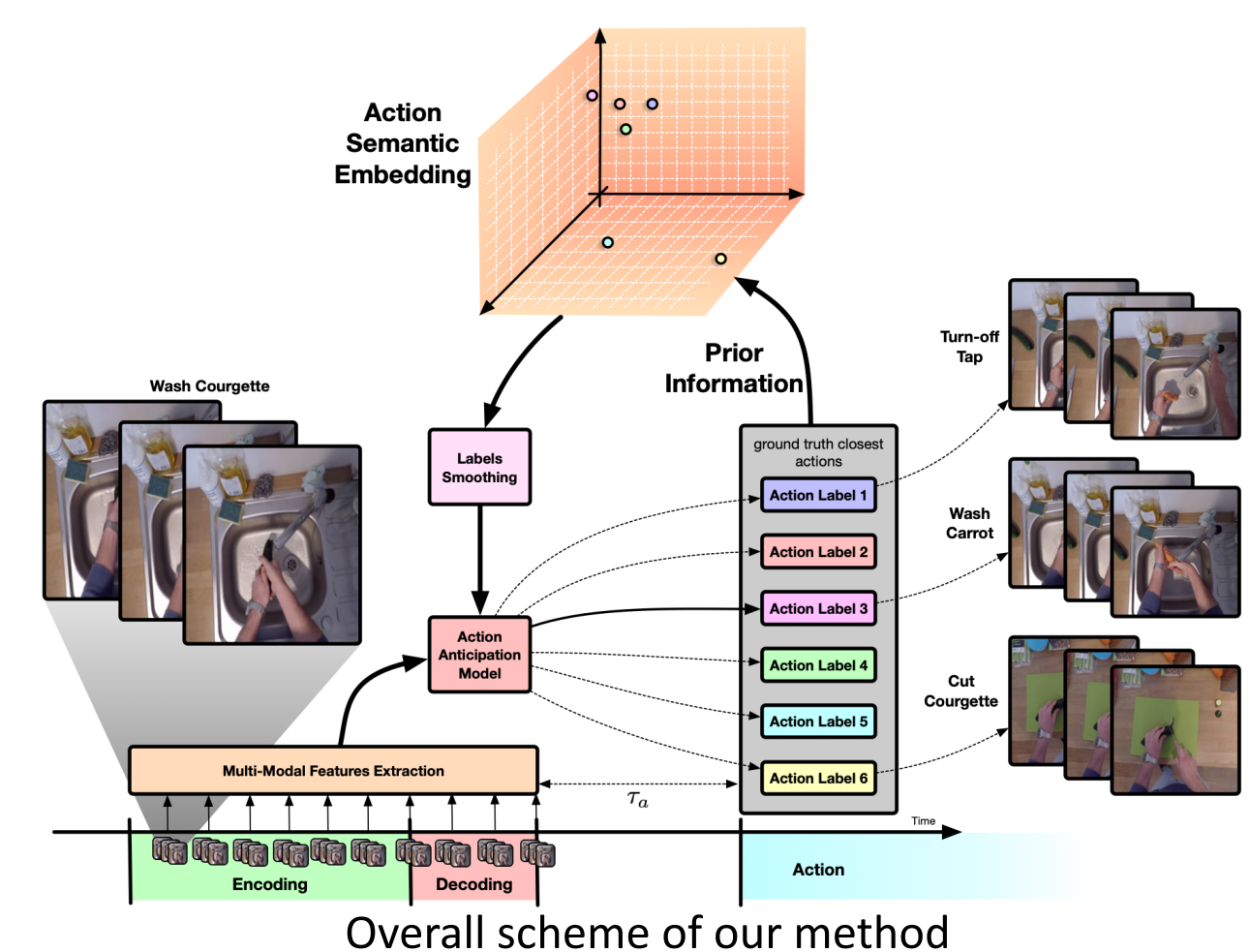
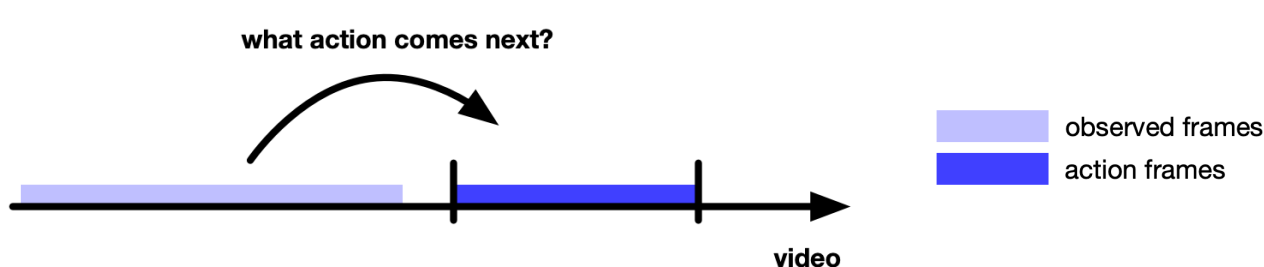


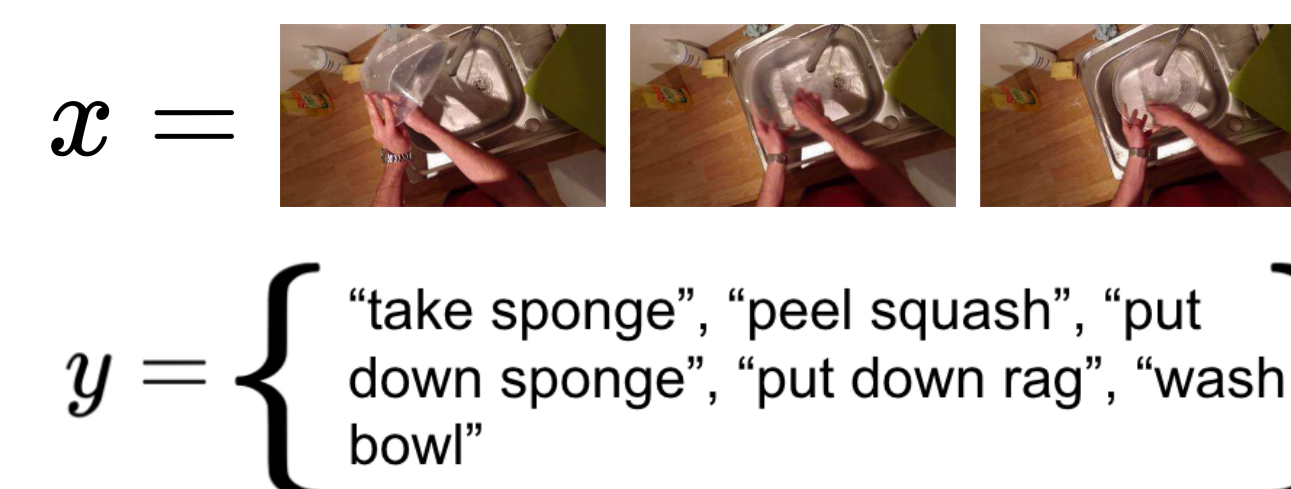
Motivation



In this work, we address the problem of anticipating egocentric human actions in an indoor scenario at several time steps. More specifically, we aim to anticipate an action by leveraging its previous video segments.



Egocentric videos exhibit an inherent uncertainty when dealing with predicting future actions. In fact, given the current state observation of an action there can be multiple, but still plausible, future scenarios that can occur. For this reason, the problem can be reformulated as a multi-label task with missing labels where, from a set of valid future realisations, only one is sampled.



The contributions of our work are the following:

- We generalise the label smoothing idea extrapolating semantic priors from the action labels to capture the multi-modal future component of the action anticipation problem,

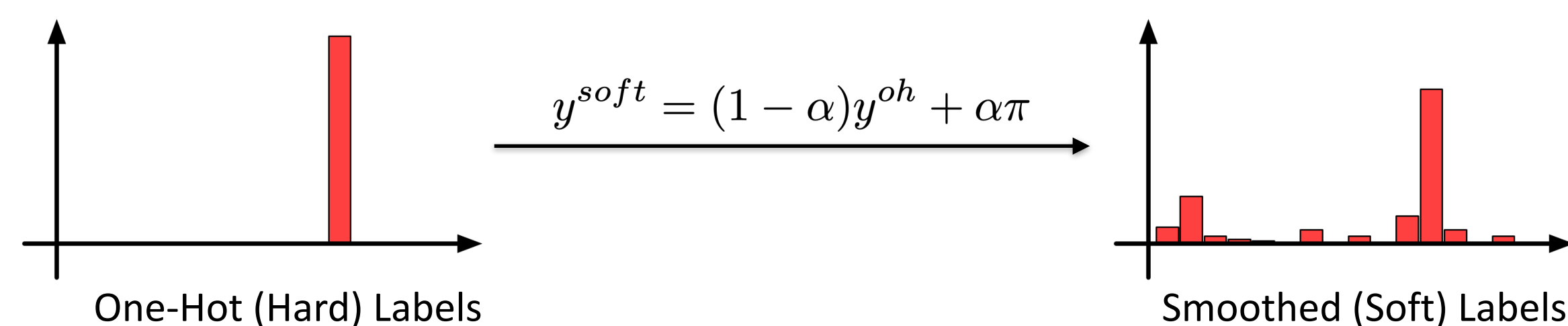
- We show that label smoothing, in this context, can be seen as a knowledge distillation process,

- We show that with our simple method we can systematically improve results of state-of-the-art models on action anticipation.

Knowledge Distillation Via Label Smoothing

All previous models designed for action anticipation are trained with cross-entropy using one-hot labels, leveraging only one of the possible future scenarios as ground truth. To overcome this issue, we smooth the target distribution enabling the chance of negative (yet still plausible) classes to be selected.

We generalise the label smoothing procedure using a custom prior distribution:



This simple method can be seen also as a knowledge distillation procedure:

$$CE[y^{soft}, p] = - \sum_i y^{soft}(i) \log p(i) = (1 - \alpha) \boxed{CE[y^{oh}, p]} + \alpha \boxed{CE[\pi, p]}$$

loss between prediction and one-hot loss between prediction and one-hot knowledge distillation term

Which Priors for Label Smoothing?

Verb-Noun Label Smoothing:

$$\mathcal{A}_v(\bar{v}) = \{(\bar{v}, n) \in \mathcal{A} \mid \forall n \in \mathcal{N}\}$$

$$\mathcal{A}_n(\bar{n}) = \{(v, \bar{n}) \in \mathcal{A} \mid \forall v \in \mathcal{V}\}$$

$$\pi_{VN}^{(k)}(i) = \mathbb{1} \left[a^{(i)} \in \mathcal{A}_v(v^{(k)}) \cup \mathcal{A}_n(n^{(k)}) \right] \frac{1}{C_k}$$

Temporal Label Smoothing:

$$\pi_{TE}^{(k)}(i) = \frac{Occ[a^{(i)} \rightarrow a^{(k)}]}{\sum_j Occ[a^{(j)} \rightarrow a^{(k)}]}$$

GloVe Label Smoothing:

$$\phi^{(k)} = Concat \left[GloVe(v^{(k)}), GloVe(n^{(k)}) \right]$$

$$\pi_{GL}^{(k)}(i) = \frac{|\phi^{(k)T} \phi^{(i)}|}{\sum_j |\phi^{(k)T} \phi^{(j)}|}$$

With this prior we reward all the actions that share either the same noun or the same verb with the ground truth.

Using such representation, we reward both the correct class and most frequent actions that precede the ground truth.

With this prior we reward not only the the correct class but also all other *similar* actions.

Results

EPIC-Kitchens-55

| | Top-5 Action Accuracy % @ different anticipation times [s] | | | | | | | |
|--------------------------------|--|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | 2 | 1.75 | 1.5 | 1.25 | 1 | 0.75 | 0.5 | 0.25 |
| LSTM One-hot Encoding | 27.71 ± 0.33 | 28.69 ± 0.34 | 29.84 ± 0.24 | 30.90 ± 0.48 | 31.93 ± 0.45 | 33.14 ± 0.36 | 34.10 ± 0.44 | 35.16 ± 0.35 |
| LSTM TE Smoothing | 27.94 ± 0.24 | 28.90 ± 0.27 | 30.06 ± 0.24 | 31.13 ± 0.19 | 32.19 ± 0.28 | 33.21 ± 0.36 | 34.17 ± 0.37 | 35.10 ± 0.25 |
| LSTM Uniform Smoothing | 28.16 ± 0.27 | 29.06 ± 0.26 | 30.23 ± 0.24 | 31.25 ± 0.27 | 32.41 ± 0.28 | 33.64 ± 0.27 | 34.69 ± 0.19 | 35.75 ± 0.13 |
| LSTM VN Smoothing | 28.43 ± 0.30 | 29.41 ± 0.31 | 30.68 ± 0.28 | 31.85 ± 0.21 | 33.08 ± 0.18 | 34.35 ± 0.19 | 35.38 ± 0.34 | 36.46 ± 0.26 |
| LSTM GL Smoothing | 28.61 ± 0.26 | 29.87 ± 0.25 | 30.97 ± 0.34 | 31.94 ± 0.34 | 33.12 ± 0.36 | 34.40 ± 0.37 | 35.51 ± 0.37 | 36.87 ± 0.25 |
| LSTM GL+VN Smoothing | 28.88 ± 0.20 | 29.94 ± 0.19 | 31.23 ± 0.32 | 32.54 ± 0.31 | 33.56 ± 0.28 | 34.92 ± 0.25 | 36.06 ± 0.33 | 37.29 ± 0.30 |
| Improv. | +1.17 | +1.25 | +1.39 | +1.64 | +1.63 | +1.78 | +1.96 | +2.13 |
| RU-LSTM | 29.44 | 30.73 | 32.24 | 33.41 | 35.32 | 36.34 | 37.37 | 38.39 |
| RU-LSTM GL+VN Smoothing | 30.37 | 31.64 | 33.17 | 34.86 | 35.90 | 37.07 | 38.96 | 39.74 |
| Improv. | +0.93 | +0.91 | +0.93 | +1.45 | +0.58 | +0.73 | +1.59 | +1.35 |




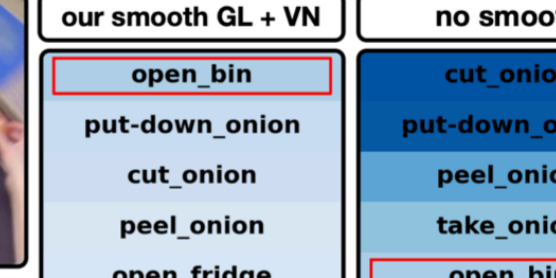

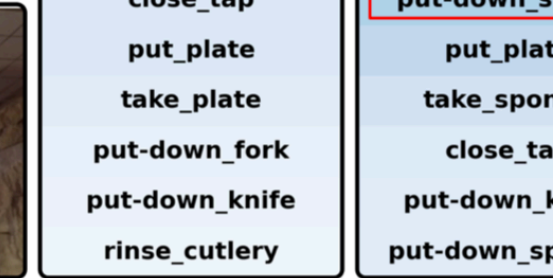

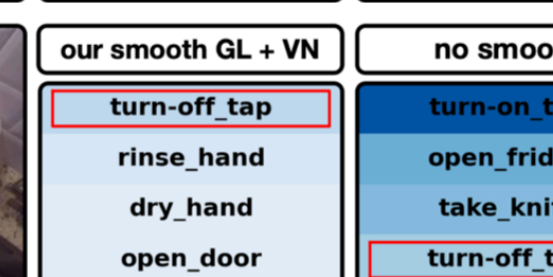

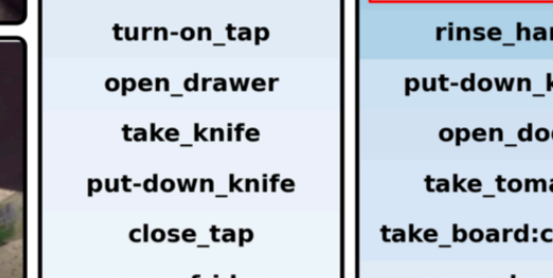

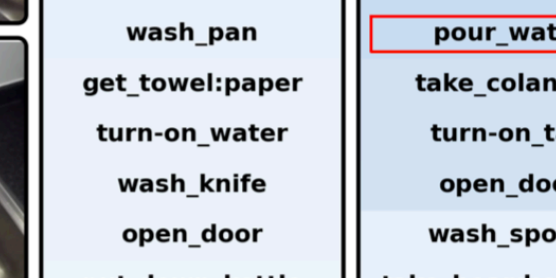
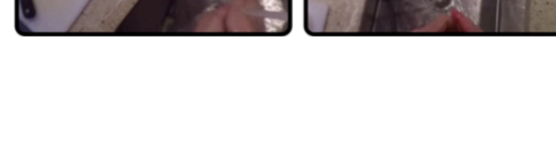
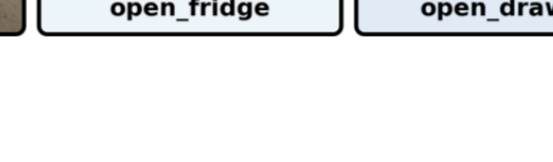
EGTEA GAZE+

| | Top-5 Action Accuracy % @ different anticipation times [s] | | | | | | | |
|-----------------------------|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | 2 | 1.75 | 1.5 | 1.25 | 1 | 0.75 | 0.5 | 0.25 |
| LSTM One-hot Encoding | 55.94 | 58.75 | 60.94 | 63.02 | 65.78 | 68.04 | 71.55 | 73.94 |
| LSTM TE Smoothing | 56.13 | 58.93 | 61.17 | 63.24 | 66.00 | 68.20 | 71.75 | 74.20 |
| LSTM Uniform Smoothing | 56.35 | 59.20 | 61.37 | 63.36 | 66.12 | 68.41 | 71.95 | 74.35 |
| LSTM VN Smoothing | 56.85 | 59.67 | 61.64 | 64.03 | 66.81 | 68.94 | 72.56 | 75.44 |
| LSTM GL Smoothing | 57.34 | 60.11 | 62.25 | 64.42 | 67.21 | 69.56 | 73.02 | 75.83 |
| Improv. | +1.4 | +1.36 | +1.31 | +1.4 | +1.43 | +1.52 | +1.47 | +1.89 |
| RU-LSTM | 56.82 | 59.13 | 61.42 | 63.53 | 66.40 | 68.41 | 71.84 | 74.28 |
| RU-LSTM GL Smoothing | 59.99 | 62.02 | 63.95 | 66.47 | 68.74 | 72.16 | 75.21 | 78.11 |
| Improv. | +3.17 | +2.89 | +2.53 | +2.94 | +2.34 | +3.75 | +3.37 | +3.83 |

Best Prior Distribution for Label Smoothing:

| | Top-5 Action Accuracy % @ 1 [s] | | | | |
|---------------------|---------------------------------|--------------|-------------------|--------------|--------------|
| | One-hot | TE Smoothing | Uniform Smoothing | VN Smoothing | GL Smoothing |
| LSTM (RGB-only) | 29.54 | 29.60 | 29.76 | 29.83 | 29.85 |
| LSTM (Flow-only) | 20.43 | 20.53 | 20.65 | 20.69 | 20.77 |
| LSTM (OBJ-only) | 29.5 | 29.64 | 29.69 | 29.79 | 29.82 |
| RU-LSTM (RGB-only) | 30.83 | 30.95 | 31.00 | 31.05 | 31.19 |
| RU-LSTM (Flow-only) | 21.42 | 21.51 | 21.51 | 21.63 | 21.73 |
| RU-LSTM (OBJ-only) | 29.89 | 29.99 | 30.07 | 30.04 | 30.19 |

Qualitative Results: Top-10 predictions

| | | | | | | | |
|---|---|--------------------|--------------------|---|---|--------------------|--------------------|
|  |  | our smooth GL + VN | no smooth |  |  | our smooth GL + VN | no smooth |
| | | put-down_spoon | wash_plate | | | open_bin | cut_onion |
| | | wash_plate | take_plate | | | put-down_onion | put-down_onion |
| | | open_tap | wash_spoon | | | cut_onion | peel_onion |
|  |  | wash_spoon | open_tap | | | peel_onion | take_onion |
| | | close_tap | put-down_spoon | | | open_fridge | open_bin |
| | | put_plate | take_sponge | | | throw_skin | open_fridge |
| | | put-down_fork | close_tap | | | close_bin | transfer_onion |
|  |  | put-down_knife | put-down_knife | | | throw_rubbish | open_drawer |
| | | rinse_cutlery | put-down_sponge | | | open_drawer | open_door |
| | | | | | | | |
| | | | | | | | |
|  |  | our smooth GL + VN | no smooth |  |  | our smooth GL + VN | no smooth |
| | | turn-off_tap | turn-on_tap | | | put-down_pan | put-down_pan |
| | | rinse_hand | open_fridge | | | wash_pan | wash_pan |
| | | dry_hand | take_knife | | | take_pan | take_pan |
|  |  | open_door | turn-off_tap | | | take_sponge | take_sponge |
| | | turn-on_tap | rinse_hand | | | pour_water | pour_water |
| | | open_drawer | put-down_knife | | | take_colander | take_colander |
| | | take_knife | open_door | | | turn-on_tap | turn-on_tap |
| | | put-down_knife | take_tomato | | | wash_knife | wash_knife |
| | | close_tap | take_board:cutting | | | open_door | open_door |
| | | open_fridge | open_drawer | | | put-down_kettle | take_board:cutting |
| | | | | | | | |