# Unsupervised Learning of Object Structure and Dynamics from Videos

Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin Murphy, Honglak Lee

Google Research

## 1 Introduction

Many vision tasks require an understanding of **object motion**.

Learning representations of object structure and dynamics from pixels, **without supervision**, is a major challenge.

We propose an object-centric model of video that lear **keypoint-based representations**.

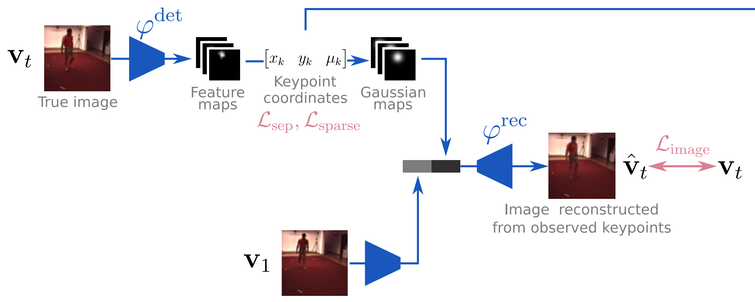

By learning dynamics in the keypoint space, we avoid accumulation of pixel errors and can make **high-quality and diverse long-term predictions**.

Our model improves both on **video prediction** and on **downstream tasks** that require an understanding of object motion.

## 2 Structured image representation

We use an autoencoder that learns to represent images as a set of **keypoints** using only an image reconstruction loss (Jakab et al., 2018).

To encourage **keypoint-object correspondence**, we add losses to make keypoints **sparse** and their trajectories **decorrelated in time.**

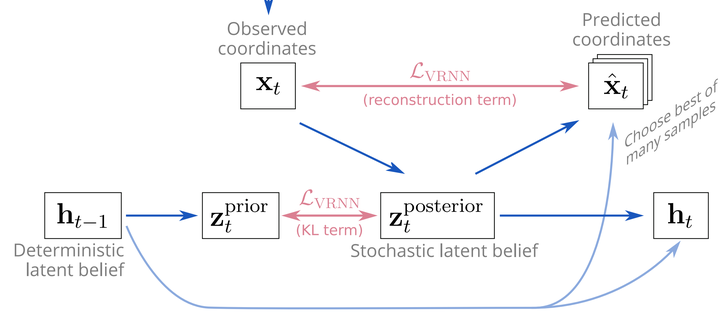



## 3 Latent dynamics model

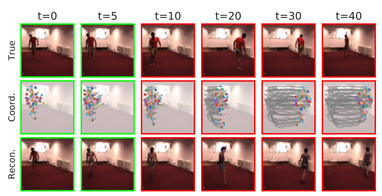
We learn dynamics in the **keypoint space**. We thus never need to condition on predicted images.

The dynamics model (VRNN) has a **deterministic** and a **stochastic** pathway to model long-term stochastic trajectories.

We use a "**best of many samples**" objective to further encourage diverse predictions.

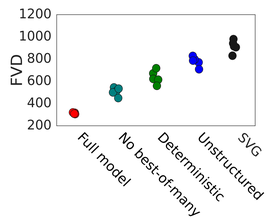



## 4 Video prediction



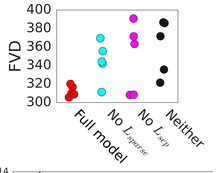

Images reconstructed from the keypoint representation (Struct-VRNN) stay sharper compared to an unstructured baseline (CNN-VRNN) and SVG (Denton and Fergus, 2018).


## 5 Metrics and ablations



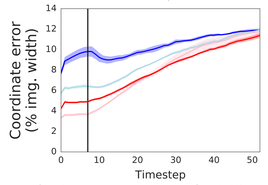

Fréchet Video Distance measures the **difference from ground-truth videos**.

Keypoint structure, stochasticity and best-of-many objective all contribute to prediction quality.




**Keypoint losses** stabilize training and improve performance.

(Each dot is one model initialization.)

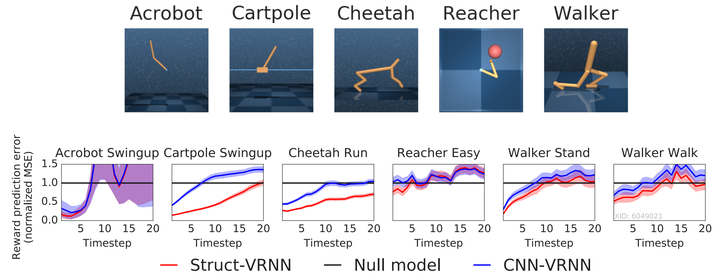



The **coordinate tracking error** of the model gets close to that of a supervised model.

## 6 Using the learned representations

A promising use case for our model are **control tasks** with spatially defined rewards, e.g. in robotics.

As a first step, we show that our model performs better at **reward prediction** than a baseline with an unstructured representation in a suite of simulated control tasks:




Our paper contains more experiments on downstream applications, exploring counterfactual scenarios by manipulating keypoints, and more.

For videos and code, see **mjlm.github.io/video_structure**