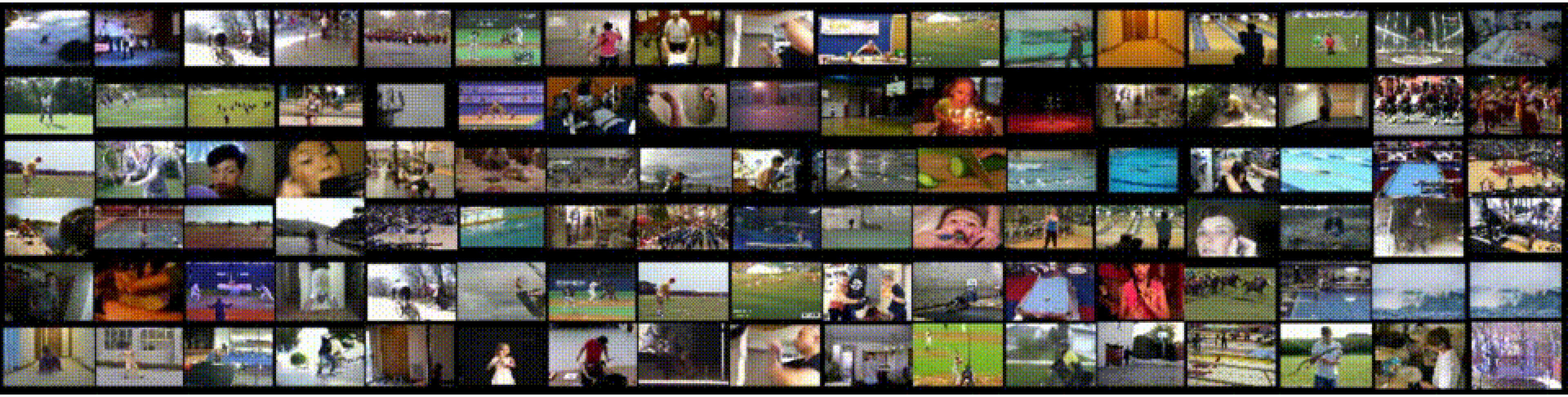
Juergen Gall

An Introduction to Temporal Action Segmentation

From Fully Supervised Learning to Weakly Supervised Learning

Action Recognition

- Large annotated datasets

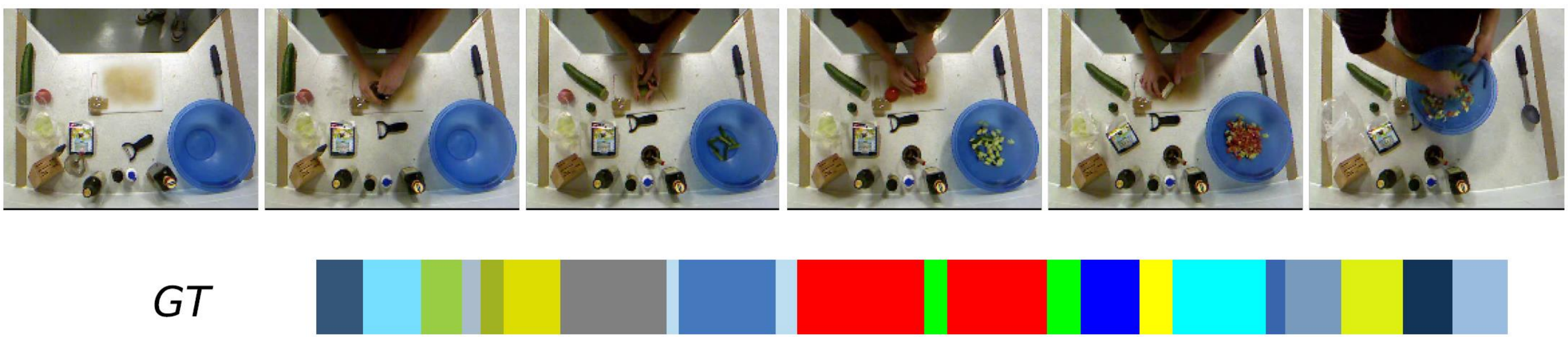


- UCF101 (98.2%), HMDB (82.5%), Kinetics-400 (82.8%), Epic-Kitchens (36.7%)
- **http://actionrecognition.net**
- Continuous data streams

Action Segmentation vs. Action Detection

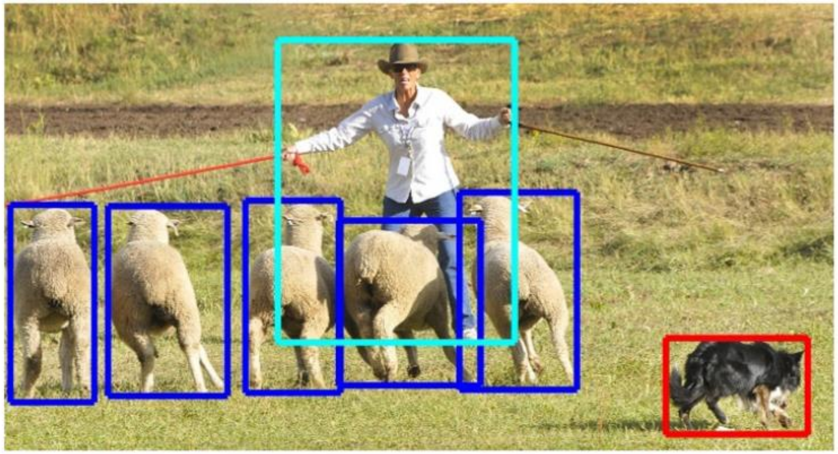
- Action Detection (THUMOS, ActivityNet)

- Action Segmentation (Breakfast, 50 Salads, GTEA)

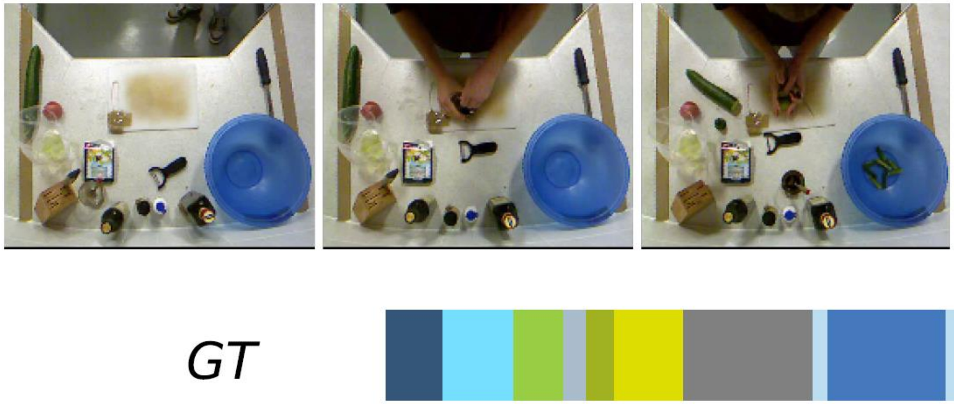


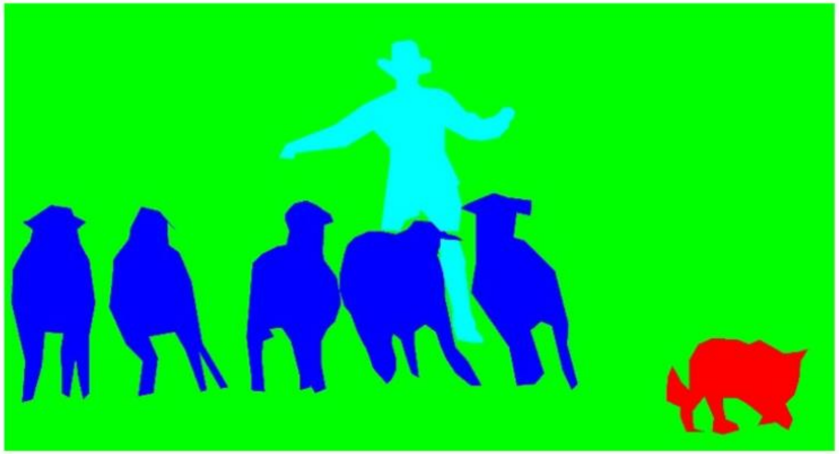
Action Segmentation vs. Action Detection

- Action Detection (Object Detection)



- Action Segmentation (Semantic Segmentation)

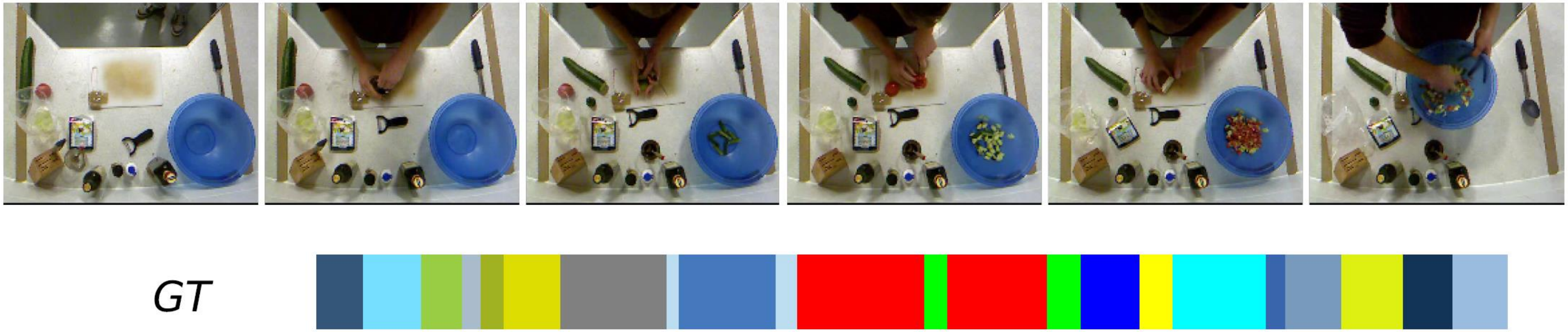




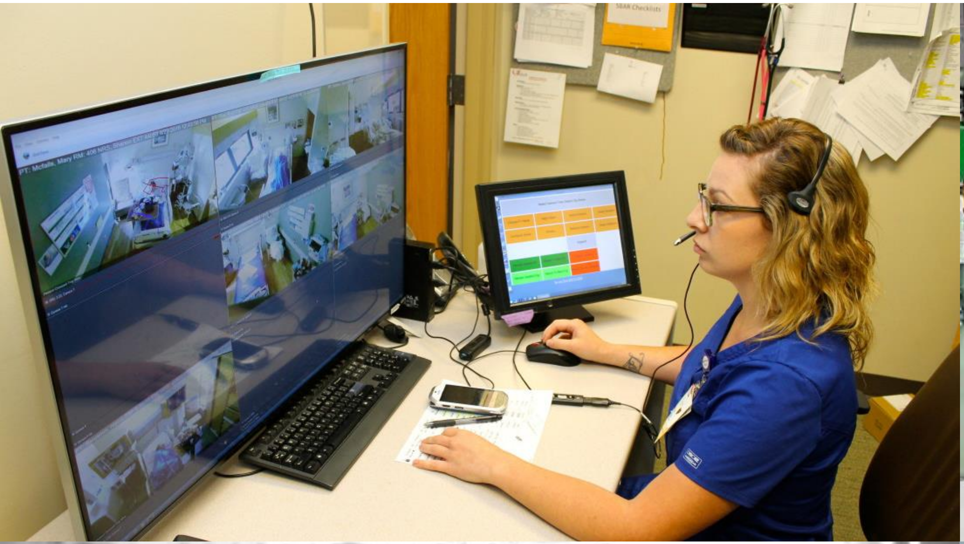
Action Segmentation vs. Action Detection

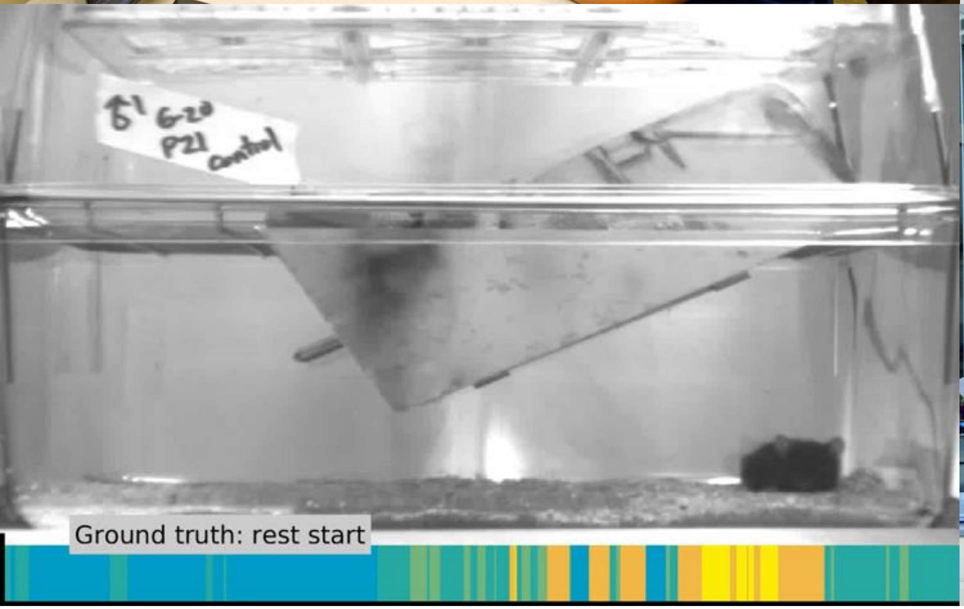
- Action Detection (THUMOS, ActivityNet)

- Action Segmentation (Breakfast, 50Salads, GTEA)



Why Action Segmentation?



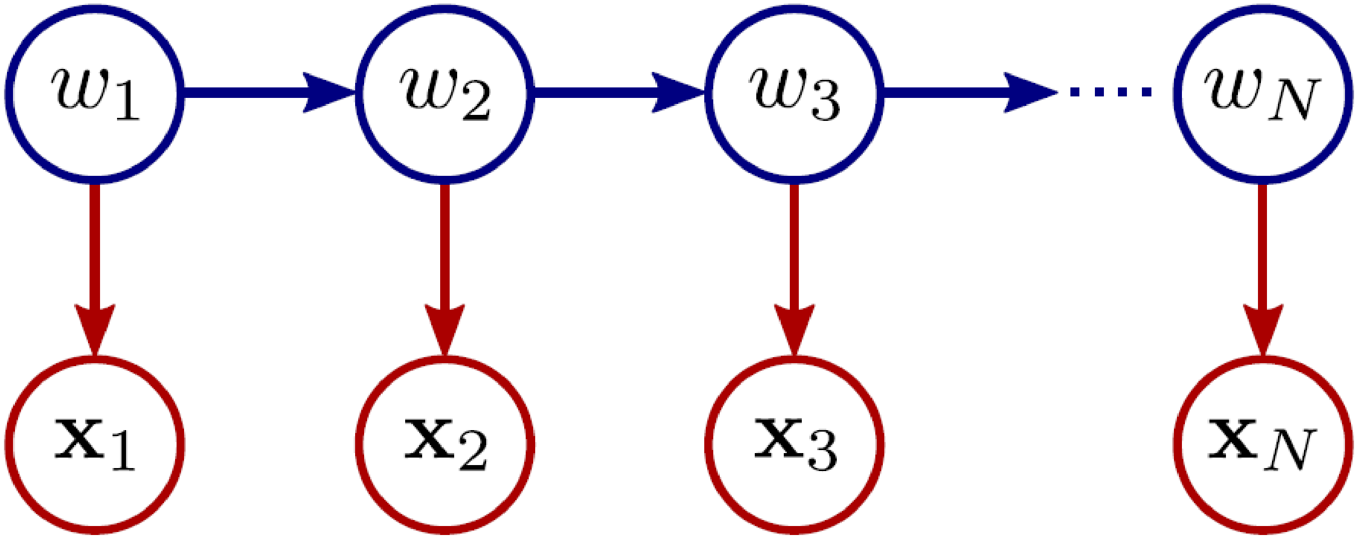


Datasets

- Breakfast
  https://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/
- 50 Salads
  https://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/
- GTEA
  http://cbs.ic.gatech.edu/fpv/#gtea
- COIN
  https://coin-dataset.github.io/

Let’s build a baseline…

Hidden Markov Model



$$Pr(\mathbf{x}_{1\ldots N}, w_{1\ldots N}) = \left(\prod_{n=1}^{N} Pr(\mathbf{x}_n | w_n)\right)\left(\prod_{n=2}^{N} Pr(w_n | w_{n-1})\right)$$

[S. Prince. **Computer Vision: Models, Learning, and Inference.** Cambridge University Press]

Simon J.D. Prince

Inference

HMM:

$$Pr(\mathbf{x}_{1\ldots N}, w_{1\ldots N}) = \left(\prod_{n=1}^{N} Pr(\mathbf{x}_n|w_n)\right)\left(\prod_{n=2}^{N} Pr(w_n|w_{n-1})\right)$$

MAP inference:

$$\begin{aligned}\hat{w}_{1\ldots N} &= \underset{w_{1\ldots N}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots N}, w_{1\ldots N})\right] \\ &= \underset{w_{1\ldots N}}{\operatorname{argmin}}\left[-\log\left[Pr(\mathbf{x}_{1\ldots N}, w_{1\ldots N})\right]\right]\end{aligned}$$

Substituting:

$$\hat{w}_{1\ldots N} = \underset{w_{1\ldots N}}{\operatorname{argmin}}\left[-\sum_{n=1}^{N}\log\left[Pr(\mathbf{x}_n|w_n)\right] - \sum_{n=2}^{N}\log\left[Pr(w_n|w_{n-1})\right]\right]$$

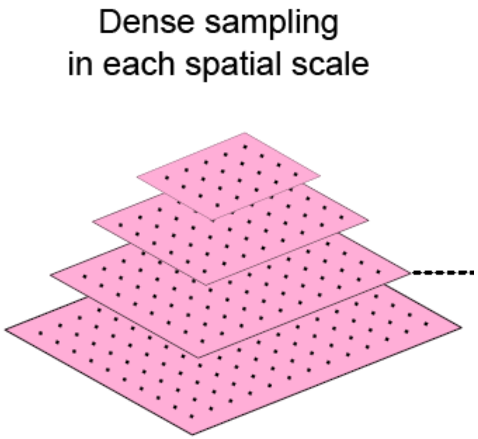
Global minimum by dynamic programming

[S. Prince. **Computer Vision: Models, Learning, and Inference.** Cambridge University Press]

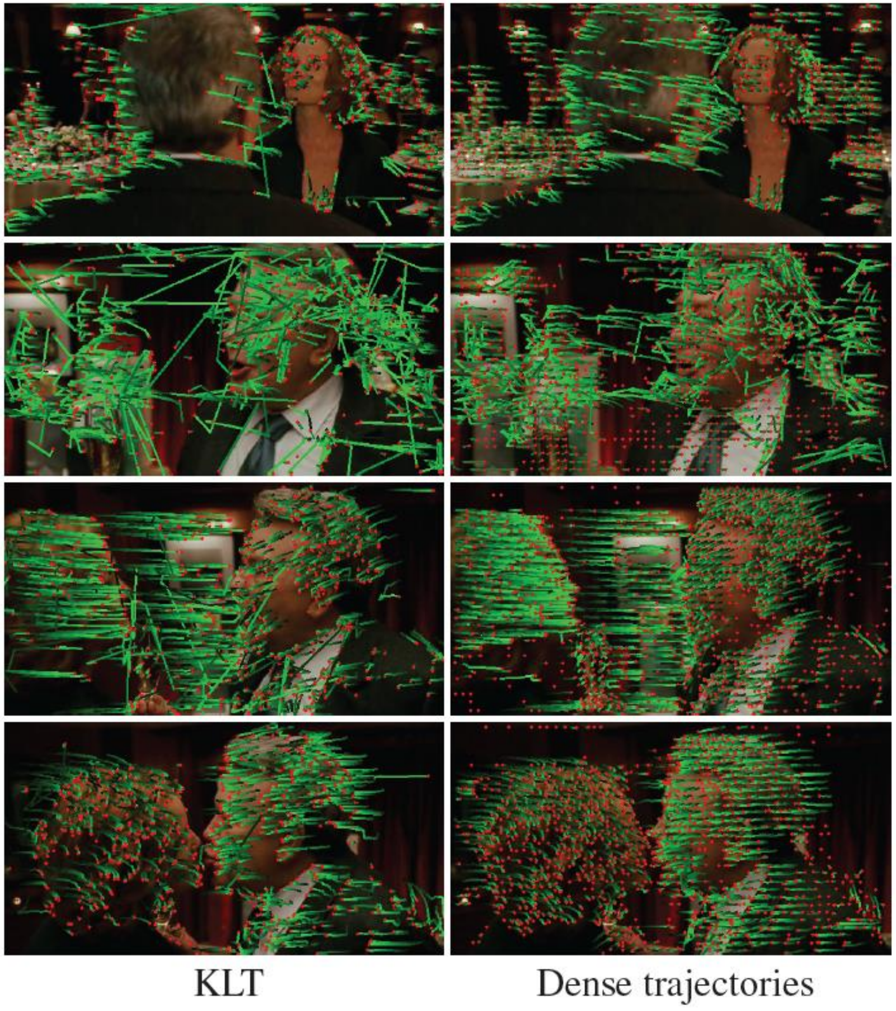
Simon J.D. Prince

Features: Dense Trajectories

- Dense sampling of features
- Feature tracking

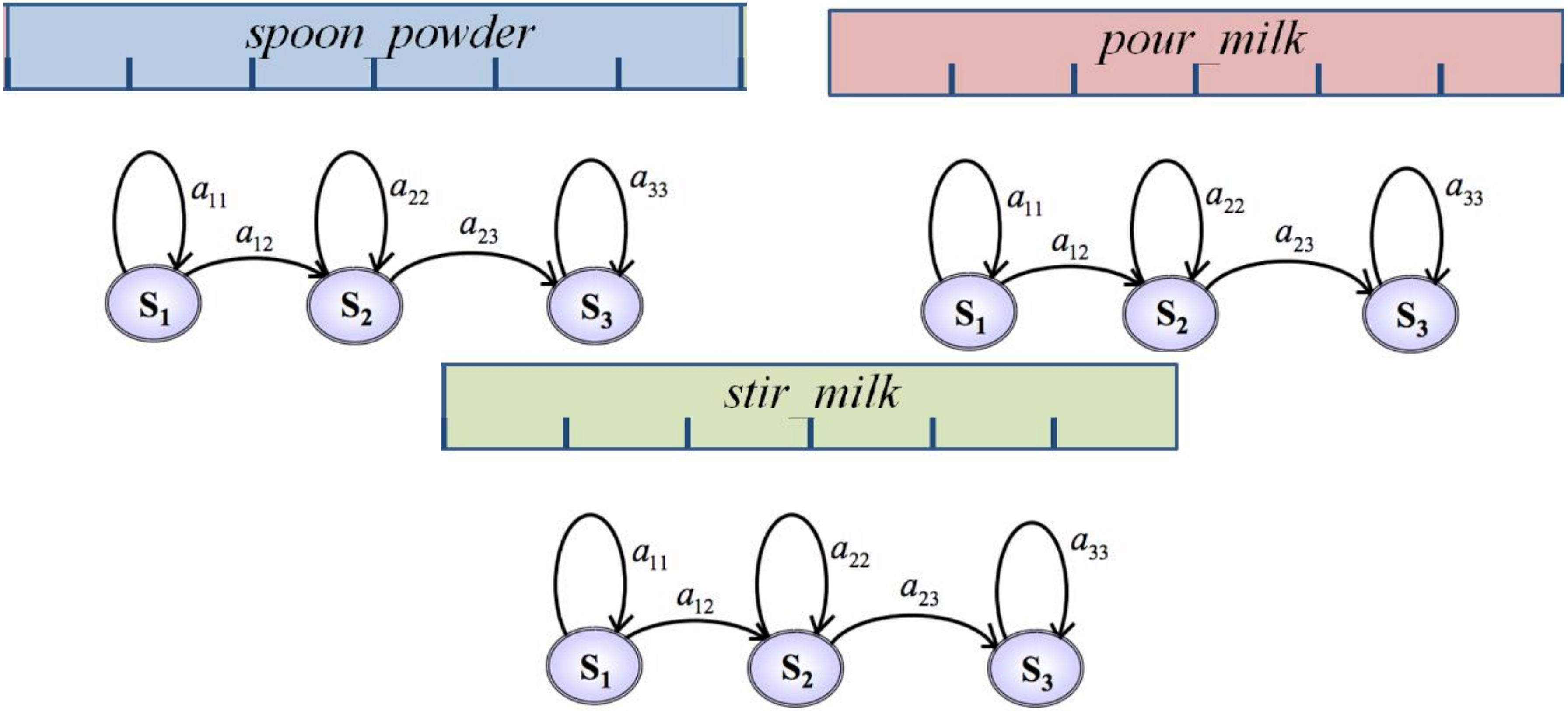


[H. Wang et al. **Dense Trajectories and Motion Boundary Descriptors for Action Recognition**. International Journal of Computer Vision 2013]



Hidden Markov Model

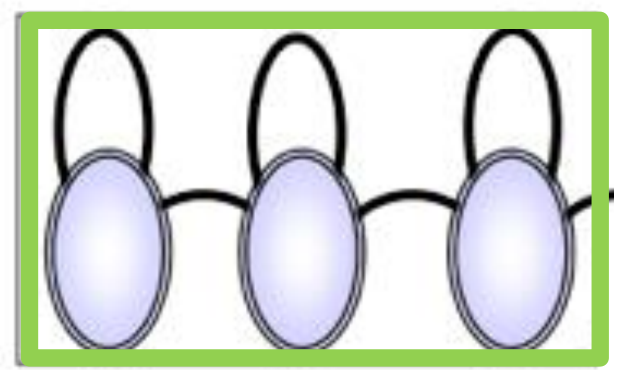
- Hidden Markov Model (HMM) for each activity

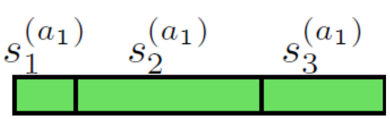


[H. Kuehne et al. **An end-to-end generative framework for video segmentation and recognition.** WACV 2016]

Baseline

- HMM + GMM (IDT)

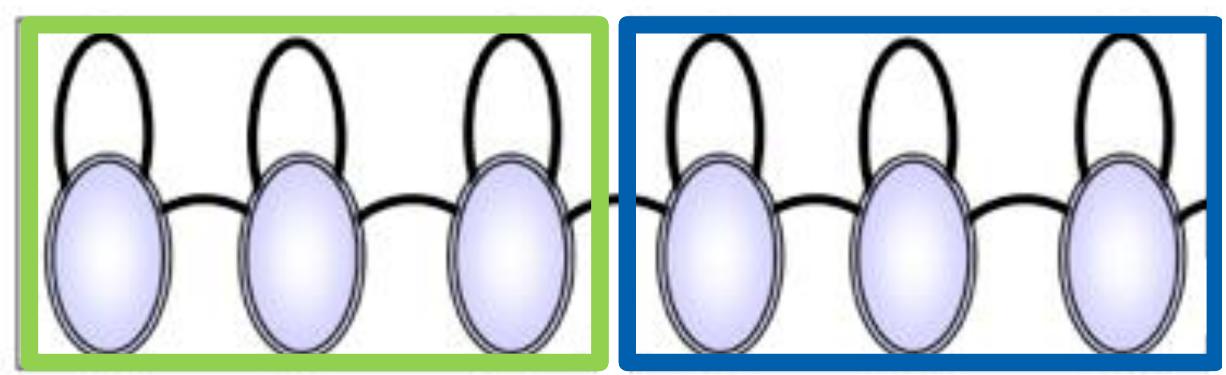


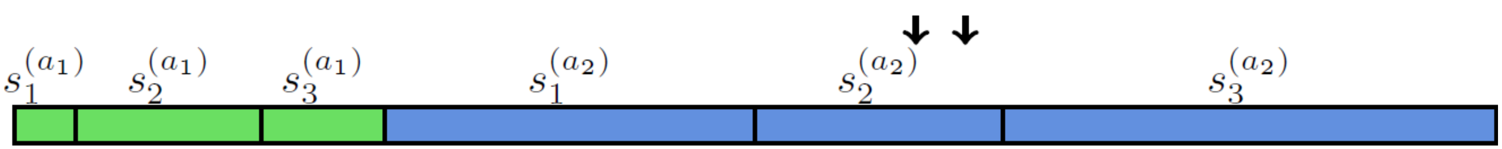


[H. Kuehne et al. **An end-to-end generative framework for video segmentation and recognition.** WACV 2016]

Baseline

- HMM + GMM (IDT)

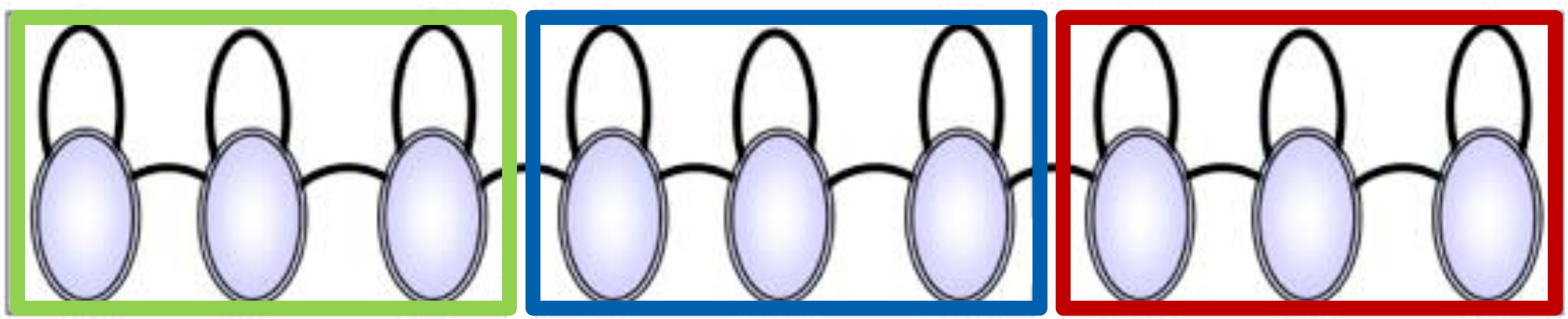


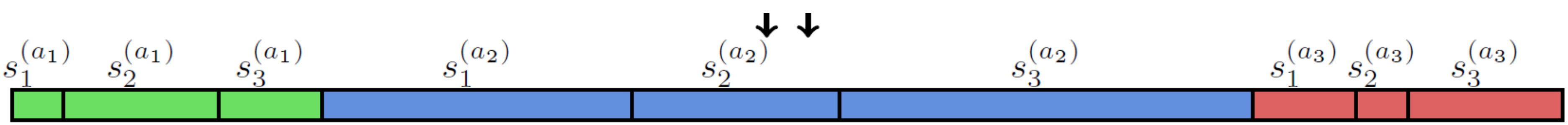


[H. Kuehne et al. **An end-to-end generative framework for video segmentation and recognition.** WACV 2016]

Baseline

- HMM + GMM (IDT)

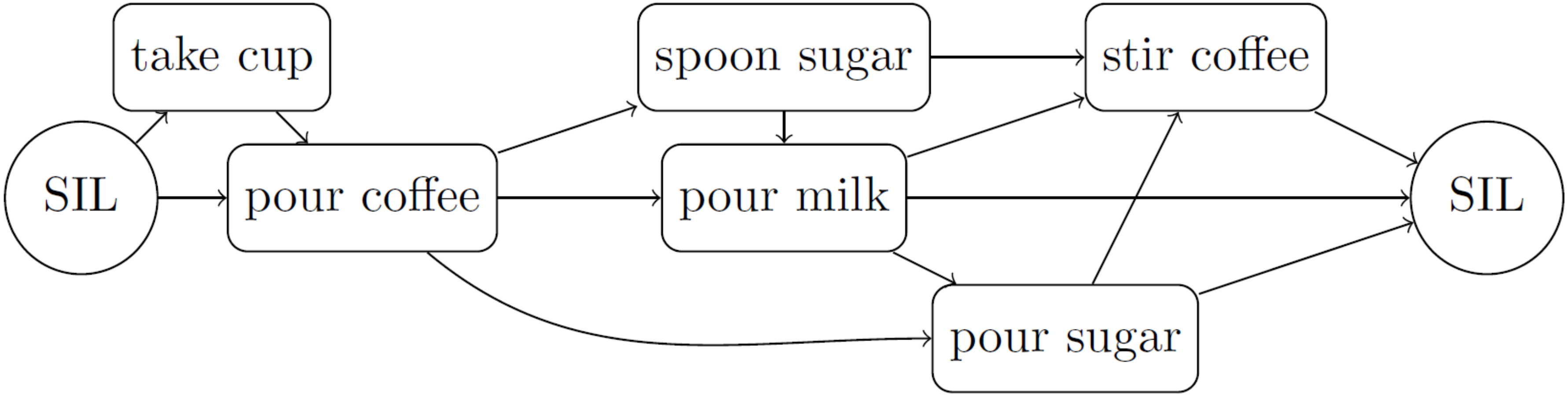




[H. Kuehne et al. **An end-to-end generative framework for video segmentation and recognition.** WACV 2016]

Grammar

- Transitions between activity HMMs are modeled by context free grammar



- SIL: start and end points
- Transition probability is 1 if connection exists otherwise 0

[H. Kuehne et al. **An end-to-end generative framework for video segmentation and recognition.** WACV 2016]

Baseline

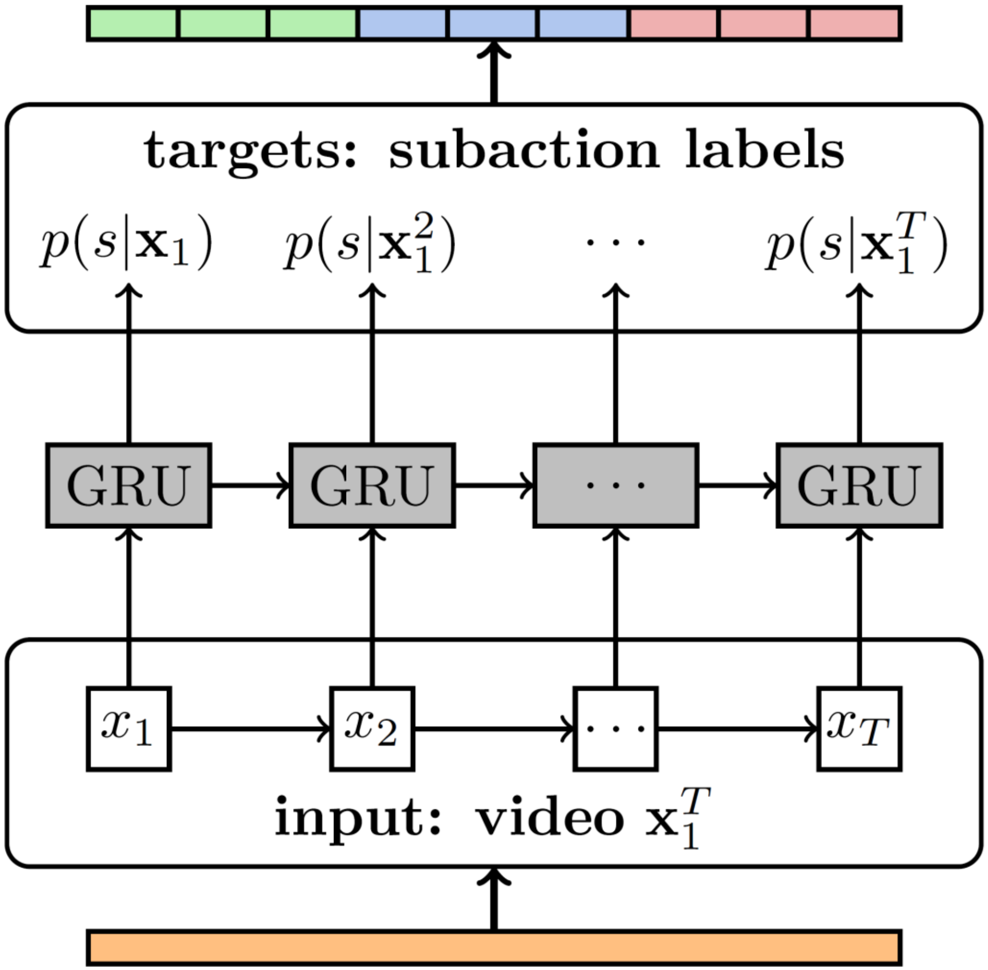
- Breakfast dataset (~65 hours)

Method	Frame-wise Accuracy (%)
Kuehne et al. 2016 (HMM+GMM)	56.3

[H. Kuehne et al. **An end-to-end generative framework for video segmentation and recognition.** WACV 2016]

Hybrid RNN-HMM

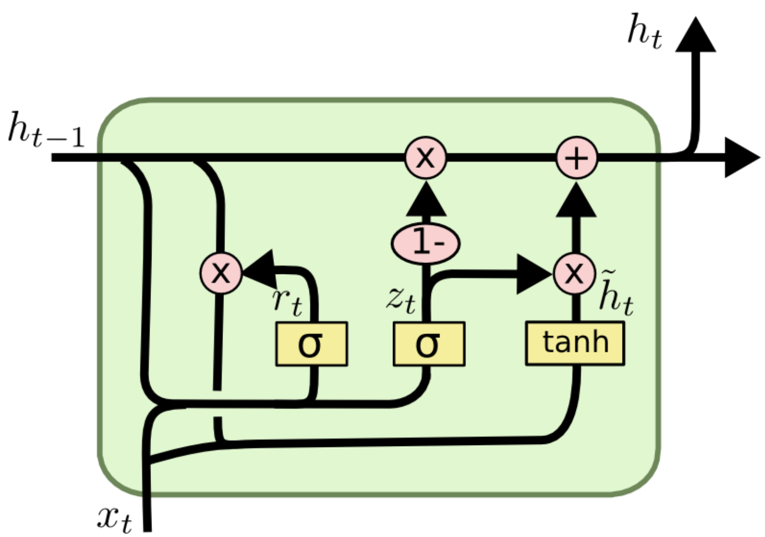
- HMM + RNN with Gated Recurrent Units (GRU)



$$p(x_t|s) = \text{const} \cdot \frac{p(s|x_t)}{p(s)}$$

Gated Recurrent Units (GRU)

- Similar to LSTM, but it does not need an additional memory cell



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

[J. Chung et al. **Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.** NIPS Workshop 2014]

[K. Cho et al. **On the Properties of Neural Machine Translation: Encoder-Decoder Approaches.** Workshop SSST 2014]

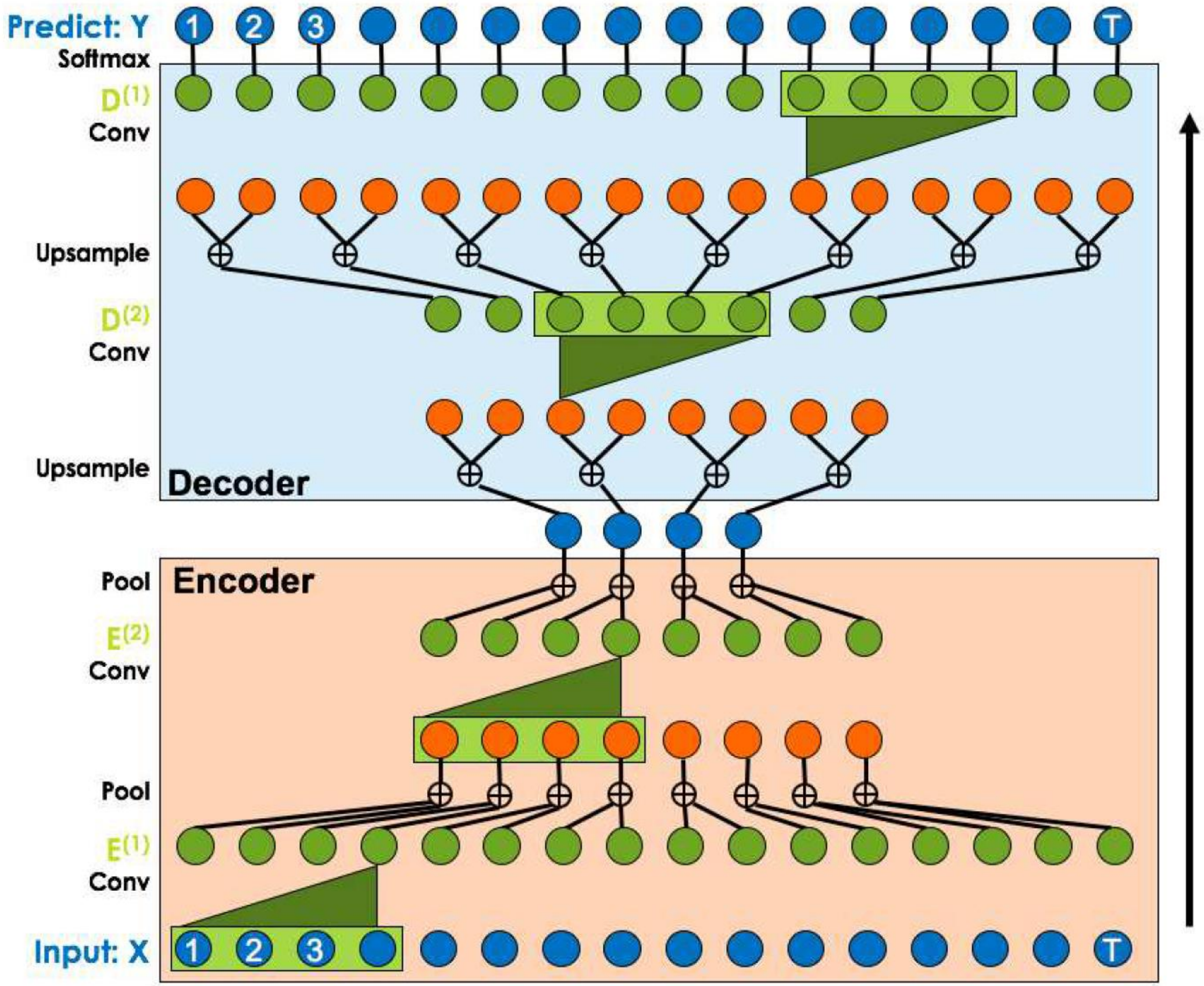
Hybrid RNN-HMM

- Breakfast dataset (~65 hours)

Method	Frame-wise Accuracy (%)
Kuehne et al. 2016 (HMM+GMM)	56.3
Richard et al. 2017 (HMM+RNN)	60.6
Kuehne et al. 2020 (HMM+RNN)	61.3

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]
[H. Kuehne et al. **A Hybrid RNN-HMM Approach for Weakly Supervised Temporal Action Segmentation.** PAMI 2020]

Temporal Convolutional Neural Network



[C. Lea et al. **Temporal Convolutional Networks for Action Segmentation and Detection.** CVPR 2017]

Temporal Convolutional Network

UNIVERSITÄT BONN

- Breakfast dataset (~65 hours)

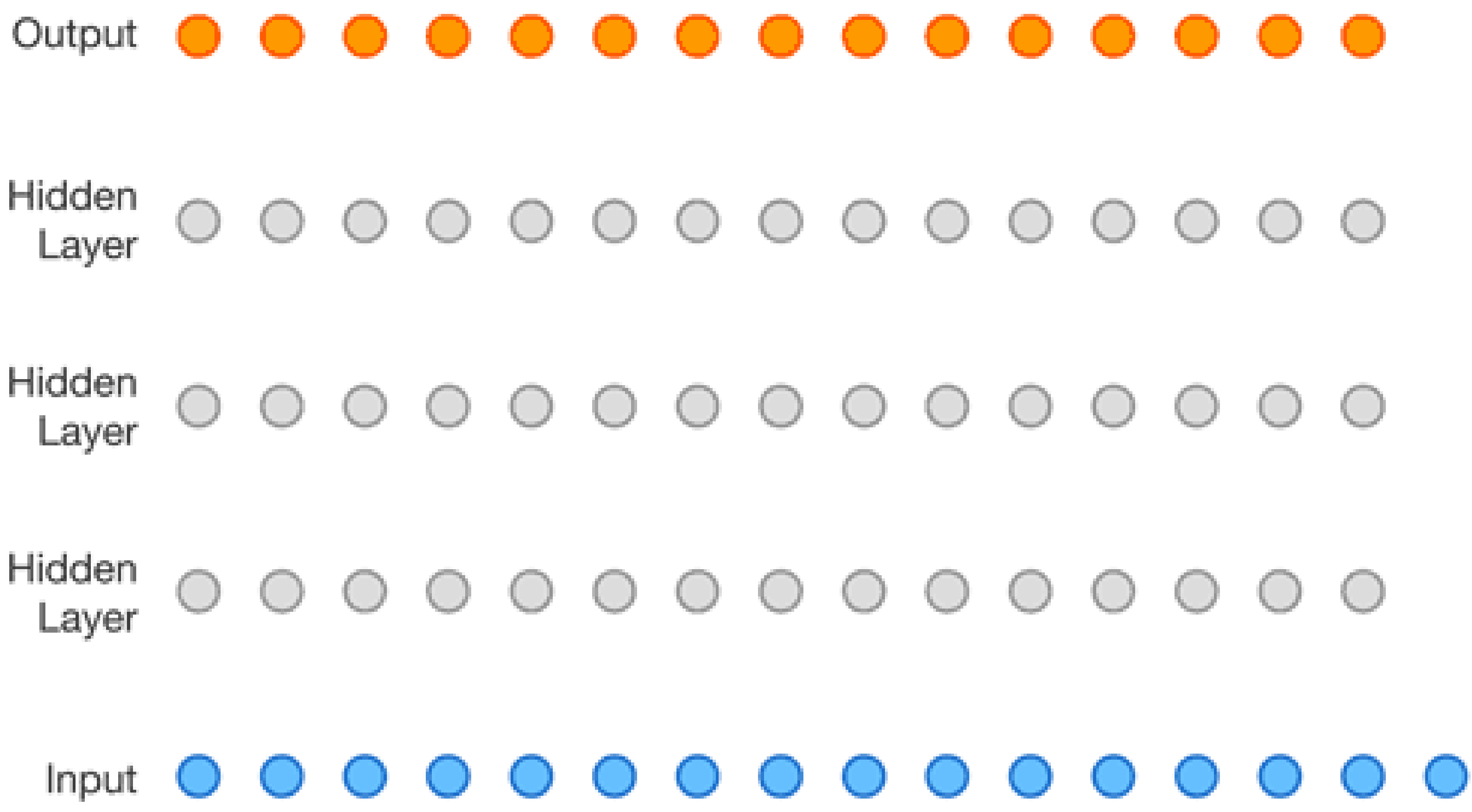
Method	Frame-wise Accuracy (%)
Lea et al. 2017 (ED-TCN)*	43.3
Kuehne et al. 2016 (HMM+GMM)	56.3
Richard et al. 2017 (HMM+RNN)	60.6
Kuehne et al. 2020 (HMM+RNN)	61.3

*[L. Ding and C. Xu. **Weakly-supervised action segmentation with iterative soft boundary assignment.** CVPR 2018]

[C. Lea et al. **Temporal Convolutional Networks for Action Segmentation and Detection.** CVPR 2017]

Temporal Convolutional Neural Network

- Dilated convolutions for audio

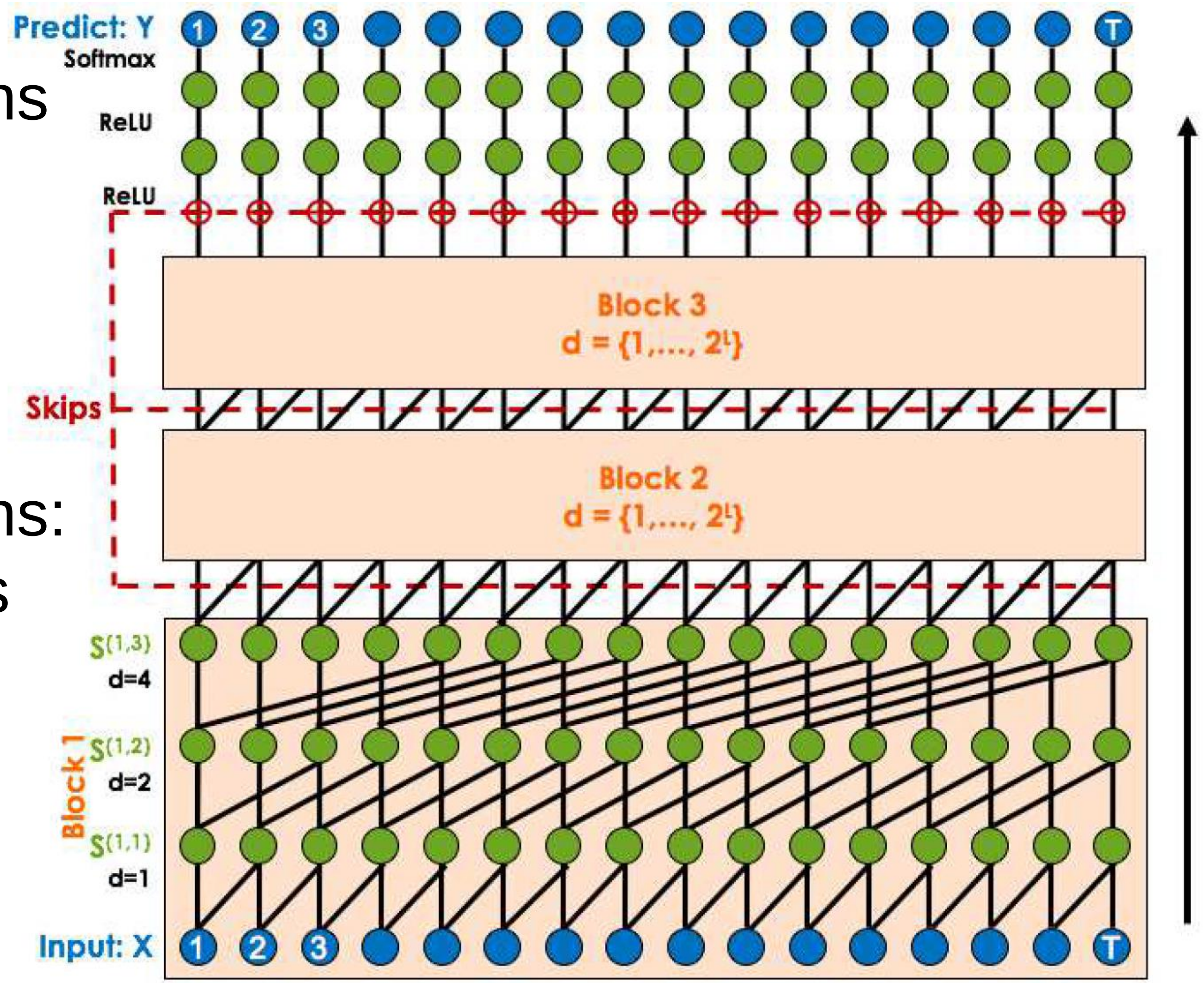


[van den Oord et al. **WaveNet: A Generative Model for Raw Audio.** SSW 2016]

Temporal Convolutional Neural Network

Dilated convolutions capture long temporal receptive field

Causal convolutions: Input for t depends only on previous observations



[C. Lea et al. **Temporal Convolutional Networks for Action Segmentation and Detection.** CVPR 2017]

Temporal Convolutional Network

- 50 Salads

Method	Frame-wise Accuracy (%)
Lea et al. 2017 (ED-TCN)	64.7
Lea et al. 2017 (Dilated TCN)	59.3

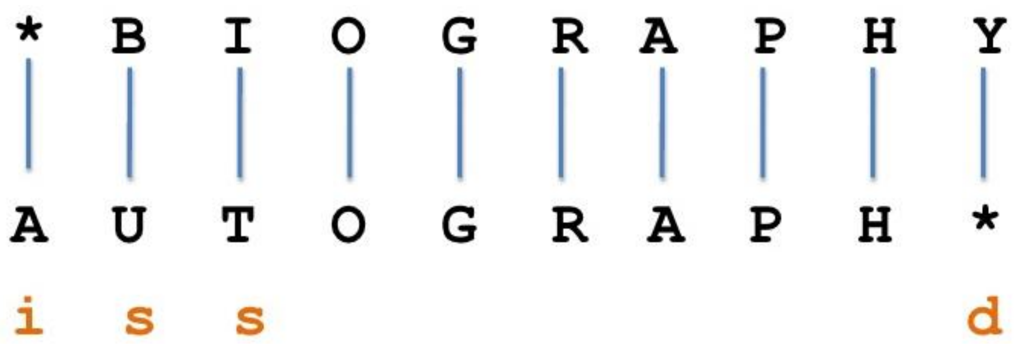
[C. Lea et al. **Temporal Convolutional Networks for Action Segmentation and Detection.** CVPR 2017]

Temporal Convolutional Network

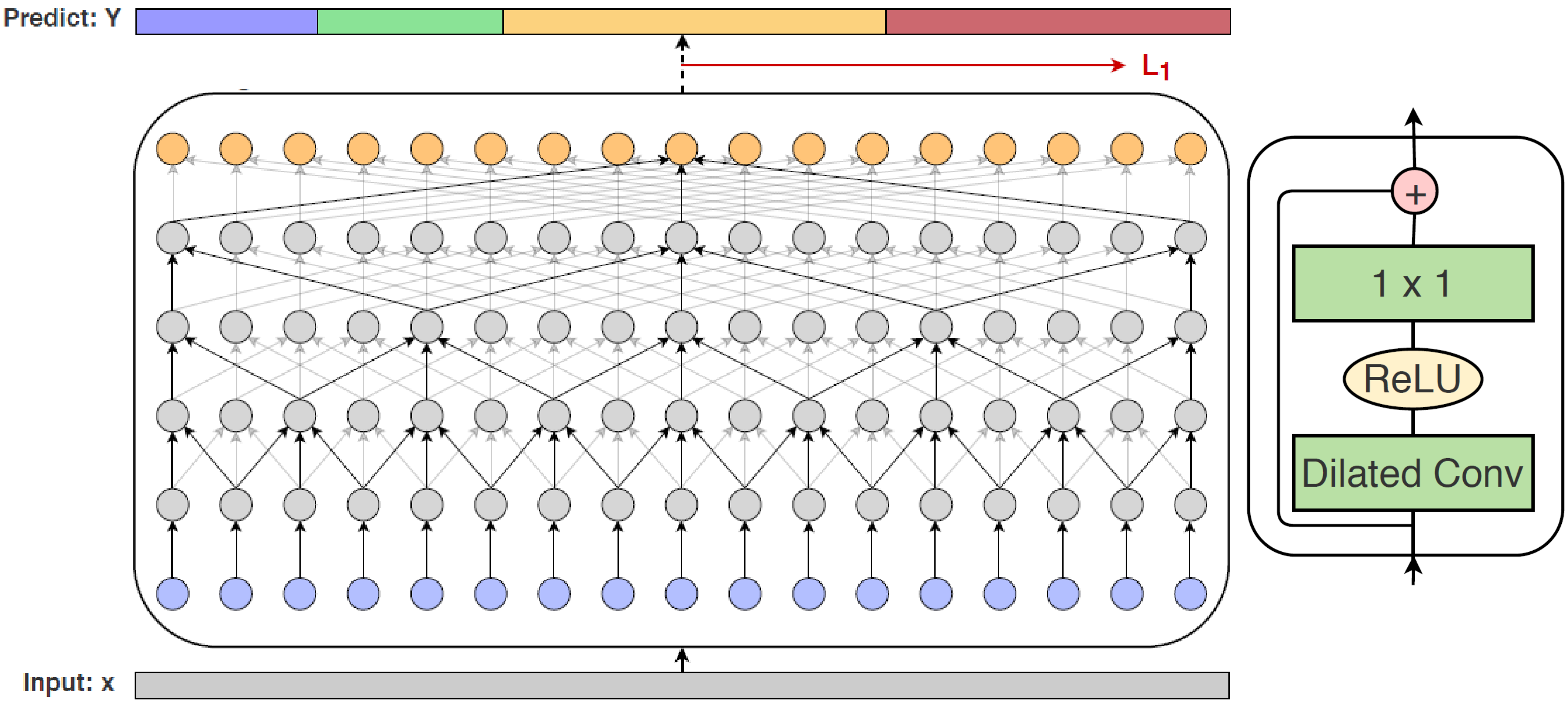
- 50 Salads

Method	1 – Norm. Edit Distance (%)	Frame-wise Accuracy (%)
Lea et al. 2017 (ED-TCN)	59.8	64.7
Lea et al. 2017 (Dilated TCN)	43.1	59.3

- Edit distance (sensitive to oversegmentation):

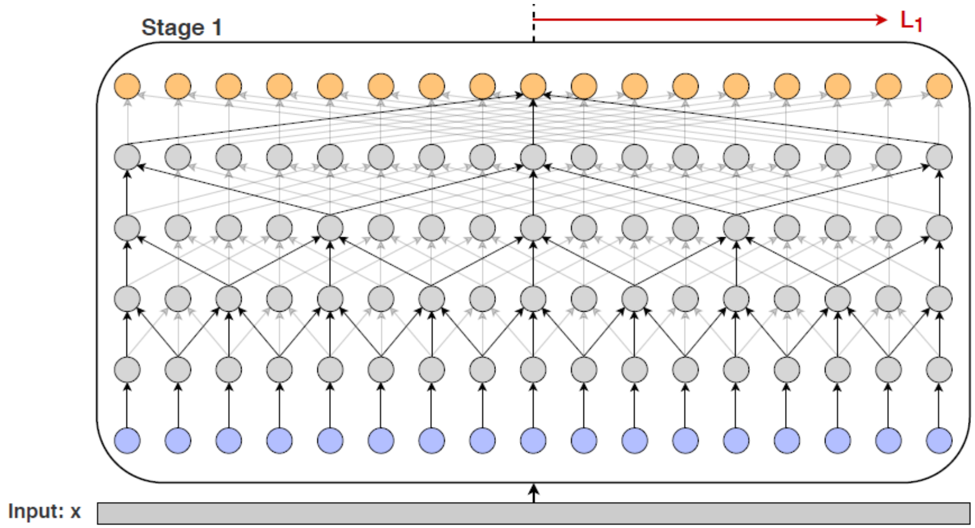


Multi-Stage Temporal Convolutional Network



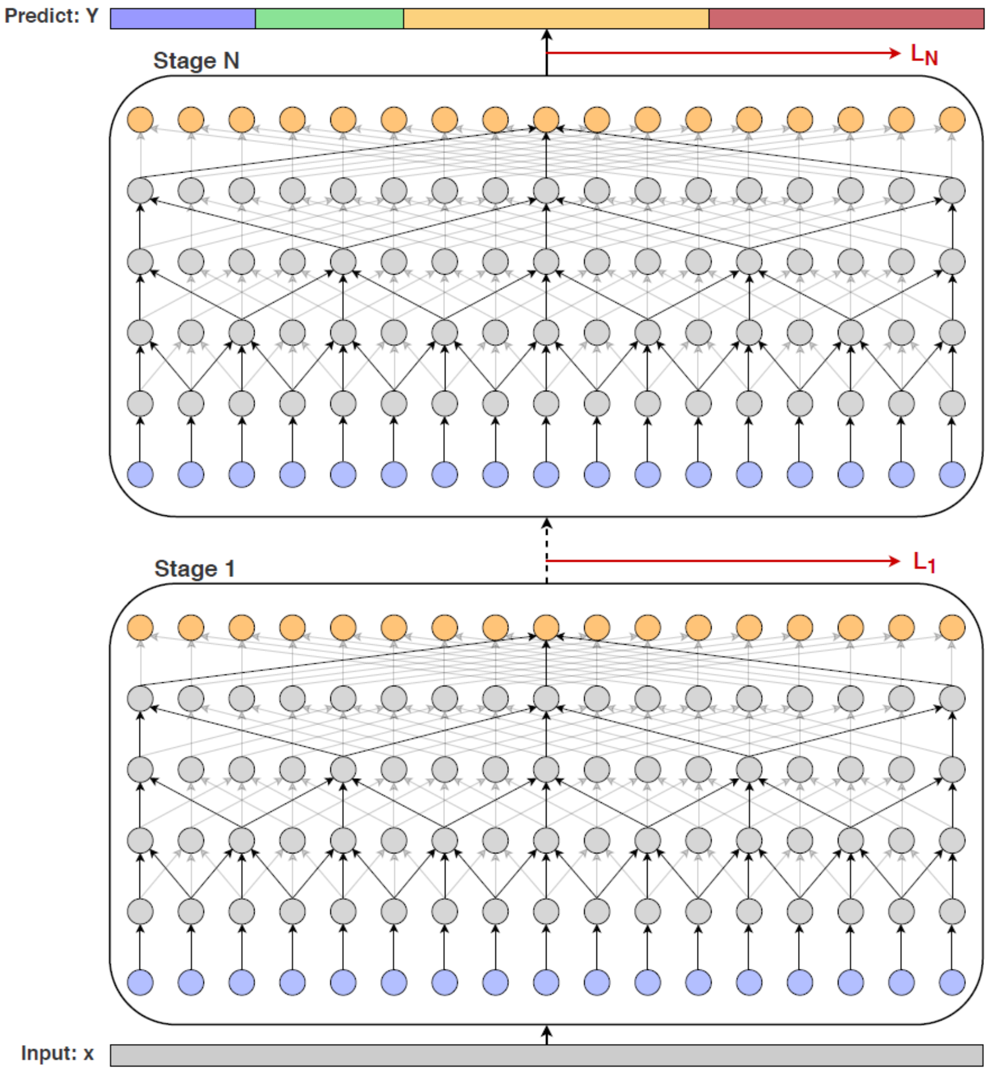
[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

Multi-Stage Temporal Convolutional Network



[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

Multi-Stage Temporal Convolutional Network



[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

Over-segmentation

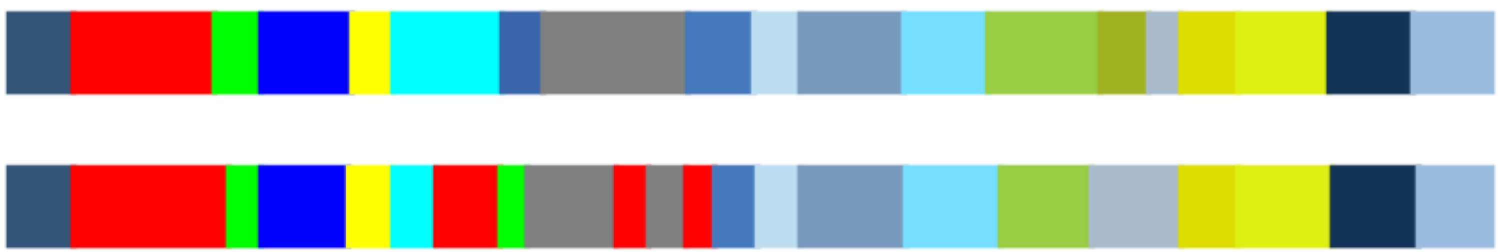
- Frame-wise classification loss:

$$\mathcal{L}_{cls} = \frac{1}{T}\sum_{t} -log(y_{t,c})$$

- Additional loss is required to avoid over-segmentation:

GT

$\mathcal{L}_{cls}$



[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

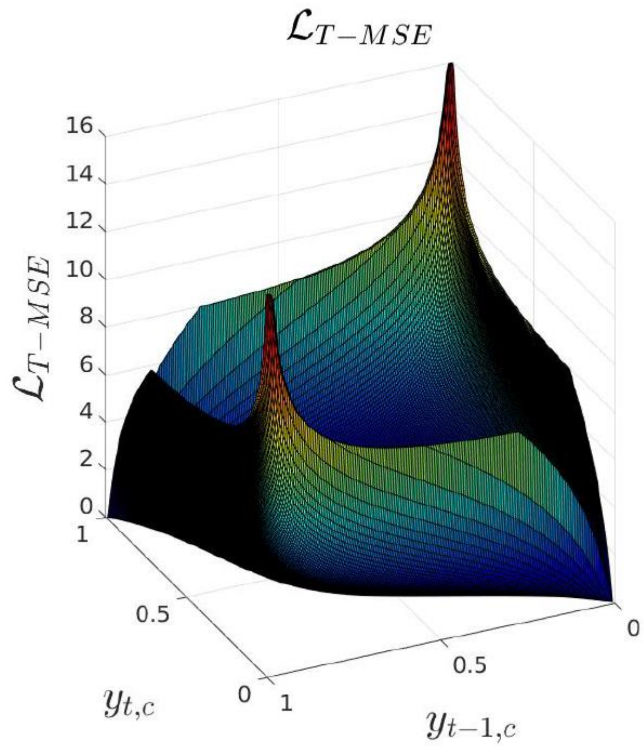
Loss

- Frame-wise classification loss $\mathcal{L}_{cls}$
- Additional loss is required to avoid over-segmentation:

$$\mathcal{L}_{T-MSE} = \frac{1}{TC}\sum_{t,c}\tilde{\Delta}_{t,c}^2$$

$$\tilde{\Delta}_{t,c} = \begin{cases} \Delta_{t,c} & : \Delta_{t,c} \leq \tau \\ \tau & : otherwise \end{cases}$$

$$\Delta_{t,c} = \left|log\ y_{t,c} - log\ y_{t-1,c}\right|$$



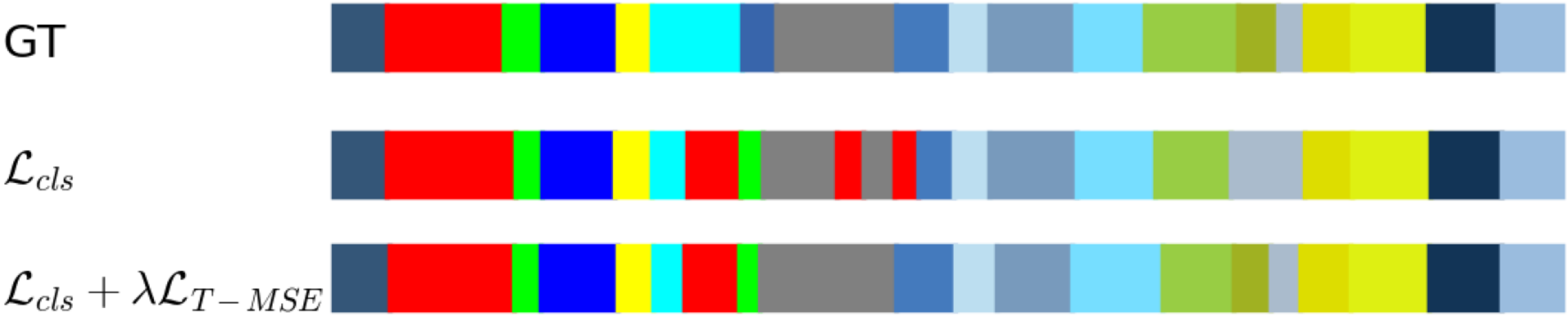
- Loss functions of all stages s:

$$\mathcal{L} = \sum_s \mathcal{L}_s \qquad \mathcal{L}_s = \mathcal{L}_{cls} + \lambda\mathcal{L}_{T-MSE}$$

[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

Loss

- Additional loss is required to avoid oversegmentation



	Edit	Acc
$\mathcal{L}_{cls}$	64.2	79.9
$\mathcal{L}_{cls} + \lambda\mathcal{L}_{T-MSE}$	**67.9**	**80.7**

Impact of stages

- Impact of stages (50 Salads)

	Edit	Acc
SS-TCN	20.5	78.2
MS-TCN (2 stages)	47.9	79.8
MS-TCN (3 stages)	64.0	78.6
MS-TCN (4 stages)	67.9	**80.7**
MS-TCN (5 stages)	**69.2**	79.5

	Edit	Acc
SS-TCN (48 layers)	40.7	78.0
MS-TCN	**67.9**	**80.7**

[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

Multi-Stage Temporal Convolutional Network

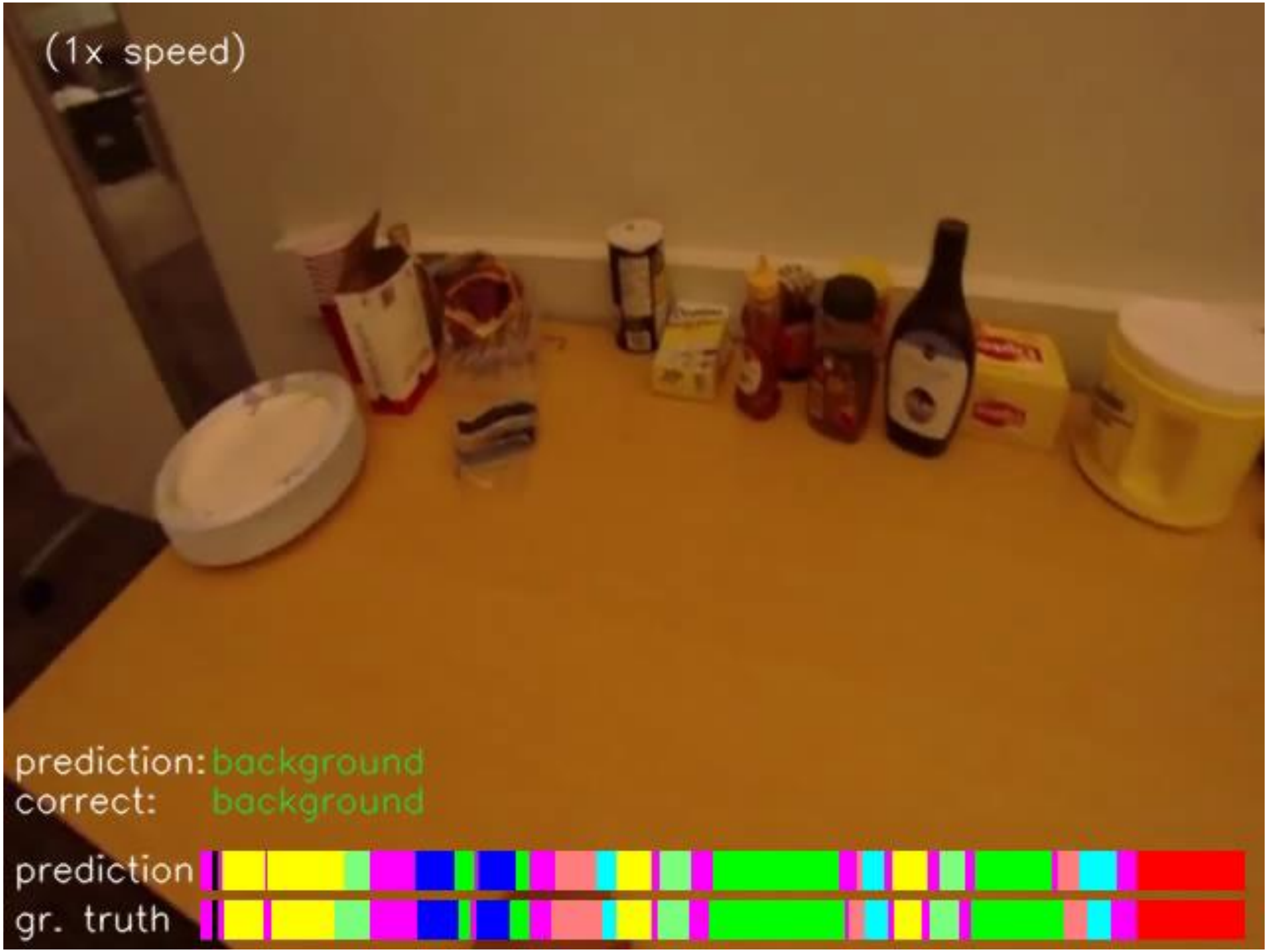
UNIVERSITÄT BONN

- Breakfast dataset (~65 hours)

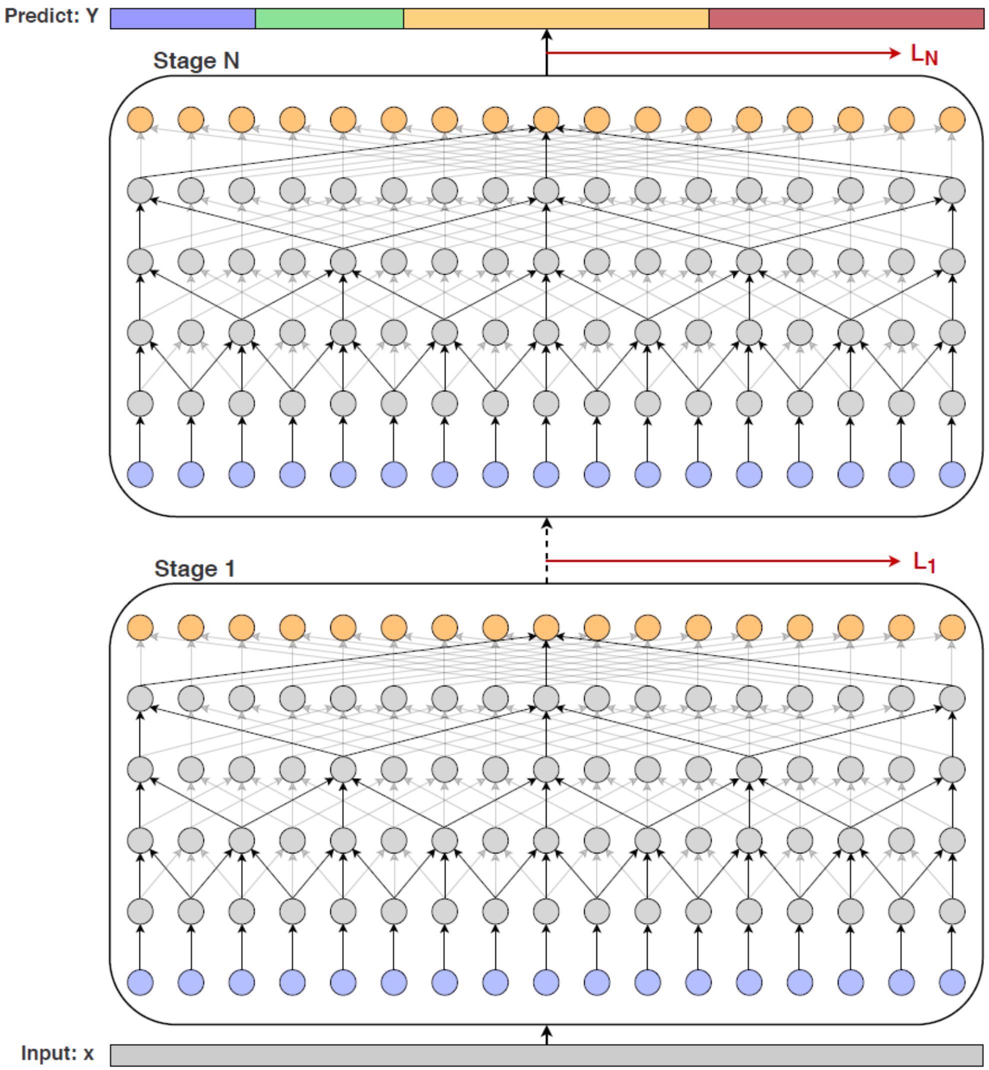
Method	Frame-wise Accuracy (%)
Lea et al. 2017 (ED-TCN)*	43.3
Kuehne et al. 2016 (HMM+GMM)	56.3
Richard et al. 2017 (HMM+RNN)	60.6
Kuehne et al. 2020 (HMM+RNN)	61.3
MS-TCN (TCN)	65.1
MS-TCN (TCN+I3D)	66.3

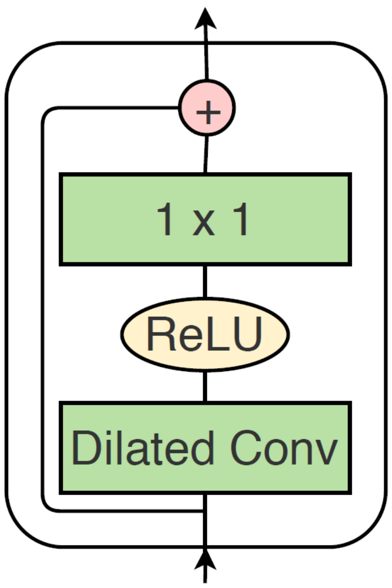
[J. Carreira and A. Zisserman. **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.** CVPR 2017]
[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.** CVPR 2019]

Temporal Action Segmentation



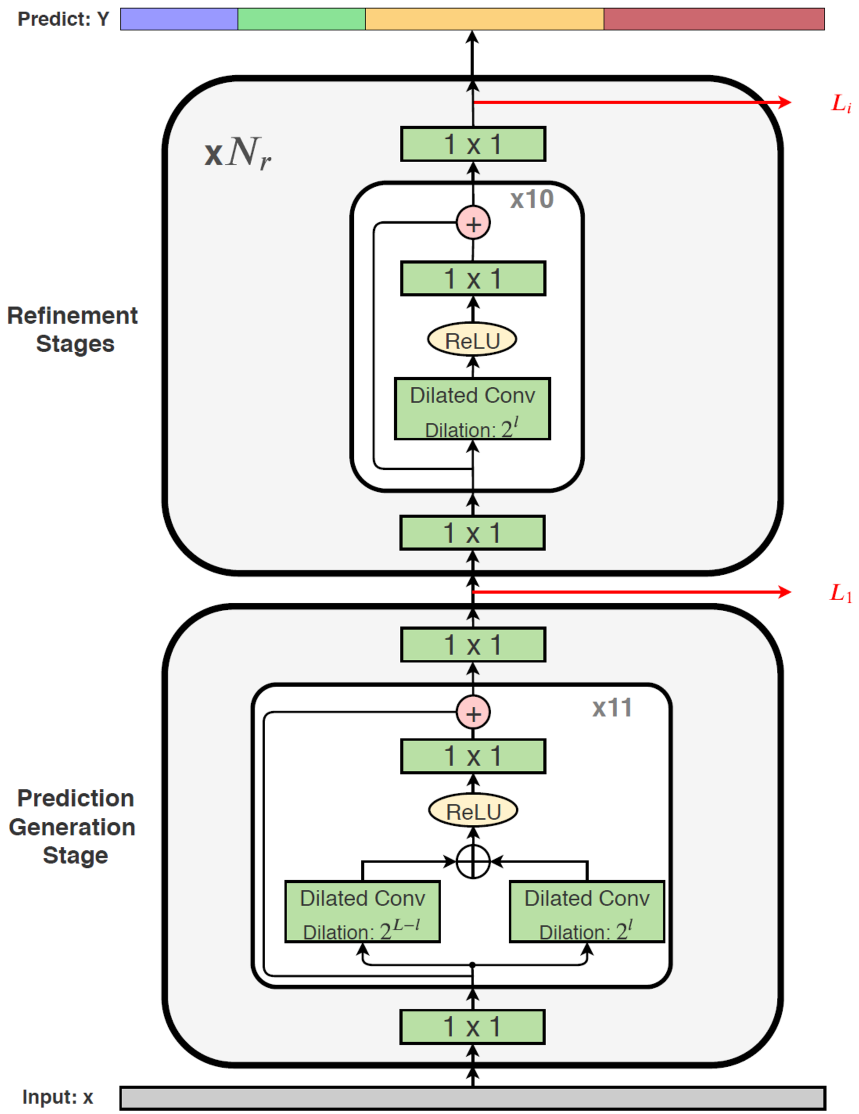
MS-TCN

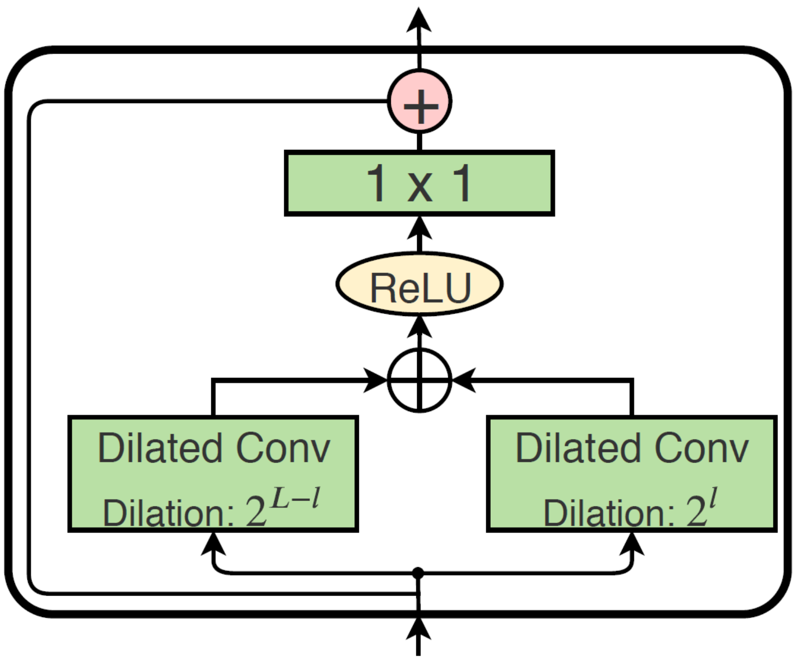




[Y. Abu Farha and J. Gall. **MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation** CVPR 2019]

MS-TCN++





[S. Li et al. **MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation.** arXiv]

MS-TCN++

- Breakfast dataset

Method	Frame-wise Accuracy (%)
Lea et al. 2017 (TCN)*	43.3
Kuehne et al. 2016 (HMM+GMM)	56.3
Richard et al. 2017 (HMM+RNN)	60.6
Kuehne et al. 2020 (HMM+RNN)	61.3
MS-TCN (TCN)	65.1
MS-TCN (TCN+I3D)	66.3
MS-TCN++ (TCN+I3D)	67.6

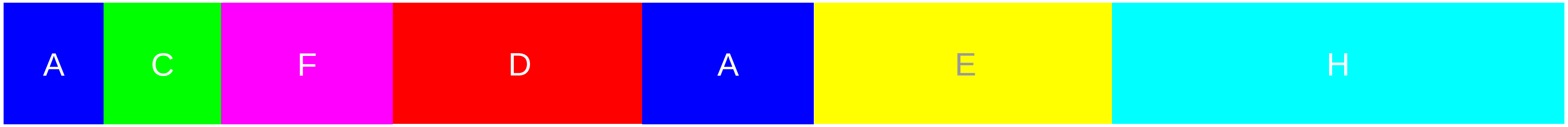
[S. Li et al. **MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation.** arXiv]

MS-TCN++ vs. MS-TCN

Weakly Supervised Learning

- Training video

- Fully supervised:



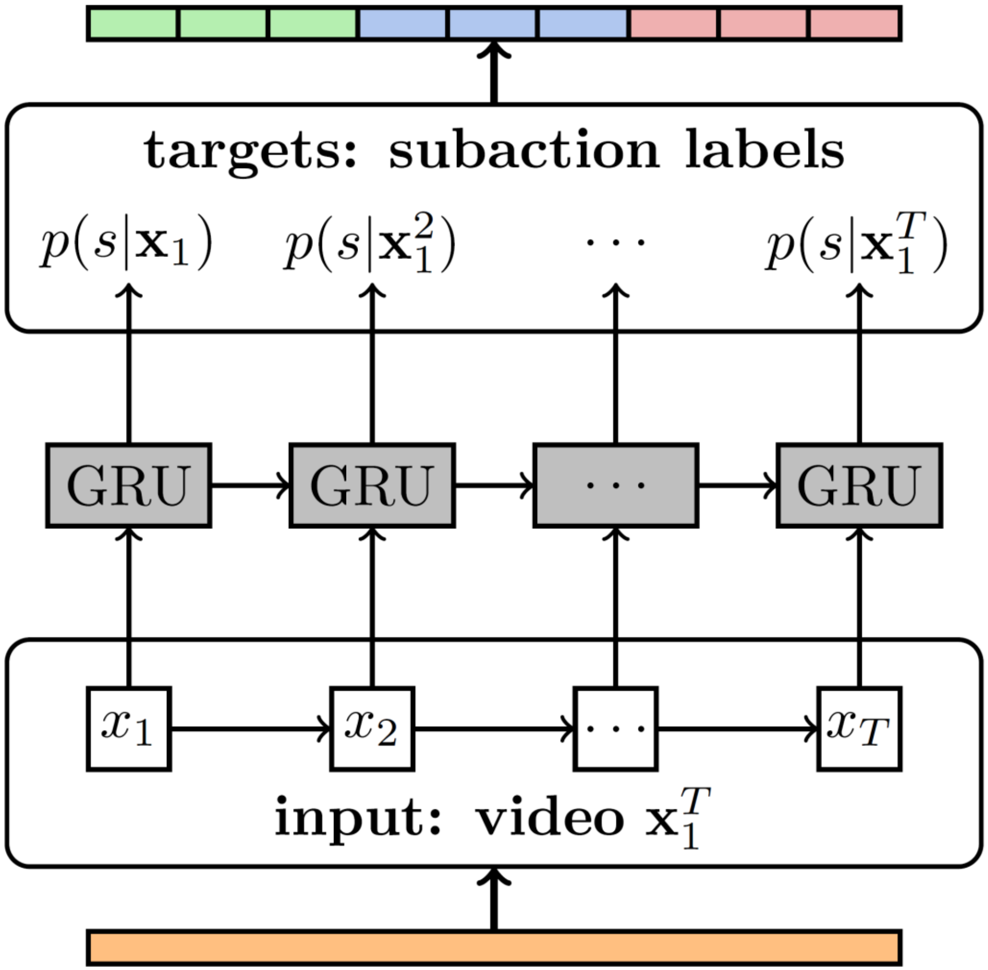
- Weakly supervised (transcripts)

A → C → F → D → A → E → H

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

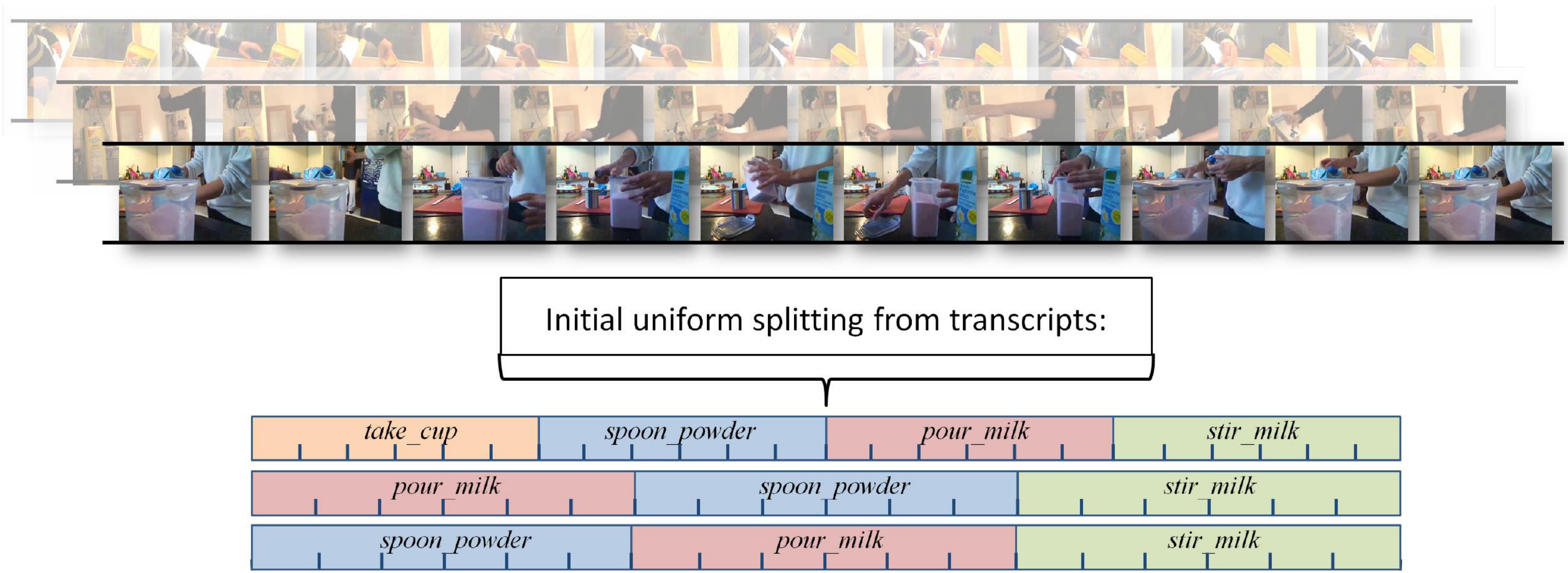
Recall: Hybrid RNN-HMM

- HMM + RNN with Gated Recurrent Units (GRU)

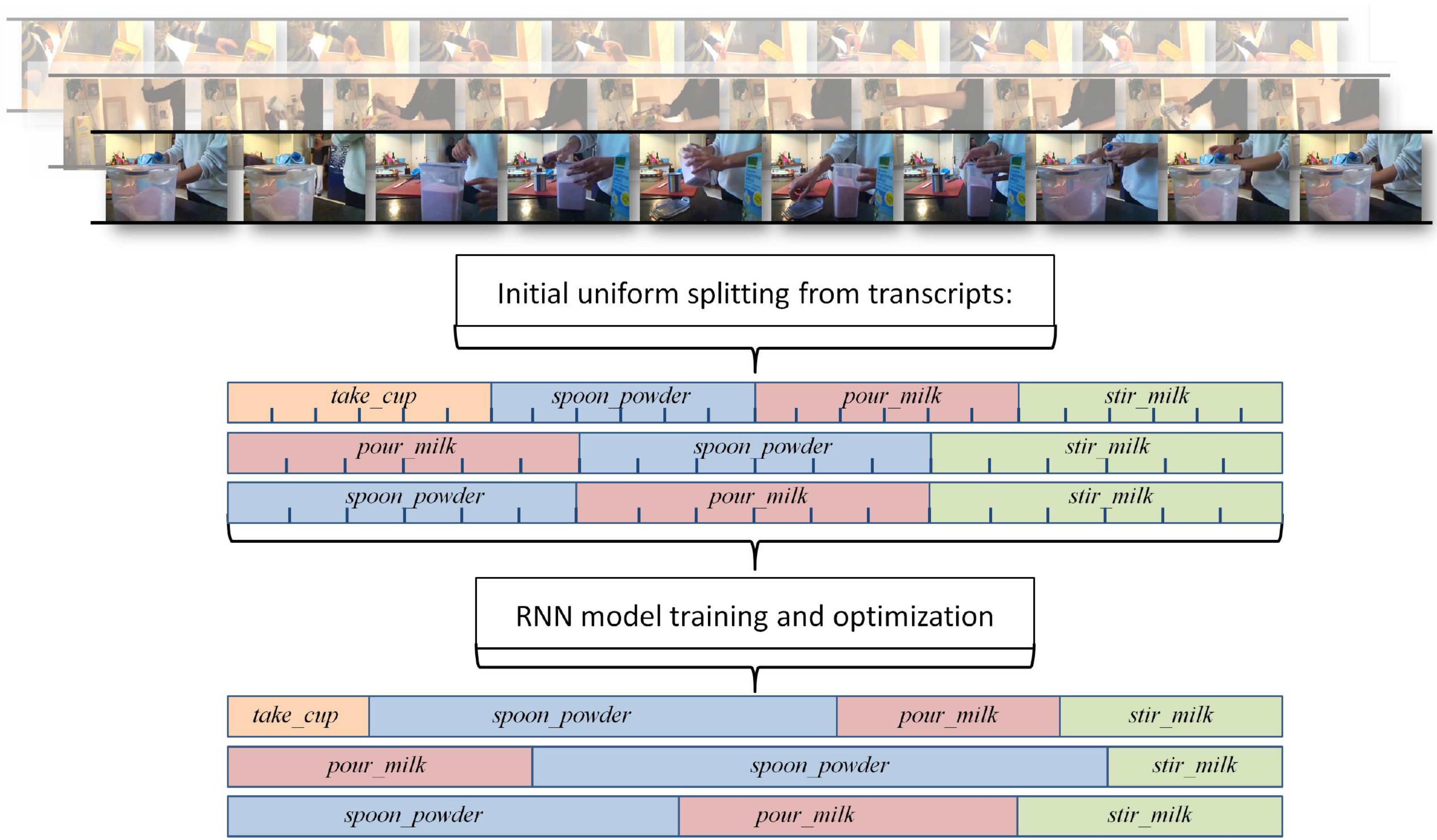


$$p(x_t|s) = \text{const} \cdot \frac{p(s|x_t)}{p(s)}$$

Weakly Supervised Learning



Weakly Supervised Learning

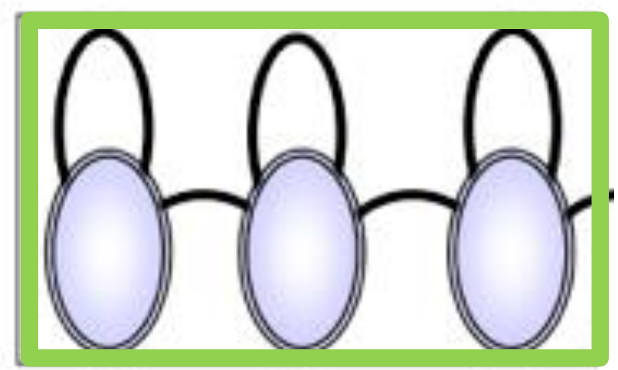


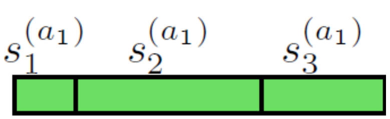
Model

- The transcripts define the order of activities:

Action transcript:

action_1 action_2 action_3



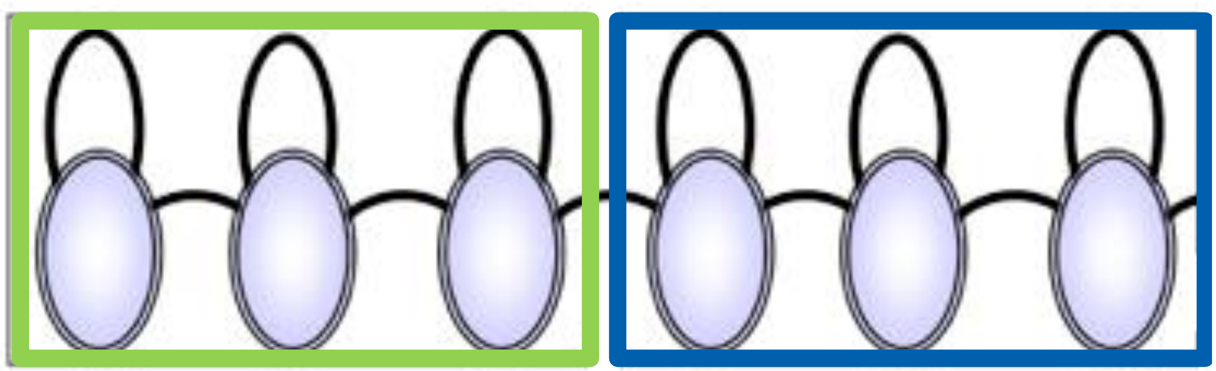


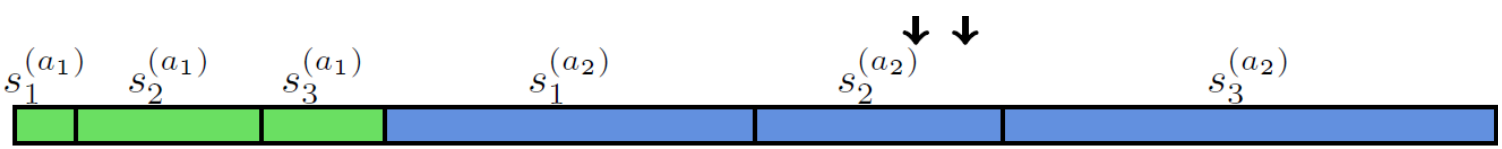
Model

- The transcripts define the order of activities:

Action transcript:

action_1 action_2 action_3



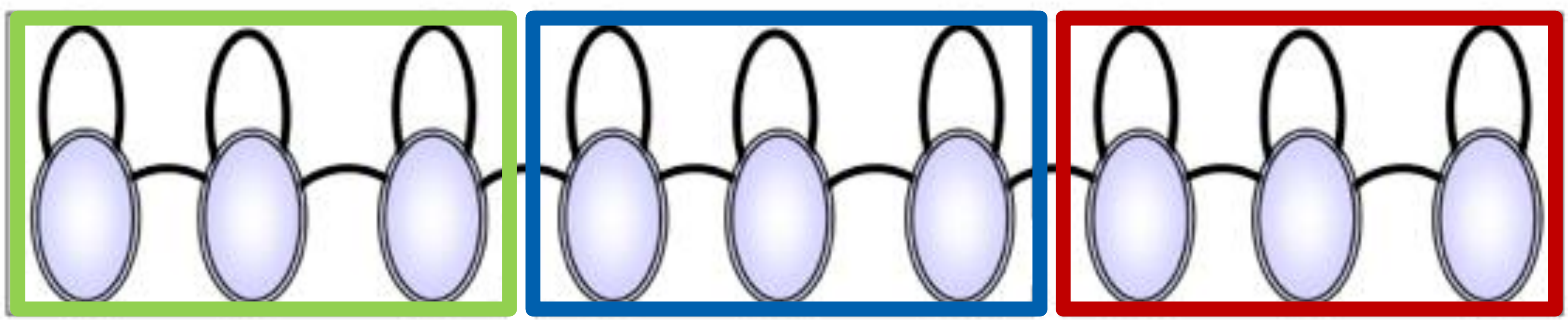


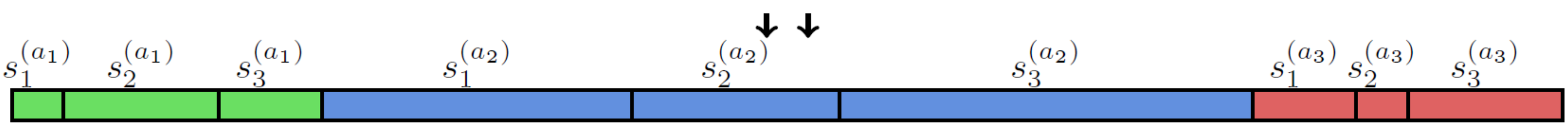
Model

- The transcripts define the order of activities:

Action transcript:

action_1 action_2 action_3





Weakly Supervised Learning

UNIVERSITÄT BONN

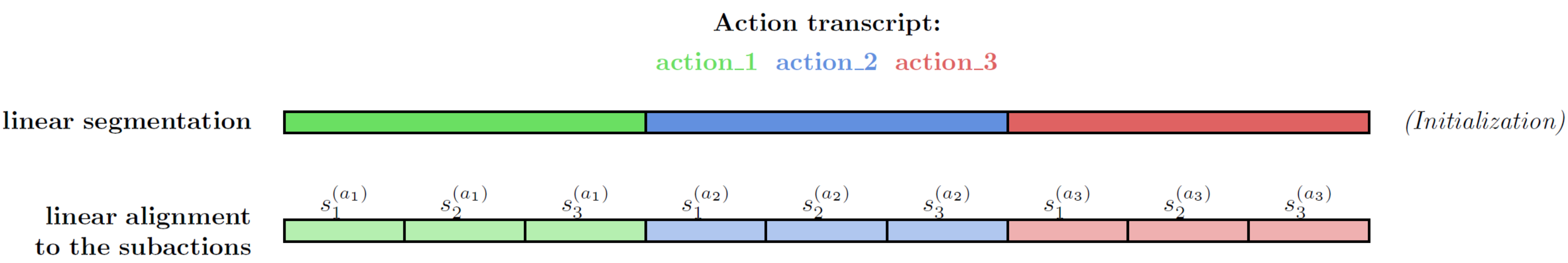
Action transcript:

action_1 action_2 action_3

linear segmentation  *(Initialization)*

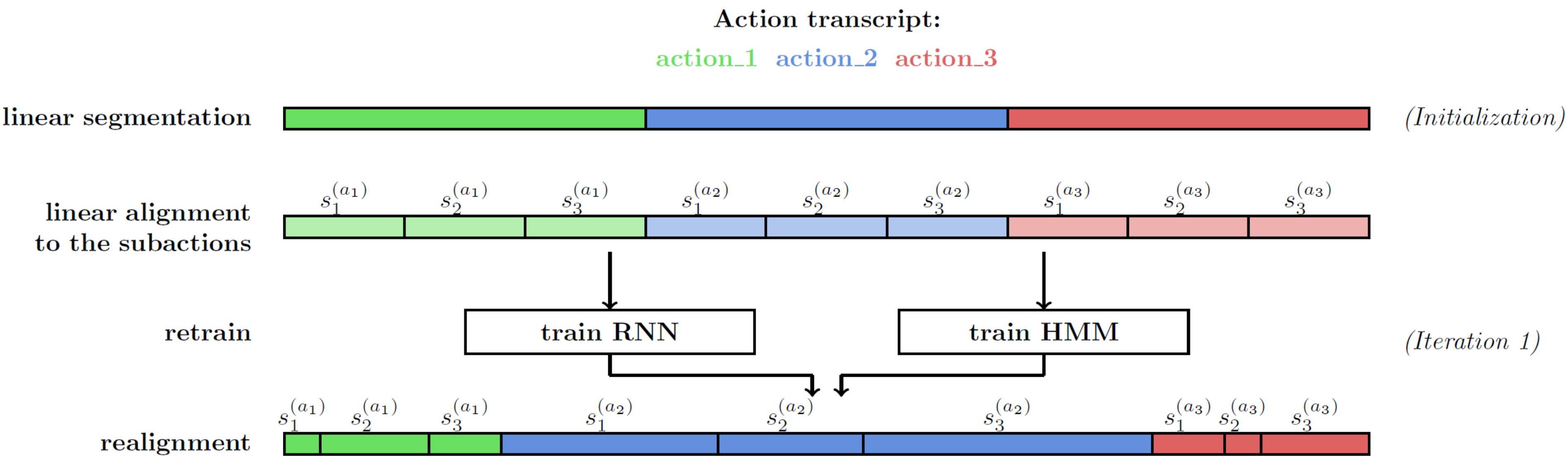
[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Weakly Supervised Learning



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

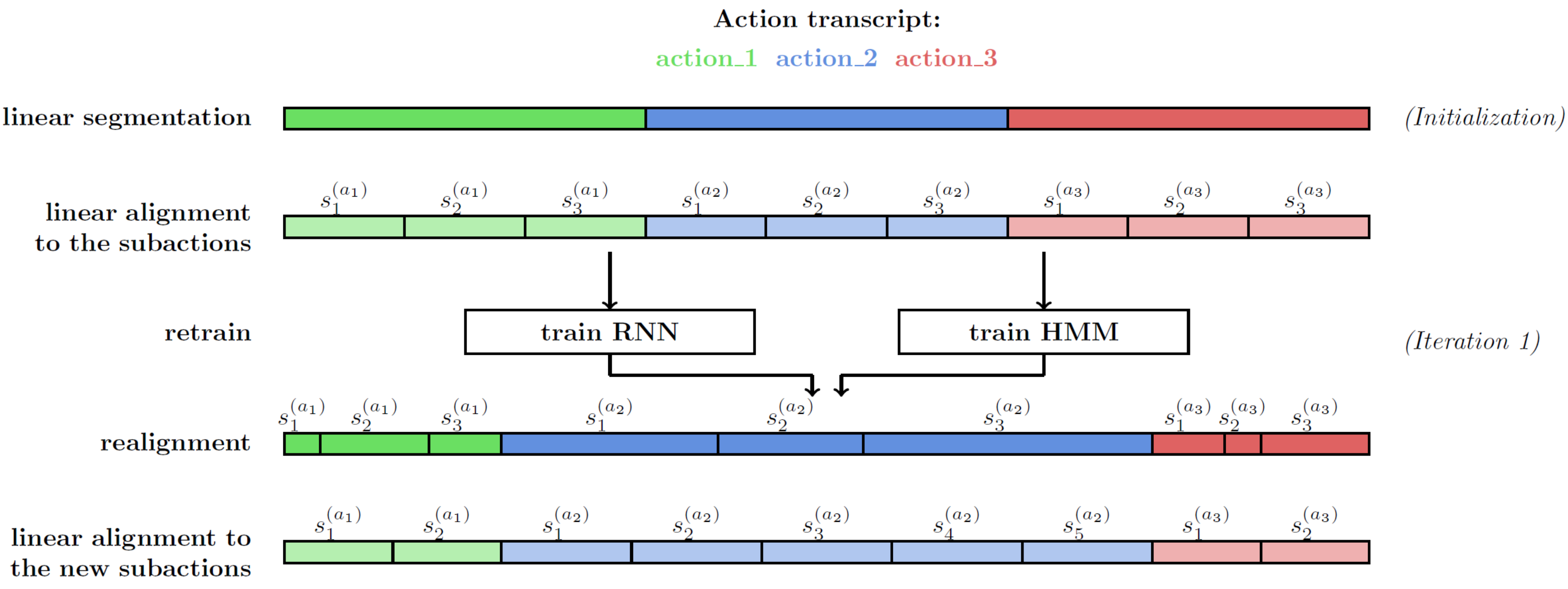
Weakly Supervised Learning



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

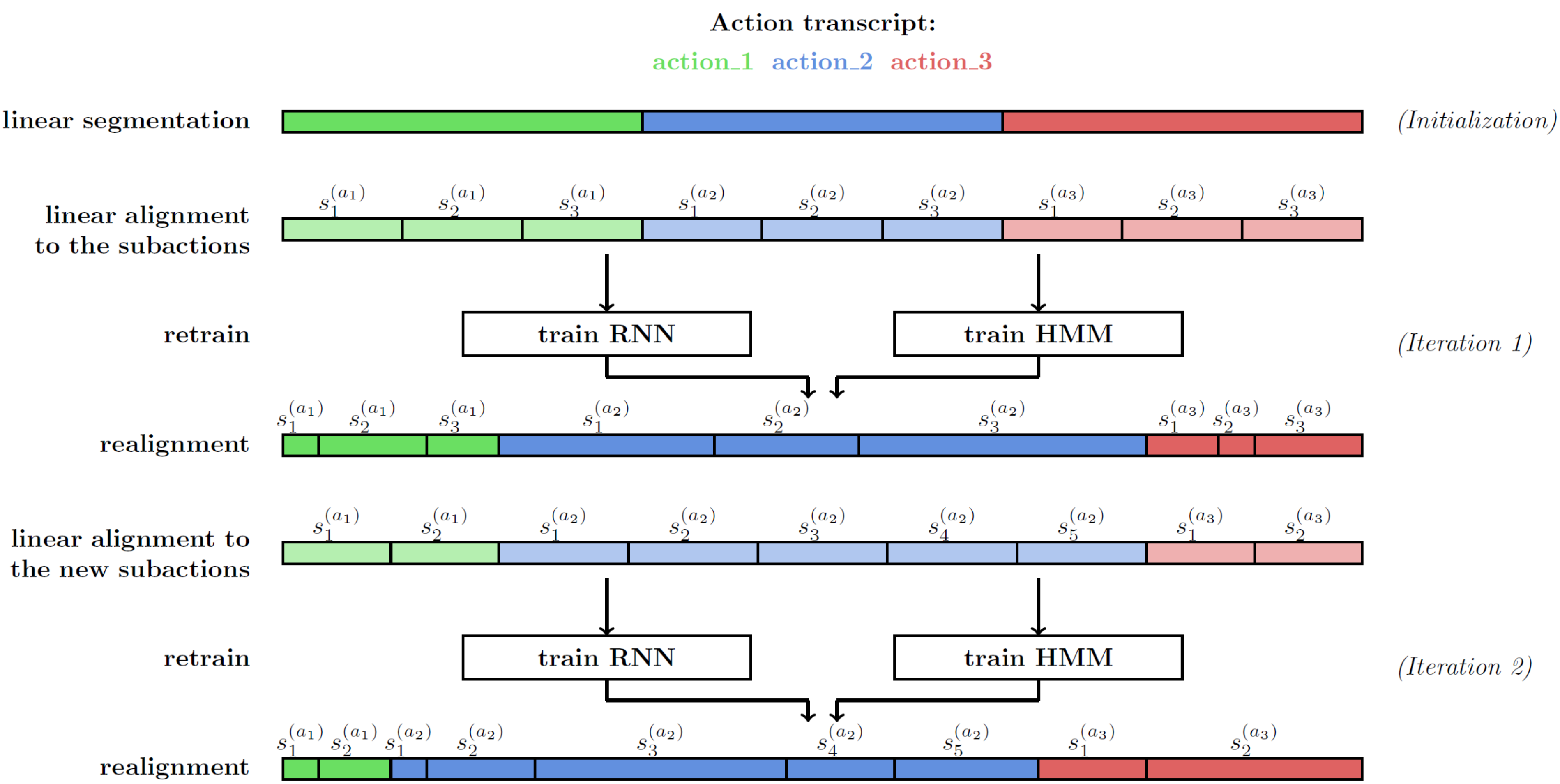
Weakly Supervised Learning

UNIVERSITÄT BONN



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Weakly Supervised Learning



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Results

	Breakfast frame accuracy (%)
pseudo-GT (HMM+RNN) Richard et al. 2017	33.3
pseudo-GT (HMM+RNN) Kuehne et al. 2020	36.7
Fully supervised (HMM+RNN) Kuehne et al. 2020	*61.3*

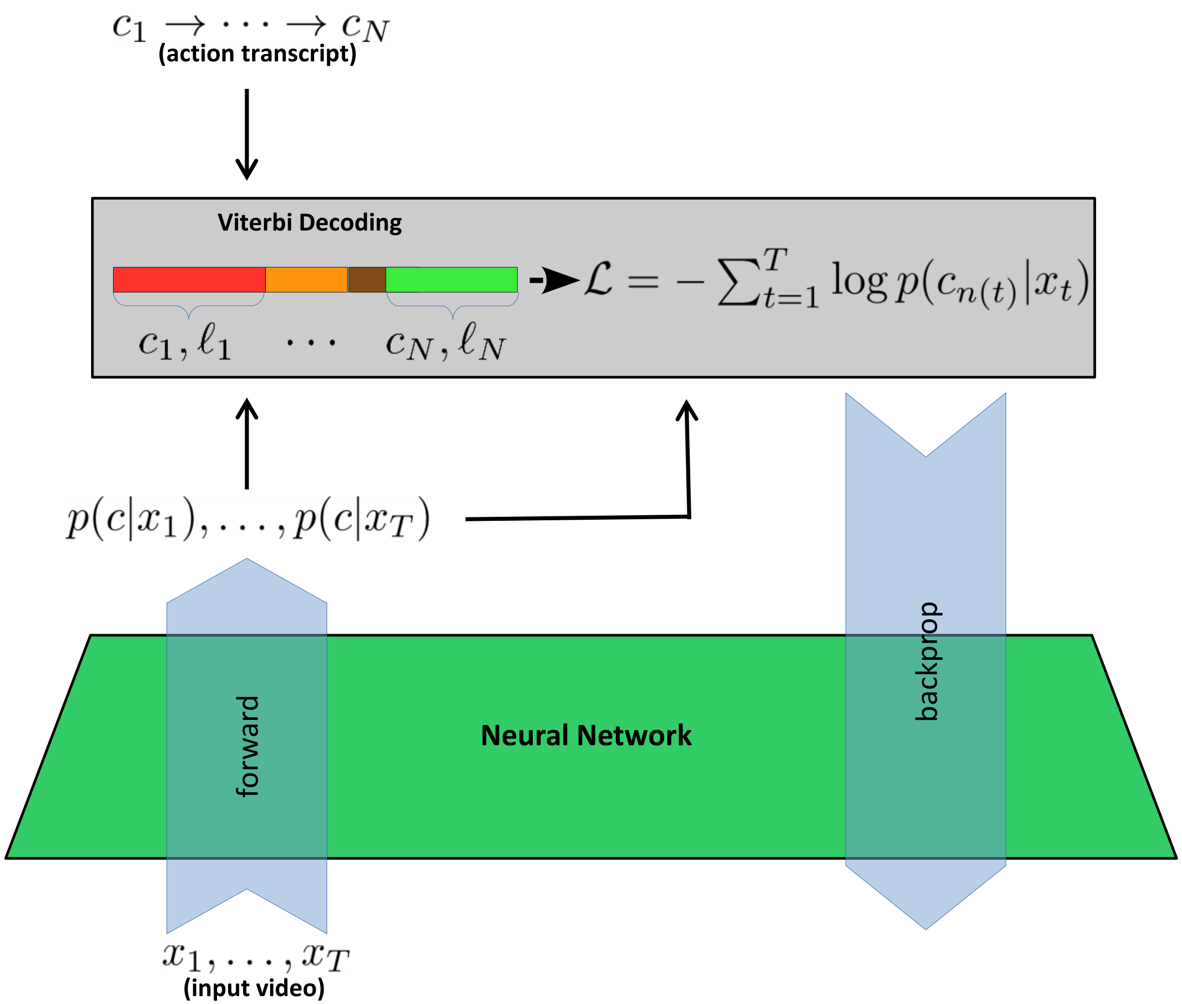
- Disadvantage: Offline and sensitive to initialization

[H. Kuehne et al. **A Hybrid RNN-HMM Approach for Weakly Supervised Temporal Action Segmentation.** PAMI 2020]

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Incremental learning

UNIVERSITÄT BONN



[A. Richard et al. **NeuralNetwork-Viterbi: A Framework for Weakly Supervised Video Learning.** CVPR 2018]

Results

	Breakfast frame accuracy (%)
pseudo-GT (HMM+RNN) Richard et al. 2017	33.3
pseudo-GT (HMM+RNN) Kuehne et al. 2020	36.7
NN-Viterbi (HMM+RNN) Richard et al. 2018	43.0
Fully supervised (HMM+RNN) Kuehne et al. 2020	*61.3*

[A. Richard et al. **NeuralNetwork-Viterbi: A Framework for Weakly Supervised Video Learning.** CVPR 2018]

Pseudo GT vs. NN-Viterbi

Evaluation Issues

UNIVERSITÄT BONN

- Weakly supervised approaches are sensitive to initialization

Model	MoF Reported	MoF Avg (± Std)	MoF Max	MoF Min
ISBA [4]	38.4	36.4 (± 1.0)	37.6	35.1
NNV [13]	43.0	39.7 (± 2.4)	43.5	37.5
CDFL [11]	50.2	48.1 (± 2.5)	50.9	44.6

[L. Ding and C. Xu. **Weakly-supervised action segmentation with iterative soft boundary assignment.** CVPR 2018]

[J. Li et al. **Weakly supervised energy-based learning for action segmentation.** ICCV 2019]

[Y. Souri et al. **On Evaluating Weakly Supervised Action Segmentation Methods.** arXiv]

Features

- Some approaches struggle with pre-trained features (I3D)
- Dimensionality is just one issue

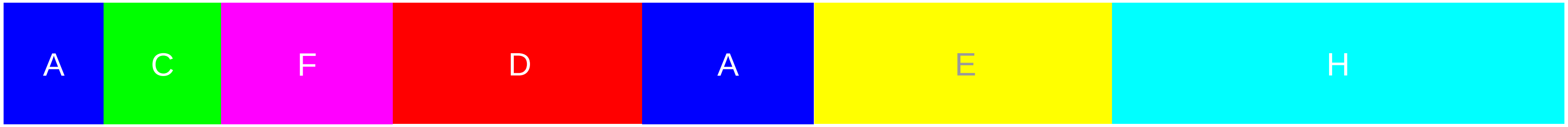
Approach	Features	Average MoF
NNV	IDT	40.6
NNV	I3D	11.4
NNV	PCA-I3D	23.2
CDFL	IDT	48.9
CDFL	I3D	34.9
CDFL	PCA-I3D	38.0

[Y. Souri et al. **On Evaluating Weakly Supervised Action Segmentation Methods.** arXiv]

Weakly Supervised Learning

- Training video

- Fully supervised:



- Weakly supervised (transcripts)

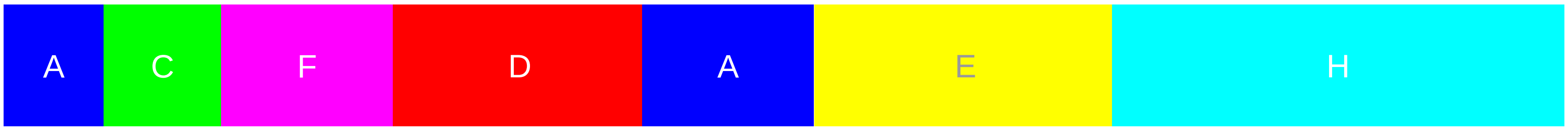
A → C → F → D → A → E → H

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Weakly Supervised Learning

UNIVERSITÄT BONN

- Fully supervised:



- Weakly supervised (transcripts)
  A → C → F → D → A → E → H
- Weakly supervised (action set)
  {A, C, D, E, F, H}
  - Order unknown
  - Number of occurrence unknown

[M. Fayyaz and J. Gall. **SCT: Set Constrained Temporal Transformer for Set Supervised Action Segmentation.** CVPR 2020]

Results

	Supervision	frame accuracy (%)
SCT (TCN+I3D) Fayyaz and Gall 2020	Action set	30.4
pseudo-GT (HMM+RNN) Kuehne et al. 2020	Transcript	36.7
NN-Viterbi (HMM+RNN) Richard et al. 2018	Transcript	43.0
HMM+RNN Kuehne et al. 2020	Full	61.3
MS-TCN++ (TCN+I3D) Li et al. arXiv	Full	67.6

Source Code

- MS-TCN: https://github.com/yabufarha/ms-tcn
- ISBA: https://github.com/Zephyr-D/TCFPN-ISBA
- NN-Viterbi: https://github.com/alexanderrichard/NeuralNetwork-Viterbi
- CDFL: https://github.com/JunLi-Galios/CDFL
- Action sets: https://github.com/alexanderrichard/action-sets
- SCT: https://github.com/MohsenFayyaz89/SCT (Codes not uploaded yet)
- Unsupervised learning: https://github.com/Annusha/unsup_temp_embed

Thank you for your attention.

European
Research
Council

Deutsche
Forschungsgemeinschaft

UNIVERSITÄT BONN