A biologically plausible route to learn 3D perception

Wanhee Lee[1]* (wanhee@stanford.edu), Jared Watrous[1]* (jwatrous2002@stanford.edu), Honglin Chen[1] (honglinc@stanford.edu), Klemen Kotar[1] (klemenk@stanford.edu), Tyler Bonnen[2] (tyler.ray.bonnen@gmail.com) Daniel L. K. Yamins[1] (yamins@stanford.edu)

[1]Stanford Neuroscience and Artificial Intelligence Laboratory, Stanford University, CA, USA

[2]Berkeley AI Research (BAIR) Lab, UC Berkeley, CA, USA

* Equal contribution

Abstract

Humans find structure in visual data; we perceive three-dimensional objects and scenes, even when viewing a static image. Here we evaluate the possibility that a simple learning objective gives rise to this ability: predicting the upcoming visual stimulus, given the current visual input and self-motion. We instantiate this hypothesis *in silico* by optimizing a transformer to predict the future images, conditioned on camera movement and the current image. This requires learning in a continuous setting (i.e., visual sequences, not standalone images), unlike standard computer vision datasets (e.g., ImageNet). To this end, we train a computational model on video datasets collected in a naturalistic 3D environment. As a proof of principle, we demonstrate how this biologically plausible optimization approach generates a visual model that can be used to infer depth, construct 3D shapes, and support cognitive process like mental rotation—all without direct supervision on these tasks. Together, our findings demonstrate how spatial perception might emerge through a biologically plausible learning objective.

Keywords: spatial perception; computational cognitive science

Introduction

Humans and other animals use two-dimensional visual inputs to navigate a three-dimensional world. Our ability to infer spatial relationships (e.g., depth, shape) is foundational to more 'complex' cognitive functions, such as planning and problem solving. For example, inferring the shape-level properties of objects, including unseen surfaces, enables us to understand grasp/interact with objects, even when they have arbitrary, unfamiliar shapes (Gibson, 2014). Remarkably, humans can infer 3D shape from a single image (Hassanin, Khan, & Tahtali, 2021), even though estimating these underlying properties from a 2D input is an ill-posed problem (Pizlo, 2001). There are many theories for how our visual system learns to represent these object-level properties. It is possible that our visual representations emerge from simple learning rules which, given the richness of our sensory environment, lead to robust object-level representations (Saffran, Aslin, & Newport, 1996). Conversely, animals might be born with 'innate' representations of objects which provide the scaffolding for our visual representations (Carey, 1991; Spelke, 1990). Deep learning frameworks offer a test bed to instantiate these hypotheses, and have become increasingly prevalent in the study of human vision (Yamins & DiCarlo, 2016). Within this framework, scientific hypotheses can be used to design and optimize computational models, such that the resulting model behaviors/representations can be used to empirically evaluate these hypotheses (Richards et al., 2019). Here we adopt this deep learning framework to evaluate the possibility that simple learning rules lead to 3D understanding of objects. We formulate this theory by considering the relationship between visual (i.e., 2D images) and proprioceptive (i.e., self-motion) signals in a continuous setting. More concretely, we use a transformer architecture to predict future visual signals (i.e., images), conditioned on self-motion (i.e., camera displacement, between two images) and prior experience (i.e., previous image). As such, our optimization framework reflects a biologically plausible learning mechanism, using an ethologically plausible data distribution. Critically, we suggest that it is possible to learn about the 3D structure of the environment without any explicit geometric prior.

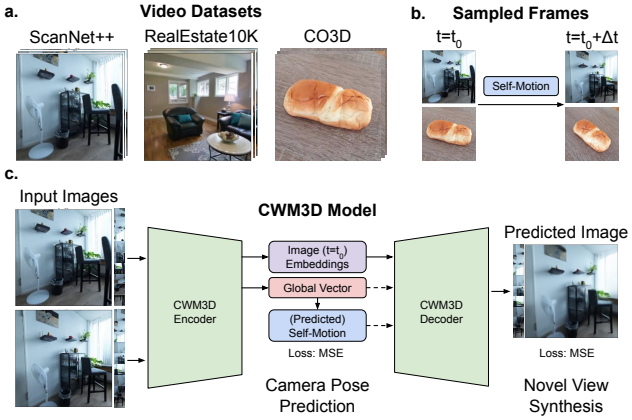

Figure 1: Using the CWM3D model to learn 3D shape understanding from natural video. (a) Ethologically plausible video datasets are used to train the model. (b) Two frames are sampled from the video. (c) The CWM3D model takes two images as input to encode image embeddings and a global vector. The global vector contains information about the global transformation between the two images. Then, the model predicts the self-motion from the global vector. Finally, either the self-motion or the global vector is used to predict the future sensory input given the current sensory input. During training, we optimize the Mean-Squared Error (MSE) loss between predictions and ground truth for both self-motion and visual input.

Method

Model The CWM3D model is designed to predict the future visual sensory input and the self-motion without any explicit geometric prior. The overall framework is illustrated in Figure 1. The model starts from Counterfactual World Modeling (CWM) (Bear et al., 2023), which predicts the future sensory input based on the current sensory input and partial information (e.g., revealed patches) of the future image. In CWM3D, we introduce two modifications to address the head-motion. First, the model encodes the global transformations between two images into an implicit vector (a global vector). This vector, combined with the current frame, is used to predict the future frame in the video and, in practice, includes the information about camera pose changes. Second, we read out a 6-DoF camera motion from the global vector using a linear projection layer. This predicted camera motion is used if the ground truth motion is not provided on a physical scale (e.g., COLMAP (Schönberger & Frahm, 2016) estimation provides arbitrary scale translation). The model has another linear layer

that converts the (predicted) camera motion into the input for the decoder, which predicts the future image using the current image and the camera motion. It is worth noting that camera pose prediction from the global vector opens the possibility of training the model on large datasets without labeled camera motion.

Dataset We used three video datasets to train our model: ScanNet++ (Yeshwanth, Liu, Nießner, & Dai, 2023), RealEstate10K (Zhou, Tucker, Flynn, Fyffe, & Snavely, 2018), and CO3D (Reizenstein et al., 2021). Both the RealEstate10K and CO3D datasets are composed of videos with COLMAP-estimated camera poses. For ScanNet++, we trained metric-scale Neural Radiance Field (Tancik et al., 2023) models to create 3D scenes using images captured from the scenes. Then, we rendered videos with random translations and rotations, simulating the exploration of an agent in static scenes.

Loss Function For both RGB prediction and camera motion prediction, we used Mean Squared Error (MSE) as the loss function to optimize the model.

Result

We illustrate how the CWM3D model can produce novel views with counterfactual self-motion, and use the prediction to infer the depth map and geometry of the object in Figure 2. In Figure 3, we describe how the model can solve the mental rotation task, demonstrating a natural extension of its capabilities.

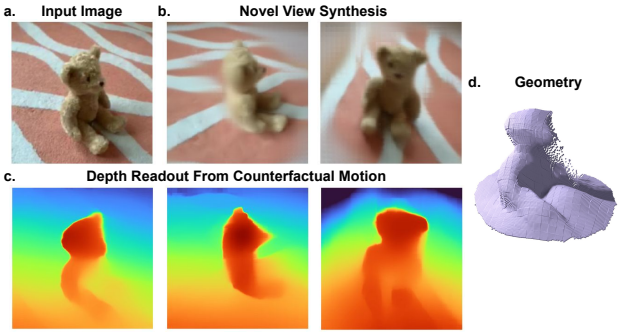

Figure 2: Predicting novel view images, depth maps, and geometry from a single image using the CWM3D model. (a) A single image is provided to the model as input. (b) The model synthesizes the novel view images corresponding to the counterfactual self-motions. (c) We extract the depth map using the predicted image from counterfactual in-plane motions and optical flow algorithm. (d) The predicted geometry of the object is presented as a point cloud. For improved visual clarity, the disparity (inverse of depth) map is visualized in (c) and points of low density are removed from the point cloud in (d).

Novel View Synthesis Given a single image, the model can generate the images from novel views with any desired motion. Rotating around the object provides useful information about the 3D shape, revealing parts that are not visible in the given image.

Depth Readout To obtain depth information from an image, we induce a counterfactual in-plane camera motion that is perpendicular to the camera's view direction. We obtain the disparity map by computing the optical flow between the original image and the predicted image using RAFT (Teed & Deng, 2020). We then invert this disparity map and scale it by the known counterfactual translation distance to obtain depth. We refer to the original CWM paper (Bear et al., 2023) for details of a method to obtain optical flow in a principled way.

Geometry Using the depth map, we construct the point cloud of the scene. It is important to note that CWM3D can obtain depth not only from the input image but also from synthesized novel view images, demonstrating the model's capability to understand the 3D shape of an object.

Mental Rotation In the mental rotation task (Shepard & Metzler, 1971), the participants are presented with two images and need to infer whether the two objects in the images share the same 3D structure. CWM3D has a natural ability to solve the mental rotation task by predicting the camera motion between two images and generating the novel view image based on the predicted motion. The difference between the prediction and the provided image can be used to determine whether the two objects have the identical 3D structure.

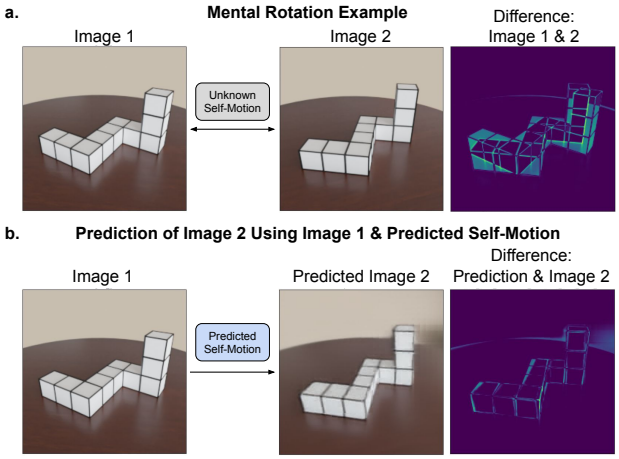

Figure 3: Mental Rotation. (a) Two images are provided with an unknown pose change. The task requires the agent to infer how to rotate the objects in three dimensions to match with each other. The difference between two images is displayed. (b) The model can predict the self-motion between two images, and use it to reconstruct image 2 from image 1. The difference between the ground-truth and the predicted image 2 is plotted.

Conclusion

In this work we introduce CWM3D, a model that learns 3D perception by optimizing a biologically plausible objective without any geometric prior. We demonstrate its capability to perform novel view synthesis, depth estimation, and geometry extraction. Furthermore, we show the model has a natural ability to solve the mental rotation task without direct supervision.

References

Bear, D. M., Feigelis, K., Chen, H., Lee, W., Venkatesh, R., Kotar, K., ... Yamins, D. L. (2023). Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change. *The epigenesis of mind: Essays on biology and cognition*, 257–291.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology press.

Hassanin, M., Khan, S., & Tahtali, M. (2021). Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, *54*(3), 1–35.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision research*, *41*(24), 3145–3161.

Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., & Novotny, D. (2021). Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10901–10911).

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schönberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. In *Conference on computer vision and pattern recognition (cvpr).*

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, *14*(1), 29–56.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., ... Kanazawa, A. (2023). Nerfstudio: A modular framework for neural radiance field development. In *Acm siggraph 2023 conference proceedings.*

Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part ii 16* (pp. 402–419).

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Yeshwanth, C., Liu, Y.-C., Nießner, M., & Dai, A. (2023). Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 12–22).

Zhou, T., Tucker, R., Flynn, J., Fyffe, G., & Snavely, N. (2018). Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*.