECCV 2018

15 Keypoints is all you need. Snower et. al, CVPR 2020

- Self-supervised understanding: Finite labelled data in the world; How do we use video data to generate its own labels?

Shuffle and Learn: Mishra et. al. 2016

UNDERSTANDING ACTIONS FROM KEYPOINTS



2-D motion perception, Gunnar Johansson. 1971

EXAMPLE VIDEO UNDERSTANDING TASK: POSE TRACKING

15 Keypoints is all you need. Snower et. al, CVPR 2020

HOW TO SOLVE POSE TRACKING TASK?

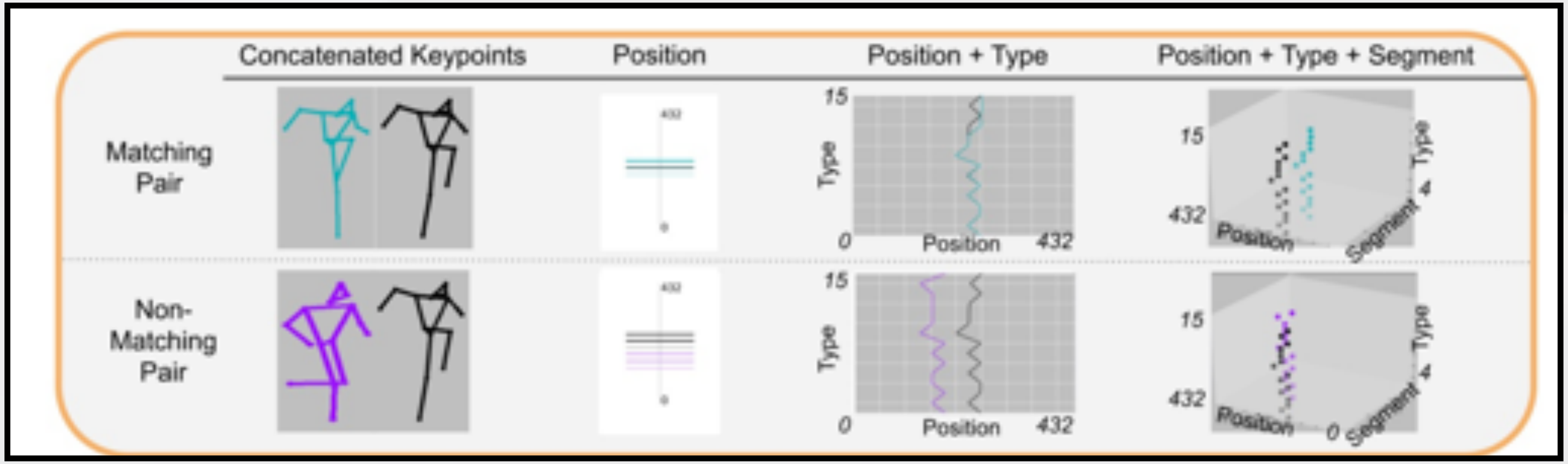
Keypoint estimation

Temporal Matching

- Optical Flow, GCN, Transformer (KeyTrack)
- Use temporal information to augment missed/poor quality detections

Assign IDs

- Match to ID from one of previous N frames



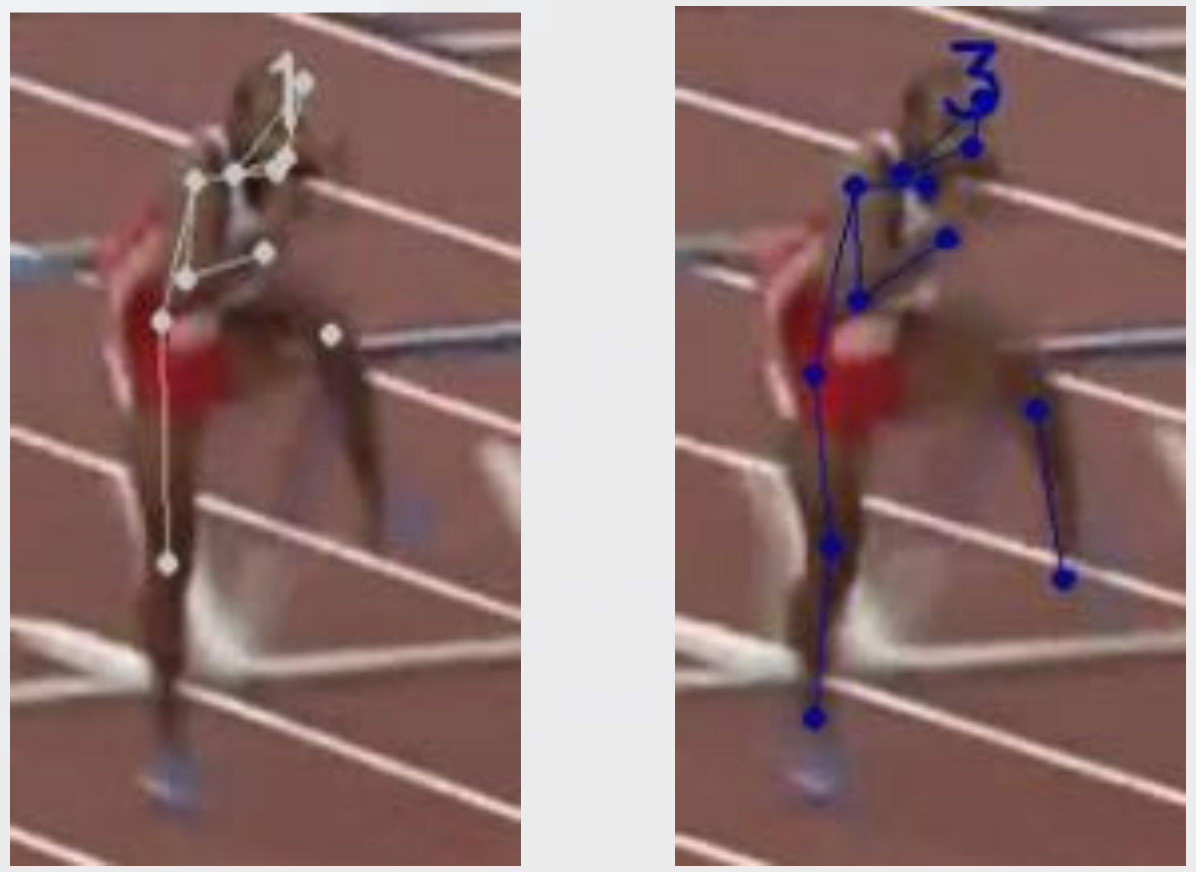
- Learn temporal pose warping using transformer

15 Keypoints is all you need. Snower et. al, CVPR 2020

MEASURING POSE TRACKING ACCURACY

FN, False Negative

FP, False Positive



IDSW, Track ID Switch

We have a set of $\mathcal{K}$ keypoints which we wish to track for a video with $\mathcal{T}$ frames, s.t. $t \in \mathcal{T}$.

The $MOTA^k$ for each keypoint $k \in \mathcal{K}$ is:

$$1 - \frac{\sum_t (FN_t^k + FP_t^k + IDSW_t^k)}{\sum_t GT_t^k}$$

Our final MOTA is the average of all $MOTA^k$:

$$\frac{\sum_k MOTA^k}{|\mathcal{K}|}$$

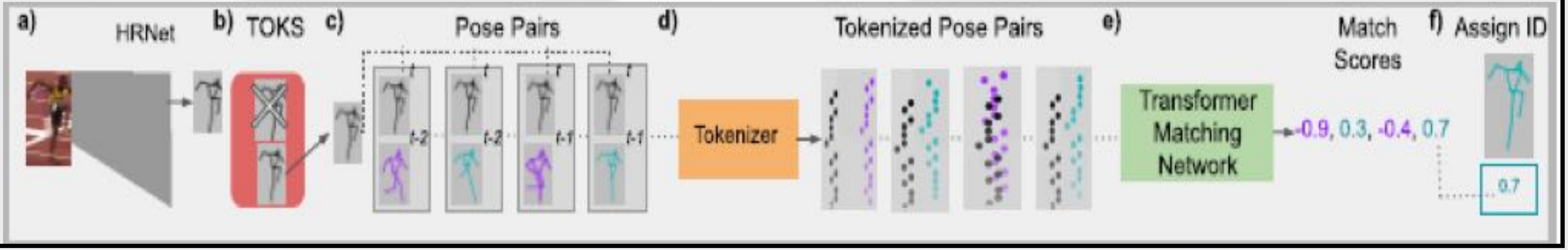
WORLD #1 IN POSE TRACKING LEADERBOARD (NOV'19 - APR'20)

PoseTrack 2018 ECCV Challenge Val Set

No.	Method	Extra Data	AP^T	AP	FPS	MOTA
1.	**KeyTrack (ours)**	✗	74.3	**81.6**	1.0	**66.6**
2.	MIPAL [27]	✗	**74.6**	-	-	65.7
3.	LightTrack (offline) [37]	✗	71.2	77.3	E	64.9
4.	LightTrack (online) [37]	✗	72.4	77.2	0.7	64.6
5.	Miracle [61]	✓	-	80.9	E	64.0
6.	OpenSVAI [38]	✗	69.7	76.3	-	62.4
7.	STAF [40]	✓	70.4	-	**3**	60.9
8.	MDPN [22]	✓	71.7	75.0	E	50.6

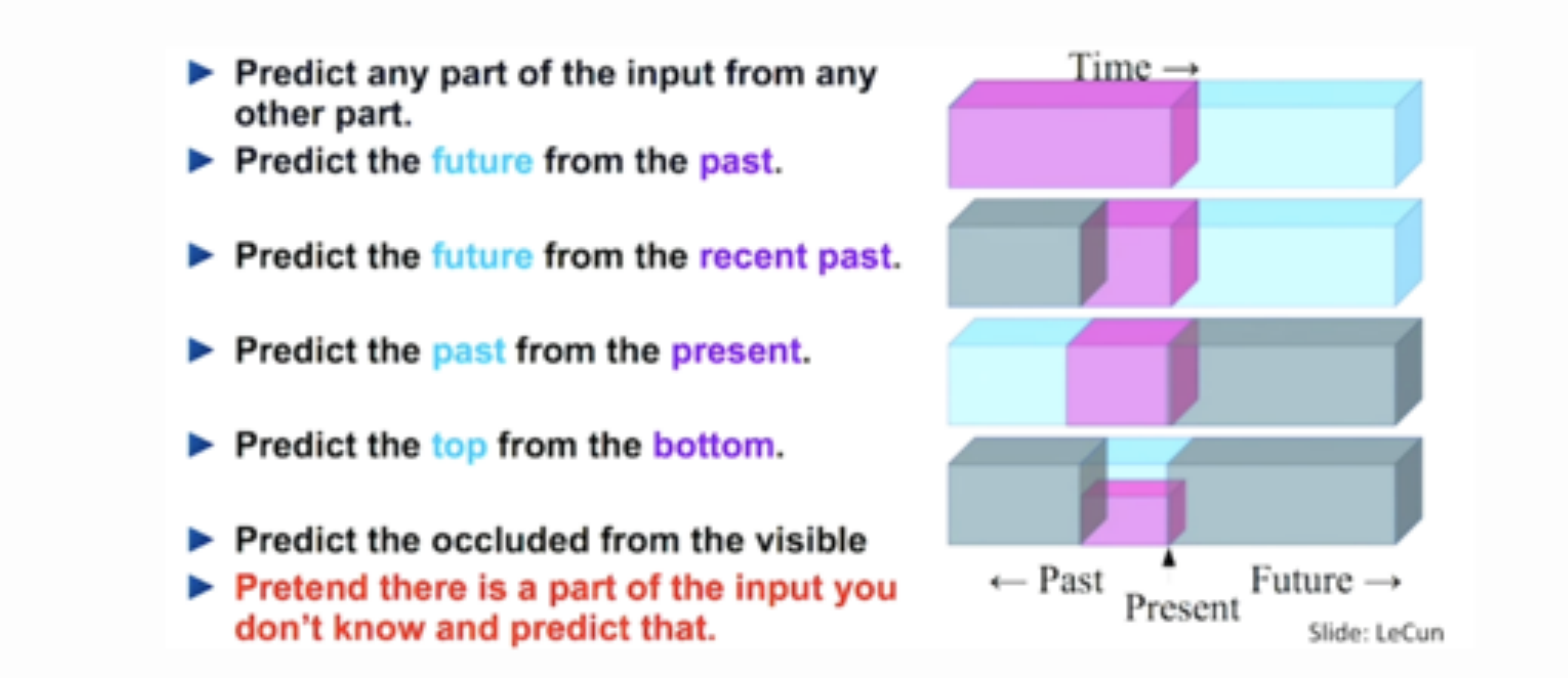
PoseTrack 2017 Test Set Leaderboard

No.	Method	Extra Data	AP^T	FPS	MOTA
1.	**KeyTrack (ours)**	✗	74.0	1.0	**61.2**
2.	POINet [42]	✗	72.5	-	58.4
3.	LightTrack [37]	✗	66.7	E	58.0
4.	HRNet [47]	✗	**75.0**	0.2	57.9
5.	FlowTrack [57]	✗	74.6	0.2	57.8
6.	MIPAL [27]	✗	68.8	-	54.5
7.	STAF [1]	✓	70.3	**2**	53.8
8.	JointFlow [17]	✗	63.6	0.2	53.1



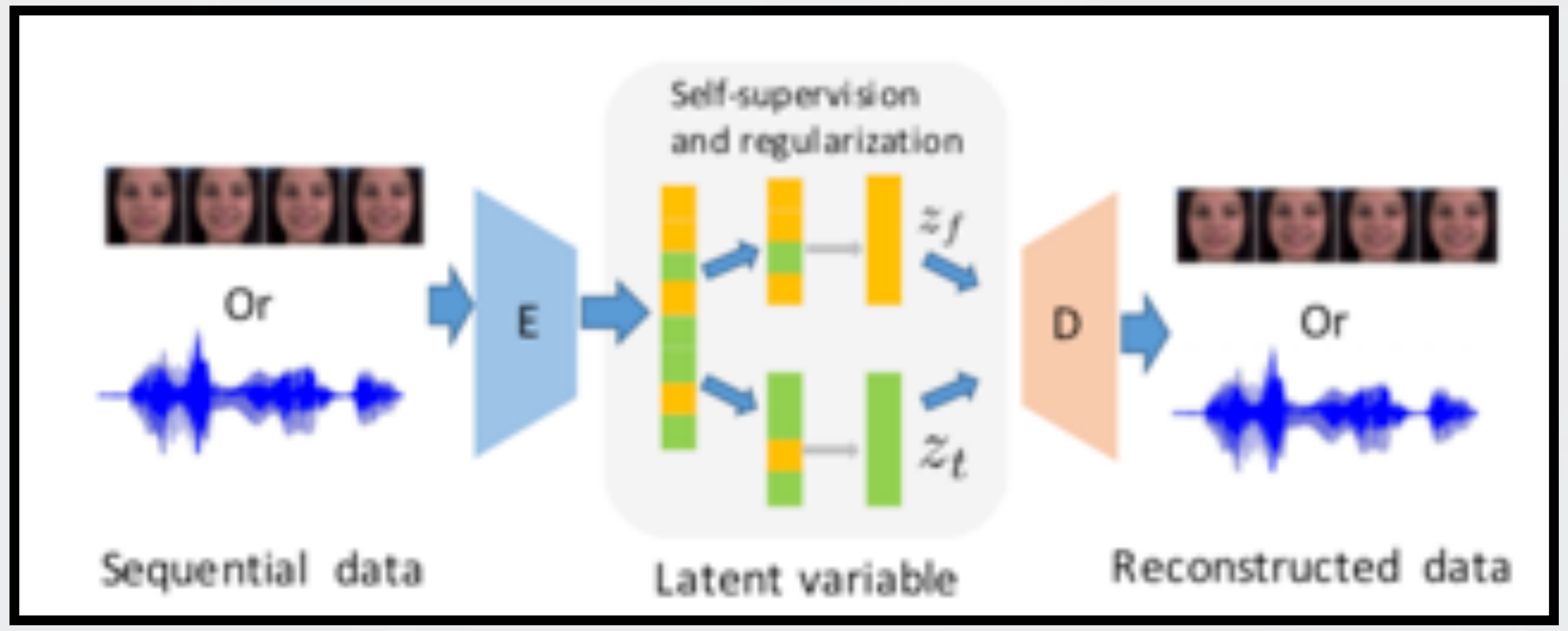
SELF-SUPERVISED METHODS FOR VIDEO UNDERSTANDING

- **Track moving objects: Wang et. al. 2015** : Track patches with motion over a small temporal window => Learns temporal motion of objects
- **Shuffle and Learn: Mishra et. al. 2016** : Validate frame order by shuffling frames => Learns temporal order of whole scene
- **Colorizing videos: Vonderick et. al. 2018** : Given two nearby frames, one in color and another in grey scale, the task is to copy colors from one frame to another nearby frame



Self-supervised approaches. Slides from Lecun, 2019

Example: Use these methods for generating disentangled representation for video generation



S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation, CVPR 2020.

SUMMARY

- Video has numerous applications in modern applications such as AR/VR, retail etc.
- Understanding video and generating a good representation is complex and computationally intensive
- Numerous opportunities with new datasets, tasks and compute platforms

QUESTIONS