JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

The Wisdom of Crowds: Temporal Progressive Attention for Early Action Prediction

Session: WED-PM-225

Alexandros Stergiou [1,2,†]

Dima Damen [3]

1 VUB VRIJE UNIVERSITEIT BRUSSEL 2 imec 3 University of BRISTOL

[†] work done whilst at the University of Bristol

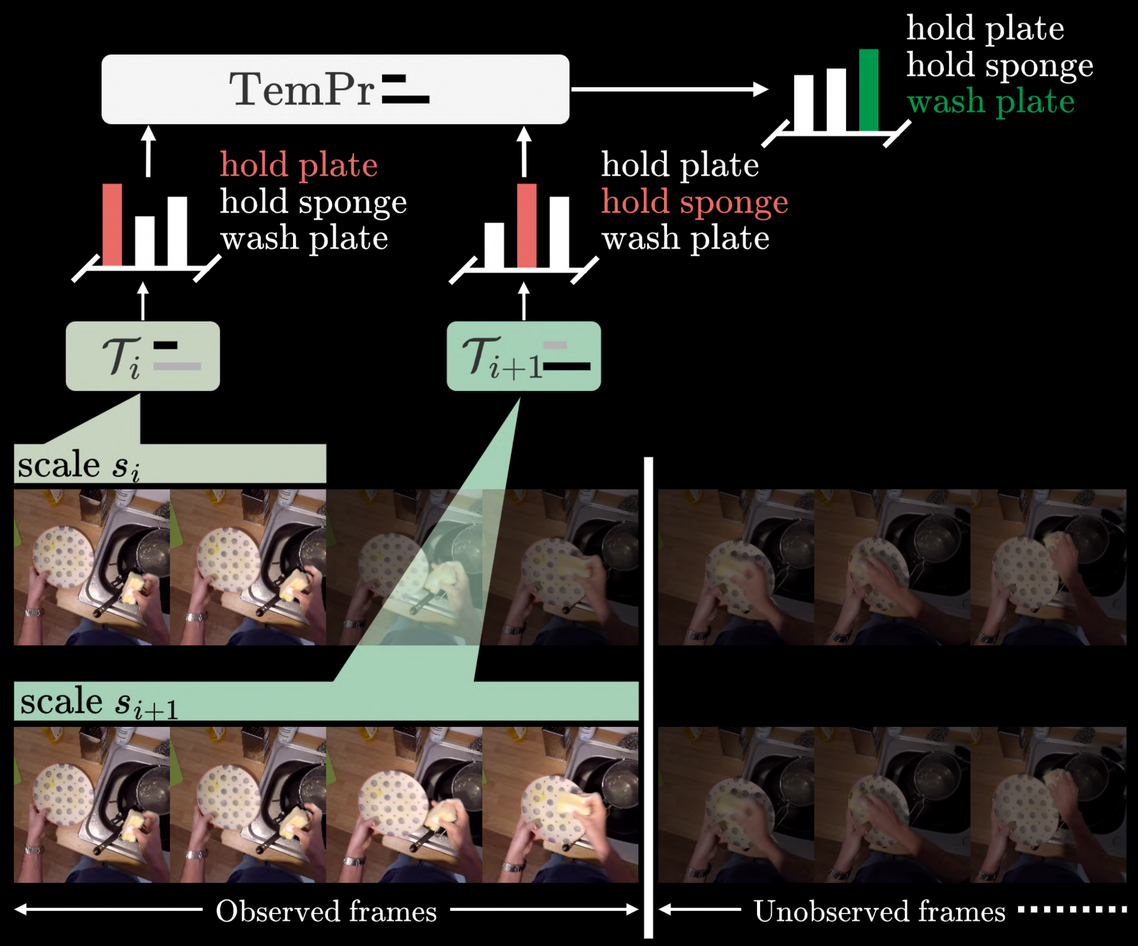
Overview

Early action recognition infers labels for partial observations of actions .

- We use a progressive fine-to-coarse temporal sampling strategy. Through this we define multiple scales over the observable part of a video.

- At each scale we use attention towers to capture discriminative representations and predict an action label. Predictions from each scale are combined adaptively based on both predictor confidence and similarity.

- We evaluate our method on UCF-101, EK-100, NTU-RGB, SSsub21, and SSv2.



Actions are not always observed in full

source: “*Roger Federer Serve Analysis by Patrick Mouratoglou”*, YouTube

Dealing with predictions — human cognition

Humans are quite good at making educated guesses.

Observed Action

Focus on kinematic information

Motor Memory

When I previously did this action the goal was…

"*drinking*"

We understand actions in a predictive and not reactive manner.

Prediction of ongoing actions

- Handshake
- Dance
- Sit

Estimations about the future rely on contextual information,

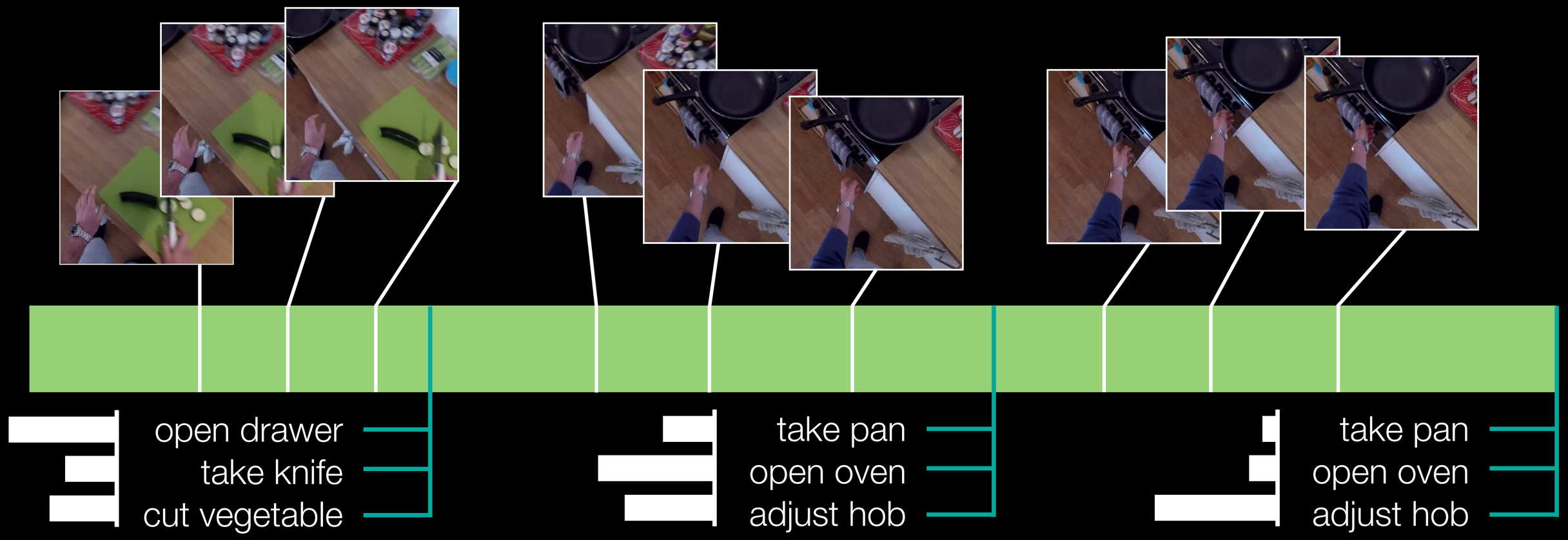
as well as partial motions.

- Hug
- Run
- Jump

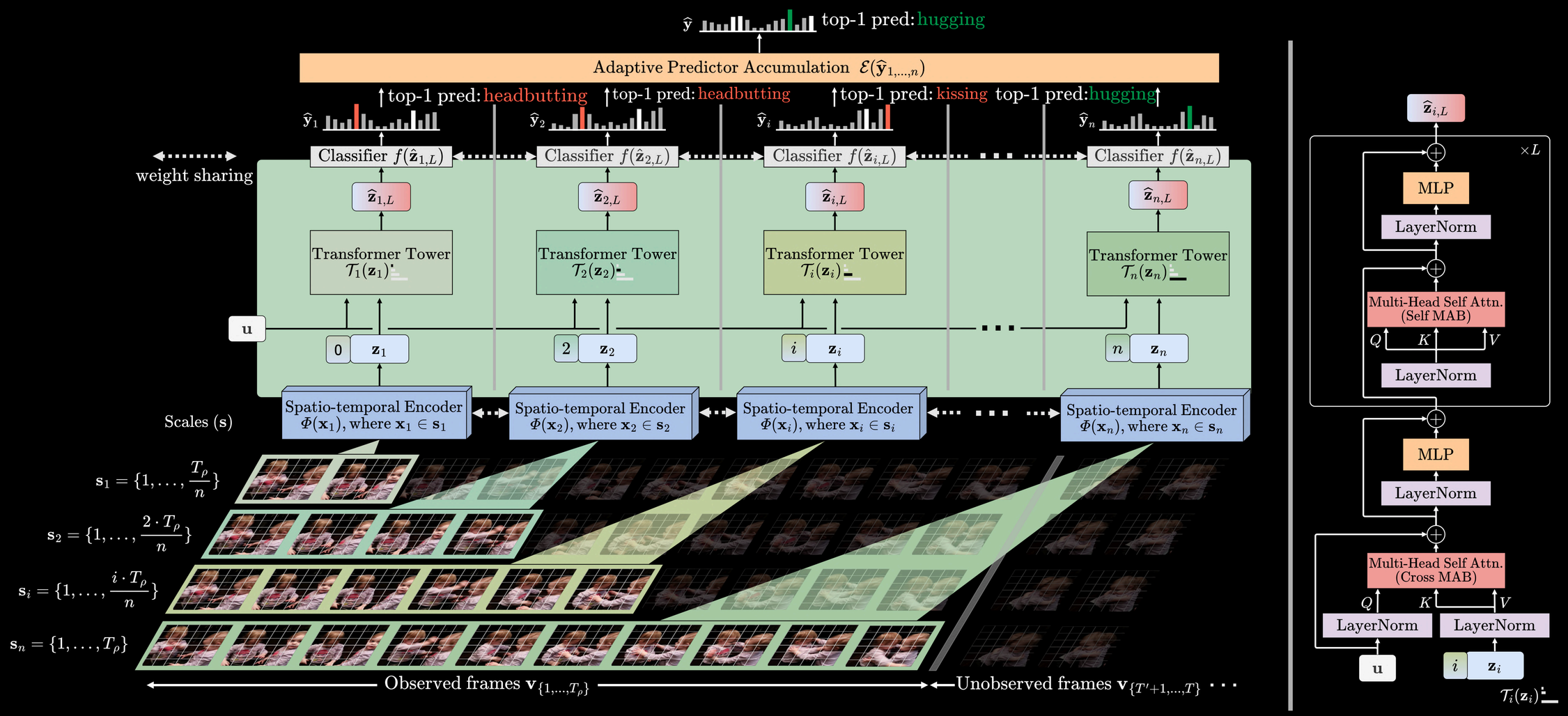
sources: *“Late Night with Seth Meyers”*, S10E75
“Parks and Recreation”, S4E1

Predictions throughout the action sequence

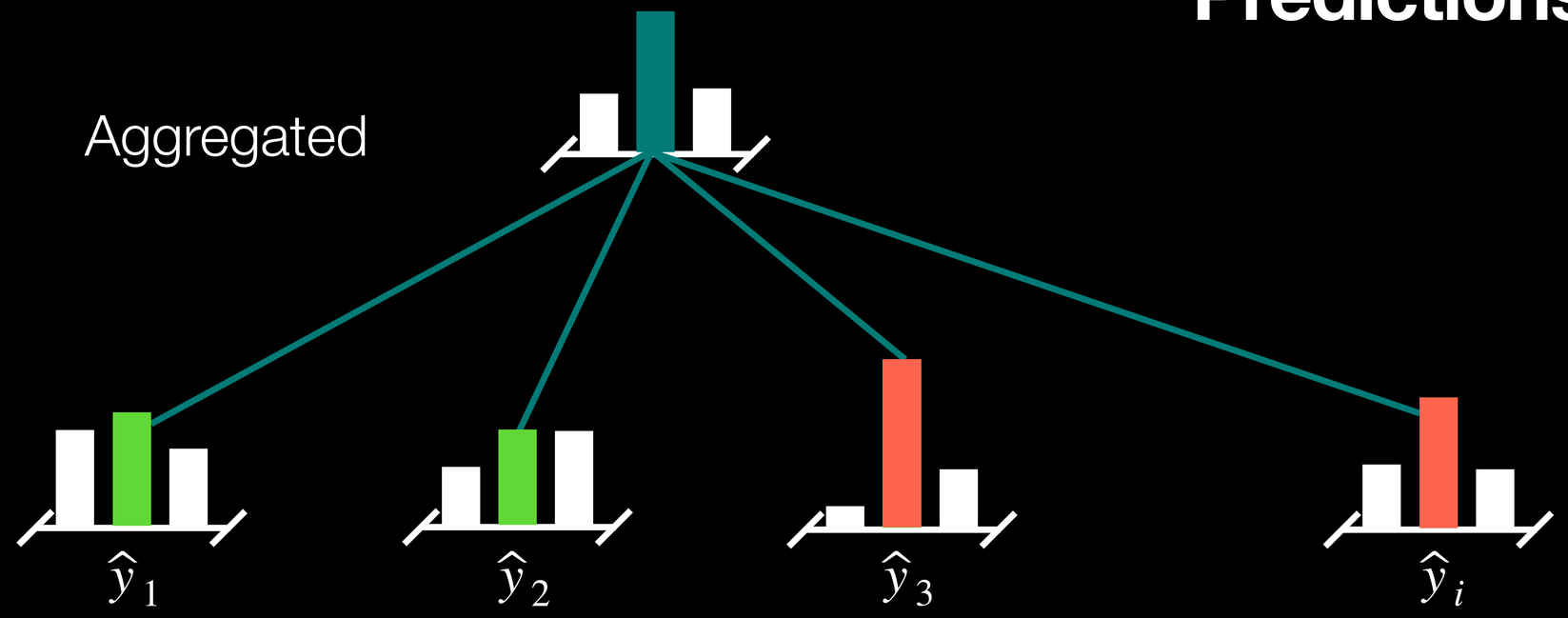
Capturing the evolution of the action at each stage.



TemPr model


top-1 pred: hugging
Adaptive Predictor Accumulation $\mathcal{E}(\hat{\mathbf{y}}_{1,\ldots,n})$
top-1 pred: headbutting
top-1 pred: headbutting
top-1 pred: kissing
top-1 pred: hugging
weight sharing
Classifier $f(\hat{\mathbf{z}}_{1,L})$
Classifier $f(\hat{\mathbf{z}}_{2,L})$
Classifier $f(\hat{\mathbf{z}}_{i,L})$
Classifier $f(\hat{\mathbf{z}}_{n,L})$
Transformer Tower $\mathcal{T}_1(\mathbf{z}_1)$
Transformer Tower $\mathcal{T}_2(\mathbf{z}_2)$
Transformer Tower $\mathcal{T}_i(\mathbf{z}_i)$
Transformer Tower $\mathcal{T}_n(\mathbf{z}_n)$
u
Spatio-temporal Encoder $\Phi(\mathbf{x}_1)$, where $\mathbf{x}_1 \in \mathbf{s}_1$
Spatio-temporal Encoder $\Phi(\mathbf{x}_2)$, where $\mathbf{x}_2 \in \mathbf{s}_2$
Spatio-temporal Encoder $\Phi(\mathbf{x}_i)$, where $\mathbf{x}_i \in \mathbf{s}_i$
Spatio-temporal Encoder $\Phi(\mathbf{x}_n)$, where $\mathbf{x}_n \in \mathbf{s}_n$
Scales (s)
$\mathbf{s}_1 = \{1, \ldots, \frac{T_\rho}{n}\}$
$\mathbf{s}_2 = \{1, \ldots, \frac{2 \cdot T_\rho}{n}\}$
$\mathbf{s}_i = \{1, \ldots, \frac{i \cdot T_\rho}{n}\}$
$\mathbf{s}_n = \{1, \ldots, T_\rho\}$
Observed frames $\mathbf{v}_{\{1,\ldots,T_\rho\}}$
Unobserved frames $\mathbf{v}_{\{T'+1,\ldots,T\}}$
$\times L$
MLP
LayerNorm
Multi-Head Self Attn. (Self MAB)
Multi-Head Self Attn. (Cross MAB)
Q
K
V

Predictions aggregation



Individual predictors

- **Individual confidence** of each predictor: $\underset{eM}{\mathscr{E}}(\hat{\mathbf{y}}_i) = \dfrac{e^{\hat{\mathbf{y}}_i}}{\sum_{k \in \mathbf{N}} e^{\hat{\mathbf{y}}_k}} \cdot \hat{\mathbf{y}}_i$

- **Collective agreement** between individual predictions $\underset{eICW}{\mathscr{E}}(\hat{\mathbf{y}}_i, \overline{\hat{\mathbf{y}}}) = \dfrac{e^{DSC(\hat{\mathbf{y}}_i, \overline{\hat{\mathbf{y}}})^{-1}}}{\sum_{k \in \mathbf{N}} e^{DSC(\hat{\mathbf{y}}_k, \overline{\hat{\mathbf{y}}})^{-1}}} \cdot \hat{\mathbf{y}}_i$

The final aggregation function takes the form : $\mathscr{E}(\hat{\mathbf{y}}_{1,\ldots,n})) = \sum_{i \in \mathbf{N}} \beta \cdot \underset{eICW}{\mathscr{E}}(\hat{\mathbf{y}}_i, \overline{\hat{\mathbf{y}}}) + (1-\beta) \cdot \underset{eM}{\mathscr{E}}(\hat{\mathbf{y}}_i)$

Accuracies over observation ratios (ρ) — UCF101

Top-1 accuracies (%) of action prediction methods on UCF-101 over different observation ratios (ρ). Methods are grouped w.r.t. the backbone used. We report **TemPr** results on 5 backbones. The best results per ρ are in **bold** and second best are underlined.

Method	Backbone	dim	Observation ratios (ρ)								
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DBDNet	ResNeXt101	3D	82.7	86.6	88.3	89.7	90.6	91.2	91.7	91.9	92.0
IGGNN			80.2	-	89.8	-	92.9	-	94.1	-	94.4
ERA			**89.1**	-	92.4	-	94.3	-	95.4	-	95.7
TemPr (ours)			85.7	91.4	92.1	92.7	93.5	93.9	94.4	94.6	94.9
TemPr (ours)	$X3D_M$	3D	87.9	93.4	94.5	94.8	95.1	**95.2**	**95.6**	96.4	**96.3**
TemPr (ours)	MoViNet-A4	3D	88.6	**93.5**	**94.9**	**94.9**	**95.4**	**95.2**	95.3	**96.6**	96.2
TemPr	MoViNet-A4	3D	87.3	93.1	94.9	94.6	95.2	94.9	94.6	95.1	95.0
TemPr			85.6	92.9	93.6	94.5	94.4	94.2	94.2	94.6	94.8
TemPr			85.2	92.1	92.5	92.9	93.3	93.7	93.5	93.8	93.7

Accuracies over observation ratios (ρ) — NTU-RGB/SSsub21/SSv2/EK-100

Top-1 accuracy (%) of EAP over different observation ratios (ρ).

(a) **NTU-RGB.**

Method	Observation ratios (ρ)					
	0.1	0.2	0.3	0.5	0.7	0.9
RankLSTM	11.5	16.5	25.7	48.0	61.0	66.1
DeepSCN	16.8	21.5	30.6	48.8	58.2	60.0
MSRNN	15.2	20.3	29.5	51.6	63.9	68.9
TS (2×L)	27.8	35.8	46.3	67.4	77.6	81.5
TemPr (ours)	**29.3**	**38.7**	**50.2**	**70.1**	**78.8**	**84.2**

(b) **SSsub21.**

Method	Observation ratios (ρ)					
	0.1	0.2	0.3	0.5	0.7	0.9
mem-LSTM	14.9	17.2	18.1	20.4	23.2	24.5
MS-LSTM	16.9	16.6	16.8	16.7	16.9	17.1
MSRNN	20.1	20.5	21.1	22.5	24.0	27.1
GGN	21.2	21.5	23.3	27.4	30.2	30.5
IGGN	22.6	-	25.0	28.3	32.2	34.1
TemPr (ours)	**28.4**	**34.8**	**37.9**	**41.3**	**45.8**	**48.6**

(c) **SSv2.**

Method	Obs. ratios (ρ)			
	0.1	0.3	0.5	0.7
Baseline (Inference)	6.9	17.6	28.9	36.0
Baseline (Fine-tuned)	14.4	23.5	31.1	39.6
TemPr (ours)	**20.5**	**28.6**	**41.2**	**47.1**

(d) **EK-100.**

Method	Verb						Noun						Action					
	Observation ratios (ρ)																	
	0.1	0.2	0.3	0.5	0.7	0.9	0.1	0.2	0.3	0.5	0.7	0.9	0.1	0.2	0.3	0.5	0.7	0.9
Baseline (Inference)	17.3	19.7	27.0	48.7	60.5	64.2	19.5	21.7	25.3	38.5	46.7	49.1	5.4	7.6	11.1	24.3	34.1	37.6
Baseline (Fine-tuned)	20.6	21.8	29.4	49.8	61.3	64.3	21.3	24.2	27.6	39.4	47.3	49.1	6.9	9.1	12.8	25.5	34.9	37.5
TemPr (ours)	**21.4**	**22.5**	**34.6**	**54.2**	**63.8**	**67.0**	**22.8**	**25.5**	**32.3**	**43.4**	**49.2**	**53.5**	**7.4**	**9.8**	**15.4**	**28.9**	**37.3**	**40.8**

Ablations on UCF-101

Ablation studies on UCF-101 with TemPr across obs. ratios. We use ♠ to denote softmax during training and ♣ for $\theta = \frac{1}{2n}$.

(a) **Video Scales Strategy.**

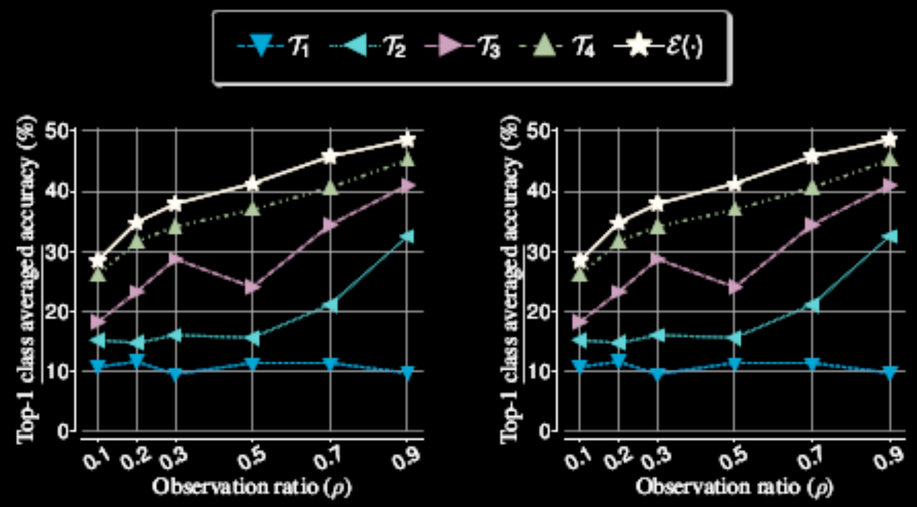
Scale strategy	Observation ratios (ρ)			
	0.2	0.4	0.6	0.8
full	86.4	88.3	88.8	89.0
equal	83.7	84.6	86.3	87.1
random	88.8	89.7	90.2	90.6
decreasing	90.0	**90.9**	91.6	**92.6**
increasing	**90.2**	**90.9**	**91.8**	92.3

(b) **Aggregation function.**

Aggregation	ρ	
	0.2	0.4
avg	89.5	90.1
softmax	87.8	89.4
top♠	84.6	87.5
gate (θ=0.1)	85.4	88.5
ICW	89.7	90.1
weighted	88.5	89.0
weighted (θ) ♣	83.4	85.8
adaptive ($\mathcal{E}(\cdot)$)	**90.2**	**90.9**

(c) **Weight sharing** over attention towers and classifiers.

Weight sharing		ρ		
MAB	$f(\cdot)$	0.2	0.4	0.6
✓	✗	73.4	76.2	79.0
✗	✗	84.7	85.8	87.3
✓	✓	89.2	90.0	90.7
✗	✓	**90.2**	**90.9**	**91.8**



Class-based ablations

Top tower predictors per class and observation ratio for TemPr . Towers $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and , $\mathcal{T}_4$ are highlighted for better readability.

class name	Observation ratios ρ					
	0.1	0.2	0.3	0.5	0.7	0.9
Putting smthng similar to other things ...	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$
Showing smthng behind smthng	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$
Holding smthng	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Poking ... smthng without ... collapsing	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Pretending to sprinkle air onto smthng	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$
Pulling two ends of smthng ... stretched	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Putting smthng into smthng	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$
Pretending to turn smthng upside down	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$
Poking a stack of smthng ... collapses	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_e$
Pulling smthng from left to right	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$
Pushing smthng from left to right	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Pretending to open smthng without ...	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$
Opening smthng	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$
Showing a photo of smthng ...	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_1$
Stuffing smthng into smthng	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_2$
Putting smthng on the edge of smthng ...	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_1$
Picking smthng up	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$
Closing smthng	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_2$
Putting smthng upright on the table	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_2$
Turning smthng upside down	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_1$
Pulling two ends of smthng ... two pieces	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_2$

We evaluate the top-performing tower for each class across observation ratios (ρ) .

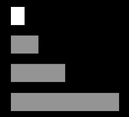
- Towers of larger scales perform better for classes that include long-term dependencies; e.g.
  Poking a stack of something without the stack collapsing or
  Pretending to sprinkle air onto something.

- Towers for smaller scales, are better suited for classes such as
  Picking something up or
  Closing something.

Qualitative results

$\rho = 0.3$

$\mathcal{T}_1$



Pulling two … gets stretched: 13.48
Pulling two … into two pieces: 11.10
Turning something … down: 5.80

$\mathcal{T}_2$

Pulling two ... two pieces: 11.31
Putting two ... gets stretched: 6.87
Stuffing ... into something: 1.30

$\mathcal{T}_3$

Putting two ... two pieces: 12.03
Putting two ... two pieces: 4.92
Putting something ... something: 2.71

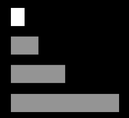
$\mathcal{T}_4$

Holding something: 8.58
Putting two ... two pieces: 8.46
Pulling something ... to right: 4.99

Qualitative results

$\rho = 0.3$

$\mathcal{T}_1$



Putting ... something: 9.25
Stuffing ... into something: 4.48
Showing ... something: 2.14

$\mathcal{T}_2$

Stuffing ... into something: 6.75
Putting ... something: 6.35
Closing something: 6.03

$\mathcal{T}_3$

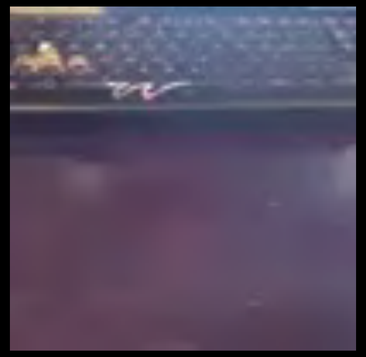
Stuffing ... into something: 9.71
Opening something: 8.19
Putting ... something: 6.75

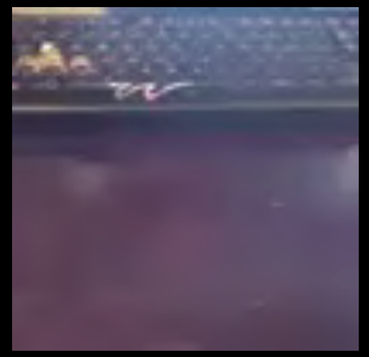
$\mathcal{T}_4$

Holding something: 5.54
Closing something: 2.24
Pulling something ... something: 2.23

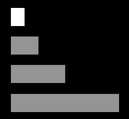
Qualitative results

$\rho = 0.3$

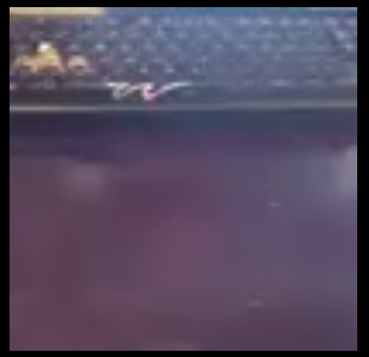




$\mathcal{T}_1$

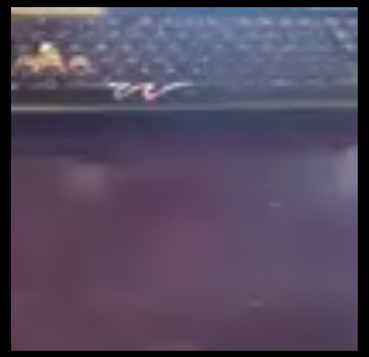


Putting something ... table: 10.11
Pretending ... onto something: 8.18
Holding something: 5.93



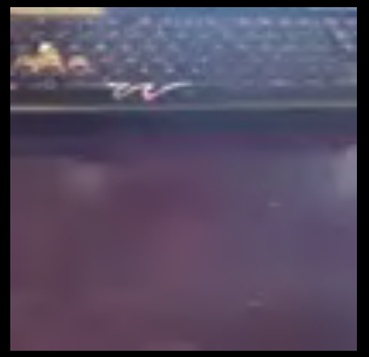
$\mathcal{T}_2$

Pulling two ... two pieces: 6.29
Putting something ... table: 5.85
Pulling two ... gets stretched: 3.36



$\mathcal{T}_3$

Putting two ... gets stretched: 9.22
Putting two ... two pieces: 5.55
Putting something ... table: 2.43

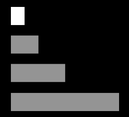


$\mathcal{T}_4$

Putting two ... gets stretched: 11.98
Putting two ... two pieces: 8.72
Putting something ... something: 4.20

Qualitative results

$\rho = 0.7$

$\mathcal{T}_1$ 

$\mathcal{T}_2$ 

$\mathcal{T}_3$ 

$\mathcal{T}_4$ 

Rock-paper-scissors: 3.43
Whisper: 3.32
Shaking hands: 3.28

Touch pocket: 6.89
Whisper: 5.11
Pat on back: 5.04

Pat on back: 7.84
Whisper: 7.45
Punch/slap: 6.23

Pat on back: 9.31
Whisper: 7.61
Knock over: 5.93

Links

Project website

GitHub code