Self-Supervised Temporal Event Segmentation Inspired by Cognitive Theories

by
Ramy Mounir

April 6th, 2021

Doorway Effect

- New scenery demands more cognitive processing causing shifting of event models.
- Shifting makes memories from past event models less accessible to the current event model.
- Evidence that continuous perceptual input is segmented into coherent units - called “events”.

Walking through doorways causes forgetting

Radvansky GA, Krawietz SA, Tamplin AK. Walking through doorways causes forgetting: Further explorations. Q J Exp Psychol (Hove). **2011** Aug;64(8):1632-45. doi: 10.1080/17470218.2011.571267. Epub 2011 May 24. PMID: 21563019.

Pettijohn KA, Radvansky GA. Walking through doorways causes forgetting: recall. Memory. **2018** Nov;26(10):1430-1435. doi: 10.1080/09658211.2018.1489555. Epub 2018 Jun 21. PMID: 29927683.

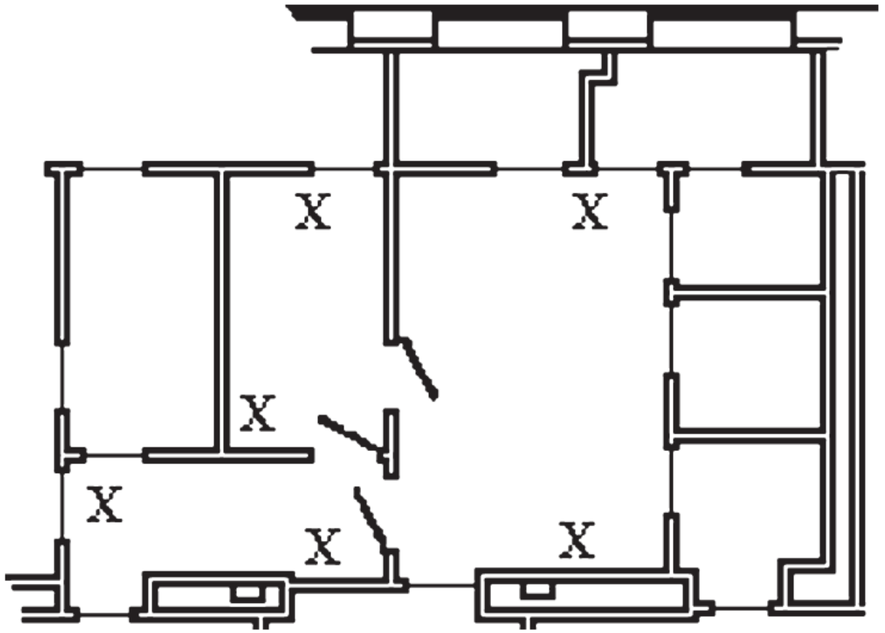
Radvansky GA, Copeland DE. Walking through doorways causes forgetting: situation models and experienced space. Mem Cognit. **2006** Jul;34(5):1150-6. doi: 10.3758/bf03193261. PMID: 17128613.

McFadyen J, Nolan C, Pinocy E, Buteri D, Baumann O. Doorways do not always cause forgetting: a multimodal investigation. BMC Psychol. **2021** Mar 8;9(1):41. doi: 10.1186/s40359-021-00536-3. PMID: 33685514; PMCID: PMC7938580.

Lawrence Z, Peterson D. Mentally walking through doorways causes forgetting: The location updating effect and imagination. Memory. **2016**;24(1):12-20. doi: 10.1080/09658211.2014.980429. Epub 2014 Nov 20. PMID: 25412111.

Seel SV, Easton A, McGregor A, Buckley MG, Eacott MJ. Walking through doorways differentially affects recall and familiarity. Br J Psychol. **2019** Feb;110(1):173-184. doi: 10.1111/bjop.12343. Epub 2018 Sep 16. PMID: 30221342.

Pettijohn KA, Radvansky GA. Walking through doorways causes forgetting: Event structure or updating disruption? Q J Exp Psychol (Hove). **2016** Nov;69(11):2119-29. doi: 10.1080/17470218.2015.1101478. Epub 2016 Feb 16. PMID: 26556012.

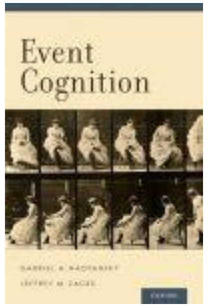


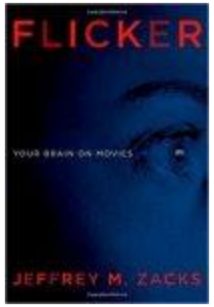
	No shift		*Shift*	
	ER	*RT*	*ER*	*RT*
Positives	.23 (.02)	2,083 (68)	.28 (.02)	2,168 (74)
Negatives	.18 (.01)	2,059 (63)	.21 (.01)	2,091 (57)

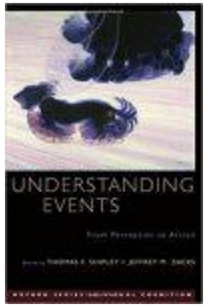
Note: ER = error rate (in proportions). RT = response time (in ms). Standard errors in parentheses.

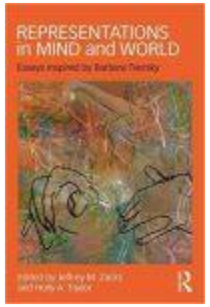
Jeffrey Zacks

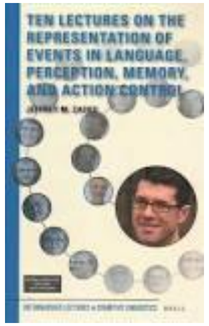
- Associate Chair, Dept. of Psychological and Brain Sciences. *Washington University, St. Louis.*
- *Ph.D. in Cognitive Psychology, Stanford University.*
- *Interests:*
    - *Perception and Cognition.*
    - *Parsing continuous stream of behavior into meaningful events.*
    - *How event segmentation affects memory and cognition.*
    - *Mental representation for reasoning about spatial relations.*
- *Author of 5 books.*





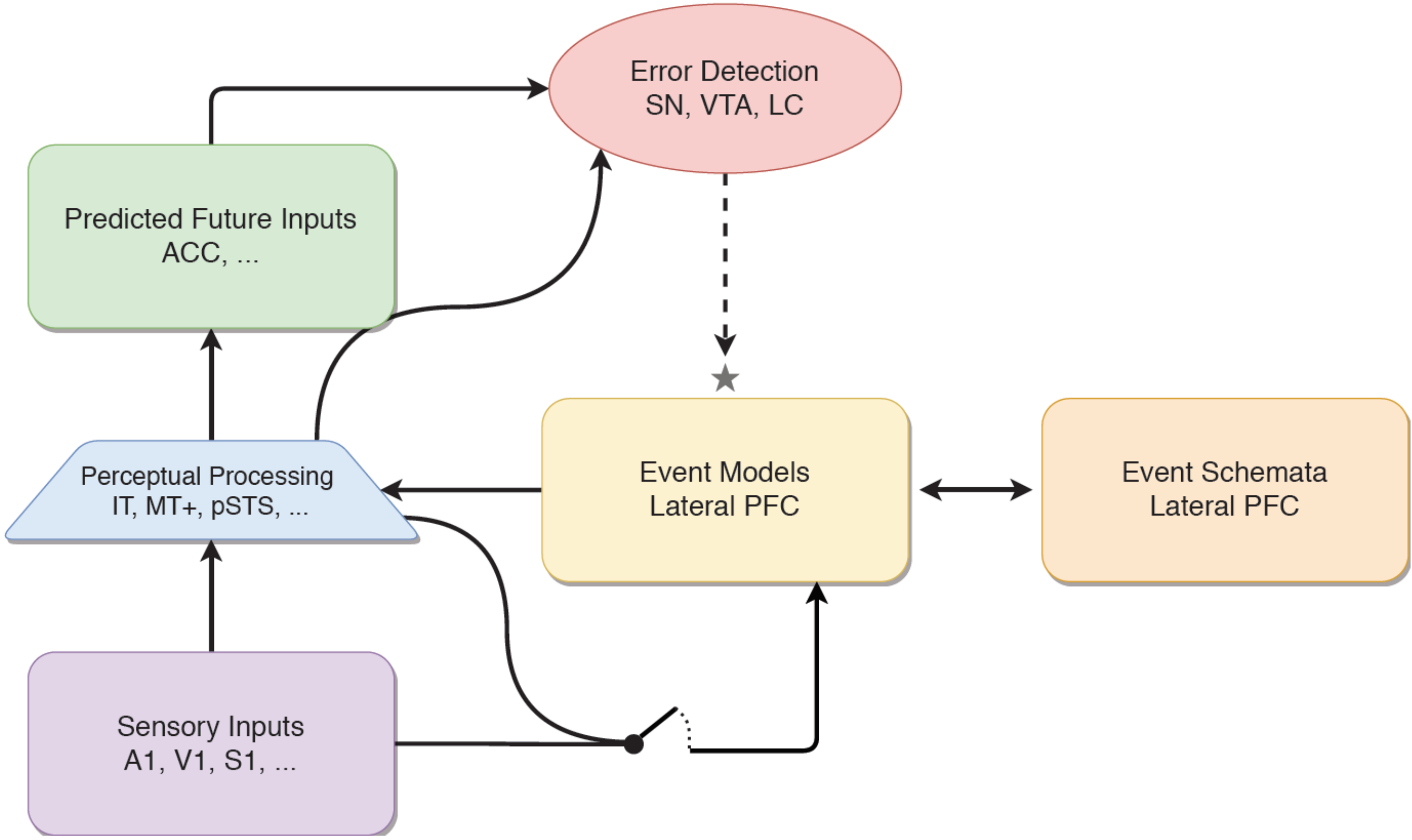






Event Segmentation Theory

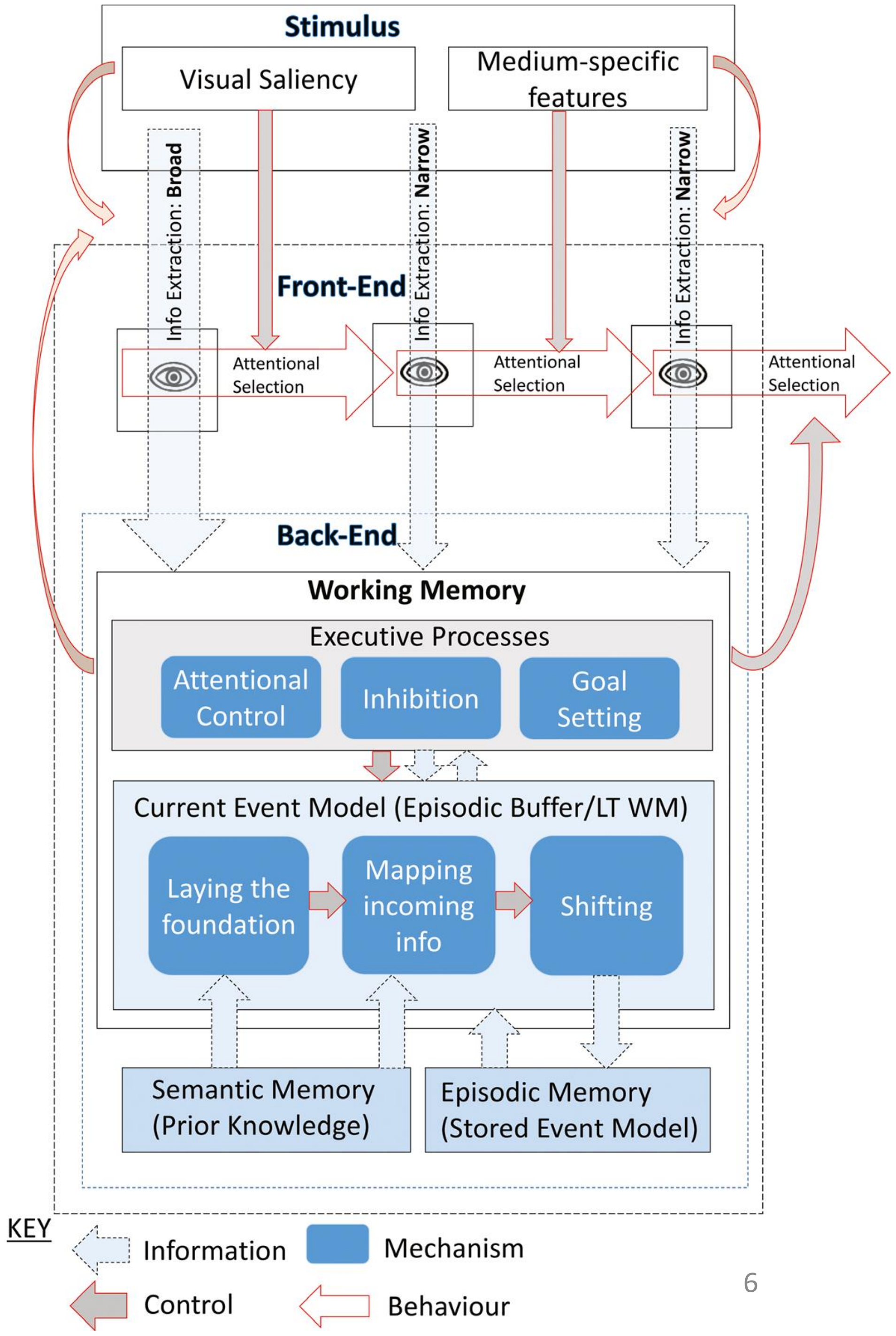
- Segmentation of ongoing activity is a spontaneous concomitant of ongoing perception.
- Event segmentation happens simultaneously on multiple timescales, though an observer may attend to a particular timescale.
- Event models are constructed through interaction of sensory input with stored knowledge



Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

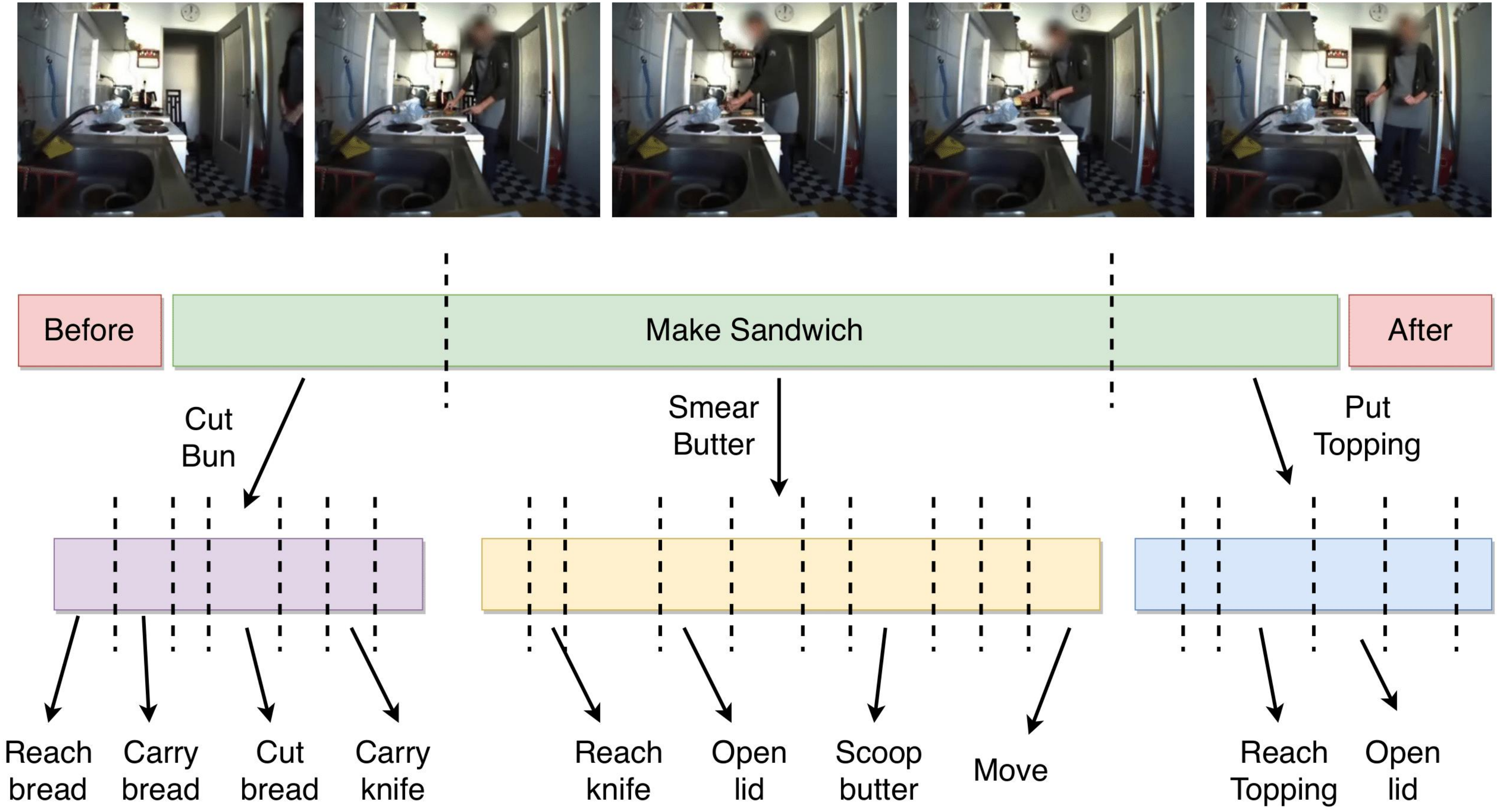
SPECT

- Front-end operates on single eye fixations, while back-end processing deals with multiple fixations
- Broad and narrow features and extracted from stimulus, which controls “Attentional Selection”
- Current event model is constructed over time in the working memory.
- Event models are stored in episodic memory (shifted) when they fail to explain or predict current features.

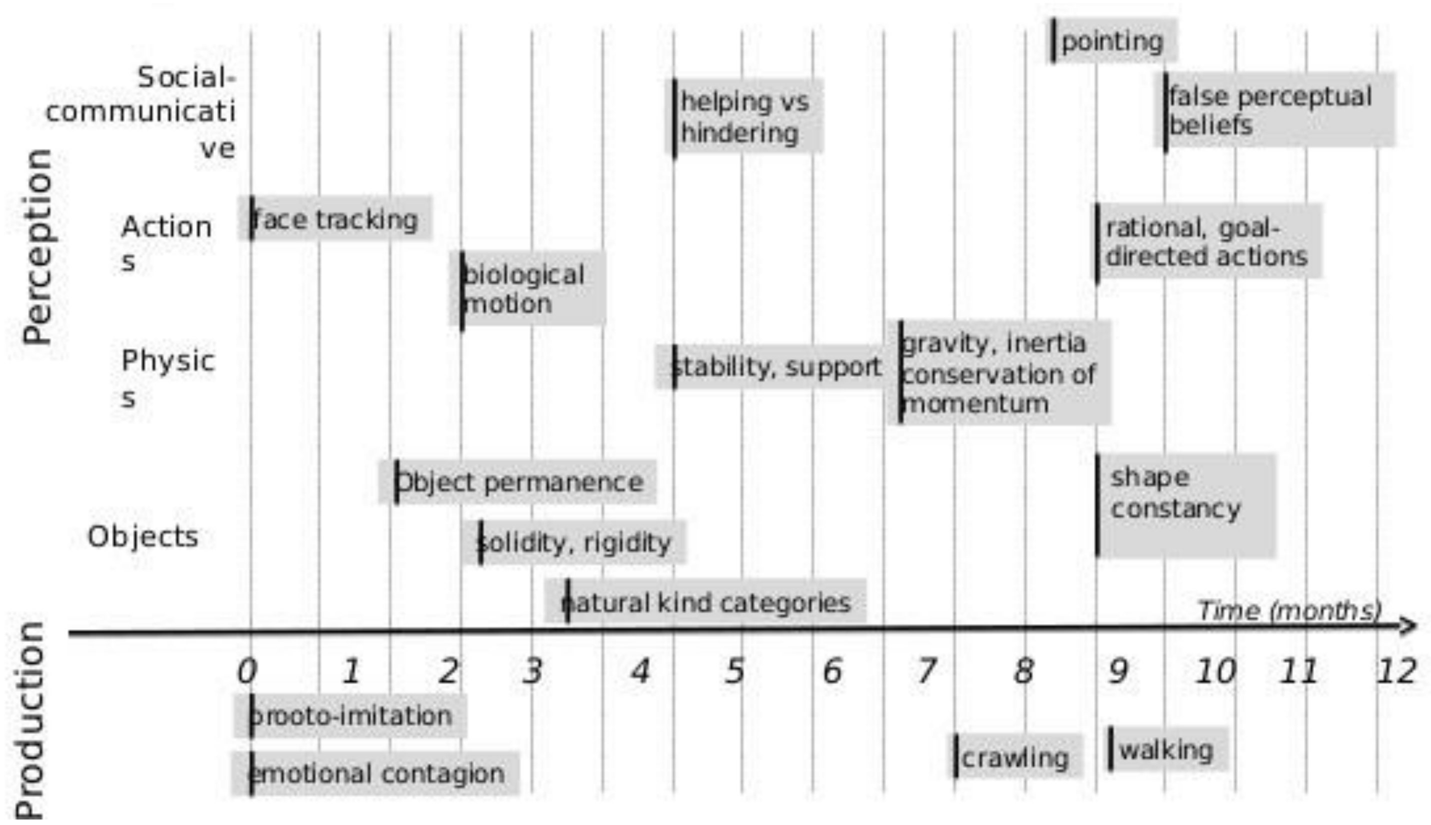


Loschky, Lester C., et al. "The scene perception & event comprehension theory (SPECT) applied to visual narratives." Topics in cognitive science 12.1 (2020): 311-351.

Hierarchy of Events



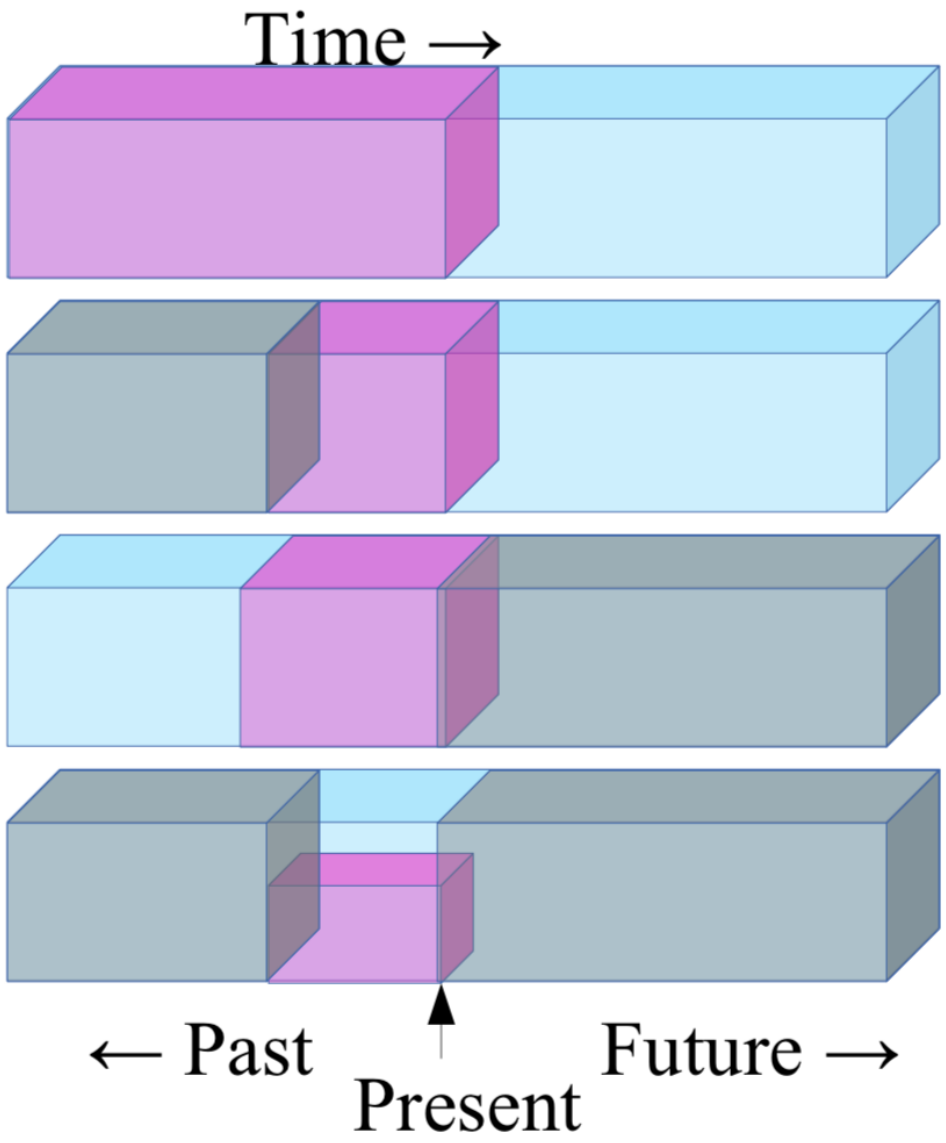
Early Conceptual Acquisition in Infants



Adopted from Yann Lecun Presentation on self-supervised learning (Image source: Emmanuel Dupoux)

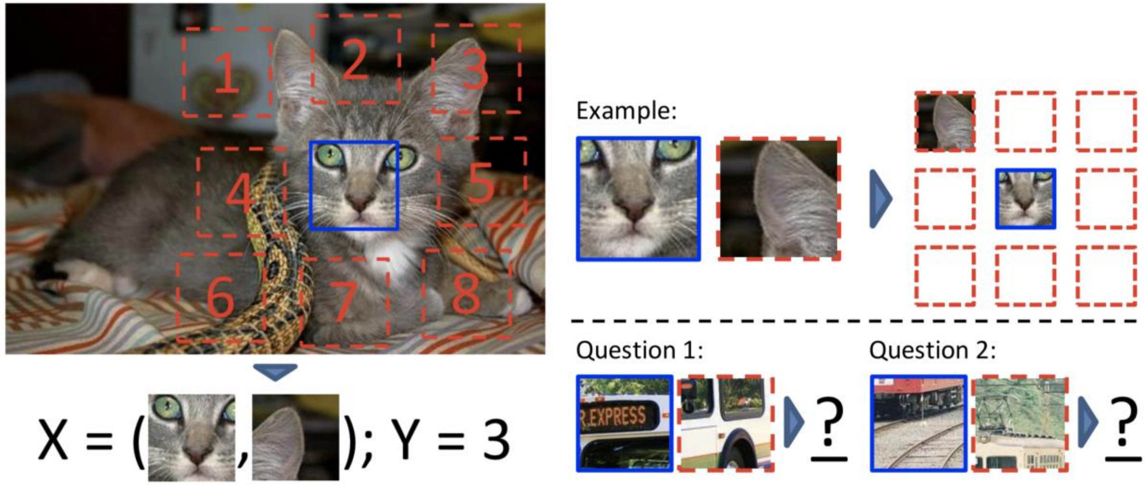
Self-supervised Learning

- Past predicts the future
- Recent past predicts future
- Present predicts the past
- Visible predicts occluded

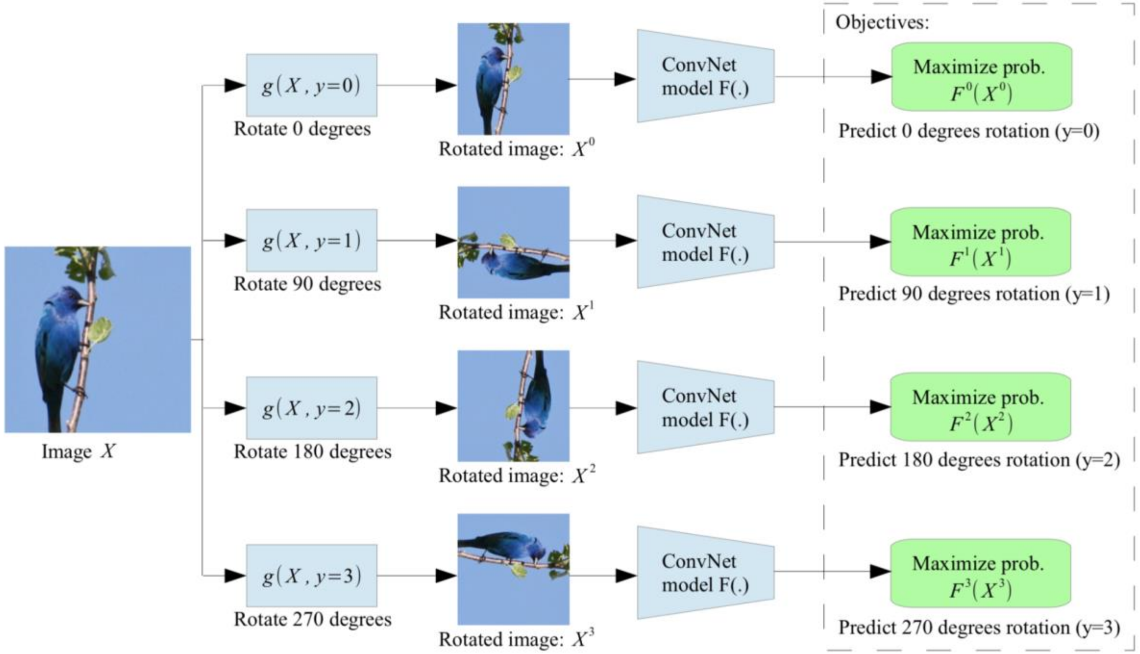


Adopted from Yann Lecun Presentation on self-supervised learning

Self-supervised Learning



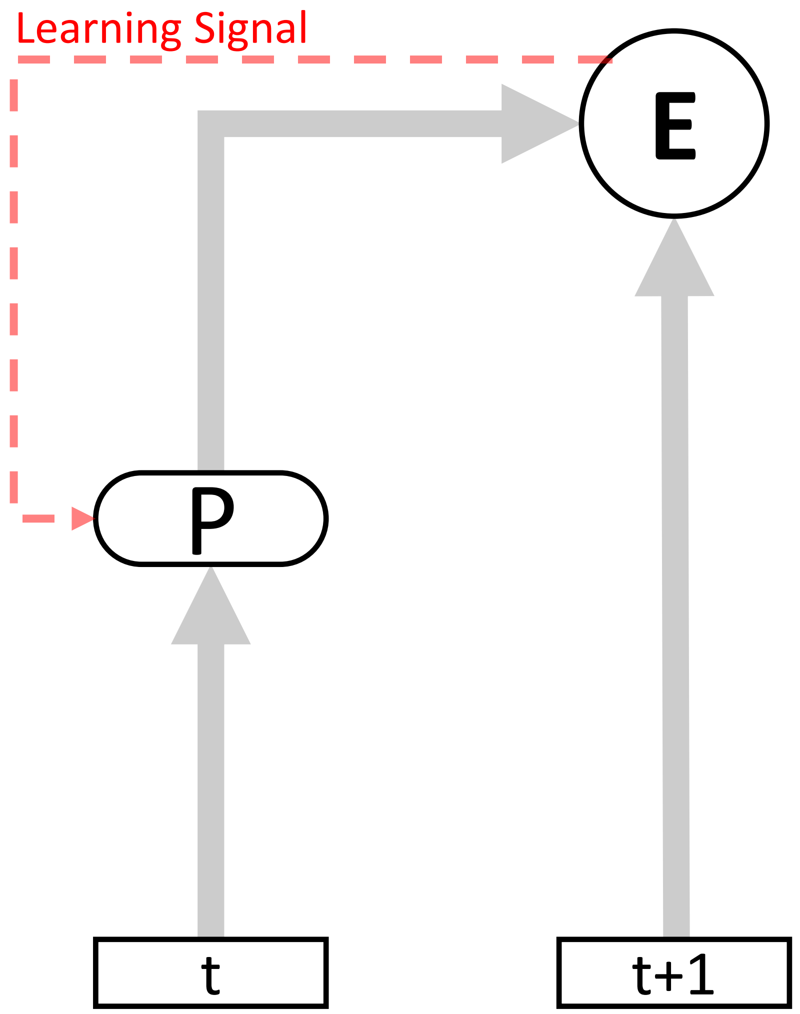
Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." Proceedings of the IEEE international conference on computer vision. 2015.



Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

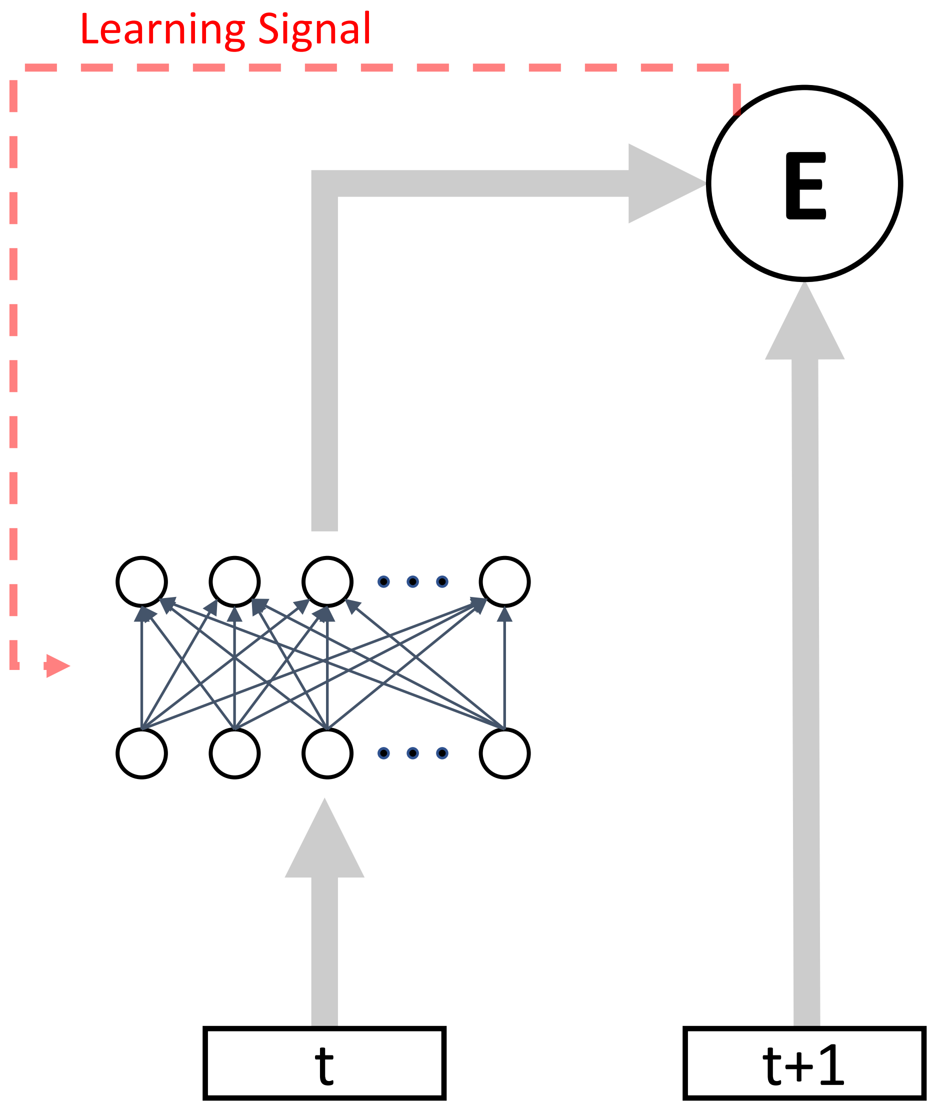
Predictive Learning Framework

- Prediction model predicts future time steps
- Prediction error improves prediction performance through gradient descend.
- Self-supervised training



Basic Predictive Learning Framework

```python
class Model(nn.Module):

    def __init__(self):

        super(Model, self).__init__()

        # ====== Define Layers ====== #
        self.conv1 = nn.Conv2d(3,16, (3,3), stride=1, padding=1)
        self.conv2 = nn.Conv2d(16,16, (3,3), stride=1, padding=1)
        self.conv3 = nn.Conv2d(16,3, (3,3), stride=1, padding=1)

        # ====== Define Loss Function ====== #
        self.loss_fn = nn.MSELoss()

    def forward(self, x):

        # ====== Define Architecture ====== #
        net = F.relu(self.conv1(x))
        net = F.relu(self.conv2(net))
        net = self.conv3(net)

        return net

model = Model()
x = torch.randn((1, 3, 224, 224)) # Input
y = torch.randn((1, 3, 224, 224)) # Label
pred = model(x)                   # Prediction (1,3,224,224)
loss = model.loss_fn(pred, y)     # Scalar loss
```



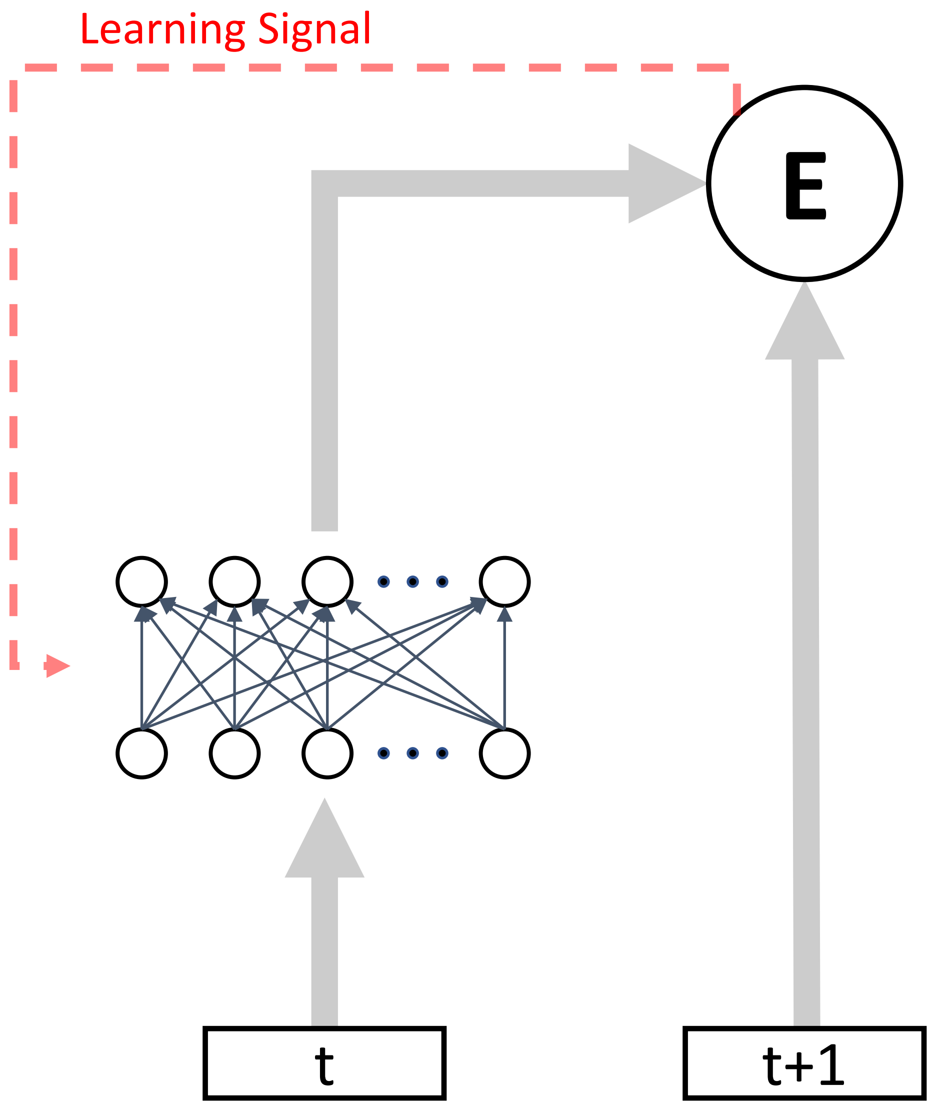
Basic Predictive Learning Framework

Naïve approach

- Use FFN (MLP/CNN) on raw input.
- Transform current perceptual input to future input.
- Loss signal trains the predictive function.

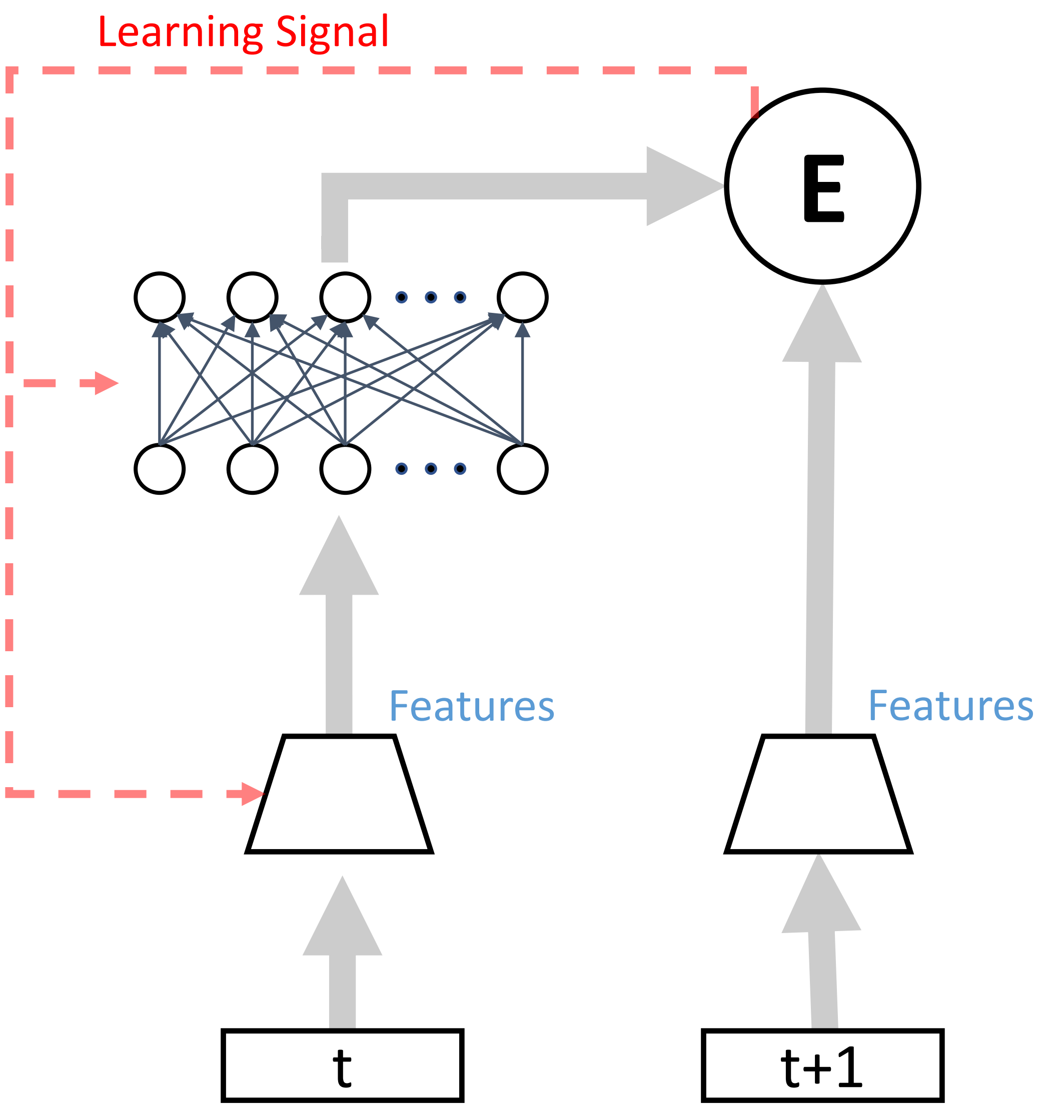
Problems

- Too much noise in input signal
- High space and time complexity



Basic Predictive Learning Framework

```python
class Model(nn.Module):

    def __init__(self):

        super(Model, self).__init__()

        # ====== Define Encoder ====== #
        self.encoder = nn.Sequential(nn.Conv2d(3,16, (3,3), 1), nn.ReLU(), nn.AvgPool2d((4,4), 4),
                                     nn.Conv2d(16,32, (3,3), 1), nn.ReLU(), nn.AvgPool2d((4,4), 4),
                                     nn.Conv2d(32,64, (3,3), 1), nn.ReLU(), nn.AvgPool2d((4,4), 4),
                                     nn.Flatten())

        # ====== Define Predictor ====== #
        self.predictor = nn.Linear(256, 256)

        # ====== Define Loss Function ====== #
        self.loss_fn = nn.MSELoss()

    def forward(self, x, y):

        # ====== Define Architecture ====== #
        x_features = self.encoder(x)
        y_features = self.encoder(y)

        pred = self.predictor(x_features)

        return pred, y_features

model = Model()
x = torch.randn((1, 3, 224, 224))          # Input
y = torch.randn((1, 3, 224, 224))          # Label
pred, y_features = model(x, y)             # Prediction (1,3,224,224)
loss = model.loss_fn(pred, y_features)     # Scalar loss
```



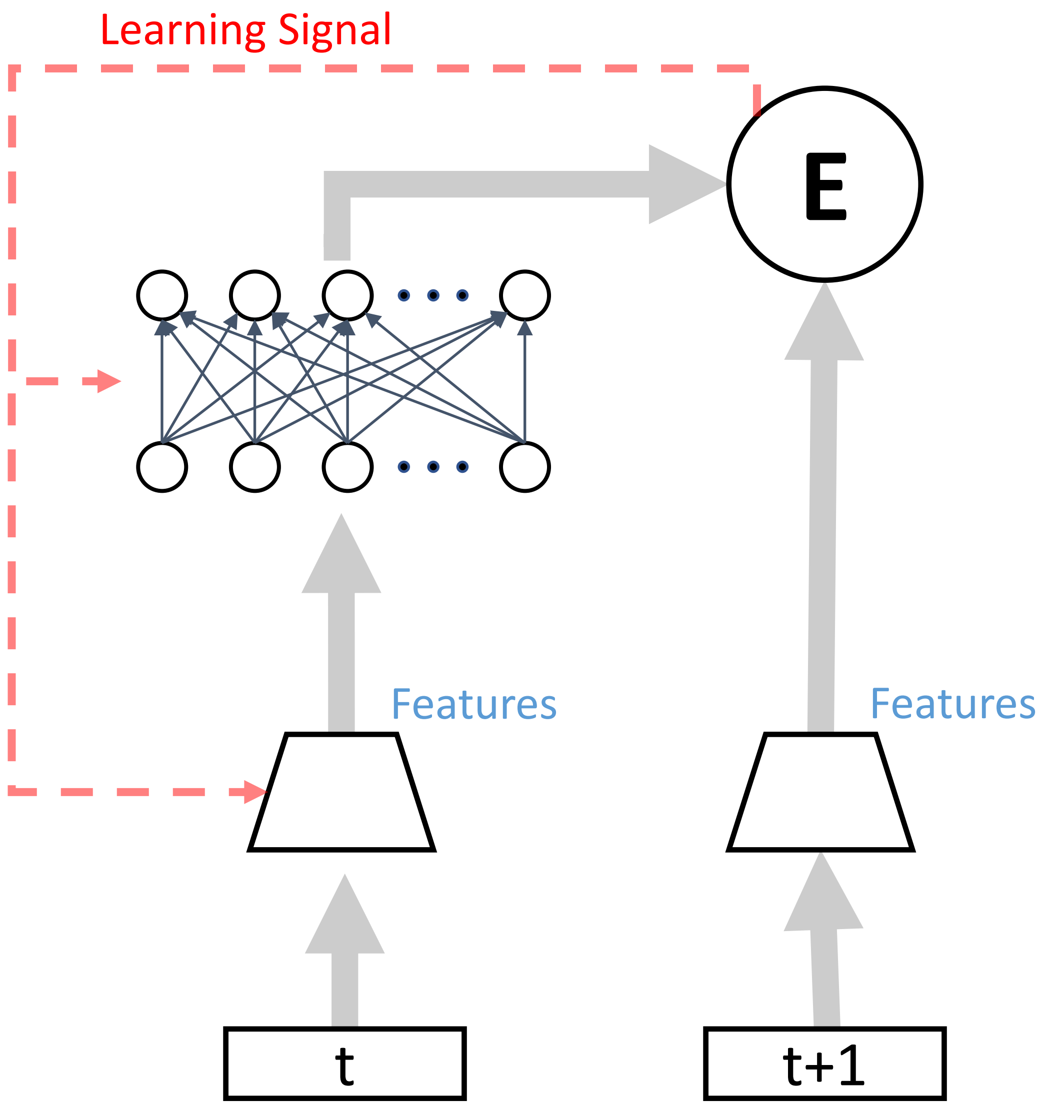
Basic Predictive Learning Framework

Solution

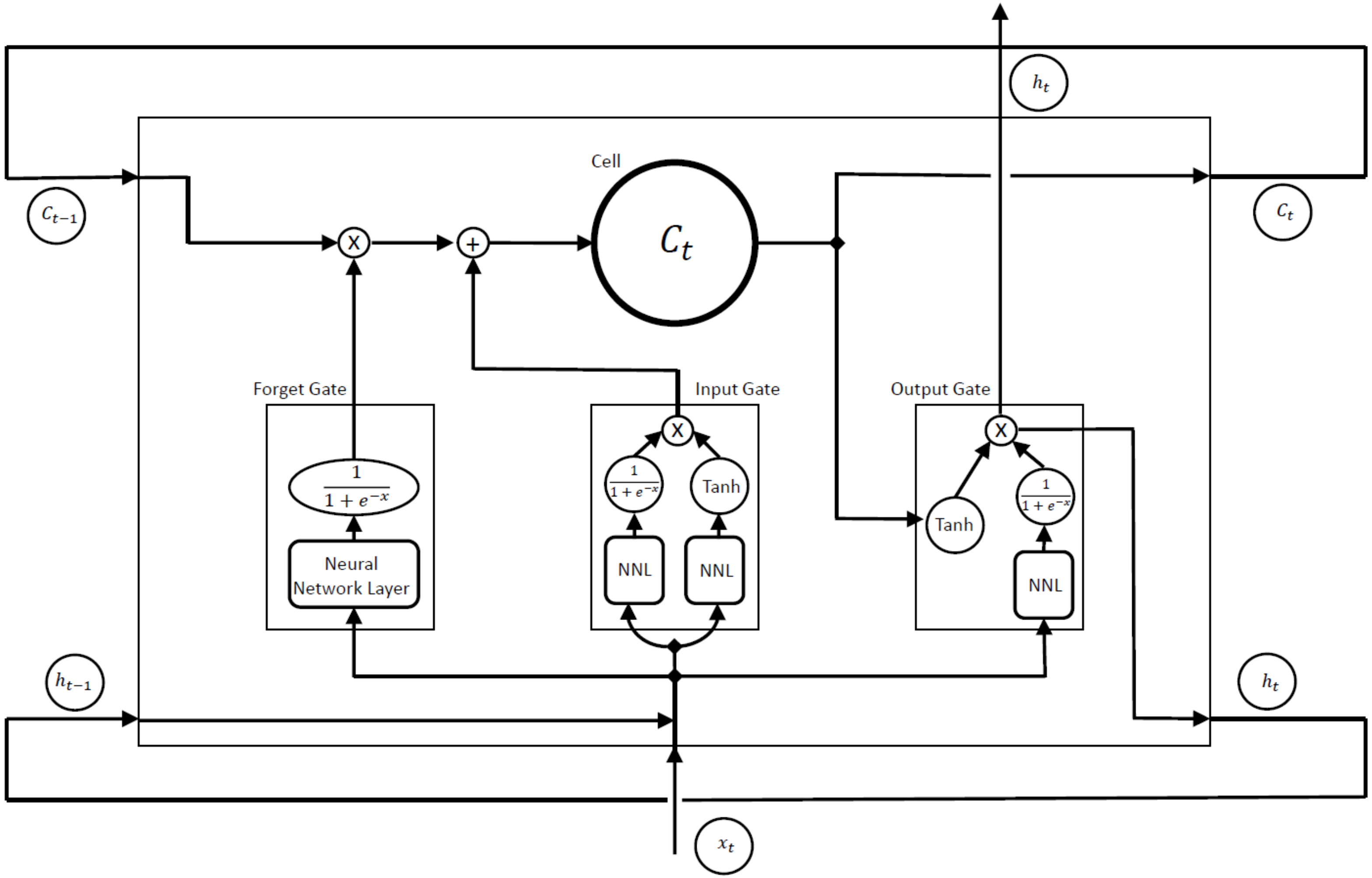
- Use a trainable feature extractor.
- Transform current features to future features.
- Loss signal trains the predictive function and feature extractor.

Problems

- Limited temporal receptive field
- Model requires features from the past to accurately predict the future

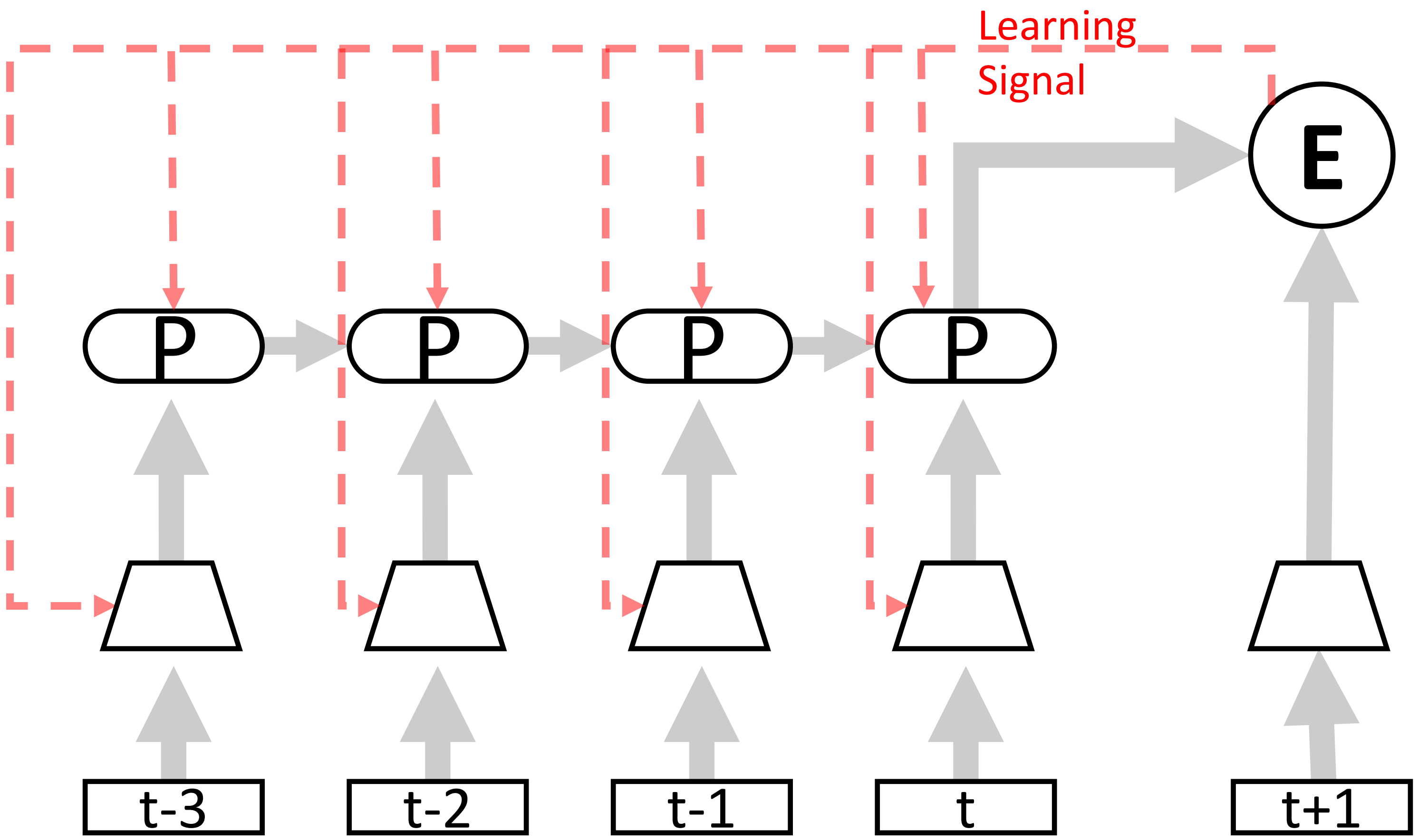


Basic Predictive Learning Framework



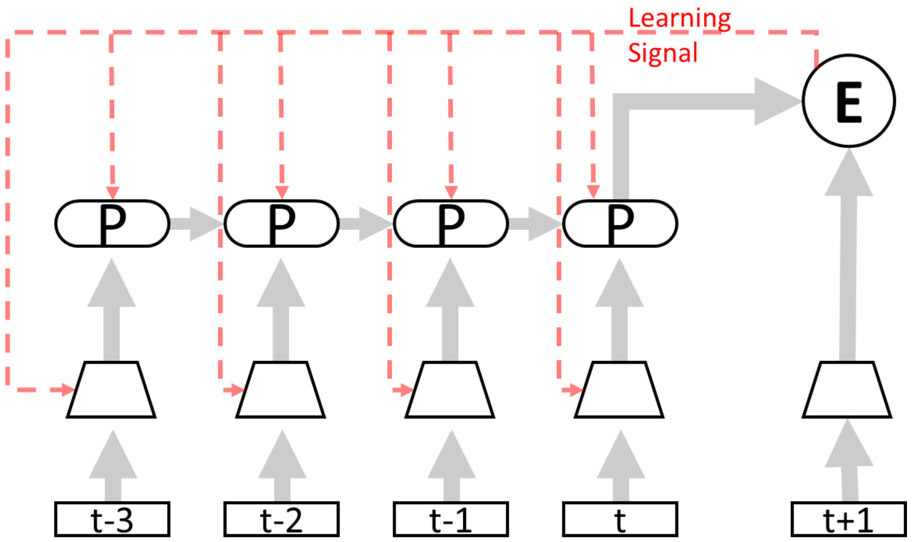
Mounir, Ramy, Redwan Alqasemi, and Rajiv Dubey. "Speech Assistance for Persons With Speech Impediments Using Artificial Neural Networks." ASME 2017 International Mechanical Engineering Congress and Exposition. American Society of Mechanical Engineers Digital Collection, 2017.

Basic Predictive Learning Framework



Basic Predictive Learning Framework

```python
class Model(nn.Module):

    def __init__(self):

        super(Model, self).__init__()

        # ====== Define Encoder ====== #
        self.encoder = nn.Sequential(nn.Conv2d(3,16, (3,3), 1), nn.ReLU(), nn.AvgPool2d((4,4), 4),
                                     nn.Conv2d(16,32, (3,3), 1), nn.ReLU(), nn.AvgPool2d((4,4), 4),
                                     nn.Conv2d(32,64, (3,3), 1), nn.ReLU(), nn.AvgPool2d((4,4), 4),
                                     nn.Flatten())

        # ====== Define Predictor ====== #
        self.predictor = nn.GRU(256, 256, 1)

        # ====== Define Loss Function ====== #
        self.loss_fn = nn.MSELoss()


    def forward(self, x, y):

        # ====== Define Architecture ====== #
        x_features = self.encoder(x).unsqueeze(1)
        y_features = self.encoder(y).unsqueeze(1)
        _, h = self.predictor(x_features)

        return h, y_features


model = Model()
x = torch.randn((4, 3, 224, 224))           # Input
y = torch.randn((1, 3, 224, 224))           # Label
pred, y_features = model(x, y)              # Prediction
loss = model.loss_fn(pred, y_features)      # Scalar loss
```



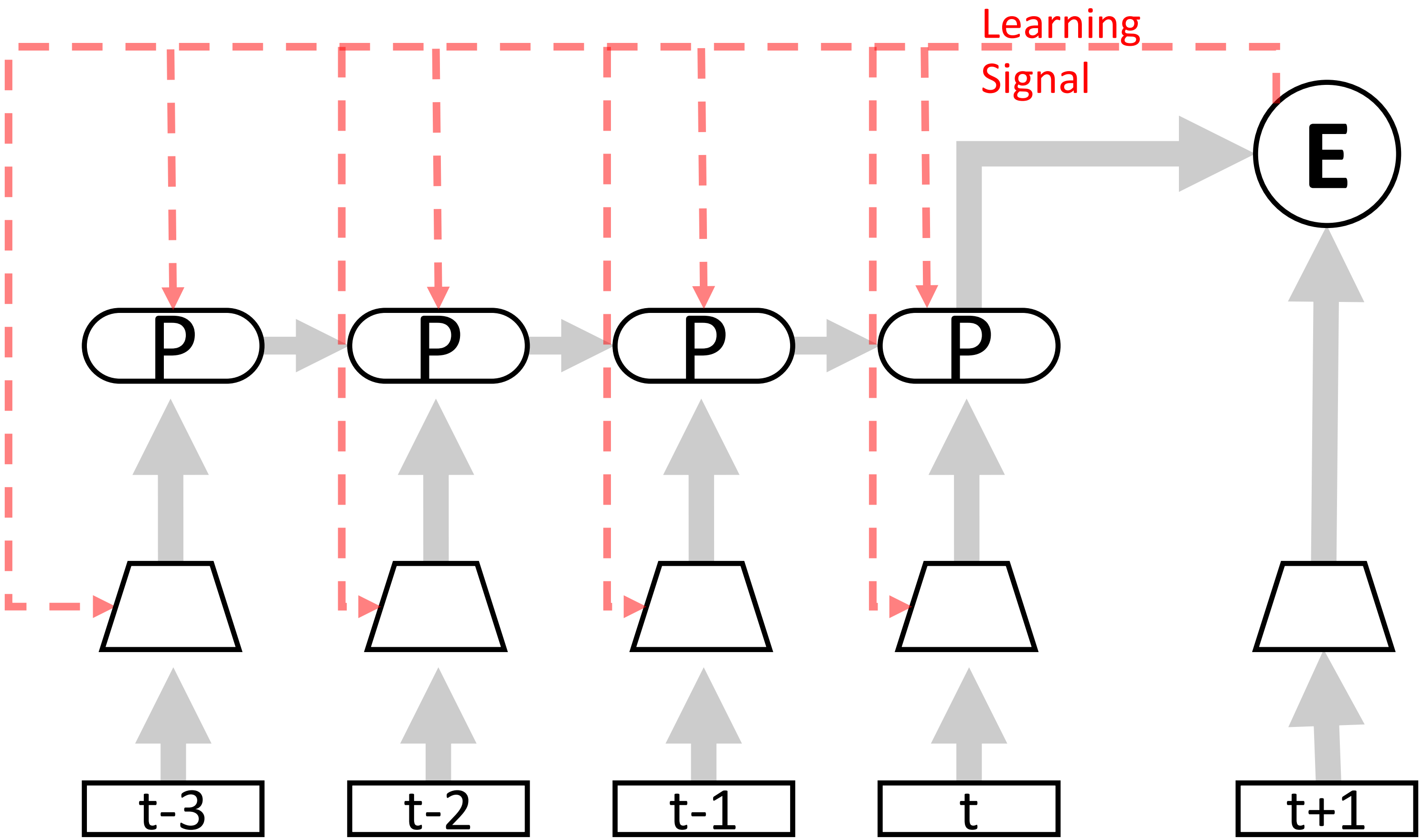
Basic Predictive Learning Framework

Solution

- Use recurrent model with internal memory.

Problems

- Unit of t is undefined.
- Different timescales.



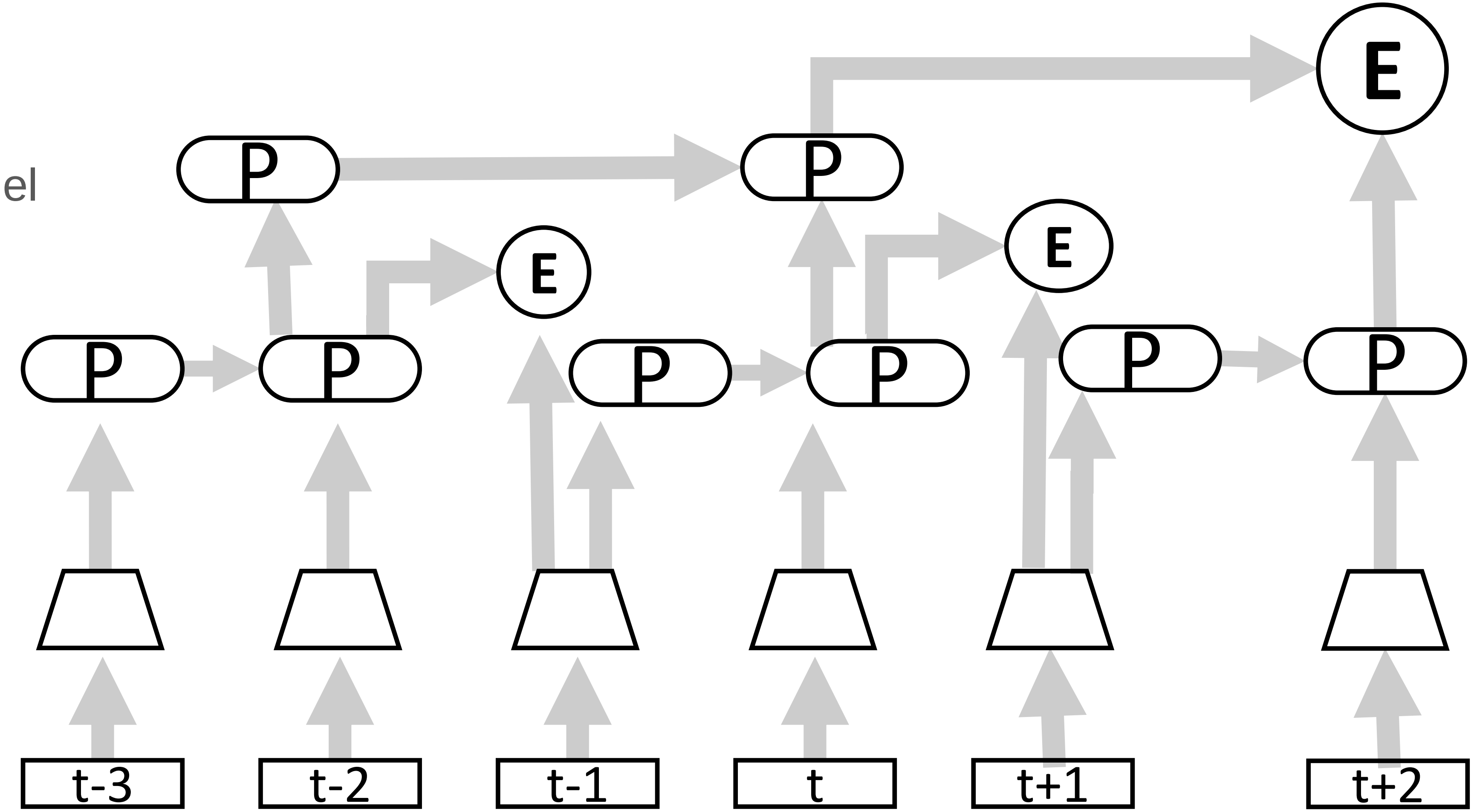
Basic Predictive Learning Framework

Solution

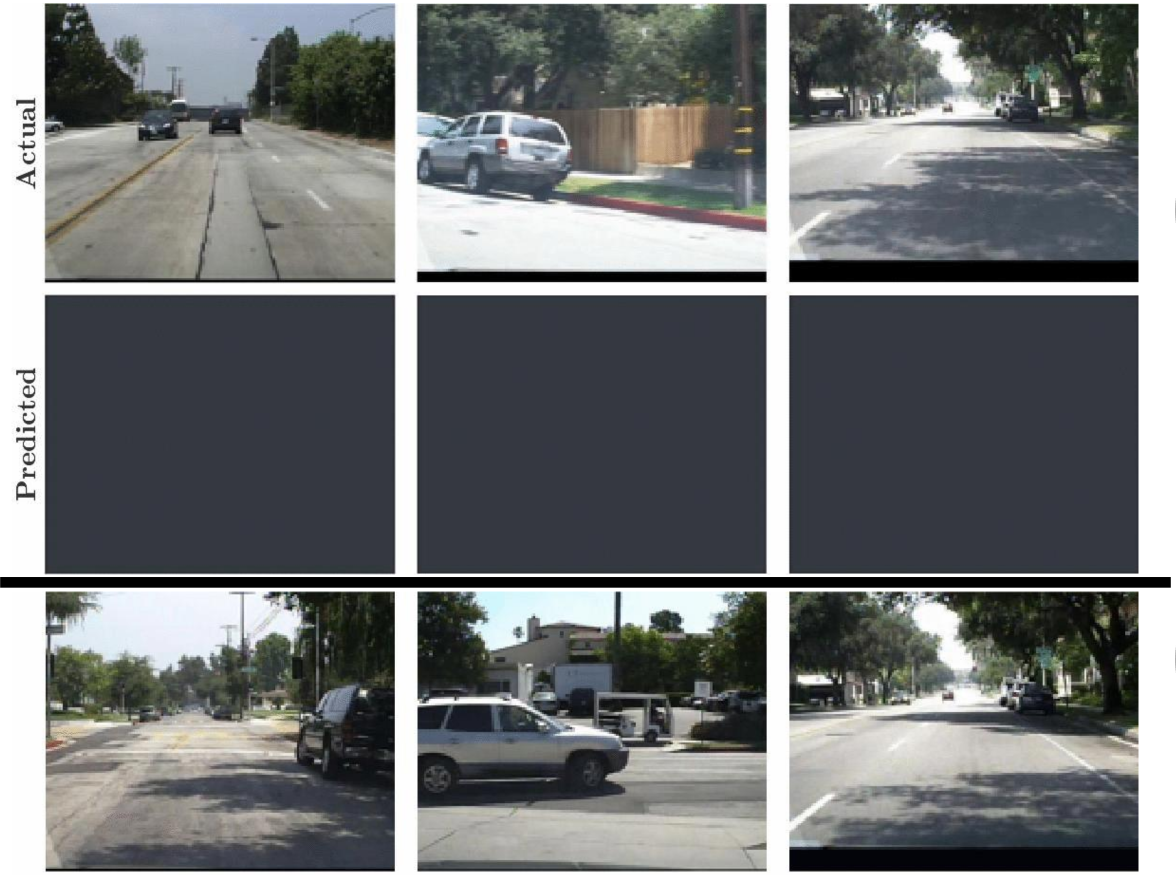
- Build hierarchical model
- Represent multiple timescales

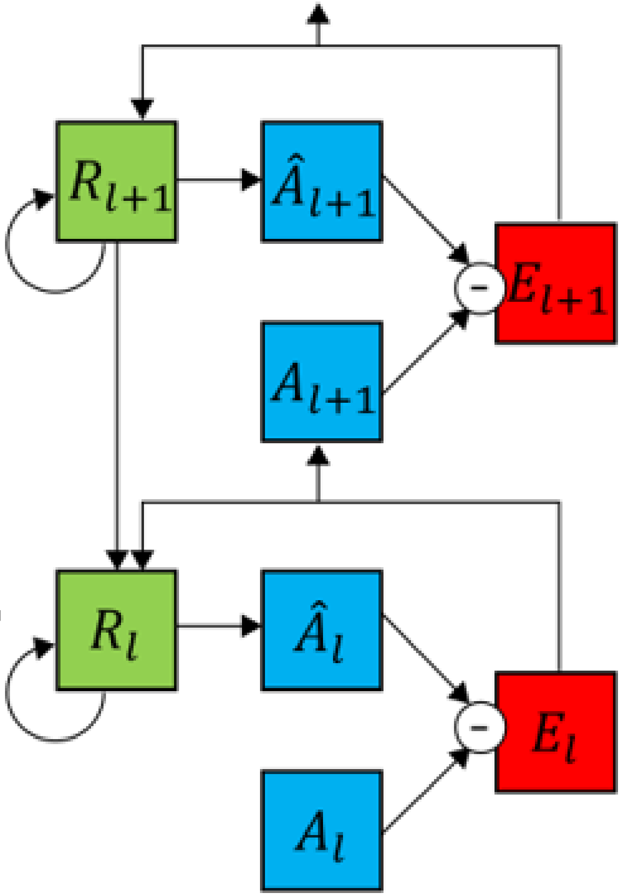
Problems

- How many levels?
- Temporal pooling?



PredNet Architecture

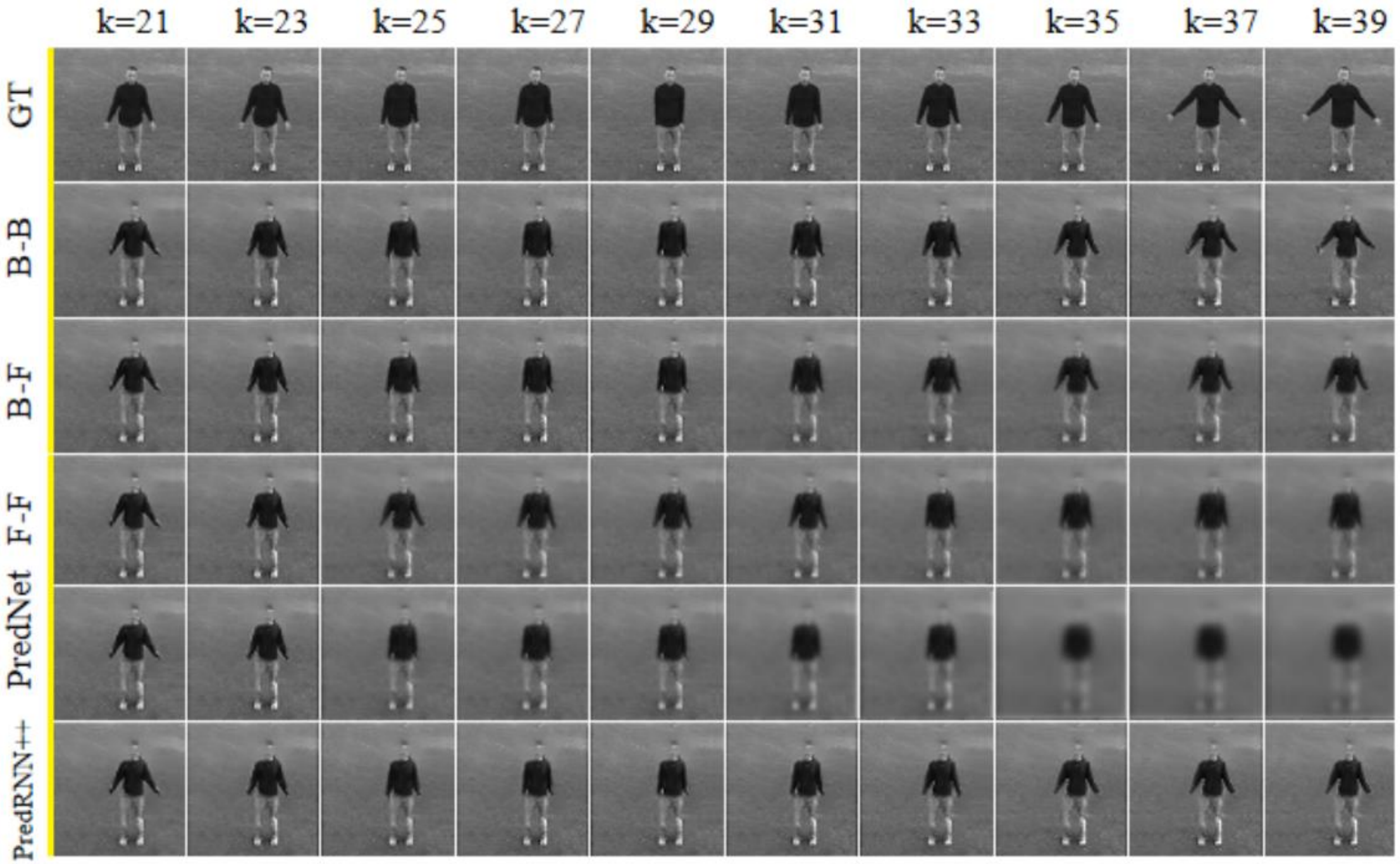


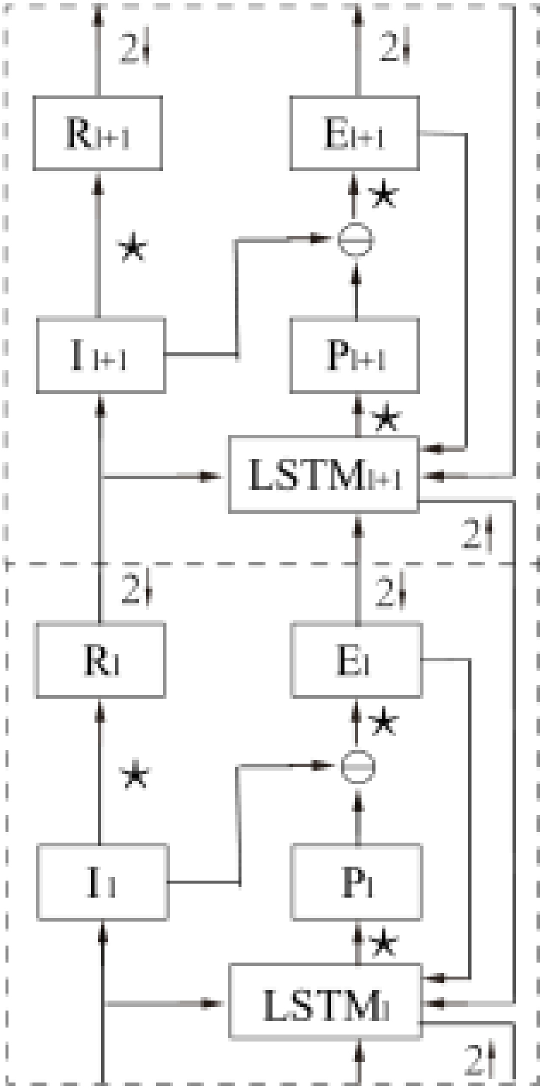


consider a two layer network

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

HPNet Architecture





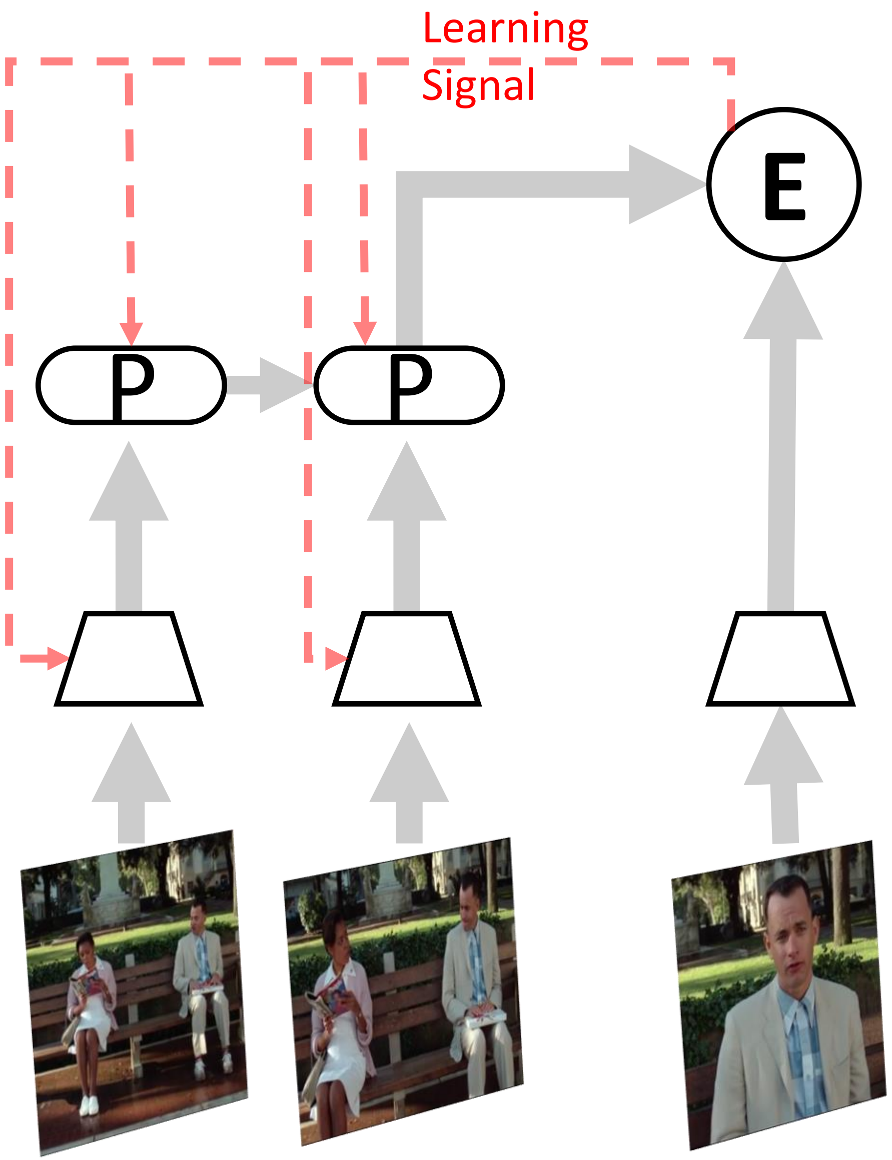
Qiu, J., Huang, G., & Lee, T. S. (2019). A neurally-inspired hierarchical prediction network for spatiotemporal sequence learning and prediction. *arXiv preprint arXiv:1901.09002*.

Importance of Prediction Error

- Prediction error is propagated backwards to train the feature extractor and the prediction function.
- Prediction error is reasonably indicative of event boundaries, pointing to the beginning and ending of events.
- The magnitude of prediction error indicates the temporal segmentation granularity.
- Spatial prediction error can be used for action localization.

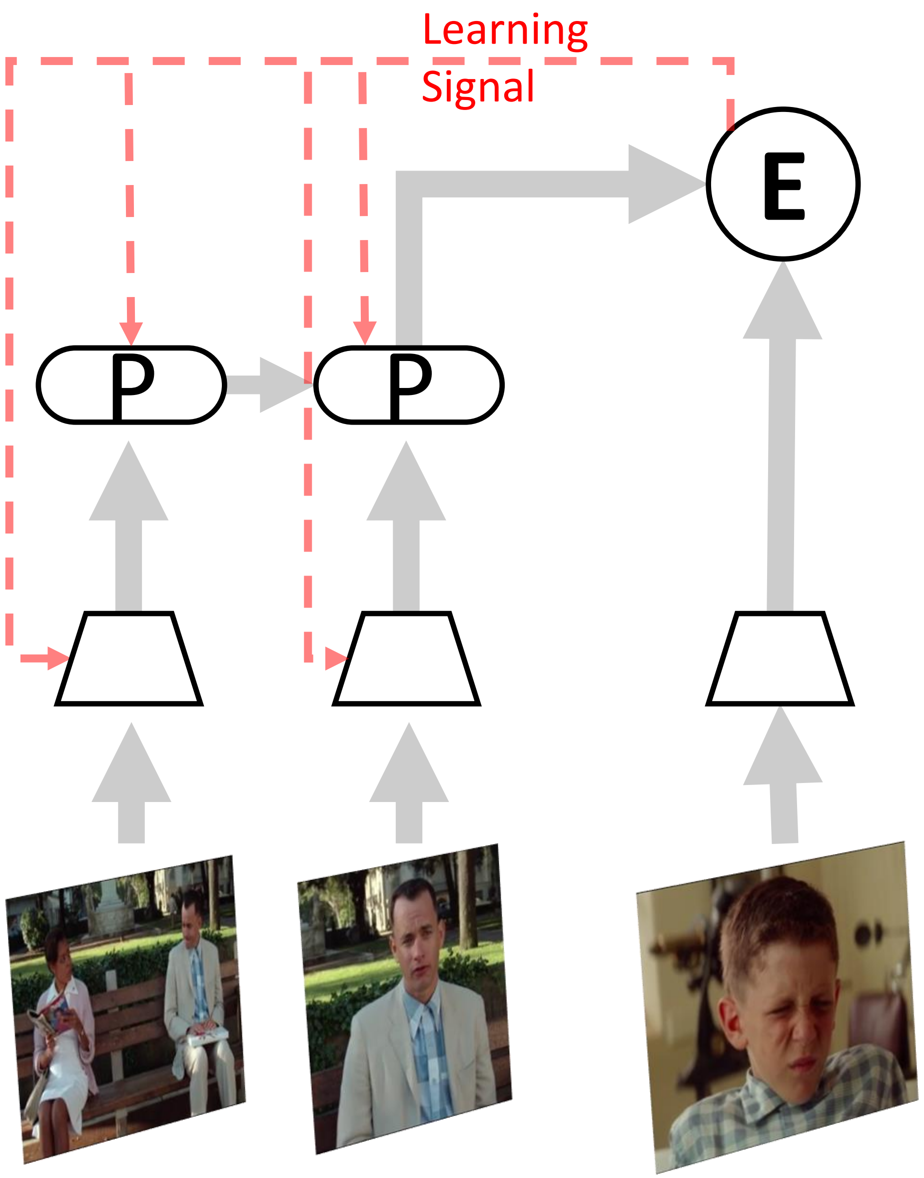
Importance of Prediction Error

Low prediction error is expected within an event, given a good prediction model.

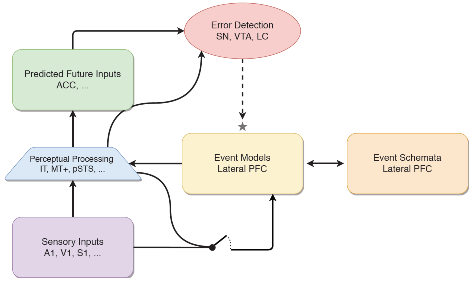


Importance of Prediction Error

High prediction error is expected within an event, given a good prediction model.



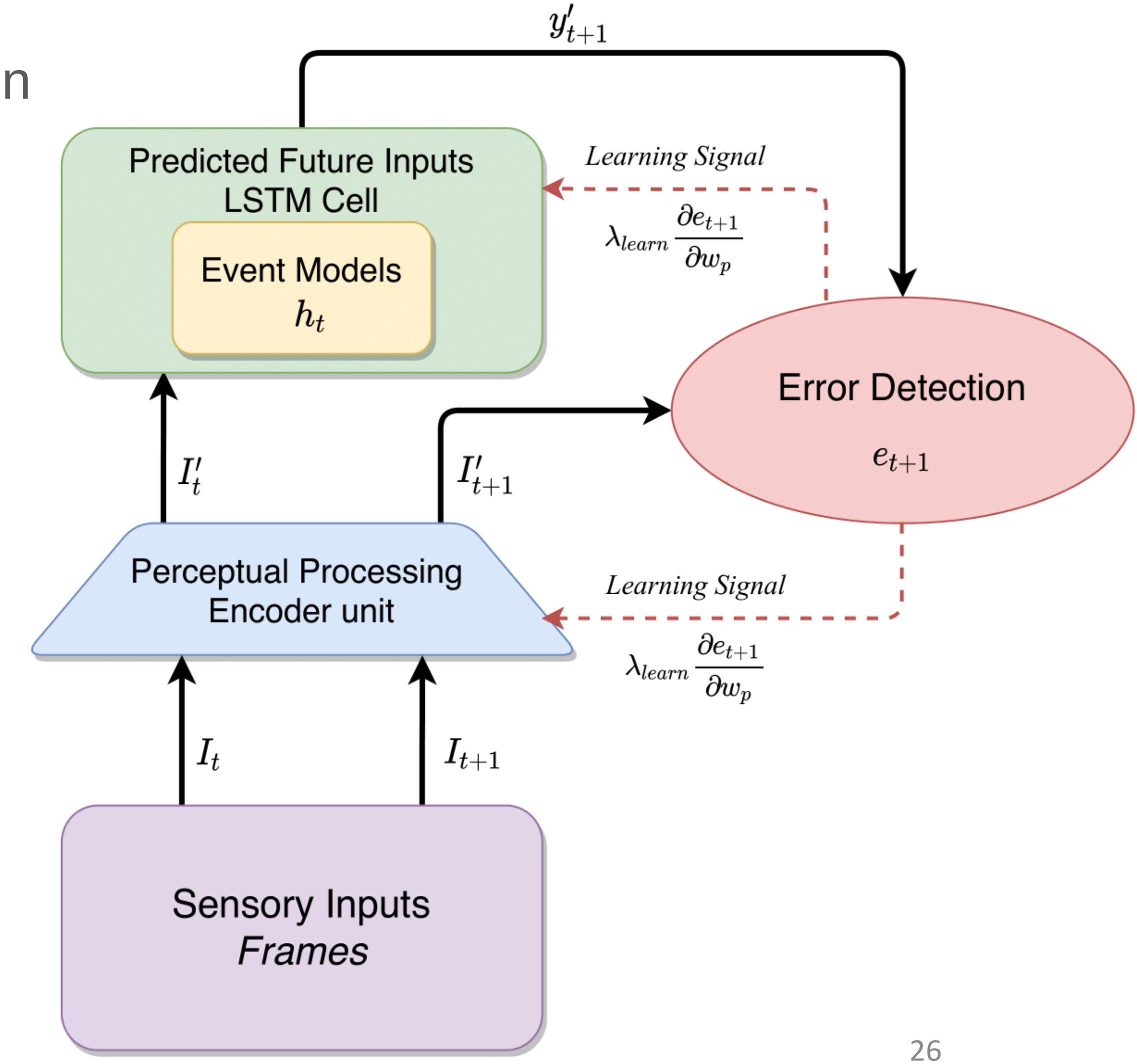
Self Supervised Event Segmentation



- LSTM cell for internal memory of event model.
- Error detection implemented as a low pass filter on prediction error running average.
- Gating signal triggered when error is above 1.5 times the running average.
- Adaptive learning controls learning rate

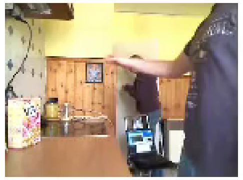
$$P_q(t) = P_q(t-1) + \frac{1}{n}(E_P(t) - P_q(t-1))$$

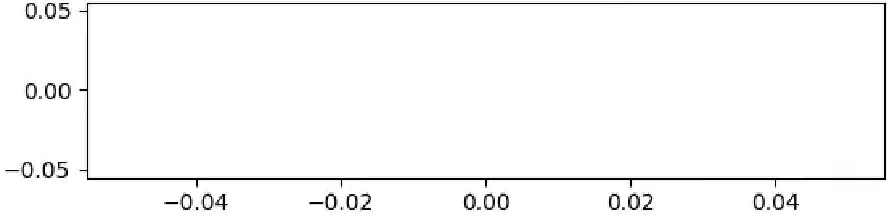
$$G(t) = \begin{cases} 1, & \frac{E_P(t)}{P_q(t-1)} > \psi_e \\ 0, & \text{otherwise} \end{cases}$$

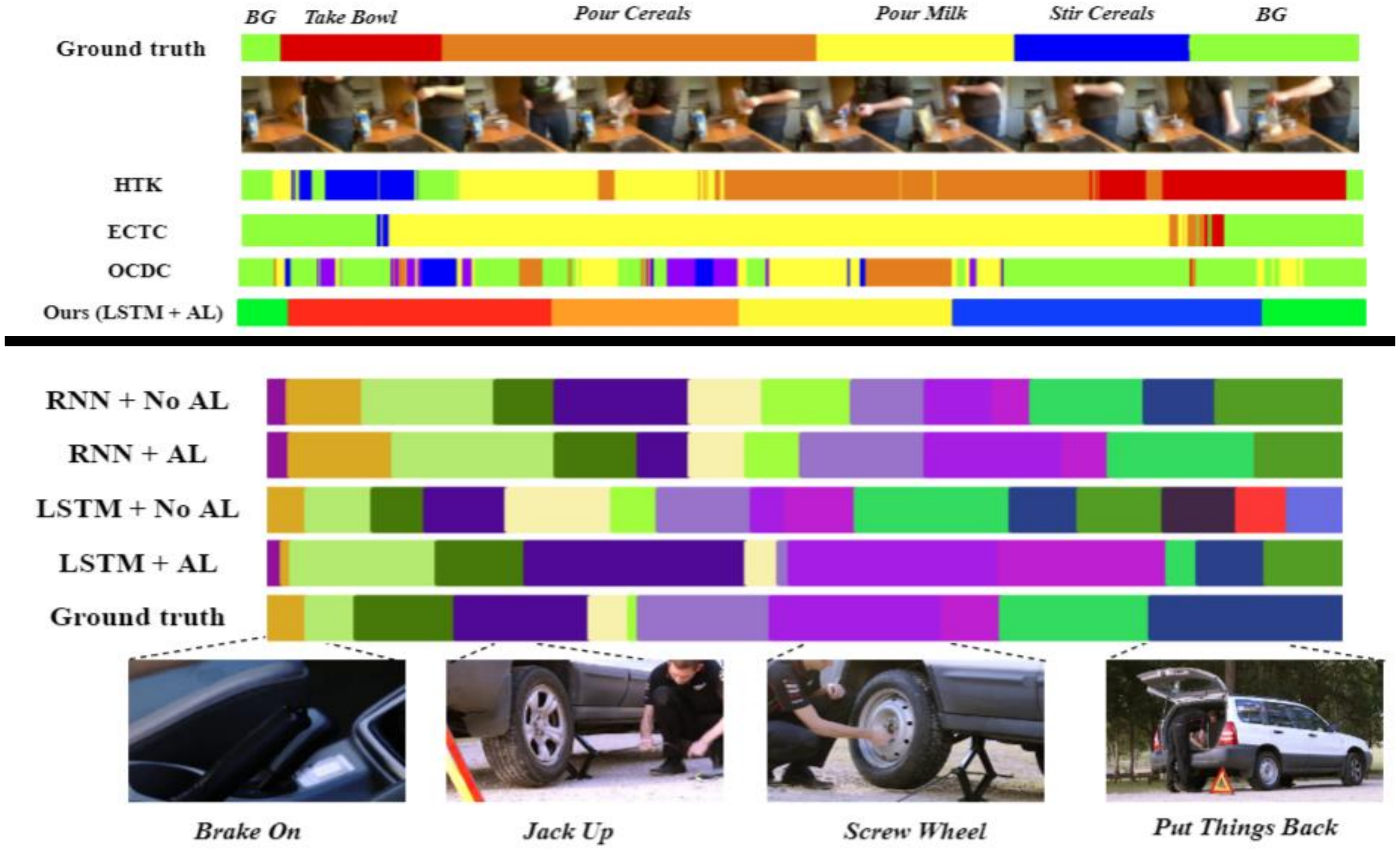


Aakur, S. N., & Sarkar, S. (2019). A perceptual prediction framework for self supervised event segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1197-1206).

Self Supervised Event Segmentation







Self Supervised Event Segmentation

Supervision	Approach	MoF	IoU
Full	SVM [19]	15.8	-
	HTK(64)[20]	56.3	-
	ED-TCN[27]	43.3	42.0
	TCFPN[10]	52.0	**54.9**
	GRU[29]	**60.6**	-
Weak	OCDC[6]	8.9	23.4
	ECTC[16]	27.7	-
	Fine2Coarse[28]	33.3	**47.3**
	TCFPN + ISBA[10]	**38.4**	40.6
None	KNN+GMM[30]	34.6	**47.1**
	Ours (LSTM + AL)	**42.9**	46.9

Table 1: Segmentation Results on the Breakfast Action dataset. MoF refers to the Mean over Frames metric and IoU is the Intersection over Union metric.

Supervision	Approach	MoF
Full	VGG**[21]	7.6%
	IDT**[21]	54.3%
	S-CNN + LSTM[21]	66.6%
	TDRN[22]	68.1%
	ST-CNN + Seg[21]	72.0%
	TCN[27]	**73.4%**
None	LSTM + KNN[4]	54.0%
	Ours (LSTM + AL)	**60.6%**

Table 2: Segmentation Results on the 50 Salads dataset, at granularity '*Eval*'. **Models were intentionally reported without temporal constraints for ablative studies.

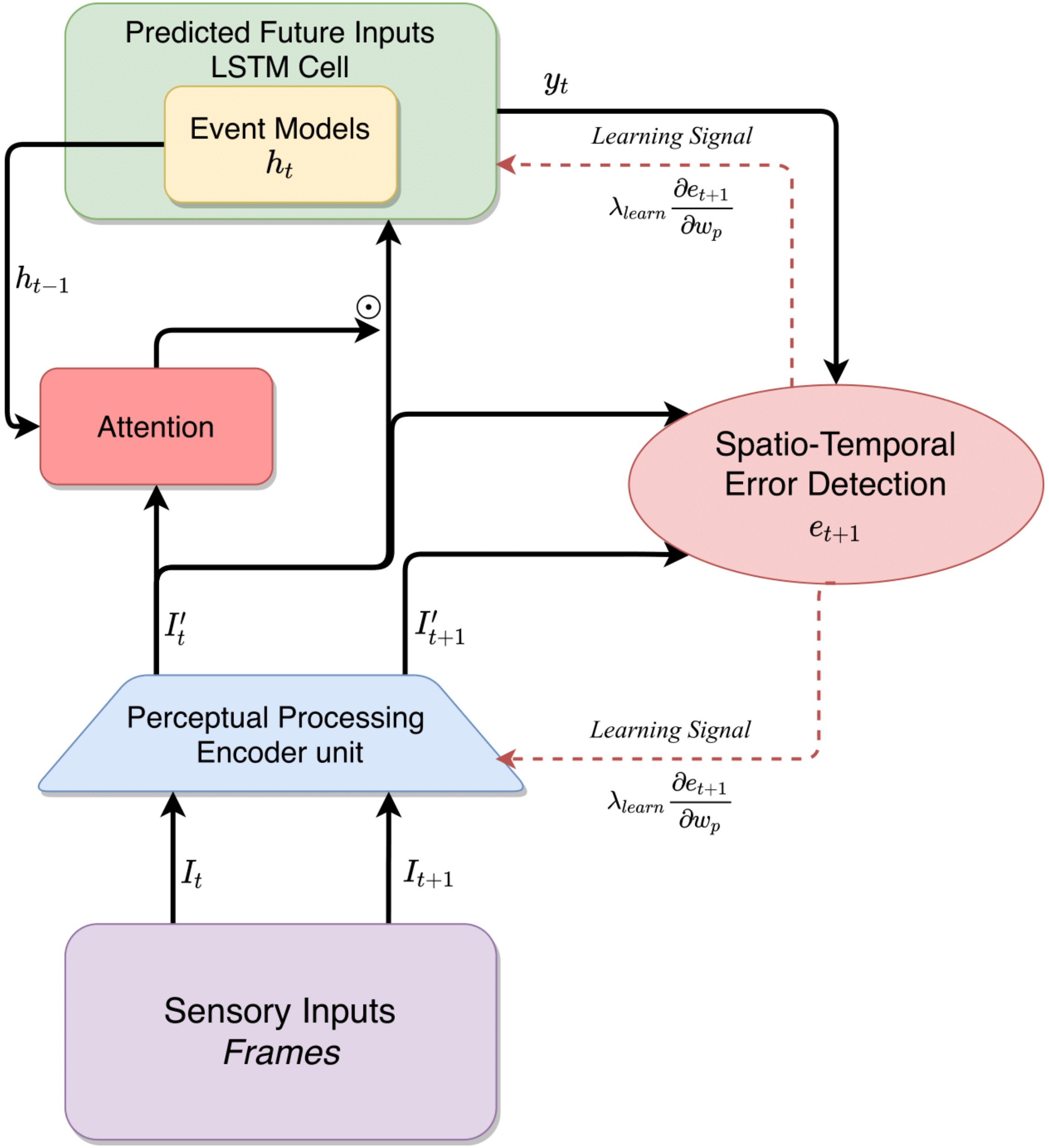
Supervision	Approach	F1
Full	HMM + Text [24]	22.9%
	Discriminative Clustering[3]	41.4%
	KNN+GMM[30] + GT	**69.2%**
Weak	OCDC + Text Features [6]	28.9%
	OCDC [6]	**31.8%**
None	KNN+GMM[30]	32.2%
	Ours (RNN + No AL)	25.9%
	Ours (RNN + AL)	29.4%
	Ours (LSTM + No AL)	36.4%
	Ours (LSTM + AL)	**39.7%**

Table 3: Segmentation Results on the INRIA Instructional Videos dataset. We report F1 score for fair comparison.

Wildlife Extended Videos

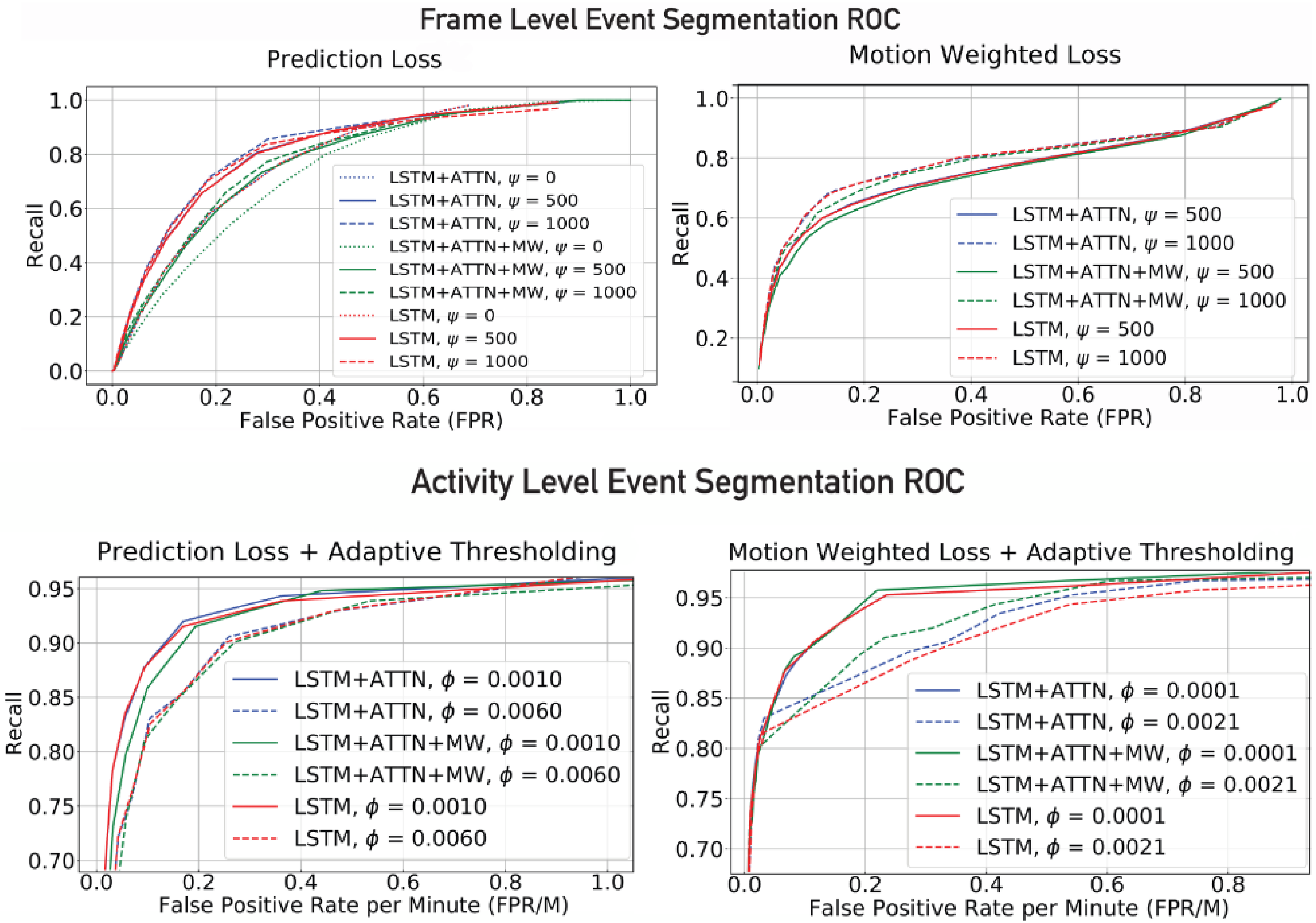
- Bahdanau attention is used to visualize the location of the bird.
- Motion-weighted loss is used instead of pure prediction loss.

$$e_t = ||(I'_{t+1} - y'_t)^{\odot 2} \odot (I'_{t+1} - I'_t)^{\odot 2}||^2$$

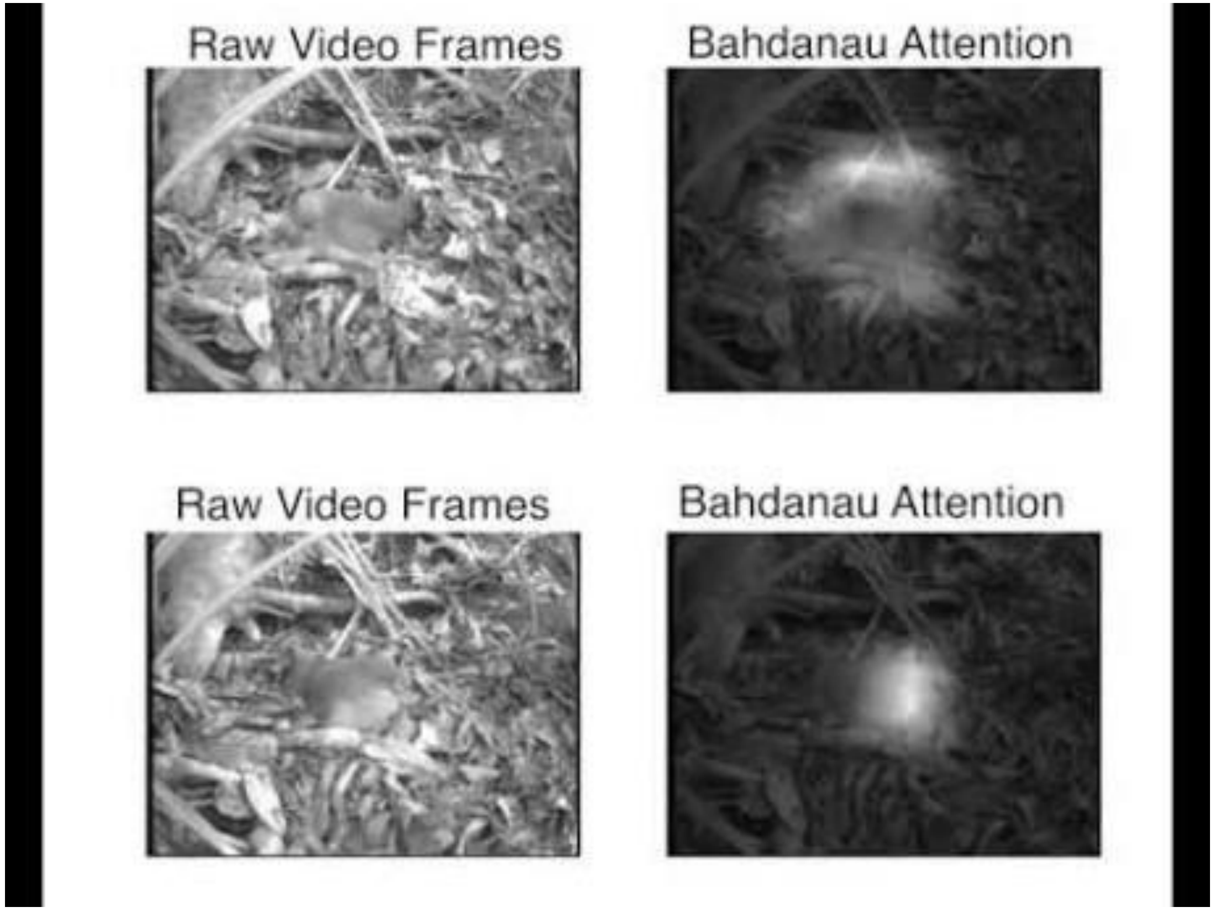


Mounir, R., Gula, R., Theuerkauf, J., & Sarkar, S. (2020). Temporal Event Segmentation using Attention-based Perceptual Prediction Model for Continual Learning. *arXiv preprint arXiv:2005.02463*.

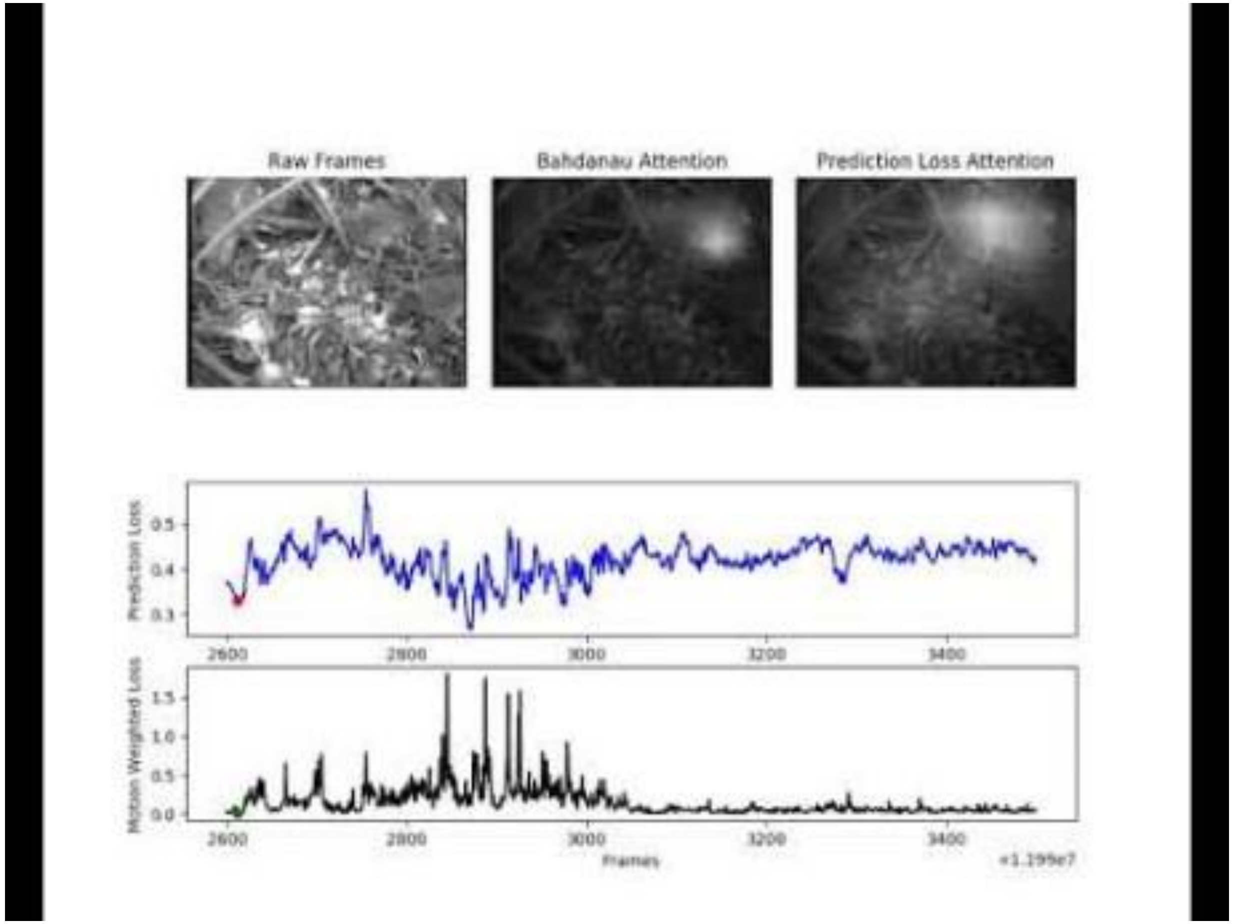
Wildlife Extended Videos

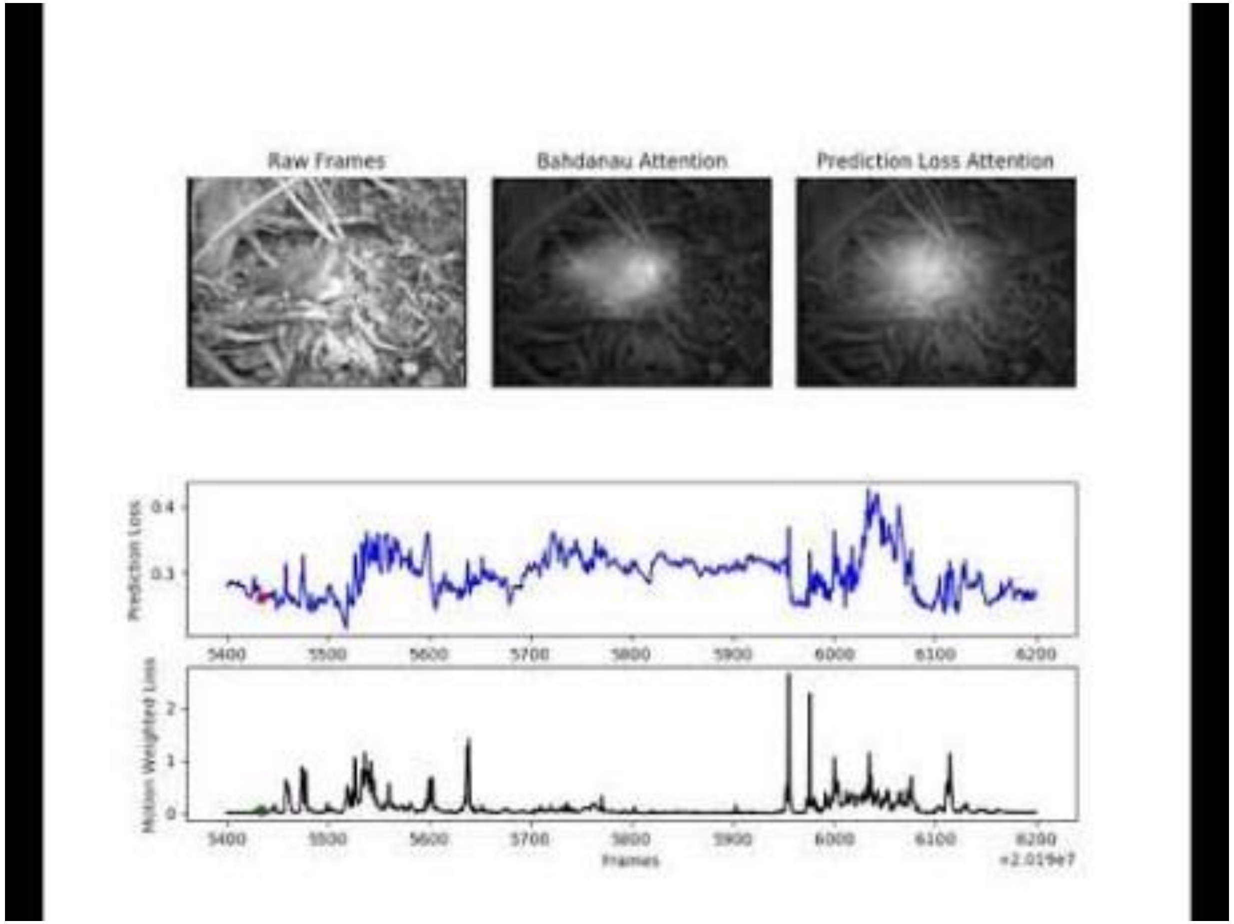


Wildlife Extended Videos



Wildlife Extended Videos

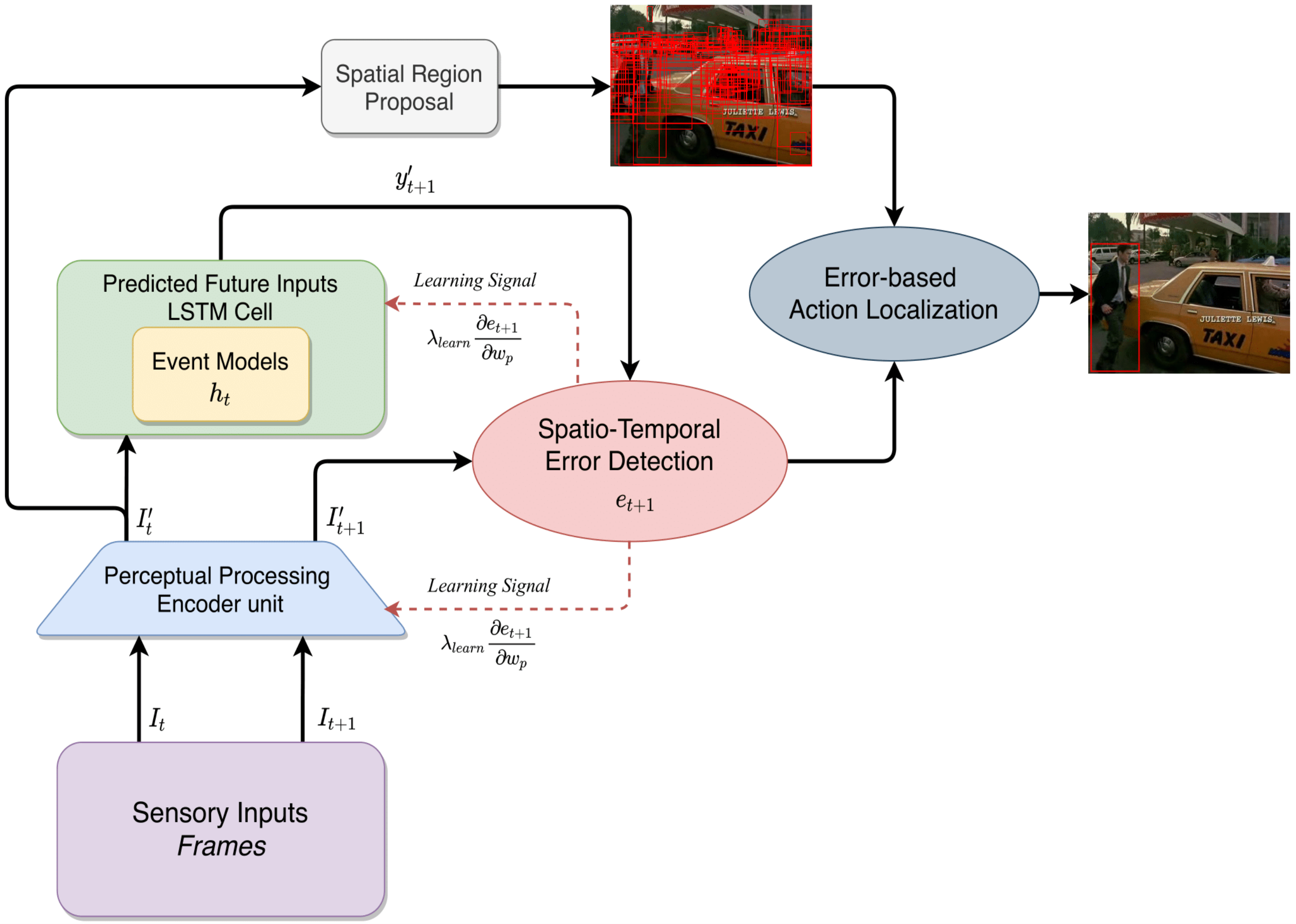




Energy-based Action Localization

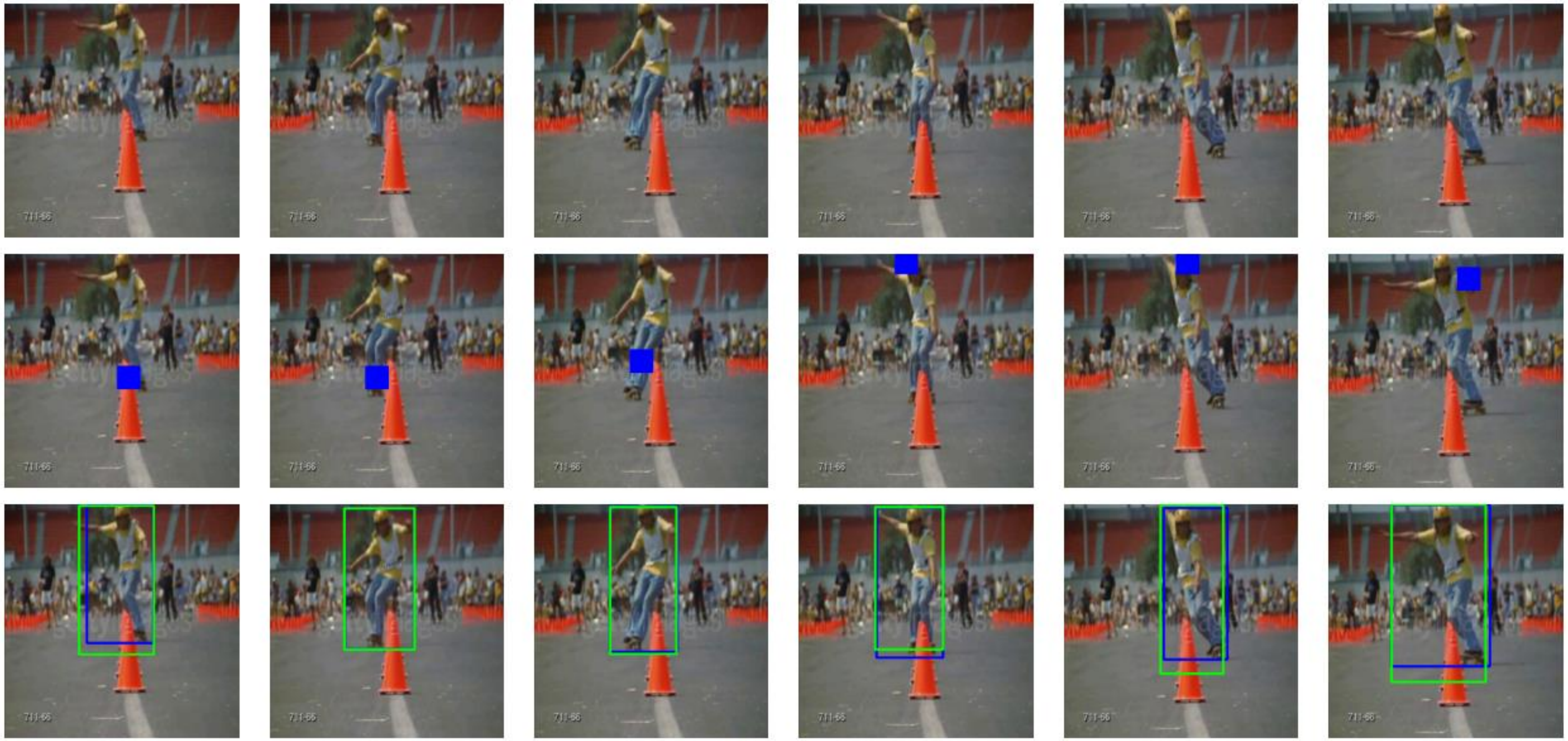
- Pretrained Spatial Region Proposal.
- Prediction error peaks filter the object proposals.
- Energy-based optimization ensures action localization and temporal consistency.

$$E(\mathcal{B}_{it}) = w_\alpha \, \phi(\alpha_{ij}, \mathcal{B}_{it}) + w_t \delta(\mathcal{B}_{it}, \{\mathcal{B}_{j,t-1}\})$$



Aakur, S. N., & Sarkar, S. (2020). Action Localization through Continual Predictive Learning. *arXiv preprint arXiv:2003.12185*.

Energy-based Action Localization

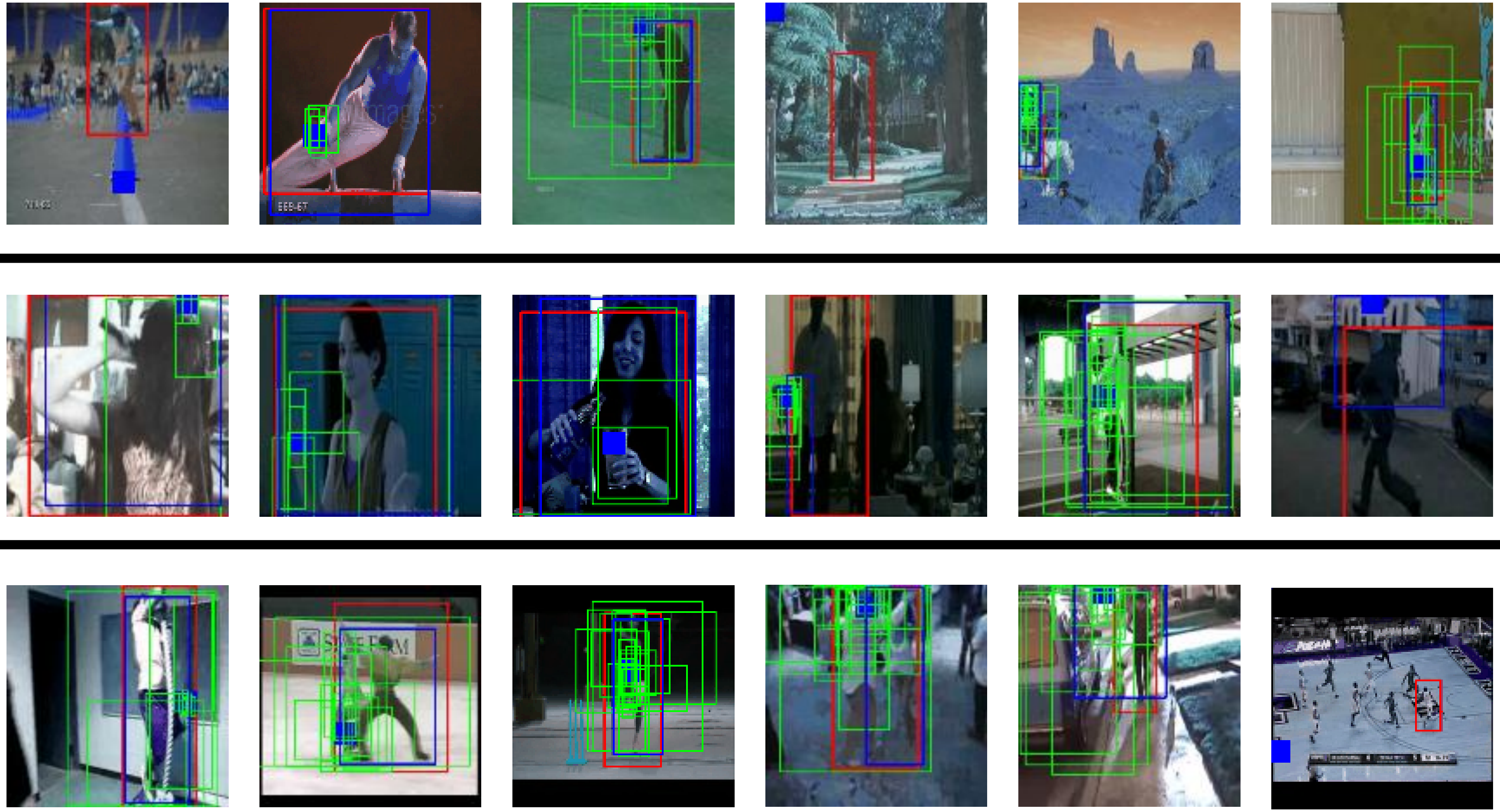


Energy-based Action Localization

UCF Sports

JHMDB

THUMOS’13



Thank you!
Questions?

INSTITUTE FOR ARTIFICIAL INTELLIGENCE

UNIVERSITY OF SOUTH FLORIDA