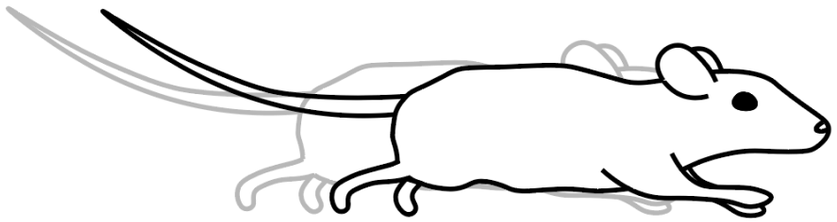
Learning Behavior Representations Through Multi-Timescale Bootstrapping

Mehdi Azabou
Georgia Institute of Technology

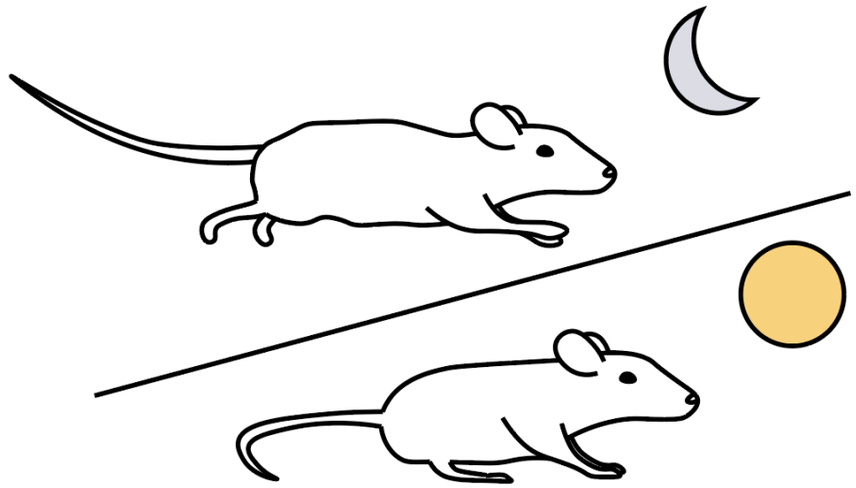
CVPR - MABe Workshop: Jun 20, 2022

Motivation

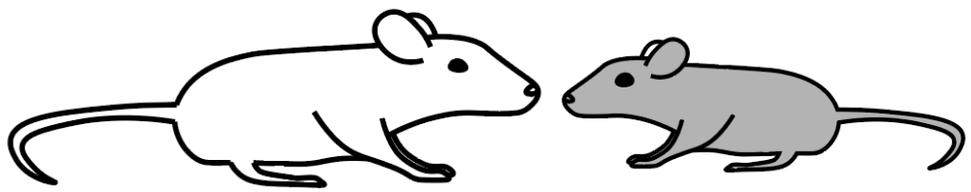
Behavior unfolds over *multiple timescales*



seconds
running, grooming or chasing



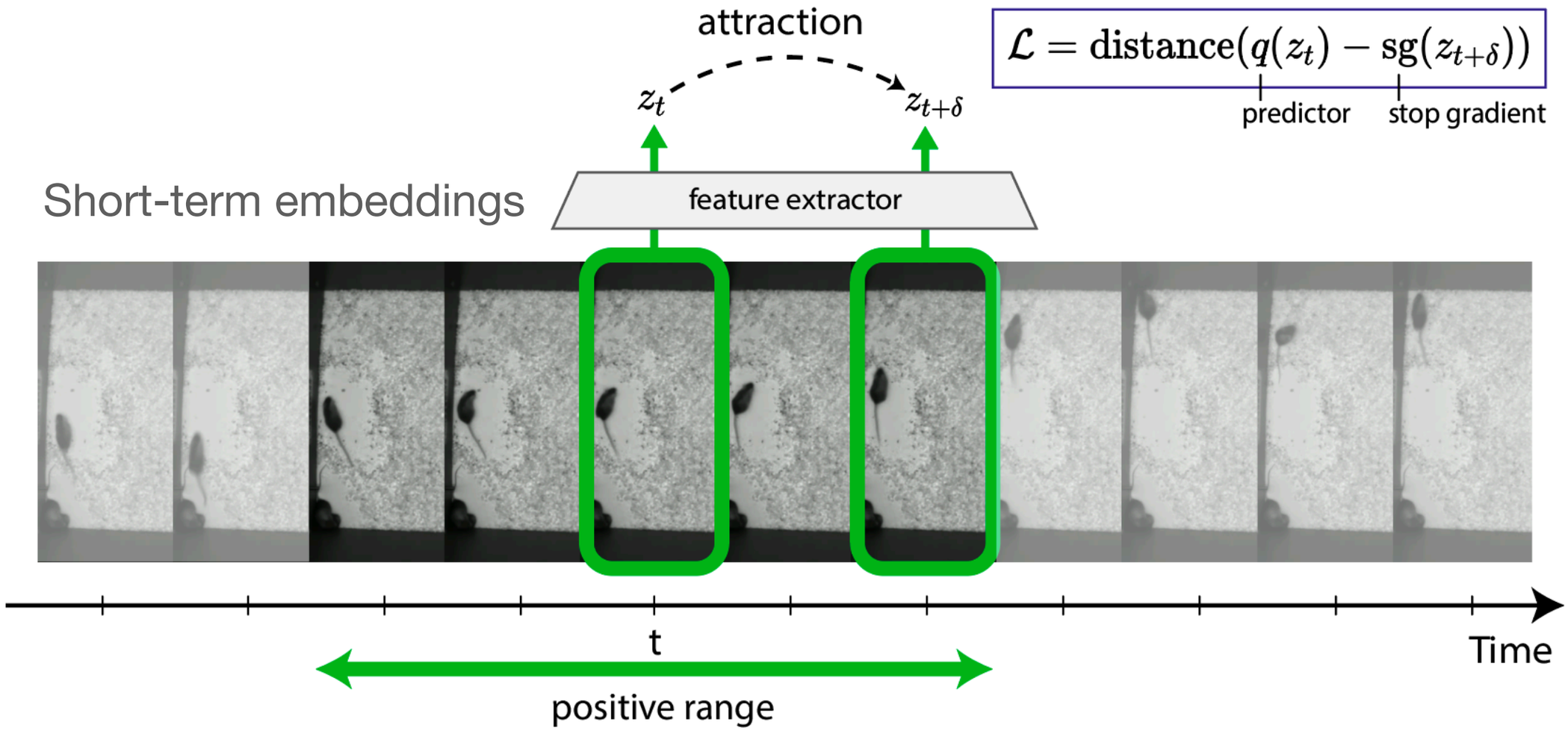
hours
mice are night active
changes based on time of day



days - ∞
age, disease progression
strain, morphology

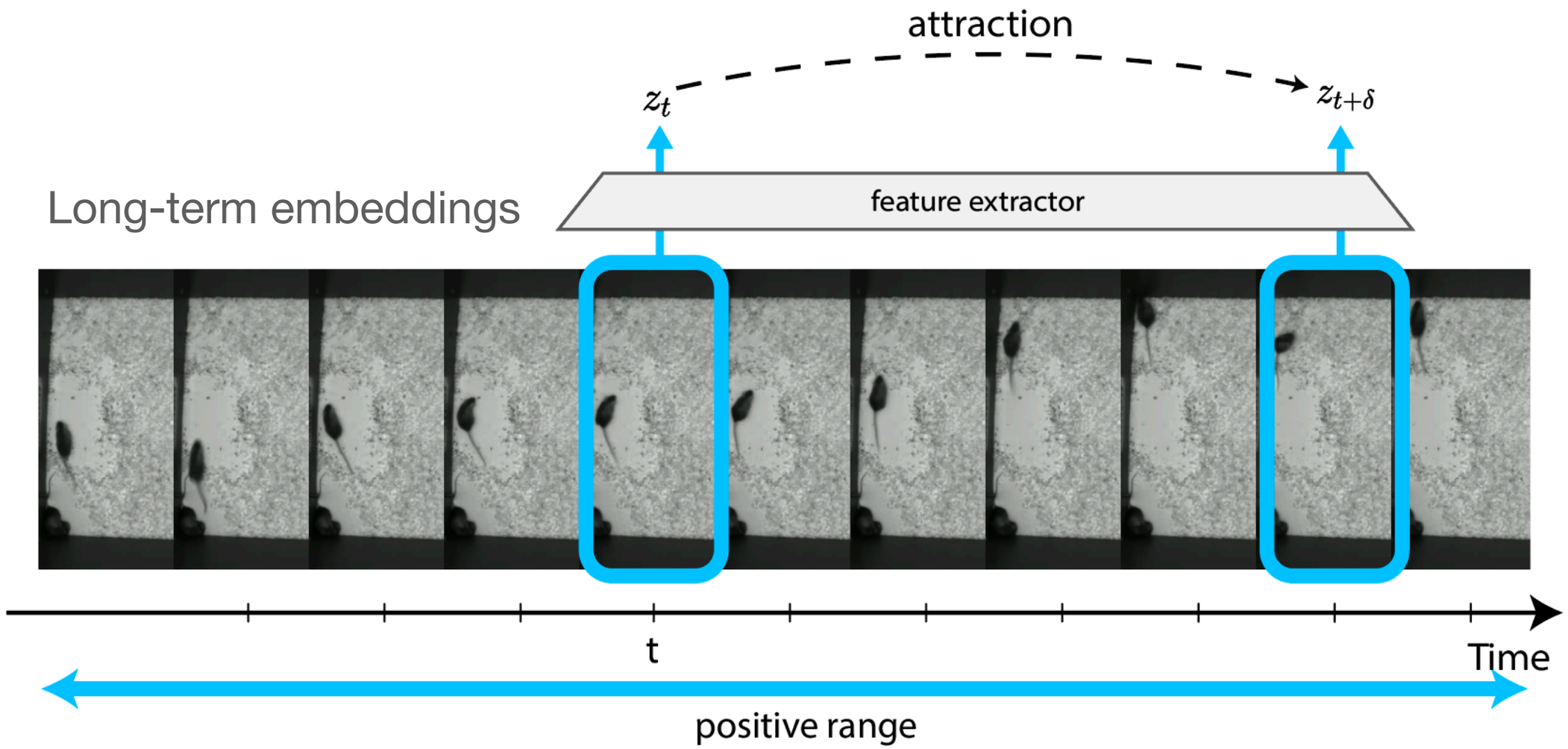
Bootstrapping Across Multiple Scales

Representation Learning: Pull the embeddings of neighboring timepoints closer to each other.

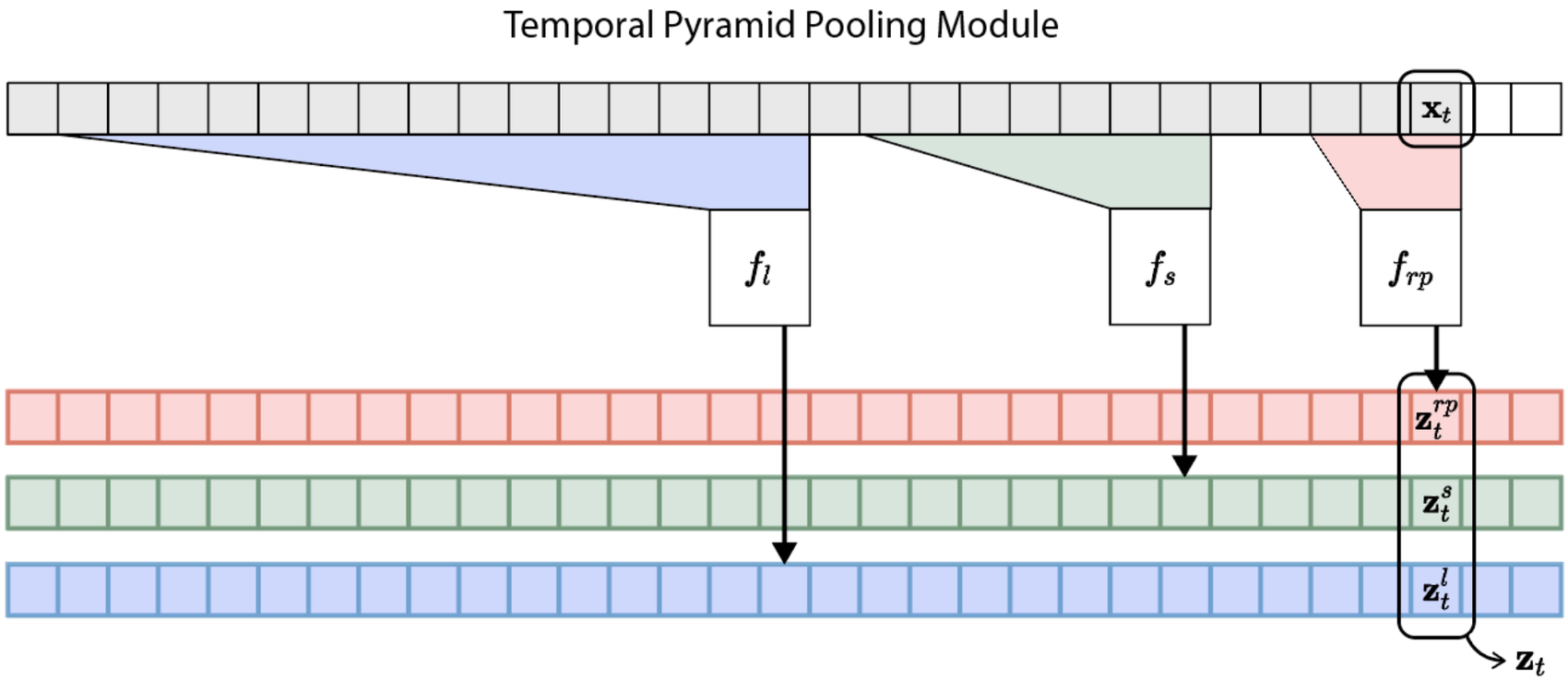


Bootstrapping Across Multiple Scales

Representation Learning: Pull the embeddings of neighboring timepoints closer to each other.



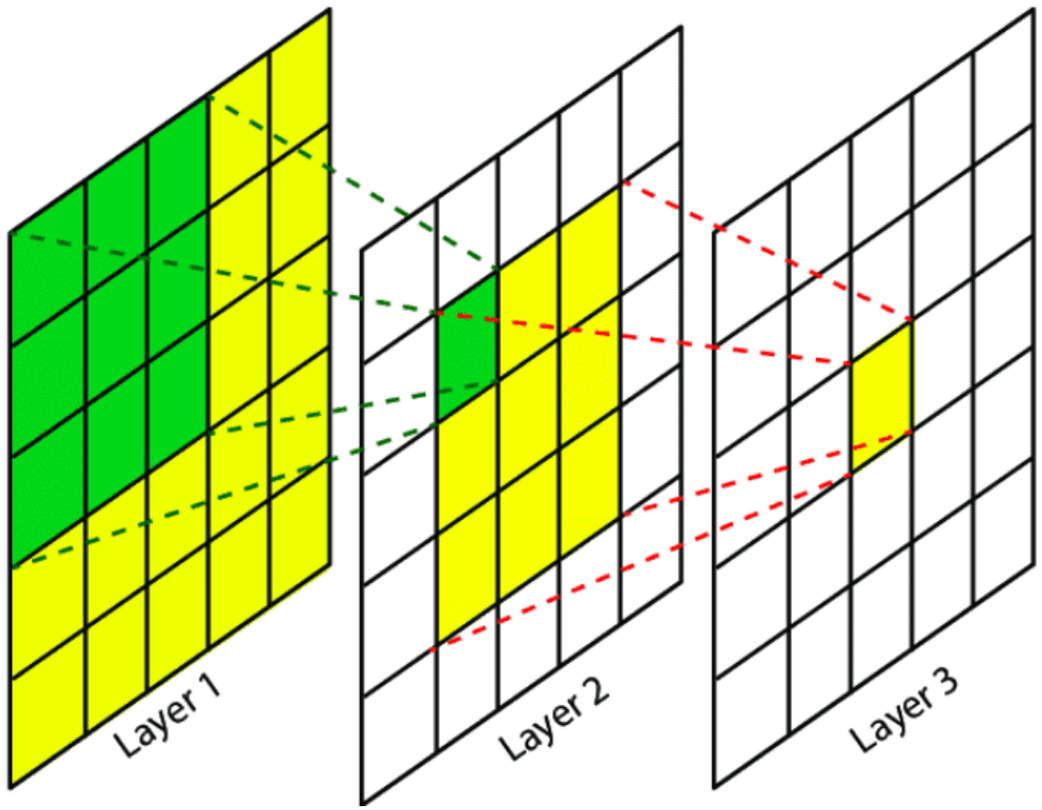
Architecture



Temporal Pyramid Pooling

Multiple TCNs with different receptive fields:
- Recent past encoder (sub-sec)
- Short-term encoder (1sec-10sec)
- Long-term encoder (minutes - hours)

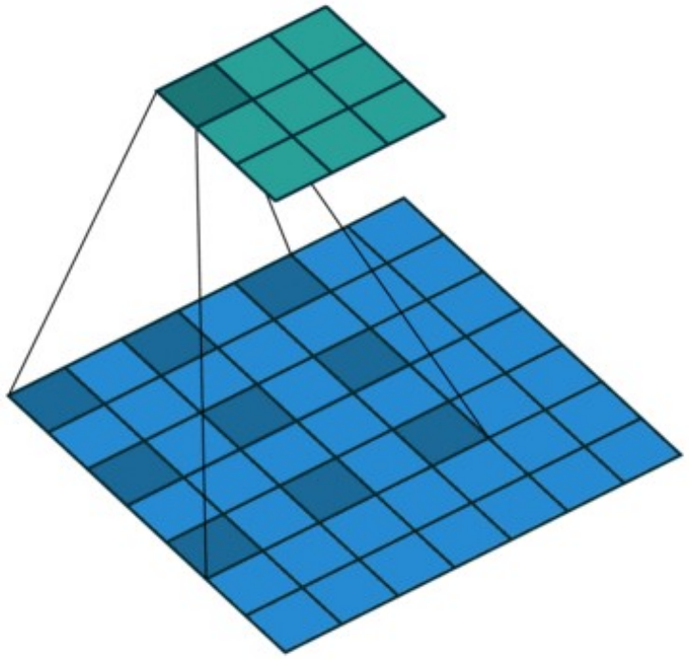
Receptive field

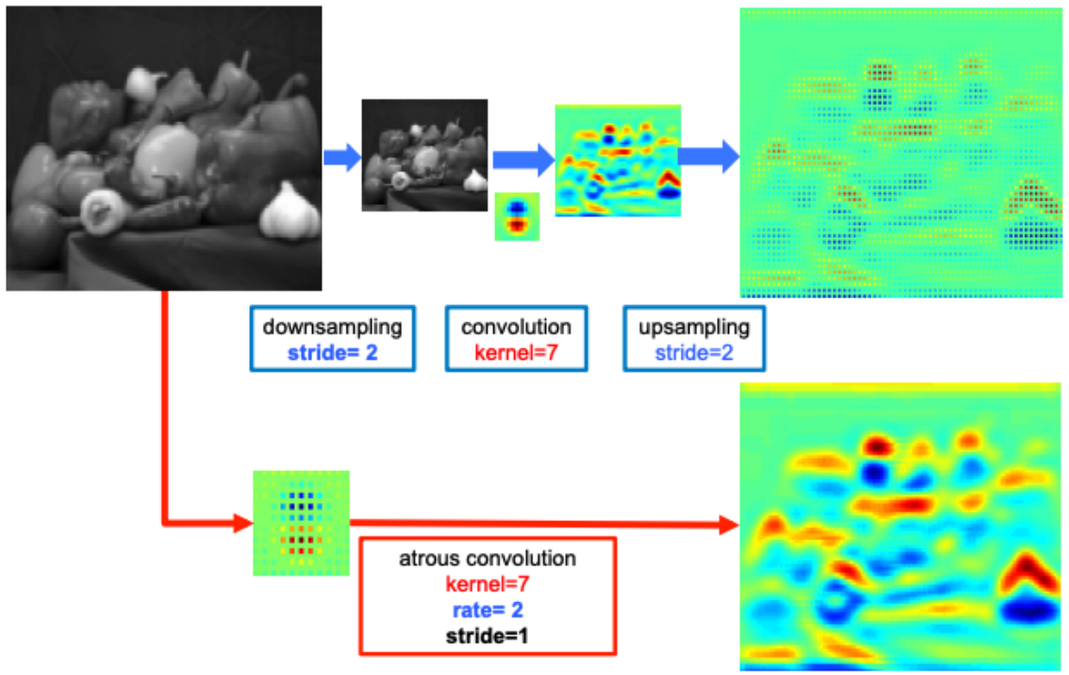


How to expand receptive field:

- Add more layers, go deeper (more parameters!)
- Downsample
- Use dilated convolutions!

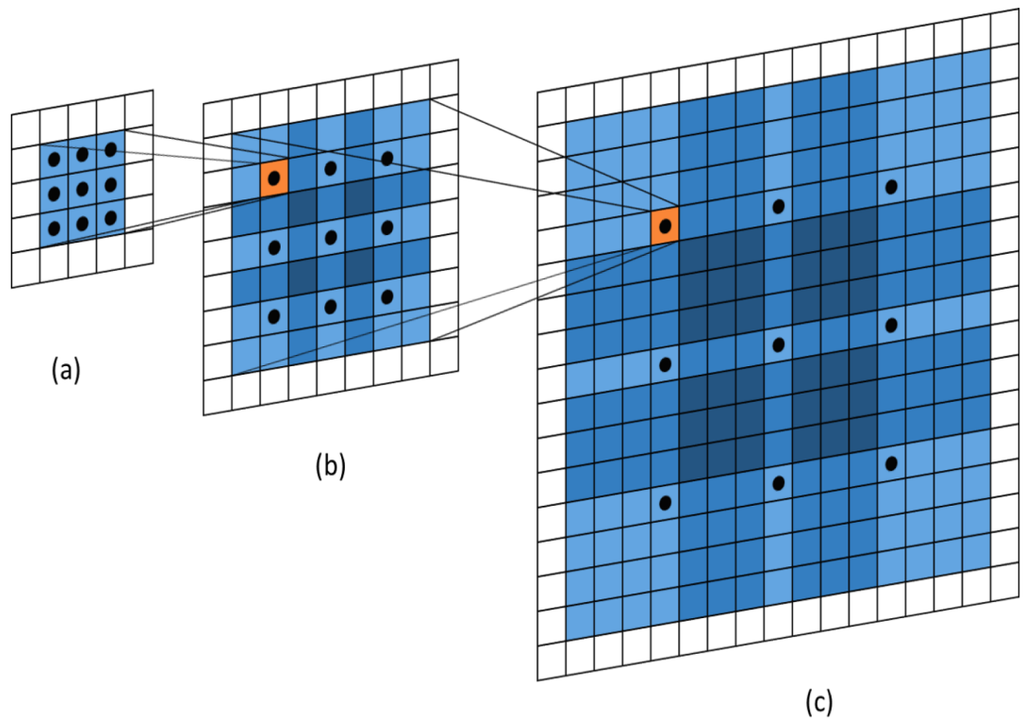
Dilated/ A trous convolutions





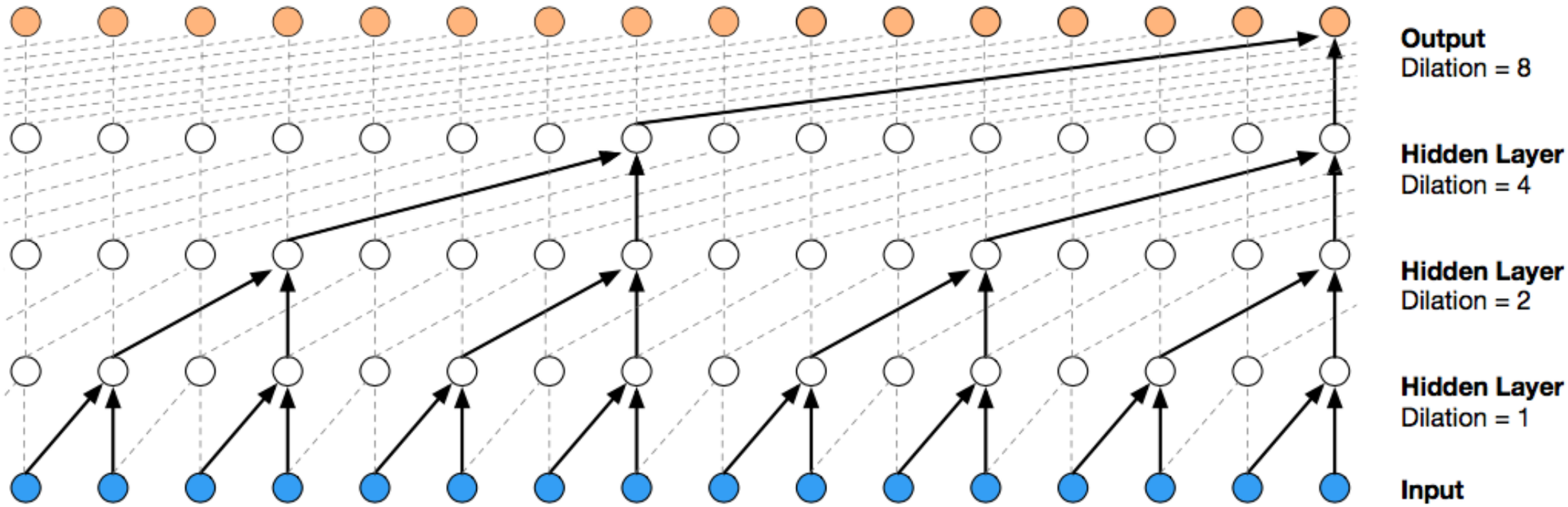
- Dilated convolutions are critical in applications that require keeping the spatial dimensions of the image.

Dilated/ A trous convolutions



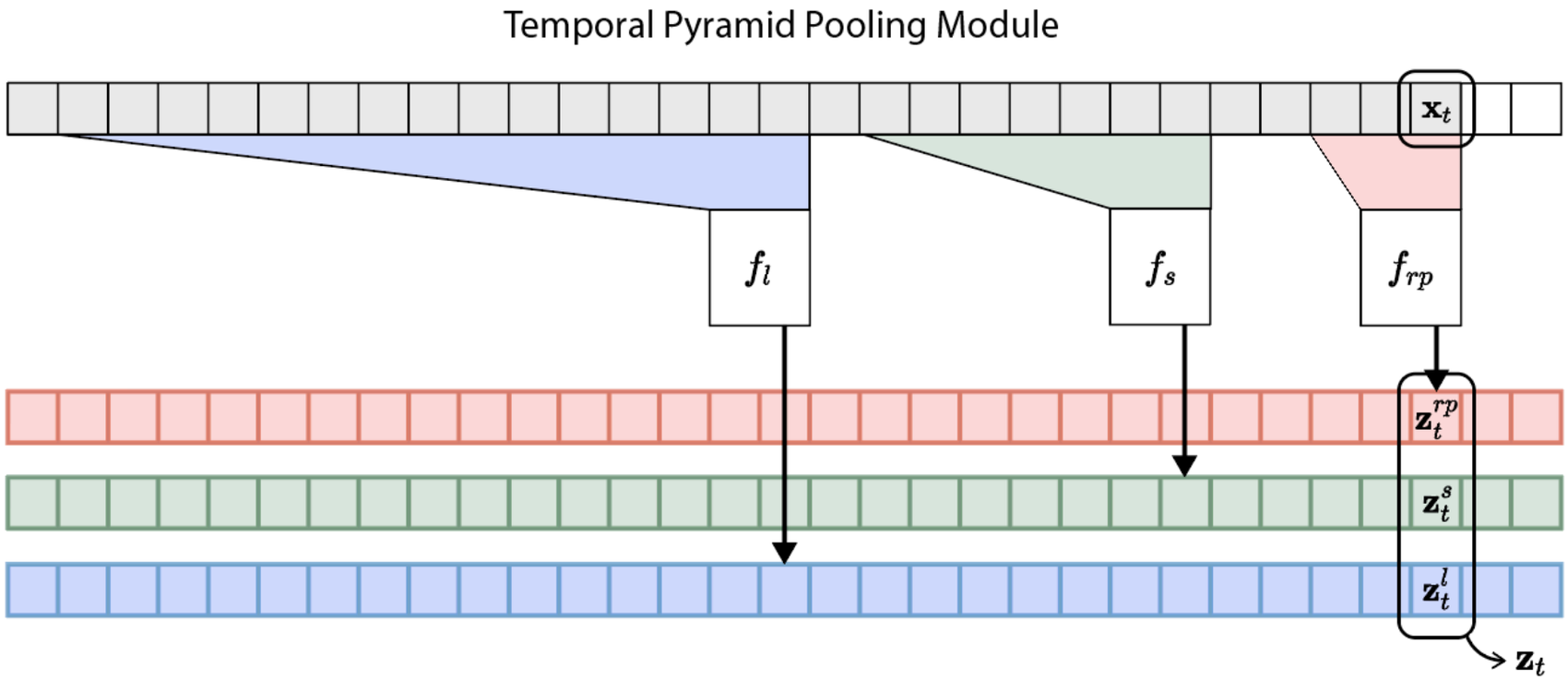
- Using dilated convolutions —> Quickly expand the receptive field

Causal convolutions



- The output at time t only depends on inputs at time t and before

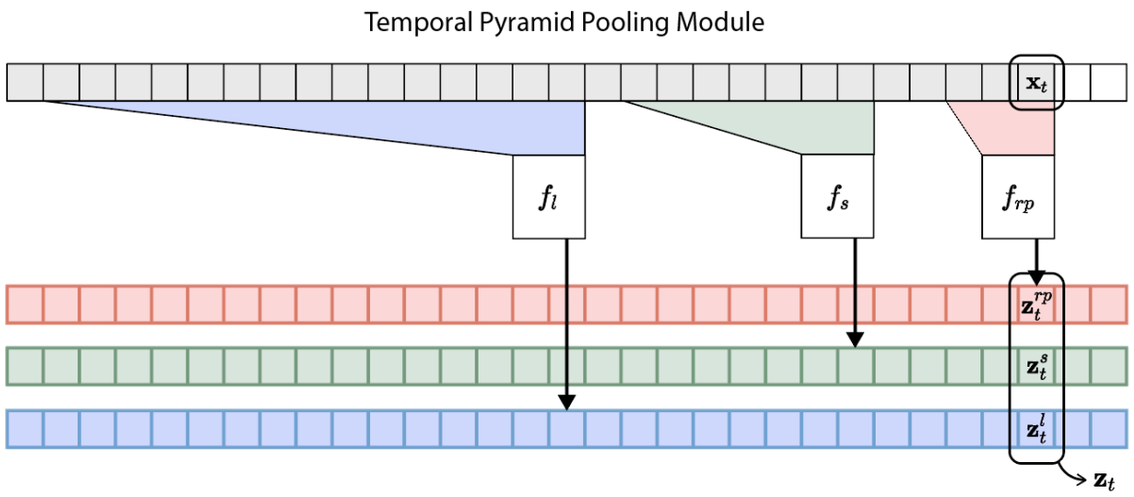
Architecture

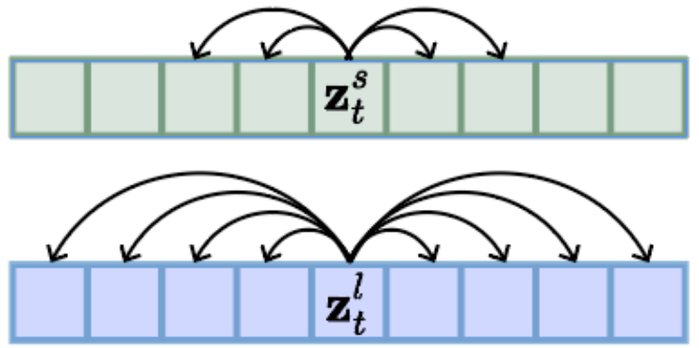


Temporal Pyramid Pooling

Multiple TCNs with different receptive fields:
- Recent past encoder (sub-sec)
- Short-term encoder (1sec-10sec)
- Long-term encoder (minutes - hours)

Learning Objective 1





Encourage similarity within each timescale!

Pull short-term embeddings from neighboring timepoints closer to each other.
Pull long-term embeddings from the same sequence closer to each other.

BYOL: Doubling the encoders

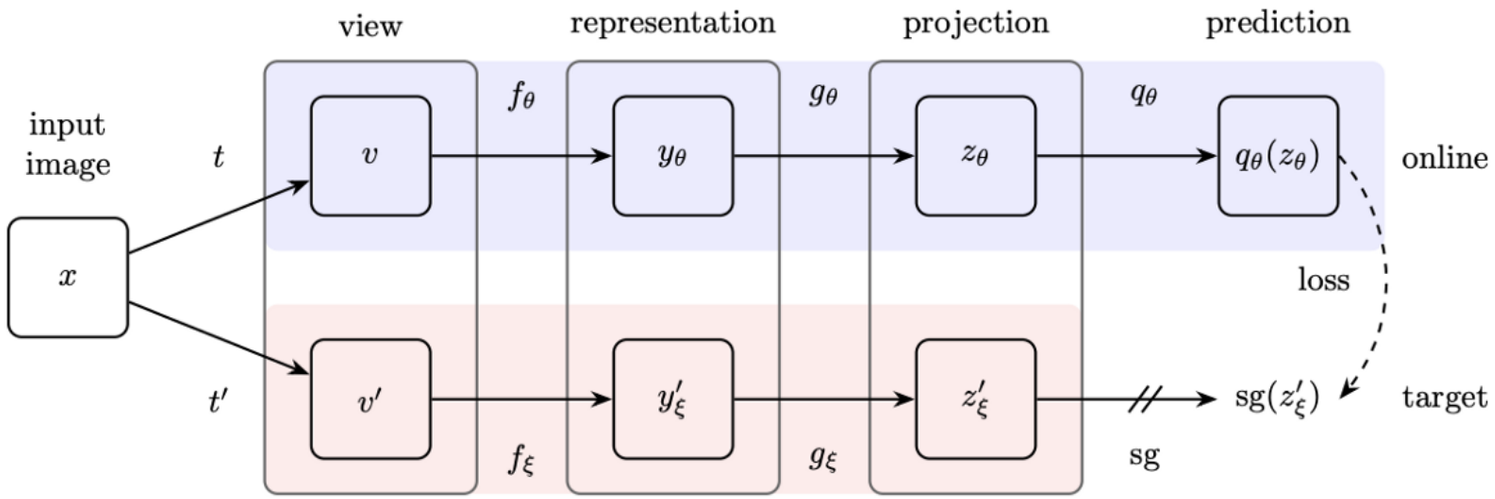


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

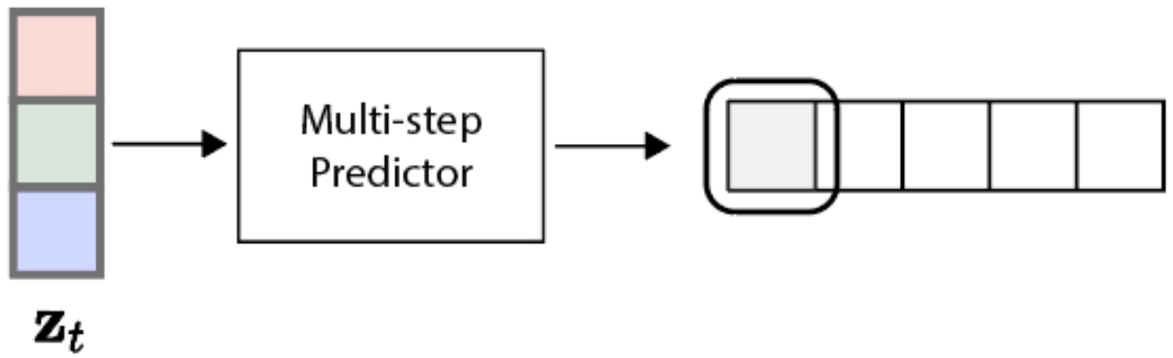
- Important components: EMA (online/target) and stop-gradient

- Alternative: Near-optimal predictor,
  Remove EMA,
  increase the learning rate of the predictor

λ	Top-1
0	0.01
1	5.5
2	$62.8_{\pm 1.5}$
10	66.6
20	$66.3_{\pm 0.3}$
Baseline	72.5

Learning Objective 2

Pretext task: Predict future actions in the next (1s) window.



By training our model to solve these pretext tasks, good representations must be learned

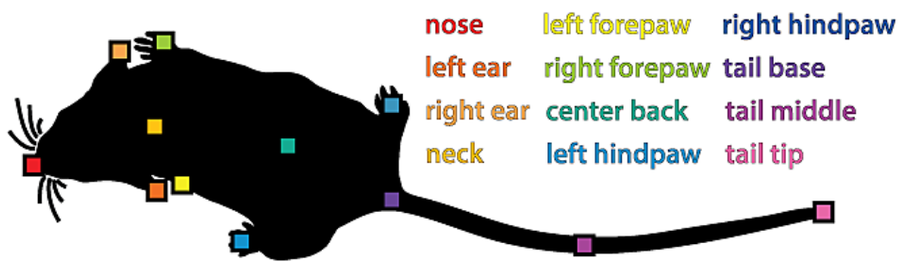
Behavioral representation learning from mouse triplets

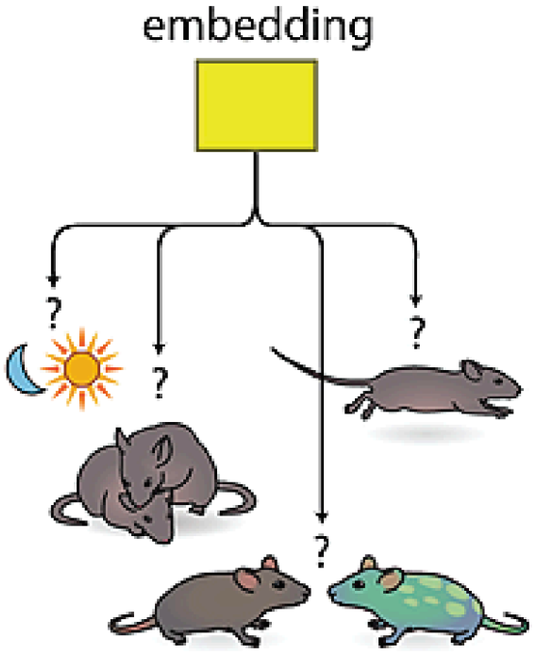
Dataset: Mouse Triplets (MABe 2022)

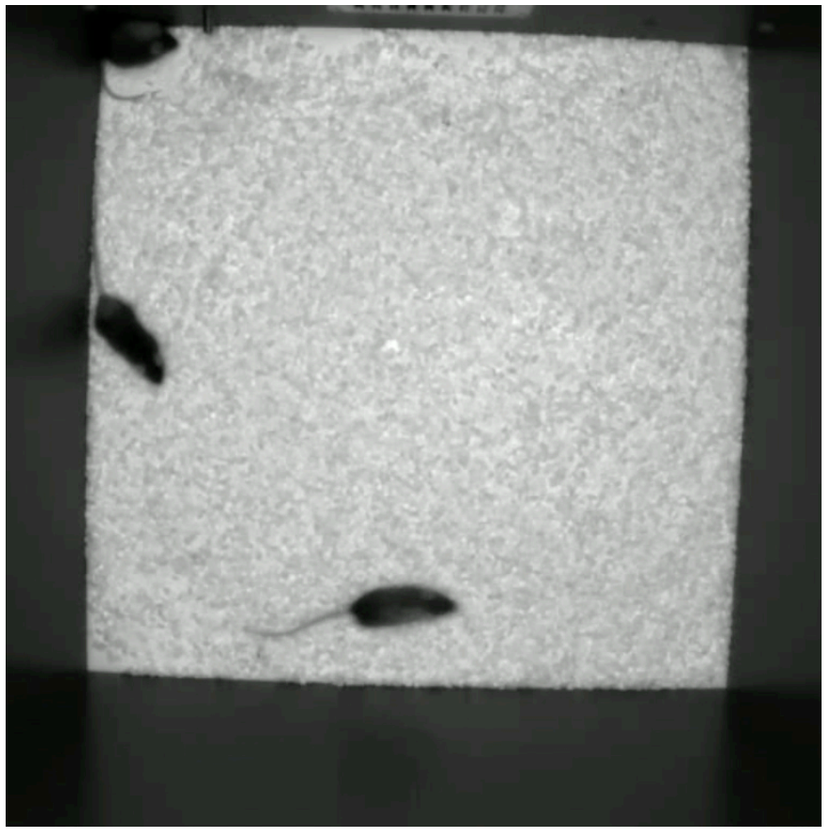
Task: Unsupervised learning from tracking data

Evaluation: Linear readout of 13 sets of labels

- time of day
- chasing behavior
- mouse strain



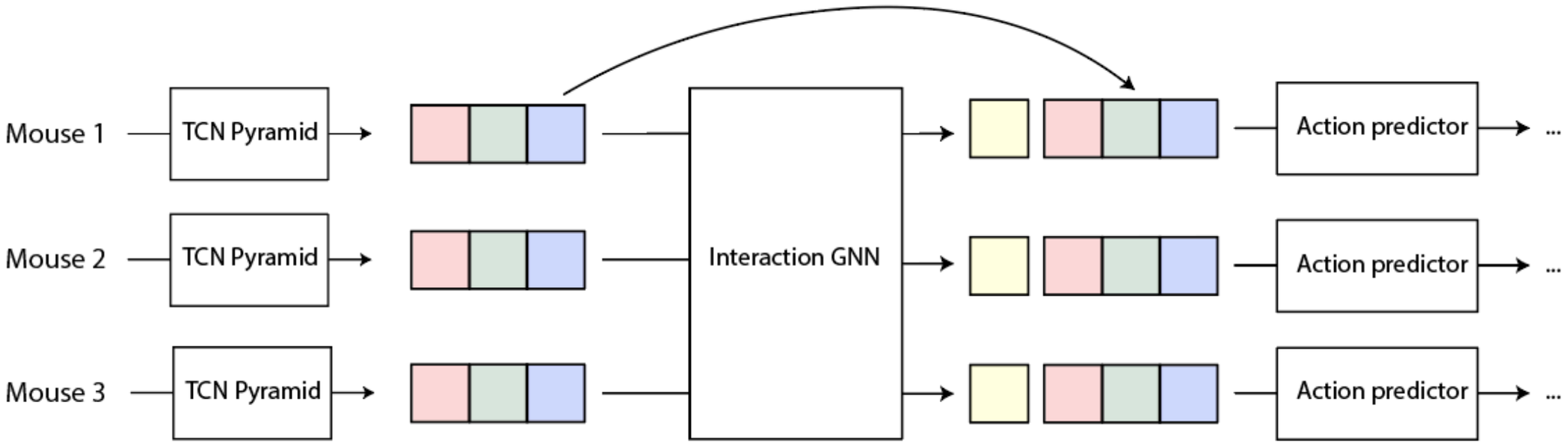




Experimental setup

We extract 36 features from the keypoint data (head orientation, body velocity, joint angles…), and select 6 to be targets for future action prediction.

In the a first stage, we process each mouse independently, then learn an *interaction embedding* to capture the mouse-mouse interactions.



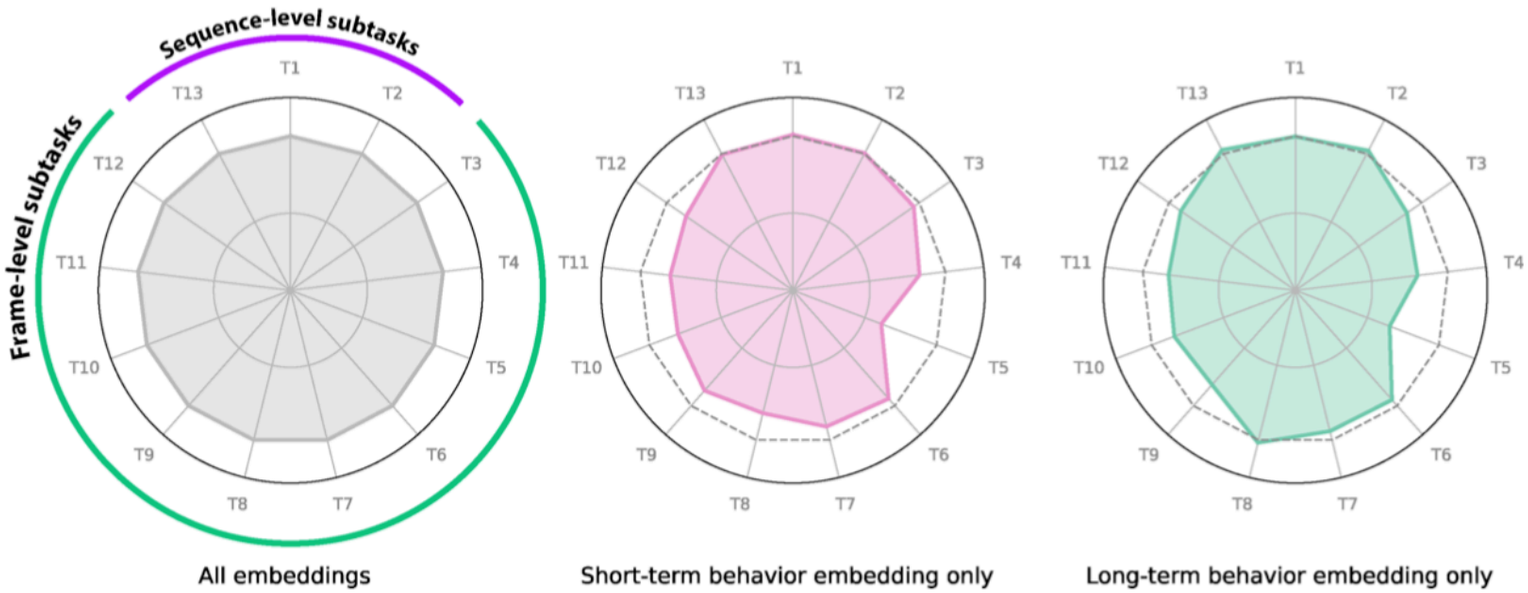
Results

Measuring representational quality

Linear readouts across 13 subtasks

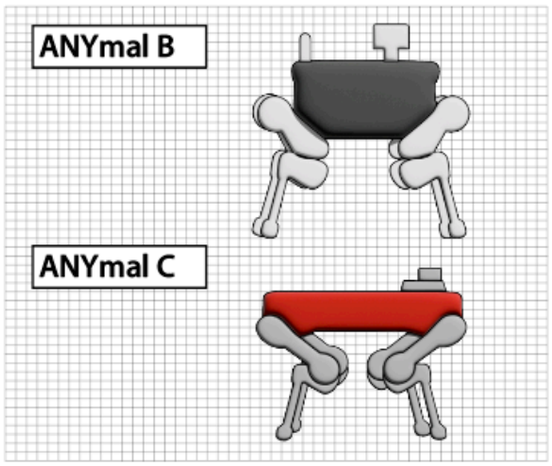
Model	F1-score	MSE	Sequence-level subtasks				Frame-level subtasks								
			T1*	T2*	T3	T13	T4	T5	T6	T7	T8	T9	T10	T11	T12
# 1	**30.3**	**0.09296**	0.09019	**0.09523**	**82.20**	69.40	1.90	1.24	**71.62**	55.52	**30.20**	0.40	1.63	1.10	**20.45**
# 2	28.3	0.09289	0.09057	**0.09513**	67.20	66.90	**2.70**	**6.60**	71.47	**54.67**	20.30	**0.68**	**3.31**	**2.37**	18.54
# 3 BAMS (Ours)	28.4	0.09298	**0.09037**	0.09513	67.10	**69.50**	2.16	2.31	66.42	53.28	30.18	0.45	1.65	1.14	19.14
PCA baseline	7.99	0.09430	0.09415	0.09449	33.83	4.13	0.00	0.00	12.69	0.08	0.00	0.00	0.00	0.00	0.00

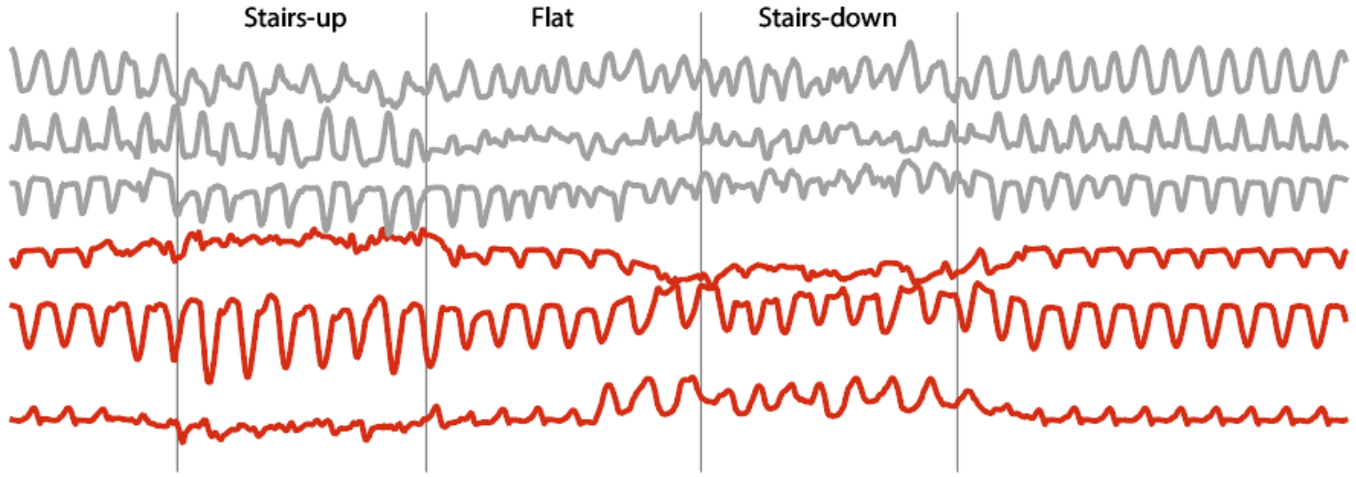
Measuring representational quality in each embedding space

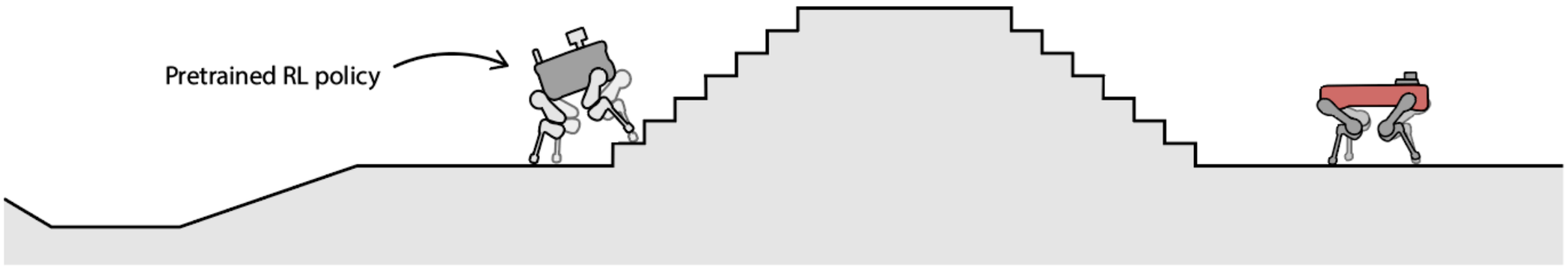


Relative decrease in accuracy from baseline (in %)

Simulated Legged Robots Experiment







We collect data from quadruped robots, with different *morphologies*, walking on procedurally generated *terrains*.

	F1-score	
Model	Terrain type	Robot type
Short-term + Long-term	**0.73**	0.98
Short-term only	0.50	0.86
Long-term only	0.62	**0.99**

Conclusion

- By separating multi-timescale features across different spaces, and designing self-supervised tasks that form these representation, our model can capture the behavioral embeddings that unfold at different rates.

- To understand and analyze behavior, it is critical to capture the factors that modulate it at different timescales.

The Team

Michael Mendelson
Georgia Tech

Maks Sorokin
Georgia Tech

Shantanu Thakoor
DeepMind

Nauman Ahad
Georgia Tech

Carolina Urzay
Georgia Tech

Eva L. Dyer
Georgia Tech

Thank you!