

---

# ORGANIZATION OF A LATENT SPACE STRUCTURE IN VAE/GAN TRAINED BY NAVIGATION DATA

---

A PREPRINT

**Hiroki Kojima**  
University of Tokyo  
kojima@sacra1.c.u-tokyo.ac.jp

**Takashi Ikegami**  
University of Tokyo  
ikeg@sacra1.c.u-tokyo.ac.jp

February 13, 2022

## ABSTRACT

We present a novel artificial cognitive mapping system using generative deep neural networks, called variational autoencoder/generative adversarial network (VAE/GAN), which can map input images to latent vectors and generate temporal sequences internally. The results show that the distance of the predicted image is reflected in the distance of the corresponding latent vector after training. This indicates that the latent space is self-organized to reflect the proximity structure of the dataset and may provide a mechanism through which many aspects of cognition are spatially represented. The present study allows the network to internally generate temporal sequences that are analogous to the hippocampal replay/pre-play ability, where VAE produces only near-accurate replays of past experiences, but by introducing GANs, the generated sequences are coupled with instability and novelty.

**Keywords** Cognitive Map · GAN · Place Cell · Prediction · Latent Space · Chaos

The cognitive map was first initially proposed by Tolman to explain the deliberate behavior of rats in a maze [1], and place cells [2] and grid cells [3] in the hippocampus are regarded as the biological implementation of it. The mechanism of the formation of these spatial representations has already been discussed [4], and today, the most prevalent explanations are based on the "path integration," which insists that the constructed spatial representation was based on its own movement and was calculated by integrating the self-motion information[5]. These views are supported by the findings that the place cells maintained stable in the dark environment[6], but lost the spatial selectivity by the suppression of the vestibular input [7] or in virtual reality (VR) environments [8].

However, some researchers questioned this path integration view, for example, claiming that grid cells are not necessary for the emergence of the place cells [9, 10], implying that the spatial metric was constructed from the self-motion cues and was not the basis of the spatial representation. Also, the spatial representations varied among different species, for example, the spatial representation of monkeys is associated with egocentric spatial views, suggesting that the cognitive map was more related to the visually centered space in some species[11]. The manner in which cognitive maps are organized differ how Google Maps for example are organized. Maps are cognitive in the sense that they are correlated with the degree and complexity of our cognitive capability [12]. Recently, many researches have reported that the hippocampal structure mapped not only the spatial structure but also the nonspatial features [13, 14, 15], such as the frequency of tones [16], the bird cartoons [17], and social relationships[18]. These findings suggest the existence of a more general underlying mapping mechanism, which is not confined to spatial processing based on the integration of self-motion signals. Furthermore, the hippocampus is not only the center of processing spatial representations, but also responsible for memories, especially episodic memories [13]. In this context, the neural basis of episodic memory is known to be related to the imagery ability [19, 20], and it has been proposed that the many aspects of the memory functions of the hippocampus are considered to be related with scene reconstruction [21], highlighting the importance of the generative nature of the hippocampus.

Herein, as an alternative mechanism to construct the cognitive map, we hypothesize that the map is self-organized at the bottleneck of the sensory reconstruction system. To confirm this hypothesis, we implemented an artificial cognitive map system using only visual information and no explicit metric information. Our hypothesis was partly inspired in a

study on generative deep neural networks, in which the generated images were smoothly mapped in the latent spaces, which are the bottleneck of the architecture [22]. Specifically, we used generative deep neural network called variational autoencoder / generative adversarial network (VAE/GAN)[23], which is a combination of VAE [22] and GAN [24, 25]. These generative deep neural networks learns to encode input images into latent vectors and generate images from the vectors. The structure of the latent space is assumed to reflect the input data structure, for example, the "World model" simulation [26] utilized this type of latent space vectors to encode visual inputs. We trained VAE/GAN on a first-person navigation task wherein the agent moves through a simple virtual environment and studies the characteristics of the internal representations.

The approach of constructing an artificial cognitive map system to understand the underlying mechanism was first proposed by Rössler [27]. The approach was subsequently realized by using recurrent neural networks (RNN). For example, Nolfi and Tani [28] trained a robot with hierarchical RNNs such that the next sensory state of the robot can be predicted and found that each layer of the neural networks encode some regularities of the environment. Noguchi et al. [29] also trained a hierarchical RNNs with recurrent gated units using visual and motor information and found that the map of the environment was self-organized at the higher layer. Banino et al. [30] used a RNNs with a long short-term memory (LSTM) architecture and showed that the grid-cell-like structure was self-organized after the path integration tasks were trained. These systems and other recent systems [31, 32, 33, 34] have different structures, but all of them used motor commands as inputs and were basically based on path integration mechanisms. This is completely different from our system, which only uses visual inputs.

Our system was constructed to predict the upcoming frame of the input videos; thus, the systems can be regarded as an example of video prediction systems [35, 36, 37]. These previous studies basically focused on the quality of video predictions, but our research interest was the resulting structure in the latent space. Thus, we kept our system as simple as possible and quantitatively characterized the latent space vectors.

The other important aspect of the place cells in the hippocampus is "replay." For instance, the cognitive map is activated not only when the mouse is actually exploring the environment but also when the mouse is replaying their past experiences or dreaming about exploring the environment [38, 39, 40]. These sequences were found to not simply be an exact replay of the past experience [41]. For example, the place cells corresponding to the locations that have not yet been traversed are also activated [42, 43]. Our system can also generate temporal sequences by iteratively predicting the next frame in a "closed-loop." Because GAN is used to produce the cognitive map, we have realistic non-existing image scenes, which might result in generating a temporal sequence that is different from the exact replay of the past experience. Hence, we investigated the dynamics of the closed-loop generation for different mapping conditions and aimed to provide a possible mechanism to generate sequences that differ from the exact replay.

## 1 Methods

### 1.1 Predictive VAE/GAN

Our system is based on VAE/GAN [23]. VAE/GAN contains three deep convolutional neural networks: the generator (Gen), discriminator (Dis), and encoder (Enc). (Table. 1) To stabilize the GAN training, we used Wasserstein GAN with gradient penalty (wGAN-gp)[44] for representing GAN.

Those components are very briefly summarized below:

VAE: This is a type of unsupervised learning that includes two convolutional neural networks: Enc and Gen. The Enc encodes the input images into low dimensional latent vectors, and the Gen reconstructs the images from the encoded latent vectors. VAE is characterized by the use of a probability distribution (normal distribution) in the latent space to encode the input image.

GAN: This is a type of unsupervised learning in which two networks compete with each other to learn. One network is the Gen, which generates data from random noise input, and the other network is the Dis, which judges whether the input image is the data generated by the Gen or the real data. Namely, the Gen tries to deceive the Dis, and the Dis tries not to be deceived by the Gen.

wGAN: This compensates for the shortcomings of the Jensen–Shanon divergence (e.g., the loss of gradient) and uses a loss function based on the Wasserstein distance (i.e., earth-moving distance) instead. Practically, wGAN stabilizes the GAN training.

Because we are interested in predicting the following scene based on the current scene, we train the network to predict an image that is  $\tau$  steps ahead (i.e.,  $\mathbf{x}(t + \tau)$ ) of the current image ( $\mathbf{x}(t)$ ), by imposing the network  $\text{Gen}(\text{Enc}(\mathbf{x}(t))) = \mathbf{x}(t + \tau)$  instead of  $\text{Gen}(\text{Enc}(\mathbf{x}(t))) = \mathbf{x}(t)$ , where  $\text{Enc}(\mathbf{x})$  and  $\text{Gen}(\mathbf{z})$  are the outputs of the Enc and the Gen respectively.

Enc	Gen	Dis
3 × 3 64 conv., leaky ReLU	fully-connected, BNorm, leaky ReLU	3 × 3 64 conv., leaky ReLU
4 × 4 128 conv., leaky ReLU	4 × 4 512 deconv., BNorm, leaky ReLU	4 × 4 128 conv., leaky ReLU
3 × 3 128 conv., leaky ReLU	4 × 4 256 deconv., BNorm, leaky ReLU	3 × 3 128 conv., leaky ReLU
4 × 4 256 conv., leaky ReLU	4 × 4 128 deconv., BNorm, leaky ReLU	4 × 4 256 conv., leaky ReLU
3 × 3 256 conv., leaky ReLU	3 × 3 64 deconv., BNorm, leaky ReLU	3 × 3 256 conv., leaky ReLU
4 × 4 512 conv., leaky ReLU		4 × 4 512 conv., leaky ReLU
3 × 3 512 conv., leaky ReLU		3 × 3 512 conv., leaky ReLU
fully-connected		fully-connected

Table 1: The architecture of the components of our networks, the encoder (Enc), the generator (Gen) and the discriminator (Dis).

The actual loss functions used for training in this experiment are presented below, starting with the loss function of the VAE:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\text{pixel}} \quad (1)$$

with

$$\mathcal{L}_{\text{prior}} = D_{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \quad (2)$$

$$\mathcal{L}_{\text{llike}}^{\text{pixel}} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x}(t))} [\log p(\mathbf{x}(t + \tau)|\mathbf{z})] \quad (3)$$

where  $q(\mathbf{z}|\mathbf{x})$  is the Enc,  $p(\mathbf{x}|\mathbf{z})$  is the Gen,  $D_{KL}(q||p)$  stands for Kullback-Leibler divergence, and  $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ .

The loss function of VAE/GAN uses the loss function from the Dis in addition to the VAE loss function  $\mathcal{L}_{\text{VAE}}$  (Eq. 1) and is represented as follows:

$$\mathcal{L}_{\text{VAE/GAN}} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\text{pixel}} + \alpha \mathcal{L}_{\text{GAN}}, \quad (4)$$

with

$$\mathcal{L}_{\text{GAN}} = \text{Dis}(\hat{\mathbf{x}}) - \text{Dis}(\mathbf{x}) + \lambda \mathbb{E}_{p(\hat{\mathbf{x}})} [(\|\nabla_{\hat{\mathbf{x}}} \text{Dis}(\hat{\mathbf{x}})\| - 1)^2], \quad (5)$$

where  $\lambda = 10$ ,  $\hat{\mathbf{x}} \sim p(\mathbf{x}|\mathbf{z})$ , and  $\alpha$  is the weight parameter for the GAN loss.  $\text{Dis}(\mathbf{x})$  represents the outputs of the Dis. We set  $\alpha = 1$ , if not stated otherwise.

In the VAE/GAN used in its first instance in the literature [23], the reconstruction error of VAE/GAN was not the pixel loss between the input image and the generated image, but was measured using the middle layer activation (Appendix.1).

In our experiments, we set the dimension of the latent space to  $d_z = 5, 10, 20$  and the prediction time step to  $\tau = 0, 5, 30$  and compared the results of VAE (Eq. 1) and VAE/GAN (Eq. 4) under each condition. We trained the network for each condition three times by changing the random seed. We trained all models with Adam[45] with  $\beta_1 = 0, \beta_2 = 0.9$ , and the learning rate of 0.0002 as an optimizer. The batch size was set to 64. The Dis was updated five times during each iteration, following the training procedure in wGAN-gp [44]. We used Chainer [46] for the actual implementation and the network structures were based on the implementation by pfnet-research[47].

## 1.2 Dataset

As a training dataset, we used a series of first-person visual inputs of an agent moving in a virtual environment. For this purpose, we built a simple 3D virtual environment similar to a figure-8 maze using Unity and captured visual images of agents moving through it at a constant speed. The environment has one junction, at the intersection of the figure-8, and the agent randomly chooses which direction to go in. The captured images consisted of  $64 \times 64$  pixels, for a total of 480 images (Fig. 1).

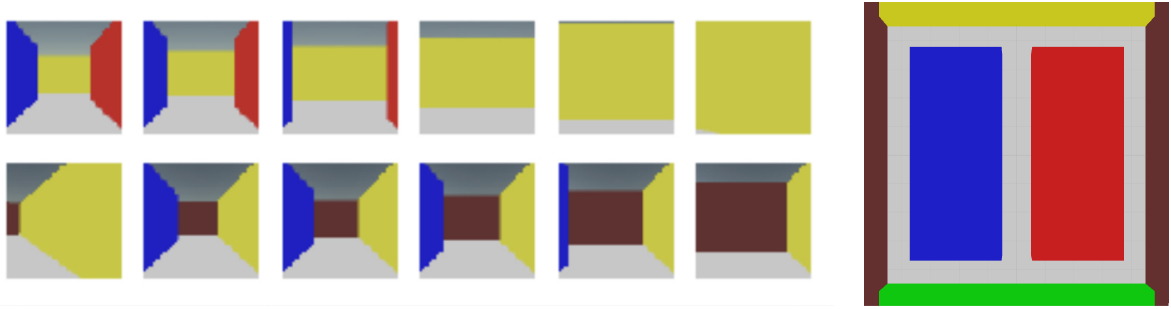


Figure 1: Examples of the temporal sequence of the images in our dataset. Right: Top view of the virtual environment.

## 2 Results

Using these networks, we tested whether the agents could predict the next scene from the current one. The results of the training (Fig. 2) show that the network can successfully predict the scene  $\tau$  steps in advance in each condition. The loss function was stabilized after around 8000 iterations of training in all conditions; thus, we will use the results after 8000 iterations in the following analysis. Unless otherwise stated, the analysis is based on the results averaged over 8000 to 10000 iterations from each seed ( $n = 3$ ). We have experimented with latent space dimensions  $d_z = 5, 10, 20$ , but in our analysis, the trend of the results did not change for varying values of  $d_z$ , so we will only describe the results for  $d_z = 10$ . We first analyzed how the images were mapped into the latent space by the Enc. Then, we introduced a "closed-loop image sequence generation" using the generated images as input recursively and characterized the sequences generated by the closed-loop method for each condition.

We first investigated the nature of the latent mappings. The following questions were addressed: Are images that are close in distance in a real space mapped into close points in the latent space (Fig. 4)? How high is the dimensionality of latent space (Fig. 5)? Is the orbit properly captured (Fig. 6, 8)? Can a clear image be generated (Fig. 7)? Then, by varying the ratio of VAE to GAN, we examine how the GAN contributes to the above properties, including the smoothness of the latent space and the similarities in the left and right pathways (Fig. 9). Furthermore, by adopting the predicted images of the network as input images in the next iteration, we can check how stable or unstable the representation of the latent space is as a deterministic dynamical system (Fig. 11).

### 2.1 Patterns in Latent Spaces

The Enc encodes the images  $\mathbf{x}(t)$  in the dataset into the latent vectors as  $\mathbf{z}(t) = \text{Enc}(\mathbf{x}(t))$ . We analyze the latent vector  $\{\mathbf{z}(t)\}$  and investigate how the input images are mapped into the latent space.

Since the agent traverses the maze while looking at the images, it is important to know the extent to which the actual image pattern is properly embedded in the latent space. For this purpose, we compared the distance matrix of the input images  $\|\mathbf{x}(t_1) - \mathbf{x}(t_2)\|$ , the target images  $\|\mathbf{x}(t_1 + \tau) - \mathbf{x}(t_2 + \tau)\|$  and the corresponding latent vectors  $\|\mathbf{z}(t_1) - \mathbf{z}(t_2)\|$  and calculated the correlation coefficients between the values of these distance matrices (Fig. 3). We found that a correlation exists and is stronger between the target images and the latent vectors than between the input images and the latent vectors, especially at  $\tau = 30$  ( $p < 0.001$ , Tukey's HSD) (Fig. 4). This indicates that the distance structure of the latent vectors reflected the distance structure of the images after  $\tau$  time steps.

To estimate the extent to which the mapped vectors are linearly projected, we compared the contribution of principal component analysis (PCA). We found that the cumulative contribution of PCA in VAE/GAN was higher than that of a single VAE (Fig. 5) or projected into a lower dimensional space. The examples of the latent mapping  $\{\mathbf{z}(t)\}$  in PCA space are shown in Fig. 6, and the generated images from the latent vectors reconstructed from the grid that is reconstructed from the first two principal components are shown in Fig. 7. On careful inspection of the images generated from the PCA space, the image generated from VAE (left) tends to be blurred compared to that generated from VAE/GAN, and there seems to be less diversity in the generated images (Fig. 7).

We further characterized the latent mapping  $\{\mathbf{z}(t)\}$  in the PCA space by calculating two measures: smoothness  $S_{\text{PCA}}$  and trajectory dissimilarity  $d_{\text{LR}}$ . Smoothness was evaluated as the average of the directional change and calculated as follows:

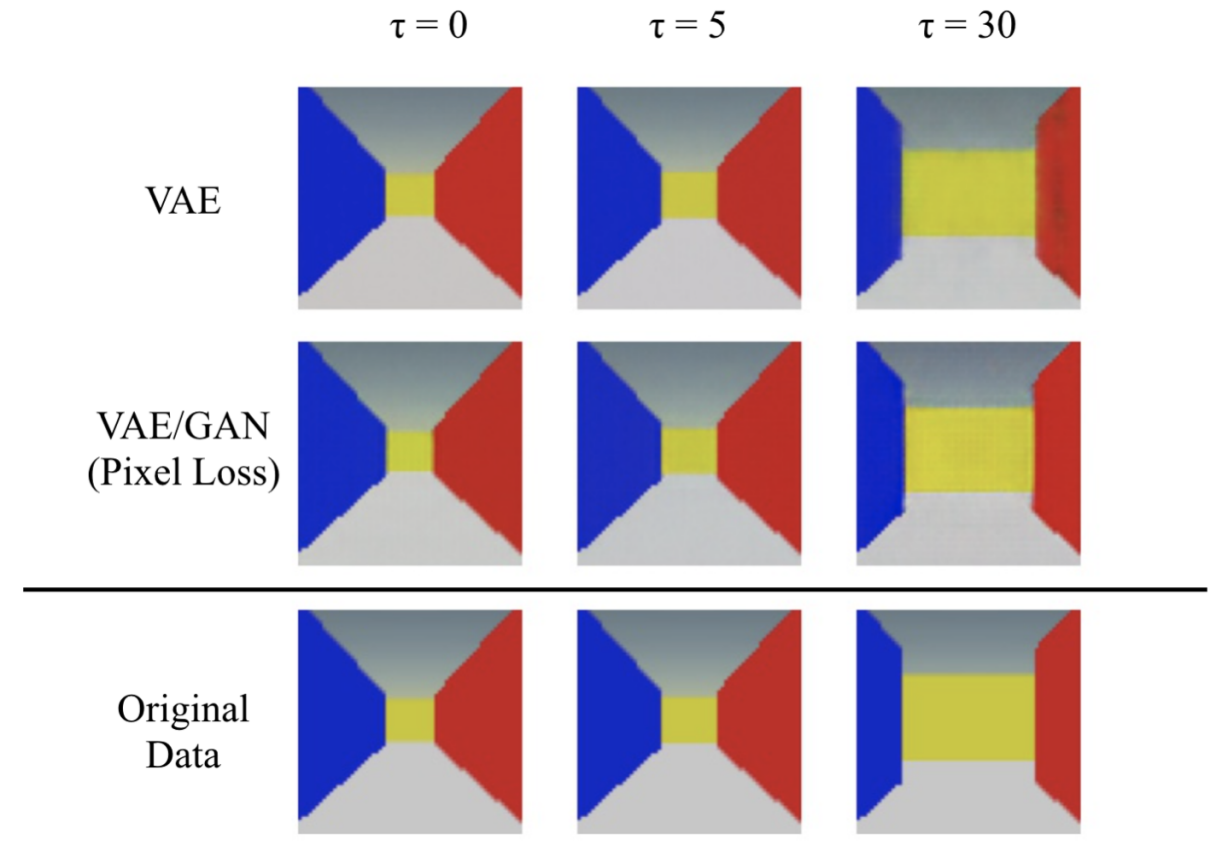


Figure 2: Examples of the generated images from the trained networks of VAE and VAE/GAN, with different prediction time steps ( $\tau = 0, 5, 30$ ). The dimension of latent space was set to  $z = 10$ . The targeted images are shown in the bottom row.

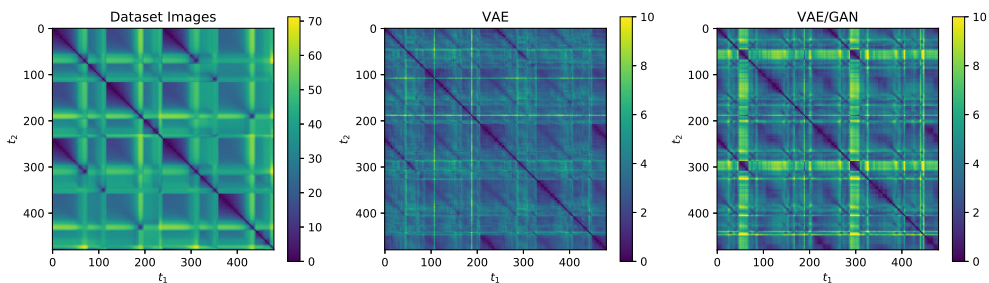


Figure 3: The distance matrix of image dataset  $\|\mathbf{x}(t_1) - \mathbf{x}(t_2)\|$  (left) and the examples of the distance matrix of the corresponding latent vectors of VAE and VAE/GAN  $\|\mathbf{z}(t_1) - \mathbf{z}(t_2)\|$  ( $\tau = 30$ ) (middle and right).

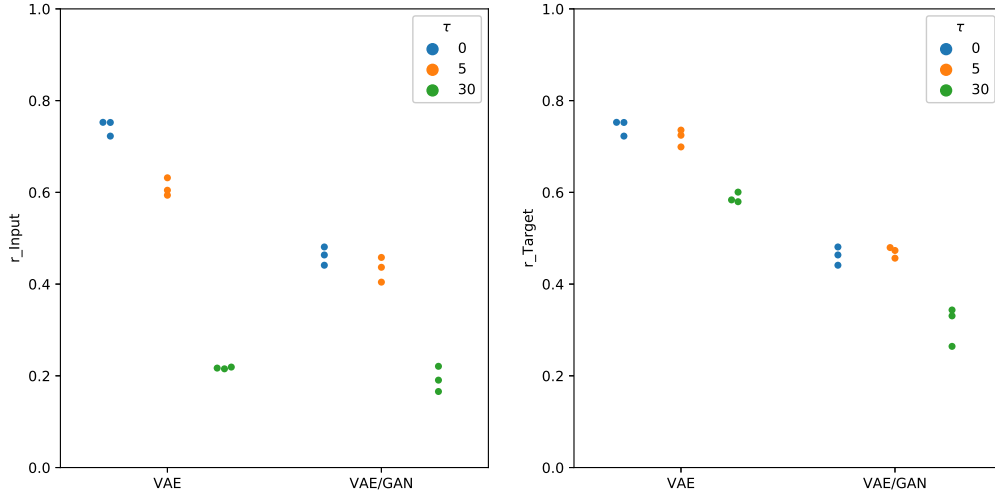


Figure 4: Correlation coefficient for distance structure of the input images/ target images and the latent vectors. Each point correspond to the result from each random seed ( $n = 3$ ). Left: Correlation coefficient between the distance matrix of the input images  $\|\mathbf{x}(t_1) - \mathbf{x}(t_2)\|$  and the distance matrix of the corresponding latent vectors  $\|\mathbf{z}(t_1) - \mathbf{z}(t_2)\|$ . Right: Correlation coefficient between the distance matrix of the target images  $\|\mathbf{x}(t_1 + \tau) - \mathbf{x}(t_2 + \tau)\|$  and the distance matrix of the latent vectors  $\|\mathbf{z}(t_1) - \mathbf{z}(t_2)\|$ .

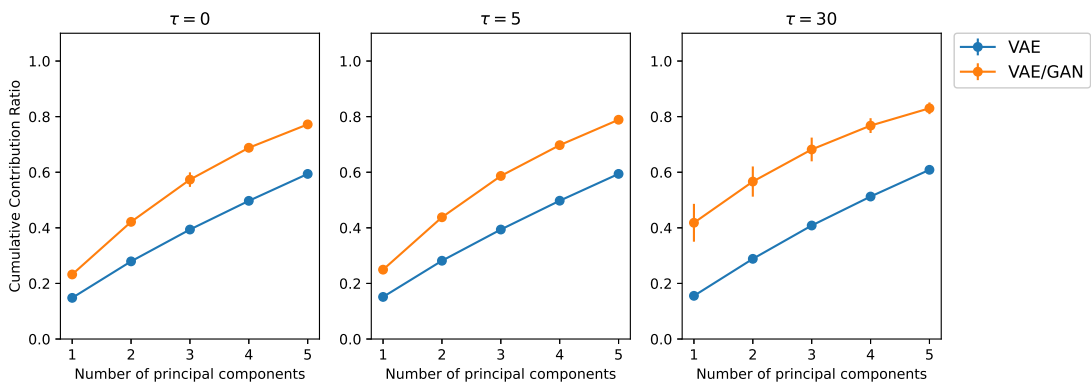


Figure 5: Cumulative contribution ratios of principal component analysis (PCA) ( $d_z = 10$ ) with different prediction timesteps,  $\tau = 0, 5, 30$ . The error bars denote the standard deviation of the results from different random seeds ( $n = 3$ ).

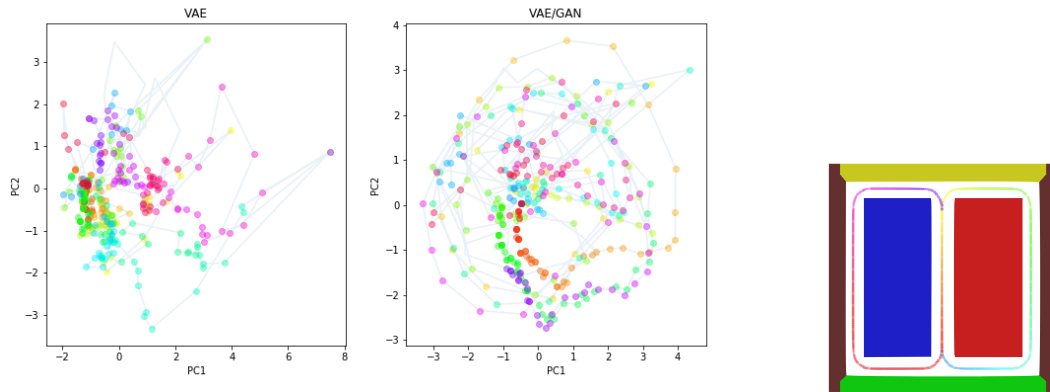


Figure 6: Example of the latent vectors  $\{z(t)\}$  mapped in the PCA space. Each color corresponds to the position in the virtual space (Right). Left: VAE ( $\tau = 5$ ), Middle: VAE/GAN ( $\tau = 5$ ).

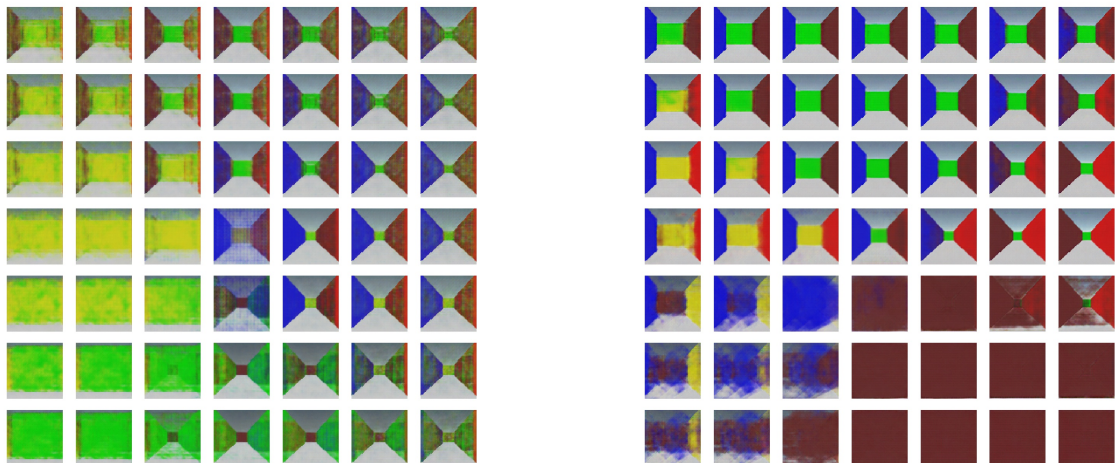


Figure 7: Example of the generated images from the latent vectors grid from the first and second principal components. Left: VAE ( $\tau = 30$ ), Right: VAE/GAN ( $\tau = 30$ ).

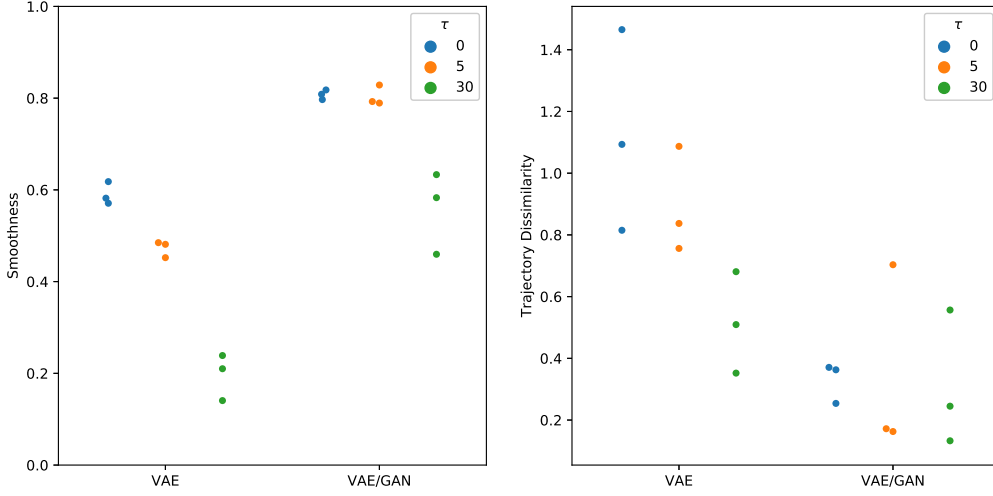


Figure 8: Characterization of the PCA maps. Left: Smoothness of the trajectory  $\{z(t)\}$  in the PCA space,  $S_{\text{PCA}}$ . Right: Trajectory dissimilarity between the right and left pathways of the 8-shaped environment,  $d_{\text{LR}}$ .

$$S_{\text{PCA}} = \left\langle \frac{(z_{\text{PCA}}(t) - z_{\text{PCA}}(t-1)) \cdot (z_{\text{PCA}}(t+1) - z_{\text{PCA}}(t))}{\|z_{\text{PCA}}(t) - z_{\text{PCA}}(t-1)\| \|z_{\text{PCA}}(t+1) - z_{\text{PCA}}(t)\|} \right\rangle, \quad (6)$$

where  $z_{\text{PCA}}(t)$  is the mapped latent vectors  $z(t)$  into the 2D PCA space.

Trajectory dissimilarity  $d_{\text{LR}}$  is the median relative distance in the 2D PCA space between the latent trajectory of the left and right pathway of the environment. When we define the distance between  $z(t_1)$  and  $z(t_2)$  in the PCA space as,

$$d_{\text{PCA}}(t_1, t_2) = \|z_{\text{PCA}}(t_1) - z_{\text{PCA}}(t_2)\|, \quad (7)$$

then the trajectory dissimilarity  $d_{\text{LR}}$  is calculated as,

$$d_{\text{LR}} = \frac{\text{Median}(\{d_{\text{PCA}}(t_{\text{L}} + \Delta t, t_{\text{R}} + \Delta t)\}_{0 \leq \Delta t \leq T_{\text{path}}})}{\text{Median}(\{d_{\text{PCA}}(t_{\text{L}} + \Delta t_1, t_{\text{R}} + \Delta t_2)\}_{0 \leq \Delta t_1, \Delta t_2 \leq T_{\text{path}}})}, \quad (8)$$

where  $\text{Median}()$  is the median value,  $t_{\text{L}}$  and  $t_{\text{R}}$  are the initial times of the left and right paths of the 8-shaped environment, respectively, and  $T_{\text{path}}$  is the duration of each pathway (Fig. 8).

We found that the smoothness of the trajectory was higher in VAE/GAN compared to that in VAE ( $p < 0.001$ , Tukey's HSD). Also, the trajectories of the left and right pathways in the latent space were closer in VAE/GAN, especially when  $\tau = 0$  ( $p < 0.05$ , Tukey's HSD) (Fig. 8). These results suggest that, with GAN, the movement direction in the latent space become stable and the two corresponding pathways, the left and the right, in the environment are similarly encoded in the latent space, which implies that the recognition of the environmental structure goes beyond the pixel similarity of the images, as shown in Fig. 3.

To further investigate the contribution of GAN, we characterized the latent vectors resulting from the training of the loss function with different weighted  $\alpha$  for the GAN loss function (Eq. 4). Here, we set  $\tau = 5$ . The results are shown in Fig. 9. First, we found that by increasing  $\alpha$ , the correlation between the distance matrix of the target images and

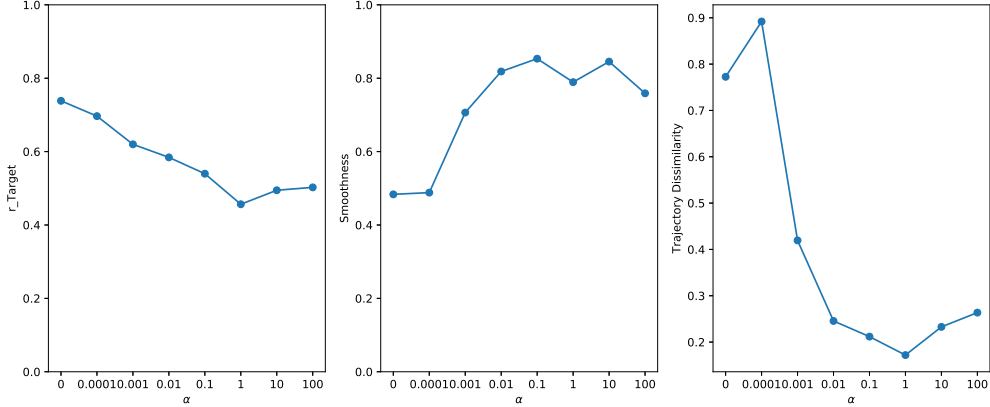


Figure 9: Dependence of the GAN weight parameter  $\alpha$  on the resulting latent map feature. Left: Correlation coefficient between the distance matrix of the target images and latent vectors. Middle: Smoothness  $S_{\text{PCA}}$  of the trajectory  $\{z(t)\}$ . Right: Trajectory dissimilarity  $d_{\text{LR}}$  between the right and left pathways of the 8-shaped environment.

the latent vectors, which corresponds the results in Fig. 4, decreased. This indicates that as the weights for GAN increases, the latent mapping does not simply reflect the pixel similarity of the images. Second, the smoothness of the trajectories  $S_{\text{PCA}}$  was enhanced with the increase in GAN weighting. Therefore, by introducing GAN, the direction of the movement in the latent space stabilized, which suggests novel coordination (i.e., global coordinates), which is different from the VAE, which encodes each image separately. Third, we observed a decrease in the trajectory dissimilarity  $d_{\text{LR}}$  with the increase in  $\alpha$ . Thus, by introducing GAN, the network recognizes the left and right pathways as alike, although not so similar in terms of the pixels of the corresponding images, which suggests a more abstract grasp of the environment. With too much weight on GAN loss,  $d_{\text{LR}}$  started to increase, because, we speculate that, the mapping becomes inaccurate because there is less weighting on  $\mathcal{L}_{\text{like}}^{\text{pixel}}$  (Eq. 3).

In our dataset, the junction in the 8-shaped environment was different from the other part of the environment in such a way that the future outcome has two possibilities depending on whether the agent turns left or right. We investigated that how each network deals with these possible multiple outcomes (Fig. A1). We found that each network has different coding schemes to deal with these multiple possibilities. In the case of VAE, the network ends up generating the superposition image of the two possibilities at the junction, which seems to minimize the pixel loss between the generated image and the images of two possibilities. VAE/GAN also shows this tendency, but due to the presence of GAN, the generated image is not a simple superposition of the two possible outcomes, but it transformed in a way that it can deceive the Dis. In contrast, VAE/GAN<sub>layerLoss</sub>, which corresponds to the original VAE/GAN architecture (Appendix.1), did not generate the superpositioned image, because the VAE/GAN evaluated the generated image based on the activation of the middle layer of the Dis, so the simple superposition did not minimize the loss function at the junction. Instead of the superposition, VAE/GAN<sub>layerLoss</sub> output the image of each possibility, but both possibilities are encoded in the neighbors of the other in the latent space (Fig. A1).

## 2.2 Sequence Generation by Closed Loop

After training the network to predict the upcoming visual inputs, we can make the network autonomous by using the following procedure. First, the initial image  $x_0$  is converted into a latent space vector by the Enc,  $z_0 = \text{Enc}(x_0)$ , and an image is generated from the latent space vector by the Gen,  $x_1 = \text{Gen}(z_0)$ . This image is now used as the input image. By repeating this process recursively, the network continues to predict the next image without receiving the actual image. We call this method a "closed loop" because it closes itself independent of any external input. We believe that this corresponds to, for example, the act of dreaming, or something like that. An example of the generated image sequence is shown in Fig. 10.

A closed loop can be regarded as a dynamical system that converts an input image  $x$  into  $\text{Gen}(\text{Enc}(x))$  in a deterministic manner. Usually the Enc outputs the latent space vectors in a stochastic manner by sampling them from the Gaussian



Figure 10: Examples of the generated image sequence by the closed loop from VAE and VAE/GAN ( $\tau = 5$ ).

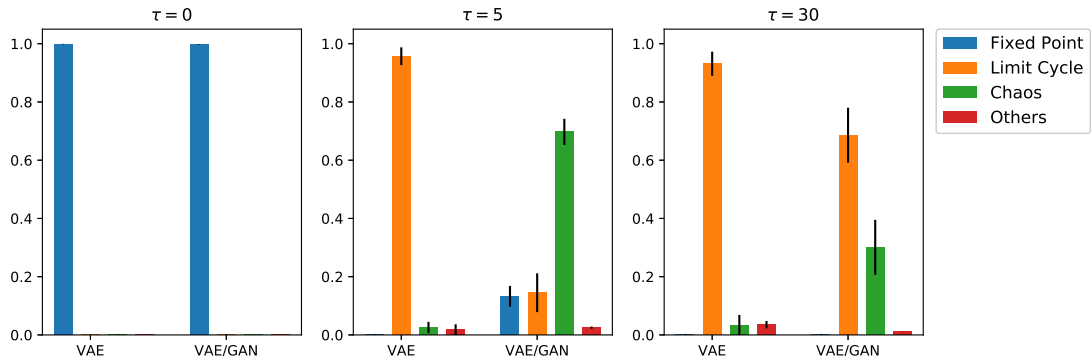


Figure 11: The ratio of each type of the closed-loop trajectories. The error bars indicate the standard deviation of the results from different random seeds ( $n = 3$ ).

distribution with the parameters from the Enc output. However, here, we used the mean value of the Gaussian distribution as the output of Enc( $\mathbf{x}$ ) to keep the closed loop procedure as deterministic. We followed this procedure from the input images for every 5 steps and iterated 200 times for each condition. The examples of the generated images from 180 to 200 steps are shown in Fig. A3. We denote the image and the latent vector at the  $i$ th iteration in the closed loop procedure as  $\mathbf{x}_i$  and  $\mathbf{z}_i$ , respectively.

The resulting trajectories generated from the closed loop can be classified into three categories: fixed point, limit cycle, and chaotic dynamics. Trajectories converged to a fixed point were detected by calculating whether the accumulated change in the latent vector after 175th iterations  $\sum_{i=175}^{199} \|\mathbf{z}_{i+1} - \mathbf{z}_i\|^2$  was below the threshold; here, we set the threshold as  $z_{thr} = 10^{-5}$ . The limit cycle was detected based on whether the latent vector similar to the latent vector in the final iteration was included in the trajectory ( $\min_i (\|\mathbf{z}_{200} - \mathbf{z}_i\|^2) < 10^{-8}$ ), and was not classified as the fixed point. We estimated the Lyapunov exponents of the rest of the trajectories by using the method by Rosenstein et al. [48] and classified the trajectories with positive Lyapunov exponents as chaotic trajectories.

We classified the closed-loop trajectories for each conditions (Fig. 11). When  $\tau = 0$ , almost all trajectories converged to fixed points. When  $\tau = 5, 30$ , the trajectories from the VAE showed mainly limit cycle dynamics, while those from VAE/GAN showed an increased fraction of chaotic trajectories, which indicates that with GAN, the closed-loop dynamics was somewhat unstable and was able to generate novel sequences.

### 3 Discussion

In this paper, we presented an artificial cognitive map system based on a generative deep neural network and only using visual inputs. We aimed to show the alternative mechanism for the cognitive map that was different from path integration (i.e., the map was calculated from its own motion). Here, we argue that some aspects of the cognitive maps can be explained without the path integration. Our view is partly inspired by the recent findings of nonspatial coding in the hippocampal structures [14], which cannot be directly explained by the integration of the self-motion information.

First, we found that the distance of the dataset images (i.e., external world) was reflected in the distance structure of the latent vectors. The encoding of the proximity is one of the important ingredients for the cognitive map. Some studies, such as those focused on the encoding of social relationships [18] and those that use life-logging data [49], focused only on the relationship between the distance structure in the data and the distance structure of the corresponding neural activity, so these mappings can be explained only by this aspect. Another important finding was that when the network was trained to predict future scenes, the latent vectors reflected the distance structure of the predicted images, not the input images. This is in accordance with the fact that the cognitive map encodes not the present position but the predicted positions [50].

Second, by using GAN, we found that the trajectory in the latent space became smooth compared to that with only VAE. This feature made the movement direction in the latent space stable, which means that when the agent moved in one direction, the corresponding latent vector also moved in certain direction. We regard this as the emergence of the global coordinates in the latent space. Also, we found that the trajectories corresponding to the left and right pathways were more alike in VAE/GAN. In the VR experiment conducted using rats [8], it was reported that when the rats always moved in a certain pathway, which they called the "systematic pillar condition," each segment of the pathway similarly activated the hippocampal neurons. This was interpreted as the encoding of the distance traversed, and this was enabled by the coupling of different modalities. Our result provides different perspectives to this. Each segment of the pathway can be mapped similarly if the agent can predict the upcoming visual input, and the latent representation was organized by GAN-like mechanisms. This phenomenon does not necessarily correspond to the "distance", which was more related to the path integration view, and does not require the coupling between different modalities.

Our system is based on a generative system, and we think that this is related to the other aspects of the hippocampus, such as the (episodic) memory. In the context of the episodic memory, Hassabis et al. [19, 20] revealed that episodic memory and imagery shared the same neural basis. To be compatible with these two functionalities, the system has to be able to generate novel scenes while retaining the specific past memories. This resembles to our system considering that it can not only generate the past scenes but also produce novel and plausible scenes, which is enabled by GAN training. In the episodic memory study [19], it was reported that the activities of a specific brain region called precuneus was related to the familiarity of the visual experience, which might suggest that it functions like the Dis in GAN.

We also investigated the properties of the closed-loop trajectories. The closed-loop trajectories from VAE were stable and almost faithfully reproduce the experienced trajectories, but the trajectories from VAE/GAN were unstable, often showing chaotic dynamics, and we claimed that this might be the origin of the novelty in the "replay" in the hippocampus [42, 43]. In this way, using a single model, we discussed the three aspects of the hippocampus: cognitive map, episodic memory and "replay". Usually, these are modeled independently using separate models, but we unified because they are actually implemented in the same region of the brain, hippocampus. The characteristics of the different functionality of the hippocampus should be correlated with each other. For example, we speculate that the smoothness in the cognitive map is related to the novelty in the "replay". We hope that our study will provide the required information for the unified understanding of functionality of hippocampus.

## 4 Conclusions

We have constructed a novel artificial cognitive mapping system using generative deep neural networks, which can map input images to latent vectors and generate temporal sequences internally. The results show that the distance of the predicted image after training is reflected in the distance of the corresponding latent vector. This indicates that the latent space is constructed to reflect the proximity structure of the data set, and may provide a mechanism through which many aspects of cognition are spatially represented [15]. The present study allows the network to internally generate temporal sequences that are analogous to hippocampal replay/pre-play ability, where VAE produces only near-accurate replays of past experiences, but by introducing GANs, the latent vectors of the temporally close images are closely aligned and the sequences acquired some instability. This may be the origin of the generation of the new sequences that are found in the hippocampus [42, 43].

## References

- [1] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.
- [2] John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.
- [3] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

- [4] Edvard I Moser, May-Britt Moser, and Bruce L McNaughton. Spatial representation in the hippocampal formation: a history. *Nature neuroscience*, 20(11):1448–1464, 2017.
- [5] Bruce L McNaughton, Carol A Barnes, Jason L Gerrard, Katalin Gothard, Min W Jung, James J Knierim, H Kudrimoti, Y Qin, WE Skaggs, M Suster, et al. Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *The Journal of experimental biology*, 199(1):173–185, 1996.
- [6] Gregory J Quirk, Robert U Muller, and John L Kubie. The firing of hippocampal place cells in the dark depends on the rat’s recent experience. *Journal of Neuroscience*, 10(6):2008–2017, 1990.
- [7] Robert W Stackman, Ann S Clark, and Jeffrey S Taube. Hippocampal spatial representations require vestibular input. *Hippocampus*, 12(3):291–303, 2002.
- [8] Zahra M Aghajani, Lavanya Acharya, Jason J Moore, Jesse D Cushman, Cliff Vuong, and Mayank R Mehta. Impaired spatial selectivity and intact phase precession in two-dimensional virtual reality. *Nature neuroscience*, 18(1):121–128, 2015.
- [9] Mark P Brandon, Andrew R Bogaard, Christopher P Libby, Michael A Connerney, Kishan Gupta, and Michael E Hasselmo. Reduction of theta rhythm dissociates grid cell spatial periodicity from directional tuning. *Science*, 332(6029):595–599, 2011.
- [10] Julie Koenig, Ashley N Linder, Jill K Leutgeb, and Stefan Leutgeb. The spatial periodicity of grid cells is not sustained during reduced theta oscillations. *Science*, 332(6029):592–595, 2011.
- [11] Edmund T Rolls and J-Z Xiang. Spatial view cells in the primate hippocampus and memory recall. *Reviews in the Neurosciences*, 17(1-2):175–200, 2006.
- [12] György Buzsáki and Edvard I Moser. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature neuroscience*, 16(2):130–138, 2013.
- [13] Daniela Schiller, Howard Eichenbaum, Elizabeth A Buffalo, Lila Davachi, David J Foster, Stefan Leutgeb, and Charan Ranganath. Memory and space: towards an understanding of the cognitive map. *Journal of Neuroscience*, 35(41):13904–13911, 2015.
- [14] Jacob LS Bellmund, Peter Gärdenfors, Edvard I Moser, and Christian F Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415), 2018.
- [15] Timothy EJ Behrens, Timothy H Muller, James CR Whittington, Shirley Mark, Alon B Baram, Kimberly L Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.
- [16] Dmitriy Aronov, Rhino Nevers, and David W Tank. Mapping of a non-spatial dimension by the hippocampal-entorhinal circuit. *Nature*, 543(7647):719–722, 2017.
- [17] Alexandra O Constantinescu, Jill X O’Reilly, and Timothy EJ Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.
- [18] Rita Morais Tavares, Avi Mendelsohn, Yael Grossman, Christian Hamilton Williams, Matthew Shapiro, Yaacov Trope, and Daniela Schiller. A map for social navigation in the human brain. *Neuron*, 87(1):231–243, 2015.
- [19] Demis Hassabis, Dharshan Kumaran, and Eleanor A Maguire. Using imagination to understand the neural basis of episodic memory. *Journal of neuroscience*, 27(52):14365–14374, 2007.
- [20] Demis Hassabis, Dharshan Kumaran, Seralynne D Vann, and Eleanor A Maguire. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731, 2007.
- [21] Demis Hassabis and Eleanor A Maguire. Deconstructing episodic memory with construction. *Trends in cognitive sciences*, 11(7):299–306, 2007.
- [22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [23] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.
- [24] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.
- [25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [26] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

- [27] Otto E Rössler. An artificial cognitive map system. *BioSystems*, 13(3):203–209, 1981.
- [28] Stefano Nolfi and Jun Tani. Extracting regularities in space and time through a cascade of prediction networks: The case of a mobile robot navigating in a structured environment. *Connection Science*, 11(2):125–148, 1999.
- [29] Wataru Noguchi, Hiroyuki Iizuka, and Masahito Yamamoto. Cognitive map self-organization from subjective visuomotor experiences in a hierarchical recurrent neural network. *Adaptive Behavior*, 25(3):129–146, 2017.
- [30] Andrea Banino, Caswell Barry, Benigno Uribe, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.
- [31] 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.
- [32] Benigno Uribe, 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.
- [33] Rajeev V Rikhye, Nishad Gothoskar, J Swaroop Guntupalli, Antoine Dedieu, Miguel Lázaro-Gredilla, and Dileep George. Learning cognitive maps as structured graphs for vicarious evaluation. *bioRxiv*, page 864421, 2020.
- [34] Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature communications*, 12(1):1–13, 2021.
- [35] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.
- [36] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.
- [37] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.
- [38] John O’keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.
- [39] György Buzsáki, Cornelius H Vanderwolf, et al. Cellular bases of hippocampal eeg in the behaving rat. *Brain Research Reviews*, 6(2):139–171, 1983.
- [40] George Dragoi and Susumu Tonegawa. Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, 469(7330):397–401, 2011.
- [41] David J Foster. Replay comes of age. *Annual review of neuroscience*, 40:581–602, 2017.
- [42] Anoopum S Gupta, Matthijs AA van der Meer, David S Touretzky, and A David Redish. Hippocampal replay is not a simple function of experience. *Neuron*, 65(5):695–705, 2010.
- [43] Federico Stella, Peter BaracsKay, Joseph O’Neill, and Jozsef Csicsvari. Hippocampal reactivation of random trajectories resembling brownian diffusion. *Neuron*, 102(2):450–461, 2019.
- [44] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.
- [45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [46] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, volume 5, pages 1–6, 2015.
- [47] pfnet research. chainer-gan-lib, 2017. <https://github.com/pfnet-research/chainer-gan-lib/>.
- [48] Michael T Rosenstein, James J Collins, and Carlo J De Luca. A practical method for calculating largest lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1-2):117–134, 1993.
- [49] Dylan M Nielson, Troy A Smith, Vishnu Sreekumar, Simon Dennis, and Per B Sederberg. Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences*, 112(35):11078–11083, 2015.
- [50] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643–1653, 2017.

## **Acknowledgement**

This work was partially supported by JSPS KAKENHI Grant, Correspondence and Fusion of AI and Brain (17H06024, 19H049790) and Chronogenesis : How the mind generates time (19H05306).

## A Appendix

### A.1 VAE/GAN with Middle Layer Loss

In the original VAE/GAN proposed in [23], the reconstruction error of VAE/GAN was not the pixel loss between the input image and the generated image but was measured using the middle layer activation, which we call  $\text{VAE/GAN}_{\text{layerLoss}}$ . In this case, the loss function is given as follows:

$$\mathcal{L}_{\text{VAE/GAN}(\text{layerLoss})} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\text{Dis}_l} + \mathcal{L}_{\text{GAN}}, \quad (9)$$

with

$$\mathcal{L}_{\text{prior}} = D_{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \quad (10)$$

$$\mathcal{L}_{\text{llike}}^{\text{Dis}_l} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x}(t))} [\log p(\text{Dis}_l(\mathbf{x}(t + \tau))|\mathbf{z})] \quad (11)$$

$$\mathcal{L}_{\text{GAN}} = \text{Dis}(\hat{\mathbf{x}}) - \text{Dis}(\mathbf{x}) + \lambda \mathbb{E}_{p(\hat{\mathbf{x}})} [(\|\nabla_{\hat{\mathbf{x}}} \text{Dis}(\hat{\mathbf{x}})\| - 1)^2], \quad (12)$$

where  $q(\mathbf{z}|\mathbf{x})$  is the Enc,  $p(\mathbf{x}|\mathbf{z})$  is the Gen,  $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\hat{\mathbf{x}} \sim p(\mathbf{x}|\mathbf{z})$ , and  $p(\text{Dis}_l(\mathbf{x})|\mathbf{z}) = \mathcal{N}(\text{Dis}_l(\mathbf{x})|\text{Dis}_l(\hat{\mathbf{x}}), \mathbf{I})$ .

### A.2 Output Image Around the Bifurcation

The example of the generated images at the bifurcation.

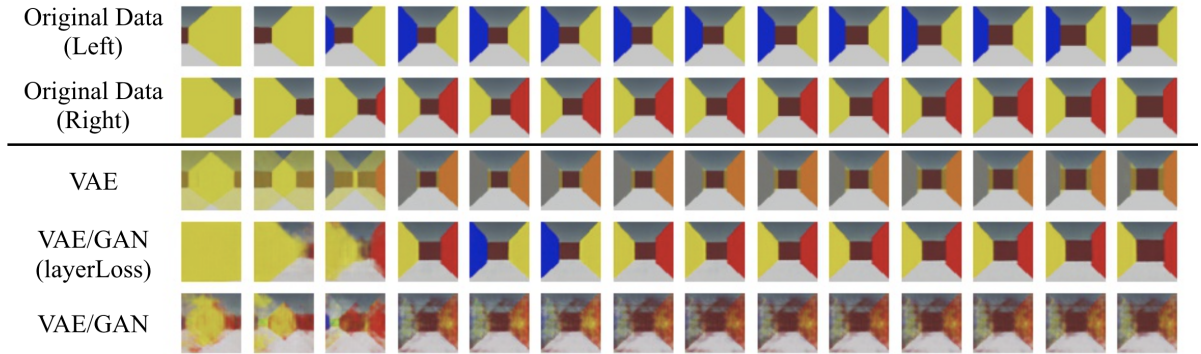


Figure A1: Examples of the output images around the bifurcation ( $\tau = 30$ ). First and second row: The target images of left and right pathways, respectively. Third row: The generated images from VAE ( $\tau = 30$ ). Fourth row : The generated images from  $\text{VAE/GAN}_{\text{layerLoss}}$  ( $\tau = 30$ ). Fifth row : The generated images from VAE/GAN ( $\tau = 30$ ).

### A.3 Variability of the generated images from the same input image

Using the encoder, we can stochastically sample the latent vectors and by putting these vectors the generator can output various outputs. Usually, this variation is relatively small and the generator outputs effectively the same image when provided the same input.

We investigated whether there were some input images which caused the large variation in the sampled output images (Fig. A2). We sampled the output images by sampling 10 latent vectors by providing the same input image and calculated the variability by taking the mean of the standard deviation of the pixel data. We found that when we feed the images before the bifurcation to the VAE/GAN with prediction timestep  $\tau = 30$ , the variation of the output images were relatively large.

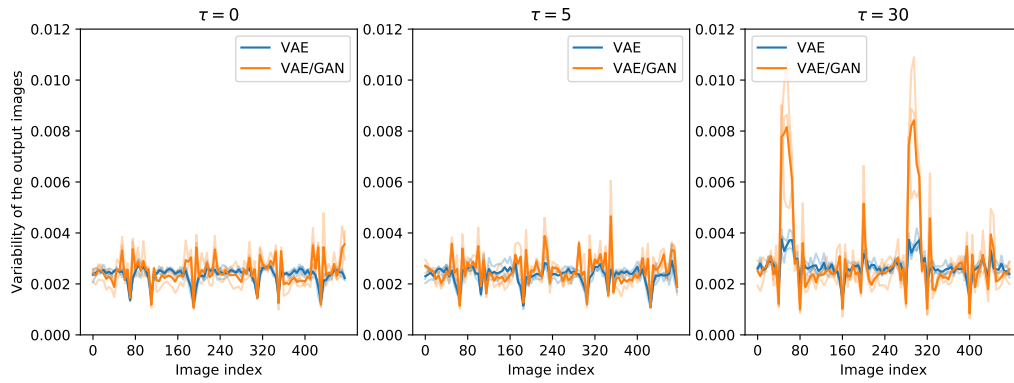


Figure A2: The standard deviation of the 10 images stochastically generated from the same input image. Image index 80 and 320 correspond to the bifurcation in the environment. The results from different random seeds were overlaid.

#### A.4 Generated images from long step closed-loop generations

The example of the generated images from the long step closed-loop generations.

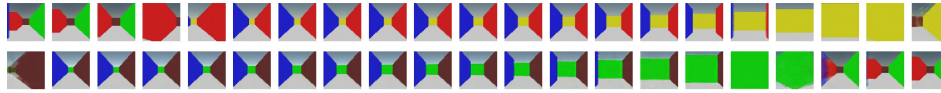


Figure A3: Examples of the generated image sequence after the 180 steps of the closed loop shown in Fig.10. Top row: VAE, Bottom row: VAE/GAN.

### **Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: