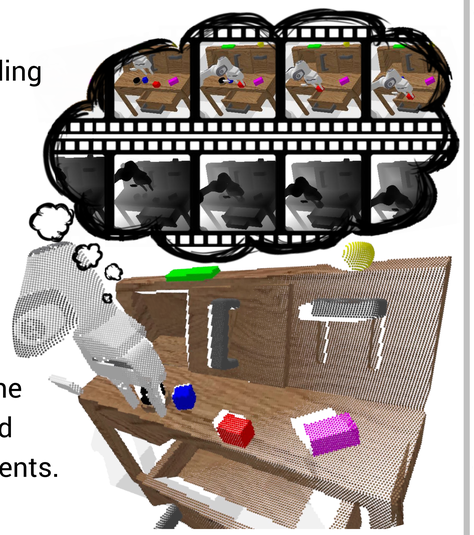
T3VIP: Transformation-based 3D Video Prediction

Iman Nematollahi, Erick Rosete-Beas, Seyed Mahdi B. Azad, Raghu Rajan, Frank Hutter, Wolfram Burgard

ICRA 2022 IEEE International Conference on Robotics and Automation

AIS Autonomous Intelligent Systems

Introduction

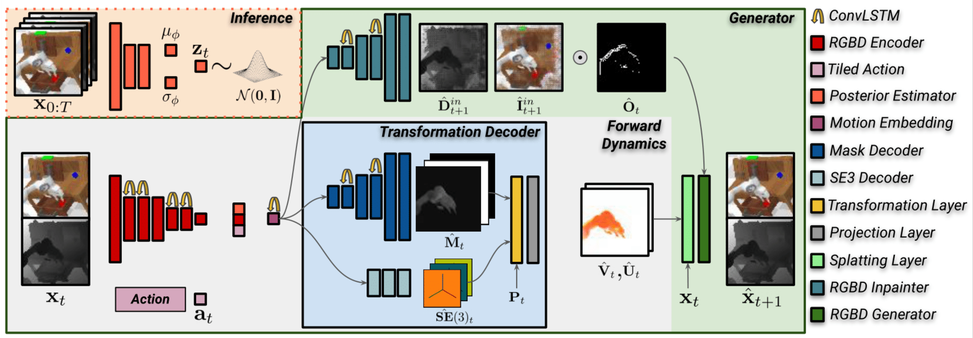


- Humans have an intuitive understanding of physics, capable of predicting the dynamics of the 3D world.
- This cognitive model enables us to generalize from past experience and predict the future observations.
- **Goal**: enable robots to comprehend the dynamics of the surrounding 3D world and predict its likely future developments.

T3VIP

T3VIP learns a 3D world model from past unlabeled experience to imagine plausible future RGB-D videos and plan the best action trajectory. T3VIP:

- Is a 3D aware world model and predicts multiple future RGB-D frames
- Decomposes a scene into salient objects
- Predicts 3D rigid body transformations of object segments
- Is fully unsupervised and learns the physical dynamics by reasoning about the visual and geometric cues
- Is equipped with AutoML techniques to find the best strategy for exploiting available observational signals
- Computes 3D scene flow, the 2D optical flow and the occlusion mask as emergent properties, leading to better interpretability
- Captures the stochastic nature of the real world



Unsupervised Formulation

We propose an unsupervised learning framework for predicting the dynamics of a scene solely based on unlabeled 3D point clouds and 2D images. T3VIP aims to:

- Reconstruct RGB images $\mathcal{L}_{rec}^{I}$
- Reconstruct Depth maps $\mathcal{L}_{rec}^{D}$
- Enforce the consistency of predicted and observed point clouds $\mathcal{L}_{knn}$
- Encourage scene flow smoothness $\mathcal{L}_{fs}^{s}$
- Encourage optical flow smoothness $\mathcal{L}_{fs}^{o}$
- Fit a prior distribution to account for stochasticity $\mathcal{L}_{kl}$

The full objective of T3VIP is:

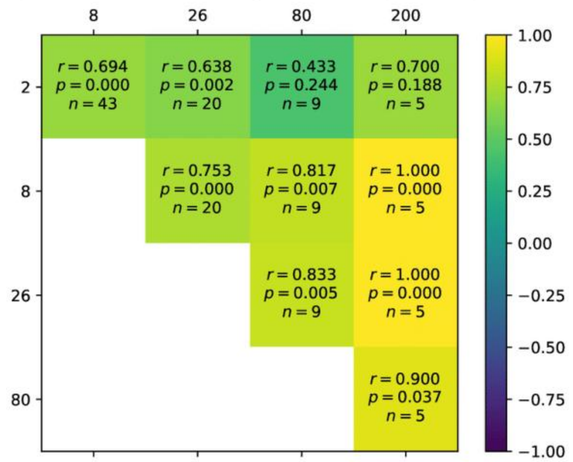
$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec}^{I} + \lambda_2 \mathcal{L}_{rec}^{D} + \lambda_3 \mathcal{L}_{knn} + \lambda_4 \mathcal{L}_{fs}^{s} + \lambda_5 \mathcal{L}_{fs}^{o} + \lambda_6 \mathcal{L}_{kl},$$

where lambdas are hyperparameters representing the relevance of each loss term.

Automated Hyperparameter Optimization

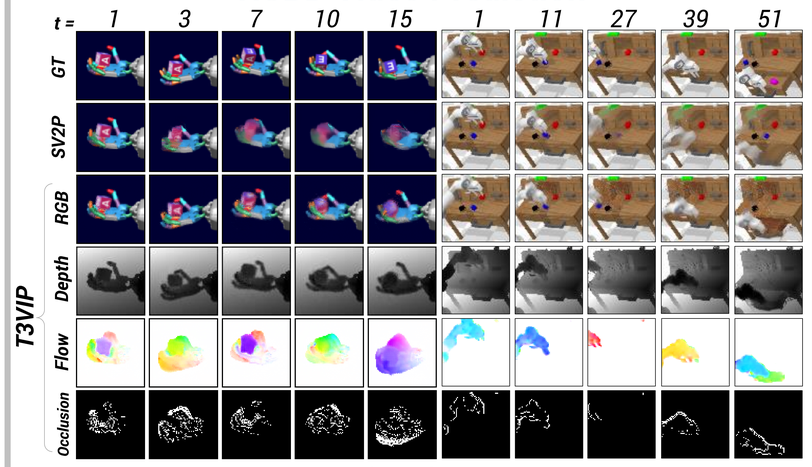
We tune T3VIP for automatically setting learning rate, alpha (L1 or L2 reconstruction), and lambdas hyperparameters:

- ASHA finds configurations leading to high-quality RGB-D predictions
- Optimization metric: Sum of PSNR scores for both the predicted RGB images and depth maps



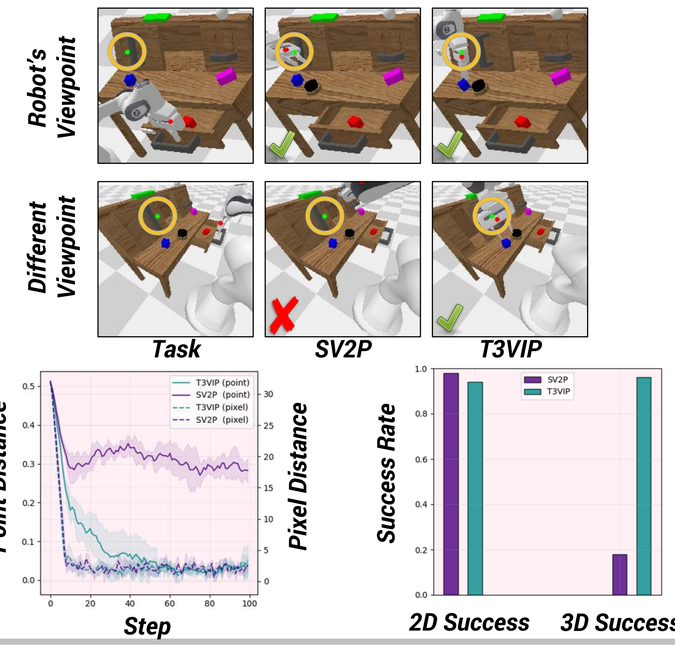
Rank Correlation of the total PSNR across the budgets

RGBD Video Prediction



- T3VIP produces sharp flow fields and sparse occlusion masks leading to sharper RGB images than baselines

Model-Predictive Control



Conclusions

- Learns an intuitive 3D world model
- Predicts multi-step RGB-D video effectively
- Models the 3D dynamics of a scene
- Leverages visual and geometric cues
- Employs AutoML to find best hyperparameters
- Enables an agent to reach 3D targets

SCAN ME