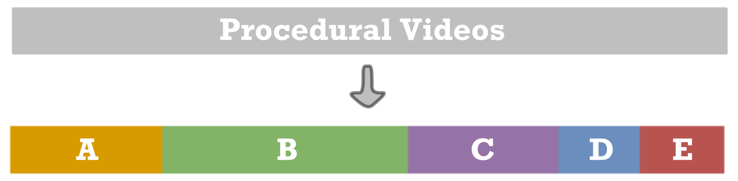
Leveraging Action Affinity and Continuity for Semi-supervised Temporal Action Segmentation

Guodong Ding and Angela Yao
National University of Singapore

Motivation

The Task

Temporally Segment procedural videos and **assign frame-wise** semantic **labels**.



Why Semi?

Frame-wise annotation is time-consuming
- Number of videos (hundreds if not thousands)
- Temporal span of videos (minutes long)

Semi-supervised only requires
- A small portion of annotated videos (as limited as 3)
- A large collection of videos unlabelled (cost free)

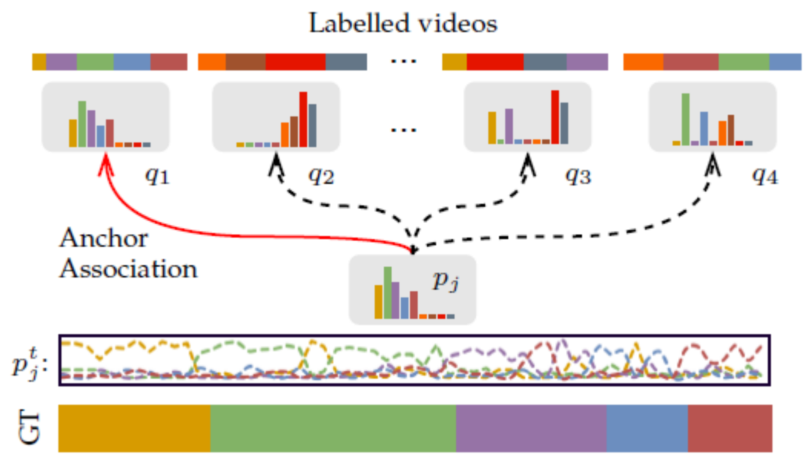
Challenges

- What **action compositions** are likely to occur?
- What is a reasonable **temporal proportion** for each action to take?
- What kind of **temporal constraints** should the action labels follow?

Approach

Action Affinity

Impose action prior induced from labelled videos to supervise unlabelled samples.



Action Prior (labelled):

$$q_i(k) = \frac{1}{T_i}\sum_{t}^{T_i} \mathbb{1}(y_i^t == k);$$

Action frequency (unlabelled):

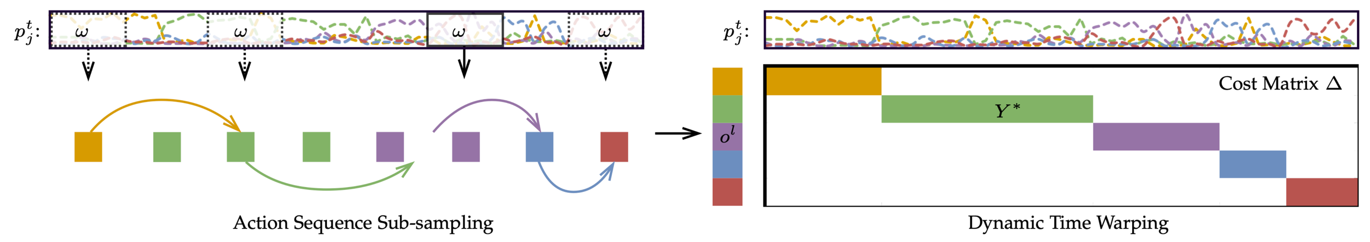
$$p_j(k) = \frac{1}{T_j}\sum_{t}^{T_j} p_j^t(k);$$

Affinity loss :

$$\boxed{\mathcal{L}_{\text{aff}} = \min_i \sum_k q_i(k) \log\left(\frac{q_i(k)}{p_j(k)}\right)}$$

Action Continuity

Mitigate the fragmentation problem of actions in network predictions.



1. Sub-sample actions in time

$$o = \arg\max_k \frac{1}{\omega} \sum_{t=t'}^{t'+\omega} p^t(k),$$

2. Remove adjacent repetitive actions

3. Using the KL-Divergence for cost calculation

$$d(l,t) = KL(o^l \| p^t) = \sum_k o^l(k) \log\left(\frac{o^l(k)}{p^t(k)}\right).$$

4. Optimize the cost along the optimal path

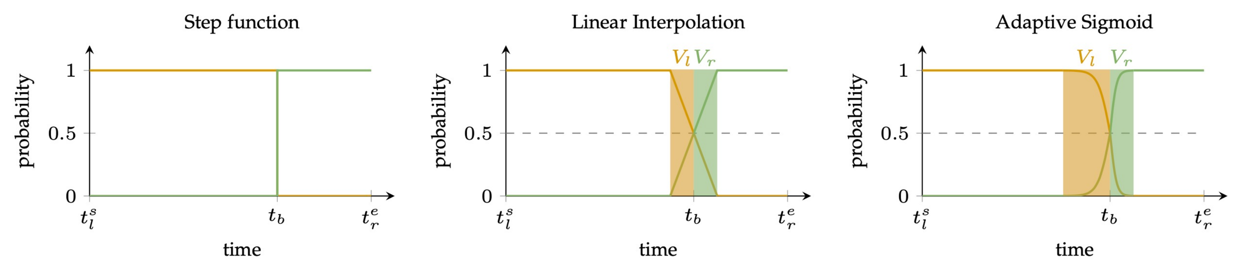
Continuity loss :

Identical to classification loss!

$$\boxed{\mathcal{L}_{\text{cont}} = \frac{1}{T}\min_Y \langle Y, \Delta\rangle = \frac{1}{T}\sum_t -\log(p^t(\tilde{y}^t)).}$$

Adaptive Boundary Smoothing

Build more robust boundaries with coarser transitional action probabilities.



(a) Standard One-hot Labels (b) Fixed-Duration Linear [18] (c) ABS (Ours)

Duration-aware Action Boundary Vicinity:

$$V_l = [t_b - (t_b - t_l^s) * v, t_b), \quad \text{and} \quad V_r = [t_b, t_b + (t_r^e - t_b) * v).$$

Action probability Assignment w/ Adaptive Sigmoid:

$$y^t(y_l) = \frac{1}{1 + e^{-\frac{\epsilon}{|V_l|}(t - t_b)}}, \quad \text{and} \quad y^t(y_r) = 1 - y^t(y_l)$$

Results

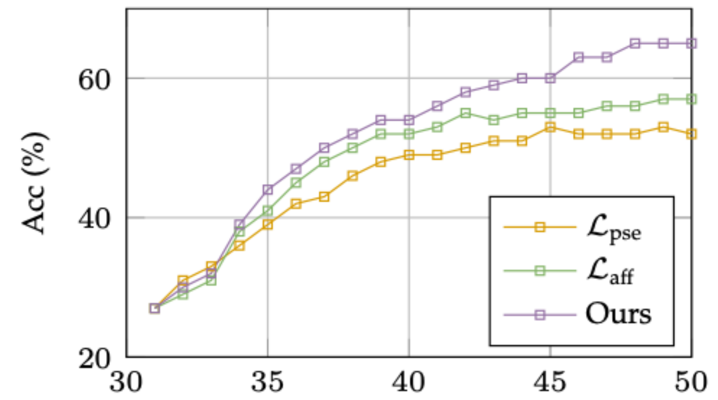
Ablation Study

$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{pse}}$	$\mathcal{L}_{\text{aff}}$	$\mathcal{L}_{\text{cont}}$	ABS	F1@10	F1@25	F1@50	Edit	Acc
✓					47.9	40.6	28.6	51.8	36.8
✓	✓				49.3	44.8	33.9	49.7	40.2
✓		✓			52.0	46.5	34.3	53.4	44.0
✓	✓	✓			54.1	46.7	34.9	54.1	47.8
✓		✓	✓		53.8	50.1	37.6	56.6	49.2
✓		✓	✓	✓	**56.9**	**51.3**	**39.0**	**57.7**	**49.5**

Action prior from affinity loss is effective:

- Stand-alone **outperforms Pseudo**
- Avoid overfitting to the incorrect pseudo labels when data annotation is rather limited.

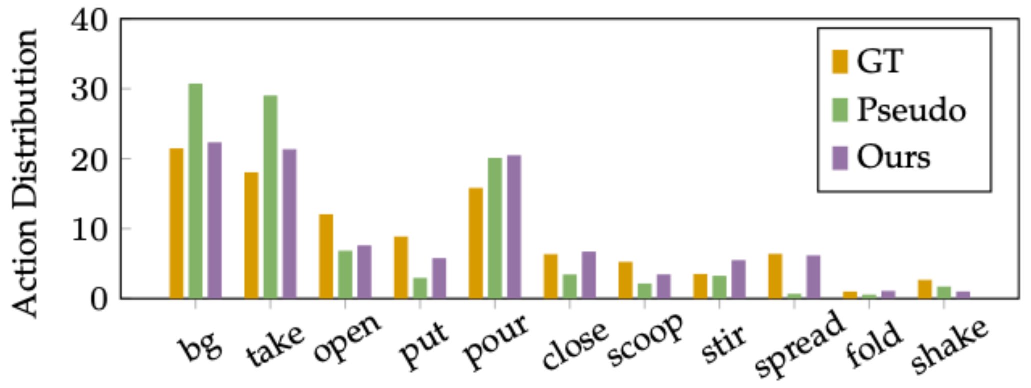
Pseudo-labels



Our proposed approach generates pseudo labels with **higher accuracy** for the unlabeled videos.

The gap in accuracy gets larger as training proceeds.

Action frequency



With action affinity loss, the pseudo labels for unlabelled data **better preserves the action prior** in the labelled data.

Performances

$\%D_L$	Method	Breakfast F1@10	F1@25	F1@50	Edit	Acc	50Salads F1@10	F1@25	F1@50	Edit	Acc	GTEA F1@10	F1@25	F1@50	Edit	Acc
5	Base	36.7	28.4	19.5	37.5	28.2	26.8	19.7	11.5	26.1	28.1	29.9	25.8	14.8	31.0	37.2
	Pseudo	40.2	28.5	20.1	41.3	20.9	22.6	17.0	12.1	22.0	24.0	48.4	42.3	30.2	45.4	48.1
	Ours	44.5	35.3	26.5	45.9	38.1	37.4	32.3	25.5	32.9	52.3	59.8	53.6	39.0	55.7	55.8
	Gain	**7.8**	**6.9**	**7.0**	**8.4**	**9.9**	**10.6**	**12.6**	**14.0**	**6.8**	**24.2**	**29.9**	**27.8**	**24.2**	**24.7**	**18.6**
10	Base	46.8	41.1	29.2	50.9	37.1	27.6	24.3	16.0	27.4	32.0	38.1	29.6	15.3	39.6	41.1
	Pseudo	49.3	44.8	33.9	49.7	40.2	36.2	32.4	24.5	33.5	41.1	65.5	60.7	45.8	59.9	57.9
	Ours	56.9	51.3	39.0	57.7	49.5	47.3	42.7	31.8	43.6	58.0	71.5	66.0	52.9	67.2	62.6
	Gain	**10.1**	**10.2**	**9.8**	**6.8**	**12.4**	**19.7**	**18.4**	**15.8**	**16.2**	**26.0**	**33.4**	**36.4**	**37.6**	**27.6**	**21.5**

Our approach

- outperforms pseudo-labelling by a large margin
- proves effectiveness with different ratios of labelled data
- overcomes **the catastrophic confirmation bias** from pseudo-labelling with very limited labels

Takeaways

Loss Functions

Two novel loss functions are proposed specifically **for the semi-supervised learning** of temporal action segmentation task.

Dense Supervision

The **densely labelled videos** do more than providing frame-wise semantic action labels, when put together at a video level, they **serve as action priors** for a specific category of procedural task.

Boundary Ambiguity

The **action boundary** itself **and** the human **annotations are ambiguous** in pinpointing exact transiting timestamps. Transitional action boundaries can be helpful.