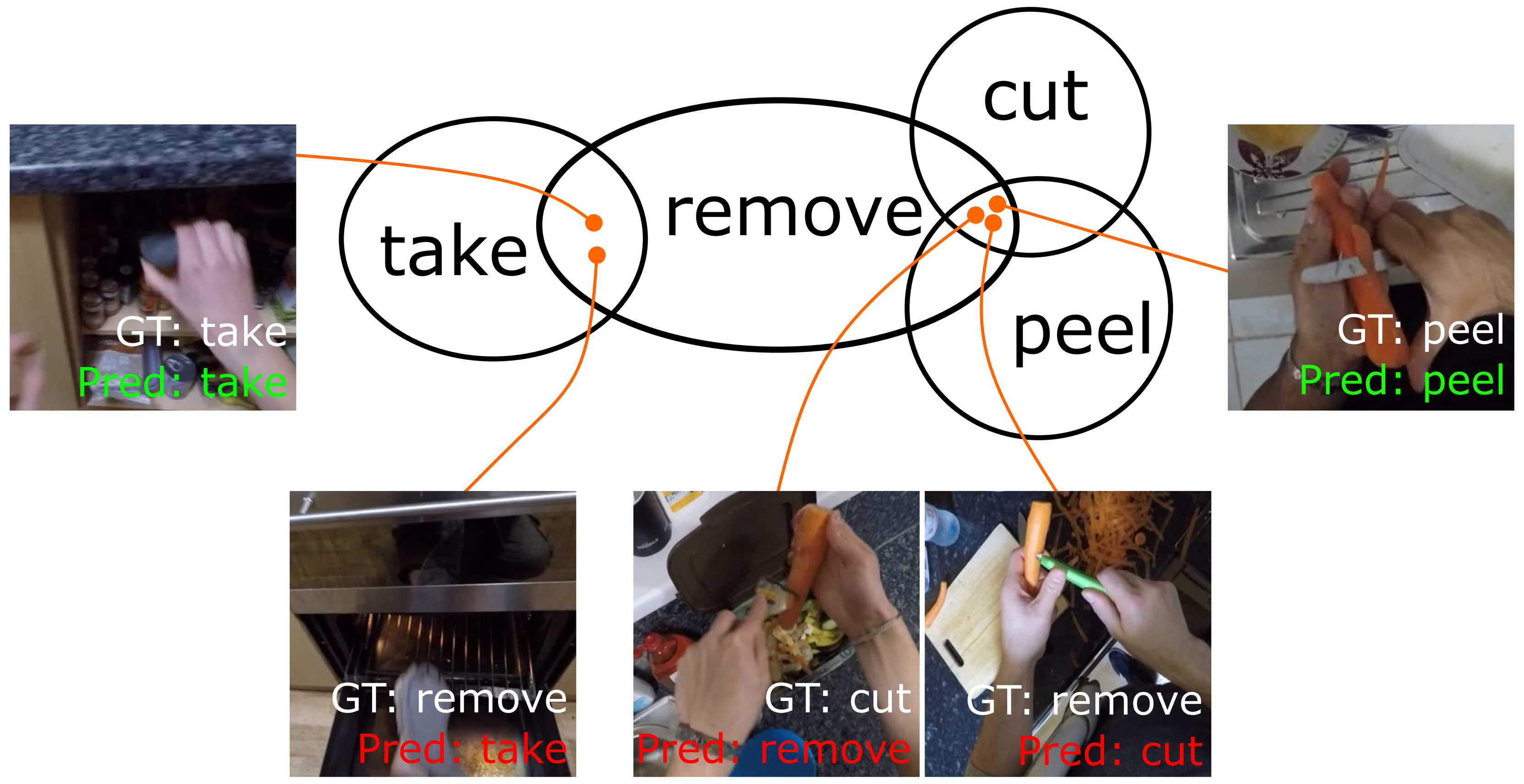
An Action Is Worth Multiple Words: Handling Ambiguity in Action Recognition

Kiyoon Kim, Davide Moltisanti, Oisin Mac Aodha, Laura Sevilla-Lara

THE UNIVERSITY of EDINBURGH

Can You Define an Action with One Verb?



- In comparison to nouns, annotators typically lack a consensus as to what constitutes an action.
- As a result, we often see interchangable labels in action datasets. More than one label can be true.
- This leads to issues in training and testing.

Single Positive Multi-label Learning!

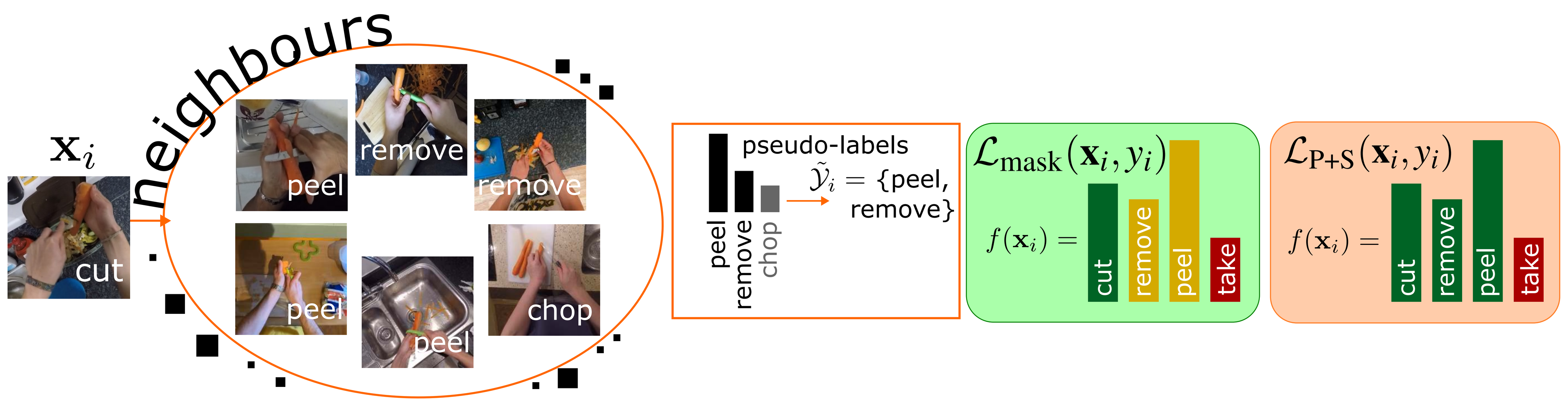
We try to disambiguate the confusing label space by generating pseudo-labels from visually similar instances that are labelled differently.

$$\mathcal{L}_{\text{AN}}(\mathbf{x}_i, y_i) = -\frac{1}{C}\sum_{c=1}^{C}\left[\mathbb{1}_{[y_i=c]}\log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i\neq c]}\log\left(1-f^{(c)}(\mathbf{x}_i)\right)\right]$$

Ours

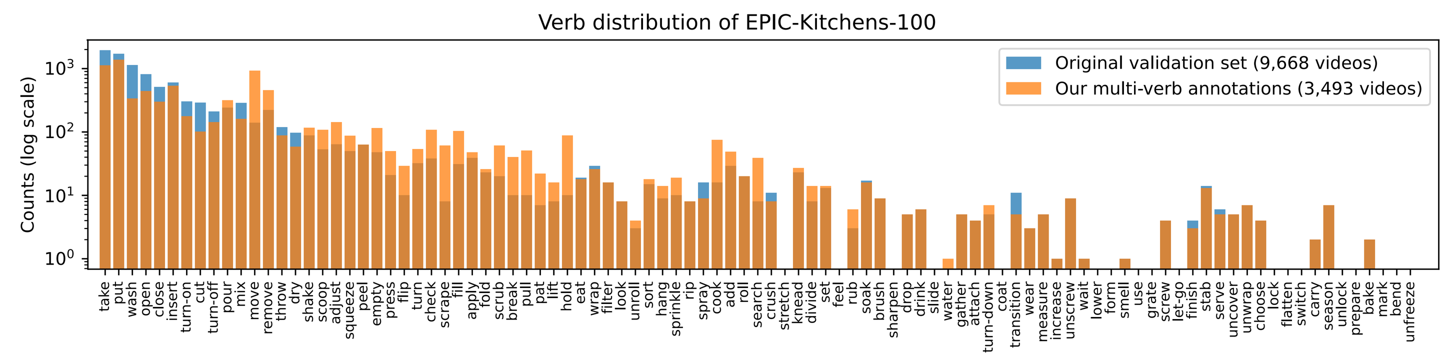
$$\mathcal{L}_{\text{mask}}(\mathbf{x}_i, y_i) = -\frac{1}{C-|\tilde{\mathcal{Y}}_i|}\sum_{\substack{c=1\\ c\notin\tilde{\mathcal{Y}}_i}}^{C}\left[\mathbb{1}_{[y_i=c]}\log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i\neq c]}\log\left(1-f^{(c)}(\mathbf{x}_i)\right)\right]$$

$$\mathcal{L}_{\text{P+S}}(\mathbf{x}_i, y_i) = -\frac{1}{C}\sum_{c=1}^{C}\Big[\mathbb{1}_{[y_i=c]}\log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[c\in\tilde{\mathcal{Y}}_i]}\log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i\neq c]}\mathbb{1}_{[c\notin\tilde{\mathcal{Y}}_i]}\log\left(1-f^{(c)}(\mathbf{x}_i)\right)\Big]$$



EPIC-Kitchens-100-SPMV (Our test set annotations)

- ~40% of the official validation set
- 3,493 videos
- 2.4 labels per video on average
- 3 annotators per video



Experiment Results

Dataset	Loss	Top-set ML	Top-1 ML	IOU Acc.	F_1	mAP
EPIC-Kitchens-100-SPMV	AN	43.7 ± 0.5	51.0 ± 0.4	11.2 ± 0.4	15.0 ± 0.6	22.8 ± 1.8
	WAN	44.8 ± 0.3	52.5 ± 0.6	15.2 ± 5.0	24.6 ± 6.6	**26.3 ± 1.3**
	LS	43.7 ± 0.9	51.8 ± 0.7	9.6 ± 0.4	12.9 ± 0.6	24.4 ± 1.5
	N-LS	43.4 ± 0.6	50.7 ± 0.4	11.0 ± 0.4	14.8 ± 0.5	22.1 ± 0.8
	Focal	43.3 ± 0.6	51.5 ± 0.3	5.7 ± 0.2	7.8 ± 0.2	23.9 ± 0.6
	EM	44.7 ± 0.4	52.9 ± 0.3	4.1 ± 0.0	7.8 ± 0.0	25.1 ± 1.0
	Mask (ours)	46.6 ± 0.2	55.2 ± 0.4	27.8 ± 0.4	36.9 ± 0.4	25.9 ± 0.8
	P+S (ours)	**46.9 ± 0.1**	**56.0 ± 0.6**	**33.5 ± 0.2**	**44.9 ± 0.3**	25.8 ± 0.8
Confusing-HMDB-102	AN	32.0 ± 1.6	38.5 ± 2.0	18.9 ± 1.6	24.8 ± 1.6	20.6 ± 3.8
	WAN	36.8 ± 0.7	40.7 ± 0.6	4.1 ± 0.1	7.9 ± 0.1	32.0 ± 1.4
	LS	32.4 ± 2.3	38.8 ± 1.8	19.3 ± 2.3	24.9 ± 2.3	19.7 ± 3.8
	N-LS	32.2 ± 1.3	38.8 ± 1.7	19.3 ± 1.7	25.3 ± 1.8	20.1 ± 3.7
	Focal	31.6 ± 1.9	38.0 ± 2.0	13.0 ± 0.8	17.6 ± 0.7	14.8 ± 3.8
	EM	31.9 ± 0.6	37.4 ± 0.9	3.2 ± 0.0	6.2 ± 0.1	18.6 ± 3.1
	Mask (ours)	41.8 ± 1.1	43.3 ± 0.9	**30.8 ± 2.5**	**36.3 ± 2.7**	40.3 ± 2.2
	P+S (ours)	**41.9 ± 0.9**	**43.4 ± 0.5**	29.9 ± 2.2	35.9 ± 2.0	**40.7 ± 2.0**

Class-level pseudo label (†) fails and we need instance-level pseudo labels.

Dataset	Loss	Top-set ML	Top-1 ML	IOU Acc.	F_1	mAP
EPIC-Kitchens-100-SPMV	Mask	46.6 ± 0.2	55.2 ± 0.4	27.8 ± 0.4	36.9 ± 0.4	**25.9 ± 0.8**
	Mask†	37.8 ± 1.3	48.4 ± 1.0	18.1 ± 0.4	23.7 ± 0.4	21.7 ± 0.9
	P+S	**46.9 ± 0.1**	**56.0 ± 0.6**	**33.5 ± 0.2**	**44.9 ± 0.3**	25.8 ± 0.8
	P+S†	23.0 ± 0.4	28.0 ± 0.5	12.4 ± 0.4	18.0 ± 0.6	20.8 ± 1.3
Confusing-HMDB-102	Mask	41.8 ± 1.1	43.3 ± 0.9	30.8 ± 2.5	**36.3 ± 2.7**	40.3 ± 2.2
	Mask*	42.6 ± 2.2	43.8 ± 2.3	30.4 ± 2.3	34.2 ± 2.4	37.4 ± 3.6
	P+S	41.9 ± 0.9	43.4 ± 0.5	29.9 ± 2.2	35.9 ± 2.0	**40.7 ± 2.0**
	P+S*	**43.2 ± 1.8**	**44.3 ± 2.1**	**31.4 ± 2.5**	35.9 ± 2.1	39.1 ± 3.1

Perfect pseudo label (*) gives minor performance improvement.