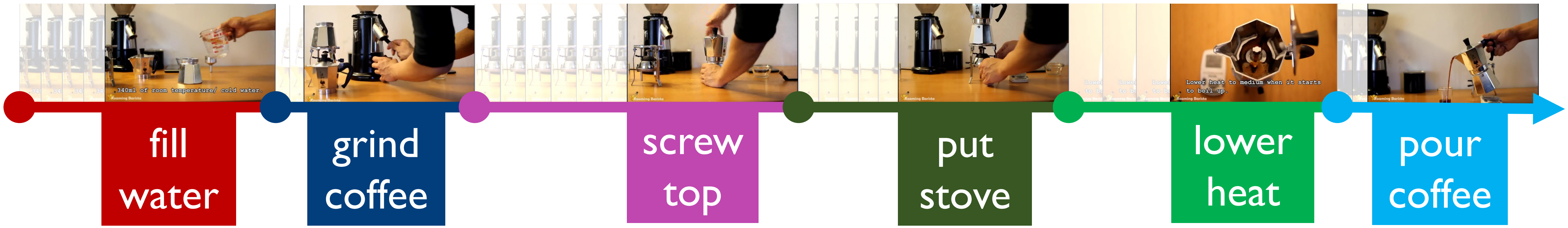


an introduction to

Temporal Action Segmentation

Angela Yao

24.10.2022

NUS | Computing
National University of Singapore

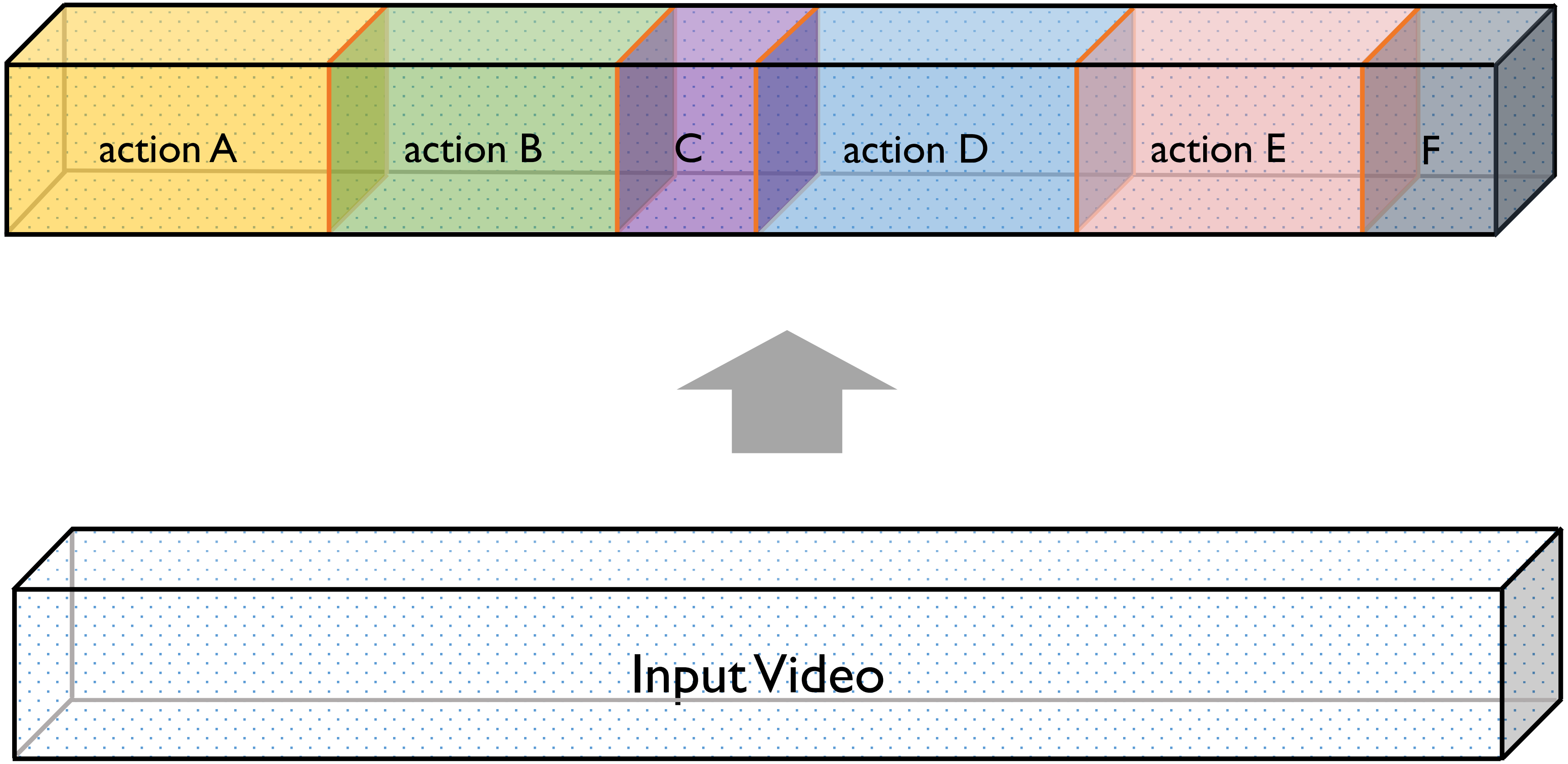
Table of Contents

NUS National University of Singapore | Computing

https://arxiv.org/pdf/2210.10352.pdf

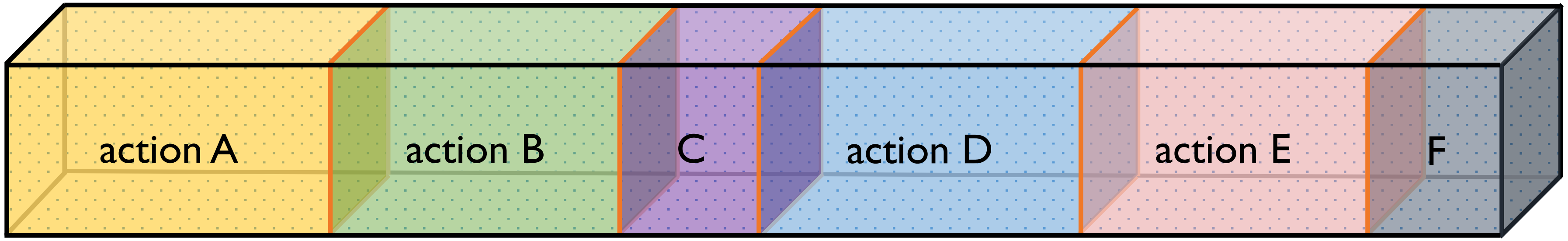
Task Definition

<u>Temporal Segmentation</u>:
Assign an action label to each frame of the video



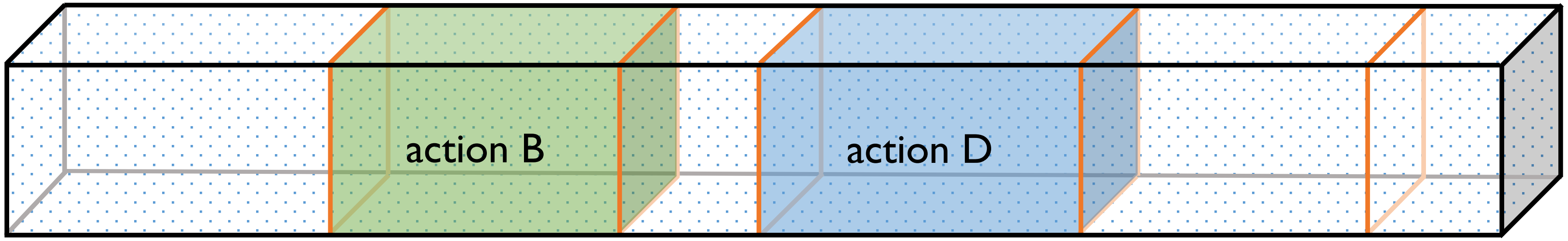
Segmentation vs. Localization

<u>Temporal Segmentation</u>:
Assign an action label to each frame of the video



Segmentation vs. Detection in Time

<u>Temporal Localization</u>:
Find temporal boundaries of specific action classes



Temporal Action Segmentation Datasets

Breakfast Actions

50 Salads

GeorgiaTech Egocentric Activities

Dataset	Year	Duration	# Videos	# Segments	# Activity	# Action	Domain	View
GTEA	'11	0.4h	28	0.5K	7	71	cooking	egocentric
50Salads	'13	5.5h	50	0.9K	1	17	cooking	top-view
Breakfast	'14	77h	1712	11K	10	48	cooking	3rd person

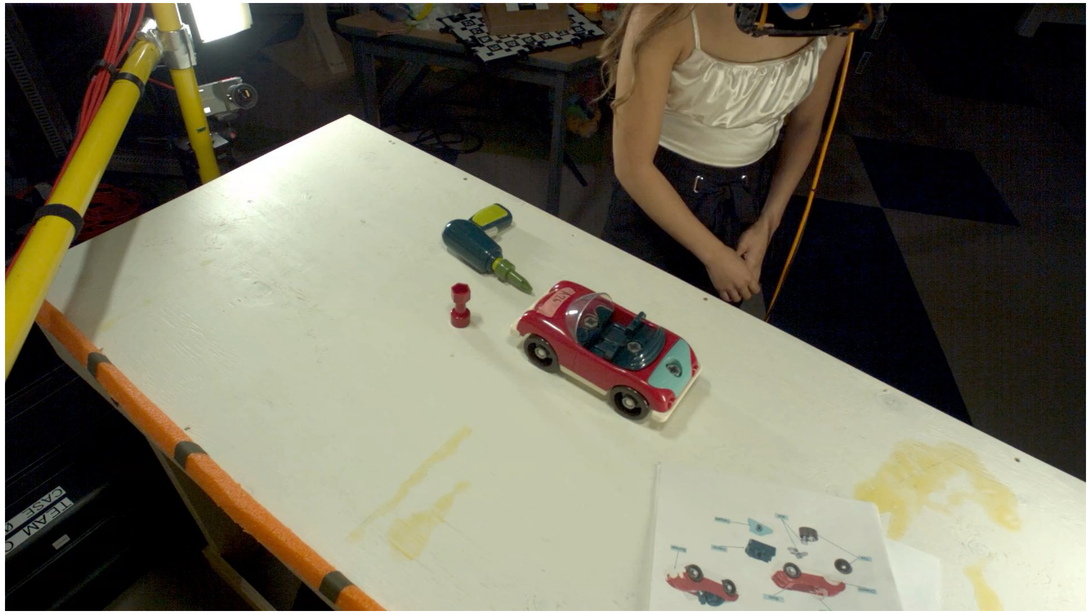
small-scale

* moving out of the kitchen domain

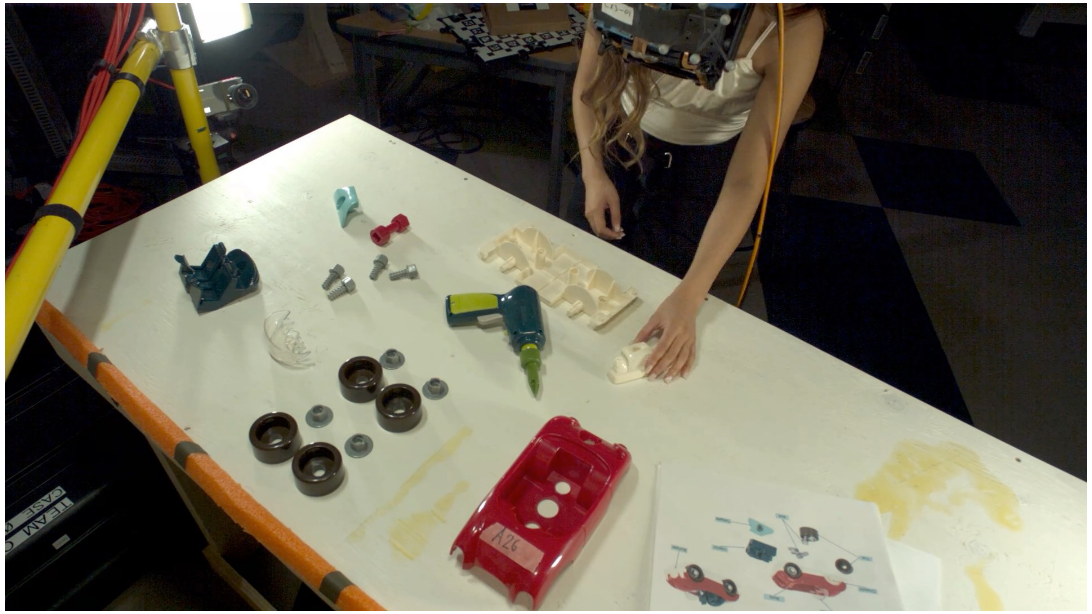
* richness in temporal variation

New Dataset – Assembly101

Disassembly



Assembly



Dataset	total hours	# videos	avg. video length (min)	# segments
50Salads [47]	4.5	50	6.4	899
Breakfast [22]	77.0	1,712	2.3	11,300
Assembly101	513.0	4,321	7.1	104,759

Temporal Variation in Datasets

NUS | Computing
National University of Singapore

Repetition score:

$$1 - \frac{u_i}{g_i},$$

number of unique actions in video i

number of total actions in video i

for video i u_i = 4, g_i → 7

Repetition score = $1 - \frac{4}{7}$ = 0.42

for video i u_i = 7, g_i → 7

Repetition score = $1 - \frac{7}{7}$ = 0

Repetition score ranges between $[0,1)$

- 0 indicates no repetition
- Closer to 1 indicates more repetitions

Temporal Variation in Datasets

NUS National University of Singapore | Computing

Order variation score:

$$1 - \frac{e(R,G)}{\max(|R|,|G|)}$$

Average edit distance $e(R,G)$ between every pair of sequences R and G, normalized w.r.t. the maximum sequence length of R and G.

for video i, $e(R,G) = 1$

$1 - \frac{1}{5} = 0.8$

for video i, $e(R,G) = 1$

$1 - \frac{5}{5} = 0$

Repetation score range $[0,1]$

- 1 → strict ordering
- closer to 0 → more deviation in action order

Temporal Variation in Datasets

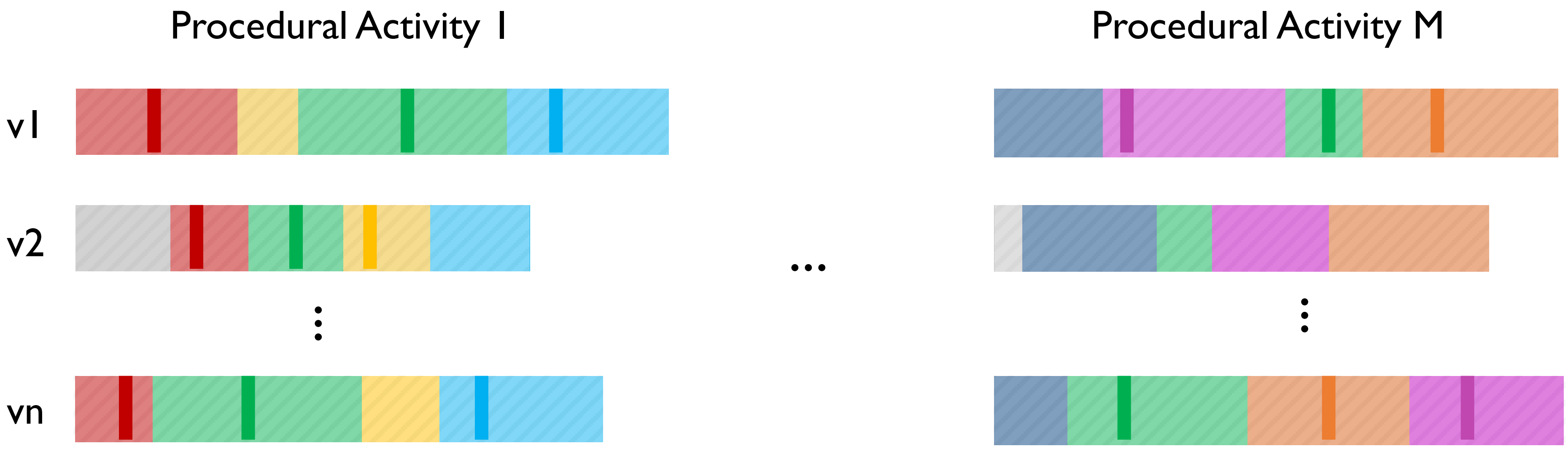
Dataset	Repetition ↑	Order Variation ↓
Breakfast	0.11	0.15
50Salads	0.08	0.02
Assembly101	0.18	0.05

High score of order variation indicates that actions follow a strict ordering.

A challenging benchmark for modelling the temporal relations between actions.

Levels of Supervision

- "Unsupervised" = Activity-Level Supervision: procedural activity label
- Fully supervised: labels of every frame in every video
- Semi supervised: labels of every frame in some videos
- Weakly supervised: transcripts, action sets, labels of some frames



Evaluation Measures

Frame-based

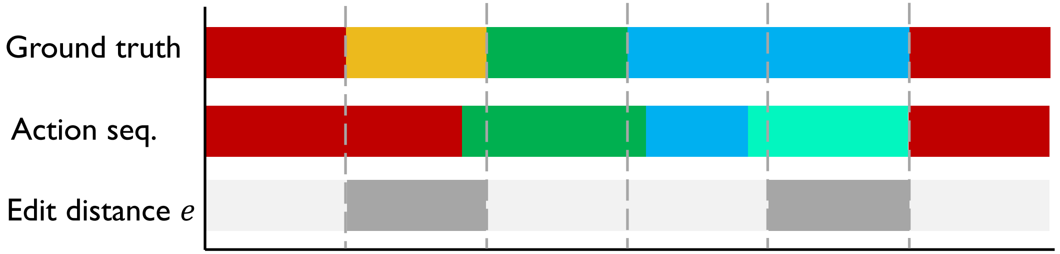
Acc/MoF

$$Acc = \frac{\#\ of\ correct\ frames}{\#\ of\ all\ frames}$$

Estimates how accurate frame wise predictions are.

Segment-based

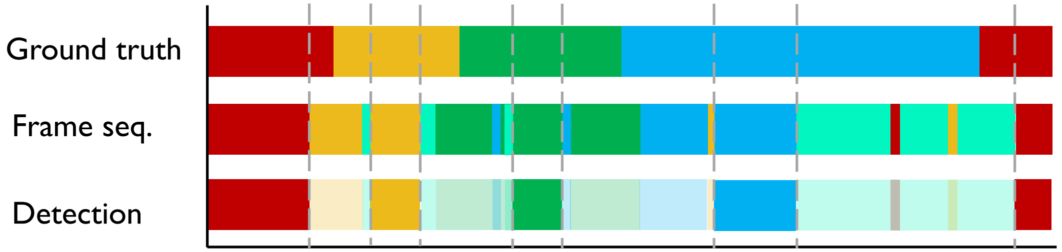
Edit score



$$Edit = \frac{1 - e(X,Y)}{\max(|X|,|Y|)} \cdot 100$$

Tolerant to small boundary shifts, as long as the sequence order is correct, the score is high

F1 score



$$F1 = 2 \cdot \frac{precision * recall}{precision + recall}$$

True positives marked by dashed line have an IOU with ground truth based on threshold τ. Remaining segments (dimmed) are false positives

Hungarian Matching in Unsupervised TAS

- Unsupervised segmentation results: segments are grouped wrt each other, groups must be assigned to action label for evaluation
- Assignment via Hungarian matching

Assignment indicator (1/0)

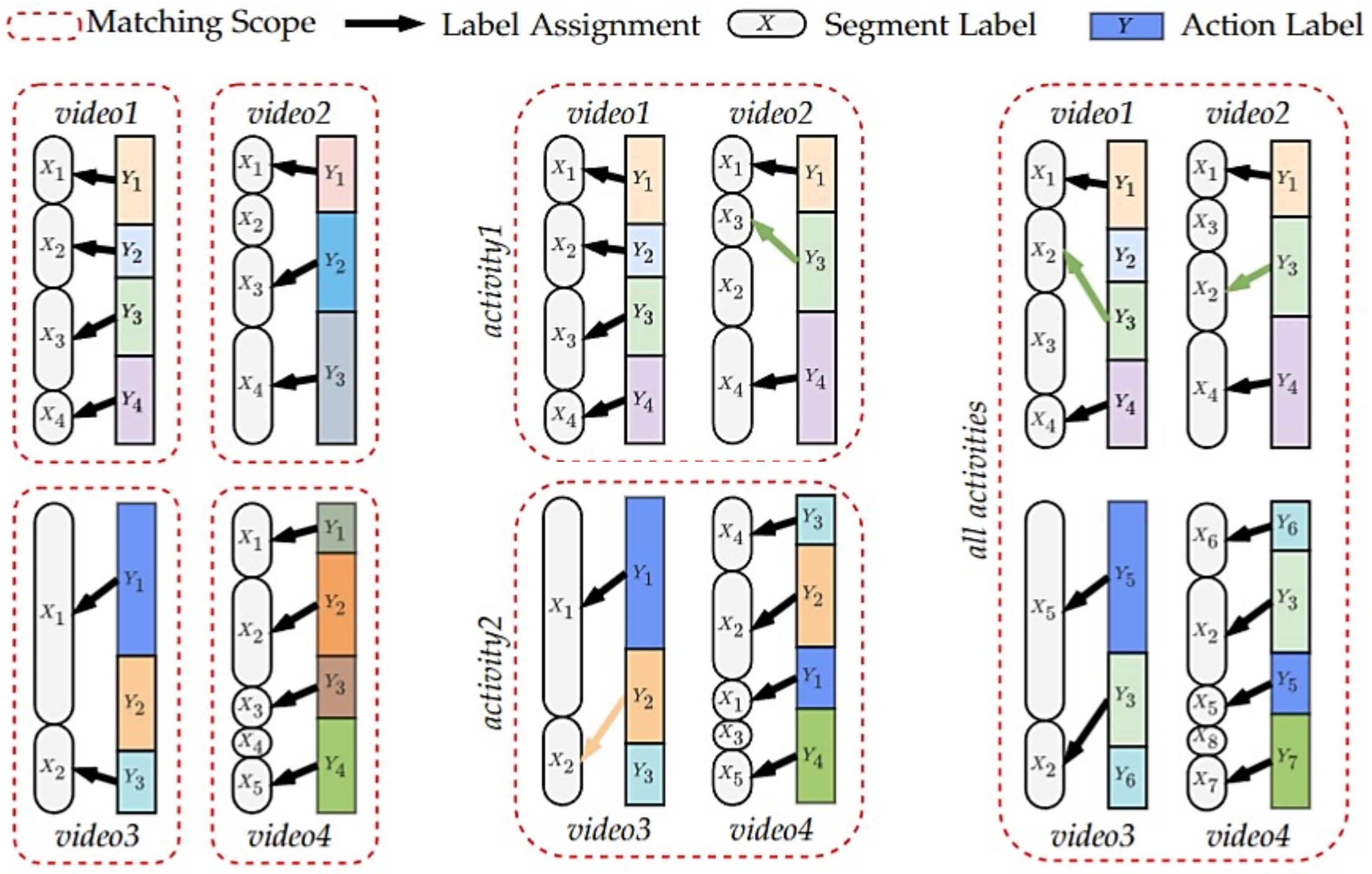
$$\hat{\mathcal{A}} = \arg\max_{\mathcal{A}} \sum_{n,m} \mathcal{A}_{n,m} \cdot \overbrace{I(X_n, Y_m)}^{\text{\# of frames with label m in cluster n}},$$

$$\text{s.t.} \quad |\mathcal{A}| = \min(N, M)$$

best match

groups of segments

semantic labels



Video-level matching

Activity-level matching

Global matching

Table of Contents

Framewise Representations

Action segmentation uses pre-computed video features as input:

- IDT: Improved Dense Trajectories[a]
  - Raw features encoded by Fisher Vectors [b] to capture 1st & 2nd order statistics
  - Reduced to 64D by PCA
- I3D: Inflated 3D ConvNet [c]
  - 2048D: concatenate RGB stream (1024D) + optical flow stream (1024D)

[a] Wang and Schmid, ICCV'13
[b] Perronnin et al, ECCV'10
[c] Carreira and Zisserman, CVPR'17

Framewise Representations: Additional Learning

- Discriminative clustering [a]
- Unsupervised contrastive learning [b]
  - Positive pairs based on K-means clustering & time threshold
- Temporal & Visual Embedding
  - same actions tend to occur in a similar temporal range
  - embedding learned w/ pretext task of frame-wise timestamp prediction to [c]
  - temporal embedding augmented to include visual cues [d]

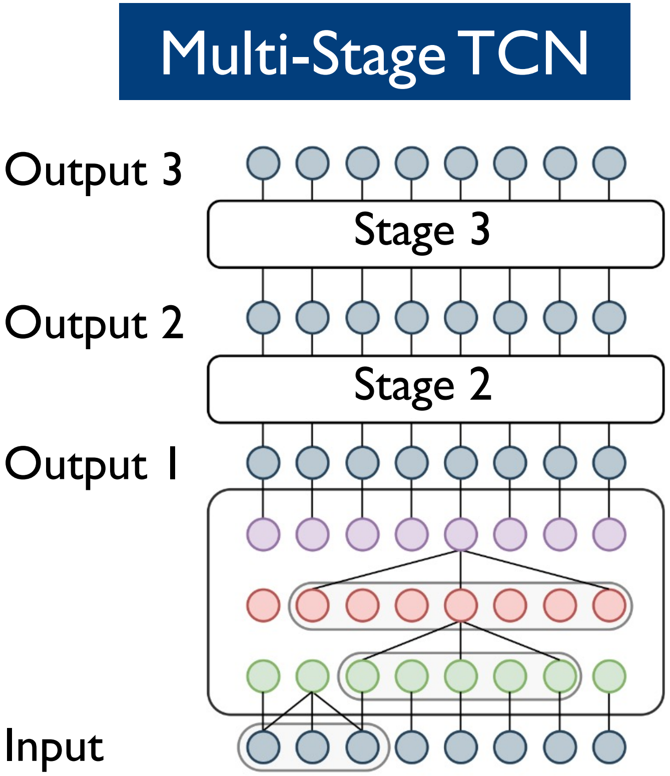
[a] Sener and Yao, CVPR'18
[b] Singhania et al, AAAI'22
[c] Kukleva et al, CVPR'19
[d] VidalMata et al, WACV'21

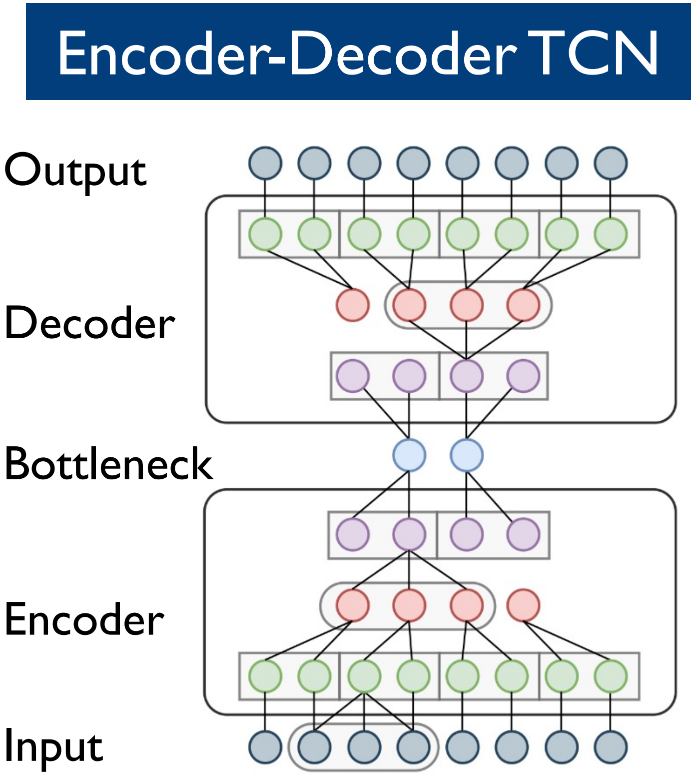
Temporal Modelling

- Recurrent Neural Networks (RNNs)
  - Bi-directional GRUs [a], [b]
- Temporal Convolutional Networks (TCNs)
  - Encoder-Decoder [c],[d]
  - Multi-stage TCNs [e],[f]
- Attention & Transformer Architectures
  - Temporal Aggregates [g]
  - ASFormer[h]
  - UVAST[i]

[a] Singh et al, CVPR'16
[b] Richard et al, CVPR'17
[c] Lea et al, CVPR'17
[d] Lei et al, CVPR'18
[e] Farha et al, CVPR '19
[f] Singhania et al, arxiv'21
[g] Sener et al, ECCV'20
[h] Yi et al, BMVC'21
[i] Behrmann, ECCV'22

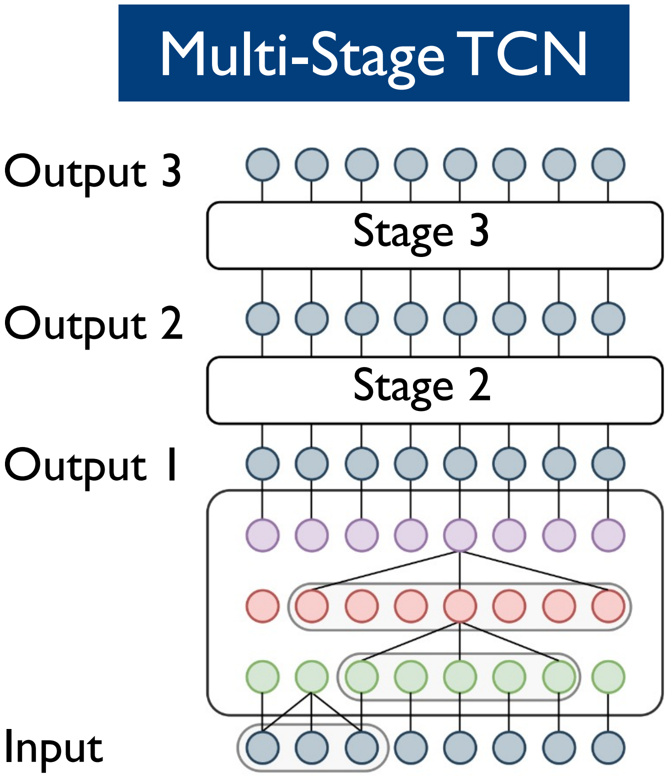
Temporal Convolutional Networks

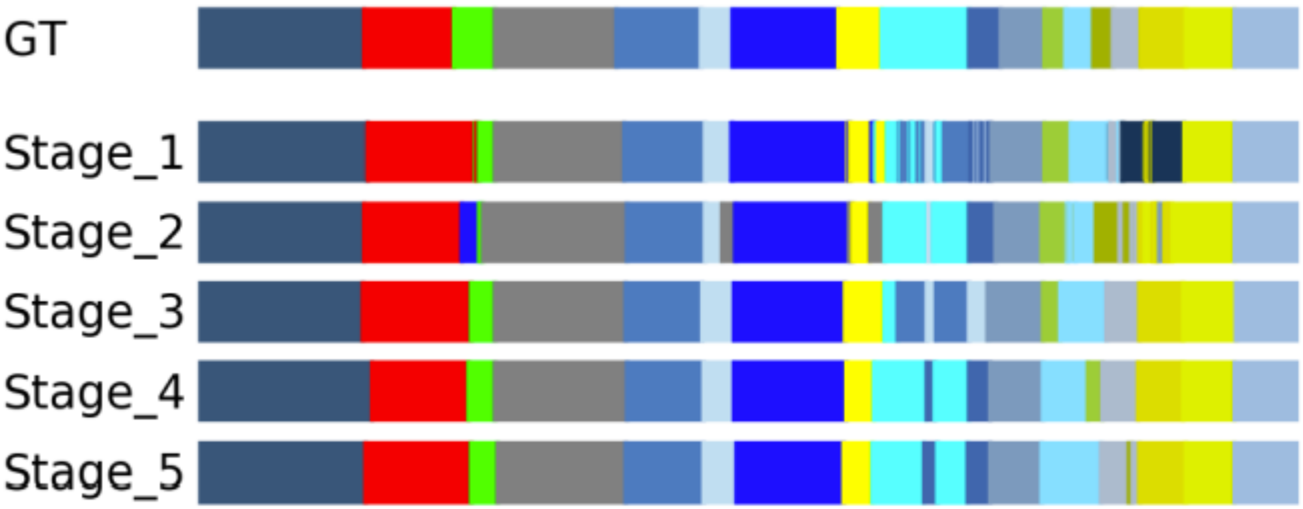




- Fixed temporal resolution vs. shrink-then-expand
- Successive probability refinement vs. decoupled representation + classification

MS-TCN





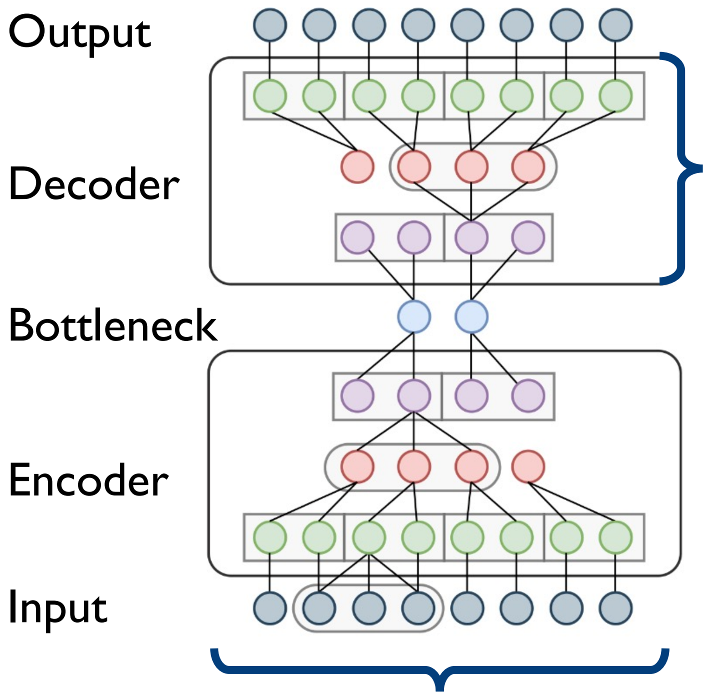
	F1@{10,25,50}			Edit	Acc
MS-TCN (2 stages)	55.5	52.9	47.3	47.9	79.8
MS-TCN (3 stages)	71.5	68.6	61.1	64.0	78.6
MS-TCN (4 stages)	76.3	**74.0**	**64.5**	67.9	**80.7**
MS-TCN (5 stages)	**76.4**	73.4	63.6	**69.2**	79.5

Effect of the number of stages on the 50Salads dataset.

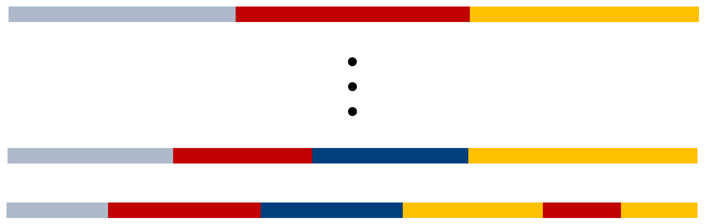
Yazan Abu, and Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation CVPR, 2019.

Encoder-Decoder TCNs

Encoder-Decoder TCN



Project each layer into an output and upsample.



Coarse-to-Fine Ensembling

$$\mathbf{p}_t^{ens} = \sum_i \alpha_i \cdot \mathrm{Up}\left[\mathbf{p}^{(i)}, t\right], \quad \sum_i \alpha_i = 1, \ \alpha_i > 0$$

Random multi-resolution max-pooling as a feature augmentation strategy

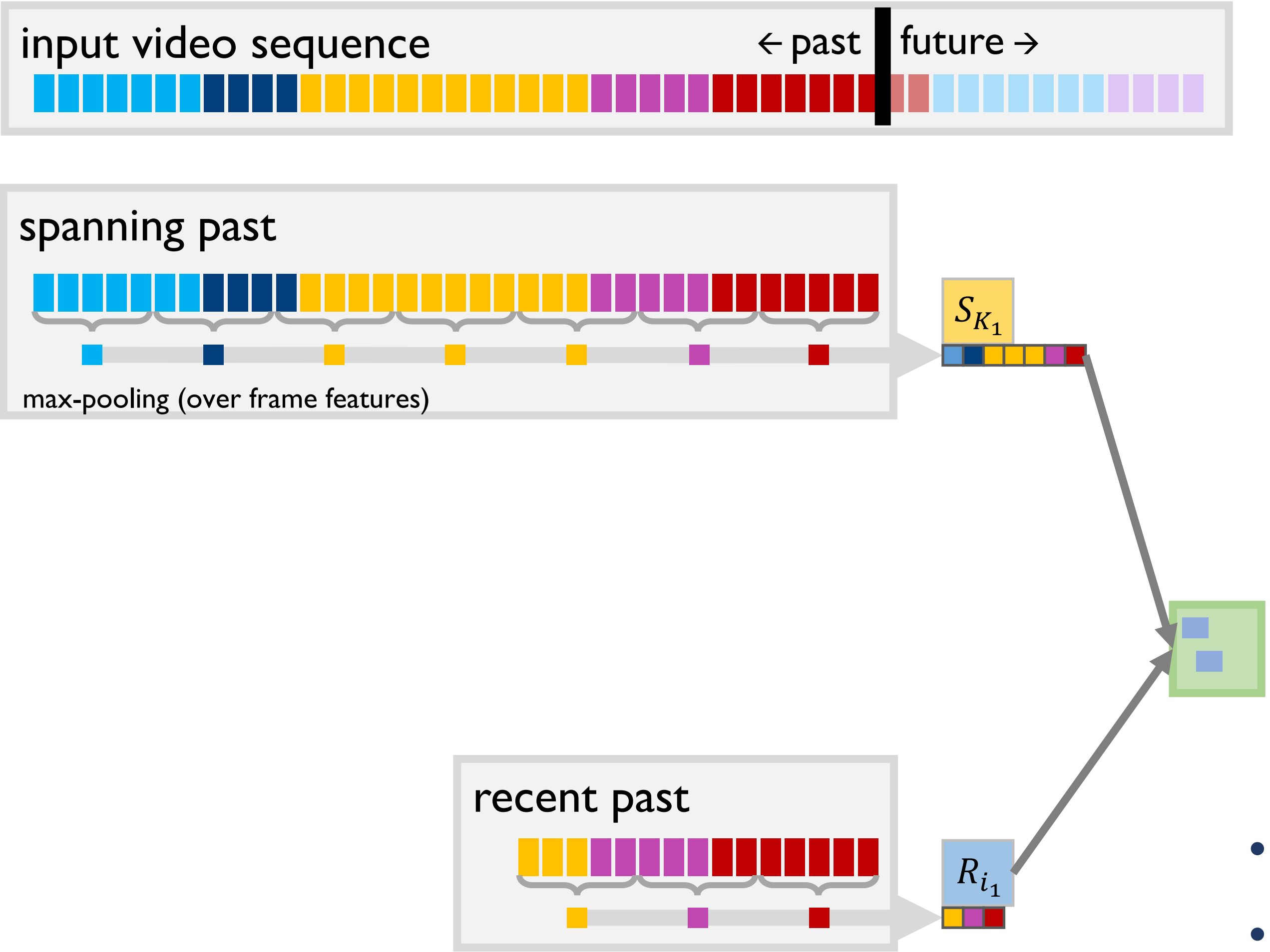
Singhania, Rahaman & Yao. Coarse-to-fine multi-resolution temporal convolution network. *arXiv:2105.10859*, 2021.

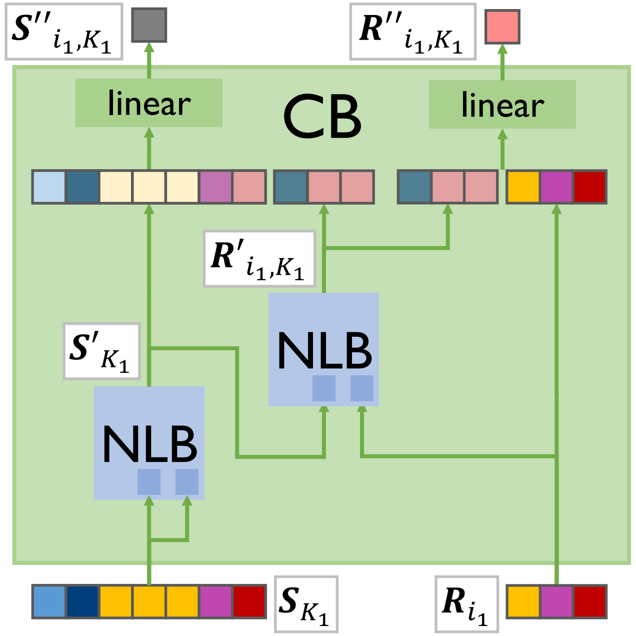
Temporal Modelling

- Recurrent Neural Networks (RNNs)
  - Bi-directional GRUs [a], [b]
- Temporal Convolutional Networks (TCNs)
  - Encoder-Decoder [c],[d]
  - Multi-stage TCNs [e],[f]
- Attention & Transformer Architectures
  - Temporal Aggregates [g]
  - ASFormer[h]
  - UVAST[I]

[a] Singh et al, CVPR'16
[b] Richard et al, CVPR'17
[c] Lea et al, CVPR'17
[d] Lei et al, CVPR'18
[e] Farha et al, CVPR '19
[f] Singhania et al, arxiv'21
[g] Sener et al, ECCV'20
[h] Yi et al, BMVC'21
[I] Behrmann, ECCV'22

Temporal Aggregates: Coupling Block (CB)



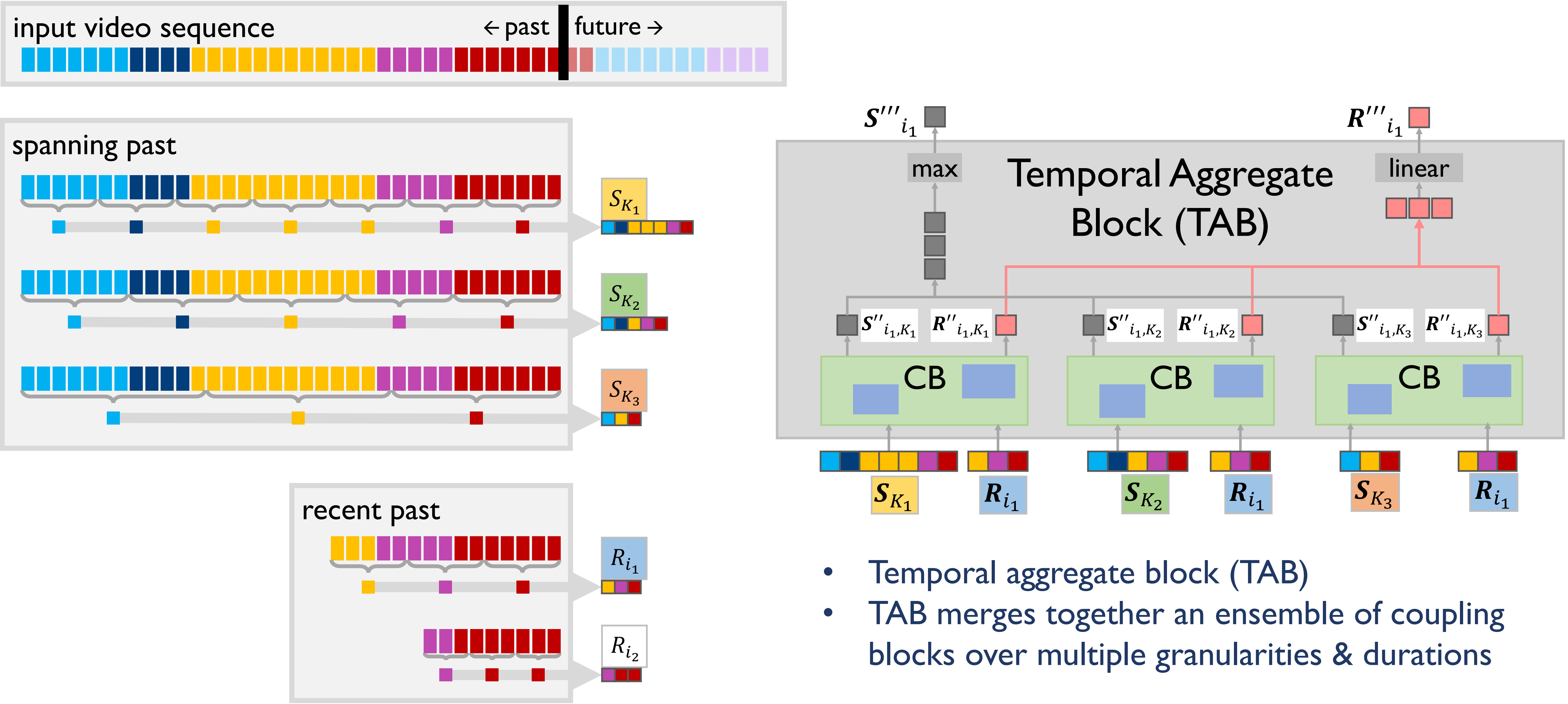


- relates recent observations to long-range past
- **attention-reweighted** spanning & recent outputs

Sener, Singhania & Yao. Temporal Aggregate Representations for Long-Range Video Understanding. *ECCV'20*

Model – Temporal Aggregate Block (TAB)

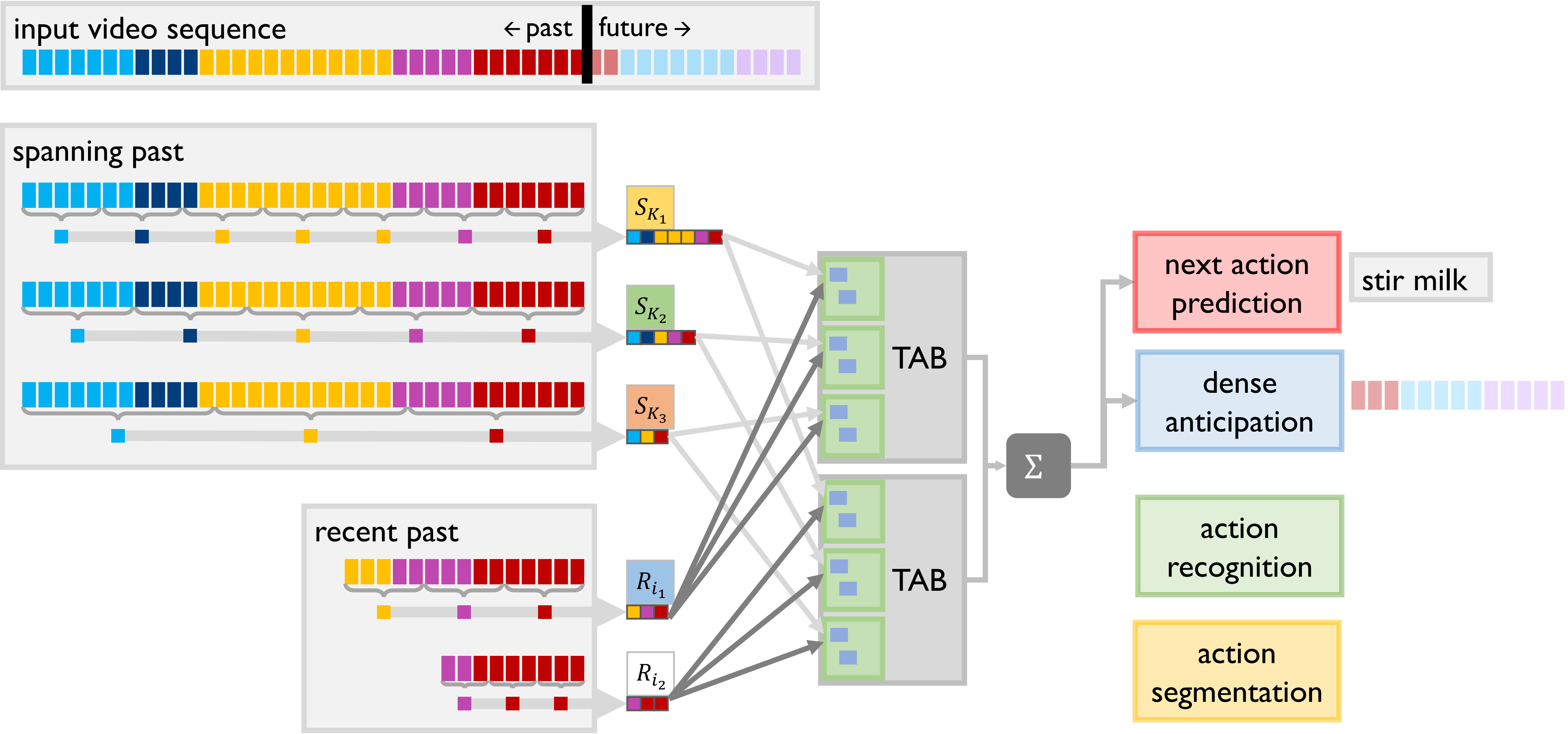
NUS National University of Singapore | Computing



- Temporal aggregate block (TAB)
- TAB merges together an ensemble of coupling blocks over multiple granularities & durations

Sener, Singhania & Yao. Temporal Aggregate Representations for Long-Range Video Understanding. *ECCV'20*

Model – Ensemble of TABs



Sener, Singhania & Yao. Temporal Aggregate Representations for Long-Range Video Understanding. *ECCV'20*

Transformer Methods

ASFormer

- encoder + several decoders for iterative refinement.
- MS-TCN-like architecture – each dilated convolutional layer replaced with a self-attention block with instance normalization

UVAST

- similar encoder to ASFormer
- decodes action segments auto-regressively
- state-of-the-art in Edit & F1-score, indicating less over-segmentation, lower MoF indicating low accuracy at the boundaries.

Sequential Modelling

- Hidden Markov Model

  Richard et al. CVPR'18
  Li and Todorovic CVPR'20
  Kukleva et al. CVPR'19
  Li and Todorovic CVPR'21

- Generalized Mallows Model

  Sener and Yao CVPR'18

- Dynamic Time Warping

  Chang et al. CVPR'19
  Ding and Yao ECCV'22

Table of Contents

NUS National University of Singapore | Computing

1. Task
- Definition
- Datasets
- Forms of Supervision
- Evaluation Measures

2. Core Techniques
- Frame-wise Representation
- Temporal Modelling
- Sequential Modelling

3. SoTA Trends
- Fully Supervised
- Weakly-Supervised
- Unsupervised
- Semi-Supervised

Fully Supervised

video i 

video j 

Methods improving representations:

Singh et al. CVPR'16
Lea et al. ECCV'16
Sener et al. ECCV'20

Temporal Convolutional Networks based solutions

ED-TCN - Lea et al, CVPR'17
MS-TCN - Farha et al, CVPR'19
C2F-TCN - Singhania et al, arxiv'21

Transformers

ASFormer - Yi et al, BMVC'21
UVAST - Behrmann, ECCV'22

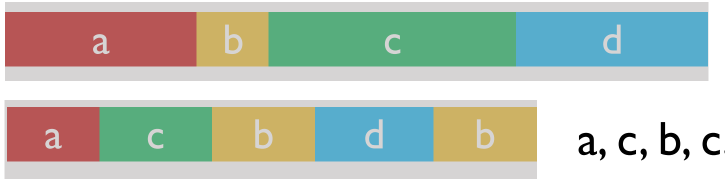
Improving existing architectures & outputs

MTDA + MS-TCN – Chen et al. CVPR'20
BCN + MS-TCN - Wang et al. ECCV'20
FIFA + UVAST - Souri et al. GCPR'21

Weak Supervision – action labels

NUS National University of Singapore | Computing

Transcripts: Ordered List of Actions



a, b, c, d

a, c, b, c, b

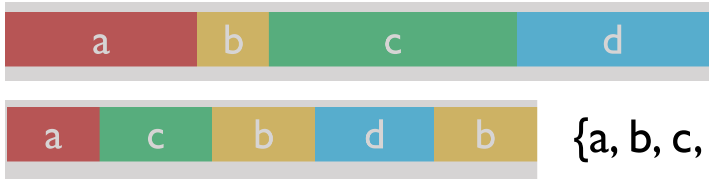
Two-Stage approaches refine segments iteratively.

HMMs - Kuehne et al. CVIU'17
RNNs - Richard et al. CVPR'17
TCNs - Ding et al. CVPR'18

Single-stage approaches directly learn segmentation.

ECTC - Huang et al. ECCV'16
NN-Viterbi - Richard et al. CVPR'16
D3TW - Chang et al. CVPR'19
CDFL - Li et al. ICCV'19

Transcripts: Set of Actions



{a, b, c, d}

{a, b, c, d}

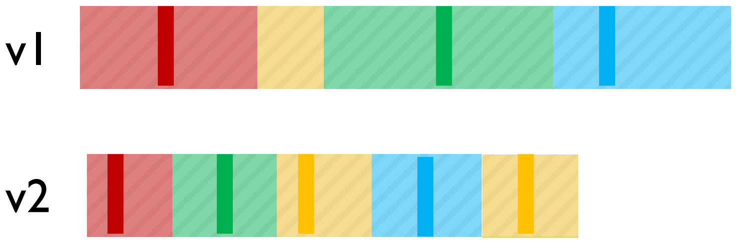
A set of actions, w/out temporal boundaries, order.
This type of labelling can arise in the form of meta-tags e.g., from video sharing platforms

Richard et al. CVPR'18
Fayyaz et al. CVPR'20
Li et al. CVPR'20
Li et al. CVPR'21

Weak Supervision

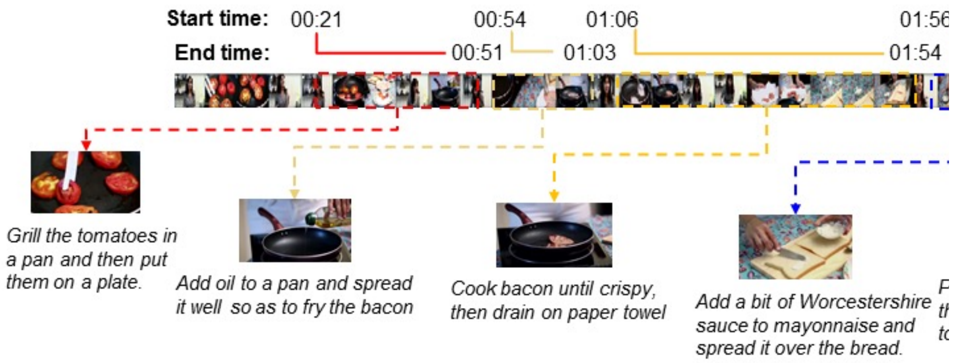
NUS National University of Singapore | Computing

Single-Frame Supervision



- Single timestamps for each / some actions to reduce annotation effort
- Provides more info than action transcripts or sets, stronger performance
  - Monotonic class probability - Li et al. CVPR’21
  - Expectation-Maximization (EM) – Rahaman et al. ECCV’22

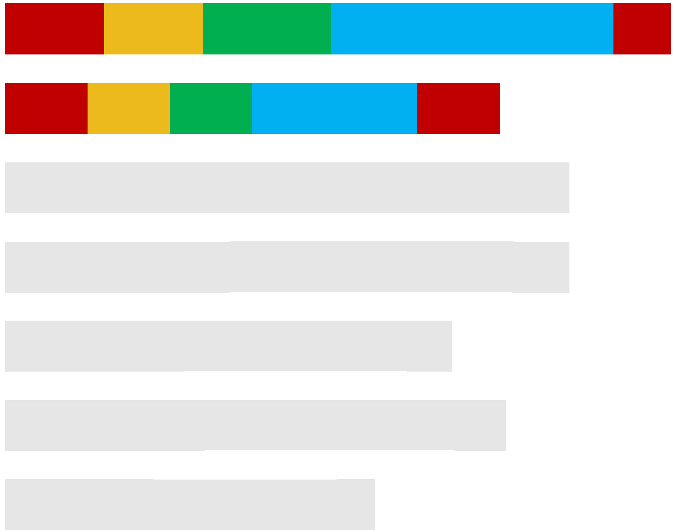
Narrations & Subtitles



- Text data from scripts, subtitles, or narrations
- Assumes temporal alignment of videos & text

Semi-Supervised

Fully labelled video for some videos of training set.



A subset of dense annotations provides more information than single-frame supervision on the entire dataset [3]

		Breakfast	50Salads	GTEA
Full-supervision	MSTCN'19 [1]	65.1	78.2	76.6
Single timestamp	Timestamp'21 [2]	64.1	75.6	66.4
Semi-supervised	SemiTAS'22 (%50) [3]	63.9	78.8	77.9

With 40% labelled data, ICC[4] performs comparably to the fully-supervised counterparts

		Breakfast	50Salads	GTEA
Full-supervision	MSTCN'19 [1]	65.1	78.2	76.6
Semi-supervised	ICC'22 (%40) [4]	71.1	78.0	78.4

[1] Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. CVPR'19
[2] Li, Z., Abu Farha, Y., Gall, J.: Temporal action segmentation from timestamp supervision. CVPR'21
[3] G. Ding and A. Yao, Leveraging action affinity and continuity for semi-supervised temporal action segmentation. ECCV'22
[4] D. Singhania, R. Rahaman, and A. Yao, Iterative contrast-classify for semi-supervised temporal action segmentation. AAAI'22

Conclusion & Outlook

- Problem formulation
  - frame-wise action labels to event-based processing?
- End-to-end learning
  - How to design? Do we even need it?
  - Integrated sequence modelling?
- New architectures
  - Transformers, graphs, etc.
- Shifting towards online video streams
  - Rethinking
- Minimizing labelling efforts

Resources

SCAN ME !

- Survey paper:

Temporal Action Segmentation: An Analysis of Modern Technique,

Guodong Ding, Fadime Sener and Angela Yao. 2022

- Work Compilation:

Awesome Temporal Action Segmentation,

Curated list of TAS works on GitHub.