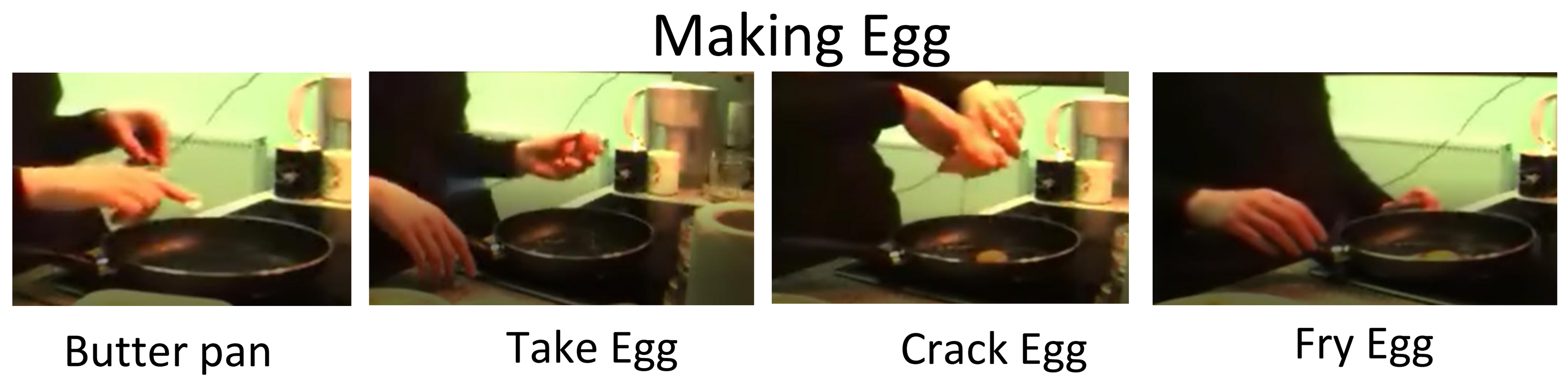
Iterative Contrast-Classify For Semi-supervised Temporal Action Segmentation

Dipika Singhania, Rahul Rahaman, Angela Yao
National University of Singapore

What is Temporal Action Segmentation(TAS)?

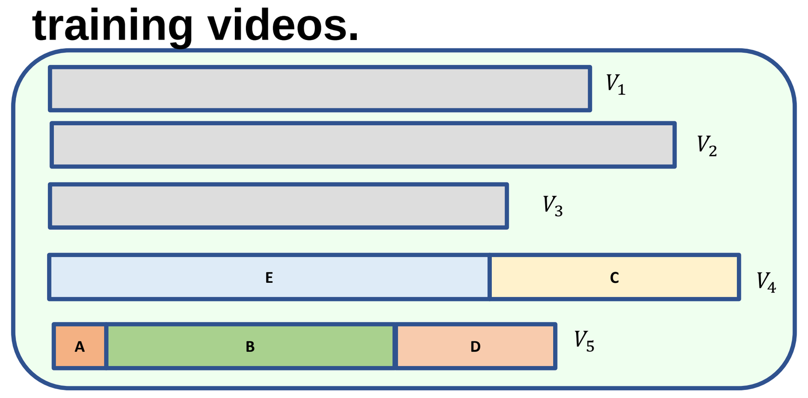
- **Input:** Takes long untrimmed video containing multiple actions in a sequence.
- **Output:** Estimates the action labels for every frame in the video. In other words, for every action segment, predicts the label and its start and end time.



Why Semi-Supervised TAS?

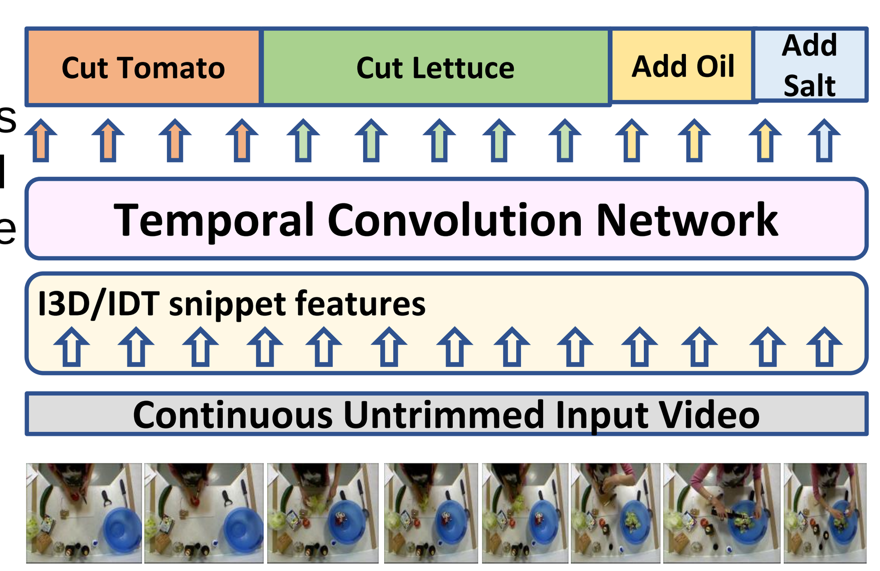
- Annotating framewise action labels for all training videos is costly as videos are long and vary in segments content, ordering and length.
- We **propose the first** semi-supervised method for TAS.
- **Semi-Supervised:** Labels only for a fraction of the videos in training set.

Semi-Supervised: Labels for few training videos.



Temporal Convolution Networks(TCNs)

- Temporal Convolution Networks (TCNs), e.g., MS-TCN++[1], ED-TCN[2], C2F-TCN[3] are base models for TAS taking extracted snippet level (IDT/I3D) features as input to produce framewise action labels in the video.
- We use representations from **Encoder-Decoder TCNs like ED-TCN[2] and C2F-TCN[3]** for our unsupervised learning framework**.**

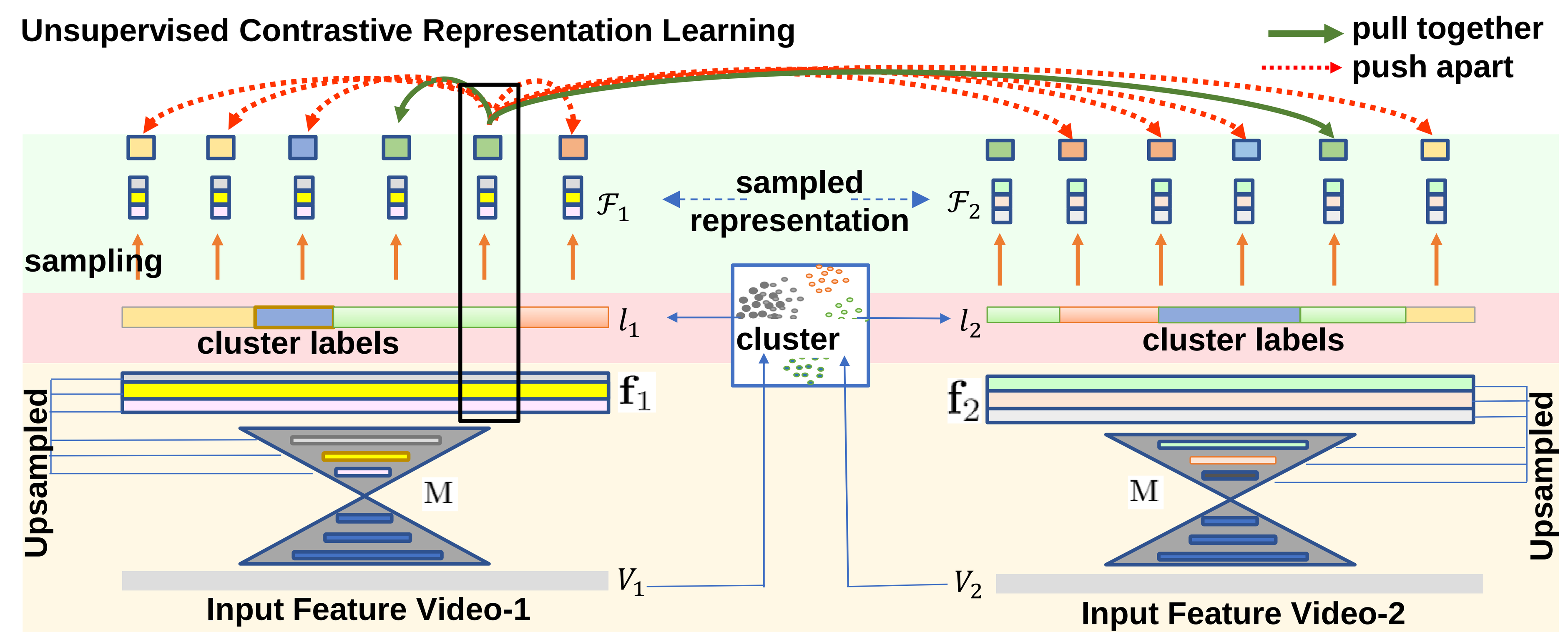


Our Semi-Supervised Learning Framework

Our overall semi-supervised learning framework has two stages.

- First, we apply an unsupervised frame-wise contrastive representation learning to learn a model M.
- Subsequently, model M is trained with linear projection layers G (action classifiers) with a small portion of the labelled training data to produce the semi-supervised model (M : G)

Frame-Wise Unsupervised Representation Learning



Step 1 (bottom yellow panel): Pass pre-trained I3D inputs V into the base TCN and generate **multi-resolution representation** f.

Step 2 (middle pink panel): Cluster the I3D inputs V within a training mini-batch and generate frame-wise cluster labels l.

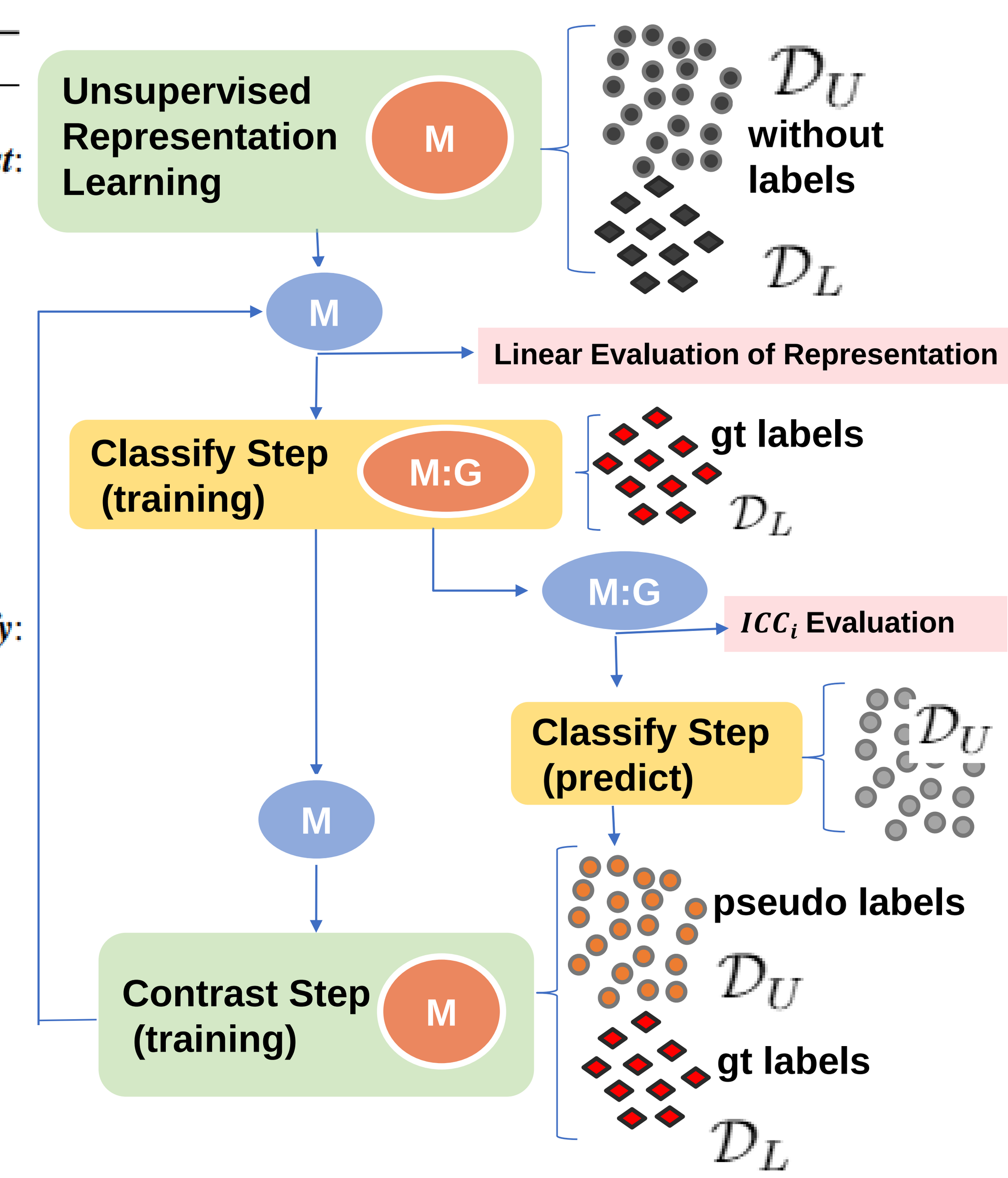
Step 3 (top green panel): Representations f and its corresponding cluster label l is sampled based on **temporal proximity sampling strategy** to form feature set F.

Step 4 Apply contrastive learning to **"pull together" (green arrows)** similar samples in the positive set and **"push apart" (red arrows)** other samples in the negative set.

Iterative-Contrast-Classify(ICC) Semi-Supervised

Algorithm 1: Iterative *Contrast-Classify* algorithm

while *iter* $\leq$ MaxIter **do**
  for *epoch* $\leq$ MaxEpoch **do** $\triangleright$ ***Contrast***:
    sample minibatch $\{V_n\}_{n=1}^N \subset \mathcal{D}_u \cup \mathcal{D}_l$
    if *iter* $= 0$ **then**
      $\{l_n\}_{n=1}^N \leftarrow Cluster(\{V_n\}_{n=1}^N)$
    end if
    $\mathcal{L}_{\text{con}} \leftarrow Contrastive(\{(V_n, l_n)\}_{n=1}^N, \mathbf{M})$
    minimize $\mathcal{L}_{\text{con}}$ and update $\mathbf{M}$
  end for
  for $V_n \in \mathcal{D}_l$ **do**
    $l_n \leftarrow y_n$
  end for
  for *epoch* $\leq$ MaxEpoch **do** $\triangleright$ ***Classify***:
    sample minibatch $\{V_n\}_{n=1}^N \subset \mathcal{D}_l$
    $\mathcal{L}_{\text{con}} \leftarrow Contrastive(\{(V_n, l_n)\}_{n=1}^N, \mathbf{M})$
    $\{\bar{p}_n\}_{n=1}^N \leftarrow Predict(\{V_n\}_{n=1}^N, \mathbf{G})$
    $\mathcal{L}_{\text{ce}} \leftarrow CrossEntropy(\{(\bar{p}_n, y_n)\}_{n=1}^N)$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{ce}}$
    minimize $\mathcal{L}$ and update $\mathbf{M}$ and $\mathbf{G}$
  end for
  for $V_n \in \mathcal{D}_u$ **do**
    $l_n \leftarrow Predict(V_n, \mathbf{G})$
  end for
  iter $\leftarrow$ *iter* $+1$
end while



Performance Unsupervised Learning

	Breakfast					50Salads				
	$F1@\{10,25,50\}$			Edit	MF	$F1@\{10,25,50\}$			Edit	MF
Input I3D Baseline	4.9	2.5	0.9	5.3	30.2	12.2	7.9	4.0	8.4	55.0
Our Representations	**57.0**	**51.7**	**39.1**	**51.3**	**70.5**	**40.8**	**36.2**	**28.1**	**32.4**	**62.5**
Improvement	52.1	49.2	38.2	46.0	40.3	28.6	28.3	24.1	24.0	7.5

Unsupervised representations outperform input I3D features in TAS scores.

Cluster	11.7	8.0	3.9	12.2	36.1	18.5	13.7	8.5	13.6	50.8
(+) Proximity	24.4	19.2	11.5	21.3	50.0	18.6	13.5	8.0	13.5	51.6
(+) Video-Level	42.9	37.6	26.6	36.4	66.1	–	–	–	–	–

Contributions from "Cluster" labels, "Proximity" and Video Level Loss.

Last-Layer($\mathbf{z}_6$)	42.9	37.6	26.6	36.4	66.1	18.6	13.5	8.0	13.5	51.6
Multi-Resolution($\mathbf{f}$)	**57.0**	**51.7**	**39.1**	**51.3**	**70.5**	**40.8**	**36.2**	**28.1**	**32.4**	**62.5**
Improvement	14.1	14.1	12.5	14.9	4.4	22.2	22.7	20.1	18.9	10.9

Multi-Resolution Features significantly contrastive representation learning.

Performance ICC Semi-Supervised

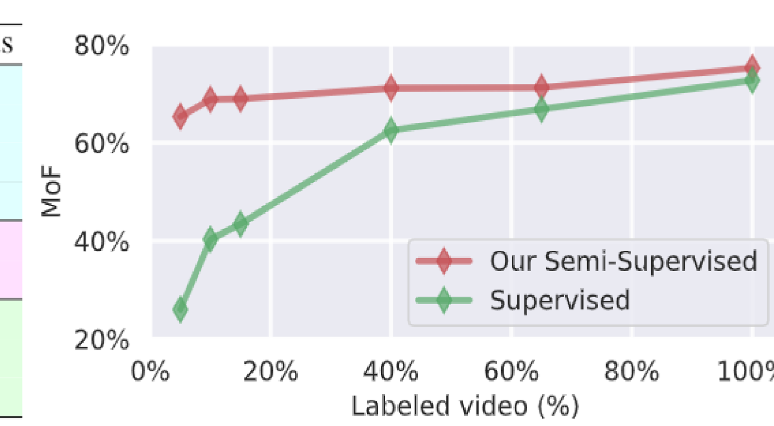
		Breakfast					50Salads				
$\%D_L$	**Method**	$F1@\{10,25,50\}$			Edit	MoF	$F1@\{10,25,50\}$			Edit	MoF
≈5	ICC$_1$	54.5	48.7	33.3	54.6	64.2	41.3	37.2	27.8	35.4	57.3
	ICC$_2$	56.9	51.9	34.8	56.5	65.4	45.7	40.9	30.7	40.9	59.5
	ICC$_3$	59.9	53.3	35.5	56.3	64.2	50.1	46.7	35.3	43.7	60.9
	ICC$_4$	60.2	53.5	35.6	56.6	65.3	52.9	49.0	36.6	45.6	61.3
	Gain	5.7	4.8	2.3	2.0	1.1	11.6	11.8	8.8	10.2	4.0

Progressive semi-supervised improvements with more iterations of ICC.

≈5	Supervised	15.7	11.8	5.9	19.8	26.0	30.5	25.4	17.3	26.3	43.1
	Semi-Super	**60.2**	**53.5**	**35.6**	**56.6**	**65.3**	**52.9**	**49.0**	**36.6**	**45.6**	**61.3**
	Gain	44.5	41.7	29.7	36.8	39.3	22.4	23.6	19.3	19.3	18.2
≈10	Supervised	35.1	30.6	19.5	36.3	40.3	45.1	38.3	26.4	38.2	54.8
	Semi-Super	**64.6**	**59.0**	**42.2**	**61.9**	**68.8**	**67.3**	**64.9**	**49.2**	**56.9**	**68.6**
	Gain	29.5	28.4	22.7	25.6	28.5	22.2	26.6	22.8	18.7	13.8

ICC improvement over supervised counterpart using same labelled data.

	Method	Breakfast	50salads
Full	MSTCN'20	67.6	83.7
	SSTDA'20	70.2	83.2
	*C2F-TCN'21	73.4	79.4
	Ours ICC (100%)	**75.2**	**85.0**
Weakly	SSTDA(65%)	65.8	80.7
	TSS'21	64.1	75.6
Semi	Ours ICC (40%)	71.1	78.0
	Ours ICC (10%)	68.8	68.6
	Ours ICC (5%)	65.3	61.3



Our ICC has impressive performance with just 5% labelled videos; at 40%, we almost match the Mean over Frames (MoF) of a 100% fully-supervised setup. ICC also improves fully-supervised scores.

References

[1] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. 2020. MS-TCN++: Multi-stage temporal convolutional network for action segmentation.

[2] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager.2017. Temporal convolutional networks for action segmentation and detection.

[3] Singhania, D.; Rahaman, R.; and Yao, A. 2021. Coarse to Fine Multi-Resolution Temporal Convolutional Network.

Our GitHub project page