Hindsight for Foresight: Unsupervised Structured Dynamics Models from Physical Interaction

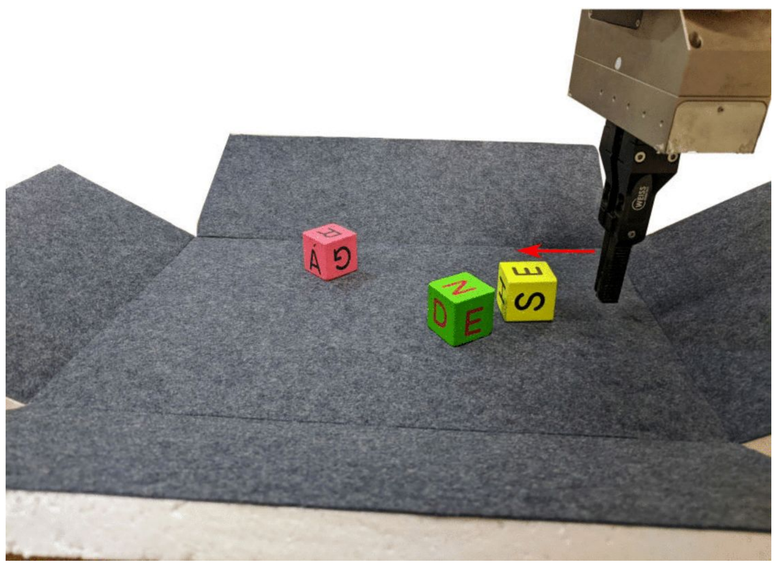
Iman Nematollahi, Oier Mees, Lukas Hermann and Wolfram Burgard

UNI FREIBURG

AIS Autonomous Intelligent Systems

Motivation

- What will happen when the robot arm moves left? Will the yellow block collide with the green block?
- We present a novel approach for modeling the dynamics of a robot's interactions in the real-world directly from unlabeled 3D point clouds and images. This formulation leads to interpretable models for visuomotor control and planning.



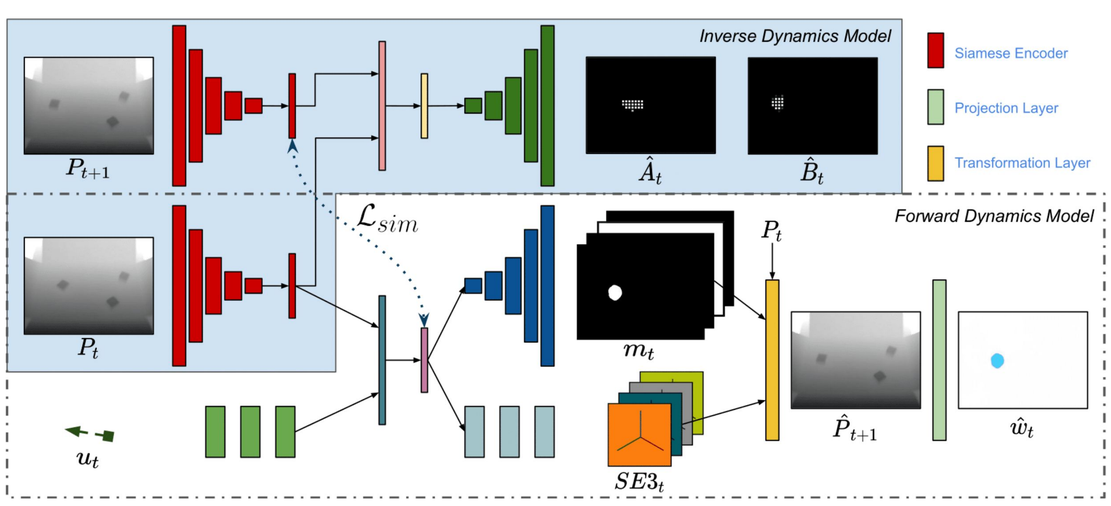
Hind4sight-Net

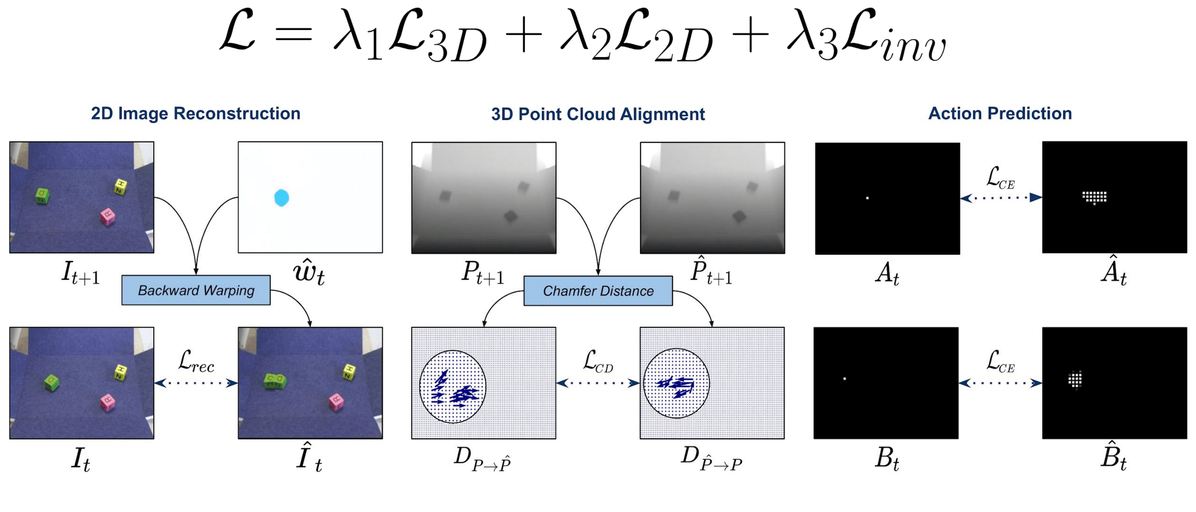
- Our dynamics model consists of both a forward and an inverse model.
- Forward Model takes a raw point cloud $P_t = (X_t, Y_t, Z_t)$ and an action u_t as inputs and decompose the scene into K objects, predict their masks m_t^k and estimate their motion as a 3D rigid body transform $[R, T] \in \mathbf{SE}(3)$ to generate the next point cloud $\hat{P}_{t+1}$.
- Inverse Model takes two consecutive raw point clouds P_t and P_{t+1} as input and predicts the corresponding poke action $\hat{u}_t$ as two heatmaps, for the start and end positions of the poke.

Training on real-world unlabeled interaction data

- The main loss functions operate on observational changes and enable learning scene dynamics in the real world without the need of data associations provided by a tracker.
- The image reconstruction loss uses the predicted 2D flow to minimize a photometric consistency error.
- The Chamfer Distance tries to enforce the geometric consistency between point clouds.
- The inverse model predicts spatial distributions of the actions that caused the scene to change.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{3D} + \lambda_2 \mathcal{L}_{2D} + \lambda_3 \mathcal{L}_{inv}$$





Results on Modeling Scene Dynamics

- We evaluate the performance of our unsupervised structured dynamics model on both simulated and real world datasets.
- Hind4sight-Net achieves the best 3D scene flow error compared to baselines even though it is fully-unsupervised and not directly trained to predict 3D scene flow:

Model	Training Paradigm	MSE (cm)
SE3-Nets	supervised	2.20
Hind4sight-Net	unsupervised	**1.63**
No Motion	x	12.6

- Hind4sight-Net's implicit action-conditioned 2D optical flow, outperforms FlowNet 2.0, despite FlowNet 2.0 having access to two consecutive images as input and having explicit optical flow supervision.

Model	Inputs	AEE
FlowNet 2.0	Images I_t and I_{t+1}	0.11
Hind4sight-Net	Point Cloud P_t and Action u_t	**0.05**

Control Performance

- We use the cross entropy method (CEM) to find poke action sequences that lead to a desired goal.
- We define the planning cost-function by a combination of the 3D and 2D domains the network has been trained on.
- We observe that in most cases we can reach the goal configuration with around 10 poke actions.