Enforcing Object Permanence using Hierarchical, Object-Centric Generative Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Object permanence is an important milestone in infant development, when the infant understands that an object continues to exist even when it no longer can be seen. However, current machine learning methods devised to build a world model to predict the future still fail at this task when having to deal with longer time sequences and severe occlusions. In this paper, we compare current machine learning with infant learning, and propose an object-centric approach on learning predictive models. This grounds object representations to an inferred location, effectively resolving the object permanence problem. We demonstrate performance on a novel object-permanence task in a simulated 3D environment.

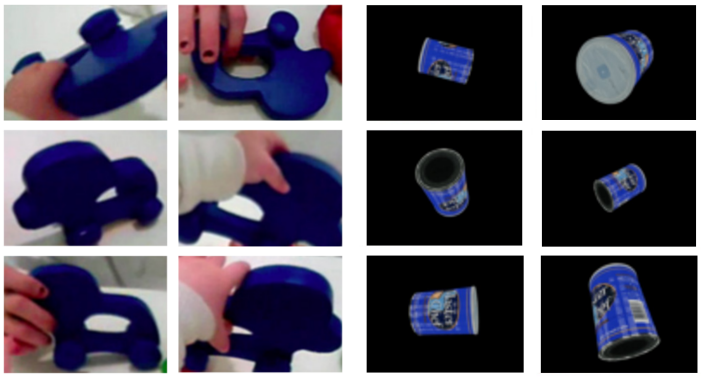
1 Introduction

An important milestone in infant development is to understand object permanence (Piaget, 1954). This is the ability to know that objects continue to exist, even when they are hidden from sight. From a neuroscience perspective, learning is often cast as the brain minimizing prediction error, also coined free energy (Friston et al., 2016). This appeals to machine learning approaches, in which parameterized models such as variational autoencoders (VAE) can be trained for predicting future sensory observations, effectively building a model of the world (Çatal et al., 2020; Hafner et al., 2020). Despite impressive results in learning representations in various domains, we found that current implementations are still unable to capture longer term dependencies, and hence fail at simple object permanence tasks.

To address this problem, we propose two novelties for learning such world models, inspired by human learning. First, we devise a hierarchical structure of the generative model, in which on the one hand, inferences are made about the kind of objects that are present, and on the other hand a global scene representation is constructed on where these objects are located in 3D space. This is consistent with recent accounts of the human vision system (Parr et al., 2021) which considers perception as inferring a scene as a factorization of separate (parts of) objects, their identity, scale and pose. Distinguishing scale and pose also matches the two stream hypothesis, which states that visual information is processed by a dorsal ("where") stream, and a ventral ("what") stream (Mishkin et al., 1983). Also, recent findings in recordings of rhesus monkey brains provide evidence that indeed 3D shape is encoded in the inferior temporal cortex (Janssen et al., 2000).

Second, we adapt the data regime with which these object-centric models are trained. In machine learning, representations are typically learned from large datasets of images (Grill et al., 2020) or videos (Qing-Yuan Jiang & Li, 2019) of a wide variety of objects and scenes. This is in stark contrast to toddlers, who learn from very object-centric views, by actively engaging with particular objects, as shown on Figure 1a (James et al., 2014; Slone et al., 2019). In similar vein, we therefore also train our object-centric models through novel view predictions of a single object instance in different

Submitted to 4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at NeurIPS 2022. Do not distribute.



(a) toddler object views (b) simulated object views

Figure 1: (a) First person views of a 15 month old toddler are typically centered around a single object (adapted with permission from (Slone et al., 2019)). (b) Object-centric views sampled from our dataset for training the generative model.

poses (Van de Maele et al., 2022), as illustrated in Figure 1b. This way, our agent can infer distinct objects and their location in space, and maintain this knowledge to predict other viewpoints, even after complete occlusions. In the remainder of this paper, we summarize our approach in Section 2, show some preliminary results of this system in Section 3, and discuss some related work and future work in Section 4.

2 Object-Centric Generative Models

Our method is based on object-centric generative models that learn latent representations of object pose and identity (Van de Maele et al., 2022). Such models consist of an encoder, which maps pixels to an object identity and pose latent, a decoder which decodes identity and pose to a novel pixel observation, and a transition model which learns how the pose latent changes after moving the camera with an action. In addition to the encoder described in Van de Maele et al. (2022), we first predict the mask using m_ϕ and mask the observation before feeding it to the encoder. Learning to filter out anything that is not part of the object makes it more resilient against occlusions and variations in the background. Specifically, we model the following components using deep neural networks:

$$\text{Mask:} \quad \mathbf{x}_{t,m} = m_\phi(\mathbf{x}_t), \qquad \text{Transition:} \quad p_\phi(\mathbf{p}_t|\mathbf{p}_{t-1}, \mathbf{a}_{t-1}),$$

$$\text{Encoder:} \quad q_\phi(\mathbf{p}_t, i|\mathbf{x}_{m,t}) = q_\phi(\mathbf{p}_t|\mathbf{x}_{t,m})q_\phi(i|\mathbf{x}_{t,m}), \qquad \text{Decoder:} \quad p_\phi(\mathbf{x}_t|\mathbf{p}_t, i),$$

These models are trained end to end using a prerecorded dataset of an agent "playing around" with an object, i.e. capturing different views of the object, and predicting the next view. The dataset consists of renders from simulated meshes from objects in the YCB dataset (Calli et al., 2015), which are augmented with random background colors and random occlusions of different negative objects. We optimize each model through the minimization of the negative evidence lower bound (ELBO). The reader is referred to the appendix for more details on optimization.

Next, we employ these models on pixel-based observations of a full scene containing multiple objects in separate distinct locations. We infer the different visible object identities and their corresponding poses in the scene. Using the predicted mask for each object, we can infer the pixel coordinates (uv-space) of the center for each object. Combining the pixel-coordinates with a pinhole model of the camera, we can now compute a belief over the object position using a 3D Multivariate Gaussian over the casted ray from the camera pose to the object center. Over multiple timesteps, information of each frame is aggregated using a Bayesian belief update rule. We call this collection of identities and their inferred positions the "Global Scene Representation" as illustrated in Fig. 2. For each of the detected objects, we also have an "Object-Centric Representation" - representing the orientation and scale of the object - acquired through encoding an object-centric crop of the object with their respective object-specific model.

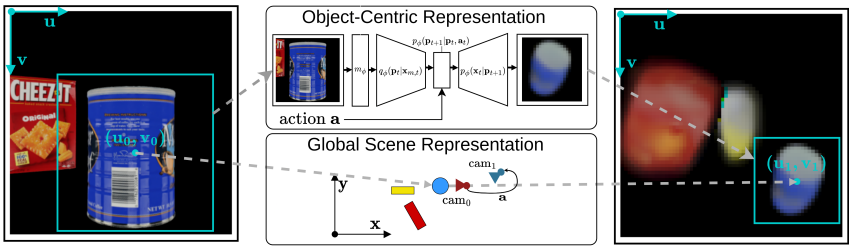


Figure 2: Flow for inferring a scene representation using object-centric models. Individual objects are detected using the object-centric models. Using the location of the object in the image and a pinhole camera model, the location of each object can be estimated. A novel viewpoint can now be rendered by estimating the pose and location of each object in the image, and compositing the predictions of the different object-centric models.

To imagine a novel view, the camera pose is first acquired by applying the action to the current camera pose. Then, for each of the objects, an inverted projection is applied to acquire the pixel coordinate where the center of the object would be observed. For each of the objects, the action is applied to their transition model, and using the likelihood, a novel observation is generated. These are then composed around the predicted object center on the novel image canvas. The z-index of each object is acquired through the estimated distance of the object to the camera.

3 Object Permanence

To evaluate the ability of object permanence, we created a 3D simulation environment in which three objects from the YCB dataset (Calli et al., 2015) are rendered at distinct locations. The model is first given the first 16 images, in which an object becomes occluded due to the camera moving (Fig. 3.a). The goal is then to predict, given the consecutive actions, the next 16 images in which the camera moves further and the initial object becomes back in view (Fig. 3.b).

We compare our approach with a sequential VAE model (SSM) that is trained directly on a dataset of sequences of actions and observations as used during evaluation (Çatal et al., 2020), as well as a recurrent state space model (RSSM) as proposed by Hafner et al. (2020). Fig. 3 shows our results. While the final object position and pose are not perfect, our model is able to estimate the object presence over time (c). In the SSM and RSSM model, this is not the case and the model does not remember any of the objects, and just imagines a novel view.

Quantitatively, we evaluate our approach for permanence using an adapted mean squared error (MSE), computed only over the pixels of the earlier occluded object on a test set consisting of 1000 randomly generated sequences, which we can compute as we acquire the object mask directly from our environment. Since the distance of the agent to the object has a large impact on the error's order of magnitude, we scale this metric with the inverse amount of visible object pixels. We show these results in Table 1. In addition, we report a detection rate, which determines whether or not the agent has captured the correct object in the final frame. For this we count the amount of pixels within the object mask, where the error is lower than an empirically determined threshold of 0.20, after which we count the objects where more than 25% of the object is correctly reconstructed. We observe that our approach has the lowest adapted MSE and the highest object detection rate of the compared models, which further supports our claims from the qualitative results.

4 Discussion

Developing commonsense is still an outstanding challenge in artificial intelligence (Davis & Marcus, 2015). Object permanence is a fundamental commonsense skill that emerges early during infant development (Piaget, 1954). Although recent advances in learning world models from sequences of observations and actions (Hafner et al., 2020), these still struggle on learning long-term dependencies. In order to improve long-term sequence predictions, one strategy is to add a hierarchy of latent sequences, where each level predicts at different time intervals (Saxena et al., 2021). This however is

Table 1: Quantitative comparison of the state space model (SSM), the recurrent state space model (RSSM) and our ensemble created from CCNs. We compare on the adapted MSE, a scaled MSE computed over the relevant pixels for the occluded object, and the percentage of scenes in which the agent is still able to reconstruct the correct object at the end of the sequence. This was run over 1000 scenes.

	adapted MSE	detected (%)
SSM	0.4029 ± 0.2653	32.9
RSSM	0.4466 ± 0.2841	37.3
Scene CCN (ours)	0.3535 ± 0.2645	60.0

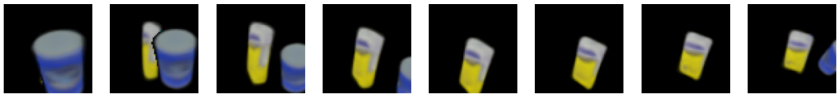
not trivial for action-conditioned scenarios, as there isn't necessarily an "action" construct at these higher, temporal levels (Zakharov et al., 2022).

A novel way of evaluating intuitive physics such as object permanence, is by evaluating whether or not an AI system renders an physically implausible trial as plausible, or vice versa (Moore et al., 2022). In similar vein, S. Punla et al. (2022) assess the Violation of Expectation (VoE) by measuring the surprise of the model. In addition, S. Punla et al. (2022) propose the PLATO model, which they demonstrate can learn a diverse set of physical concepts, among which is object permanence. Similar to our approach, PLATO also uses an object-centric representation, with a learned object code for each object. In contrast to our approach, they learn the object codes in an end-to-end fashion from sequential data, but using more simple primitives and a fixed camera viewpoint. Still, this illustrates that an hierarchical, object-centric scene representation might be key to develop spatial commonsense understanding.

Our approach uses a crop of the object to compute the pose, after which the action with respect to the center of the observation is applied. There will automatically be an offset as this action does not entirely match the movement with respect to the object. In cases the agent requires a precise estimate over the object pose, it is thus crucial to first move to a viewpoint where the object is central. While our approach does have a higher detection rate than the compared baselines, it uses an inductive bias

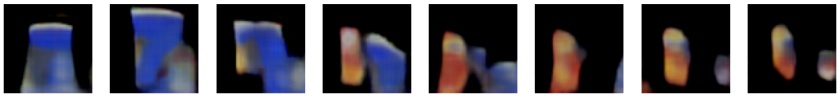
(a) input sequence

(b) output sequence



(c) prediction with our model

(d) prediction with the SSM



(e) prediction with the RSSM model

Figure 3: Sequence of observations where an object is occluded, together with predictions by the different models.

that the scene is static and that the objects do not change over time. A potential extension would be to have a belief over object presence, which can be updated as well when evidence shows that an object has been removed from the expected position. In future work we plan to address these limitations and will work on methods where the higher level of the hierarchy is also learned end-to-end.

References

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pp. 510–517, 2015. doi: 10.1109/ICAR.2015.7251504.

Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, aug 2015. ISSN 0001-0782. doi: 10.1145/2701413. URL `https://doi.org/10.1145/2701413`.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John ODoherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68: 862–879, September 2016. ISSN 01497634. doi: 10.1016/j.neubiorev.2016.06.022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Karin H. James, Susan S. Jones, Linda B. Smith, and Shelley N. Swain. Young children's self-generated object views and object recognition. *Journal of Cognition and Development*, 15(3): 393–401, 2014. doi: 10.1080/15248372.2012.749481.

Peter Janssen, Rufin Vogels, and Guy A Orban. Three-dimensional shape coding in inferior temporal cortex. *Neuron*, 27(2):385–397, August 2000. doi: 10.1016/s0896-6273(00)00045-3. URL `https://doi.org/10.1016/s0896-6273(00)00045-3`.

Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, January 1983. doi: 10.1016/0166-2236(83)90190-x. URL `https://doi.org/10.1016/0166-2236(83)90190-x`.

David S. Moore, Lisa M. Oakes, Victoria L. Romero, and Koleen C. McCrinck. Leveraging developmental psychology to evaluate artificial intelligence. In *IEEE International Conference on Development and Learning (ICDL)*, 2022.

Thomas Parr, Noor Sajid, Lancelot Da Costa, M. Berk Mirza, and Karl J. Friston. Generative Models for Active Vision. *Frontiers in Neurorobotics*, 15:651432, April 2021. ISSN 1662-5218. doi: 10.3389/fnbot.2021.651432. URL `https://www.frontiersin.org/articles/10.3389/fnbot.2021.651432/full`.

Jean Piaget. *The Construction of Reality in the Child*. Routledge, 1954.

Gen Li Jian Lin Lei Li Qing-Yuan Jiang, Yi He and Wu-Jun Li. SVD: A large-scale short video dataset for near-duplicate video retrieval. In *Proceedings of International Conference on Computer Vision*, 2019.

Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv:1810.00597 [cs, stat]*, October 2018. URL `http://arxiv.org/abs/1810.00597`. arXiv: 1810.00597.

Candida S. Punla, https://orcid.org/ 0000-0002-1094-0018, cspunla@bpsu.edu.ph, Rosemarie C. Farro, https://orcid.org/0000-0002-3571-2716, rcfarro@bpsu.edu.ph, and Bataan Peninsula State University Dinalupihan, Bataan, Philippines. Are we there yet?: An analysis of the competencies of BEED graduates of BPSU-DC. *International Multidisciplinary Research Journal*, 4(3): 50–59, September 2022.

Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=fU7-so5RRhW`.

Lauren K. Slone, Linda B. Smith, and Chen Yu. Self-generated variability in object images predicts vocabulary growth. *Developmental Science*, 22(6):e12816, 2019. doi: https://doi.org/10.1111/desc.12816. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12816`.

Toon Van de Maele, Tim Verbelen, Ozan Çatal, and Bart Dhoedt. Embodied object representation learning and recognition. *Frontiers in Neurorobotics*, 16, 2022. ISSN 1662-5218. doi: 10.3389/fnbot.2022.840658.

Alexey Zakharov, Qinghai Guo, and Zafeirios Fountas. Variational predictive routing with nested subjective timescales. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=JxFgJbZ-wft`.

Ozan Çatal, Samuel Wauthier, Cedric De Boom, Tim Verbelen, and Bart Dhoedt. Learning Generative State Space Models for Active Inference. *Frontiers in Computational Neuroscience*, 14:574372, November 2020. ISSN 1662-5188. doi: 10.3389/fncom.2020.574372.

A Appendix

A.1 Optimization of the Object-Centric Models

These models are optimized using the minimization of the negative ELBO over a trajectory of the agent. In our implementation, we only consider sequences of one transition, and thus two timesteps ($T = 1$), as the environment is static and the agent can teleport/move to any viewpoint of the object. For each timestep, a reconstruction error (MSE) is applied to both the reconstruction of the observation using the encoder model p_ϕ and the predicted mask $m_\phi(\mathbf{x}_t)$ directly, as well as a binary cross entropy over the predicted identity, which is equivalent to a KL-divergence between the approximate posterior and the marginal probability over identity i. The approximate posterior is conditioned on the masked input $\mathbf{x}_{t,m} = m_\phi(\mathbf{x}_t) \odot \mathbf{x}_t$. Starting from $t = 1$, the model can predict future observations and transition the latent pose variable. A reconstruction loss is also computed for the likelihood of each transitioned latent. Finally, a KL-divergence term draws the belief over the transitioned latent towards the belief over pose acquired through the approximate posterior. Additionally, all learned distributions are also regularized using a KL-divergence with the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$. This is optimized using constraint optimization, where the Lagrange multiplier weights λ_i are scaled automatically for the terms to stay below predefined constraints (Rezende & Viola, 2018).

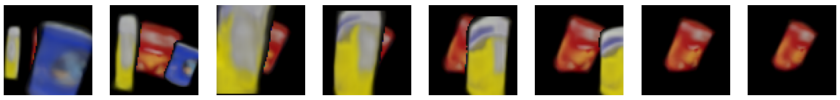
$$
\begin{aligned}
L = &\sum_{t=0}^{T} \lambda_{x,t} ||\mathbf{x}_t - p_\phi(\mathbf{x}_t|\mathbf{p}_t, i)||_2 + \lambda_{i,t} D_{KL}[q_\phi(i|\mathbf{x}_{t,m})||p(i)] + \lambda_{m,t} ||m_\phi(\mathbf{x}_t) - \mathbf{m}_t||_2 + \\
&\quad D_{KL}[q_\phi(\mathbf{p}_t|\mathbf{x}_{t,m})||\mathcal{N}(\mathbf{0}, \mathbf{1})] + \\
&\sum_{t=1}^{T} \lambda_{\text{trans},t} ||\mathbf{x}_t - p_\phi(\mathbf{x}_t|\mathbf{p}_{t,\text{trans}})||_2 + D_{KL}[q_\phi(\mathbf{p}_t|\mathbf{x}_{t,m})||p_\phi(\mathbf{p}_{t,\text{trans}}|\mathbf{p}_{t-1}, \mathbf{a}_{t-1})] + \\
&\quad D_{KL}[p_\phi(\mathbf{p}_{t,\text{trans}}|\mathbf{p}_{t-1})||\mathcal{N}(\mathbf{0}, \mathbf{1})]
\end{aligned}
\tag{1}
$$

A.2 Additional Qualitative Results

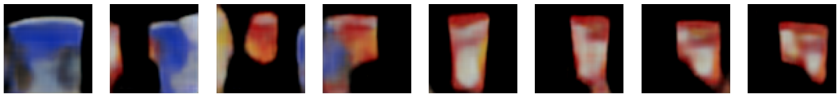
In Figure 4, we show a prediction of a sequence where the agent has seen multiple observations of the object without occlusion. We can first observe that in the CCN-based model, correct inference

(a) input sequence

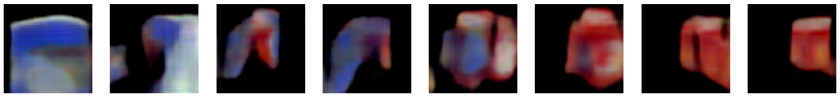
(b) output sequence



(c) prediction with our model



(d) prediction with the sequential VAE



(e) prediction with the RSSM model

Figure 4: Sequence of observations where an object is only partially occluded in the final frame, together with predictions by the different models.

has been made about the presence of the sugar box, even though it has only been partially observed, while the other models do not have enough information to start reconstructing the sugar box. It can also be observed that when the agent has a large set of evidence about the occluded object and/or the object is not completely occluded, it will have a better prediction and does not forget the object, also in case of the SSM and RSSM models.