Probing neural representations of scene perception in a hippocampally dependent task using artificial neural networks

Markus Frey[1,2] Christian F. Doeller[1,2] Caswell Barry[3]

[1]Kavli Institute for Systems Neuroscience, NTNU, Norway [2]Max-Planck-Insitute for Human Cognitive and Brain Sciences, Germany
[3]Cell & Developmental Biology, UCL, United Kingdom

Abstract

Deep artificial neural networks (DNNs) trained through backpropagation provide effective models of the mammalian visual system, accurately capturing the hierarchy of neural responses through primary visual cortex to inferior temporal cortex (IT) [41, 43]. However, the ability of these networks to explain representations in higher cortical areas is relatively lacking and considerably less well researched. For example, DNNs have been less successful as a model of the egocentric to allocentric transformation embodied by circuits in retrosplenial and posterior parietal cortex. We describe a novel scene perception benchmark inspired by a hippocampal dependent task, designed to probe the ability of DNNs to transform scenes viewed from different egocentric perspectives. Using a network architecture inspired by the connectivity between temporal lobe structures and the hippocampus, we demonstrate that DNNs trained using a triplet loss can learn this task. Moreover, by enforcing a factorized latent space, we can split information propagation into "what" and "where" pathways, which we use to reconstruct the input. This allows us to beat the state-of-the-art for unsupervised object segmentation on the CATER and MOVi-A,B,C benchmarks.

1. Introduction

Recently, it has been shown that neural networks trained with large datasets can produce coherent scene understanding and are capable of synthesizing novel views [12, 21]. These models are trained on egocentric (self-centred) sensory input and can construct allocentric (world-centred) responses. In animals, this transformation is governed by structures along the hierarchy from the visual cortex to the hippocampal formation, an important model system related to navigation and memory [20, 30]. Notably, the hippocampus is a necessary component of the network supporting memory and perception of places and events and is one of the first brain regions compromised during the progression of Alzheimer's disease (AD) [3, 34]. However, experimental knowledge regarding the interplay across multiple interacting brain regions is limited and new computational models are needed to better explain the single-cell responses across the whole transformation circuit.

Here, we developed a scene recognition model to better understand the intrinsic computations governing the transformation from egocentric to allocentric reference frames, which controls successful view synthesis in humans and other animals. For this, we developed a novel hippocampally dependent task, inspired by the 4-Mountains-Test [18], which is used in clinics to predict early-onset Alzheimer's disease [40]. We tested this task by creating a biologically realistic model inspired by recent work in scene perception, in which scenes need to be re-imagined from several different viewpoints.

The main contributions of our paper are the following:

- We introduce and open-source the allocentric scene perception (ASP) benchmark for training view synthesis models based on a hippocampally dependent task which is frequently used to predict AD.
- We show that a biologically realistic neural network model trained using a triplet loss can accurately distinguish between hundreds of scenes across many different viewpoints and that it can disentangle object information from location information when using a factorized latent space.
- Lastly, we show that by using a reconstruction loss combined with a pixel-wise decoder we can perform unsupervised object segmentation, outperforming the state-of-the-art models on the CATER, MOVi-A,B,C benchmarks.

2. Related work

2.1. Related work in neuroscience

The study of scene perception in neuroscience was first systematically explored in behavioural experiments in

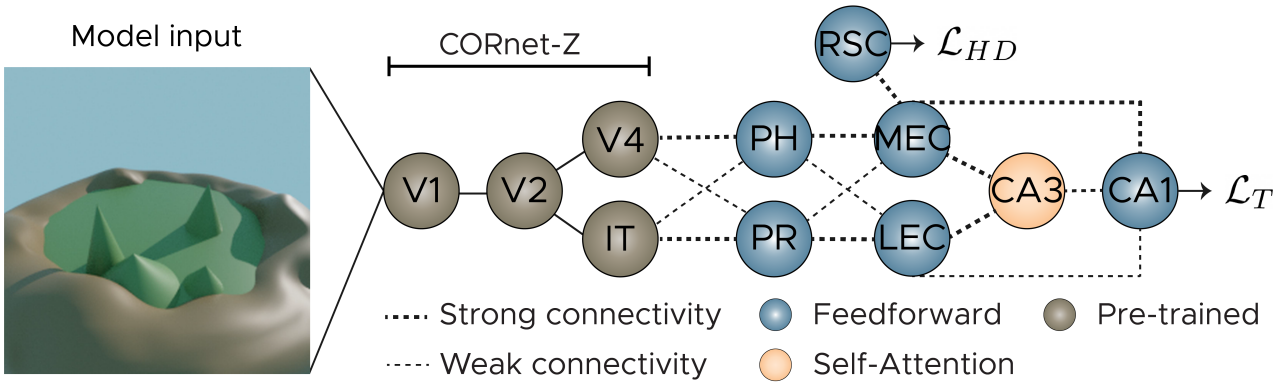

Figure 1. **Task design and model architecture**
We adapt the 4-Mountains-Test [18] using a simplified task design with one to four objects with circular symmetry and a distal landmark, depicting a mountain range. We then design a biologically inspired model architecture, for which we take visual cortex responses from pre-trained CORnet-Z [24] and feed them through perirhinal (PR) and parahippocampal (PH) cortices. The retrosplenial cortex uses an auxiliary loss to decode head direction. Medial entorhinal (MEC) and lateral entorhinal cortex (LEC) use a disentangled latent space to separate object from location information. Both are integrated within the hippocampus, with CA3 using a self-attention layer across both time and space.

1972, by showing for the first time scenes in real-world contexts to human participants [2]. These outdoor scenes were either shown intact or scrambled into six randomly arranged pieces. It was shown, that the correct identification of cued objects was drastically lower when the scenes were scrambled, indicating that the various parts of a scene are perceived as a whole.

Neural recordings as a response to scenes were first reported in fMRI experiments that showed that the parahippocampal place area (PPA) is involved in perceiving the visual environment, being more active for outdoor scenes than single objects [9]. Moreover, activity is reduced when scrambling the picture into random pieces, indicating that neurons in PPA are sensitive to the structure of the entire scene [10]. A second scene-selective region was shown to be also active during mental imagination of scenes, subsequently labelled as the retrosplenial complex (RSC). The RSC also acts as a hub to integrate sensory, motor and visual information and is crucial for the transformation from egocentric to allocentric reference frames [8, 29].

Computational modelling has suggested that the transformation in RSC is governed by head direction cells [5], which encode the world-centred facing direction of the head of the animal in the azimuthal plane [32]. When egocentric information from the parietal cortex reaches RSC, the head direction signal aligns these cells to a common reference, creating complementary allocentric cell types. These include object vector cells in the medial entorhinal cortex [19], which are active at spatially confined objects and boundary vector cells (BVCs) in the subiculum [25] which encode boundary information in their neuronal activity. The combined activity of several BVCs giving rise to allocentric place cells in the hippocampus [1, 5].

In recent years, normative machine-learning approaches have investigated visual information processing in an egocentric reference frame while translating it into an allocentric representation for downstream task performance. Several types of these models have been explored, most notably the Tolman-Eichenbaum-Machine (TEM) [38] and the Spatial Memory Pipeline (SMP) [37]. These normative models have shown the emergence of allocentric spatial cells in two separate tasks, using a similar objective function: predicting the next sensory observation. However, both of these models have not been tested on their abilities to generate images from novel viewpoints, with TEM only taking in abstract sensory observations and SMP being employed in a reinforcement-learning (RL) environment using a spatial navigation task.

2.2. Related work in computer science

Beyond TEM and SMP, specially designed scene perception models have tackled the problem of novel view synthesis in several distinct ways. For example, in robotics, SLAM (Simultaneous localization and mapping) algorithms have been used extensively to represent scenes and navigate within them, but mainly in a supervised setting on partially occluded scenes and with a focus on the navigational abilities of robots, which are endowed with additional sensors [6].

Neural radiance field (NERF) networks [27, 28, 35] try to mimic the image synthesis based on real-world physics, namely the way light is reflected from certain materials and

how these rays end up in the camera sensor. For this, the neural network must infer all the scene's physical properties. The actual reconstruction of an image is done analytically by a traditional rendering engine, which is typically not changed during training [36]. The advantage of these rendering methods lies in the fact that arbitrary resolutions can be achieved. This is possible as it is not using image or voxel space, but a continuous neural representation, where coordinates are mapped through a neural network to their corresponding value - representing colour, occupancy or material properties [27, 31]. However, a common shortfall of neural rendering methods is the use of separate neural networks for each scene, making it hard to gauge the generalization abilities beyond its training scenes.

Traditional scene decomposition methods beyond NERFs are able to generalize across many scenes, while still relying on pixel-level information. Most commonly, the model takes several snapshots of a scene from different viewpoints which are joined in a latent space. To produce new views, the latent space is conditioned with a new set of camera coordinates, reconstructing a novel viewpoint [12]. Recent models have focused on architectures using a slot structure which disentangles objects within a scene from each other using self-attention [16, 21, 26]. This allows the model to learn a fully unsupervised factorized latent space, which is used to synthesize novel viewpoints and scenes with different compositionality [21].

3. Task design

None of the above-mentioned biologically inspired models is equipped to deal with randomly sampled egocentric sensory observations. Therefore we built a model inspired by novel view synthesis tasks while considering the particularities of the hippocampally dependent 4-Mountains-Test [7, 18]. This test is used in the clinic and requires allocentric topographical processing for successful task performance, thus is sensitive to hippocampal damage including that which accumulates during the early stages of AD. The participant first views an image of a scene with four mountains in it and after a two-second delay is asked to match the same-scene image out of four images (three distractors and one target image, see Figure S1 for the same test using our adapted design). Most importantly, the correct image is the same scene, seen from a different viewpoint. In contrast, the distractor images show scenes where the objects are located in different allocentric configurations compared to the original scene.

We simplified the task into its core components by rendering four objects with circular symmetry and a global reference frame given by the surrounding landscape. Novel scenes were rendered by randomly changing the world-centred location of each object, thereby changing the relative distances of the objects to themselves and the boundary. We acquire egocentric sensory observations by rendering the scene across different viewpoints, varying both azimuth and elevation (Figure 2). We also render views from the 'inside' of the environment, simulating the views of an agent navigating the scene. We render six different versions of this task, by varying both the distal landmark $(with|without)$ and the colours of the objects $(mix|green|white)$, while randomly varying the number of objects within the scene from 1 to 4, with 10% of the scenes having one object, 20% two, 30% three, and 40% containing four objects. We sampled different colours, as traditional scene perception models struggle to differentiate objects with the same colour or the floor's colour. During rendering, we also store segmentation masks and depth profiles for each rendered viewpoint (Figure 2 & Figure S2). Moreover, we store positional information for analysis of neural activity with respect to different reference frames. This includes the positions of the objects in an egocentric reference frame (position 'on the screen'), angular information such as azimuth and elevation and the allocentric positions of each object within the environment (Figure 4).

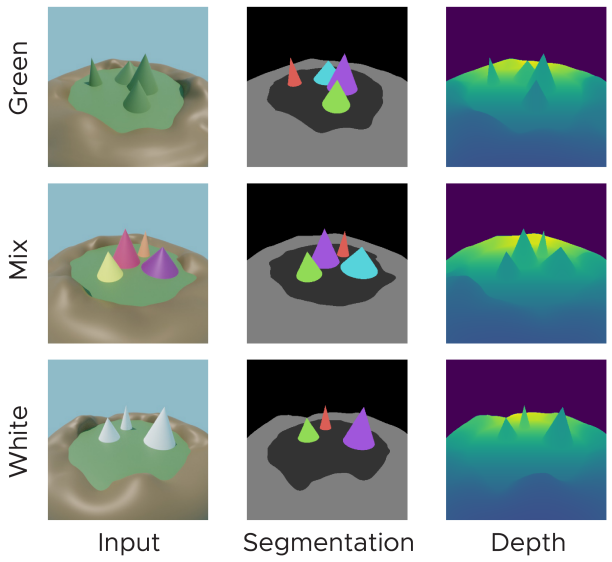

Figure 2. **Variations of task design**
We render the task in six different versions (three shown here), varying both the number of objects within the scenes from 1 to 4 as well as object colour. We provide segmentation masks, including objects, backgrounds, and depth maps. We also render each task without a distal landmark; see Figure S2.

4. Model architecture

The overall architecture of the model follows known biological connectivity between visual cortices and the hippocampal formation [3] (Figure 1). To simulate the responses of the visual cortex, we use a pre-trained convo-

lutional neural network which accurately predicts neural responses from the macaque visual system [24, 33] (CORNet-Z). This ensures that visual cortex responses do not overfit the objects, colours and backgrounds we use but generalize onto natural images. We extract activations from visual cortex area 4 (V4) and inferior temporal (IT) cortex for each rendered viewpoint.

The information from V4 and IT is then routed through perirhinal (PR) and parahippocampal (PH) cortices using either weak or strong connectivity. We implement the strong connections using a feedforward layer with non-linearity between two areas, and the weak connections by adding a residual connection consisting of a linear feedforward layer that uses the summed input of all previous activations. Both areas receive information mainly from the ventral visual processing stream, with PR being crucial for the representation of objects ('what'), while PH predominantly processes visuospatial information [3]. We implement this split of information by averaging the visual cortex representation either across the temporal or spatial axis, which also enforces disentangled latent representations in subsequent layers, namely the medial (MEC) and lateral (LEC) entorhinal cortex. MEC receives input which is averaged across time, therefore being trained to keep spatial information, while LEC receives input averaged across the spatial dimension, thereby retaining temporal information. We implement the hippocampus consisting of CA3 and CA1, with the former integrating information using a self-attention layer across different snapshots of the same scene and the latter integrating both temporal and spatial information using the outer product, which is fed through a simple feedforward layer, similar to how TEM integrates information in the hippocampal formation [39].

To reconstruct the input across novel views, we use a pixel-wise decoder similar to the one used in SIMONe [21]. For the feedback connections, we use the CA1 layer which is split into temporal (LEC) and spatial (MEC) information exactly as the forward pass splits the visual information at the level of the PR/PH layers. We then sample from LEC and MEC for each object and pixel and combine them with periodic positional encodings. These are fed through a 5-layer MLP decoder using 128 units each, decoding RGB values for each pixel and object which are combined with alpha masks to produce the full reconstructed image (Figure S3).

5. Model optimization

During model training, the model receives egocentric sensory observations which are transformed into allocentric representations in the hippocampal formation. The model is trained by minimizing either a triplet loss on this hippocampal latent space or an L2-reconstruction loss in pixel space

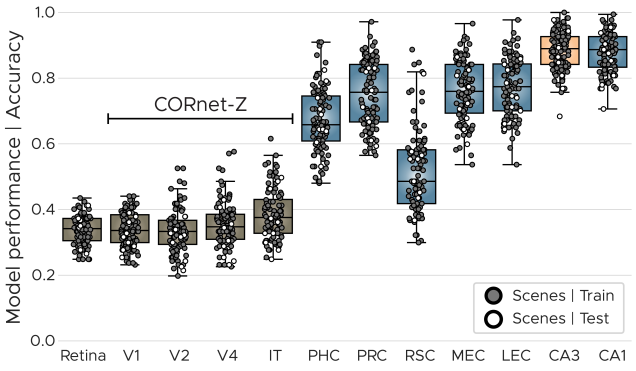

Figure 3. **Model performance on scene perception task.** Model performance across layers. For each layer, we calculate the accuracy by sampling triplets and quantifying the number of correctly classified same-scene images. Scenes used for training are depicted as grey circles and test scenes as white circles. Most world-centred responses are measured in late layers, indicating that hippocampal responses are similar for images from the same scene from different viewpoints.

if the model is tasked to reconstruct the image:

$$\mathcal{L}_{a,p,n} = \sum_i \max\left(\sqrt{(x_i^a - x_i^p)^2} - \sqrt{(x_i^a - x_i^n)^2} + \alpha, 0\right) \tag{1}$$

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_n (x_i - \hat{x}_i)^2 \tag{2}$$

where $x^{a,p,n}$ denotes the anchor, positive or negative representation from a given layer. If the image layer is used, these correspond to the original image in pixel space. The L2-reconstruction loss is calculated between the predicted reconstruction $\hat{x}_i$ and the original image x_i on the full-resolution image. The final loss is summed across the width, height and timesteps and averaged across batches.

The predictive objective function used in previous models [37, 38] is similar to the reconstruction loss in our model as the latent space enforces the reconstruction of scenes from different viewpoints, which can be understood as a prediction of not only the next observation but all possible observations. Note that many recent models capable of novel view synthesis use variational inference [4, 21], for which it is unclear how individual distribution statistics would be sampled in biological tissue. We, therefore, do not sample from the distribution and only use the mean, similar to how TEM constructs its latent space [38].

6. Results

6.1. Performance on adapted 4-Mountains-Test

We first train our scene perception model to separate between different scenes, closely linked to the original task in

the 4-Mountains-Test (Figure S1). We use a triplet loss in which we sample an anchor and a positive image from the same scene, together with a negative image from a different scene. This contrastive loss function allows us to disentangle the hippocampal representation between different scenes maximally. Model performance is evaluated by randomly sampling triplets (anchor, positive, negative) from the whole dataset, calculating a cross-correlation matrix and using the smallest off-diagonal value as the correct image. If a layer is able to distinguish between scenes, the correlation between the same-scene pair seen from different viewpoints should be high, while the correlation between different-scene pairs should be low.

This performance measure allows us to evaluate accuracy across all model layers. We observe that the pre-trained layers and the image itself show performance levels around chance, with IT performing best (Chance: 33%, IT: 39%, Figure 3). This means that even though IT is thought to contain object-specific information, it lacks crucial information to differentiate between scenes containing the same objects but in different allocentric positions. Nevertheless, these scenes can be distinguished by late layers in our model, with CA3 and CA1 showing performance close to 90% (CA3 89%, CA1 88%), indicating that these layers construct a world-centred representation of the environment, which is needed for separating the scenes in the adapted 4-Mountains-Test.

6.2. Neural representations within network layers

To quantify the amount of allocentric information contained in each trained layer, we calculate an allocentricity measure. We define allocentricity as the coefficient of variation of the activation of each artificial neuron across several images of the same scene. A neuron that fires similarly across images of the same scene from a different viewpoint has a high allocentric score, while a neuron with a high variance in its activations for the same stimuli has a low allocentric score. We observe that the hippocampal layers show the highest allocentricity score (PH, -5.5$\pm$0.03; PR, -2.6$\pm$0.05; CA3, 5.0$\pm$2.2; CA1, 3.5 $\pm$ 1.7), indicating that these layers have learned to be active for the same scene across different viewpoints, i.e. have formed an allocentric representation of the environment (Figure S4).

Having established a differentiation across layers between egocentric and allocentric information, we sought to investigate which scene properties are represented in each layer. We use a linear readout to investigate the separation of scene properties into low-level features like colours or the position on the screen, mid-level elements like object size and high-level features like allocentric position and scene identity [11]. We observe a trend toward later layers incorporating more high-level information. However, the overall structure is less clear than previously reported in the visual cortex [17,22,42]. Information regarding the position of objects in allocentric coordinates can only be effectively read out from hippocampal layer CA3, while egocentric information - object's position in screen coordinates - is reduced drastically in the layers beyond MEC, with CA1 seemingly only retaining allocentric information (Figure S4).

We next sought to investigate individual neural responses to obtain a fine-grained understanding of the computations being performed within each layer and, importantly, the ways in which they relate to known biological data. For this purpose, we visualize the activity of a subset of neurons within each layer with regard to different reference frames (Figure 4). We use the egocentric reference frame for pixel coordinates and the allocentric reference frame for objects or locations in the environment. To visualize angular allocentric responses, we visualize the activity using polar coordinates. We observe allocentric boundary and place-like activity in the CA1 layers as a function of the position of an object (Figure 4). This indicates that single neurons within the model layer are highly activated whenever an object is close to the boundary or occupies a certain allocentric position within the environment, similar to the boundary and place-like activity observed in biological organisms [25, 30]. These spatial responses also show similar mechanisms to biological cells, as they tend to remap across different scenes (S5. We observe these responses as a result when using only the triplet loss on the CA1 layer. As we describe further below, we can reconstruct and segment the image into distinct objects using an additional reconstruction loss.

6.3. Reconstructing the input through feedback connections

Having established that the model is able to discern between novel scenes from arbitrary viewpoints, we next explored its ability to reconstruct the scene across these viewpoints. For this, we added an additional reconstruction loss in pixel space (Figure S3). Reconstructing the input image is more challenging than just differentiating between scenes, as complete scene information has to be retained across layers or reinstated from a latent representation. Therefore, we used a factorized latent space to sample objects and frame information for each individual pixel and time point, similar to recent scene perception models [21]. It is assumed that mental imagination (and similarly novel view synthesis) is guided by a viewpoint-changing signal likely provided by grid cells in the medial entorhinal cortex, which is combined with object information in the lateral entorhinal cortex and is then further disentangled into egocentric information inside the retrosplenial cortex, which together with visual cortex establishes the mental image [1].

We first test the reconstruction loss across our six variations of the task, in which we varied object colour and back-

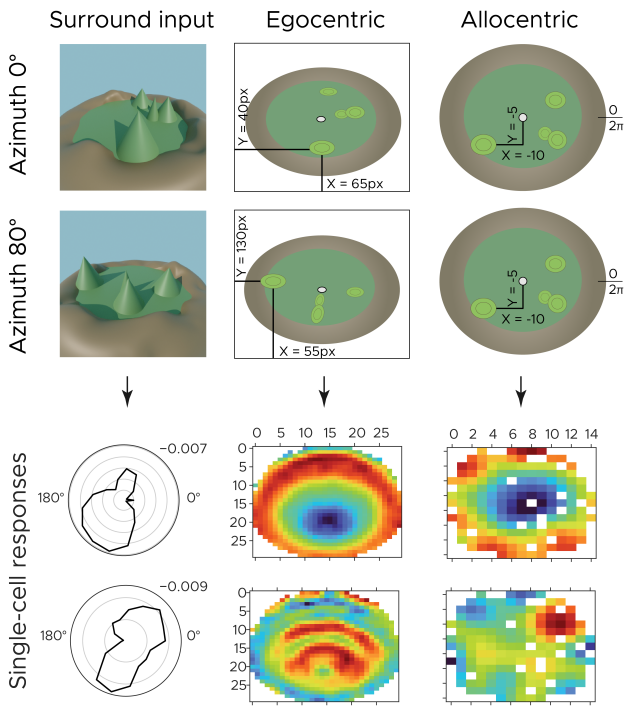

Figure 4. **Single-cell representations across different reference frames.** (Top) Illustration of egocentric and allocentric schemas. The left side shows two snapshots from the same scene using different viewpoints. The panels to the right depict the egocentric and allocentric schema for the respective snapshot. Black lines indicate screen coordinates for the same object in the egocentric view and world coordinates in the allocentric view. Note that by definition the allocentric, world-centred schema is the same for both snapshots. (Bottom) Example representations from neurons within the network, across each reference frame, showing activity of two neurons for azimuth, egocentric object position and allocentric object position.

ground information (Figure S2). We observe a difference in the segmentation performance of the model depending on the object colours used, with the green and white colours performing worse than the mixed objects, likely because it is harder to differentiate between objects of the same colour (Table 1). This has also been noticed in traditional scene decomposition models, which struggle to disentangle objects of the same colour and objects having a similar colour to the background [12]. Interestingly, if we do not enforce the disentanglement in CA1 via the triplet loss, we can still see a differentiation between scenes in the late layers of the model by just training on the reconstruction loss.

6.4. Model performance on CATER and MOVi

Lastly, we evaluated our neural network architecture on the Compositional Actions and TEmporal Reasoning (CATER) benchmark [13] as well as the MOVi-A,B,C

		Distal landmark	No landmark
MSE	Green	**27.912 ± 1.289**	46.411 ± 2.777
	Mix	45.908 ± 4.405	60.159 ± 1.923
	White	52.713 ± 4.119	**19.167 ± 0.575**
FG-ARI	Green	0.137 ± 0.122	0.046 ± 0.010
	Mix	**0.207 ± 0.148**	**0.297 ± 0.119**
	White	0.086 ± 0.035	0.140 ± 0.023
ARI	Green	0.122 ± 0.063	**0.067 ± 0.066**
	Mix	**0.383 ± 0.122**	0.016 ± 0.016
	White	0.136 ± 0.089	0.039 ± 0.033

Table 1. **Reconstruction and unsupervised segmentation performance on ASP**
We compare the mean-squared error for reconstructing the input images (MSE), the foreground adjusted rand index (FG-ARI) and the full adjusted rand index (ARI) across all six variations of our dataset. We use random seeds to report the mean and standard deviation across five model runs. The lowest MSE values and highest ARI values are displayed in bold.

	MONet	SIMONe	SAVi	Ours
CATER	0.412 ± 0.012	0.918 ± 0.036	0.928 ± 0.008	**0.939 ± 0.013**
MOVi-A	-	0.618 ± 0.200	0.820 ± 0.030	**0.790 ± 0.017**
MOVi-B	-	0.307 ± 0.330	0.615 ± 0.030	**0.460 ± 0.033**
MOVi-C	-	0.198 ± 0.005	0.470 ± 0.030	**0.318 ± 0.009**

Table 2. **Segmentation performance across unsupervised models on CATER and MOVi.**
We compare the foreground adjusted rand index (FG-ARI) across different unsupervised models, quantifying how well the model segmentation matches the real masks. Note that MONet is a static-frame model which predicts segmentation for each frame separately, likely causing the model to fail to track objects stably. SAVi is not a fully unsupervised model, using optical flow as a supervision signal. Baseline scores for MONet, S-IODINE and SIMONe were taken from [21], SAVi score was taken from [23]. To obtain the final model performance, we first train a model for 200000 steps with a fixed learning rate and subsequently present the efficacy of five models initialized with these weights and trained utilizing an annealing learning rate.

datasets [14]. These datasets consist of three to eleven randomly placed objects, with a fixed camera location but moving objects. We use the modified CATER dataset from [21], which includes segmentation masks for each object in order to explore the model's ability to perform object segmentation fully unsupervised. We use the same model architecture for higher-level cortices as described above, but replace the pre-trained visual representation with four convolutional layers, using a kernel size of four and a stride of 2. We additionally increase the number of units in the pixel-wise decoder from 128 to 512 and train the model for 300000 steps using a batch size of 1. As shown in Figure 5, our model is able to reconstruct the input frames and segment the ob-

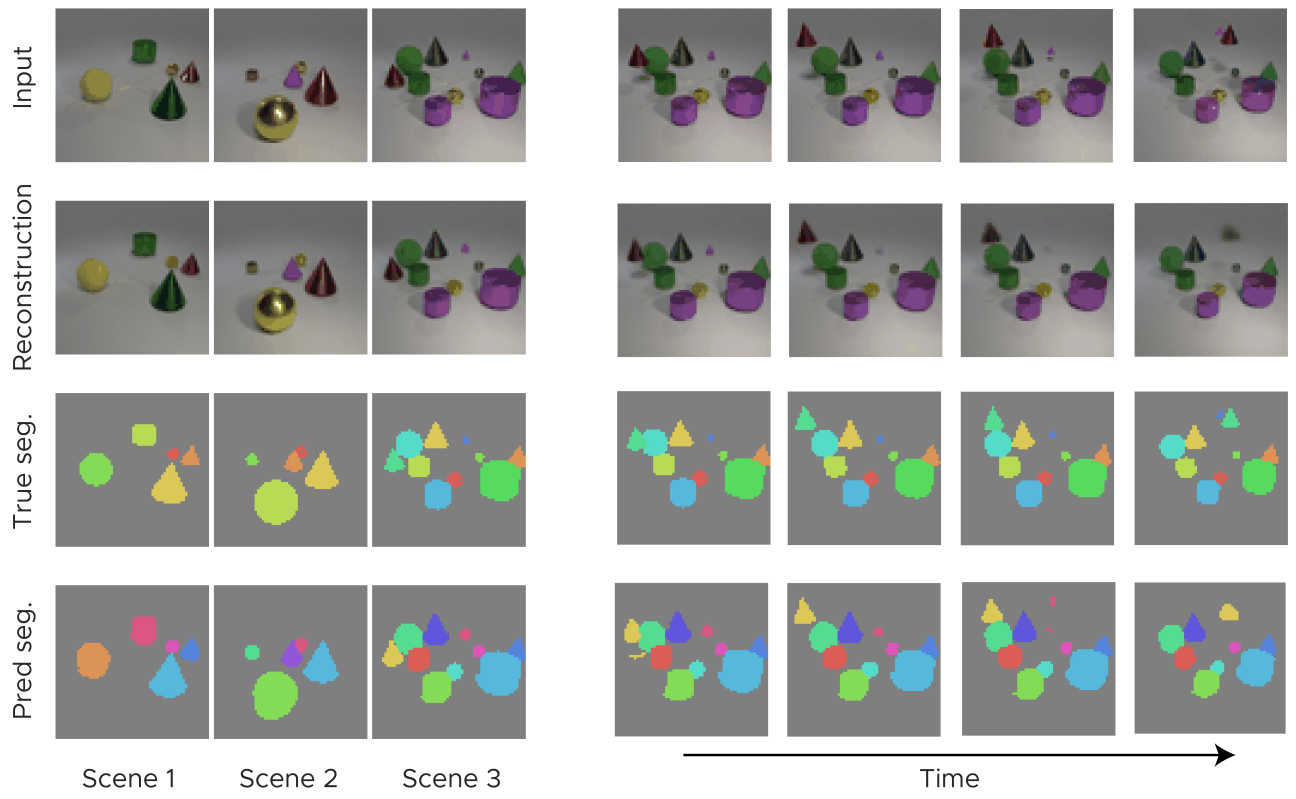

Figure 5. **Model performance for reconstructing and segmenting novel scenes.** (Left) Model performance across three scenes taken from the test set of CATER. The top two rows show model input and reconstruction of input, the bottom two rows show the true segmentation of objects within the scene and the predicted segmentation of the model. (Right) Reconstruction and segmentation performance across time for Scene 3. The top left object (red in the original input) moves up and to the left of the frame. The model is able to accurately segment and track the object across all 16 frames (4 intermediate frames shown here), albeit classifying the shadow as part of the object.

jects on the CATER dataset, by using only a reconstruction loss. We compare our model against other unsupervised segmentation models like MONet [4], S-IODINE [15], SIMONe [21] and SAVi [23] (see Table 2). We achieve comparable or better performance to the best baseline models on CATER (FG-ARI: SAVi 0.928 $\pm$ 0.008; Ours 0.939 $\pm$ 0.013) while outperforming all unsupervised models (SAVi is trained on optical flow information). Based on visual examination of background segmentation in [21], we note that the full ARI score (taking into account both background and foreground) of our model likely outperforms many of the baseline models, which do not report the full score (Ours, FG-ARI: 0.939 $\pm$ 0.013, ARI: 0.825 $\pm$ 0.04). For the more challenging MOVi datasets, we further increase the number of units in the decoding layer to 1024 and train the model for 400000 steps using a batch size of 1. We observe a decline in performance with increasing scene complexity (FG-ARI, MOVi-A 0.790, MOVi-B 0.460, MOVi-C 0.318), while still outperforming SIMONe on all three datasets. Taken together, these comparisons show that our biologically inspired model is not only able to reconstruct images from our novel benchmark but shows comparable performance to state-of-the-art unsupervised scene segmentation models.

7. Discussion

Here we explored the neural representations of an artificial neural network which was trained to perform allocentric topographical processing, similar to how the Four-Mountains-Test is used to predict the early onset of Alzheimer's disease. The model uses visual representations from V4 and IT and uses higher-level areas like the entorhinal and hippocampus to successfully differentiate between scenes from different viewpoints. We show that this biologically inspired model can discern between hundreds of scenes and generalize beyond its training set. Moreover, it is able to reconstruct the visual input through a factorized latent space [21, 24], disentangling object from spatial information.

Our model is able to perform novel view synthesis by imagining scenes from different viewpoints (4MT) or dif-

ferent moments in time (CATER & MOVi) and can segment objects on par with recent state-of-the-art models. One shortcoming of this approach is the relatively small model size which likely prevents it from performing well on more challenging real-world datasets like MOVi-C,D,E or COCO. More powerful visual representations might help for these datasets, for which our visual cortex can be easily replaced with features from models that have a higher similarity to visual cortices [33].

In the future, we want to further explore the difference in neural representations across the objective functions used. We observed that spatially modulated cells also arise in the model using a reconstruction loss, but only in the temporally averaged pathway (MEC) and only in a small subset of neurons. This likely arises from the use of LEC & MEC as a bottleneck, forcing the model to retain scene information in order to fully reconstruct the image, which is not needed for the disentangling of the latent representation.

References

[1] Andrej Bicanski and Neil Burgess. A neural-level model of spatial memory and imagery. *Elife*, 7:e33752, 2018. 2, 5

[2] Irving Biederman. Perceiving real-world scenes. *Science*, 177(4043):77–80, 1972. 2

[3] Chris M Bird and Neil Burgess. The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, 9(3):182–194, 2008. 1, 3, 4

[4] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 4, 7

[5] Patrick Byrne, Suzanna Becker, and Neil Burgess. Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological review*, 114(2):340, 2007. 2

[6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016. 2

[7] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, Mar. 2016. ISSN: 2379-190X. 3

[8] Benjamin J Clark, Christine M Simmons, Laura E Berkowitz, and Aaron A Wilber. The retrosplenial-parietal network and reference frame coordination for spatial navigation. *Behavioral neuroscience*, 132(5):416, 2018. 2

[9] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998. 2

[10] Russell A Epstein. Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences*, 12(10):388–396, 2008. 2

[11] Russell A Epstein and Chris I Baker. Scene perception in the human brain. *Annual review of vision science*, 5:373, 2019. 5

[12] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 1, 3, 6

[13] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019. 6

[14] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022. 6

[15] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019. 7

[16] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. 3

[17] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. 5

[18] Tom Hartley, Chris M Bird, Dennis Chan, Lisa Cipolotti, Masud Husain, Faraneh Vargha-Khadem, and Neil Burgess. The hippocampus is required for short-term topographical memory in humans. *Hippocampus*, 17(1):34–48, 2007. 1, 2, 3, 10

[19] Øyvind Arne Høydal, Emilie Ranheim Skytøen, Sebastian Ola Andersson, May-Britt Moser, and Edvard I Moser. Object-vector coding in the medial entorhinal cortex. *Nature*, 568(7752):400–404, 2019. 2

[20] David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154.2, Jan. 1962. 1

[21] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. *Advances in Neural Information Processing Systems*, 34:20146–20159, 2021. 1, 3, 4, 5, 6, 7

[22] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014. 5

[23] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*, 2021. 6, 7

[24] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018. 2, 4, 7

[25] Colin Lever, Stephen Burton, Ali Jeewajee, John O'Keefe, and Neil Burgess. Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, 29(31):9771–9777, 2009. 2, 5

[26] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 3

[27] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2, 3

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[29] Matthias Nau, Tobias Navarro Schröder, Markus Frey, and Christian F Doeller. Behavior-dependent directional tuning in the human visual-navigation network. *Nature communications*, 11(1):1–13, 2020. 2

[30] John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, Nov. 1971. 1, 5

[31] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

[32] James B Ranck. Head direction cells in the deep cell layer of dorsolateral pre-subiculum in freely moving rats. *Electrical activity of the archicortex*, 1985. 2

[33] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020. 4, 8

[34] Goran Šimić, Ivica Kostović, Bengt Winblad, and Nenad Bogdanović. Volume and number of neurons of the human hippocampal formation in normal aging and alzheimer's disease. *Journal of Comparative Neurology*, 379(4):482–494, 1997. 1

[35] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2

[36] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 3

[37] Benigno Uria, Borja Ibarz, Andrea Banino, Vinicius Zambaldi, Dharshan Kumaran, Demis Hassabis, Caswell Barry, and Charles Blundell. The spatial memory pipeline: a model of egocentric to allocentric understanding in mammalian brains. *BioRxiv*, 2020. 2, 4

[38] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020. 2, 4

[39] James CR Whittington, Joseph Warren, and Timothy EJ Behrens. Relating transformers to models and neural representations of the hippocampal formation. *arXiv preprint arXiv:2112.04035*, 2021. 4

[40] Ruth A Wood, Kuven K Moodley, Colin Lever, Ludovico Minati, and Dennis Chan. Allocentric spatial memory testing predicts conversion from mild cognitive impairment to dementia: an initial proof-of-concept study. *Frontiers in neurology*, 7:215, 2016. 1

[41] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. 1

[42] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014. 5

[43] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021. 1