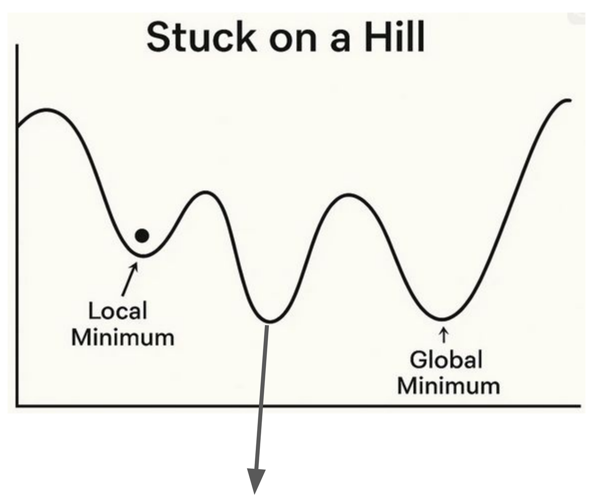
# Learning Probabilistic Embeddings for Unsupervised Action Segmentation

Shuai Li, Duc Manh Vu, Juergen Gall
Computer Vision Group, University of Bonn

ECCV 2026 | MALMÖ | SEPT 8-12

## Motivation

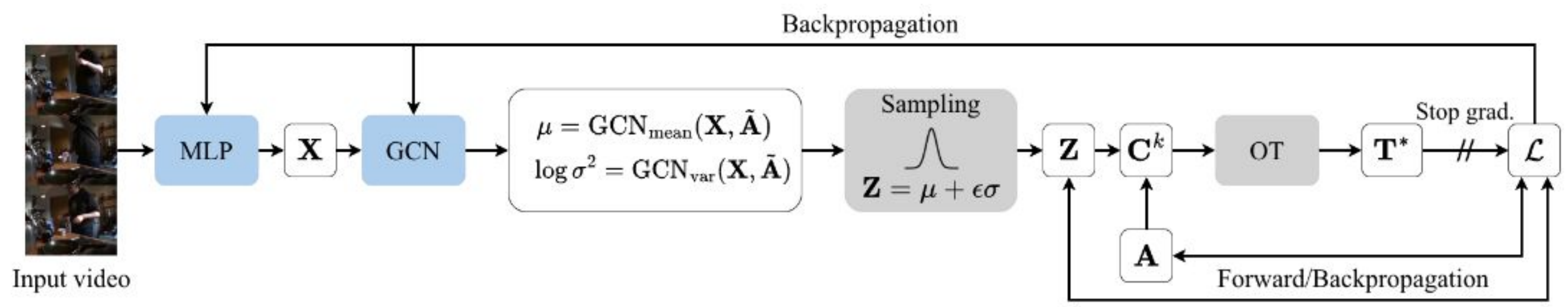
1. SOTA methods follow EM paradigm: OT as Pseudo-labels, learning representations given pseudo labels.
2. Deterministic embeddings tend to overfit the wrong pseudo labels, and stuck in poor local optimums.




A better local optimum.

We want to get out of poor local optimums and find "Better" local optimums during training.

## Methodology




During training

1. Use OT to generate pseudo labels.
2. Reparametrization-trick enables end-to-end training with uncertainty-based cross-entropy loss.

During inference

3. Take the mean of Gaussian as the embedding, use OT for segmentation.
4. Hungarian matching to tag semantic labels for the inferred segmentations.

$$\min_{\mathbf{T}\in\mathcal{T}} \alpha\mathcal{F}_{\mathrm{GW}}(\mathbf{C}^v, \mathbf{C}^a, \mathbf{T}) + (1-\alpha)\mathcal{F}_{\mathrm{KOT}}(\mathbf{C}^k, \mathbf{T}) + \lambda D_{\mathrm{KL}}(\mathbf{T}^\top \mathbf{1}_N \,\|\, \mathbf{q}) - \epsilon H(\mathbf{T}),$$

$$\mathcal{L}_{\mathrm{uncer}} \approx -\frac{1}{MN}\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{K} t_{ij}^{(m)} \log p_{ij}^{(m)}$$

## Quantitative results

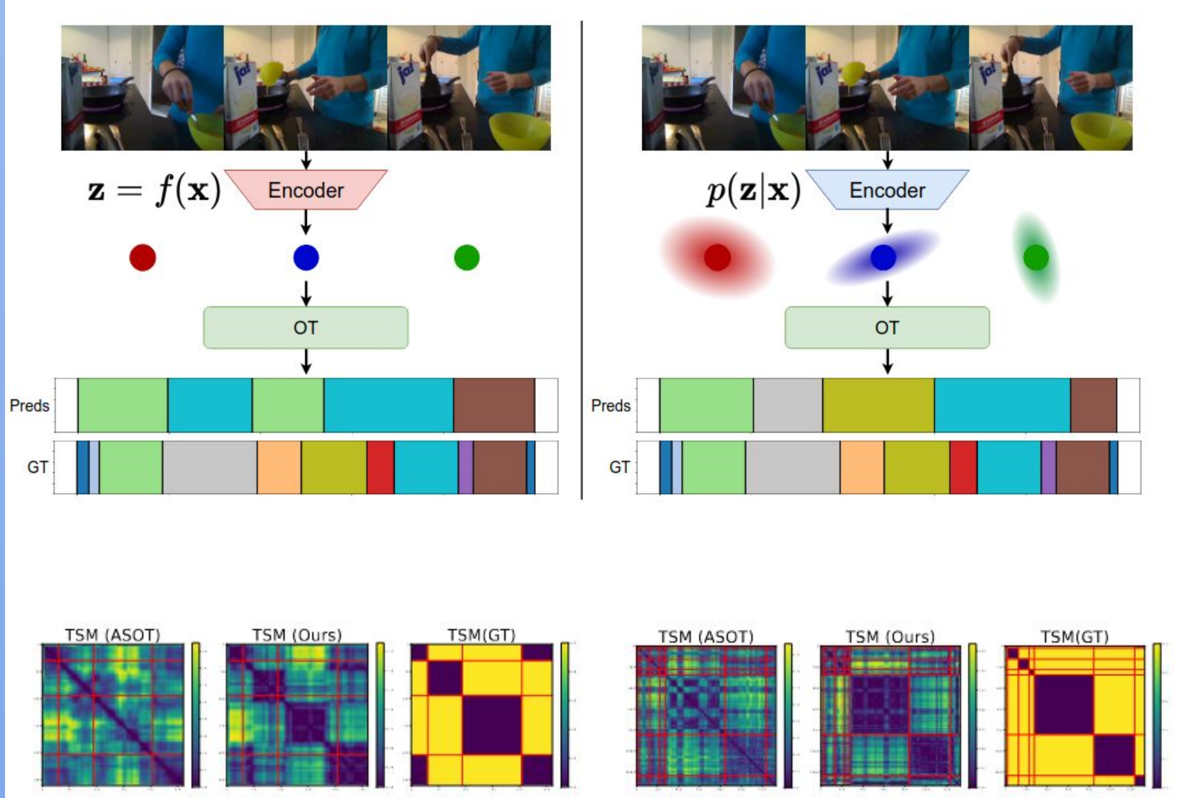
| Methods | Breakfast | | | YTI | | | 50Salads (Eval) | | | DA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MoF | F1 | mIoU | MoF | F1 | mIoU | MoF | F1 | mIoU | MoF | F1 | mIoU |
| CTE [16] | 41.8 | 26.4 | - | 39.0 | 28.3 | - | 35.5 | - | - | 47.6 | 44.9 | - |
| VTE [34] | 48.1 | - | - | - | 29.9 | - | 30.6 | - | - | - | - | - |
| UDE [30] | 47.4 | 31.9 | - | 43.8 | 29.6 | - | 42.2 | 34.4 | - | - | - | - |
| ASAL [18] | 52.5 | 37.9 | - | 44.9 | 32.1 | - | 39.2 | - | - | - | - | - |
| TOT [17] | 47.5 | 31.0 | - | 40.6 | 30.0 | - | 47.4 | 42.8 | - | 56.3 | 51.7 | - |
| TOT+ [17] | 39.0 | 30.3 | - | 45.3 | 32.9 | - | 44.5 | 48.2 | - | 58.1 | 53.4 | - |
| UFSA [32] | 52.1 | 38.0 | - | 49.6 | 32.4 | - | 55.8 | 50.3 | - | 65.4 | 63.0 | - |
| ASOT [38] | 56.1 | 38.3 | 18.6 | 52.9 | 32.1 | 24.7 | 59.3 | 53.6 | 30.1 | 70.4 | 68.0 | 45.9 |
| HVQ [27] | 54.4 | 39.7 | - | 50.3 | 35.1 | - | - | - | - | - | - | - |
| VASOT [2] | 57.5 | 39.0 | 18.8 | 53.2 | 35.7 | **25.2** | 60.6 | 57.4 | 34.5 | 70.9 | 75.1 | 49.3 |
| CLOT [3] | 60.1 | 40.1 | 18.5 | 54.4 | 36.7 | 23.4 | 59.4 | **63.2** | **38.8** | 68.8 | 72.6 | 48.1 |
| PEOT (Ours) | **60.7** | **40.5** | **19.0** | **55.4** | **37.4** | 22.9 | **64.9** | 58.9 | 30.2 | **71.2** | **75.8** | **51.7** |

| Methods | Breakfast | | | YTI | | | 50Salads (Eval) | | | DA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MoF | F1 | mIoU | MoF | F1 | mIoU | MoF | F1 | mIoU | MoF | F1 | mIoU |
| ASOT [38] | 56.4 | 35.7 | 17.2 | 49.0 | 34.1 | 23.7 | 59.4 | 56.7 | 25.6 | 59.0 | 63.7 | 40.6 |
| ASOT [38] + Prob. | 60.7 | 40.5 | 19.0 | 55.4 | 37.4 | 22.9 | 64.9 | 58.9 | 30.2 | 71.2 | 75.8 | 51.7 |
| VASOT [2] | 54.5 | 35.3 | 15.5 | 47.5 | 30.4 | 18.0 | 53.4 | 51.9 | 26.7 | 67.9 | 70.2 | 48.0 |
| VASOT [2] + Prob. | 57.2 | 36.1 | 15.8 | 52.5 | 32.5 | 18.8 | 62.2 | 53.6 | 26.4 | 71.1 | 76.1 | 51.8 |

1. SOTA on BF/DA, competitive on 50Salads/DA.
2. PEs is effective on top of ASOT/VASOT baselines.

## Contributions

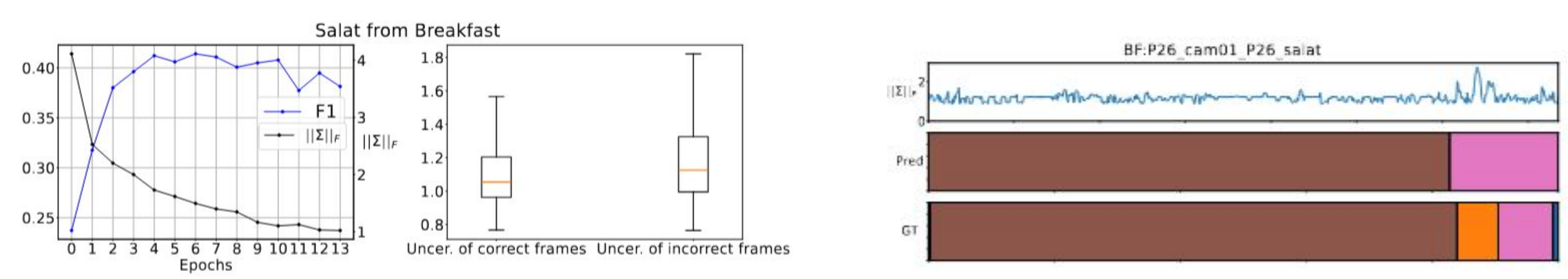
1. Novel probabilistic treatment of frame embeddings.
2. SOTA results on Breakfast, DesktopAssembly,



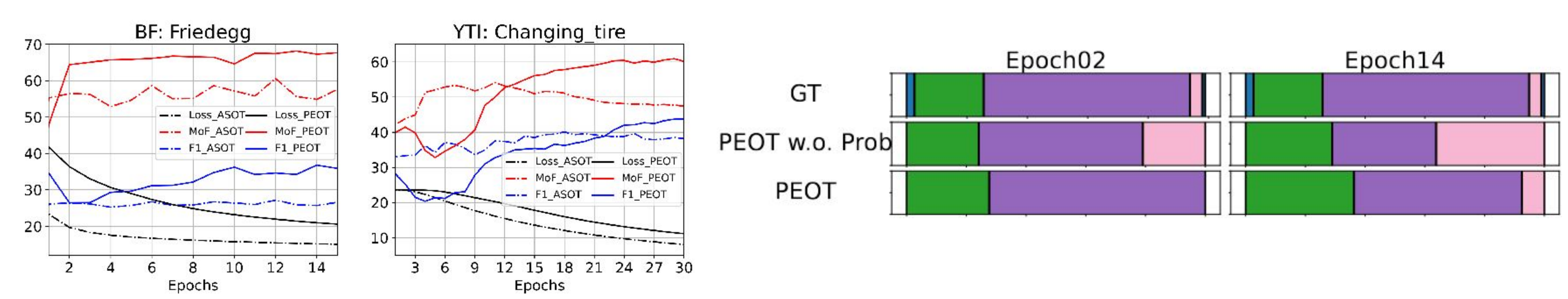

(a) Breakfast: P12_webcam01_P12_tea (b) Breakfast: P15_stereo01_P15_juice

Our method generates better features compared to ASOT baseline that benefits further action segmentation.

## Analysis of the learned uncertainty




1. Uncertainty decreases during training but converges as training finishes.
2. Uncertainty is in general higher for incorrectly predicted frames.

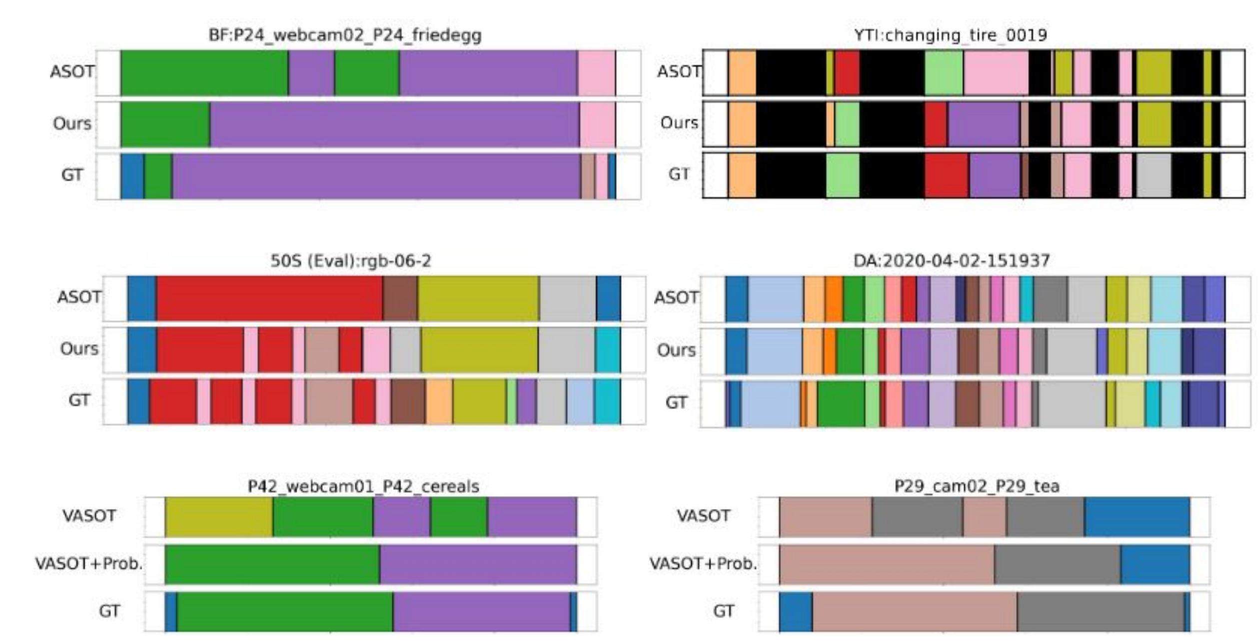



Our approach recovers missing actions as training progresses that is closer to G.T. compared to baseline using deterministic embeddings.

| Stoch. | Breakfast | | | YTI | | | 50Salads (Eval) | | | DA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MoF | F1 | mIoU | MoF | F1 | mIoU | MoF | F1 | mIoU | MoF | F1 | mIoU |
| Gauss | 58.3 | 38.7 | 18.1 | 53.5 | 34.9 | 21.3 | 61.0 | 58.8 | 30.2 | 61.9 | 63.2 | 42.0 |
| Drop | 58.5 | 40.7 | 18.4 | 48.5 | 34.3 | 22.3 | 62.0 | 58.6 | 29.6 | 65.3 | 73.8 | 48.6 |
| Ours | 60.7 | 40.5 | 19.0 | 55.4 | 37.4 | 22.9 | 64.9 | 58.9 | 30.2 | 71.2 | 75.8 | 51.7 |

Our approach works better than other stochastic embedding baselines.

## Qualitative results




Our approach works in videos containing reverse/repetitive actions, and can detect short actions, it also works in videos with large number of background frames.

## Acknowledgements

Behave erc

Code is available!