Semi-Weakly-Supervised Learning of Complex Actions from Instructional Task Videos

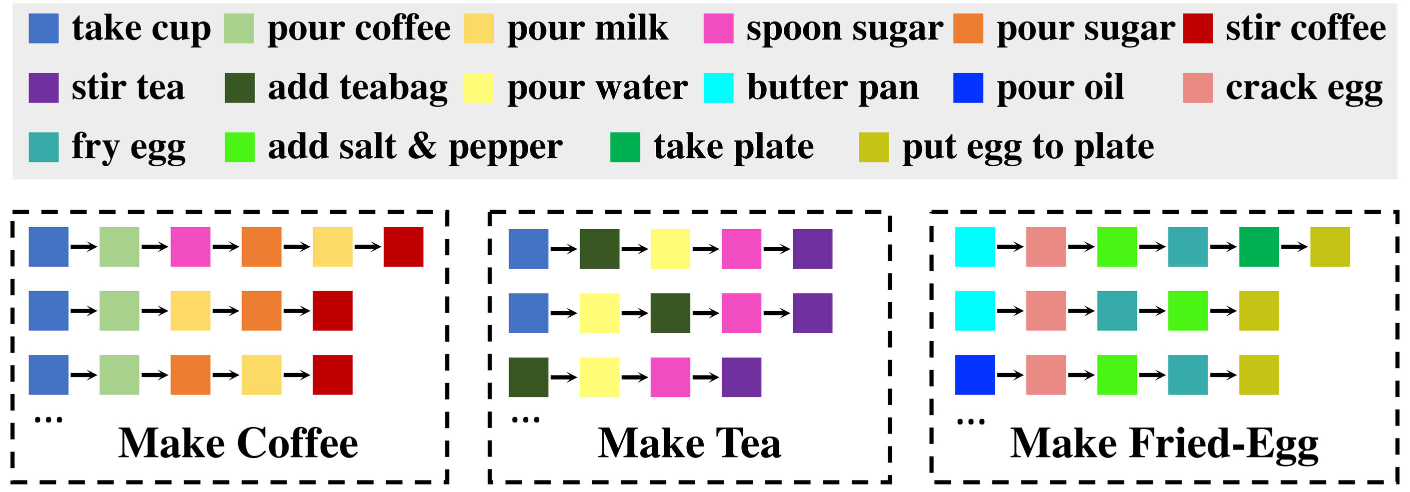
Yuhan Shen
e-mail: shen.yuh@northeastern.edu

Ehsan Elhamifar
e-mail: e.elhamifar@northeastern.edu

Khoury College of Computer Sciences, Northeastern University, Boston, USA

Motivation

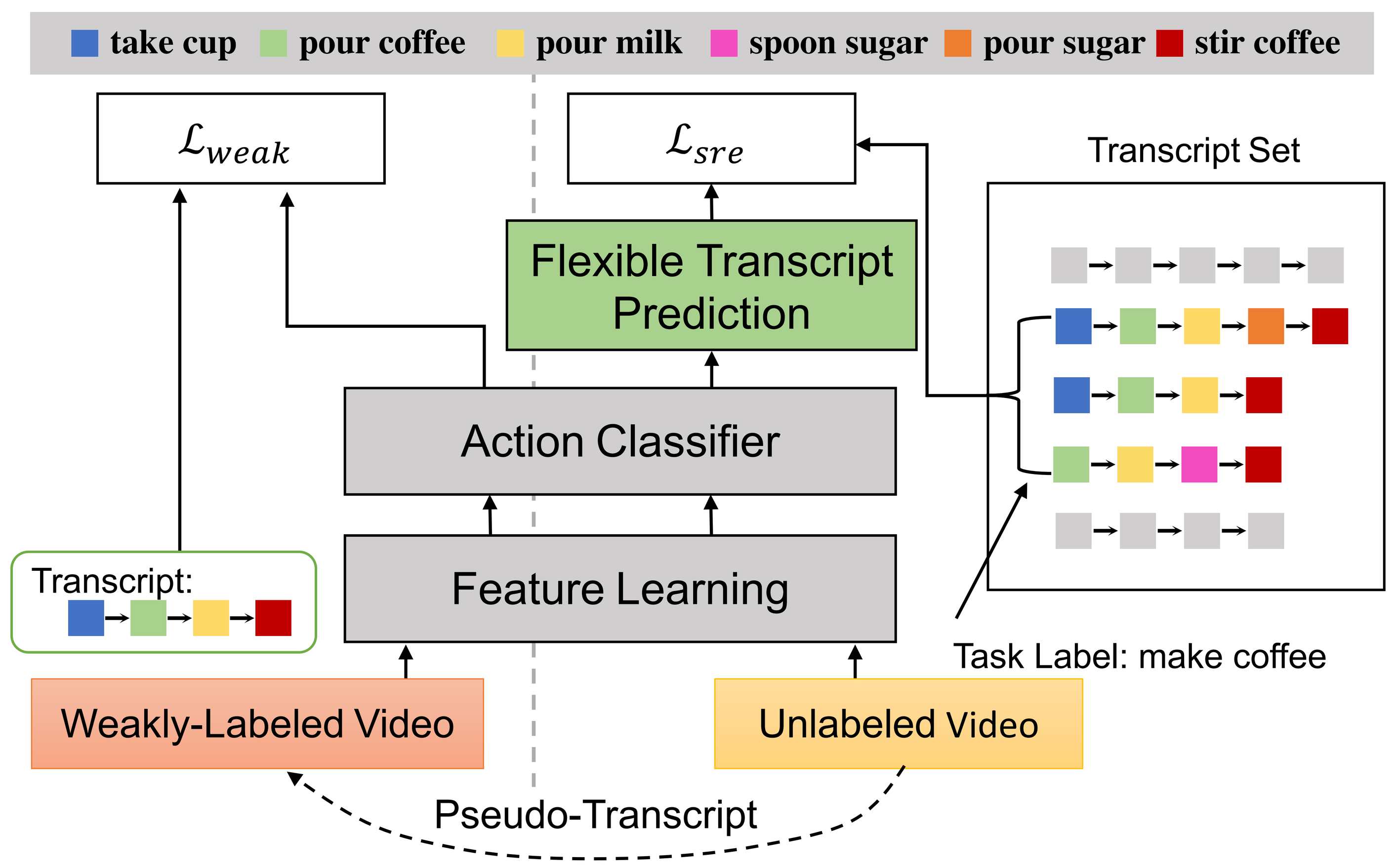
We introduce **Semi-Weakly-Supervised Learning (SWSL)** of complex actions: train an action segmentation model using a small number of *weakly-labeled* videos and a large number of *unannotated* videos of procedural tasks (with task labels).



- **Key Observation**: transcripts of unlabeled videos often have small but nonzero distances to the transcripts of the weakly-labeled videos from the same task.

Contributions

- Propose a **Semi-Weakly-Supervised Learning** scheme for action segmentation from instructional videos by using a small set of weakly-labeled videos and a large set of unlabeled videos.
- It can **work with any weakly-supervised method**.
- Develop a **Flexible Transcript Prediction** method to recover the transcripts of unlabeled videos given the predicted probabilities.
- Develop a **Soft Restricted Edit** loss to find weak alignment between transcripts while allowing insertion, deletion, substitution, and **adjacent transposition**.
- Significantly improve the performance by adding unlabeled videos for training. Code is available on `https://github.com/Yuhan-Shen/SWSL`



- **Training Objective**: $\mathcal{L}_{swsl} = \mathcal{L}_{weak} + \rho\mathcal{L}_{sre}$.
- **Self Training**: iteratively **generate pseudo-transcripts** for unlabeled videos.

Prior Work

- Fully-supervised methods: framewise annotation; expensive annotation cost.
- Weakly-supervised methods: an ordered (or unordered) list of actions in each video; reduce annotation cost, but still require watching the whole videos.
- Unsupervised methods: remove annotation cost; limited capability.

Flexible Transcript Prediction

- Predict the transcript of unannotated videos from **maximizing the likelihood**:

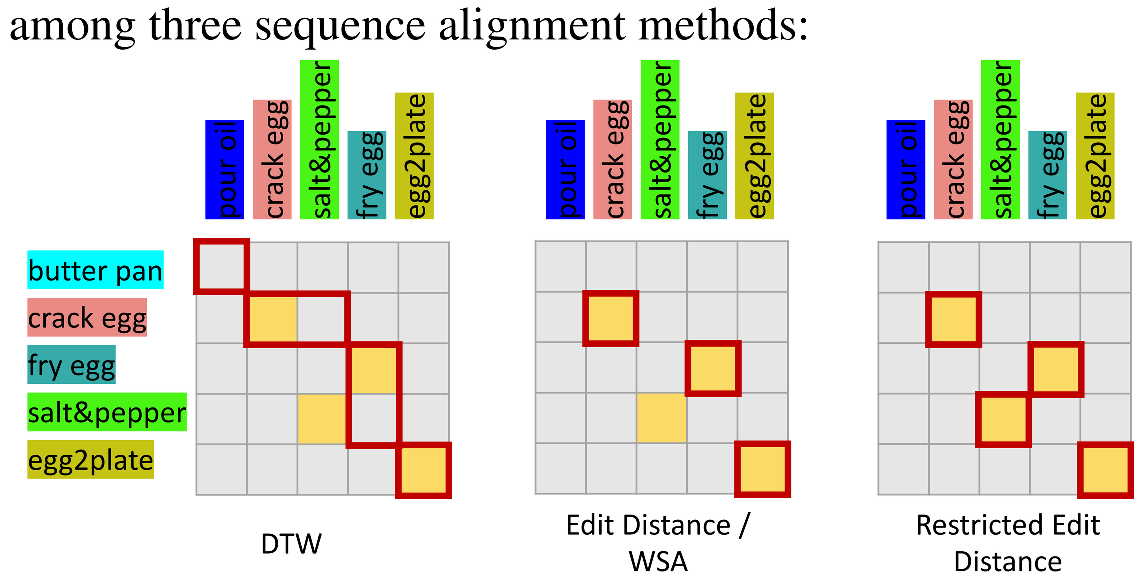
$$\max_{\{t_i\},\{a_i\}} \prod_{i=1}^{L} \prod_{j=t_{i-1}+1}^{t_i} p_{j,a_i} \quad \text{s.t.} \quad t_0 = 0,\ t_L = T. \tag{1}$$

- Predict **transcript with flexible length** by **dynamic programming**:

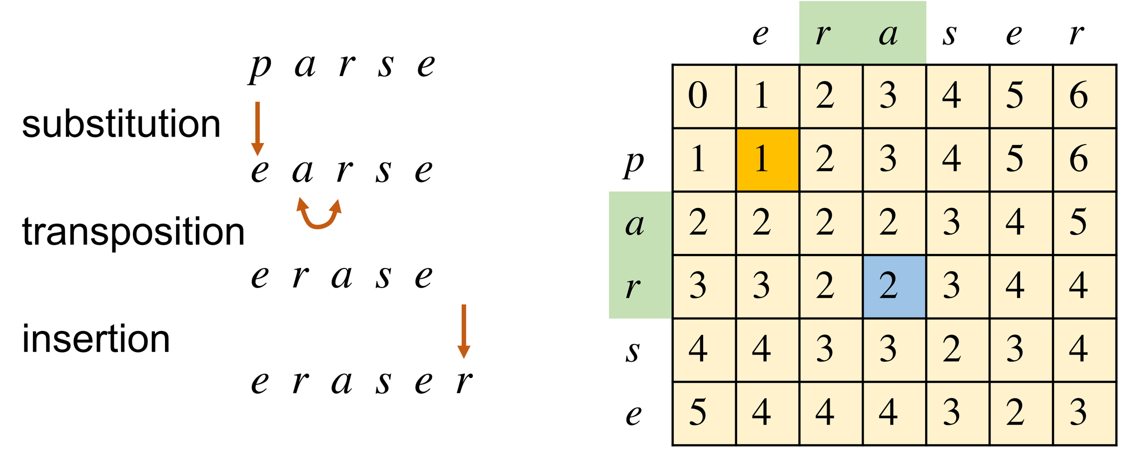
$$\min_{\{t_i\},\{a_i\},L} -\frac{1}{T}\sum_{i=1}^{L} \sum_{j=t_{i-1}+1}^{t_i} \log p_{j,a_i} + \lambda L \ \text{s.t.}\ t_0 = 0,\ t_L = T,\ \ell_{min} \leq L \leq \ell_{max}, \tag{2}$$

Soft Restricted Edit (SRE) Distance

- A **differentiable** loss function that measures the distance between two sequences.
- Allow insertion, deletion, substitution and **adjacent transposition**.
- **Minimize the distance** between the transcript of weakly-labeled videos and the transcript of unlabeled videos predicted by **Flexible Transcript Prediction**.
- Difference among three sequence alignment methods:



- Illustration of Restricted Edit Distance:



- **Soft Restricted Edit Distance**:

$$e_{i,j} = \min_\beta \begin{cases} e_{i-1,j} + c_D, & \text{(deletion)} \\ e_{i,j-1} + c_I, & \text{(insertion)} \\ e_{i-1,j-1} + \delta_{i-1,j-1}, & \text{(substitution)} \\ e_{i-2,j-2} + \delta_{i-2,j-1} + \delta_{i-1,j-2} + c_T\ (\forall i,j \geq 3). & \text{(transposition)} \end{cases} \tag{3}$$

We develop efficient **forward and backward** algorithms to allow **end-to-end learning**:

Algorithm 1: Forward Propagation for SRE

input : Pairwise cost matrix $\Delta = [\delta_{i,j}] \in \mathbb{R}^{L' \times L}$; $c_D, c_I, c_T, \beta \geq 0$.

$e_{i,1} = (i-1) \cdot c_D, i \in \{1, 2, \ldots, L'+1\}$
$e_{1,j} = (j-1) \cdot c_I, j \in \{2, \ldots, L+1\}$
for $i \leftarrow 2$ **to** $L'+1$ **do**
  for $j \leftarrow 2$ **to** $L+1$ **do**
    update $e_{i,j}$ via (3);

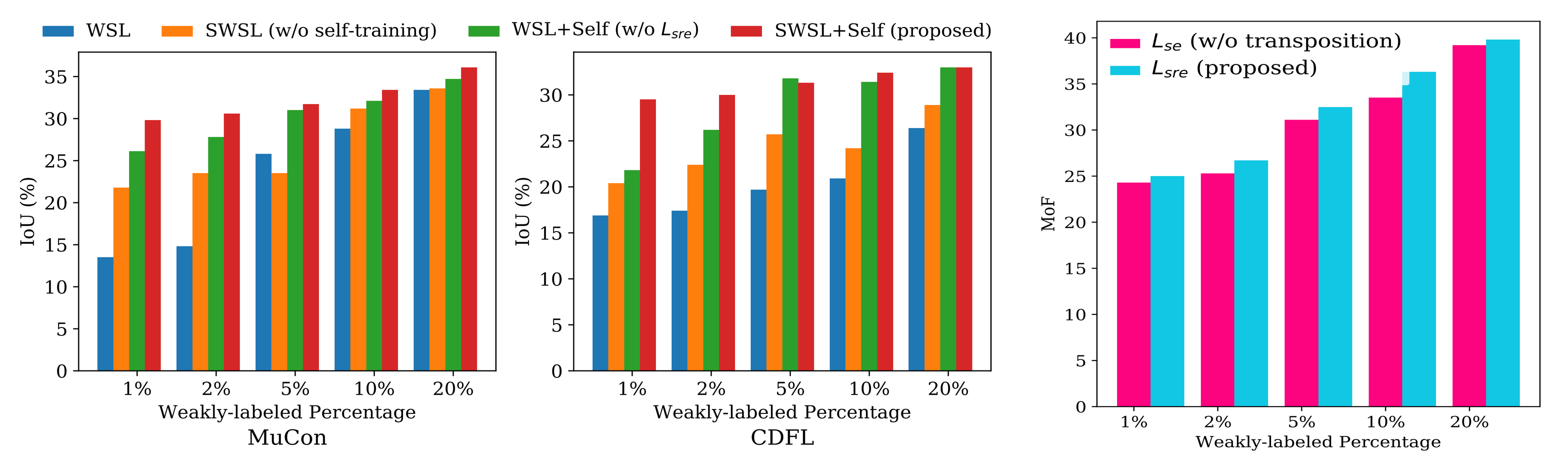
output : SRE Loss, $\mathcal{L}_{sre} = e_{L'+1,L+1}$

Experiments

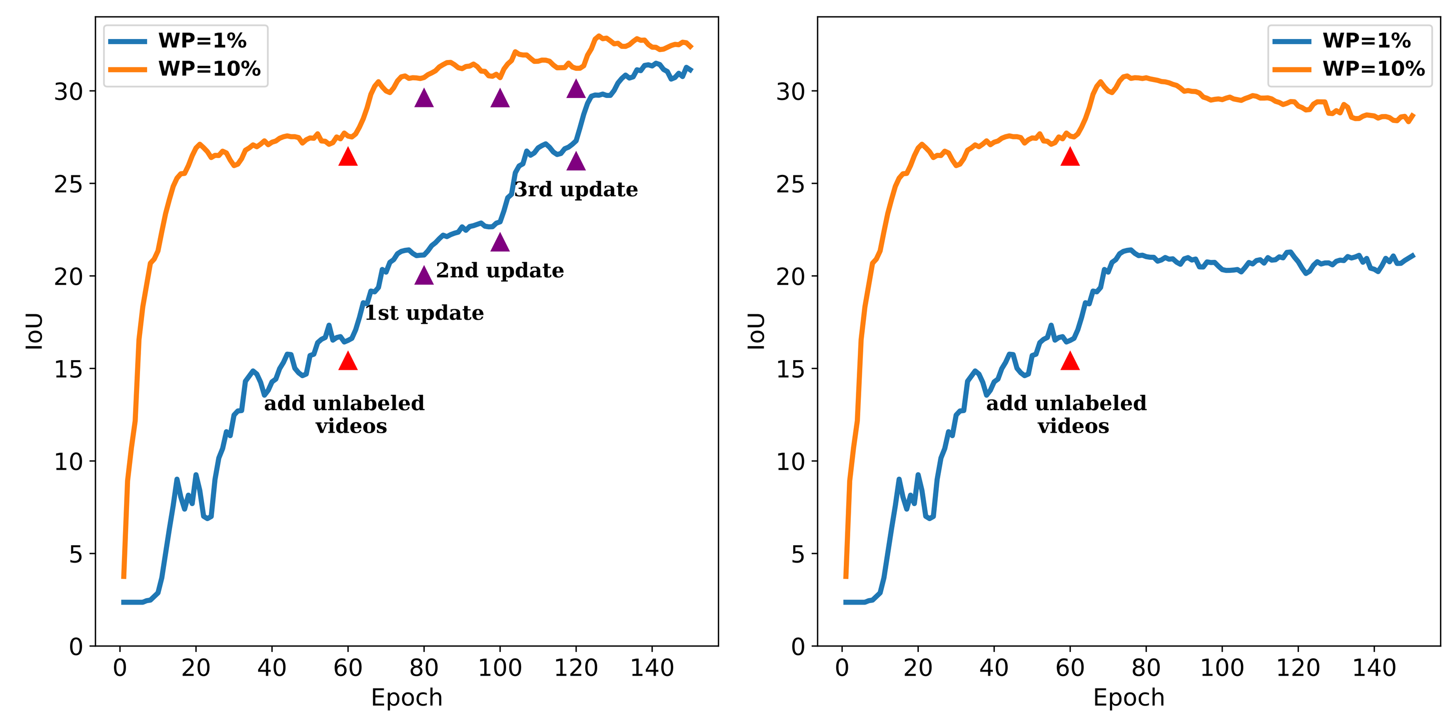
- **Networks:** MuCon (Souri et al. TPAMI'21) and CDFL (Li et al. ICCV'19).
- **Quantitative Results:** WP: **W**eakly-labeled video **P**ercentage. UP: **U**nlabeled video **P**ercentage.

	WP	UP	MuCon					CDFL				
			Breakfast		CrossTask			Breakfast		CrossTask		
			MoF	IoU	MoF	IoU	F1	MoF	IoU	MoF	IoU	F1
WSL	1%	0	11.0	13.5	38.2	14.6	2.6	10.9	16.9	20.7	8.6	3.0
SWSL+Self	1%	99%	**25.0**	**29.8**	**48.1**	**17.9**	**8.9**	**32.4**	**29.5**	**21.8**	**9.2**	**9.9**
WSL	2%	0	12.9	14.8	44.0	15.8	5.3	10.9	17.4	20.5	8.6	5.3
SWSL+Self	2%	98%	**26.7**	**30.6**	**44.6**	**17.8**	**11.3**	**35.4**	**30.0**	**21.4**	**9.1**	**10.1**
WSL	5%	0	23.1	25.8	42.3	16.1	8.3	13.4	19.7	20.4	8.7	5.1
SWSL+Self	5%	95%	**32.5**	**31.7**	**50.6**	**18.3**	**11.5**	**39.6**	**31.3**	**22.6**	**9.1**	**11.3**
WSL	10%	0	28.0	28.8	42.1	16.7	9.9	20.4	20.9	23.2	9.0	7.8
SWSL+Self	10%	90%	**36.3**	**33.4**	**49.0**	**18.0**	**12.1**	**40.4**	**32.4**	**24.0**	**9.3**	**11.7**
WSL	20%	0	35.2	33.4	44.4	17.7	11.0	31.7	26.4	23.6	9.0	8.1
SWSL+Self	20%	80%	**39.8**	**36.1**	**54.5**	**19.3**	**11.8**	**43.5**	**33.0**	**24.8**	**9.0**	**13.2**
WSL	*100%*	*0*	*48.5*†	*39.1**	*48.4**	*21.0**	*16.7**	*50.2*†	*35.9**	*31.5**	*13.2**	*18.8**

- **Ablation Studies**: left: effect of SRE loss and self-training (left); right: comparison between SRE loss and SE loss (without allowing adjacent transposition).



- **Training Process**: IoU of different methods on the Breakfast test set as a function of the number of training epochs. Left: SWSL+Self. Right: SWSL.



- **Qualitative Results**: visualization of a test video about 'making scrambled egg'