Learning 3D object-centric representation through prediction with inductive bias of infants

John Day[*] **(johnmday@umich.edu)**
International Research Center for Neurointelligence, The University of Tokyo
Bunkyo-ku, Tokyo, 113-0033, Japan

Tushar Arora[*]**(tushar@bu.edu)**
Boston University
Boston, MA 02215, USA

Jirui Liu (liujirui2000@outlook.com)
Laboratory of Brain and Intelligence, Tsinghua University
Beijing, 100084, China

Li Erran Li (erranli@gmail.com)
AWS AI, Amazon

Ming Bo Cai (mingbo.cai@miami.edu)
Department of Psychology, University of Miami
Coral Gables, FL 33146, USA

Abstract

As part of human core knowledge, object is the building block of mental representation that supports high-level concepts and symbolic reasoning. Infants develop the notion of objects situated in 3D environments without supervision. Towards understanding the minimal set of assumptions needed to learn object perception, we investigate a predictive learning approach to learn three key abilities without supervision: a) segmenting objects from images, b) inferring objects' locations in 3D and c) perceiving depth. Critically, we restrict the input signals to those available to infants, namely, only streams of visual input and information of self-motion, mimicking the efference copy in the brain. In our framework, objects are latent causes of scenes constructed by the brain that facilitate efficient prediction of the future sensory input. All the three abilities are by-products of learning to predict. The model includes three networks that learn jointly to predict the next-moment visual input based on two previous scenes. This work demonstrates a new approach to learning symbolic representation grounded in sensation.

Keywords: object perception; unsupervised learning; infant development; predictive learning

Object perception inspired by infant learning

Modern computer vision enjoys many advantages in its learning materials unavailable to infants: labels of objects in millions of images (Deng et al., 2009), pixel-level annotation of object boundaries for object segmentation task and depth information from Lidar to learn 3D perception. Although deep networks powered by such labeled data excel in many vision tasks, the reliance on supervision limits their ability to generalize to objects of unlabeled categories. In contrast, the brain is able to acquire general perceptual abilities such as object segmentation and 3D perception in the first few months of life without supervision or knowledge of object categories (Spelke, 1990), allowing it to adapt to new environment with unknown objects. Motivated by this contrast, there has been a surge of object-centric representational learning (OCRL) models with unsupervised or self-supervised learning in recent years. Nonetheless, most of them only achieve object segmentation in 2D images, lacking 3D representation. Some require additional information not *directly* available to the brain (such as depth, optical flow and object bounding boxes (Elsayed et al., 2022)).

Decades of developmental research suggests that infants likely honor a few principles reflecting basic constraints of physical objects to perceive objects long before language acquisition (Spelke, 1990). Among these principles, we examine whether *rigidity*, an assumption that objects move rigidly, together with the principle of predictive learning, are sufficient to allow neural networks to learn object perception with only access to signals similar to those available to infant brains: only streams of visual input and efference copy (a copy of self-motion information from motor cortex).

We reason that with the mental construct of objects as discrete entities, the brain can utilize the properties of rigid bodies to efficiently predict the cohesive motion of all visible points on an object (both in 3D space and 2D retinal image) by keeping track of only a few motion parameters of the whole object after perceiving its shape. Depth perception, 3D localization and object segmentation are all skills needed for such prediction. We hypothesize that they may arise jointly as byproducts

[*] equal contribution

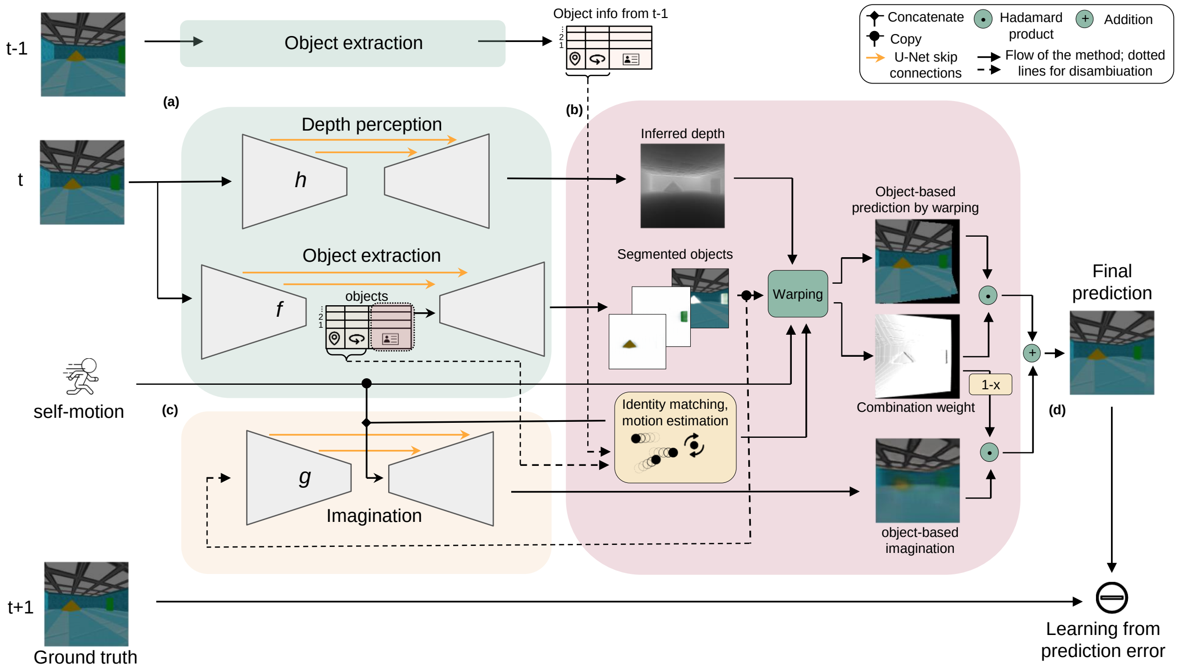

Figure 1: Architecture for the *Object Perception by Predictive LEarning (OPPLE)* network: The model includes three networks for depth perception, object extraction and imagination to predict the image at $t+1$ using those at $t-1$ and t . All three are convolutional networks with U-Net structure(Ronneberger et al., 2015). **(a)** From the image at time t, the object extraction network f outputs the location, pose, an identity code (from the encoder) and a probabilistic segmentation map (from the decoder) for each object. This is achieved by a recurrent neural network inserted between the encoder and decoder of f. Depth perception network h infers a depth map. **(b)** A matching score between any objects of two frames based on the distances between their identity codes is used to weight the velocities estimated for each object at t using its inferred location and that of any candidate object at $t-1$ to obtain an estimated velocity for that object. Self motion and estimated object motion are used together with object segmentation and depth maps to predict the image at $t+1$ by shifting each pixel at t. **(c)** The segmented object images and depth at t, together with all motion information, are used by the imagination network to 'imagine' the regions not predictable by warping. **(d)** The predictions based on warping and by imagination are merged according to the extent that each pixel is predictable by warping. The error between the actual image and the prediction provides major teaching signals for all networks.

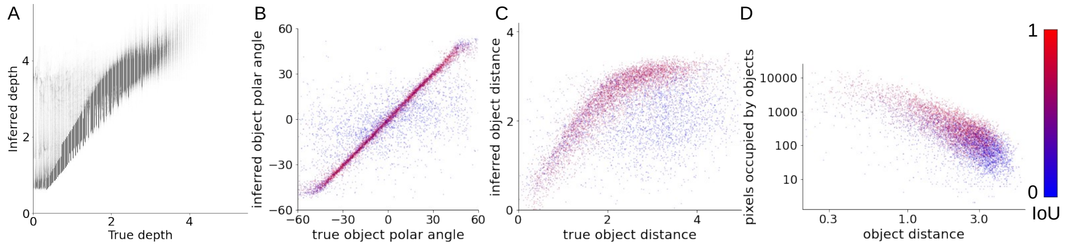

Figure 2: Performance of depth inference (A) and 3D object localization (B,C). D: Nearer and larger objects are segmented better.

of learning to improve the accuracy of predicting the future visual inputs. In addition, due to occlusion and self motion, some parts of a new scene are invisible before. Those parts may be predicted based on the statistical regularity of scenes learned from experience. Based on this reasoning, we implemented a model which integrates two approaches of prediction: warping current visual input based on predicted optical flow and 'imagining' regions unpredictable by warping based on statistical regularity in environments. We name the model as Object Perception by Predictive LEarning (OPPLE).

Performance on 3D object perception

Our model contains two convolutional neural networks tasked to explicitly infer from an image all objects' 3D locations, poses, probabilistic segmentation map, a latent code representing each object, and the distance of each pixel from the camera (depth). Their outputs for two consecutive images are used to predict the next image by combining the above-mentioned two approaches of prediction: warping the current image based on the predicted optical flow with the assumption of rigid objects, and implicit 'imagination' of the newly

Model	ARI-fg	IoU
MONet	0.36	0.20
slot-attention-128	0.34	0.38
slate	0.30	0.20
AMD	0.19	0.02
O3V	0.37±0.01(3)	0.22±0.10(3)
OPPLE (our model)	**0.58**±**0.07**(6)	**0.45**±**0.02**(6)

Table 1: Performance of models on object segmentation. (**A**) The inferred depth exhibits high correlation with ground truth ($r = 0.92$). (**B,C**) Inference of object 3D locations (correlations of $r = 0.86$ for bearing angle and $r = 0.51$ for distance against ground truth), (**D**) Closer and bigger objects (red dots) are segmented better.

visible parts based on the statistical regularity of the environment by a third convolutional network. We train all three networks jointly by minimizing the prediction error and a few losses that encourage the consistency between quantities *inferred* from the new image and those *predicted* based on the first two images (Fig **??**). We evaluate our networks on a dataset of scenes with moving objects and camera with complex surface texture against a few state-of-the-arts object-centric representational learning models: MONet(Burgess et al., 2019), slot-attention(Locatello et al., 2020) (a scaled-up version), SLATE(Singh, Deng, & Ahn, 2021), AMD(Liu, Wu, Yu, & Lin, 2021) and O3V(Henderson & Lampert, 2020). Table 1 illustrates that our model outperforms several state-of-the-arts unsupervised object-centric learning models in object segmentation based on two common metrics.

References

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., & Lerchner, A. (2019). Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Elsayed, G. F., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M. C., & Kipf, T. (2022). Savi++: Towards end-to-end object-centric learning from real-world videos. *arXiv preprint arXiv:2206.07764*.

Henderson, P., & Lampert, C. H. (2020). Unsupervised object-centric video generation and decomposition in 3d. *Advances in Neural Information Processing Systems*, *33*, 3106–3117.

Liu, R., Wu, Z., Yu, S., & Lin, S. (2021). The emergence of objectness: Learning zero-shot segmentation from videos. *Advances in Neural Information Processing Systems*, *34*, 13137–13152.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., ... Kipf, T. (2020). Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Singh, G., Deng, F., & Ahn, S. (2021). Illiterate dall-e learns to compose. In *International conference on learning representations.*

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