Developing Object Permanence from Videos

Frederic Becker[1] (frederic.becker@uni-tuebingen.de),
Manuel Traub[1] (manuel.traub@uni-tuebingen.de),
Sebastian Otte[2] (sebastian.otte@uni-luebeck.de),
Martin V. Butz[1] (martin.butz@uni-tuebingen.de)
[1] Cognitive Modeling, Department of Computer Science and Department of Psychology, University of Tübingen
Sand 14, 72076 Tübingen, Germany
[2] Adaptive AI Lab, Institute for Robotics and Cognitive Systems, University of Lübeck
Ratzeburger Allee 160, 23562 Lübeck, Germany

Abstract

Humans learn that temporarily occluded objects continue to exist within the first months of their lives. Deep learning models, on the other hand, struggle to generalize such concepts from observations, due to missing proper inductive biases. Here, we introduce the first self-supervised interpretable machine learning model that learns about object permanence directly from video data without supervision. We augment a slot-based autoregressive deep learning system with the ability to adaptively and selectively fuse latent imaginations with pixel-based observations into consistent object-specific 'what' and 'where' encodings over time. We show that (i) Loci-Looped tracks objects through occlusions and anticipates their reappearance while outperforming state-of-the-art baseline models, (ii) Loci-Looped shows signs of surprise when the principle of object permanence is violated, and (iii) Loci-Looped's internal latent loop is key for learning object permanence.

Keywords: intuitive physics learning; machine learning; object-centric cognition; compositional scene representation; information fusion

Introduction

In infancy, humans develop an impressive intuitive understanding of the fundamental principles governing our physical world, manifested through expectations of how objects behave and interact (Aguiar & Baillargeon, 1996; Y. Lin, Stavans, & Baillargeon, 2022; Summerfield & Egner, 2009). These expectations have been explicitly probed with the Violation-of-Expectation (VoE) paradigm (Baillargeon, Spelke, & Wasserman, 1985). In VoE experiments, infants are presented with videos that either adhere to (e.g., an occluded object reappears) or violate (e.g., an occluded object vanishes) a physical concept, while their gaze behavior is monitored. If they look longer at physical violations compared to similar plausibly unfolding scenes, we conclude that they have developed an understanding of the physical concept that was violated. By means of the VoE paradigm it was demonstrated that infants as young as 2.5 months become able to reason about hidden objects and their behavior (Baillargeon & DeVos, 1991). The understanding of object persistence, object inertia, and object solidity, is indeed believed to be part of a core knowledge system that is building the foundations for more complex cognitive processes (Butz, 2021; Lake, Ullman, Tenenbaum, & Gershman, 2016; Spelke & Kinzler, 2007; Spelke, Breinlinger, Macomber, & Jacobson, 1992; Y. Lin et al., 2022).

In comparison to human cognition, state-of-the-art machine learning (ML) models lack this understanding (Weihs et al., 2022). Building on a long tradition (Munakata, Mcclelland, Johnson, & Siegler, 1997), there has been a recent surge of interest in both benchmarking and constructing ML systems that learn to develop intuitive physics (Weihs et al., 2022; Piloto, Weinstein, Battaglia, & Botvinick, 2022; Smith et al., 2019; Riochet et al., 2022). For example, Piloto et al. (2022) recently trained a deep learning model on next-frame prediction tasks using videos, and subsequently evaluated its physical knowledge through the VoE paradigm. However, approaches like theirs, as well as those by Smith et al. (2019) and Riochet et al. (2022), rely on supervised information that includes object-respective ground truth masks, offering specific details about the location and identity of each object in the scene. The challenge of learning object permanence was thus partially side-stepped. Addressing the segregation problem (Greff et al., 2020), i.e., segmentation and tracking, while simultaneously developing object permanence in one model remains an open challenge.

In this work, we learn the concept of object permanence with a recently introduced self-supervised segmentation and tracking model named Loci-v1 (Traub, Otte, et al., 2023), which incorporates cognitively inspired inductive biases. The model design is motivated by the interplay of two cognitive systems, which together enable the development of intuitive physics in infants (Y. Lin et al., 2021). The Object File System (OFS) constructs temporary representations of the 'where' and 'what' of objects. The Physical Reasoning System accesses the OFS to predict how object interactions will unfold based on acquired physical knowledge. Similarly, Loci-v1 represents a scene as a composition of objects while disentangling the position and the appearance of each object into expressive latent codes. The model then produces next time step predictions by explicitly modeling per-object dynamics using recurrent units, as well as inter-object interactions using attention mechanisms. Learning signals are computed from the next-frame prediction error resembling the idea of predictive coding (Clark, 2013; Butz, Achimova, Bilkey, & Knott, 2021; Butz, 2008; Den Ouden, Kok, & De Lange, 2012; Lotter, Kreiman, & Cox, 2017).

Concretely, we hypothesize that object permanence may emerge from building a strong latent world model that continually generates predictions about next world states, while only making sparse use of sensory observations. Key to this is an efficient and adaptive information fusion process of obser-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

vations and predictions. We test our hypothesis by augmenting Loci-v1 with the ability to fuse latent temporal imaginations with pixel-space observations into consistent compositional scene percepts. While an outer sensory loop allows our augmented model, named Loci-Looped, to build and update representations of visible objects, the novel inner-loop allows to imagine object-centric latent state dynamics—much like the dreamer architecture (Hafner, Lillicrap, Ba, & Norouzi, 2020; P. Wu, Escontrela, Hafner, Abbeel, & Goldberg, 2023), but on an explicit object-oriented level. We show that the inner-loop enables Loci-Looped to simulate the state of temporarily hidden objects over time. Importantly, Loci-Looped learns without supervision to adaptively fuse external, sensory information with internal, anticipated information for each object individually via a parameterized percept gate. As a result, we show in our experiments that Loci-Looped learns to imagine the trajectory of temporarily hidden objects, thereby developing the principles of object permanence and directional inertia. Our ablation studies confirm that the inner-loop and the flexible control of it are key for learning this behavior.

Method

We give a brief introduction to Loci-v1 (Traub, Otte, et al., 2023) including its formalization. We then introduce our novel developments defining Loci-Looped.

Loci-v1

Loci-v1 consists of three main components: an encoder module that parses visual information into object representations, a transition module that projects these representations into the future, and a decoder module that reconstructs a visual scene from this prediction. Each of the three components comprises k slots that share their weights. Each slot is dedicated to process one object. It may stay empty when more slots than objects are available.

The ResNet-based, slotted encoder module receives the current frame I^t, the previous prediction error E^t, a background mask $\hat{M}^t_{bg}$ as well as slot-specific predictions of position $\hat{Q}^t_k$, visibility mask $\hat{M}^{t,v}_k$, RGB object reconstruction $\hat{R}^t_k$, and the summed visibility mask of the remaining slots $\hat{M}^{t,s}_k$. Positions are encoded as isotropic Gaussians in pixel space, masks as grayscale images. The encoder produces gestalt codes $\tilde{G}^t_k$ and positional codes $\tilde{P}^t_k$ as output. Gestalt codes encode shape and surface patterns, while positional codes include object location (x_k, y_k), size (σ_k), and priority (ρ_k).

The transition module predicts the encodings at the next timestep, namely $\hat{G}^{t+1}_k$ and $\hat{P}^{t+1}_k$. It implements a combination of slot-wise recurrent layers to model object dynamics and across-slot attention layers to model object interactions. In contrast to the PLATO model (Piloto et al., 2022), the recurrent layers do not receive a history of object states depicting previous object dynamics. Following the transition module, the gestalt codes are binarized, creating an information bottleneck that biases the slots to develop factorized compositional encodings of entities.

The decoder module then reconstructs the predicted scene from $\hat{G}^{t+1}_k$ and $\hat{P}^{t+1}_k$ and a provided image of the scene background, which marks the only supervised model input. For each slot, a ResNet architecture produces the predicted RGB object reconstruction $\hat{R}^{t+1}_k$, visibility mask $\hat{M}^{t+1,v}_k$, and position $\hat{Q}^t_k$. All slot outputs are unified in the prediction $\hat{R}^{t+1}$ by taking the sum over the RGB object reconstructions weighted by the visibility masks and the background mask. Along with the next input frame I^{t+1} the prediction serves to generate prediction error E^{t+1}. This process repeats in each timestep.

Loci-Looped

Object mask Visibility masks outputted by most compositional scene representation models exclusively depict visible components of objects. To enable a holistic scene understanding, we introduce an extra mask that is designed to encode entire object shapes, which serves as an additional input to the encoder of Loci-Looped. To compute this mask, we assume that only slot-object k is in the scene, ignoring the remaining slots. Consequently, in the decoding process slot k only competes with the background for visibility yielding object mask

$$M^{t,o}_k = \frac{\exp(M^t_k)}{\exp(M^t_k) + \exp(M^t_{bg})}, \tag{1}$$

where M is generated by the decoder. Figure 1 illustrates the difference between the visibility mask and the object mask. The latter encodes the complete 2D object shape, while the former only depicts those parts that are currently visible.

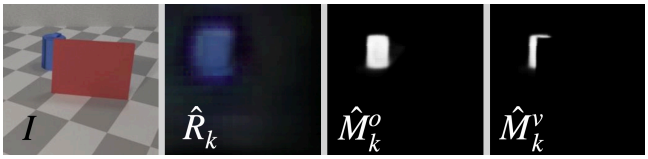

Figure 1: The object and visibility mask enable an interpretable holistic scene understanding in Loci-Looped. *From left to right*: Current video frame, reconstructed RGB object, object mask and visibility mask of slot k depicting the blue object.

Occlusion state The introduction of the object mask enables Loci-Looped to determine the degree of occlusion for each object. We calculate the occlusion state O^t_k as follows:

$$O^t_k = 1 - \frac{\sum_{i,j}[M^{t,v}_k(i,j) > \theta]}{\sum_{i,j}[M^{t,o}_k(i,j) > \theta] + c}, \tag{2}$$

where θ is a threshold value, which we set to 0.8, and c is a small constant. By counting the number of pixels larger than threshold θ, the denominator determines the total area of the object, while the numerator determines the visible area of the object. The occlusion state ranges from 0 (fully visible) to

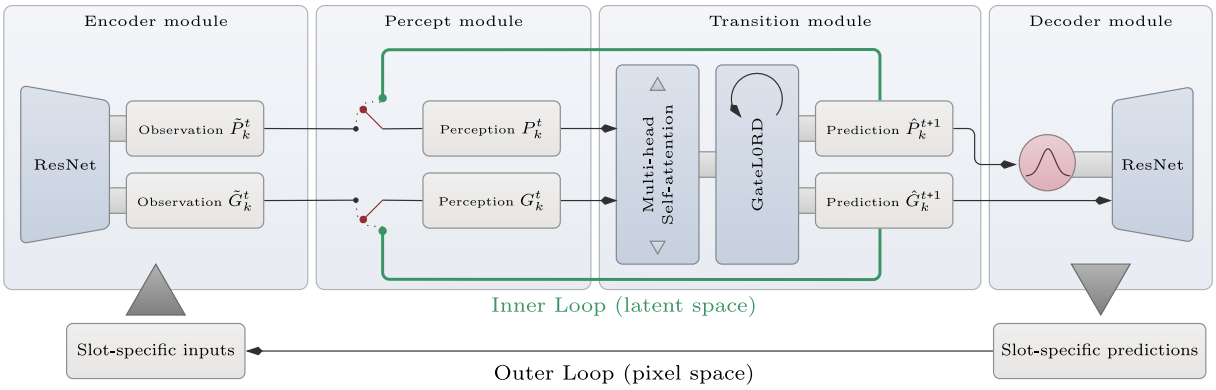

Figure 2: The slot-wise processing architecture of Loci-Looped. Predictions are made available on two routes. First, through an outer-loop in pixel-space enabling consistent object tracking over time. Second, through an inner-loop allowing for latent imaginations.

1 (fully occluded), allowing Loci-Looped to explicitly represent the state of occlusion, increasing interpretability and serving as input to the percept gate controller.

Percept gate Loci-v1's object tracking approach draws inspiration from Kalman filtering, which iteratively predicts object state changes and then adaptively fuses these predictions with current observations (Kalman, 1960). Accordingly, Loci-v1 predicts the next object states, decodes them into pixel space and then uses these predictions along with the current frame to produce new object states (see Figure 2; outer-loop). While the Kalman filter separates the steps of observation and information fusion, Loci-v1 observes and fuses jointly and implicitly during the encoding process. This is advantageous when fusing pixel-based information (e.g., combining hidden and visible object parts). However, when the model needs to fully maintain its own predictions because the current frame does not provide new information (e.g., during full occlusion), the encoding process via the outer-loop becomes disruptive. As an alternative, recent work from model-based reinforcement learning advocates the efficiency and precision of predicting directly in latent space (Hafner et al., 2019, 2020; Ha & Schmidhuber, 2018). Latent world models can be used to imagine how a scene will unfold while not being provided with new observations, which is the case during temporary occlusions or blackouts. Therefore, we introduce an inner processing loop in Loci-Looped, which enables the model to propagate internal imaginations over time in latent space (see Figure 2; inner-loop).

Similar to the Kalman filter, we equip the model with the ability to linearly interpolate between the current observations and the last predictions. Formally, the current object states G_k^t, P_k^t become a linear blend of the observed object states $\tilde{G}_k^t$, $\tilde{P}_k^t$ and the predicted object states $\hat{G}_k^t$, $\hat{P}_k^t$:

$$G_k^t = \alpha_k^{t,G}\tilde{G}_k^t + (1-\alpha_k^{t,G})\hat{G}_k^t \tag{3}$$

$$P_k^t = \alpha_k^{t,P}\tilde{P}_k^t + (1-\alpha_k^{t,P})\hat{P}_k^t \tag{4}$$

The weighting α is specific for each gestalt and position code in each slot k. Importantly, Loci-Looped learns to regulate the two percept gates on its own in a fully self-supervised manner. It learns an update function g_θ, which takes as input the observed state $\tilde{S}_k^t$, the predicted state $\hat{S}_k^t$, and the last positional encoding P_k^{t-1}:

$$(z_k^{t,G}, z_k^{t,P}) = g_\theta(\tilde{S}_k^t, \hat{S}_k^t, P_k^{t-1}) + \varepsilon, \tag{5}$$

where a state comprises the gestalt encoding, the positional encoding, and the occlusion state. By adding noise ε sampled from a Gaussian with a fixed standard deviation to the function g_θ, the gates tend to be either close or open, rather than remaining partially open (Gumbsch, Butz, & Martius, 2022). We model g_θ with a feed-forward neural network. To be able to fully rely on its own predictions, Loci-Looped needs to be able to fully close the gate by setting α exactly to zero. We therefore use a rectified hyperbolic tangent to compute α:

$$(\alpha_k^{t,G}, \alpha_k^{t,P}) = \max(0, \tanh((z_k^{t,G}, z_k^{t,P}))). \tag{6}$$

To encourage robust world models without the reliance on continuous external updates, we impose an L_0 loss on gate openings encouraging the sparse use of observations. The introduction of the percept gate enables Loci-Looped to control its perception flexibly fusing predictions with observations, essentially estimating their relative information values.

Table 1: Training objectives used by Loci-v1 and Loci-Looped.

Name	Type	Loci-v1	Loci-Looped
Next-Frame Prediction	Loss	✓	✓
Input-Frame Reconstruction	Loss	-	✓
Gestalt Change	Reg	✓	✓
Position Change	Reg	✓	✓
Perceptgate Openings	Reg	-	✓
Object Permanence	Reg	✓	-

Loss functions

A complete list of the training losses used is presented in Table 1. Compared to Loci-v1, we dispense the use of an object permanence loss, which explicitly facilitated the maintenance of object representations in case of occlusions. Instead, Loci-Looped learns the concept of object permanence autonomously. Furthermore, it is worth noting that the percept gates do not only control the forward information flow, but also the backward flow of gradients. When the percept gates are closed, the error signal is only backpropagated to the transition module but not to the encoder module, which could lead to its degeneration. To avoid this, we additionally compute a reconstruction loss in Loci-Looped, which is directly derived from the current observations.

Training

We adopt the training procedure of Loci-v1. Loci-Looped is trained in a fully unsupervised manner, except that the background of each scene is provided. The model is trained end-to-end, utilizing the rectified Adam optimizer (Liu et al., 2021) with a learning rate of 0.0001 in conjunction with truncated backpropagation through time. The model was trained for 1150k updates at a resolution of 480×320.

Experiments and results

In this section, we evaluate Loci-Looped in a VoE experiment, demonstrating that it learns (i) to reliably track objects through occlusion and (ii) the concept of object permanence by anticipating the reappearance of occluded objects. In addition, (iii) we provide ablation studies examining the role of the percept gate for learning object permanence.

Baselines We compare Loci-Looped against three compositional scene representation models. SAVi (Kipf et al., 2022) is state-of-the-art in self-supervised scene segmentation and also makes use of a slot-wise encoder-transition-decoder architecture. The base model Loci-v1 is state-of-the-art in the CATER benchmark (Girdhar & Ramanan, 2019). G-SWM (Z. Lin et al., 2020) learns an object-centric world model and is state-of-the-art in the task of video prediction.

Tracking objects through occlusions

Trainingset We train on the ADEPT (Smith et al., 2019) dataset. The training set contains 1000 synthetic videos displaying up to 7 solid objects traversing the scene with constant speed and direction. It shows physically plausible dynamics including partial and full object occlusions, while excluding any other object interactions (e.g., collisions).

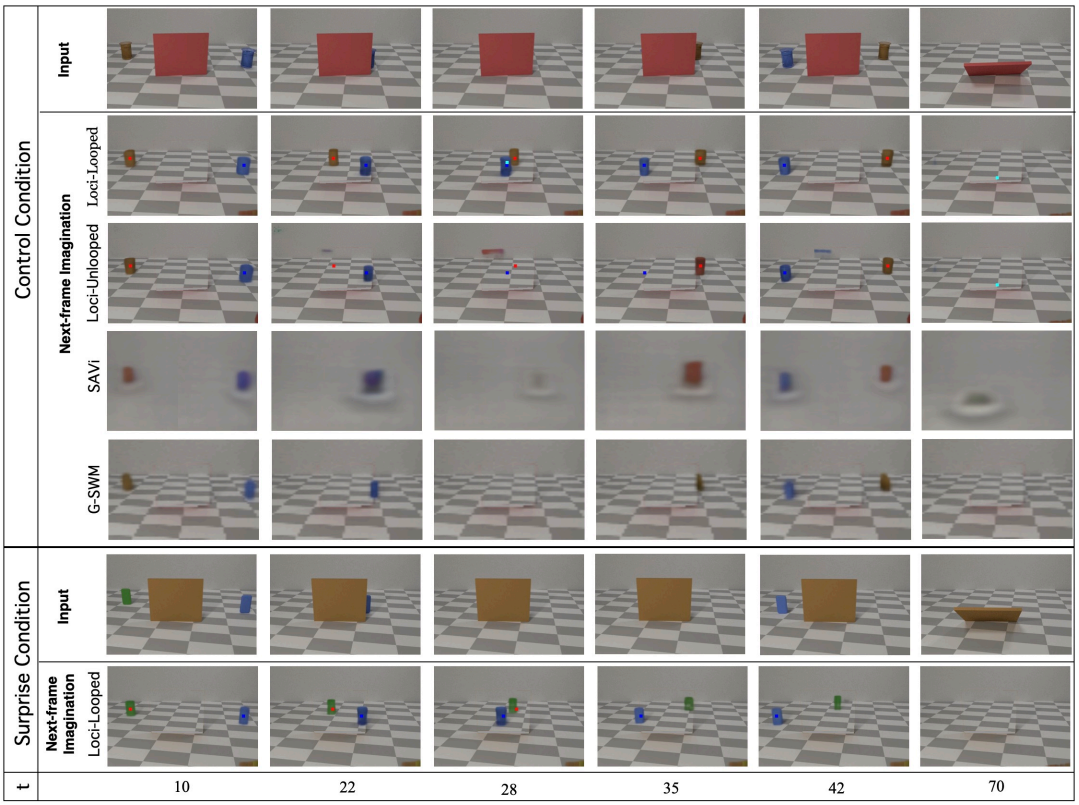

Figure 3: Loci-Looped maintains stable object percepts of the occluded objects. *Control Condition:* Two objects traverse the scene and both objects reappear. *Surprise Condition:* Two objects traverse the scene, the blue object reappears while the green object vanishes. *Next-frame Imagination:* The model's perception on how the scene unfolds behind the occluder, generated by applying layer summation without the occluder slot. The colored dots show the GT positions of the objects.

Table 2: Tracking results.

Model	Mean Tracking Error	Successful Trackings (%)	MOTA
Loci-Looped	**2.6** ± 2.7	**96.6**	**0.84**
Loci-Visibility	7.7 ± 10.6	43.6	0.64
Loci-Unlooped	12.4 ± 14.8	7.4	0.76
Loci-v1	12.5 ± 10.3	38.4	-1.34
G-SWM	26.8 ± 14.5	7.1	0.23
SAVi	26.7 ± 12.6	3.2	-0.67

Testset We use 35 videos of the ADEPT vanish scenario as test set. This scenario starts with a large screen placed in the center of the scene. Then one or two objects enter the scene from opposite directions, disappear behind the screen, traverse the area behind the screen while hidden, reappear on the other side of the screen, and finally exit the scene. The traversing objects are not visible for 10.3 frames on average which equals 25.0% of their total time being present.

Metric We evaluate the performance of the models with respect to two key capabilities. First, we quantify how well the models detect objects and identify them temporally consistently using Multiple Object Tracking Accuracy (MOTA) (Bernardin & Stiefelhagen, 2008). Second, we quantify the model's tracking error as the distance between estimated object positions and the true object positions. The estimated object positions can be extracted directly from Loci-Looped, which represents positional information explicitly. To extract object positions from the SAVi model, we first calculate object masks for each slot and then determine the center of them. Importantly, temporarily occluded objects are included in both metrics.

Results As shown in Figure 3, only Loci-Looped maintains stable object representations throughout the occlusion phase and precisely imagines the trajectory of the occluded objects. The average tracking error and the MOTA are listed in Table 2. Loci-Looped outperforms both baseline models by a large margin. At this point, allow us to emphasize that this precision is remarkable seeing that Loci-Looped was never informed about the location or existence of neither visible nor occluded objects. Importantly, 96.6% of slots that were recruited before the occlusion phase achieved a final tracking error (i.e., the tracking error in the moment the objects exit the scene) smaller than 10%, indicating that these slots tracked their assigned objects successfully throughout the entire scene.

Violation of expectation

Having seen that Loci-Looped tracks objects successfully through occlusion, we now test whether it has also learned to anticipate their reappearance.

Surprise scenario We focus on the ADEPT's vanish scenario that tests the concept of object permanence and directional inertia. The surprise condition (11 videos) features two objects that again traverse the scene behind the occluder this time, however, only one object reappears from behind the screen whereas the other vanishes while behind the screen. In the control condition both objects reappear. This scenario is designed to test the model's anticipation about the reappearance of the occluded object.

Slot error We compute an object- and thus slot-specific slot error as follows:

$$E_k^t = \frac{\sum_{i,j}\left[(I^{t+1} - \hat{R}^{t+1}) \odot \hat{M}_k^{t,v}\right]^2}{\sum_{i,j} \hat{M}_k^{t,v}}, \tag{7}$$

where the overall prediction error is simply masked by the visibility mask of slot k. In addition, we divide the error by the sum of the visibility mask values to make the error invariant to the size of the object. For the following analysis we only consider slots that represent non-occluder objects and that achieved a final tracking error smaller than 10%.

Results The model's response indicates a significantly greater prediction error when hidden objects fail to reappear showing a clear violation of expectation. Notably, this is the case for both time points: when the object should reappear after having slid past the occluder (around frame 20) and when the occluder falls over (around frame 60) after having not re-appeared before (cf., Figure 4a; $t(75) = 1.69, p = .047$; $t(75) = 3.68, p < .001$; as well as error peaks in Figure 4b around frames 30 and 65). As shown in Figure 3, Loci-Looped tends to park the object behind the occluder if it did not reappear until the occluder falls over. Note that this behavior is fully emergent, as Loci-Looped is never trained on objects that permanently disappear behind occluders, and shows the model's strong bias to maintain stable, consistent object representations.

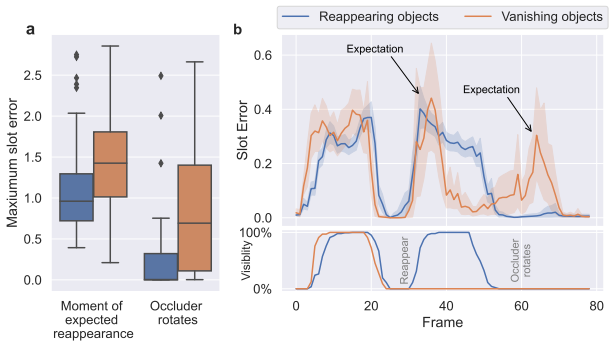

Figure 4: Results on the VoE experiment.

Percept gate facilitates object permanence learning

Ablation studies To further investigate the effect of the percept gate we train a version of Loci-Looped that can only make use of the outer-loop, labelling this variant Loci-Unlooped. As shown in Figure 3, we find that this model does not maintain stable object representations of occluded objects, suggesting that the inner-loop is crucial for this ability. In addition, we ablated the parameterized update function g_θ controlling the percept gate by switching to the inner-loop directly proportionally to the perceived occlusion state of each object (i.e. $\alpha_k^t = 1 - O_k^t$) at test time using the Loci-Looped model. Consequently, this model version utilizes the inner-loop when objects are occluded and the outer-loop when objects are visible. As reported in Table 2, Loci-Visibility performs worse than Loci-Looped indicating that the adaptive fusion mechanism improves tracking performance.

Influence of percept gate regularization Our evaluation shows that during object occlusions Loci-Looped learned to rely on its inner-loop (mean inner-loop integration: 99.2%), essentially switching to a latent imagination mode. When objects were visible, the model made only sparse usage of observations (mean inner-loop integration: 91.1%). This is due to the percept gate opening regularization, which imposes the inductive bias to predict the visible world while only glimpsing at it. As a result, the model trains itself on simulated occlusions besides the encountered ones during training generating further error signals that encourage the learning of more accurate latent temporally predictive models. This may have enabled the model to easily generalise to extended occlusion scenarios. To test this hypothesis, we ran another experiment varying the strength of the percept gate regularization (β_{Gate}). As illustrated in Figure 5, we indeed find that the development of object permanence improves with higher percept gate regularization suggesting that the model's inductive bias to anticipate the next latent perception and hereby build a robust world model is key for learning object permanence. However, this comes at the price of less accurate object perception, as less information from the actual observation is integrated, leading to higher reconstruction errors.

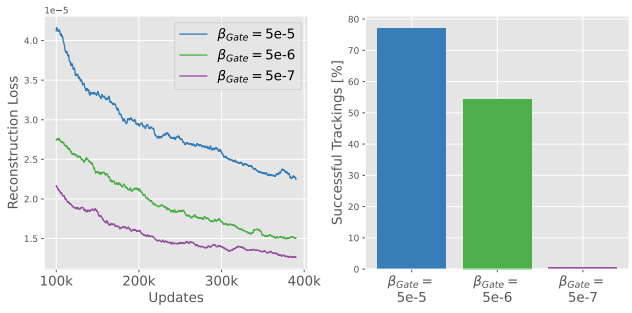

Figure 5: Reconstruction Loss and Successful Tracking rate for different percept gate regularization β_{Gate} strengths.

Discussion

We have addressed the question how a temporally predictive, autoregressive machine learning system may learn about inertia and object permanence. Our fully unsupervised learning system Loci-Looped can learn to track objects through occlusions, showing surprise when objects do not reappear where expected—just like the data from infant studies suggest.

Learning success depended on numerous design choices. We relied on a temporally predictive model that encourages visual segmentation into compressed, slot-based object encodings (Piloto et al., 2022; Traub et al., 2022). In contrast to other slot-oriented processing architectures (Z. Wu, Dvornik, Greff, Kipf, & Garg, 2023; Locatello et al., 2020; Yuan, Chen, Li, & Xue, 2023; Weihs et al., 2022), our Loci-Looped model processes visual information in a slot-oriented manner starting directly from the pixel level. It integrates prediction error information, thus following the predictive coding principle (Rao & Ballard, 1999; Friston, 2009). Furthermore, as indicated by our ablation studies and prior work with Loci-v1 (Traub et al., 2022), modeling success relied on further inductive learning and information processing biases. First, the slots had to be informed bottom-up by the slot-respective mask predictions about both the currently visible object parts (relevant during occlusions) and the fully visible object. Second, the system required an adaptive slot-specific object fusion mechanism, which we termed a percept gate. This mechanism enables Loci-Looped to fully rely on its internal predictions when external information is not available. Third, the mechanism had to be encouraged to prefer relying on its internal predictions by penalizing the integration of external information. Fourth, the experiences themselves had to include temporary occlusions with varying duration.

Currently, Loci-Looped relies on a static camera pose as well as on a provided background image. Concurrent research work (Traub, Becker, Sauter, Otte, & Butz, 2023) has shown, though, that Loci-v1 is extendable to real-world datasets and moving cameras. We are currently merging these system abilities to enable Loci-Looped's percept gate evaluation in more diverse scenarios. Furthermore, we probe Loci-Looped's ability to learn about object continuity and object solidity and enable it to generate probabilistic, generative predictions. Another current limitation is that Loci-Looped—as all other slot-based processing systems for that matter—has equal processing resources for each slot. We are currently improving encoding efficiency by enabling the system to selectively probe the visual information in a task-oriented, selective manner. Our aim is to develop a resource-rational system that learns to distribute its visual processing resources optimally given current task and context (Bhui, Lai, & Gershman, 2021; Butz, 2022; Heald, Lengyel, & Wolpert, 2023; Lieder & Griffiths, 2020; Schwöbel, Marković, Smolka, & Kiebel, 2021). Overall, we hope that the presented model contributes to advance both the development of more human-like, interpretable artificial intelligence and our computational understanding of human cognitive development.

Acknowledgements

This work received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645 as well as from the Cyber Valley in Tübingen, CyVy-RF-2020-15. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Manuel Traub and Frederic Becker, and the Alexander von Humboldt Foundation for supporting Martin Butz and Sebastian Otte.

References

Aguiar, A., & Baillargeon, R. (1996). 2.5-Month-old reasoning about occlusion events. *Infant Behavior and Development*, *19*, 293.

Baillargeon, R., & DeVos, J. (1991, December). Object permanence in young infants: further evidence. *Child Development*, *62*(6), 1227–1246. (https://doi.org/10.1111/j.1467-8624.1991.tb01602.x)

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985, January). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208. (10.1016/0010-0277(85)90008-3)

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, *2008*, 1–10. (10.1155/2008/246309)

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15-21. (Value based decision-making) doi: 10.1016/j.cobeha.2021.02.015

Butz, M. V. (2008). How and why the brain lays the foundations for a conscious self. *Constructivist Foundations*, *4*(1), 1-42.

Butz, M. V. (2021, March). Towards Strong AI. *KI - Künstliche Intelligenz*, *35*(1), 91–101. (10.1007/s13218-021-00705-x)

Butz, M. V. (2022). Resourceful event-predictive inference: The nature of cognitive effort. *Frontiers in Psychology*, *13*. doi: 10.3389/fpsyg.2022.867328

Butz, M. V., Achimova, A., Bilkey, D., & Knott, A. (2021). Event-Predictive Cognition: A Root for Conceptual Human Thought. *Topics in Cognitive Science*, *13*(1), 10–24. (10.1111/tops.12522)

Clark, A. (2013, June). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. (10.1017/S0140525X12000477)

Den Ouden, H., Kok, P., & De Lange, F. (2012). How Prediction Errors Shape Perception, Attention, and Motivation. *Frontiers in Psychology*, *3*.

Friston, K. (2009, July). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. (10.1016/j.tics.2009.04.005)

Girdhar, R., & Ramanan, D. (2019). CATER: A diagnostic dataset for compositional actions and temporal reasoning. *CoRR*, *abs/1910.04744*.

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., ... Lerchner, A. (2020, July). *Multi-Object Representation Learning with Iterative Variational Inference.* arXiv. (10.48550/arXiv.1903.00450)

Gumbsch, C., Butz, M. V., & Martius, G. (2022, January). *Sparsely Changing Latent States for Prediction and Planning in Partially Observable Domains.* arXiv. (10.48550/arXiv.2110.15949)

Ha, D., & Schmidhuber, J. (2018, March). World Models. *arXiv:1803.10122 [cs, stat]*. (10.5281/zenodo.1207631)

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2020, March). *Dream to Control: Learning Behaviors by Latent Imagination.* arXiv. (10.48550/arXiv.1912.01603)

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019, May). Learning Latent Dynamics for Planning from Pixels. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2555–2565). PMLR. (ISSN: 2640-3498)

Heald, J. B., Lengyel, M., & Wolpert, D. M. (2023). Contextual inference in learning and memory. *Trends in Cognitive Sciences*, *27*(1), 43-64. doi: https://doi.org/10.1016/j.tics.2022.10.004

Kalman, R. E. (1960, March). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. (10.1115/1.3662552)

Kipf, T., Elsayed, G. F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., ... Greff, K. (2022, March). *Conditional Object-Centric Learning from Video.* arXiv. doi: 10.48550/arXiv.2111.12594

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016, November). *Building Machines That Learn and Think Like People.* arXiv. (10.48550/arXiv.1604.00289)

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. doi: 10.1017/S0140525X1900061X

Lin, Y., Li, J., Gertner, Y., Ng, W., Fisher, C. L., & Baillargeon, R. (2021, March). How do the object-file and physical-reasoning systems interact? Evidence from priming effects with object arrays or novel labels. *Cognitive Psychology*, *125*, 101368. doi: 10.1016/j.cogpsych.2020.101368

Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. *Cambridge handbook of cognitive development*, 168–194.

Lin, Z., Wu, Y.-F., Peri, S., Fu, B., Jiang, J., & Ahn, S. (2020, 13–18 Jul). Improving generative imagination in object-centric world models. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 6140–6149). PMLR.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J.

(2021, October). *On the Variance of the Adaptive Learning Rate and Beyond.* arXiv. (10.48550/arXiv.1908.03265)

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., ... Kipf, T. (2020). Object-Centric Learning with Slot Attention. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 11525–11538).

Lotter, W., Kreiman, G., & Cox, D. (2017). *Deep predictive coding networks for video prediction and unsupervised learning.* doi: doi.org/10.48550/arXiv.1605.08104

Munakata, Y., Mcclelland, J., Johnson, M., & Siegler, R. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*(4), 686-713. doi: 10.1037/0033-295x.104.4.686

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022, September). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267. (10.1038/s41562-022-01394-8)

Rao, R. P. N., & Ballard, D. H. (1999, January). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. (10.1038/4580)

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2022, September). IntPhys 2019: A Benchmark for Visual Intuitive Physics Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5016–5025. (10.1109/TPAMI.2021.3083839)

Schwöbel, S., Marković, D., Smolka, M. N., & Kiebel, S. J. (2021). Balancing control: A bayesian interpretation of habitual and goal-directed behavior. *Journal of Mathematical Psychology*, *100*, 102472. doi: 10.1016/j.jmp.2020.102472

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J. B., & Ullman, T. D. (2019, January). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Neural Information Processing Systems (NIPS)*.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992, October). Origins of knowledge. *Psychological Review*, *99*(4), 605–632. doi: 10.1037/0033-295x.99.4.605

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. (10.1111/j.1467-7687.2007.00569.x)

Summerfield, C., & Egner, T. (2009, September). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409.

Traub, M., Becker, F., Sauter, A., Otte, S., & Butz, M. V. (2023, October). *Loci-Segmented: Improving Scene Segmentation Learning.* arXiv. doi: 10.48550/arXiv.2310.10410

Traub, M., Otte, S., Menge, T., Karlbauer, M., Thümmel, J., & Butz, M. V. (2022, October). *Learning What and Where – Unsupervised Disentangling Location and Identity Tracking.* arXiv. (10.48550/arXiv.2205.13349)

Traub, M., Otte, S., Menge, T., Karlbauer, M., Thümmel, J., & Butz, M. V. (2023, February). *Learning What and Where: Disentangling Location and Identity Tracking Without Supervision.* arXiv. (10.48550/arXiv.2205.13349)

Weihs, L., Yuile, A., Baillargeon, R., Fisher, C., Marcus, G., Mottaghi, R., & Kembhavi, A. (2022). Benchmarking progress to infant-level physical reasoning in AI. *Transactions on Machine Learning Research*. Retrieved from https://openreview.net/forum?id=9NjqD9i48M

Wu, P., Escontrela, A., Hafner, D., Abbeel, P., & Goldberg, K. (2023, 14–18 Dec). Daydreamer: World models for physical robot learning. In K. Liu, D. Kulic, & J. Ichnowski (Eds.), *Proceedings of the 6th conference on robot learning* (Vol. 205, pp. 2226–2240). PMLR.

Wu, Z., Dvornik, N., Greff, K., Kipf, T., & Garg, A. (2023). Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *The eleventh international conference on learning representations.* (10.48550/arXiv.2210.05861)

Yuan, J., Chen, T., Li, B., & Xue, X. (2023, February). *Compositional Scene Representation Learning via Reconstruction: A Survey.* arXiv. (10.48550/arXiv.2202.07135)