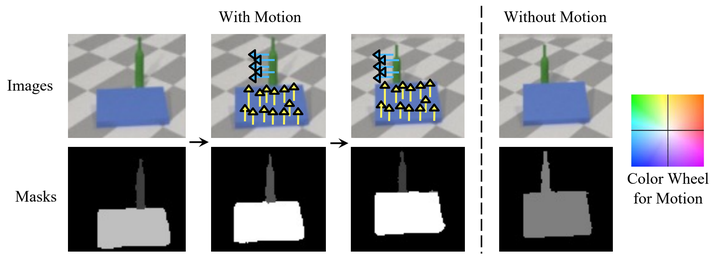
Unsupervised Discovery of 3D Physical Objects From Video

Yilun Du, Kevin Smith, Tomer Ulman, Joshua B. Tenenbaum, Jiajun Wu

Massachusetts Institute of Technology · Harvard · Stanford University

Unsupervised 3D Physical Object Discovery

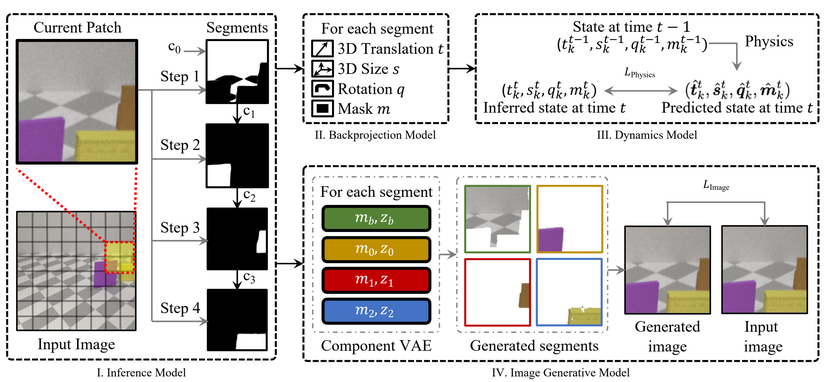
- Discover 3D physical objects from video by following cognitive development
- Motion as a universal cue of objects to segment individual objects from a video
- Local foveation (submap) decomposition of images to faithfully recover each component object in a scene
- Show that resultant discovered 3D physical objects can be utilized for physical reasoning



Motion is an important cue for object segmentation from early in development. We combine motion with an approximate understanding of physics to discover 3D objects that are physically consistent across time. In the video above, motion cues (shown with colored arrows) enable our model to modify our predictions from a single large incorrect segmentation mask to two smaller correct masks

Method

- Utilize motion cues to segment 3D objects from video
- Enforce that masks of segmented 3D physical objects change consistently according to the rules of physics.
- Learning through joint optimization over a back projection model a image generative model, dynamics model



POD-Net contains four modules for discovering physical objects from video. (I) An inference model auto-regressively infers a set of candidate object masks and latents to describe each patch of an image; (II) A backprojection model maps each mask to a 3D primitive; (III) A dynamics model captures the motion of 3D physical objects; and (IV) An image generative model decodes latents and masks to reconstruct the image.

- Infer objects in a foveated manner, where individual segments of objects are slowly pieced together.

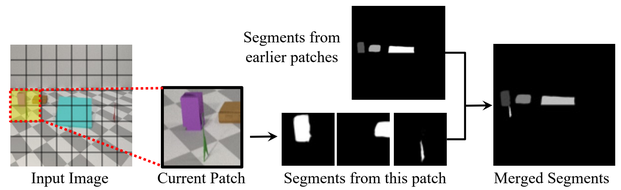
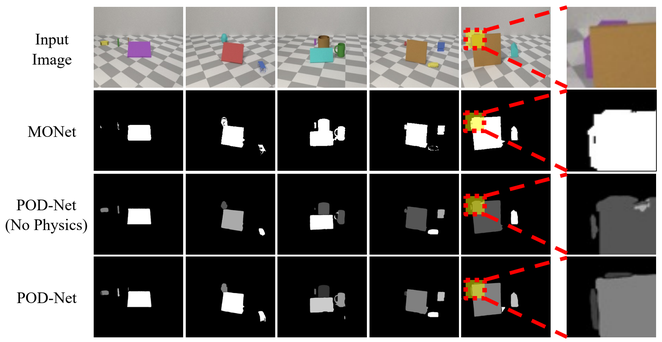


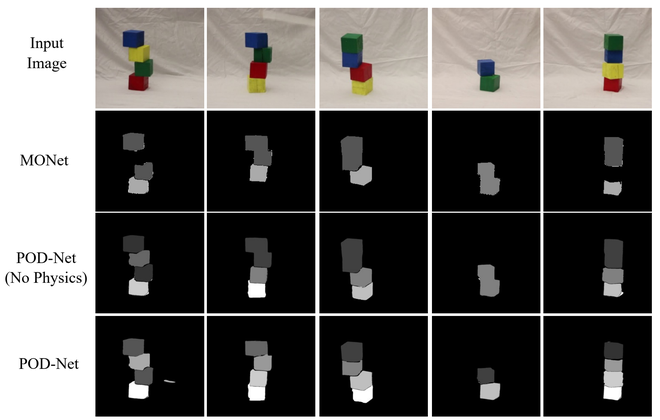
Illustration of sub-patch decomposition for image inference. An image is divided in a 8 x 8 grid, with inference applied to each 2 x 2 subgrid. To generate a global segmentation mask, object masks are sequentially inferred for each subpatch. Each object mask is either matched to an existing object or used to create a new object.

Object Segmentation

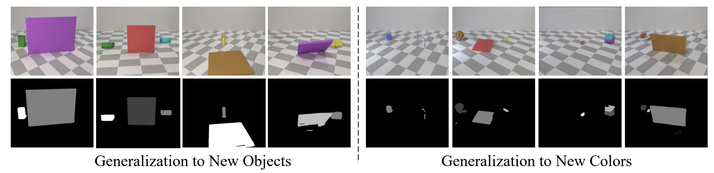
- Foveated inference enables segmentation of both large and small objects



- Motions enables segmentations to separate individual blocks

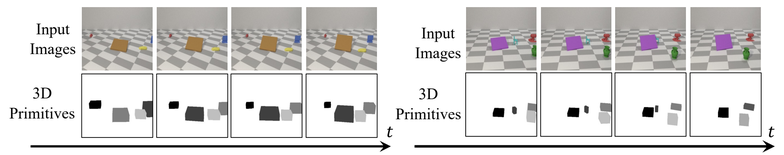


- Segmented objects generalizes to new colors and shapes of objects

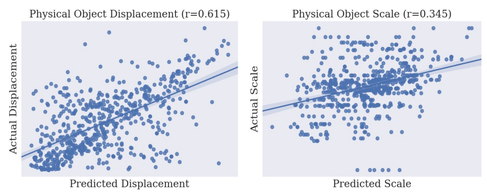


3D Object Discovery

- Backprojection model infers a set of 3D primitives representing the overall scene



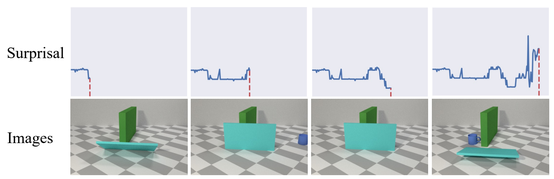
- Inferred 3D primitives have good correlation with ground truth 3D objects



Plot of predicted translation of 3D primitives vs ground truth translation of 3D primitives (left) and plot of predicted scale of 3D primitives vs ground truth scale of 3D primitives.

Physical Reasoning

- Inferred 3D physical objects can be used to reason about the underlying physical of a scene
- Infer physical plausibility of 3D objects teleporting over barriers



Surprise over time in a “Block“ scene. POD-net has relatively low surprisal throughout most of the video. But when the occlude falls and the object appears to “teleport” across the wall, POD-Net recognizes this abnormal shift position and becomes surprised.

Additional Information

- Website at https://yilundu.github.io/podnet/
- Code at https://github.com/yilundu/podnet