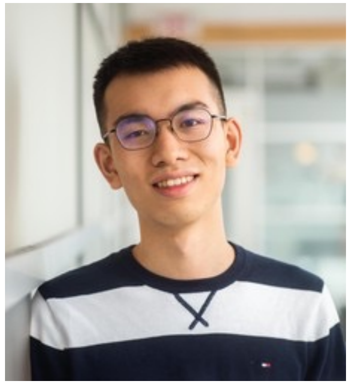
Semi-Weakly-Supervised Learning of Complex Actions from Instructional Task Videos



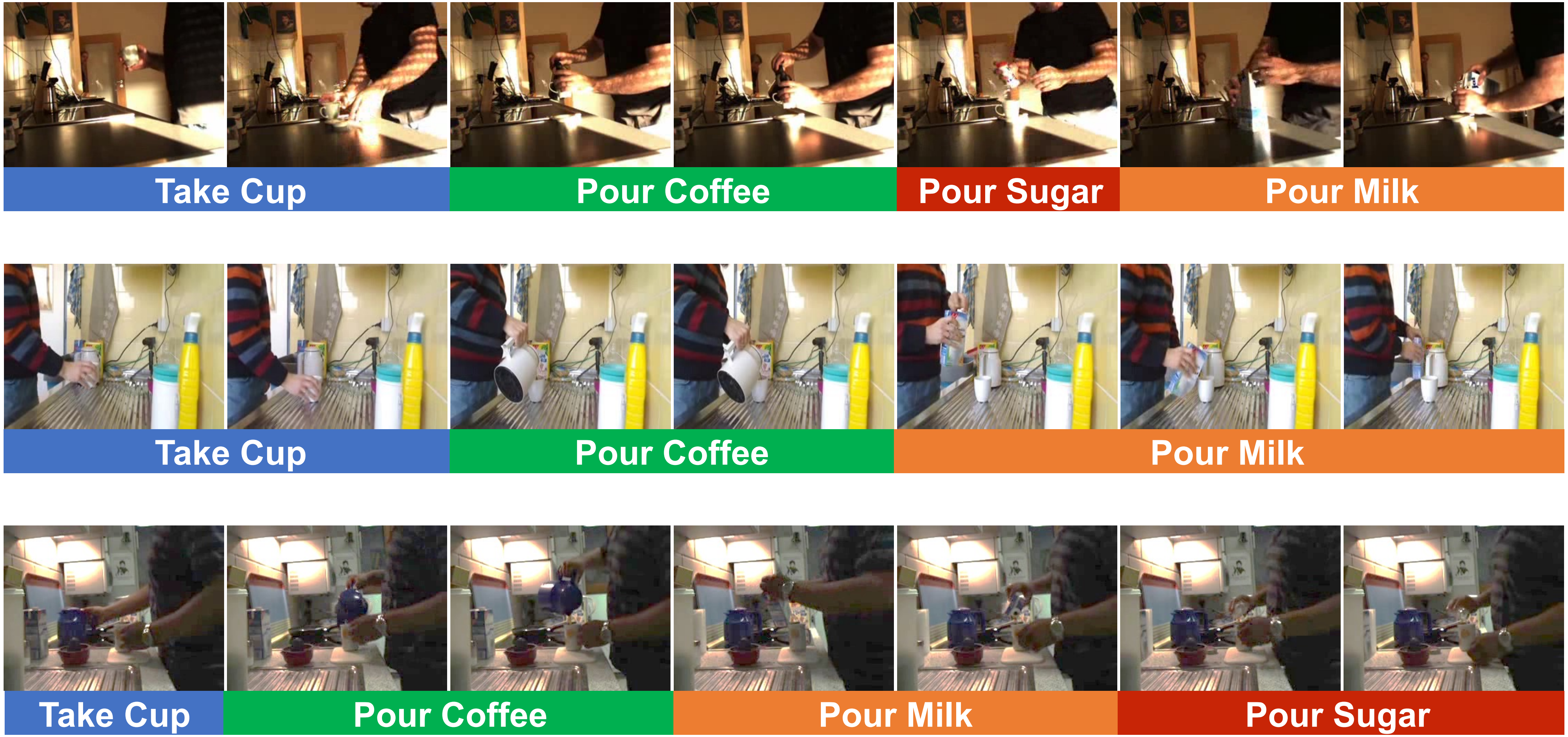
Yuhan Shen
Khoury College of Computer Sciences
Northeastern University

Ehsan Elhamifar
Khoury College of Computer Sciences
Northeastern University

Action Segmentation

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Goal: action segmentation in instructional task videos



Prior Work

- **Fully Supervised** [Kuehne et al'16, Lea et al'17, GoelBrunskill'19, Singh et al'19, Farha'19]: **framewise annotation;** **costly**

- **Weakly-Supervised** [Bojanowski et al'14, Ding-Xu'18, Chang et al'19, Li et al'19, Fayyaz-Gall'20, Souri et al'21, Lu-Elhamifar'21]: **an ordered or unordered list of actions** in each video; **still costly**

Transcript: **Take Cup** → **Pour Coffee** → **Pour Sugar** → **Pour Milk**

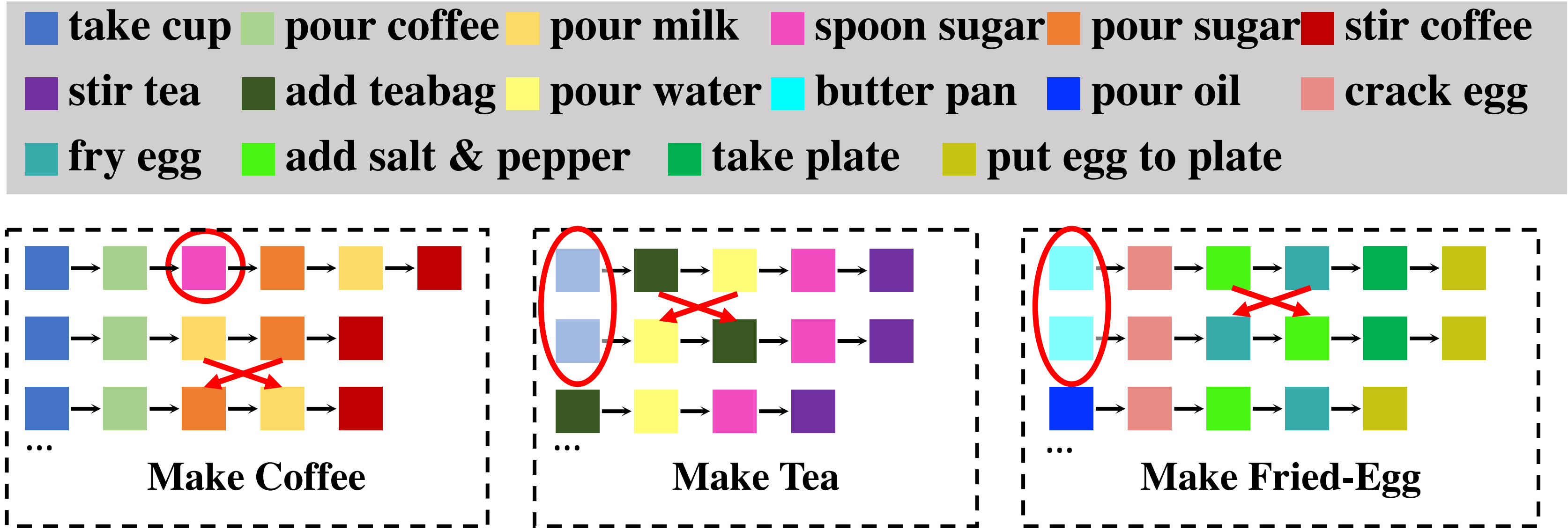
- **Unsupervised** [Alayrac et al'16, Sener-Yao'18, Elhamifar-Naing'19, Kukleva et al'19, Shen et al'21]: **no action-level annotation**; **task label** for each video;

limited performance

Task Label: **Make Coffee**

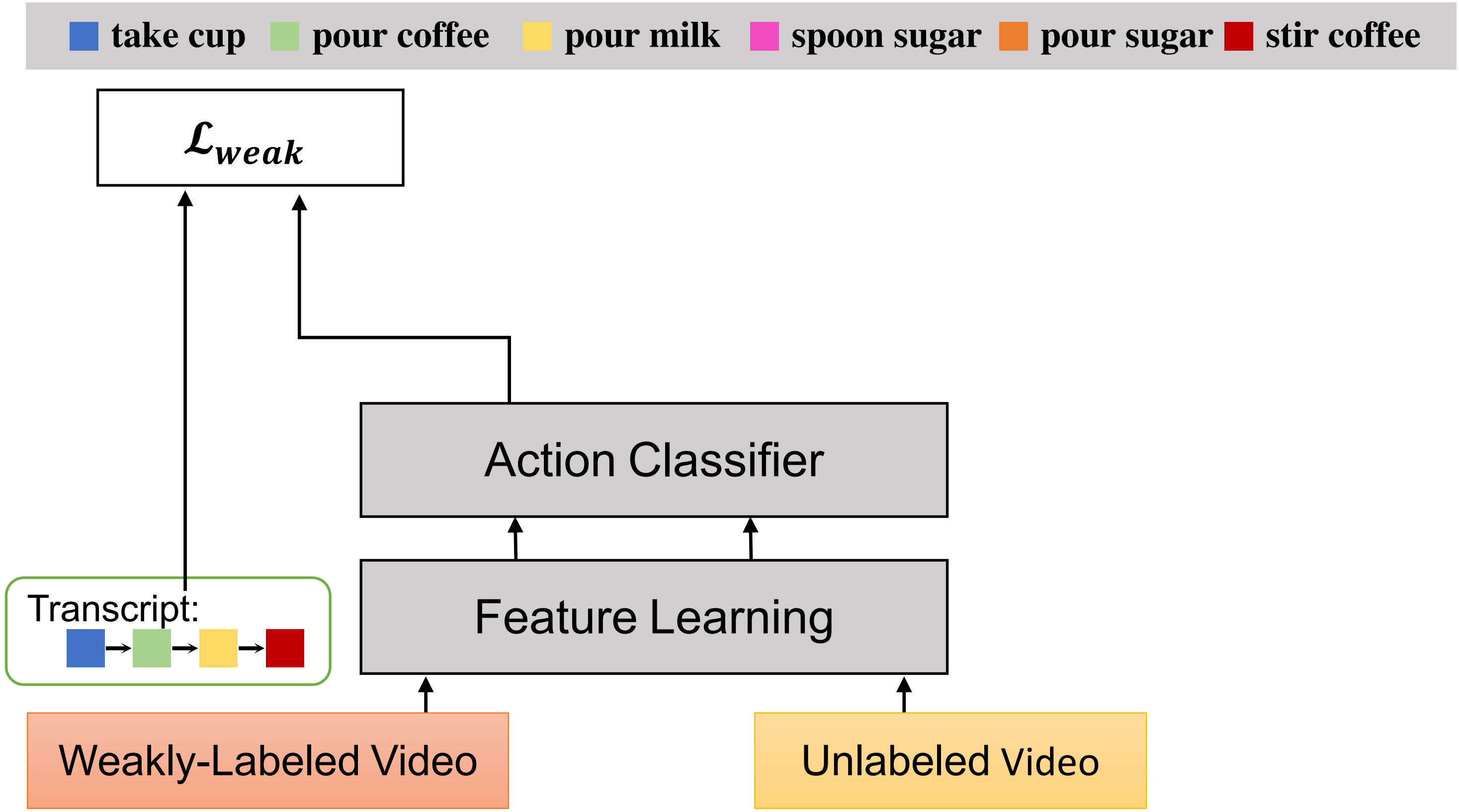
Contribution

- **Semi-Weakly Supervised Learning (SWSL) of complex actions**
  - Weakly-labeled videos (small) and unlabeled videos (large)
  - Unlabeled videos have **task labels**
- **Observation:** transcripts within the same task usually have **a small distance**
  - Missing actions
  - Adjacent actions are **swapped**



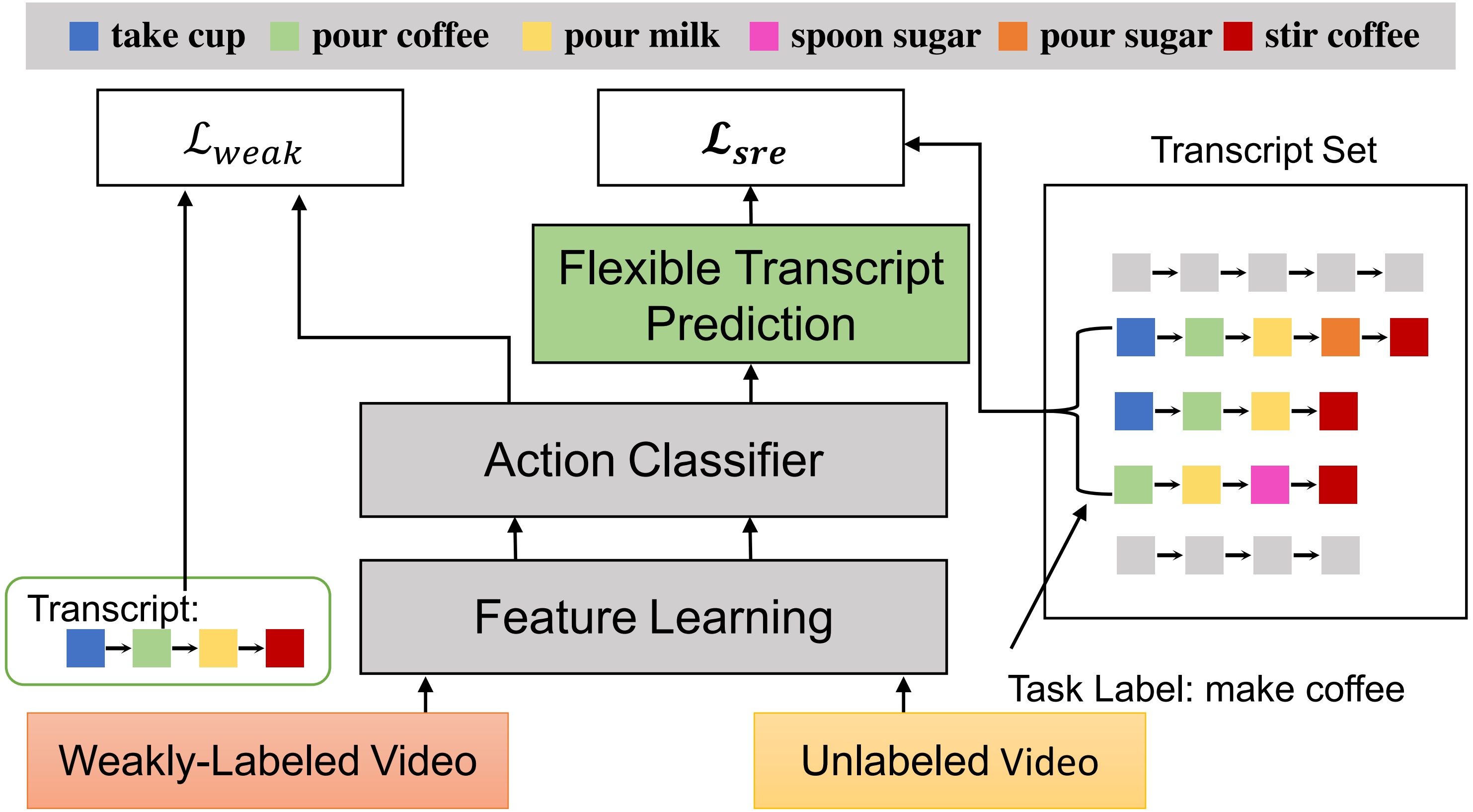
Proposed Approach

- L_{weak} : apply **weakly-supervised action segmentation** methods on weakly-labeled videos



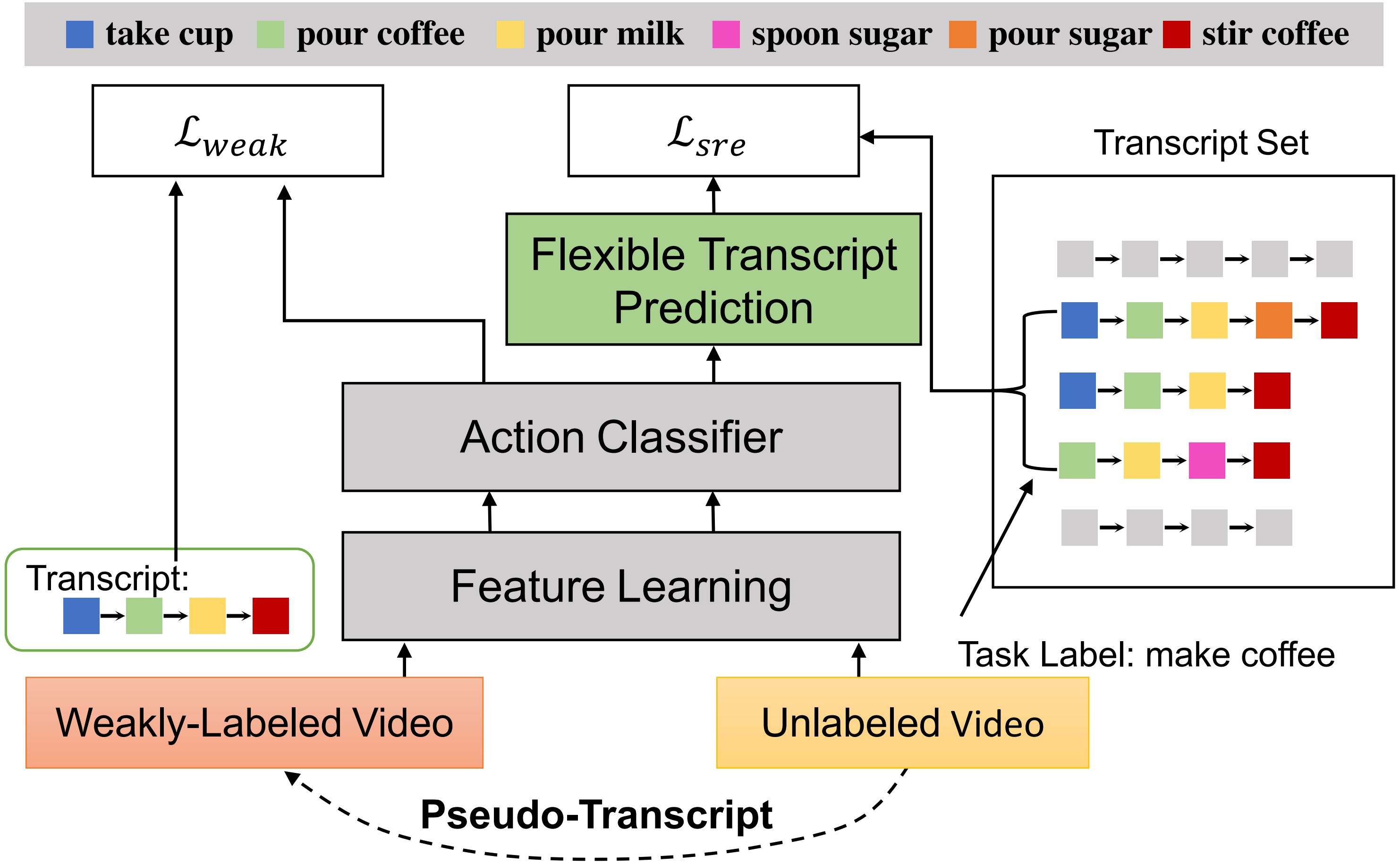
Proposed Approach

- **Flexible Transcript Prediction**: predict the transcript of unlabeled videos
- **Soft Restricted Edit Loss** ($\boldsymbol{L_{sre}}$): encourage a small distance between transcripts of the same task



Proposed Approach

- **Self training**: iteratively generate pseudo-transcripts for unlabeled videos in the order of confidence



Soft Restricted Edit Loss

- Motivated by **Restricted Edit Distance**
- Allow insertion, deletion, substitution, and **adjacent transposition**
- Consider two words: “parse” and “eraser”

p a r s e

substitution

e a r s e

deletion

e r s e

insertion

e r a s e

insertion

e r a s e r

Edit Distance: 4

p a r s e

substitution

e a r s e

transposition

e r a s e

insertion

e r a s e r

Restricted Edit Distance: 3

	e	r	a	s	e	r
	0	1	2	3	4	5	6
p	1	1	2	3	4	5	6
a	2	2	2	2	3	4	5
r	3	3	2	2	3	4	4
s	4	4	3	3	2	3	4
e	5	4	4	4	3	2	3

Soft Restricted Edit Loss

- **Dynamic programming** for Restricted Edit Distance
- Make it **differentiable**:
  - Replace indicator function with a **continuous distance function**
  - Replace minimum operation with **soft minimum**

$$e_{i,j} = \min \begin{cases} e_{i-1,j} + 1 & \text{(deletion)} \\ e_{i,j-1} + 1 & \text{(insertion)} \\ e_{i-1,j-1} + \mathbb{1}(\boldsymbol{x}_{i-1} \neq \boldsymbol{y}_{j-1}) & \text{(substitution)} \\ e_{i-2,j-2} + 1, \text{if } (\boldsymbol{x}_{i-2} = \boldsymbol{y}_{j-1}, \boldsymbol{x}_{i-1} = \boldsymbol{y}_{j-2}) & \text{(transposition)} \end{cases}$$

$$e_{i,j} = \min_\beta \begin{cases} e_{i-1,j} + c_D, \\ e_{i,j-1} + c_I, \\ e_{i-1,j-1} + \delta_{i-1,j-1}, \\ e_{i-2,j-2} + \delta_{i-2,j-1} + \delta_{i-1,j-2} + c_T \quad (\forall i,j \geq 3), \end{cases}$$

$$\boldsymbol{\min_\beta(a_1, a_2, \cdots) = -\beta log \sum_k e^{-\frac{a_k}{\beta}}}$$

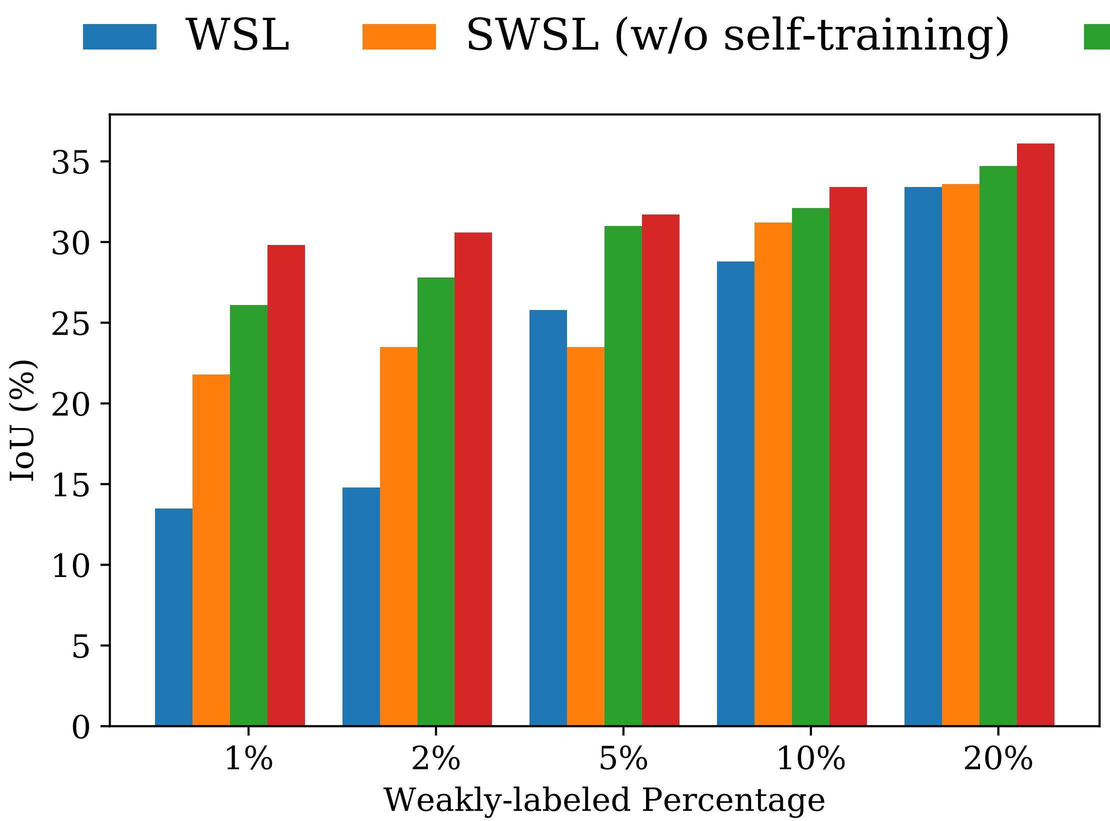
Experiments

- **Datasets:** Breakfast (Kuehne et al. CVPR’14) and CrossTask (Zhukov et al. CVPR’19)
- **Methods:** MuCon (Souri et al. TPAMI’21) and CDFL (Li et al. ICCV’19)
- Improve performance by **a large margin,** especially in the case of **limited labeled data**

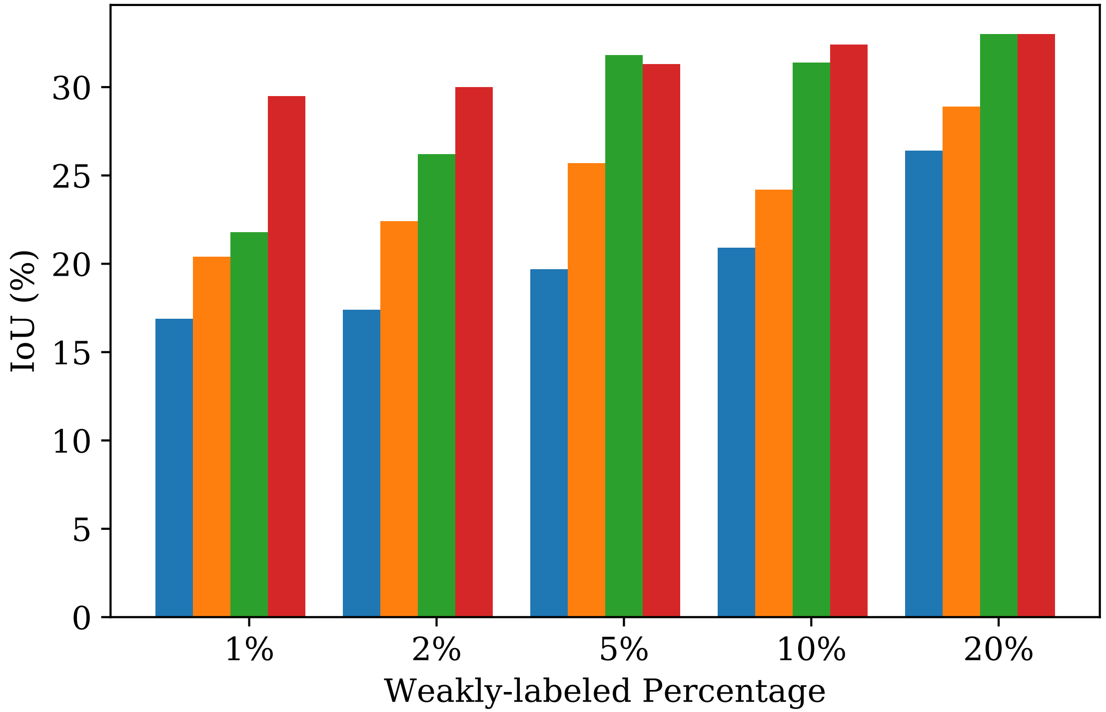
	WP	UP	MuCon					CDFL				
			Breakfast		CrossTask			Breakfast		CrossTask		
			MoF	IoU	MoF	IoU	F1	MoF	IoU	MoF	IoU	F1
WSL	1%	0	11.0	13.5	38.2	14.6	2.6	10.9	16.9	20.7	8.6	3.0
SWSL+Self	1%	99%	**25.0**	**29.8**	**48.1**	**17.9**	**8.9**	**32.4**	**29.5**	**21.8**	**9.2**	**9.9**
WSL	2%	0	12.9	14.8	44.0	15.8	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**

Experiments

- **Ablation studies:** both **self-training** and **SRE loss** improve performance



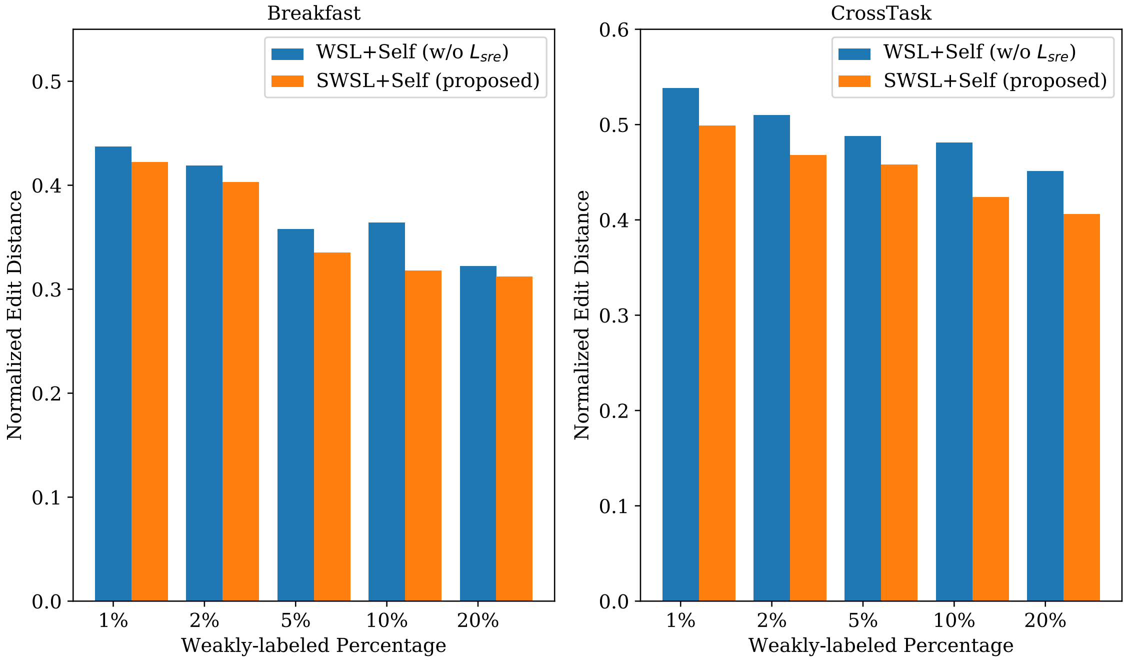
Using MuCon as the network

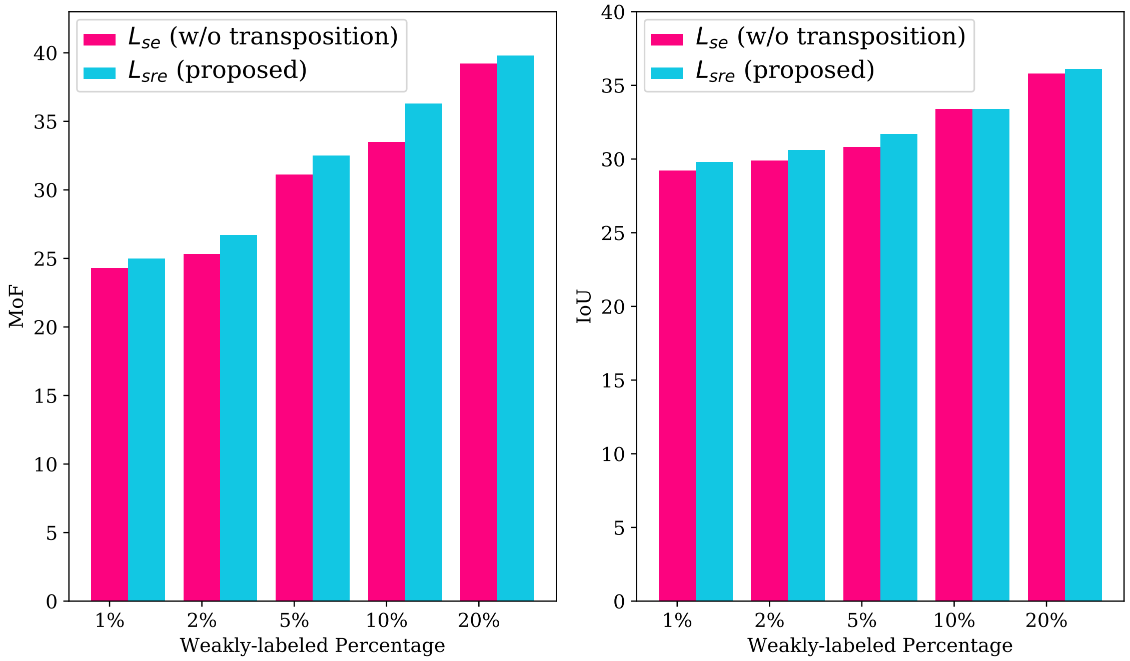


Using CDFL as the network

Experiments

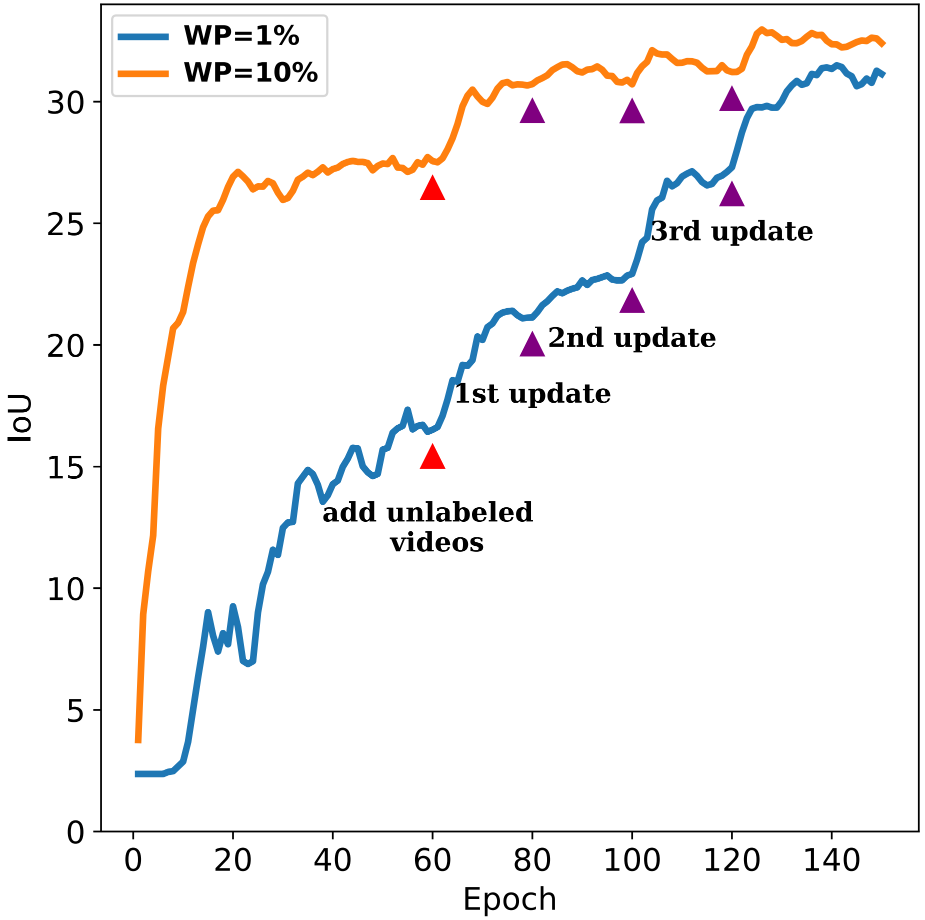
- **Effects of SRE Loss:** predict more accurate transcripts
- **Comparison between SRE Loss and SE Loss** (without adjacent transposition): more flexible transcripts, higher performance



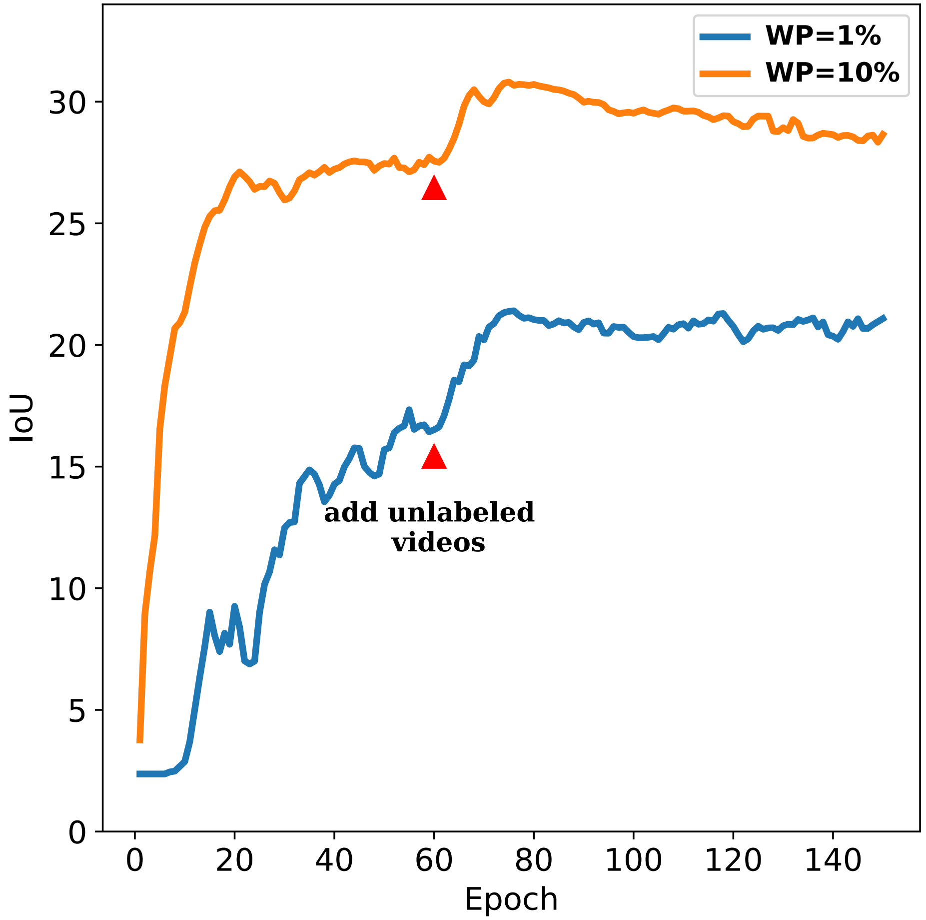


Experiments

- **Training Progress:** performance gain after each update



SWSL+Self (Proposed)



SWSL (w/o self training)

Thanks!