**Objective**

Goal: Learn unsupervised predictors of physical states directly from raw observations and without relying on a simulator in two steps:

(i) **Unsupervised learning** of dynamically-salient objects from videos.

(ii) Train a predictor using the tracker’s detection as supervisory signal.

We validate our method on synthetic data and real data of scenarios of balls rolling on various surfaces.

**Single Object Detection**

Key ideas:

1. **Causality** ($L_{\text{disc}}$): Inspired by [1]. The discriminator $D$ ensures that extracted positions are plausible trajectories and identify temporal reshuffling.

2. **Equivariance** ($L_{\text{siamese}}$): Detection should be equivariant w.r.t random rotation $g$, i.e. $\Phi(gx_T) = g\Phi(x_T)$.

3. **Low entropy** ($L_{\text{ent}}$): Makes sure that detection is spatially localized and locks properly onto one single object.

**Unsupervised Detection and Tracking of Dynamic Objects**

- **1118 videos containing balls rolling on complex terrains.**
- **Dataset split into three types of terrain:**
  - **POOLR**: Flat pool table; 151 videos (1 ball)
  - **BOWLR**: Paper mâché Ellipsoidal Bowl; 216 videos (1 ball)
  - **HEIGHTR**: Paper mâché heightfield; 543 videos (1 b.), 208 (2 b.)
- 8 different types of balls used across all scenarios.
- **Annotations** of objects positions are provided for every test set.

**Evaluation of Our Method**

- **Extension to Multiple Objects**
  - Even when multiple objects are present, our tracker is always able to consistently track one object thanks to the entropy constraint.
  - After learning the first objects, we sequentially train a new tracker where we mask previously detected objects on the extracted heatmaps.

**Tracker Error on Different Dataset**

- **Our tracker performs well across synthetic and real datasets and different types of objects and terrains.**
- **Variance of the error is low, tracking never fails.**

**Ablation Study**

- **BOWLR (1b.) Ablation Study**

**Extrapolation with Unsupervised Data**

- **We use our tracker to train an extrapolator such as IFS [2] and {Pos, Disp, Prob}Net[3].**

- **Models are trained to predict the next $T=\{15,20\}$ steps observing $T_0=4$ frames.**

- **Best results are obtained with *Net models which use tensor state representations.**

**References**

