Iterative Contrast-Classify For Semi-supervised Temporal Action Segmentation

Dipika Singhania
Rahul Rahaman
Angela Yao

Temporal Action Segmentation

- **Input:** Takes untrimmed video containing multiple actions in a sequence.
  These videos are longer in time and last up to 10 minutes long.

Making Egg

(pictures taken from Breakfast Actions Dataset, CVPR 2014)

Temporal Action Segmentation

- **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, estimate the label, start and end time for every action in the video.

Making Egg

Butter pan | Take Egg | Crack Egg | Fry Egg

(pictures taken from Breakfast Actions Dataset, CVPR 2014)

Use of Temporal Action Segmentation

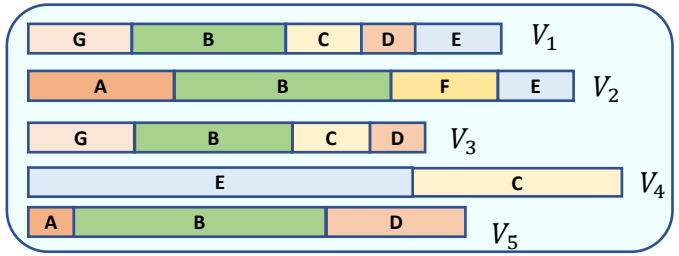
Automatic video temporal segmentation can help to automatically interpret and summarize videos over time. Untrimmed videos helps in understanding relationship across time.

Picture of COIN dataset : https://coin-dataset.github.io/

- Unprecedented amount of data from multiple domains
  - Nursing and caring
  - Household daily routines
  - Science and craft
  - Planting
  - Recipe preparation

Existing Popular Supervision Types

Full Supervision[1, 2, 3] : Requires action labels for all frames of all training videos.



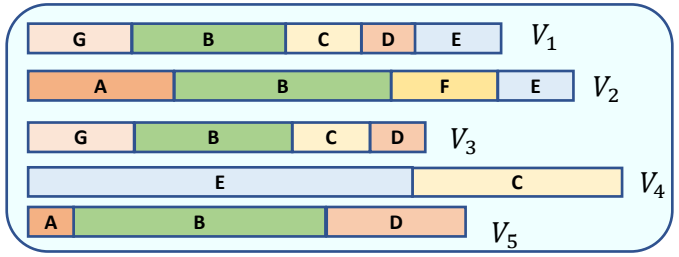
Labelling every frame is computationally expensive task, as per video can contain up to 10K frames.

[1] MSTCN++, Li et al., 2020
[2] ED-TCN, Lea et al., 2017
[3] C2F-TCN, Singhania et. al. 2021

Different colors denote different actions in the video. Grey color denotes unlabeled video parts.

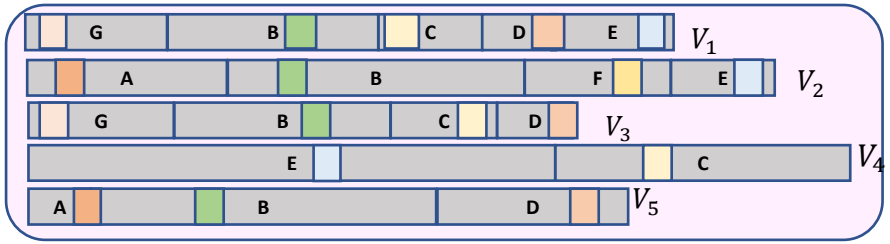
Existing Popular Supervision Types

Full Supervision: Requires action labels for all frames in all training videos.



Weak Supervision: Some label for every action in all training videos

Transcripts [4,5], Single [6] or Few [7] frames



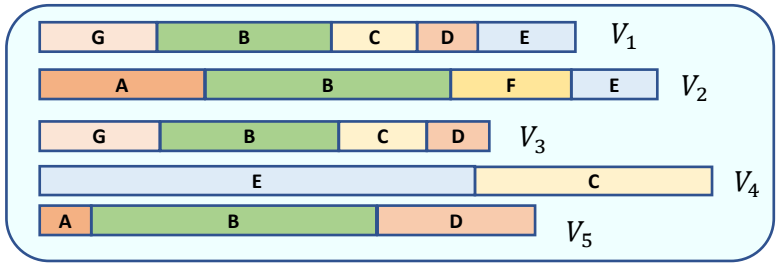
Weak Supervision still has High Annotation Efforts:
Annotators still require to watch all training videos to avoid missing labels of any action label segment.

Different colors denote different actions in the video. Grey color denotes unlabeled video parts.

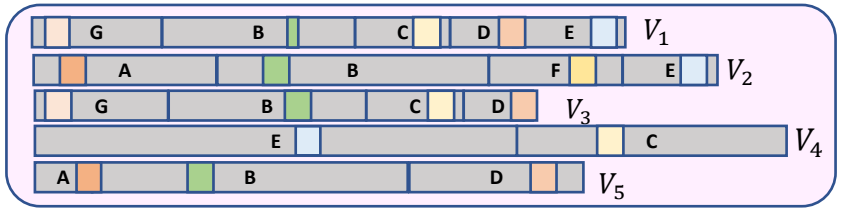
[4] MuCon, Souri et al., 2019
[5] CDFL, Li et al., 2019
[6] TSS, Li et al., 2021
[7] SSTDA, Chen et al., 2020

Our Proposal: Semi-Supervised

Full: All labels for all training videos

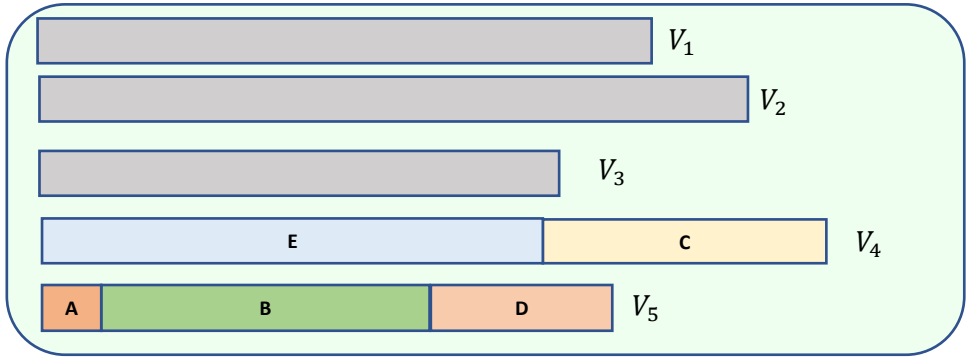


Weakly: Some label for every action in all training videos



Semi-Supervised: Having labels only for a fraction of the videos in the training set.

All labels for few training videos



Ours is first to show results for semi-supervised temporal action segmentation.

Different colors denote different actions in the video. Grey color denotes unlabeled video parts.

Proposed Semi-Supervised Setup

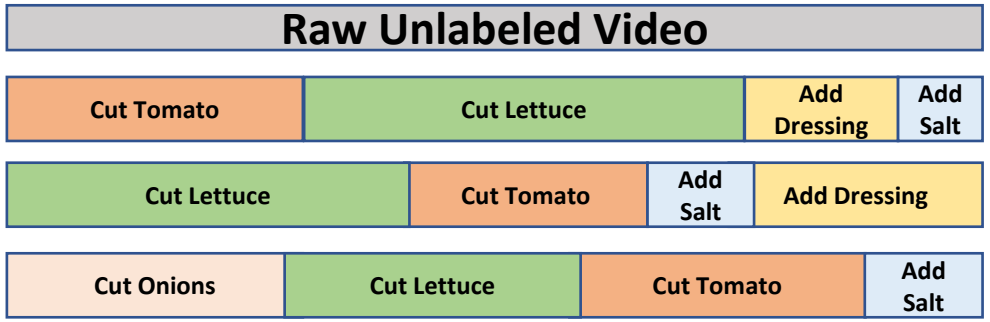
- Semi-Supervised = Unsupervised Representation + Very Few Labeled Videos Conventional Supervised

Proposed Semi-Supervised Setup

- Semi-Supervised = Unsupervised Representation + Very Few Labeled Videos Conventional Supervised

Challenges in designing unsupervised representation task for temporal action segmentation

- Untrimmed unlabeled videos vary in content and timing of action segments.
  - Varying order of actions
  - Varying length of actions
  - Missing actions

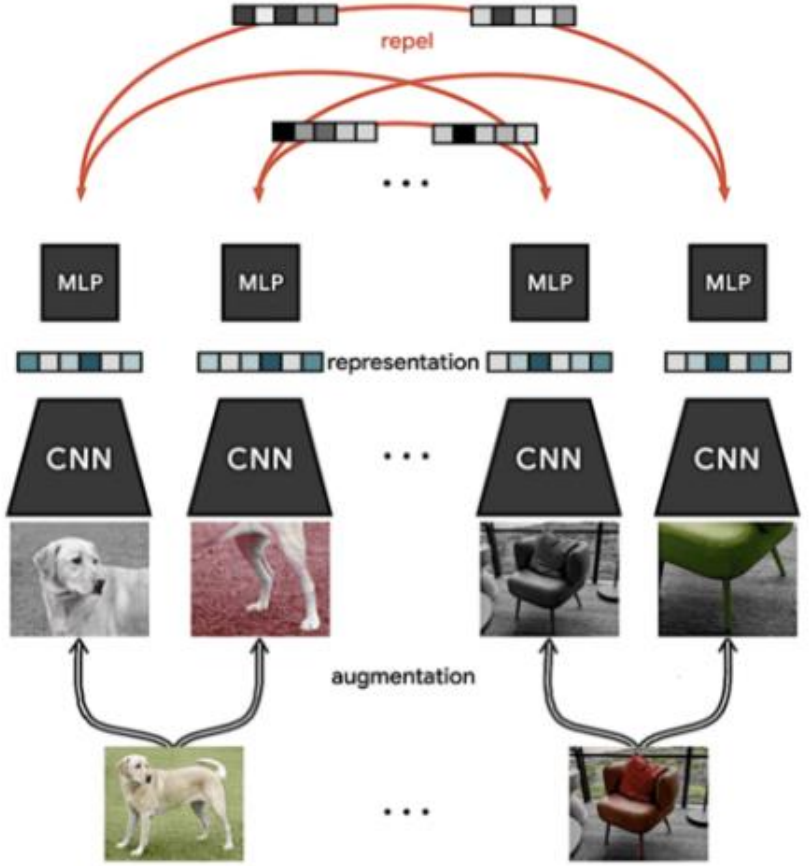


Unsupervised Representation Learning

- We are inspired by the success of **SimCLR Contrastive Framework in bringing high** semi-supervised classification (or recognition) scores in fields of both images[8] (or videos [9])**.**

- For classification (or recognition) task, standard **SimCLR** technique is to bring image [8] (or video[9]) and its augmentations near to aid in learning to classify (or recognize) the images (or video).

- However, Temporal Action Segmentation is a **frame-wise (and not video-wise)** classification task, so directly making video and its augmentation near would not likely help in bringing inter-video segmentation.

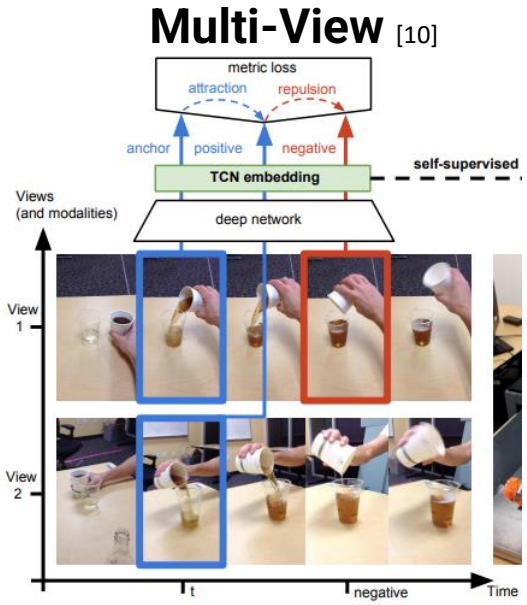


SimCLR Framework [8]

[8] SimCLR, Chen et al., 2020
[9] TCL, Singh et al.,2021

Framewise SimCLR for Longer Videos

Multi-View: Pulling together different views temporally aligned frames.



[10] Sermanet et al. 2018
[13] Kuehne et al. 2014
[14] Stein et al. 2013
[15] Fathi et al., 2011

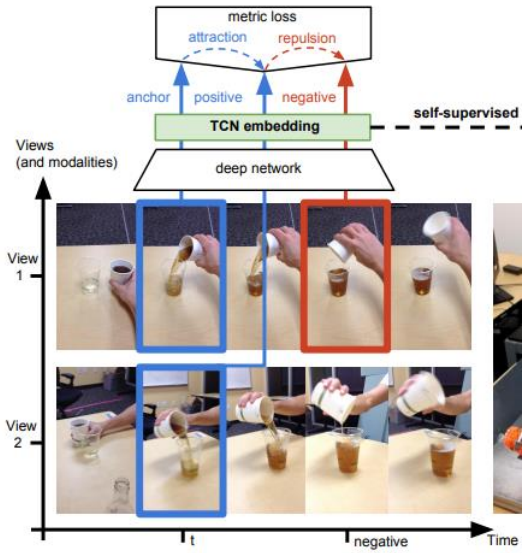
Above properties does not hold for temporal segmentation dataset like Breakfast [13], 50salads [14], GTEA [15] etc.

Previous SimCLR for Longer Videos

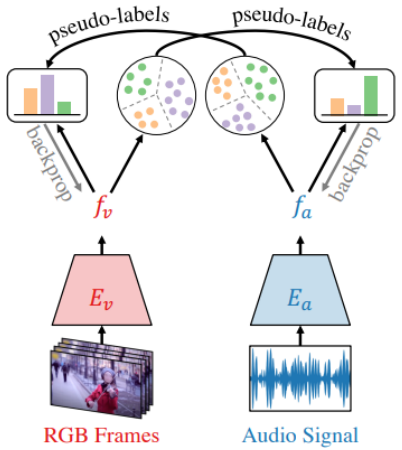
Multi-View: Pulling together different views temporally aligned frames.

Multi-Modal: Pulling together different Modalities **(Video-Audio) or (Video-Text)** together.

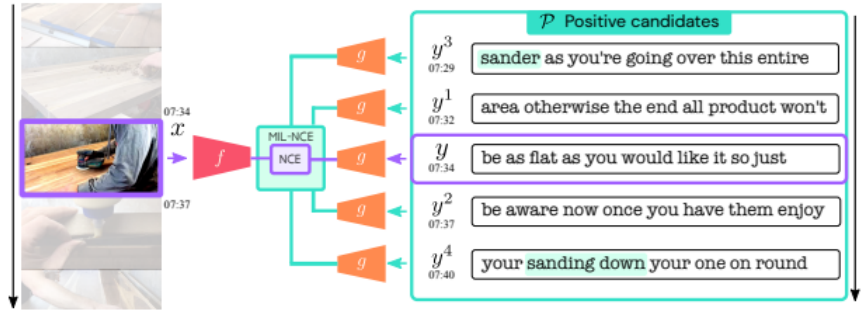
Multi-View



Video-Audio [11]



Video-Text [12]

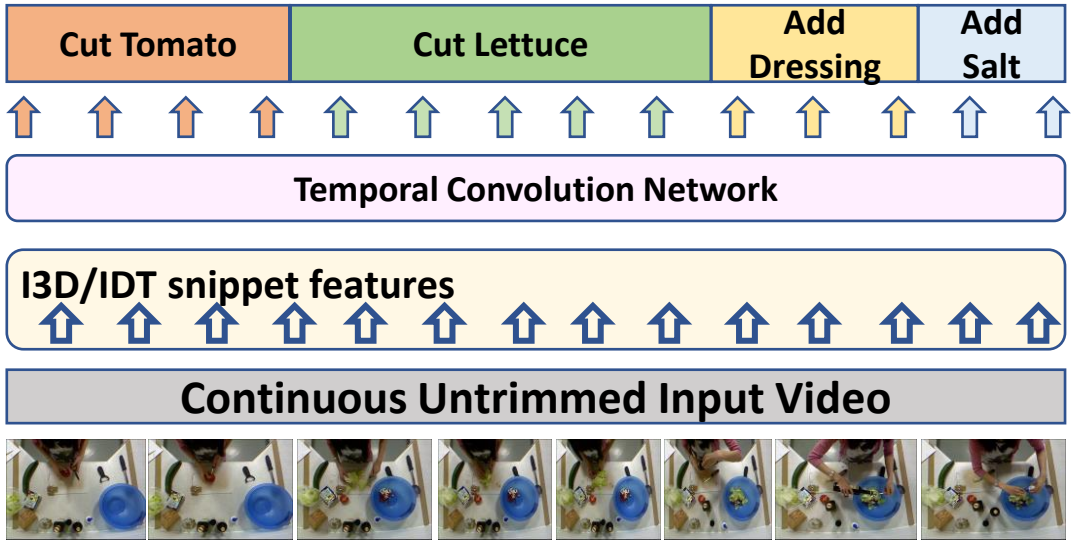


[11] Alwassel et al. 2019
[12] Miech et al. 2020

Above properties does not hold for temporal segmentation dataset like Breakfast, 50salads, GTEA etc.

Overview Temporal Action Segmentation Steps

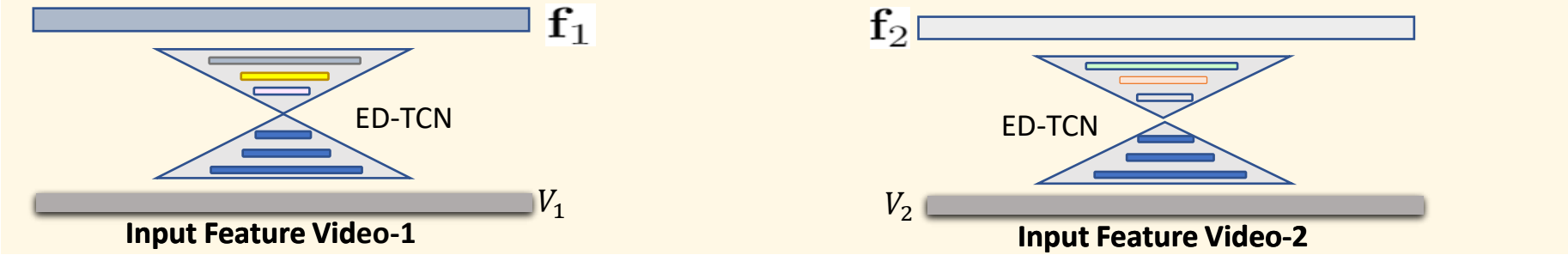
- Temporally segmenting actions in long video requires both local motion and global long-range dependencies information.
- It is standard to extract snippet level (IDT/I3D) features to capture local temporal motion and to reduce computational expense of joint end to-end training i.e., 3D video clips to 1D features as input.
- Exacted features are passed through **Temporal Convolution Networks (TCNs)** [1, 2, 3] to produce framewise action labels for all frames in the video.
- TCNs helps captures global action compositions and long-range dependencies.
- **Representations from TCNs** in Supervised Setup is trained with frame-wise cross-entropy loss.



[1] MSTCN++, Li et al., 2020
[2] ED-TCN, Lea et al., 2017
[3] C2F-TCN, Singhania et. al. 2021

Representation from TCN

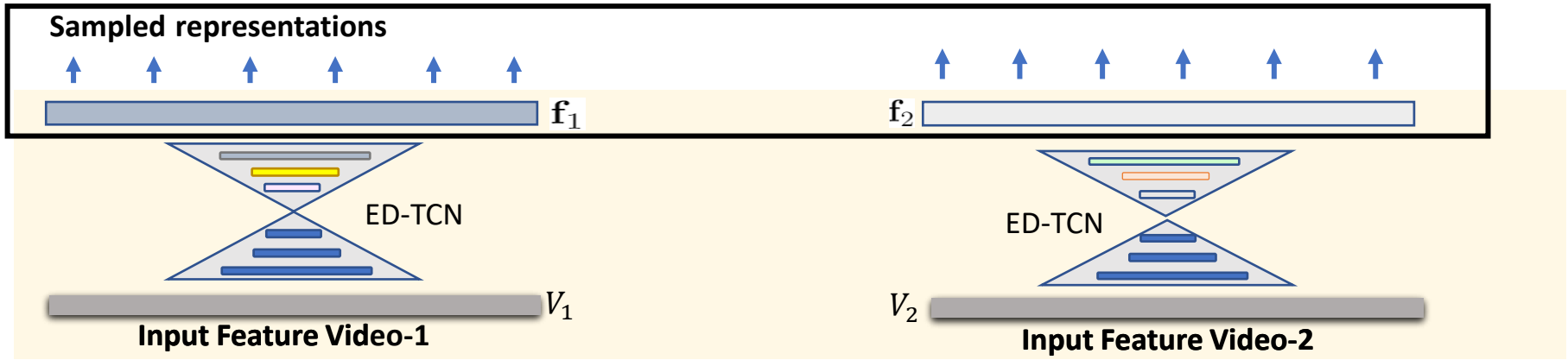
- In unsupervised learning we train the frame-wise representation from TCNs with Contrastive Framewise loss.
- We use Encoder-Decoder like TCNs, ED-TCN[2] and C2F-TCN [3] as our base TCNs.



[2] ED-TCN, Lea et al., 2017
[3] C2F-TCN, Singhania et. al. 2021

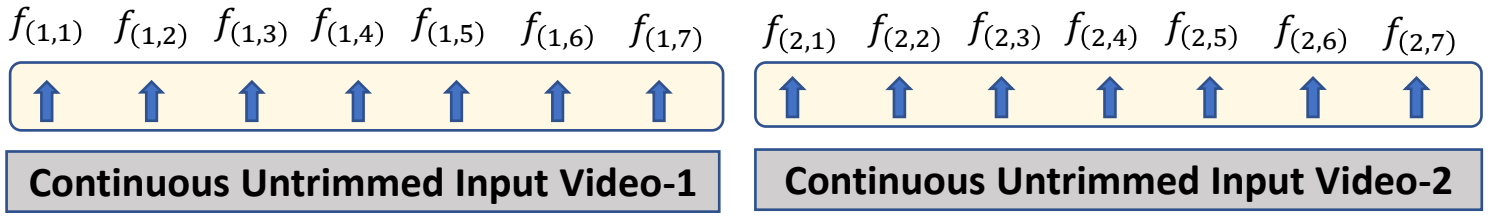
Sampling representations

- The videos for action segmentation are long, i.e., 1-18k frames and thus representations from TCNs.
- Considering all representations of every video in a batch in unsupervised contrastive learning is too computationally expensive to consider.
- Therefore, we sample fixed number of representation from each video in a batch for contrastive learning.



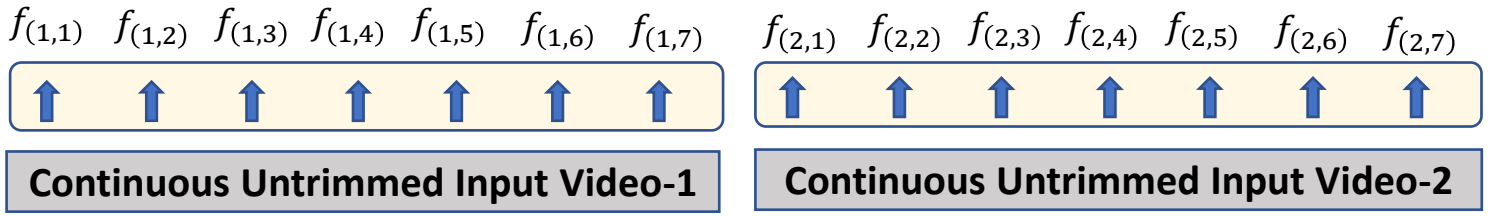
Frame-Level Contrastive Formulation

A set of representation
$F \coloneqq \{f_{(n,i)}, (n,i) \in I\}$
indexed by set I, n is video-id, i is representation id within video n.

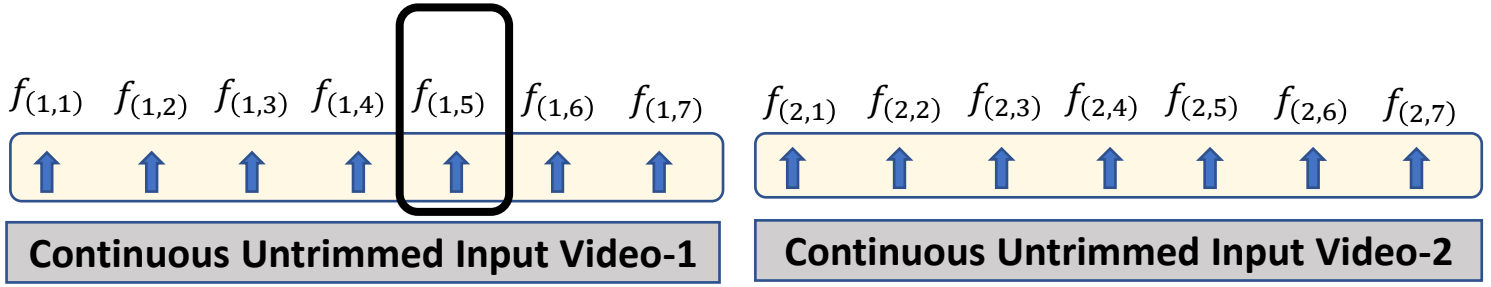


Frame-Level Contrastive Formulation

A set of representation
$F \coloneqq \{f_{(n,i)}, (n,i) \in I\}$
indexed by set I, n is video-id, i is representation id within video n.

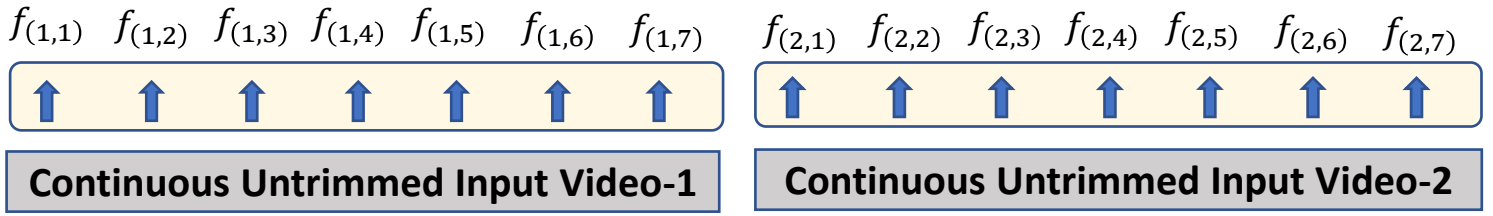


Each representation $f_{(n,i)} \in F$ of index (n,i) needs to be associated with two disjoint sets of indices:



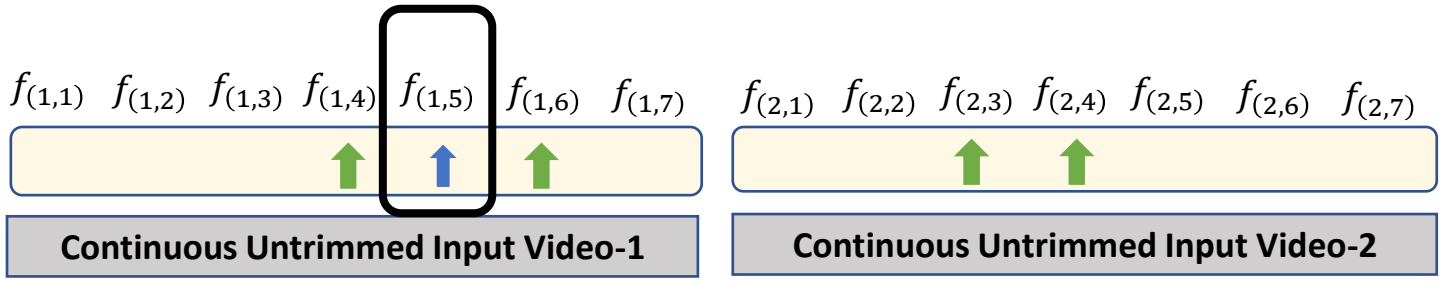
Frame-Level Contrastive Formulation

A set of representation
$F \coloneqq \{f_{(n,i)}, (n,i) \in I\}$
indexed by set I, n is video-id, i is representation id within video n.



Each representation $f_{(n,i)} \in F$ of index (n,i) is associated with two disjoint sets of indices:

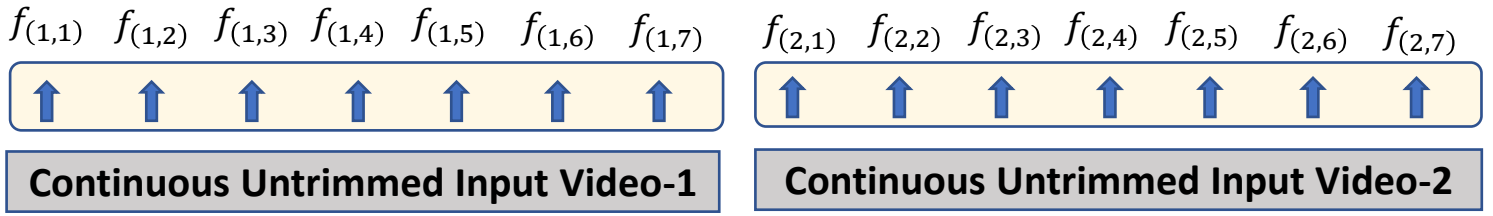
Positive Set Indices: $P_{(n,i)} \subset I \setminus \{(n,i)\}$.



Positive Set containing indices of representation which must be **pull together** with Contrastive Loss.

Frame-Level Contrastive Formulation

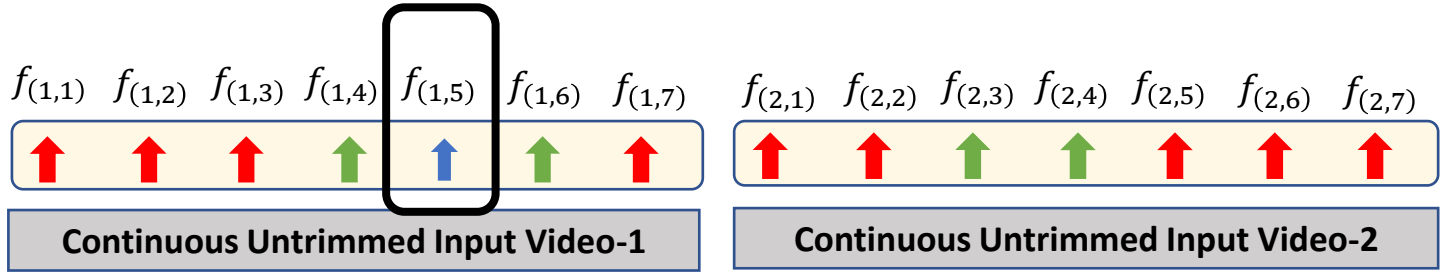
A set of representation
$F \coloneqq \{f_{(n,i)}, (n,i) \in I\}$
indexed by set I, n is video-id, i is representation id within video n.



Each representation $f_{(n,i)} \in F$ of index (n,i) is associated with two disjoint sets of indices:

Positive Set Indices: $P_{(n,i)} \subset I \setminus \{(n,i)\}$.

Negative Set Indices: $N_{(n,i)} \subset I \setminus \{(n,i)\}$.



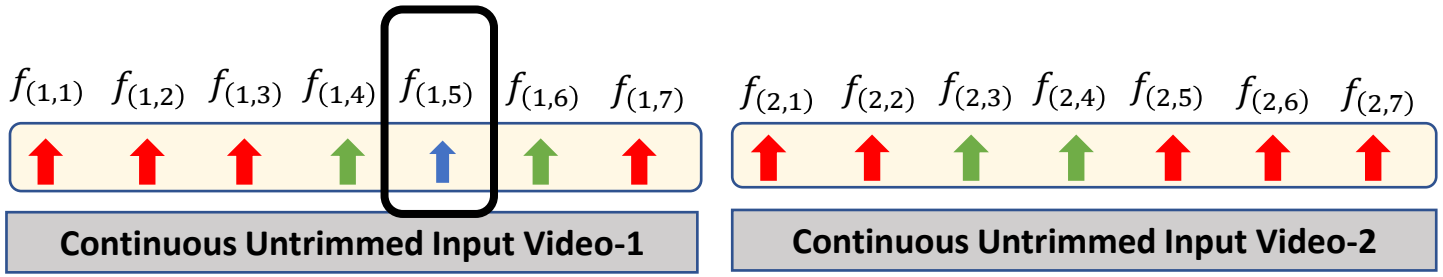
Negative Set containing indices of representation which must be **push apart** with Contrastive Loss.

Frame-Level Contrastive Formulation

Each representation $f_{(n,i)} \in F$ of index (n,i) is associated with two disjoint sets of indices:

Positive Set Indices: $P_{(n,i)} \subset I \setminus \{(n,i)\}$.

Negative Set Indices: $N_{(n,i)} \subset I \setminus \{(n,i)\}$.



Considering $f_{(1,5)}$ as our sampled representation and $(1,6) \in P_{(1,5)}$, **Contrastive Learning** maximizes the probability such that $f_{(1,6)}, f_{(1,5)}$ is more similar than any feature $(r,k) \in N_{(1,5)}$.
e_τ is the τ scaled exponential cosine similarity.

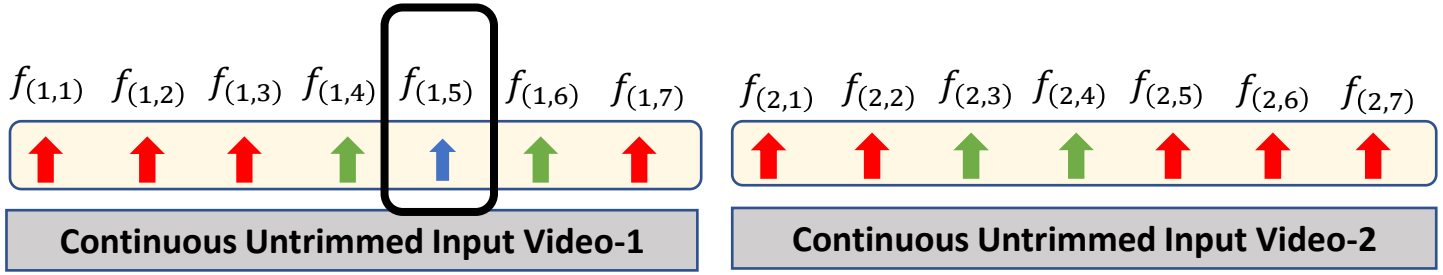
$$p = \frac{e_\tau(\uparrow,\uparrow)}{e_\tau(\uparrow,\uparrow) + \Sigma_{\uparrow} e_\tau(\uparrow,\uparrow)}$$

Frame-Level Contrastive Formulation

Each representation $f_{(n,i)} \in F$ of index (n, i) is associated with two disjoint sets of indices:

Positive Set Indices: $P_{(n,i)} \subset I \setminus \{(n, i)\}$.

Negative Set Indices: $N_{(n,i)} \subset I \setminus \{(n, i)\}$.



For each $(m, j) \in P_{(n,i)}$, **Contrastive Learning** maximizes the probability ensures that $f_{(m,j)}, f_{(n,i)}$ is more similar than any feature $(r, k) \in N_{(n,i)}$. e_τ is the τ scaled exponential cosine similarity.

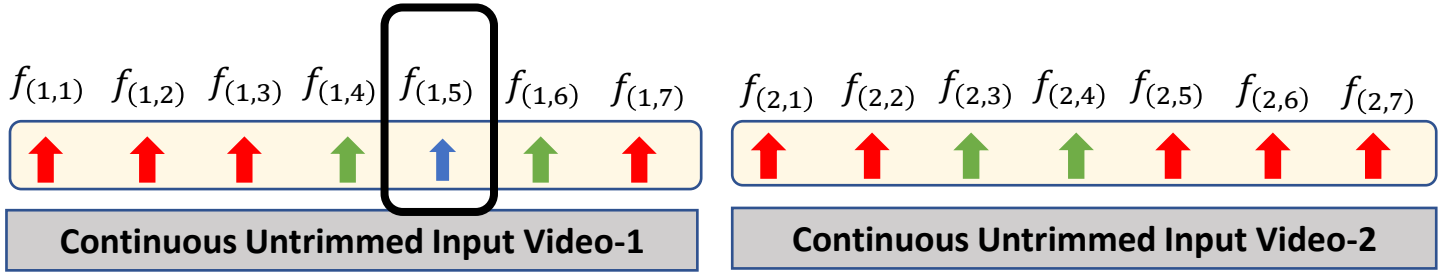
$$p_{nm}^{ij} = \frac{e_\tau(f_{(m,j)}, f_{(n,i)})}{e_\tau(f_{(m,j)}, f_{(n,i)}) + \Sigma_{(r,k)\in N_{(n,i)}} e_\tau(f_{(r,k)}, f_{(n,i)})}$$

Frame-Level Contrastive Formulation

Each feature $f_{(n,i)} \in F$ of index (n, i) is associated with two disjoint sets of indices:

Positive Set Indices: $P_{(n,i)} \subset I \setminus \{(n, i)\}$.

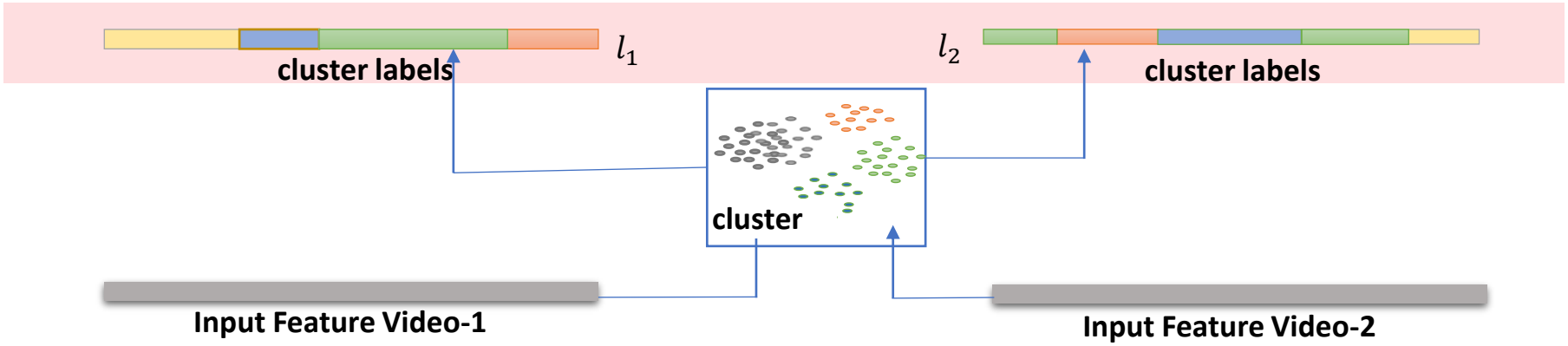
Negative Set Indices: $N_{(n,i)} \subset I \setminus \{(n, i)\}$.



- The key to effective unsupervised contrastive feature learning is to identify the relevant positive and negative sets to perform the targeted segmentation task.

Frame-level positive and negative sets

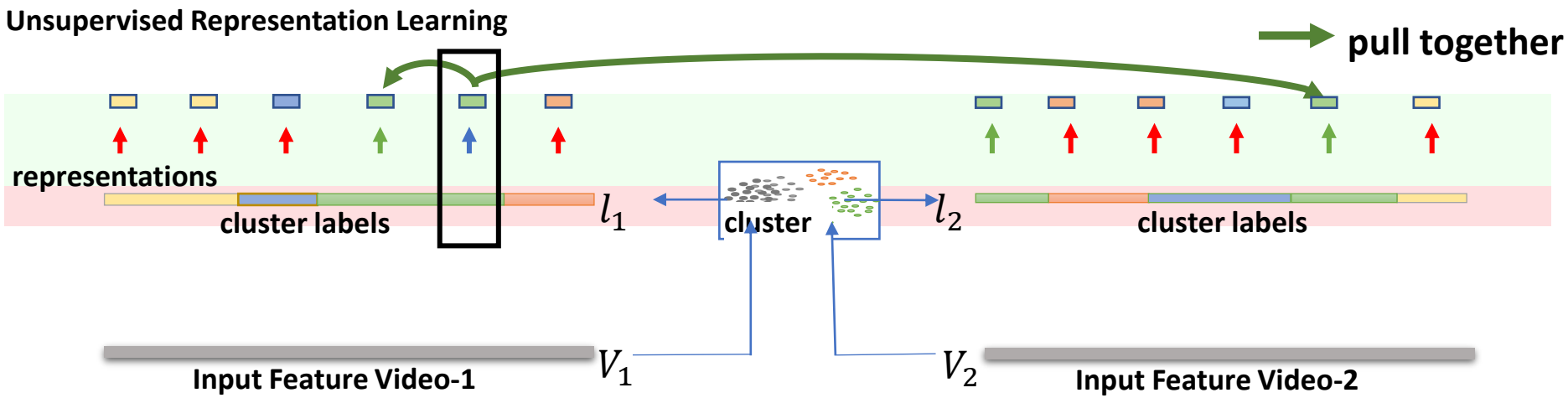
- Unsupervised setting has no labels to guide formation of these positive/negative sets and I3D/IDT pre-extracted features captures local discriminative motion.
- Therefore, we **Cluster the Input Features to get pseudo segmentation labels.**
- After clustering, each feature $f_{(n,i)}$ is assigned the cluster label $l[t_i^n]$.



Frame-level positive set

- Representation belong to same cluster and in close temporal proximity forms the positive set.
- Time proximity is added to reduce the false positives of clustering.

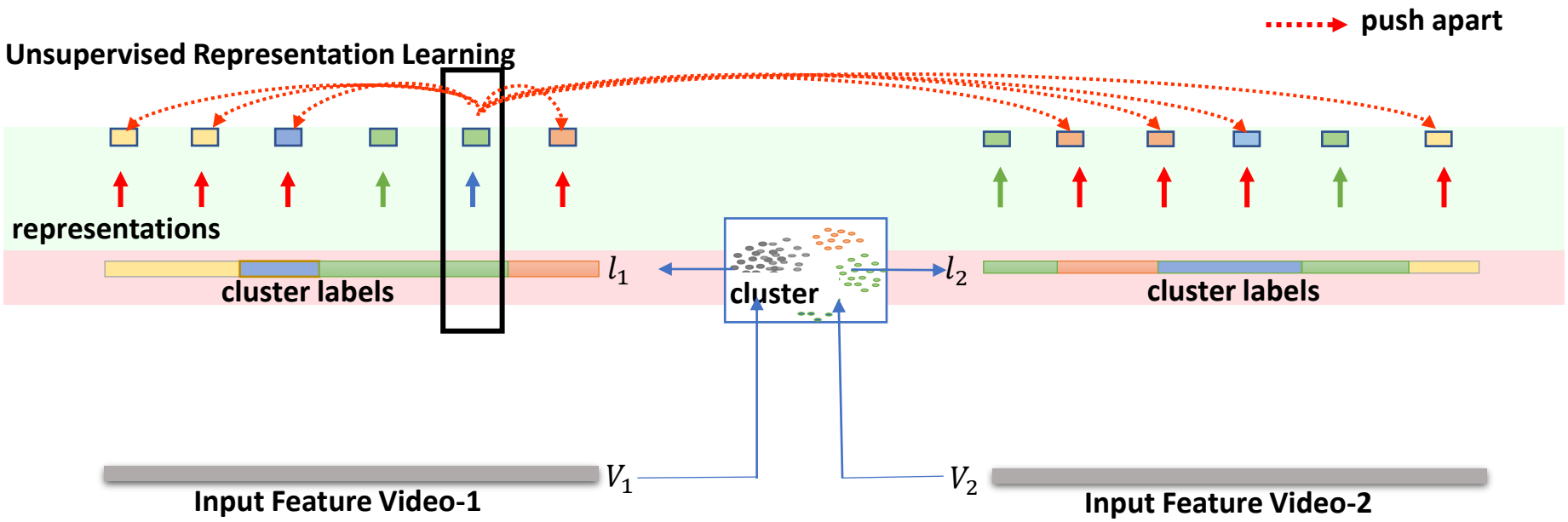
$$P_{(n,i)} = \{(m,j): |\mathrm{t_i^n} - \mathrm{t_j^m}| \leq \delta, l_n[t_i^n] = l_\mathrm{m}[t_j^m]\}$$



Frame-level negative set

- Negative Set: Representation not having the same cluster.

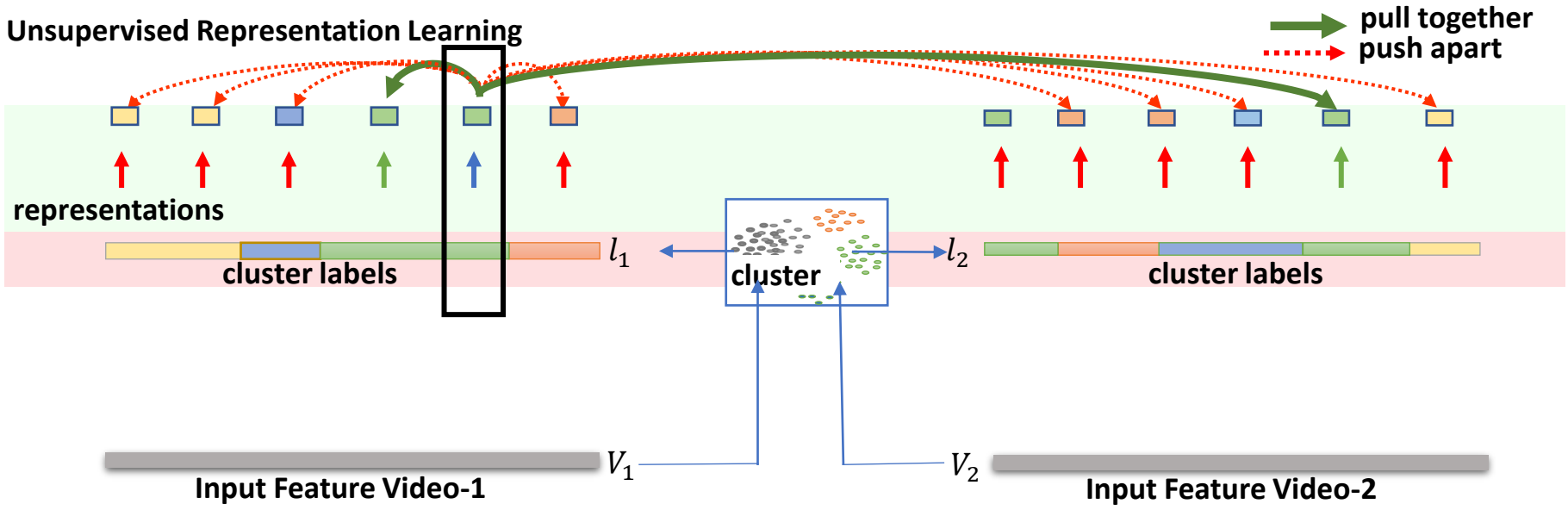
$$N_{(n,i)} = \{(m,j): l_n[t_i^n] \neq l_m[t_j^m]\}$$



Frame-Level Contrastive Formulation

For each $(m, j) \in P_{(n,i)}$, **Contrastive Feature Learning** maximizes the probability ensures that $f_{(m,j)}, f_{(n,i)}$ is more similar than any feature $(r, k) \in N_{(n,i)}$. e_τ is the τ scaled exponential cosine similarity

$$p_{nm}^{ij} = \frac{e_\tau(f_{(m,j)}, f_{(n,i)})}{e_\tau(f_{(m,j)}, f_{(n,i)}) + \Sigma_{(r,k)\in N_{(n,i)}} e_\tau(f_{(r,k)}, f_{(n,i)})}$$

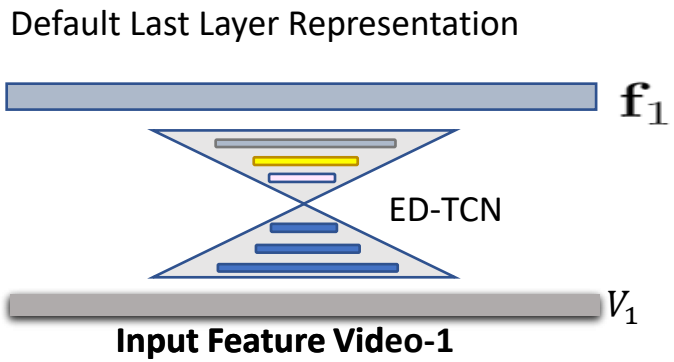


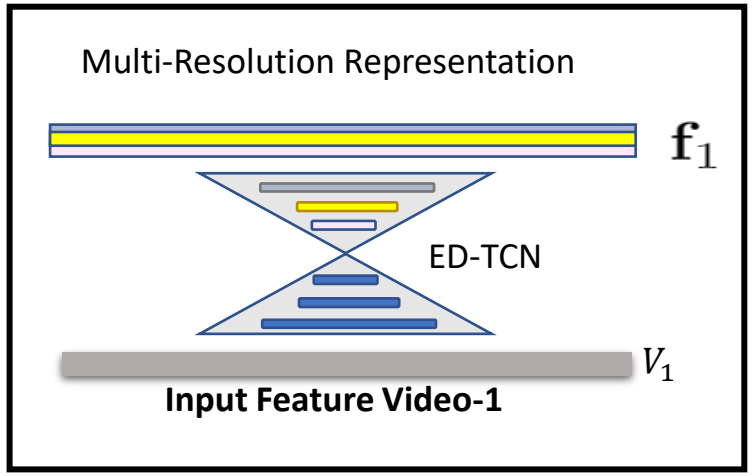
Additional Video-Level Contrastive Loss

- Datasets like Breakfast contains multiple complex activities videos examples, Making Coffee, Making Sandwich etc., and thus we need to separate representations from different complex activities.
- We construct video-level summary features by max-pooling the frame-level features along the temporal dimension.
- With max-pooled video level representation, we pull videos representation from same complex activity together and push apart representation from different complex activity.

Multi-Resolution Representation

- Default representation of Encoder-Decoder TCN's is last layer of decoder.
- However, we propose to use multi-resolution representation from multiple layers of decoder.
- We show using multi-resolution feature representation significantly improves contrastive learning.

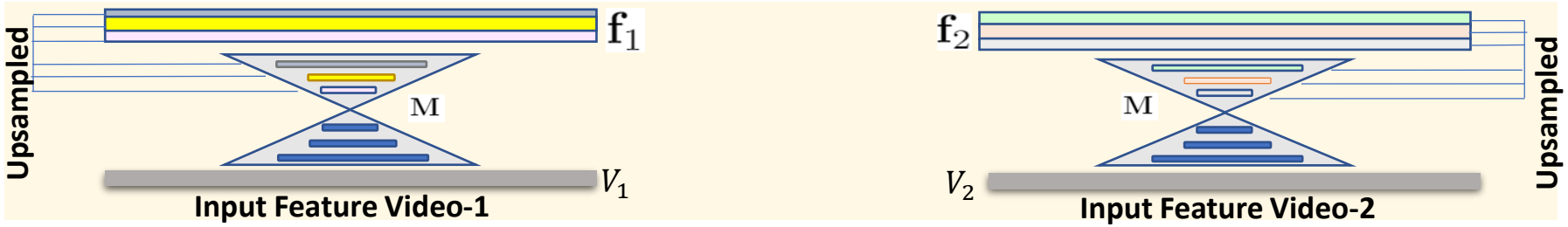




Multi-Resolution Representation

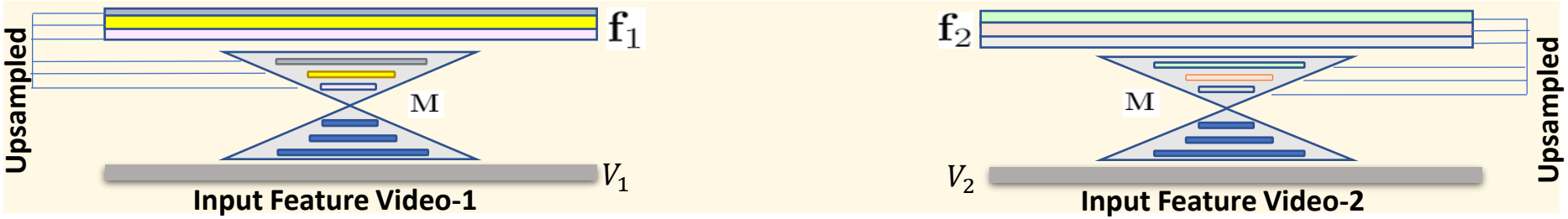
Multi-Resolution features are obtained by

1. Up sampling decoder's inherent multiple resolution features to have common length using a temporal linear interpolation.
2. Normalization and then concatenation along the latent dimension for each time.



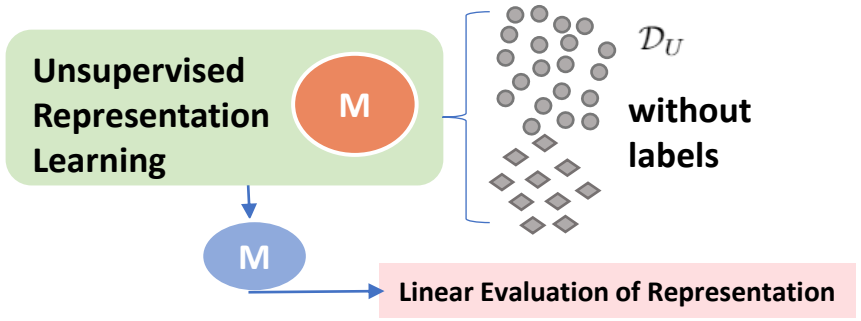
Multi-Resolution Representation

- Procedure of forming multi-resolution representation using Upsampling, Normalization and Concatenation encodes a degree of temporal continuity implicitly by design.
- Therefore, multi-resolution representations makes temporal segmentation less prone to over-segmentation and significantly improves representation learning.



Evaluating the Learned Representation

- We evaluate learnt representation using linear classifier.
- The assumption is if the unsupervised learned features are sufficiently strong, a simple linear classifier is sufficient to separate the action classes.



Improvement with Representation Learning

	Breakfast					50Salads					GTEA				
	$F1@\{10, 25, 50\}$			Edit	MF	$F1@\{10, 25, 50\}$			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	48.5	42.2	26.4	40.2	61.9
Our Representations	**57.0**	**51.7**	**39.1**	**51.3**	**70.5**	**40.8**	**36.2**	**28.1**	**32.4**	**62.5**	**70.8**	**65.0**	**48.0**	**65.7**	**69.1**
Improvement	52.1	49.2	38.2	46.0	40.3	28.6	28.3	24.1	24.0	7.5	22.3	22.8	21.6	25.5	7.2
Our unsupervised learning gives a large improvement in segmentation compared to input features.															

Significant gains over the input I3D, verifies the ability of the base TCN to perform the task of temporal segmentation with our designed unsupervised contrastive representation learning.

Ablations for Representation Learning

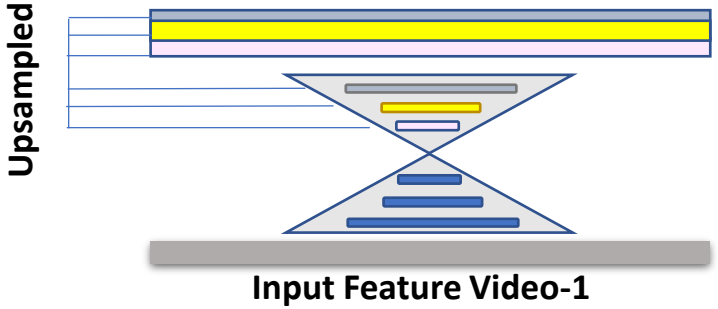
	Breakfast					50Salads					GTEA				
	$F1@\{10,25,50\}$			Edit	MF	$F1@\{10,25,50\}$			Edit	MF	$F1@\{10,25,50\}$			Edit	MF
Cluster	11.7	8.0	3.9	12.2	36.1	18.5	13.7	8.5	13.6	50.8	57.3	48.6	31.6	52.4	60.5
(+) Proximity	24.4	19.2	11.5	21.3	50.0	18.6	13.5	8.0	13.5	51.6	62.9	56.6	38.0	52.6	62.2
(+) Video-Level	42.9	37.6	26.6	36.4	66.1	–	–	–	–	–	–	–	–	–	–

Contribution of clustering and time-proximity conditions and video-level constraints for contrastive learning (with z_6).

- *'Cluster' row applies cluster labels condition,* $l_n[t_i^n] = l_m[t_j^m]$
- *' (+) Proximity' adds the condition* $|t_i^n - t_j^m| \leq \delta$.
- *Adding time proximity is more effective for Breakfast and GTEA likely because their videos follows a more rigid sequencing of actions than salads preparation videos of 50Salads.*
- *Adding the Video-Level contrastive loss used in Breakfast only gives further boosts.*

Improvement with Multi-Resolution Representation

	Breakfast					50Salads					GTEA				
	$F1@\{10,25,50\}$			Edit	MF	$F1@\{10,25,50\}$			Edit	MF	$F1@\{10,25,50\}$			Edit	MF
Last-Layer(z_6)	42.9	37.6	26.6	36.4	66.1	18.6	13.5	8.0	13.5	51.6	62.9	56.6	38.0	52.6	62.2
Multi-Resolution(f)	**57.0**	**51.7**	**39.1**	**51.3**	**70.5**	**40.8**	**36.2**	**28.1**	**32.4**	**62.5**	**70.8**	**65.0**	**48.0**	**65.7**	**69.1**
Improvement	14.1	14.1	12.5	14.9	4.4	22.2	22.7	20.1	18.9	10.9	7.9	8.4	10.0	13.1	6.9
Using Multi-Resolution(f) representation instead of final decoder z_6 significantly improves learned representation scores.															



- *Multi-Resolution feature encodes some degree of similarity bringing temporal continuity in temporal segmentation output.*

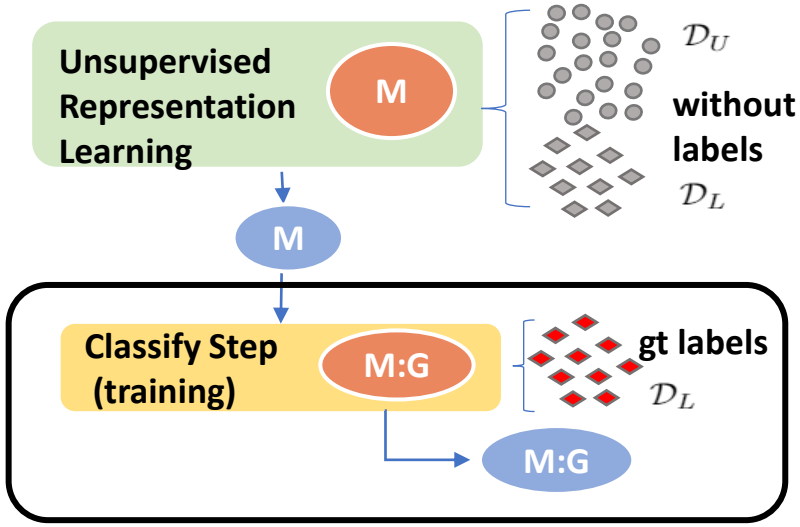
Semi-Supervised Temporal Segmentation

Unsupervised learning is used as pretraining method however it does not have information of the action classes. We train a linear classifier with few labelled videos for semi-supervised temporal segmentation results.

Iterative Contrast-Classify(ICC) Semi-Supervised

Classify Step: Learning G,M with $\mathbf{D_L}$

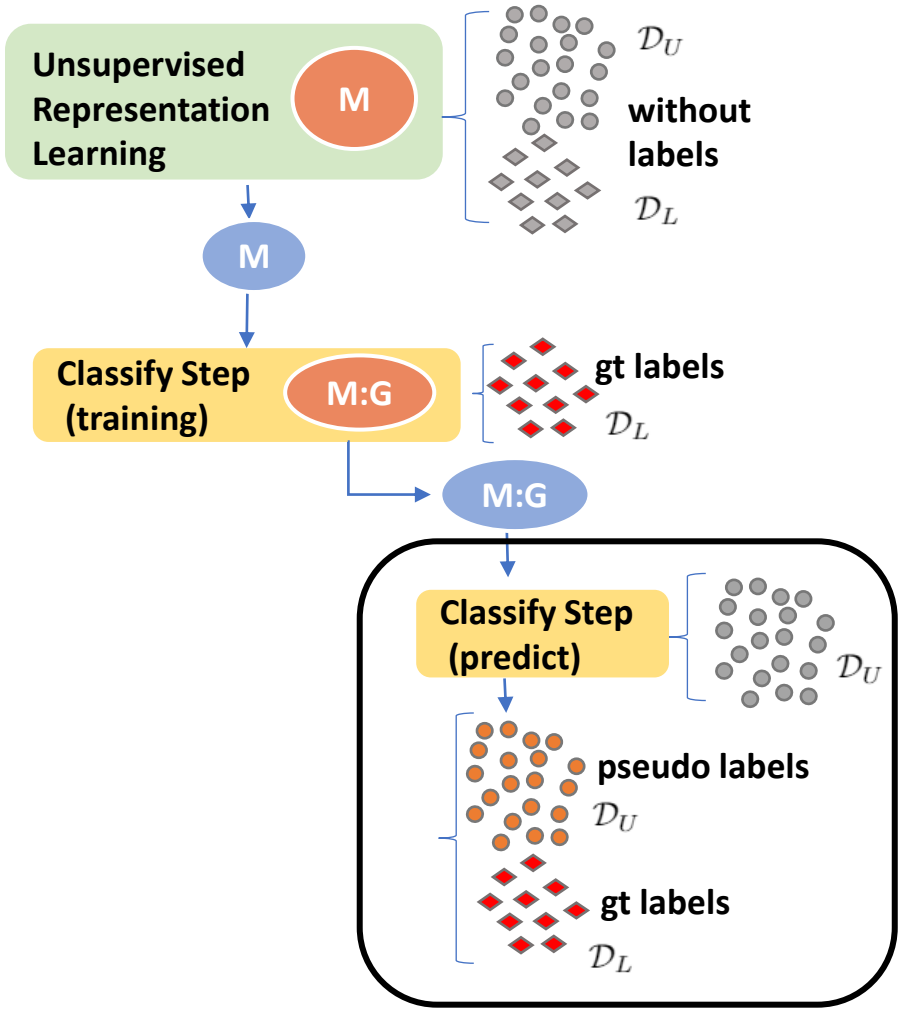
- Model M is coupled with a linear projection G and a softmax to generate the actual segmentation.
- G can only be learned using few labelled training videos, i.e., from D_L.
- In addition to learning G, D_L is leveraged to finetune M as well.
- The learning rate used for fine-tuning the parameters of the model M is significantly lower than linear projection layers G.



Iterative Contrast-Classify(ICC) Semi-Supervised

Pseudo Labels: Label prediction of $\mathbf{D_U}$ using M:G

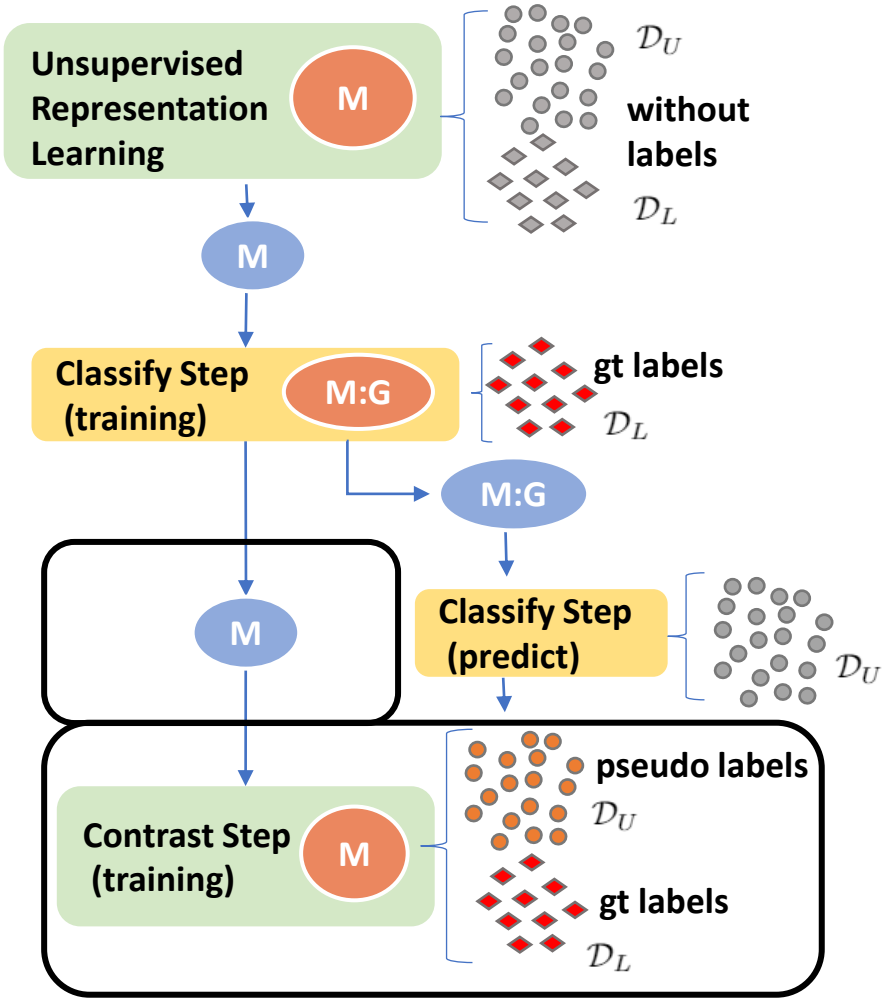
- After classification fine-tuning, we can use M and G to predict frame-level action labels for all unlabeled training videos, i.e., generates pseudo-labels for D_U.
- **Intuition:** The pseudo labels for D_U is significantly better than the input features' clusters labels used initially in the unsupervised representation learning.



Iterative Contrast-Classify(ICC) Semi-Supervised

Contrast Step: Update M with $\mathbf{D_L} \cup \mathbf{D_U}$

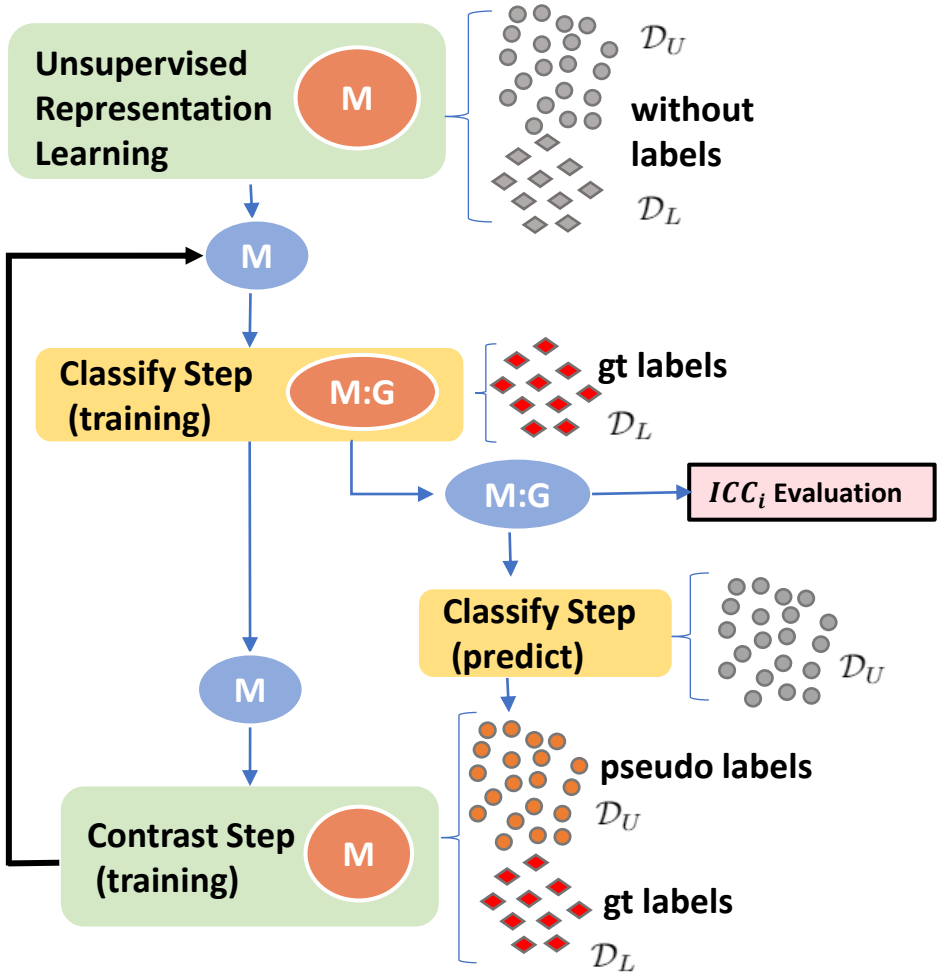
- Therefore, with stronger pseudo labels of $D_U \cup D_L$, contrastive representations in base model M is updated by replacing cluster labels with the pseudo-labels.



Iterative Contrast-Classify(ICC) Semi-Supervised

Iterate between Contrast and Classify Steps (ICC)

- Pseudo labels refined M in turn can help in finding better pseudo labels through another following classify step.
- By iterating between the contrast and classify, we can thus progressively improve the performance of the semi-supervised segmentation.
- The segmentation performance is evaluated at the end of the classify step after the training of G.
- Initial unsupervised representation learning can be considered the ``contrast" step of ICC_1.
- Performance saturates after 4 iterations of contrast-classify and ICC_4 is our final semi-supervised result.



Improvements with ICC

		Breakfast					50Salads					GTEA				
$\%D_L$	**Method**	$F1@\{10,25,50\}$			Edit	MoF	$F1@\{10,25,50\}$			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	70.3	66.5	49.5	64.7	66.0
	ICC_2	56.9	51.9	34.8	56.5	65.4	45.7	40.9	30.7	40.9	59.5	77.0	70.6	54.1	67.8	68.0
	ICC_3	59.9	53.3	35.5	56.3	64.2	50.1	46.7	35.3	43.7	60.9	77.6	71.2	54.2	71.3	68.0
	ICC_4	60.2	53.5	35.6	56.6	65.3	52.9	49.0	36.6	45.6	61.3	77.9	71.6	54.6	71.4	68.2
	Gain	5.7	4.8	2.3	2.0	1.1	11.6	11.8	8.8	10.2	4.0	7.6	5.1	5.1	6.7	2.2

Progressive semi-supervised improvement with more iterations of ICC.

- *We see progressive improvements with increase in number of iterations of our proposed ICC algorithm.*
- *The gain in performance is especially noticeable for Edit and F1 scores, which suggest contrastive representation learning bring temporal continuity with segmentation results.*

Qualitative Example Showing Improvement in ICC

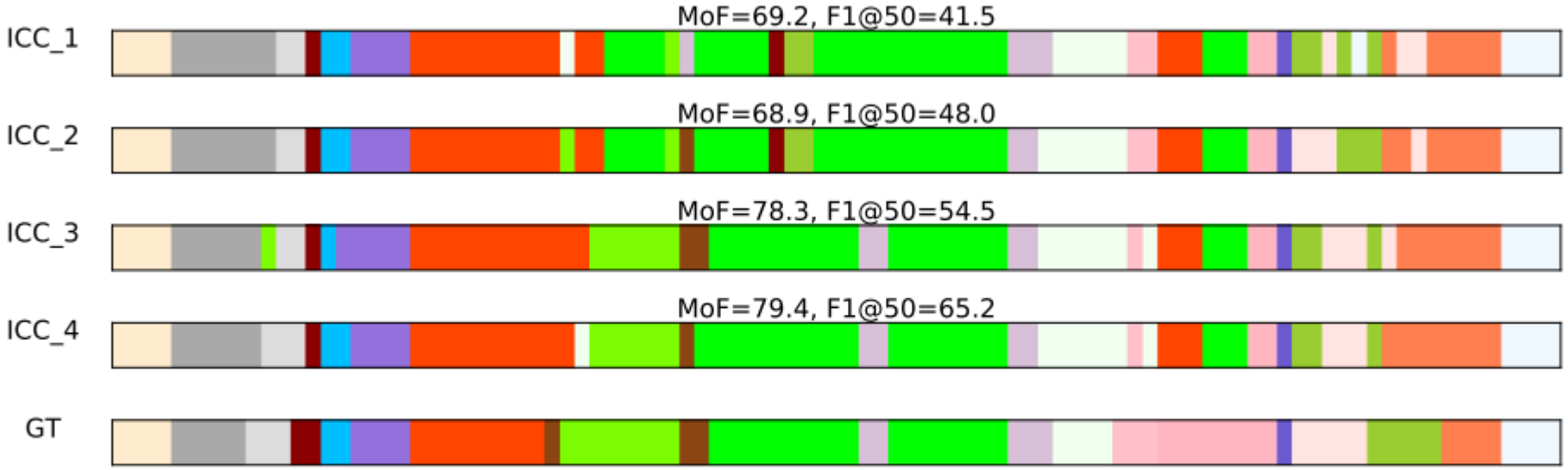


Figure 6: A qualitative example taken from of 50salads, showing progressive improvement in segmentation results with number of iterations of ICC. Some segments becomes more aligned to ground truth(GT) leading improved MoF and F1@50 scores.

ICC Semi-Supervised Results

$\%D_L$	Method	Breakfast					50Salads					GTEA				
		$F1@\{10,25,50\}$			Edit	MoF	$F1@\{10,25,50\}$			Edit	MoF	$F1@\{10,25,50\}$			Edit	MoF
≈5	Supervised	15.7	11.8	5.9	19.8	26.0	30.5	25.4	17.3	26.3	43.1	64.9	57.5	40.8	59.2	59.7
	Semi-Super	**60.2**	**53.5**	**35.6**	**56.6**	**65.3**	**52.9**	**49.0**	**36.6**	**45.6**	**61.3**	**77.9**	**71.6**	**54.6**	**71.4**	**68.2**
	Gain	44.5	41.7	29.7	36.8	39.3	22.4	23.6	19.3	19.3	18.2	13.0	14.1	13.8	12.2	8.5
≈10	Supervised	35.1	30.6	19.5	36.3	40.3	45.1	38.3	26.4	38.2	54.8	66.2	61.7	45.2	62.5	60.6
	Semi-Super	**64.6**	**59.0**	**42.2**	**61.9**	**68.8**	**67.3**	**64.9**	**49.2**	**56.9**	**68.6**	**83.7**	**81.9**	**66.6**	**76.4**	**73.3**
	Gain	29.5	28.4	22.7	25.6	28.5	22.2	26.6	22.8	18.7	13.8	17.5	20.2	21.4	13.9	12.7
Semi-Super (our ICC_4) significantly improves supervised counterpart using same labelled data amount. See also Fig. 1																

We shows our final 'Semi-Super' results, i.e., ICC_4 for various percentages of labelled data.
We compare with the 'Supervised' case of training the base model with the same labelled dataset.

SOTA comparison

	Method	Breakfast	50salads	GTEA
Full	MSTCN'20	67.6	83.7	78.9
	SSTDA'20	70.2	83.2	79.8
	*C2F-TCN'21	73.4	79.4	79.5
	Ours ICC (100%)	**75.2**	**85.0**	**82.0**
Weakly	SSTDA(65%)	65.8	80.7	75.7
	TSS'21	64.1	75.6	66.4
Semi	Ours ICC (40%)	71.1	78.0	78.4
	Ours ICC (10%)	68.8	68.6	73.3
	Ours ICC (5%)	65.3	61.3	68.2

Table 3: Segmentation MoF comparison with *SOTA* on 3 benchmark datasets. Our ICC can improve its fully-supervised counterpart. Our semi-supervised results is competitive in MoF with different levels of supervision.

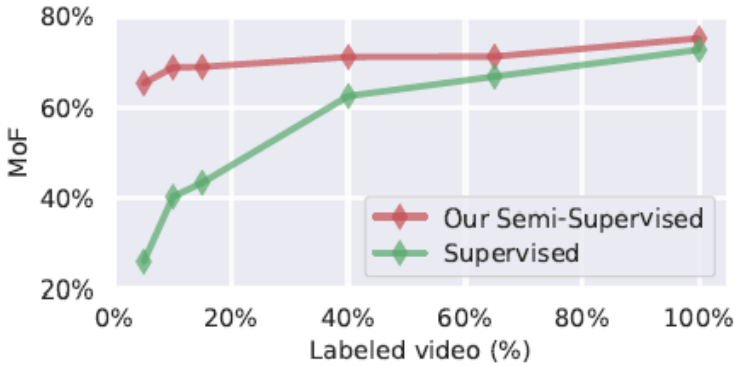


Figure 1: **Frame accuracy on Breakfast dataset:** Our semi-supervised approach has impressive performance with just 5% labelled videos; at 40%, we almost match the Mean over Frames (MoF) of a 100% fully-supervised setup.

Conclusion

- We show that pre-trained input features that capture semantics and motion of short-trimmed video segments can be used to learn higher-level representations to interpret long video sequences.
- Our proposed multi-resolution representation formed with outputs from multiple decoder layers, implicitly bring temporal continuity and consequently large improvements in unsupervised contrastive representation learning.
- Our final semi-supervised learning algorithm ICC can significantly reduce the annotation efforts, with 40% labelled videos approximately achieving fully-supervised (100%) performance.
- Furthermore, ICC also improves performance even when used with 100% labels.