Neuroscience

The inevitability and superfluousness of cell types in spatial cognition

Reviewed Preprint
v1 • August 28, 2024
Not revised

Xiaoliang Luo, Robert M Mok, Bradley C Love

Department of Experimental Psychology, University College London, London, UK • MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK • Department of Psychology, Royal Holloway, University of London Royal Holloway, University of London, Egham, UK • The Alan Turing Institute, London, UK

Abstract

Summary

Discoveries of functional cell types, exemplified by the cataloging of spatial cells in the hippocampal formation, are heralded as scientific breakthroughs. We question whether the identification of cell types based on human intuitions has scientific merit and suggest that "spatial cells" may arise in non-spatial computations of sufficient complexity. We show that deep neural networks (DNNs) for object recognition, which lack spatial grounding, contain numerous units resembling place, border, and head-direction cells. Strikingly, even untrained DNNs with randomized weights contained such units and support decoding of spatial information. Moreover, when these "spatial" units are excluded, spatial information can be decoded from the remaining DNN units, which highlights the superfluousness of cell types to spatial cognition. Now that large-scale simulations are feasible, the complexity of the brain should be respected and intuitive notions of cell type, which can be misleading and arise in any complex network, should be relegated to history.

eLife assessment

This study uses a deep neural network approach to challenge the role of spatially selective neurons like place, head or border cells for position decoding. The findings are **important** as they suggest that such functional cell types may emerge naturally from object recognition in complex visual environments, but are neither necessary, nor particularly critical for position decoding. However, direct evidence supporting this conclusion remains **incomplete**.

https://doi.org/10.7554/eLife.99047.1.sa4

1 Introduction

Spatial cognition encompasses our cognitive abilities to know where we are and navigate in complex environments, which includes self-localization, comprehension of environmental layouts, and efficient navigation toward distant goal locations. The prevailing view of the field attributes

these functions to neural machinery within the hippocampal formation which contain a diverse array of "spatial" cell types including place cells, head-direction cells, border cells, and grid cells (e.g., Hartley et al. 2014; Moser et al. 2008; Grieves and Jeffery 2017), and that these cells collectively form the foundation of an internal cognitive map of space that underlie our spatial abilities (O'Keefe and Dostrovsky, 1971; Taube et al., 1990; Lever et al., 2009; Hafting et al., 2005). As such, a substantial proportion of the field focuses on characterizing particular functional "cell types" based on neural activity patterns in these regions. However, these "spatial cells" were identified due to their intriguing firing patterns that piqued the interest of neuroscientists and later defined based on subjective criteria – place cells appear to encode particular locations, border cells respond to borders, head-direction cells are sensitive to a heading direction, grid cells exhibit firing fields in a grid-like pattern. In practice, the field's initial fascination with these cells has led to a stubborn tendency to focus on "cell type" classification and overlook the fact that the criteria for distinct cell types seldom align seamlessly with empirical data and often exhibit mixed selectivity across task and environmental variables (e.g., Hollup et al. 2001; McKenzie et al. 2013; Dupret et al. 2010; Eichenbaum 2015; Grieves and Jeffery 2017; Latuske et al. 2015; Tang et al. 2015), raising questions about their precise roles, if any.

Could these cells simply arise from any systems with complex computations, unrelated to processes specific to spatial or navigation? There are strong hints that "spatial" cells may arise simply from domain-general learning algorithms that can lead to both concept and spatial neural representations (Mok and Love, 2019, 2023; Stachenfeld et al., 2017; Whittington et al., 2020). Indeed, models show that spatial firing patterns can arise from factors entirely unrelated to spatial cognition, such as sparseness constraints (Kropff and Treves, 2008; Franzius et al., 2007a,b).

Another question is whether these cells are particularly important for spatial cognition. Research has shown that "spatial" cells may not be functionally more critical than those with other cells with multiplexed or hard-to-interpret firing patterns for spatial localization (Diehl et al., 2017) or for explaining neuronal response profiles in the medial entorhinal cortex (Nayebi et al., 2021). Even lesion studies in the hippocampal formation do not consistently result in specific navigation impairments (Hales et al., 2014; Whishaw and Tomie, 1997; Whishaw et al., 1997).

Considering that both biological and artificial "cells" rarely perfectly align with human-defined criteria with many cells exhibit "uninterpretable" firing patterns, one unsettling possibility is that the neuroscience may have inadvertently constrained its pursuit by fixating on the search for interpretable "cell types". By adopting this top-down, perhaps näıve approach, the field may be investing a large proportion of its resources in a potentially fruitless quest, ignoring other explanations for the emergence of spatial cells and the foundations of spatial cognition.

A harsh appraisal of the field is that it looks for whatever cell type superficially reflects the answer sought. For example, how do animals localize themselves? The naive answer is that there must be place cells. Setting aside this answer provides zero insight into how such a cell came to be, there is no reason for the brain to perfectly align with our intuitions and, equally problematic, cells with these properties might arise but not serve the top-down function neuroscientist ascribe to them. Ascribing interpretable functions to cells and naming them may be good for neuroscience careers, but is it good for neuroscience?

In this contribution, we offer an alternative to the prevailing notion that cells with readily interpretable receptive fields in regions like hippocampus underlie spatial systems. We propose that what we commonly refer to as "spatial cells" might be inevitable derivatives of general computational mechanisms, and are in no way intrinsically spatial. Our results demonstrate that these cells could manifest within any sufficiently rich information processing system, such as

perception. Remarkably, not only do we find spatial signals within a non-spatial system but also demonstrate that these purported signals, ostensibly representing spatial knowledge, do not appear to have a privileged role for tangible downstream tasks related to spatial cognition as previously claimed.

To evaluate these possibilities, we turned to deep neural networks (DNNs) that were information processing systems devoid of components dedicated to spatial processing. Specifically, we analyzed deep learning models of perception with hundreds of millions of parameters. DNNs trained on natural images have achieved human-level performance in recognizing real-world objects in photographs (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). These models exhibit strong parallels with representations in the primate ventral visual stream (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Eickenberg et al., 2017; Zeman et al., 2020). Our investigation aims to determine whether processing egocentric visual information within such complex – yet non-spatial – models can lead to brain-like representations of allocentric spatial environments, and whether these representations are essential for spatial cognition. Additionally, we considered whether untrained DNNs (absent any experiences) can account for the emergence of spatial cells and spatial cognition.

To evaluate functional spatial information in perceptual DNN models, we began by creating a three-dimensional virtual environment reminiscent of a laboratory used in animal studies in neuroscience (**Fig. 1**). In this virtual setting, an agent randomly forages in a two-dimensional square area, much like an animal would explore an enclosure. The agent "sees" images of the room within its field of view over many locations and heading directions. Images from the first-person perspective are processed by a DNN, and we assess whether the internal representations of the model contain various kinds of spatial knowledge. We adopted a decoding framework which parallels experimental techniques employed by neuroscientists. Specifically, we trained linear regression models using various levels of representations (unit activations from visual inputs) generated by DNNs of different architectures to decode navigation-relevant variables, including self-localization, heading direction, and distance to the closest wall on locations and heading directions they have not encountered before. To test whether DNN units carry information like "spatial" cells in the brain, we classified and visualized model units based on standard criteria for place, direction, and border cells (Tanni et al., 2022; Banino et al., 2018). To evaluate whether these cells play a privileged role in spatial cognition, we excluded units classified as spatial units based on traditional spatial cell criteria and assessed their spatial knowledge.

To foreshadow our findings, we discovered that DNNs, regardless of their architectural variations, representation levels, or training states, possess noteworthy proficiency in allocentric spatial understanding. Spatial variables including location, heading direction and distance to boundaries can be accurately decoded from the DNN models. Additionally, a substantial number of DNN units exhibit spatial firing patterns that align with the criteria used in neuroscience to classify place, head-direction, and border cells. Notably, a significant portion of these units employs mixed-selectivity and conjunctive coding, akin to characteristics found in hippocampal neurons (Eichenbaum, 2015). We further revealed that excluding units meeting the classical criteria of spatial cell types has minimal impact on spatial decoding performance, reinforcing the perspective that the code underpinning spatial cognition is more distributed and diverse than previously thought, relying on neural population coding rather than individual and specialized cell types (Hebb, 1949; Eichenbaum, 2018; Ebitz and Hayden, 2021). In summary, our analyses suggest that "spatial" cell types may be inevitable and superfluous in complex information processing systems.

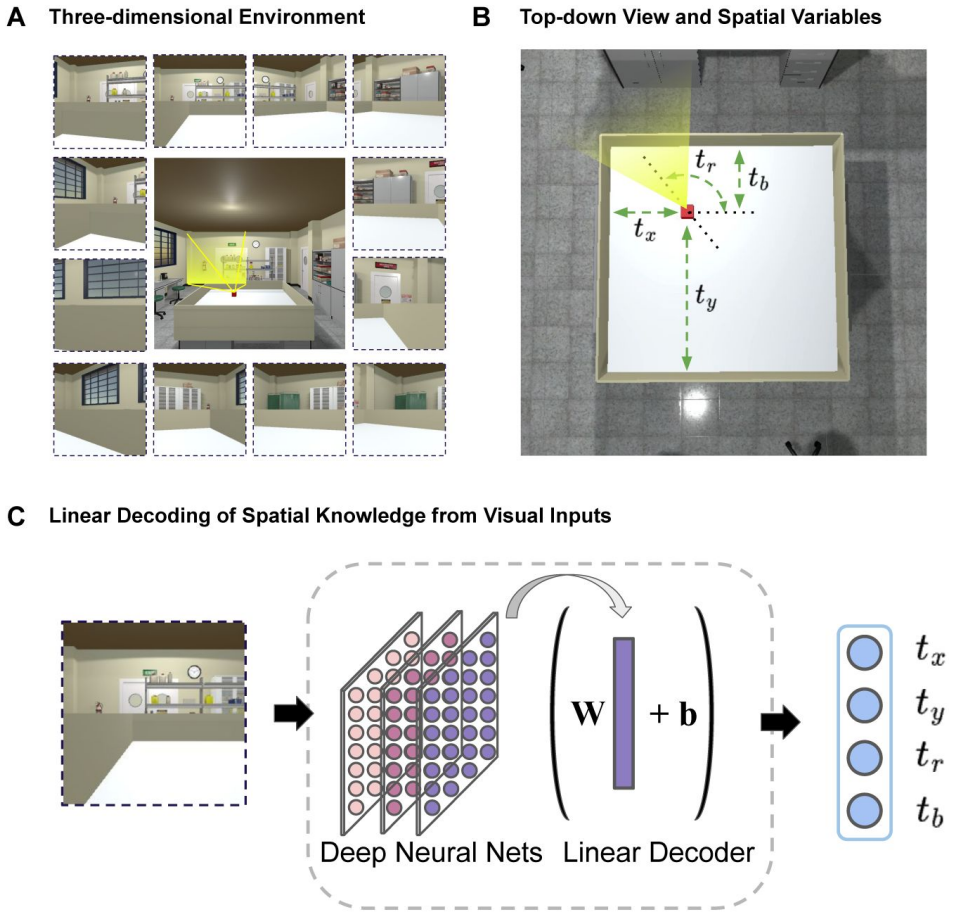


Figure 1

Assessing spatial knowledge in non-spatial perception systems using a linear decoding approach in a virtual environment.

(A) A three-dimensional virtual space is created to resemble a realistic laboratory environment with a variety of visual features. An agent moves randomly in a two-dimensional area which is within a three-dimensional space and processes first-person views of the environment. Central image shows the threedimensional environment. Surrounding images are example views taken by the agent at different locations and heading directions. (B) Top-down view of the area where the agent can explore. We define spatial knowledge of the agent with four values. The agent's location is denoted by the Cartesian coordinates t_x, t_y. The agent's heading direction is denoted by the angle t_r. The distance between the agent and the nearest wall is t_b. (C) Individual views are processed by perception models (deep neural networks of object recognition). We train linear regression models with various levels of internal representations from these networks to assess spatial knowledge related to self-location (t_x, t_y), heading direction (t_r) and distance to the closest wall (t_b).

2 Results

2.1 Spatial Knowledge through Non-spatial Systems

We predict that spatial knowledge can arise in computational systems with sufficient complexity absent a spatial grounding. We test this hypothesis by analyzing the percepts of an agent freely moving in a virtual environment. The percepts take the form of viewpoints fed into a deep neural network with its millions of weights either randomly initialized or trained for object recognition, a non-spatial task. From various network layers, we attempted to decode spatial information such as location, from this complex computational system.

To assess information latent in complex networks, we trained linear regression models to decode the agent's location, head-direction and distance to the closet border using fixed unit activity extracted at various levels of the networks. While sequential information is important for spatial systems, our decoding framework does not utilize sequential information of sampled views during training, making it a stricter test for the central claim that spatial knowledge can form out of general information processing. We expect that one's location, heading-direction and distance from border may be decodable from visual information alone due to the inherent continuity of perception across spatial locations and/or directions. All results reported here are tested on out-of-sample locations and views using the perception model VGG-16 (Simonyan et al., 2014), unless noted otherwise. For details referring to training and testing procedure, see Methods. For results of other models such as Resnet-50 (He et al., 2016) and Vision Transformers (ViT; Dosovitskiy et al. 2020), see Appendix.

We find that spatial knowledge, namely one's own location (spatial coordinates; **Fig. 2B**, left panel), heading direction (**Fig. 2B**, middle panel) and distance to the nearest wall (**Fig. 2B**, right panel) can be successfully decoded from unit activity in a perception model. We quantified the degree of spatial knowledge by the decoding error, that is, the deviation of the model's prediction of its own location, heading direction, or distance to the nearest wall relative to the ground truth (i.e., the physical unit of space or angle in the virtual environment). We normalized decoding error with respect to the maximal unit length of the moving area (see Methods). For example, a normalized error of 0.05 for location prediction means the decoded location is 5% off from the true location relative to the maximal width/depth of the entire moving area. As more locations and views are sampled for training the agent (x-axis), decoding performance improves. The agent demonstrates impressive decoding performance across all spatial knowledge tests, even when trained on just 30% of randomly selected independent locations and associated viewpoints within the entire space. The moving area encompasses 289 distinct locations, each featuring 24 evenly spaced views covering the full 360-degree spectrum (for more details, please see Methods).

Decoding performance varied across layers of the network where deeper layers typically achieved lower decoding error (apart from the penultimate layer in VGG-16, i.e., fc2). Crucially, decoders across all sampling rates and levels of representation show substantially better performance than chance (green and blue lines), determined by strategies invariant to visual inputs (i.e., random or predicting the center of possible choices; see Methods). Our results demonstrate there is sufficient spatial information to achieve remarkably low decoding error in a perception model based on a few visual snapshots of the virtual space, which could be further improved through the integration of information across views in a perception-based navigation system.

To test the generality our claim that spatial knowledge can arise from complex computational systems irrespective of their architectural variations or training states, we examined several computational systems including deep convolutional neural networks (DCNN) and Vision Transformers (ViT). We evaluated these models in both their pre-trained form (optimized for

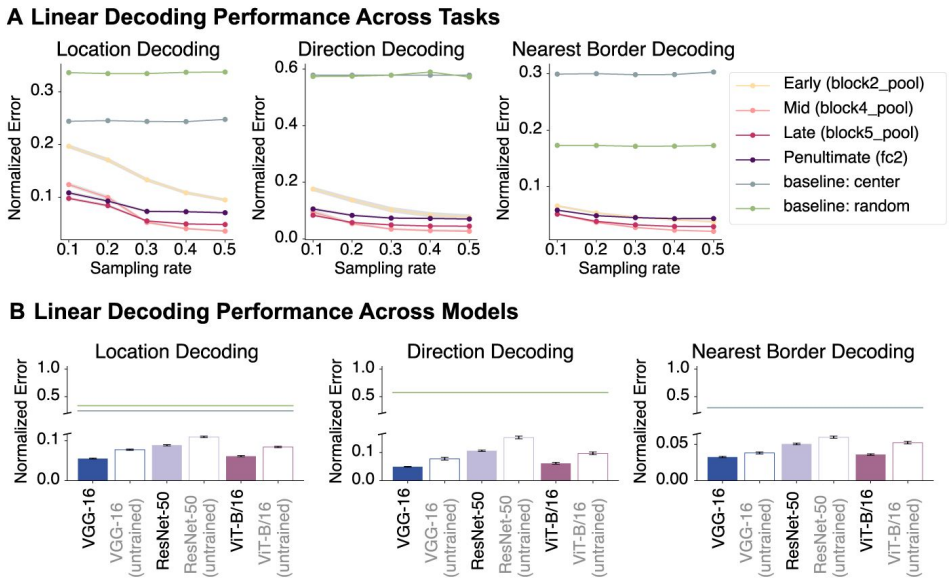


Figure 2

Perception models absent a spatial basis possess extensive spatial knowledge.

(A) Decoding performance across tasks and model layers. Across three tasks, mid-to-late layers exhibited lower decoding errors compared to early layers and the penultimate layer of VGG-16 (fc2). All layers outperformed chance (green and blue lines). (B) Decoding performance across a number of deep neural network architectures (penultimate layer; see Appendix for full results), including convolutional networks and vision transformers. All pre-trained and untrained models outperformed baseline measures. Error is in normalized virtual environment units (see Methods). See full results across models, layers and sampling rates in the Appendix. Shaded areas in B and error bars in C represent 95% confidence intervals of the mean decoding error (bootstrapped across locations).

visual object recognition), and in an untrained state (randomly initialized parameters), following the same decoding approach described earlier. Here, we present the results for the penultimate layer representation and uniformly sampled 30% of all locations and their views for training the decoder, but similar results were observed across layers (for full results across various models, layers, and sampling rates, see Appendix.). We found that all models possessed latent spatial knowledge, exhibiting lower errors far below chance (**Figure 2C**). Notably, this was true even in untrained networks. While one might contend that deep convolutional neural networks, with their reliance on local convolution operations, inherently possess a predisposition for extracting spatial knowledge from visual information, the revelation that an untrained ViT—comprising nothing more than a hierarchy of fully-connected layers (i.e., self-attention) and non-linear operations—can achieve superior decoding performance illustrates the inevitability of computational complexity in facilitating the extraction of spatial information necessary for cognitive spatial processes, even in non-spatial systems.

2.2 "Spatial Cells" Inevitably Arise in Complex Computational Systems

How was it possible to decode spatial information from non-spatial networks? The field's common conception is that cells exhibiting spatial firing profiles play a pivotal role in supporting spatial cognition, as these profiles appear intuitively useful to spatial tasks such as navigation. Could spatial cells be responsible for the impressive decoding of spatial information within a perception system like we have shown? If a perception system devoid of a spatial component demonstrates classically spatially-tuned unit representations, such as place, head-direction, and border cells, can "spatial cells" truly be regarded as "spatial"? Might they be inevitable derivatives of any complex system? If so, do they drive spatial knowledge, or are they in fact superfluous in spatial cognition?

To determine if the model's spatial knowledge is primarily supported by spatially-tuned units like those found in the brain, we classified every hidden unit in each model based on criteria used to identify spatial cells in neuroscience for place cells (Tanni et al., 2022), head-direction cells (Banino et al., 2018), and border cells (Banino et al., 2018), respectively (see Methods). The composition of cell types across layers is shown in **Fig. 3A**. Contrary to the predominant view that "spatial cells" are unique properties of spatial systems and are grounded in space-related tasks, we found many units in the non-spatial model VGG-16 that satisfied the criteria of place, head-direction, and border cells (see Methods for criteria; for other models, see Appendix). The majority of units show mixed selectivity irrespective of layer depth. Notably, a significant proportion of units show both place and directional tuning (P+D), matching a common observation in the literature (e.g., McNaughton et al. 1991; Tang et al. 2015; Grieves and Jeffery 2017). We selected example units and plotted their spatial activation patterns in the two-dimensional virtual space (irrespective of direction), and their direction selectivity in polar plots (**Fig. 3B**-**3E**; for more examples, see Appendix). These examples show spatially-tuned units without directional tuning that match hippocampal place cells (**Fig. 3B**), direction-selective units with strong direction selectivity accompanied by minor spatial selectivity matching head-directional cells (**Fig. 3C**), and units with boundary cell-like tuning (**Fig. 3D**). There were also many units that exhibited both strong place and directional tuning (**Fig. 3E**). These results support our theory that "spatial cells" might arise in any computational systems, even in systems designed for nonspatial tasks (e.g., object recognition). Consistent with our view, we found no clear relationship between cell type distribution and spatial information in each layer. This raises the possibility that "spatial cells" do not play a pivotal role in spatial tasks as is broadly assumed.

2.3 "Spatial Cells" Are Superfluous in Spatial Cognition

We have establishd that"spatial cells" can arise in non-spatial systems. Here, we consider whether units with spatial properties are necessary to decode spatial information. It may very well be that spatial units, akin to place cells, might not only arise in complex (including random) networks but

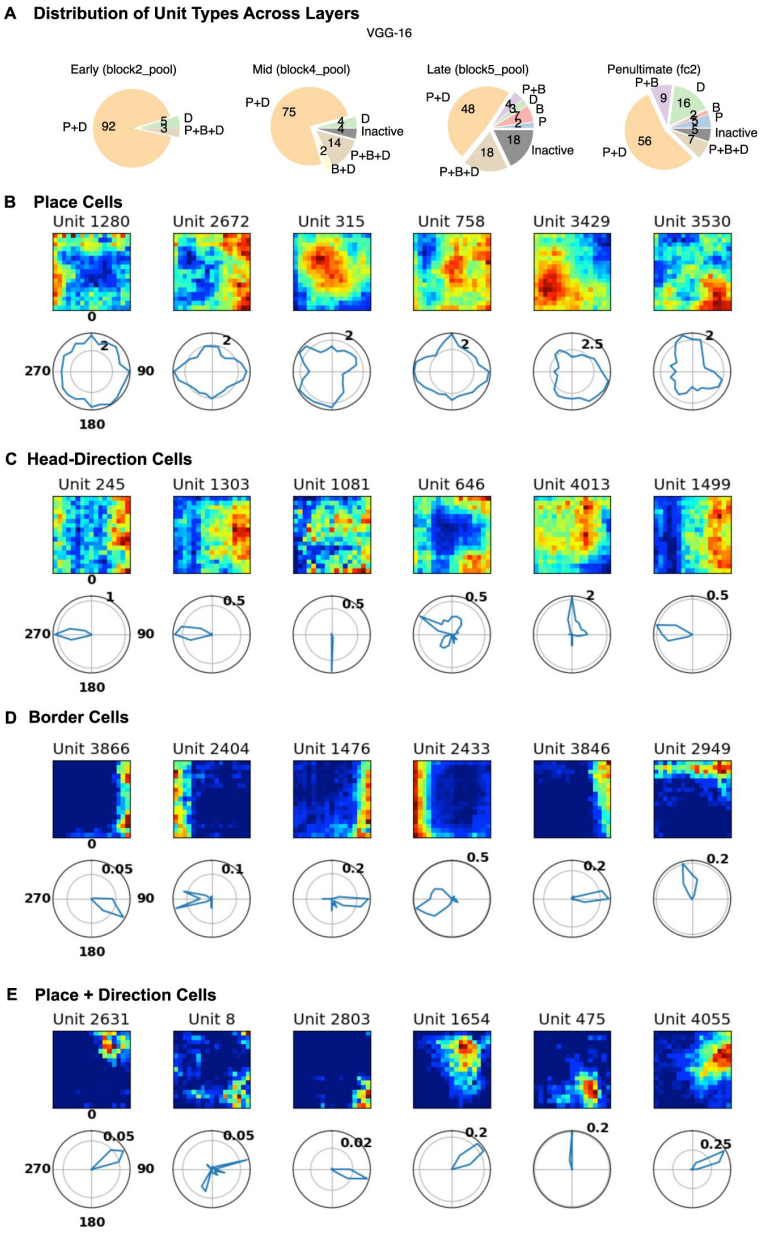


Figure 3

Representations of perception models developed for object recognition exhibit typical spatial cell-like firing profiles.

Active units were classified across model layers based on standard criteria used to identify place cells (P), head-direction cells (D), and border cells (B). Example units from VGG-16 (see Appendix for more examples from other models). (A) Pie charts illustrating the proportion of "spatial" cell types identified in DNNs, including units that are inactive. Many units satisfied the criteria for place, head-direction and border cells irrespective of layer depth. Many units exhibited mixed selectivity, with a significant amount of units displaying strong place and directional tuning. (B-E) Spatial firing profiles of model units in spatial activation maps and polar plots . For activation maps, each unit's activation was plotted at each location in the two-dimensional area irrespective of heading direction. For direction selectivity, polar plots show the average activity of the activity map across location at a given angle, which reflects the tuning magnitude to each heading direction). (B) Example place cell units that show strong spatial selecitivity with little direction selectivity. (C) Examples of head-direction cell units that show strong direction selectivity but weak location selectivity. (D) Examples of border cell units that respond strongly to boundaries of the environment. (E) Examples of mixed-selective place and head-direction cell units with strong spatial and directional tuning. Examples presented are from VGG-16. See Appendix for more examples.

are also superfluous to spatial cognition. To test this possibility, we performed a systematic exclusion analysis to assess whether model units that exhibit the strongest spatial firing properties contribute to spatial knowledge. First, we scored and ranked each unit classified as place, head-direction, and border cell units (see Methods), and then re-trained linear decoders without the top n units of a specific cell type and evaluated the model's corresponding spatial knowledge of the environment. That is, we test model's location, heading direction and distance to closet border decoding ability by excluding place, head-direction and border cell units, respectively. We repeated this procedure with a progressively higher exclusion ratio (see Methods).

In line with our hypothesis, excluding spatial units in the models that scored highest on each of the corresponding criteria (place field activity, number of place fields, directional tuning, and border tuning), had minimal effect on decoding performance even with a large proportion of the highest ranked spatial units being excluded (**Fig 4B**, top panel). To assess whether any detriment was observed was significant and specific to excluding highly-ranked spatial units, we randomly excluded the same number of units without regard to the spatial score (**Fig. 4B**, bottom panel) and observed a similar pattern across four exclusion scenarios, meaning that highly-tuned spatial units contributed no more to spatial cognition than a random selection of units.

Finally, we performed an additional analysis where we excluded model units that contributed most to each of the tasks. Specifically, we identified each unit's importance for each task, based on the magnitudes of the coefficients learned by the linear decoders (trained in **Section 2.1**), excluded the top n units, and re-trained linear decoders for each respective task using the remaining units. This approach had a greater impact on decoding error compared to spatial unit criteria or random exclusion, with performance suffering more as the exclusion ratio rose (see **Figure 4C**). Nevertheless, decoding was still possible at high exclusion ratios, suggesting that spatial information is widely distributed across network units.

3 Discussion

The prevailing perspective in neuroscience conflates spatial knowledge and the variety of spatial cognitive abilities in animals with "spatial cells" in the hippocampal formation. In this work, we proposed an alternative perspective where "spatial cells" can simply arise in non-spatial computational systems with sufficient complexity, and that these units may, in fact, be by-products of such systems absent any leading role in spatial cognition.

Deep neural networks (DNNs) for object recognition served as an ideal testbed for our hypotheses. Object recognition DNNs, designed to generate translation-invariant visual features for successful categorization, lack inherent spatial elements. Unlike spatial navigation systems (e.g., Banino et al. 2018; Cueva and Wei 2018), DNNs in object recognition do not purposefully incorporate information about velocity, heading direction, or event sequences. Exploring various DNN architectures for object recognition allows a comprehensive examination of the connection between spatial representations and non-spatial complex systems.

We applied DNNs to a three-dimensional virtual environment to simulate an agent foraging in a two-dimensional space, akin to how an animal explores an enclosure. Detailed analyses revealed that DNN's representations contained sufficient spatial information to decode key variables, despite these DNNs being non-spatial perceptual systems. We found that many DNN units passed established criteria for classification as place, head-direction, and border cells, suggesting such units can be easily found in various non-spatial computational systems. Furthermore, excluding the "spatial" units had minimal impact on spatial decoding tasks, suggesting a non-essential role for spatial cognition. These results call into question utility of labelling these cell "types".

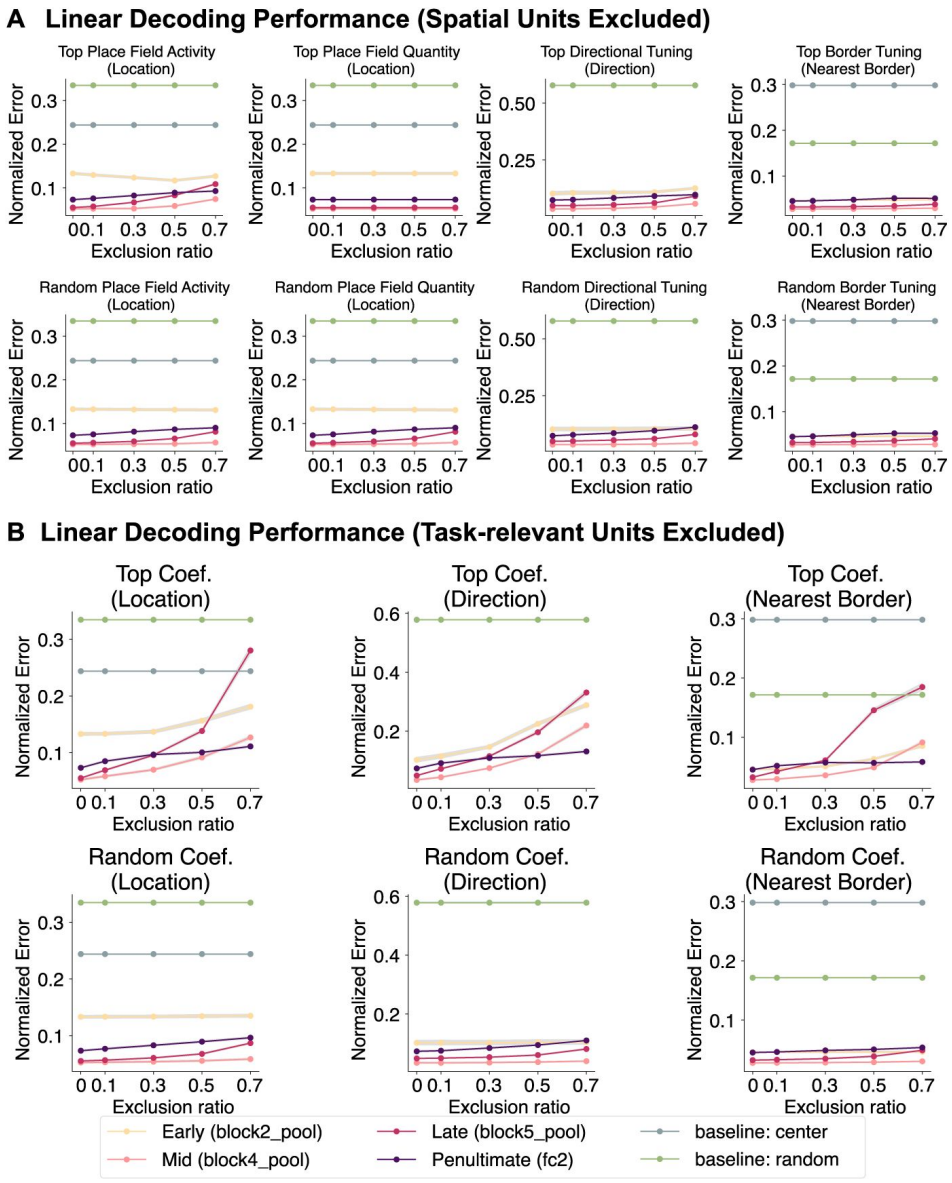


Figure 4

"Spatial" cells do not play a privileged role in spatial cognition.

Exclusion analyses showed that units exhibiting traditional spatial firing profiles do not form the basis of the model's spatial knowledge. (A) Units in each layer were ranked based on standard criteria for place (maximum place field activity, number of place fields), head-direction (strength of directional tuning), and border (strength of border tuning) cells. As more highly-ranked spatial units were excluded, the overall decoding performance remained relatively stable across all tasks (top). Excluding an equivalent number of units randomly yielded similar performance (bottom). (B) Model units' contribution to spatial knowledge based on their contribution to decoding performance (magnitude of regression coefficients). More highly-ranked task-relevant units excluded results in a marked deterioration of decoding performance by decoders trained on the remaining units (top). Randomly excluding an equivalent number of units had minimal impact on performance (bottom). Results here are from VGG-16. For other models, please refer to the Appendix.

We considered object recognition models from two prominent DNN families: deep convolutional networks, drawing inspiration from the mammalian visual system (Fukushima, 1980), and visual transformers, adapted from the Transformer architecture originally designed for natural language processing (Vaswani et al., 2017). Spatial information was readily decoded from models from both families, including untrained models with randomly initialized weights. The decoding results from the untrained visual transformer are particularly striking because transformers incorporate minimal inductive biases, such as constraints that respect spatial proximity (e.g., convolutions). In summary, complex networks that are not spatial systems, coupled with environmental input, appear sufficient to decode spatial information.

Our work raises the question of whether focusing on single cells and their "type" is the correct approach to understanding spatial cognition, or indeed any cognitive function. The field has maintained a persistent inclination towards identifying cell types based on firing profiles that are subjectively interpreted to be useful for the behavior of interest (Hartley et al., 2014; Moser et al., 2008; Sarel et al., 2017; Tsao et al., 2018; Høydal et al., 2019; Grieves et al., 2020; Ormond and O'Keefe, 2022). This is likely due to the historical backdrop in neuroscientific research where pioneers in visual neuroscience discovered and named individual cells in V1 by observing their firing profiles (Hubel and Wiesel, 1959) which won the Nobel Prize (1981) and marked a foundational era for the field. This paradigm, led to discoveries of place and grid cells (Nobel Prize in Physiology or Medicine, 2014), and the "Jennifer Aniston" neuron or "concept cells" (Quiroga et al., 2005), which were intuitively compelling and attractive as an apparent explanation. These cells may in fact play a role, but our work suggests that this must be critically assessed rather than assumed, as they may not play a privileged role compared to other cells in the population, and could even be superfluous for the context at hand. The general assumption that the simplicity and interpretability of firing profiles somehow lends support to a crucial role of these cells should be questioned. With a growing recognition of the imperative to move beyond mere cell type categorization, as underscored by Olshausen and Field (2006) on the insufficient characterization of a majority of V1 cells, there is a new shift to focus on a broader perspective where neural assemblies or populations form the fundamental computational unit (Hebb, 1949; Eichenbaum, 2018; Ebitz and Hayden, 2021).

We pose a broader question: do our preconceptions of how complex systems *should* work hinder our aim to understand the brain's inner workings? Notably, spatial representations conventionally linked to the hippocampal formation have been unveiled in sensory areas (Saleem et al., 2018; Long et al., 2021; Long and Zhang, 2021). While prior attributions pointed to modulatory signals from the hippocampus, our study raises concerns about potential biases stemming from preconceived notions and an underestimation of the complexity inherent in the sensory system—it may be shouldering more extensive responsibilities than initially presumed. Work in other domains are reaching similar conclusions (cf. McMahon and Isik 2023).

One upshot of our contribution is that the pursuit of identifying and cataloging cell types aligned with subjective notions of how the brain works might impede scientific progress. While searching for cell types was a reasonable strategy in an era before it was straightforward to conduct large-scale simulations, this quest should now be relegated to history. Otherwise, apparent discoveries of cell types are likely to reflect the activity of broad classes of complex networks that implement a variety of functions or none at all in the case of the random networks we considered. It is far too easy for neuroscientists, including computational neuroscientists who design their models to manifest a variety of cell types, to fool themselves when the underlying complexity of the systems considered is not taking into account.

4 Methods

4.1 Virtual environment

We setup our virtual environment using the Unity3D Engine (editor version: 2021.3.16f1; Silicon). The three-dimensional laboratory environment was adapted from an existing template purchased on the Unity Store (3DEverything, 2022). Our virtual agent's movement is constrained in a two-dimensional square placed inside the three-dimensional space, much like a moving area for an animal in real-world experiments. The agent randomly moves around the square area and captures first-person pictures of the environment across locations and heading directions.

Environment specifications

The specifications for the environment, measured in Unity units, are as follows: the lab room has a height of 3.17 units, a width of 6.29 units, and a depth of 10.81 units. The moving area, which is a distinct part of the environment, has a height of 0.25 units, a width of 2 units, and a depth of 2 units. Additionally, the moving area is located at a specific location whose relative distances from the center to the lab room boundaries are: 3.59 units to the left wall, 5.35 units to the right wall, 5.55 units to the bottom wall, and 2.80 units to the front wall.

The agent is only allowed to move within the squared moving area. For simplicity, we discretize reachable locations by the agent in the area as a 2D grid where the agent can only move between points on the grid. We use a 17×17 grid (289 unique locations).

Visual input

To collect visual inputs from the virtual environment, the agent randomly moves along the specified n-by-n grid in the squared area. We denote l as the total number of unique locations on the grid. The grid is evenly spaced. At each location, the agent turns around and captures m number of frames, each separated by a fixed angle (e.g., 15 degrees per frame). Each frame obtained by the agent is processed by a DNN (up to a given layer). Resulting outputs at the same location are considered independent data-points, each forming a long 1D vector. Repeating this procedure for all locations yields a M-by-F data matrix $\mathbf{X}$ where M = l × m and F is the number of feature variables. Each feature variable represents a neuron on the output layer of the DNN.

4.2 Spatial decoding

To evaluate whether representations in the deep neural networks (DNN) optimized for object recognition on natural images encode spatial information, we set up a series of spatial decoding tasks where we use fixed network representations across various hidden layers to decode spatial location (i.e., coordinates), head-direction (i.e., rotation degrees) and distance (i.e., Euclidean) to the nearest border using visual inputs of views from unvisited locations and/or directions. We formulate spatial decoding as a prediction problem where we fit a set of linear regression models on representations of visual views to predict location, direction and distance to nearest borders. We define the nearest border to the agent as the border that has the shortest Euclidean distance to the agent's location regardless of the agent's heading direction.

Decoder training and testing

To fit and test spatial decoders, we randomly sample locations in the environment (subject to a sampling rate). When a location is sampled for training, we consider views of different rotations as independent data-points. For example, a location decoder is fit to predict a vector $\mathbf{t_i} = [t_x, t_y, t_r,$

t_y, t_r, t_b] where t_x, t_y are the x, y coordinates, t_r is the direction, t_b is the shortest distance to a border, given a view respectively. For the entire training set ($\mathbf{X_i}$, $\mathbf{T_i}$) where $\mathbf{T_i}$ = {$\mathbf{t_1}$, $\mathbf{t_2}$, ..., $\mathbf{t_i}$}, we can write the decoder as

$$\mathbf{T_i} = \mathbf{X_i}\mathbf{W_i} + \mathbf{b_i} \quad (1)$$

where $\mathbf{W_i}$ are the coefficients and $\mathbf{b_i}$ are the intercepts. Decoder weights (both coefficients and intercepts) are optimized to minimize the mean squared error between the true and predicted targets.

Feature representation and selection

We consider a number of representations for visual inputs. Given a DNN model, we train separate decoders using representations across different layers. We apply L_2 regularization on the decoder weights to reduce overfitting. Additionally, we consider different exclusion strategies (see Sec. 4.4) where unit representations are selectively removed based on their firing profiles (e.g., the number of place fields a unit contains).

Performance measure

We report the normalized decoding error on the heldout set of location, head-direction and nearest border decodings for each representation at a given sampling rate. The normalized decoding error reflects the deviation of the model's prediction on location, heading direction or distance to the nearest wall relative to the ground truth (i.e., in terms of physical unit of space or angle in the virtual environment) and normalized with respect to the maximal length of the moving area (2 units in both depth and width). To reflect variance of the average error, we compute a two-sided 95% confidence interval of the average using a bootstrapping approach over locations.

Baselines

We compare decoding performance achieved by our agent to baselines where the agent decodes location, head-direction and distance to border based on fixed policies independent to visual inputs. Specifically, we establish two baselines where the agent simply either predict randomly within the legal bound or predict the centre position of the room (when predicting location), predict 90 degrees (when predicting rotation) and predict the distance from the centre of the room (when predicting nearest border distance).

4.3 Profiling spatial properties of model units

To gain a more intuitive understanding of the kinds of model units that are most useful for spatial decoding tasks, we profile all model units from the layers we consider based on their spatial patterns of firing. We consider the most common cell types that are found to encode spatial information as defined in the neuroscience literature, which include place cells, border cells and head-direction cells. For each cell type, we track a number of measures defined below.

Place cells

In our investigation, the identification of place-cell-like model units hinges on several key characteristics of unit firing. Firstly, we examine the number of place fields exhibited by each unit, defining a qualified field as one that spans 2D space ranging from 10 pixels to half the size of the environment, as outlined in Tanni et al. (2022). Additionally, we consider the maximal activation of each field, providing insights into the intensity and significance of unit firing.

Head-direction cells

Following Banino et al. (2018), the degree of directional tuning exhibited by each unit was assessed using the length of the resultant vector of the directional activity map. In our case, there is an activity map spanning entire 2D space for each direction. Vectors corresponding to each direction of an activity map were created:

$$r_i = [\beta_i \cos \alpha_i, \beta_i \sin \alpha_i]^T \tag{2}$$

where α and β are, respectively, the angle and average intensity (irrespective of spatial locations) of direction i in the activity map. These vectors were averaged to generate a mean resultant vector:

$$\vec{r} = \frac{\sum_{n=1}^{N} r_i}{\sum_{n=1}^{N} \beta_i} \tag{3}$$

and the length of the resultant vector calculated as the magnitude of'r. We used 24 angular directions uniformly spaced.

Border cells

Border score definition is based on Banino et al. (2018) where for each of the four walls in the square enclosure, the average activation for that wall, b_i, was compared to the average centre activity c obtaining a border score for that wall, and the maximum was used as the border-score for the unit:

$$b_s = max_{i \in \{1,2,3,4\}} \frac{b_i - c}{b_i + c} \tag{4}$$

where b_i is the mean activation for bins within d_b distance from the i-th wall and c the average activity for bins further than d_b bins from any wall (d_b = 3). Units with border score > 0.5 are considered border-like.

4.4 Spatial decoding with unit exclusion

To delve deeper into how different units within the DNN models contribute to the linear decoding of spatial information, we employ two complementary analyses. Model units are selectively excluded subject to criteria detailed below. Spatial decoders are re-trained on the rest of the model units following the same procedure as before (sec. 4.2).

Excluding units by spatial profile

In this first analysis, we selectively exclude units based on their firing profiles, which fall into the various spatial unit categories mentioned earlier. We initially rank all units in descending order according to their specific spatial measures (e.g., field count) and then systematically increase the exclusion rate, assessing the impact on the corresponding downstream tasks as we go. For example, we progressively exclude the top n% of units with the most place-like characteristics for varying values of n, examining the relationship between exclusion and performance of location decoding. This process is carried out separately for each unit type and their associated tasks we have defined, and we also include a control group where an equivalent number of units are excluded at random.

Excluding units by task-relevance

Our second exclusion analysis focuses on identifying the task-relevance of each model unit by utilizing previously learned decoders and their coefficient magnitudes. Essentially, units with larger coefficients are deemed more crucial for decoding, while those with smaller or zero coefficients are considered less relevant. Similar to the exclusion by spatial profile, we gradually increase the exclusion rate by selectively removing units based on the magnitude of their decoder coefficients. Additionally, we conduct a control analysis where an equal number of units are randomly chosen for exclusion. We then retrain linear decoders with the remaining units and assess their performance on downstream spatial tasks.

Data and code availability

The code for simulations and analyses are publicly available at *https://github.com/don-tpanic/Space*.

Acknowledgements

This work was supported by ESRC (ES/W007347/1), Wellcome Trust (WT106931MA), and a Royal Society Wolfson Fellowship (18302) to B.C.L., and the Medical Research Council UK (MC UU 00030/7) and a Leverhulme Trust Early Ca-reer Fellowship (Leverhulme Trust, Isaac Newton Trust: SUAI/053 G100773, SUAI/056 G105620, ECF-2019-110) to R.M.M.

Figure S.1

VGG-16 (untrained).

Linear decoders trained with representations from various layers of the untrained VGG-16 achieve low errors across sampling rates; though not as good as its trained version. Mid-to-advanced layers show superior performance than early layers. As more locations are sampled for training the linear decoders, overall decoding performance improves. All model layers can decode better than two visual-invariant baselines.

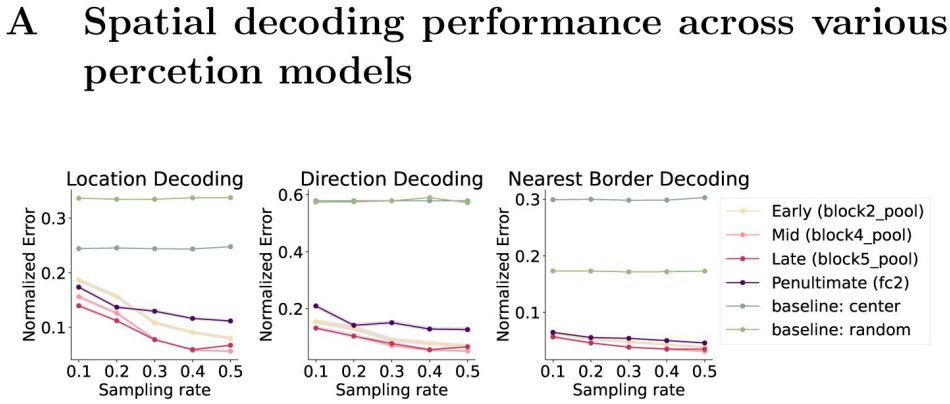


Figure S.2

ResNet-50 (trained).

Linear decoders trained with representations from various layers of the ResNet-50 pretrained on object recognition achieve low errors across sampling rates. Mid-to-advanced layers show superior performance than early layers. The penultimate layer decoding performance did not improve as much as the intermediate layers as more locations are sampled for training the linear decoders. All model layers can decode better than two visual-invariant baselines.

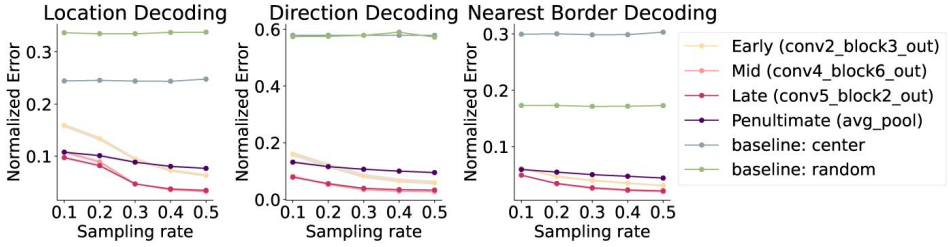


Figure S.3

ResNet-50 (untrained).

Similar to ResNet-50 pretrained on images, the untrained counterpart can effectively decode spatial knowledge related to location, heading direction and distance to borders. All layers decoder better than the baseline decoders which do not rely on visual signals of the environment.

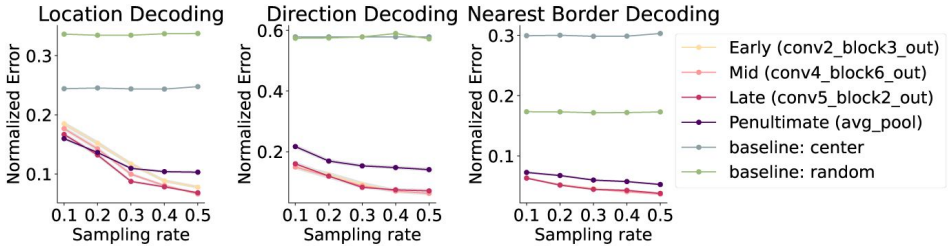


Figure S.4

ViT-B/16 (trained).

Linear decoders trained on different layers of the pretrained ViT model show very similar decoding performance. Overall the decoding performance is much better than the two baseline decoders which do not incorporate visual signals.

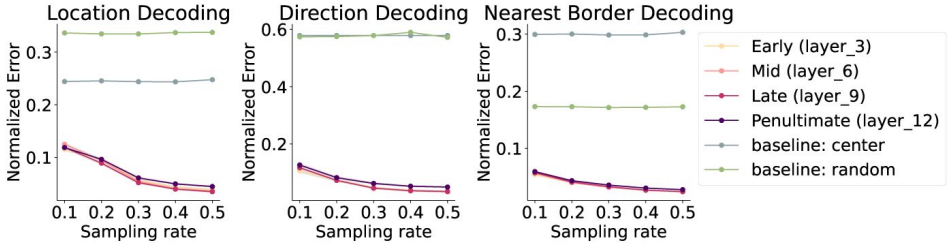
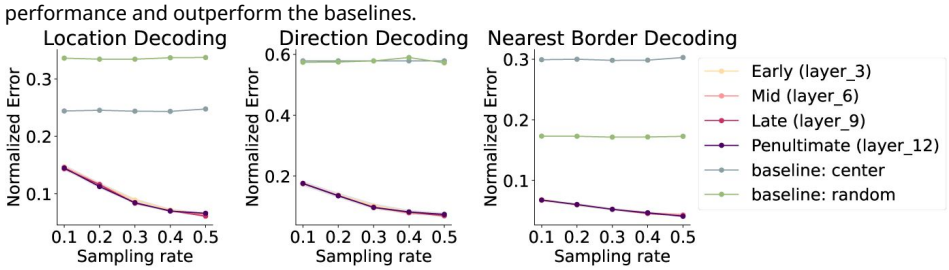


Figure S.5

ViT-B/16 (untrained).

Linear decoders trained on the untrained ViT model also achieve accurate decoding performance on location, heading direction and distance to the nearest border. Similar to the trained version, all layers considered in our analysis achieve comparable performance and outperform the baselines.



B Distribution of cell types of spatial units across models and layers

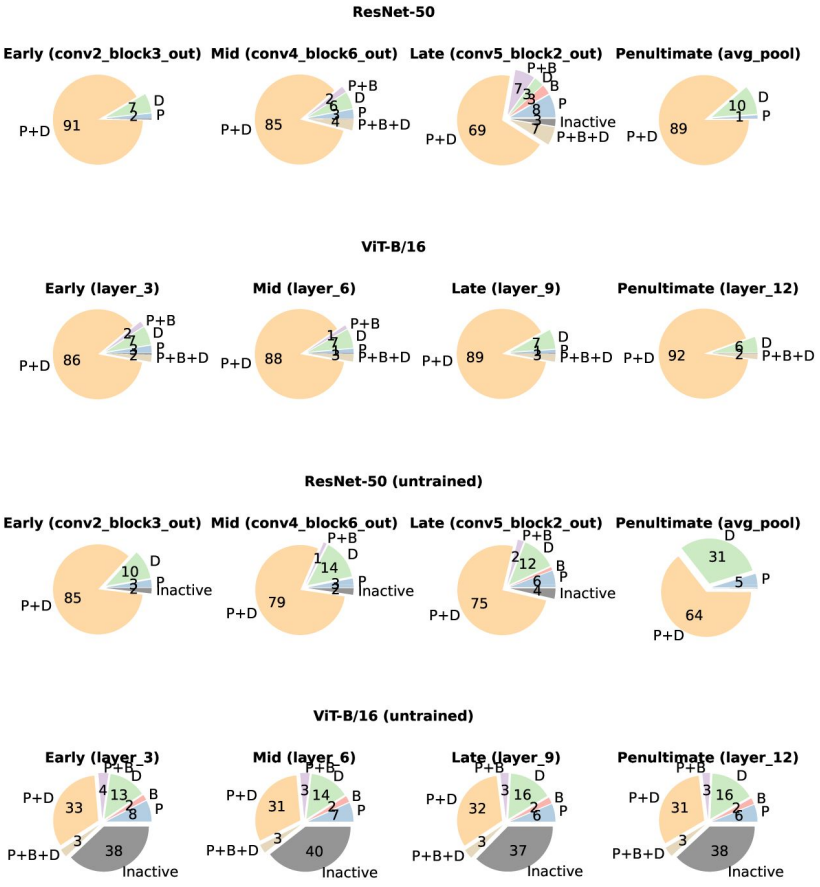


Figure S.6

Distribution of different spatial unit types across layers of perceptual models of object recognition.

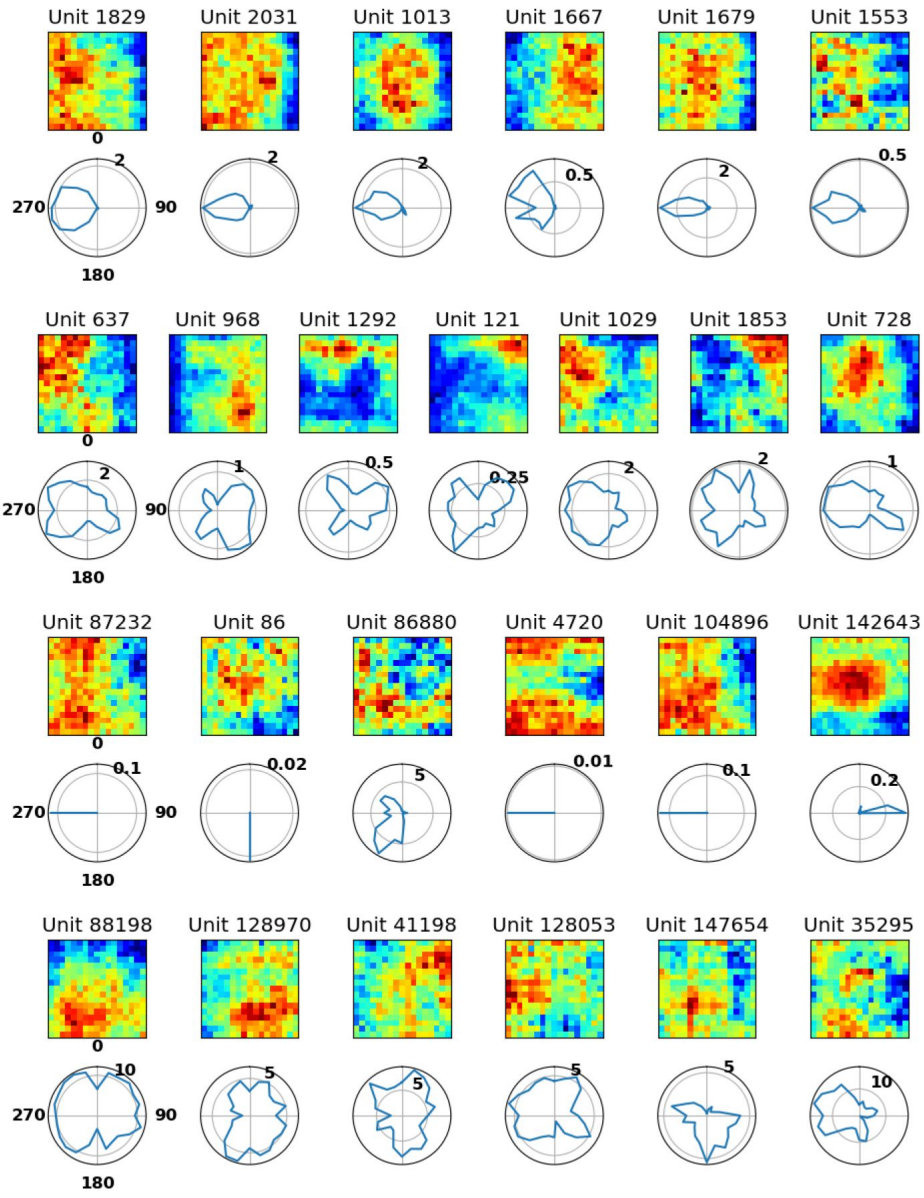


Figure S.7

Examples of model units exhibiting spatial characteristics.

Figure S.8

Excluding units with the strongest spatial profiles had minimal impact on spatial knowledge (ResNet-50).

D Spatial decoding performance under unit exclusion across various perception models

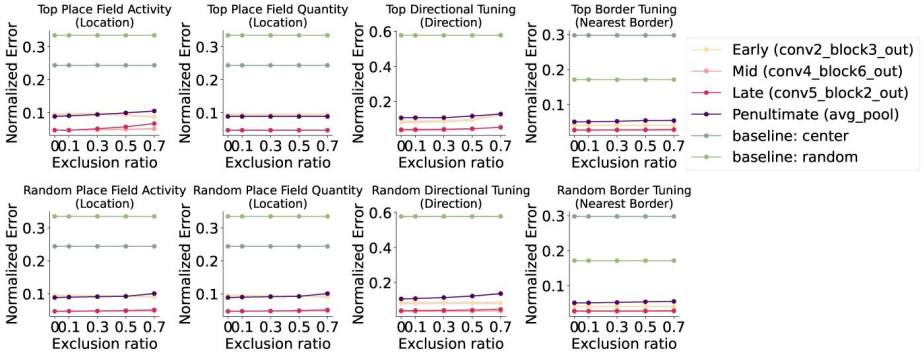


Figure S.9

Excluding units by task-relevance affects spatial decoding performance (ResNet-50).

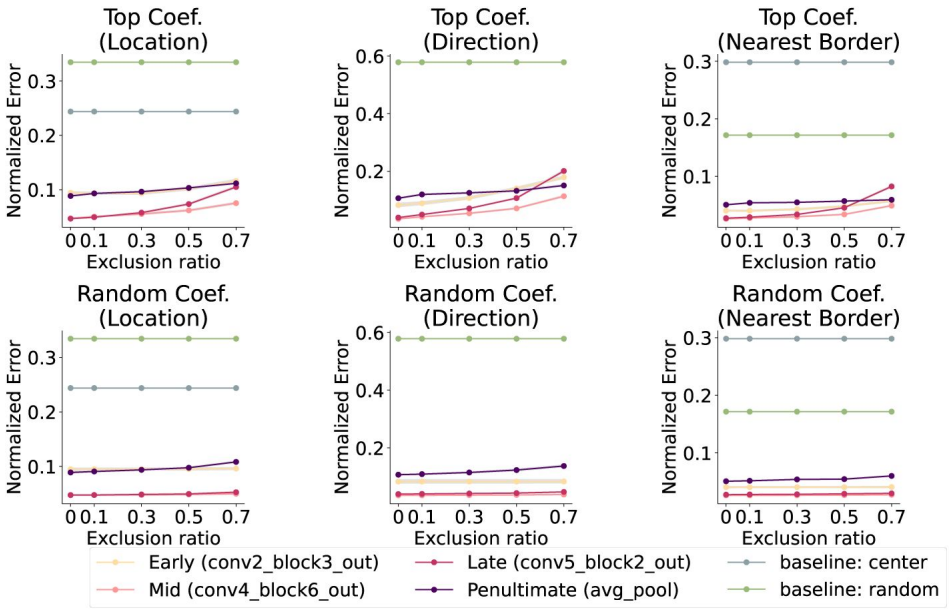


Figure S.10

Excluding units with the strongest spatial profiles had minimal impact on spatial knowledge (ViT-B/16).

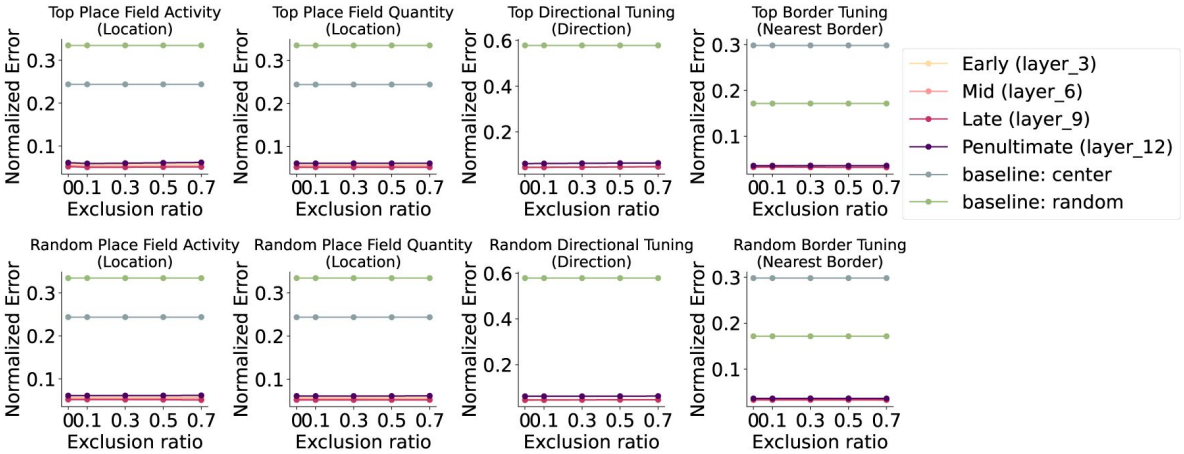
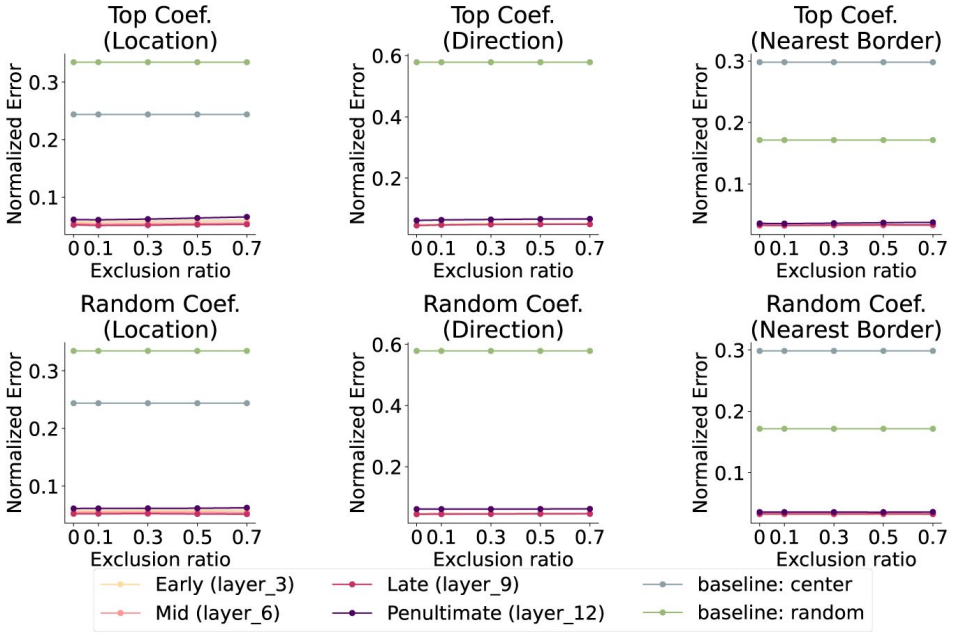


Figure S.11

Excluding units by task-relevance affects spatial decoding performance (ViT-B/16).



References

1. 3DEverything (2022) **Hospital Laboratory | 3D Interior | Unity Asset Store**

2. Banino A *et al.* (2018) **Vector-based navigation using grid-like representations in artificial agents** *Nature* **557**:429–433 https://doi.org/10.1038/s41586-018-0102-6

3. Cueva CJ, Wei XX (2018) **Emergence of grid-like representations by training recurrent neural networks to perform spatial localization** *arXiv*

4. Diehl GW, Hon OJ, Leutgeb S, Leutgeb JK (2017) **Grid and Nongrid Cells in Medial Entorhinal Cortex Represent Spatial Location and Environmental Features with Complementary Coding Schemes** *Neuron* **94**:83–92 https://doi.org/10.1016/j.neuron.2017.03.004

5. Dosovitskiy A *et al.* (2020) **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale** *arXiv* https://doi.org/10.48550/arxiv.2010.11929

6. Dupret D, O'Neill J, Pleydell-Bouverie B, Csicsvari J (2010) **The reorganization and reactivation of hippocampal maps predict spatial memory performance** *Nature Neuroscience* **13**:995–1002 https://doi.org/10.1038/nn.2599

7. Ebitz RB, Hayden BY (2021) **The population doctrine in cognitive neuroscience** *Neuron* **109**:3055–3068 https://doi.org/10.1016/j.neuron.2021.07.011

8. Eichenbaum H (2015) **Perspectives on 2014 Nobel Prize** *Hippocampus* **25**:679–681 https://doi.org/10.1002/hipo.22445

9. Eichenbaum H (2018) **Barlow versus Hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition?** *Neuroscience Letters* **680**:88–93 https://doi.org/10.1016/j.neulet.2017.04.006

10. Eickenberg M, Gramfort A, Varoquaux G, Thirion B (2017) **Seeing it all: Convolutional network layers map the function of the human visual system** *NeuroImage* **152**:184–194 https://doi.org/10.1016/J.NEUROIMAGE.2016.10.001

11. Franzius M, Sprekeler H, Wiskott L (2007) **Slowness and Sparseness Lead to Place, Head-Direction, and Spatial-View Cells** *PLoS Computational Biology* **3** https://doi.org/10.1371/journal.pcbi.0030166

12. Franzius M, Vollgraf R, Wiskott L (2007) **From grids to places** *Journal of Computational Neuroscience* **22**:297–299 https://doi.org/10.1007/s10827-006-0013-7

13. Fukushima K (1980) **Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position** *Biological Cybernetics* **36**:193–202 https://doi.org/10.1007/BF00344251

14. Grieves RM, Jeffery KJ (2017) **The representation of space in the brain** *Behavioural Processes* **135**:113–131 https://doi.org/10.1016/j.beproc.2016.12.012

15. Grieves RM, Jedidi-Ayoub S, Mishchanchuk K, Liu A, Renaudineau S, Jeffery KJ (2020) **The place-cell representation of volumetric space in rats** *Nature Communications* **11** https://doi.org/10.1038/s41467-020-14611-7

16. Güçlü U, van Gerven MAJ (2015) **Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream** *Journal of Neuroscience* **35**:10005–10014 https://doi.org/10.1523/JNEUROSCI.5023-14.2015

17. Hafting T, Fyhn M, Molden S, Moser MB, Moser EI (2005) **Microstructure of a spatial map in the entorhinal cortex** *Nature* **436**:801–806

18. Hales J, Schlesiger M, Leutgeb J, Squire L, Leutgeb S, Clark R (2014) **Medial Entorhinal Cortex Lesions Only Partially Disrupt Hippocampal Place Cells and Hippocampus-Dependent Place Memory** *Cell Reports* **9**:893–901 https://doi.org/10.1016/j.celrep.2014.10.009

19. Hartley T, Lever C, Burgess N, O'Keefe J (2014) **Space in the brain: how the hippocampal formation supports spatial cognition** *Philosophical Transactions of the Royal Society B: Biological Sciences* **369** https://doi.org/10.1098/rstb.2012.0510

20. He K, Zhang X, Ren S, Sun J (2016) **Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition** *IEEE Computer Society* https://doi.org/10.1109/CVPR.2016.90

21. Hebb D (1949) **The Organizationof Behavior: A Neuropsychological Theory, 2002nd edn**

22. Hollup SA, Molden S, Donnett JG, Moser MB, Moser EI (2001) **Accumulation of Hippocampal Place Fields at the Goal Location in an Annular Watermaze Task** *The Journal of Neuroscience* **21**:1635–1644 https://doi.org/10.1523/JNEUROSCI.21-05-01635.2001

23. Hubel DH, Wiesel TN (1959) **Receptive fields of single neurones in the cat's striate cortex** *The Journal of Physiology* **148**:574–591 https://doi.org/10.1113/JPHYSIOL.1959.SP006308

24. Høydal Skytøen ER, Andersson SO, Moser MB, Moser EI (2019) **Objectvector coding in the medial entorhinal cortex** *Nature* **568**:400–404 https://doi.org/10.1038/s41586-019-1077-7

25. Khaligh-Razavi SM, Kriegeskorte N (2014) **Deep Supervised, but Not Unsupervised** *Models May Explain IT Cortical Representation. PLoS Computational Biology* **10** https://doi.org/10.1371/journal.pcbi.1003915

26. Krizhevsky A, Sutskever I, Hinton GE (2012) **ImageNet classification with deep convolutional neural networks**

27. Kropff E, Treves A (2008) **The emergence of grid cells: Intelligent design or just adaptation?** *Hippocampus* **18**:1256–1269 https://doi.org/10.1002/hipo.20520

28. Latuske P, Toader O, Allen K (2015) **Interspike Intervals Reveal Functionally Distinct Cell Populations in the Medial Entorhinal Cortex** *Journal of Neuroscience* **35**:10963–10976 https://doi.org/10.1523/JNEUROSCI.0276-15.2015

29. Lever C, Burton S, Jeewajee A, O'Keefe J, Burgess N (2009) **Boundary Vector Cells in the Subiculum of the Hippocampal Formation** *The Journal of Neuroscience* **29**:9771–9777 https://doi.org/10.1523/JNEUROSCI.1319-09.2009

30. Long X, Zhang SJ (2021) **A novel somatosensory spatial navigation system outside the hippocampal formation** *Cell Research* **31**:649–663 https://doi.org/10.1038/s41422-020-00448-8

31. Long X, Deng B, Cai J, Chen ZS, Zhang SJ (2021) **A compact spatial map in V2 visual cortex. preprint** *Neuroscience* https://doi.org/10.1101/2021.02.11.430687

32. McKenzie S, Robinson NTM, Herrera L, Churchill JC, Eichenbaum H (2013) **Learning Causes Reorganization of Neuronal Firing Patterns to Represent Related Experiences within a Hippocampal Schema** *Journal of Neuroscience* **33**:10243–10256 https://doi.org/10.1523/JNEUROSCI.0879-13.2013

33. McMahon E, Isik L (2023) **Seeing social interactions** *Trends in Cognitive Sciences* **27**:1165–1179 https://doi.org/10.1016/j.tics.2023.09.001

34. McNaughton L, Chen LL, Markus J (1991) **DeadReckoning,"kmdmark Learning, and the Sense of Direction: A Neurophysiological and Computational Hypothesis** *Journal of Cognitive Neuroscience* **3**:190–202

35. Mok RM, Love BC (2019) **A non-spatial account of place and grid cells based on clustering models of concept learning** *Nature Communications* **10** https://doi.org/10.1038/S41467-019-13760-8

36. Mok RM, Love BC (2023) **A multilevel account of hippocampal function in spatial and concept learning: Bridging models of behavior and neural assemblies** *Science Advances* **9** https://doi.org/10.1126/sciadv.ade6903

37. Moser EI, Kropff E, Moser MB (2008) **Place Cells, Grid Cells, and the Brain's Spatial Representation System** *Annual Review of Neuroscience* **31**:69–89 https://doi.org/10.1146/annurev.neuro.31.061307.090723

38. Nayebi A *et al.* (2021) **Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. In: Advances in Neural Information Processing Systems** *Neuroscience* https://doi.org/10.1101/2021.10.30.466617

39. O'Keefe J, Dostrovsky J (1971) **The hippocampus as a spatial map** *Preliminary evidence from unit activity in the freely-moving rat. Brain Research* **34**:171–175 https://doi.org/10.1016/0006-8993(71)90358-1

40. Olshausen BA, Field DJ, Van Hemmen JL, Sejnowski TJ (2006) **What Is the Other 85 Percent of V1 Doing?** *23 Problems in Systems Neuroscience* :182–212 https://doi.org/10.1093/acprof:oso/9780195148220.003.0010

41. Ormond J, O'Keefe J (2022) **Hippocampal place cells have goaloriented vector fields during navigation** *Nature* **607**:741–746 https://doi.org/10.1038/s41586-022-04913-9

42. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) **Invariant visual representation by single neurons in the human brain** *Nature* **435**:1102–1107 https://doi.org/10.1038/nature03687

43. Saleem AB, Diamanti EM, Fournier J, Harris KD, Carandini M (2018) **Coherent encoding of subjective spatial position in visual cortex and hippocampus** *Nature* **562**:124–127 https://doi.org/10.1038/s41586-018-0516-1

44. Sarel A, Finkelstein A, Las L, Ulanovsky N (2017) **Vectorial representation of spatial goals in the hippocampus of bats** *Science* **355**:176–180 https://doi.org/10.1126/science.aak9589

45. Simonyan K, Zisserman A (2015) **Very Deep Convolutional Networks for LargeScale Image Recognition** *arXiv*

46. Simonyan K, Vedaldi A, Zisserman A (2014) **Deep inside convolutional networks: Visualising image classification models and saliency maps** *2nd International Conference on Learning Representations, ICLR 2014 Workshop Track Proceedings pp 1–8*

47. Stachenfeld KL, Botvinick MM, Gershman SJ (2017) **The hippocampus as a predictive map** *Nature Neuroscience* **20**:1643–1653 https://doi.org/10.1038/nn.4650

48. Tang Q *et al.* (2015) **Anatomical Organization and Spatiotemporal Firing Patterns of Layer 3 Neurons in the Rat Medial Entorhinal Cortex** *Journal of Neuroscience* **35**:12346–12354 https://doi.org/10.1523/JNEUROSCI.0696-15.2015

49. Tanni S, De Cothi W, Barry C (2022) **State transitions in the statistically stable place cell population correspond to rate of perceptual change** *Current Biology* **32**:3505–3514 https://doi.org/10.1016/j.cub.2022.06.046

50. Taube J, Muller R, Ranck J (1990) **Head-direction cells recorded from the postsubiculum in freely moving rats** *II. Effects of environmental manipulations. The Journal of Neuroscience* **10**:436–447 https://doi.org/10.1523/JNEUROSCI.10-02-00436.1990

51. Tsao A, Sugar J, Lu L, Wang C, Knierim JJ, Moser MB, Moser EI (2018) **Integrating time from experience in the lateral entorhinal cortex** *Nature* **561**:57–62 https://doi.org/10.1038/s41586-018-0459-6

52. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) **Attention is all you need**

53. Whishaw IQ, Tomie JA (1997) **Perseveration on place reversals in spatial swimming pool tasks: Further evidence for place learning in hippocampal rats** *Hippocampus* **7**:361–370 https://doi.org/10.1002/(SICI)1098-1063(1997)7:4<361::AID-HIPO2>3.0.CO;2-M

54. Whishaw IQ, McKenna JE, Maaswinkel H (1997) **Hippocampal lesions and path integration** *Current Opinion in Neurobiology* **7**:228–234 https://doi.org/10.1016/S0959-4388(97)80011-6

55. Whittington JC, Muller TH, Mark S, Chen G, Barry C, Burgess N, Behrens TE (2020) **The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation** *Cell* **183**:1249–1263 https://doi.org/10.1016/J.CELL.2020.10.024

56. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) **Performance-optimized hierarchical models predict neural responses in higher visual cortex** *Proceedings of the National Academy of Sciences of the United States of America* **111**:8619–8624 https://doi.org/10.1073/pnas.1403112111

57. Zeman AA, Ritchie JB, Bracci S, de Beeck H Op (2020) **Orthogonal Representations of Object Shape and Category in Deep Convolutional Neural Networks and Human Visual Cortex** *Scientific Reports* **10**:1–12 https://doi.org/10.1038/s41598-020-59175-0

Editors

Reviewing Editor
Panayiota Poirazi
FORTH Institute of Molecular Biology and Biotechnology, Heraklion, Greece

Senior Editor
Panayiota Poirazi
FORTH Institute of Molecular Biology and Biotechnology, Heraklion, Greece

Reviewer #1 (Public Review):

Summary:
This study investigated spatial representations in deep feedforward neural network models (DDNs) that were often used in visual tasks. The authors create a three-dimensional virtual environment, and let a simulated agent randomly forage in a smaller two-dimensional square area. The agent "sees" images of the room within its field of view from different locations and heading directions. These images were processed by DDNs. Analyzing model neurons in DDNs, they found response properties similar to those of place cells, border cells and head direction cells in various layers of deep nets. A linear readout of network activity can recover key spatial variables. In addition, after removing neurons with strong place/border/head direction selectivity, one can still decode these spatial variables from the remaining neurons in the DNNs. Based on these results, the authors argue that that the notion of functional cell types in spatial cognition is misleading.

Strengths:
This paper contains interesting and original ideas, and I enjoy reading it. Most previous studies (e.g., Banino, Nature, 2018; Cueva & Wei, ICLR, 2018; Whittington et al, Cell, 2020) using deep network models to investigate spatial cognition mainly relied on velocity/head rotation inputs, rather than vision (but see Franzius, Sprekeler, Wiskott, PLoS Computational Biology, 2007). Here, the authors find that, under certain settings, visual inputs alone may contain enough information about the agent's location, head direction and distance to the boundary, and such information can be extracted by DNNs. If confirmed, this is potentially an interesting and important observation.

Weaknesses:
While the findings reported here are interesting, it is unclear whether they are the consequence of the specific model setting, and how well they would generalize. Furthermore, I feel the results are over-interpreted. There are major gaps between the results actually shown and the claim about the "superfluousness of cell types in spatial cognition". Evidence directly supporting the overall conclusion seems to be weak at the moment.

Major concerns:

(1) The authors reported that, in their model setting, most neurons throughout the different layers of CNNs show strong spatial selectivity. This is interesting and perhaps also surprising. It would be useful to test/assess this prediction directly based on existing experimental results. It is possible that the particular 2-d virtual environment used is special. The results will be strengthened if similar results hold for other testing environments.

In particular, examining the pictures shown in Fig. 1A, it seems that local walls of the 'box' contain strong oriented features that are distinct across different views. Perhaps the response of oriented visual filters can leverage these features to uniquely determine the spatial variable. This is concerning because this is a very specific setting that is unlikely to generalize.

(2) Previous experimental results suggest that various function cell types discovered in rodent navigation circuits persist in dark environments. If we take the modeling framework presented in this paper literally, the prediction would be that place cells/head direction cells should go away in darkness. This implies that key aspects of functional cell types in the spatial cognition are missing in the current modeling framework. This limitation needs to be addressed or explicitly discussed.

(3) Place cells/border cell/ head direction cells are mostly studied in the rodent's brain. For rodents, it is not clear whether standard DNNs would be good models of their visual systems. It is likely that rodent visual system would not be as powerful in processing visual inputs as the DNNs used in this study.

(4) The overall claim that those functional cell types defined in spatial cognition are superfluousness seems to be too strong based on the results reported here. The paper only studied a particular class of models, and arguably, the properties of these models have a major gap to those of real brains. Even though, in the DNN models simulated in this particular virtual environment, (i) most model neurons have strong spatial selectivity; (ii) removing model neurons with the strongest spatial selectivity still retain substantial spatial information, why this is relevant to the brain? The neural circuits may operate in a very different regime. Perhaps a more reasonable interpretation of the results would be: these results raise the possibility that those strongly selective neurons observed in the brain may not be essential for encoding certain features, as something like this is observed in certain models. It is difficult to draw definitive conclusions about the brain based on the results reported.

https://doi.org/10.7554/eLife.99047.1.sa3

Reviewer #2 (Public Review):

Summary:
The authors aim at challenging the relevance of cell populations with characteristic selectivity for specific aspects of navigation (e.g. place cells, head direction and border cells) in the processing of spatial information. Their claim is that such cells naturally emerge in any system dealing with the estimation of position in an environment, without the need for a special involvement of these cells in the computations. In particular the work shows how when provided with spatial error signals, networks designed for invariant object recognition spontaneously organize the activity in their hidden layers into a mixture of spatially selective cells, some of them passing classification criteria for place, head direction or border cells. Crucially, these cells are not necessary for position decoding, nor are they the most informative when it comes to the performance of the network in reconstructing spatial position from visual scenes. These results lead the authors to claim that focusing on the classification of specific cell types is hindering rather than helping advancement in the understanding of spatial cognition. In fact they claim that the attention should rather be pointed at understanding highly-dimensional population coding, regardless of its direct interpretability or its appeal to human observers.

Strengths:
Methodologically the paper is consistent and convincingly support the author claims regarding the role of cell types in coding for spatial aspects of cognition. It is also interesting how the authors leverage on established machine learning systems to provide a sort of counter-argument to the use of such techniques to establish a parallel between artificial and biological neural representations. In the recent past similar applications of artificial neural networks to spatial navigation have been directed at proving the importance of specific neural substrates (take for example Banino et al. 2018 for grid cells), while in this case the

same procedure is used to unveil them as epiphenomena, so general and unspecific to be of very limited use in understanding the actual functioning of the neural system. I am quite confident that this stance regarding the role of place cells and co. could gather large sympathy and support in the greater part of the neuroscience community, or at least among the majority of theoretical neuroscientists with some interest in the hippocampus and higher cognition.

Weaknesses:
My criticism of the paper can be articulated in three main points:
- What about grid cells? Grid cells are notably not showing up in the analyses of the paper. But they surely can be considered as the 'mother' of all tailored spatial cells of the hippocampal formation. Are they falling outside the author's assessment of the importance of this kind of cells? Some discussion of the place grid cells occupy in the vision of the authors would greatly help.
- The network used in the paper is still guided by a spatial error signal, and the network is trained to minimize spatial decoding error. In a sense, although object classfication networks are not designed for spatial navigation, one could say that the authors are in some way hacking this architecture and turning it into a spatial navigation one through learning. I wonder if their case could be strengthened by devising a version of their experiment based on some form of self-supervised or unsupervised learning.
- The last point is more about my perception of the community studying hippocampal functions, rather than being directed at the merits of the paper itself. My question is whether the paper is fighting an already won battle. That is whether the focus on the minute classification of response profiles of cells in the hippocampus is in fact already considered an 'old' approach, very useful for some initial qualitative assessments but of limited power when asked to provide deeper insight into the functioning of hippocampal computations (or computations of any other brain circuit).

https://doi.org/10.7554/eLife.99047.1.sa2

Reviewer #3 (Public Review):

Summary:
In this paper, the authors demonstrate the inevitably of the emergence of some degree of spatial information in sufficiently complex systems, even those that are only trained on object recognition (i.e. not "spatial" systems). As such, they present an important null hypothesis that should be taken into consideration for experimental design and data analysis of spatial tuning and its relevance for behavior.

Strengths:
The paper's strengths include the use of a large multi-layer network trained in a detailed visual environment. This illustrates an important message for the field: that spatial tuning can be a result of sensory processing. While this is a historically recognized and often-studied fact in experimental neuroscience, it is made more concrete with the use of a complex sensory network. Indeed, the manuscript is a cautionary tale for experimentalists and computational researchers alike against blindly applying and interpreting metrics without adequate controls.

Weaknesses:
However, the work has a number of significant weaknesses. Most notably: the degree and quality of spatial tuning is not analyzed to the standards of evidence historically used in studies of spatial tuning in the brain, and the authors do not critically engage with past work that studies the sensory influences of these cells; there are significant issues in the authors' interpretation of their results and its impact on neuroscientific research; the ability to linearly decode position from a large number of units is not a strong test of spatial

information, nor is it a measure of spatial cognition; and the authors make strong but unjustified claims as to the implications of their results in opposition to, as opposed to contributing to, work being done in the field.

The first weakness is that the degree and quality of spatial tuning that emerges in the network is not analyzed to the standards of evidence that have been used in studies of spatial tuning in the brain. Specifically, the authors identify place cells, head direction cells, and border cells in their network and their conjunctive combinations. However, these forms of tuning are the most easily confounded by visual responses, and it's unclear if their results will extend to forms of spatial tuning that are not. Further, in each case, previous experimental work to further elucidate the influence of sensory information on these cells has not been acknowledged or engaged with.

For example, consider the head direction cells in Figure 3C. In addition to increased activity in some directions, these cells also have a high degree of spatial nonuniformity, suggesting they are responding to specific visual features of the environment. In contrast, the majority of HD cells in the brain are only very weakly spatially selective, if at all, once an animal's spatial occupancy is accounted for (Taube et al 1990, JNeurosci). While the preferred orientation of these cells are anchored to prominent visual cues, when they rotate with changing visual cues the entire head direction system rotates together (cells' relative orientation relationships are maintained, including those that encode directions facing AWAY from the moved cue), and thus these responses cannot be simply independent sensory-tuned cells responding to the sensory change) (Taube et al 1990 JNeurosci, Zugaro et al 2003 JNeurosci, Ajbi et al 2023).

As another example, the joint selectivity of detected border cells with head direction in Figure 3D suggests that they are "view of a wall from a specific angle" cells. In contrast, experimental work on border cells in the brain has demonstrated that these are robust to changes in the sensory input from the wall (e.g. van Wijngaarden et al 2020), or that many of them are not directionally selective (Solstad et al 2008).

The most convincing evidence of "spurious" spatial tuning would be the emergence of HD-independent place cells in the network, however, these cells are a small minority (in contrast to hippocampal data, Thompson and Best 1984 JNeurosci, Rich et al 2014 Science), the examples provided in Figure 3 are significantly more weakly tuned than those observed in the brain, and the metrics used by the authors to quantify place cell tuning are not clearly defined in the methods, but do not seem to be as stringent as those commonly used in real data. (e.g. spatial information, Skaggs et al 1992 NeurIPS).

Indeed, the vast majority of tuned cells in the network are conjunctively selective for HD (Figure 3A). While this conjunctive tuning has been reported, many units in the hippocampus/entorhinal system are *not* strongly hd selective (Muller et al 1994 JNeurosci, Sangoli et al 2006 Science, Carpenter et al 2023 bioRxiv). Further, many studies have been done to test and understand the nature of sensory influence (e.g. Acharya et al 2016 Cell), and they tend to have a complex relationship with a variety of sensory cues, which cannot readily be explained by straightforward sensory processing (rev: Poucet et al 2000 Rev Neurosci, Plitt and Giocomo 2021 Nat Neuro). E.g. while some place cells are sometimes reported to be directionally selective, this directional selectivity is dependent on behavioral context (Markus et al 1995, JNeurosci), and emerges over time with familiarity to the environment (Navratiloua et al 2012 Front. Neural Circuits). Thus, the question is not whether spatially tuned cells are influenced by sensory information, but whether feed-forward sensory processing alone is sufficient to account for their observed turning properties and responses to sensory manipulations.

These issues indicate a more significant underlying issue of scientific methodology relating to the interpretation of their result and its impact on neuroscientific research. Specifically, in

order to make strong claims about experimental data, it is not enough to show that a control (i.e. a null hypothesis) exists, one needs to demonstrate that experimental observations are quantitatively no better than that control.

Where the authors state that "In summary, complex networks that are not spatial systems, coupled with environmental input, appear sufficient to decode spatial information." what they have really shown is that it is possible to decode *some degree* of spatial information. This is a null hypothesis (that observations of spatial tuning do not reflect a "spatial system"), and the comparison must be made to experimental data to test if the so-called "spatial" networks in the brain have more cells with more reliable spatial info than a complex-visual control.

Further, the authors state that "Consistent with our view, we found no clear relationship between cell type distribution and spatial information in each layer. This raises the possibility that "spatial cells" do not play a pivotal role in spatial tasks as is broadly assumed." Indeed, this would raise such a possibility, if 1) the observations of their network were indeed quantitatively similar to the brain, and 2) the presence of these cells in the brain were the only evidence for their role in spatial tasks. However, 1) the authors have not shown this result in neural data, they've only noticed it in a network and mentioned the POSSIBILITY of a similar thing in the brain, and 2) the "assumption" of the role of spatially tuned cells in spatial tasks is not just from the observation of a few spatially tuned cells. But from many other experiments including causal manipulations (e.g. Robinson et al 2020 Cell, DeLauilleon et al 2015 Nat Neuro), which the authors conveniently ignore. Thus, I do not find their argument, as strongly stated as it is, to be well-supported.

An additional weakness is that linear decoding of position is not a strong test, nor is it a measure of spatial cognition. The ability to decode position from a large number of weakly tuned cells is not surprising. However, based on this ability to decode, the authors claim that "'spatial' cells do not play a privileged role in spatial cognition". To justify this claim, the authors would need to use the network to perform e.g. spatial navigation tasks, then investigate the network's ability to perform these tasks when tuned cells were lesioned.

Finally, I find a major weakness of the paper to be the framing of the results in opposition to, as opposed to contributing to, the study of spatially tuned cells. For example, the authors state that "If a perception system devoid of a spatial component demonstrates classically spatially-tuned unit representations, such as place, head-direction, and border cells, can "spatial cells" truly be regarded as 'spatial'?" Setting aside the issue of whether the perception system in question does indeed demonstrate spatially-tuned unit representations comparable to those in the brain, I ask "Why not?" This seems to be a semantic game of reading more into a name then is necessarily there. The names (place cells, grid cells, border cells, etc) describe an observation (that cells are observed to fire in certain areas of an animal's environment). They need not be a mechanistic claim (that space "causes" these cells to fire) or even, necessarily, a normative one (these cells are "for" spatial computation). This is evidenced by the fact that even within e.g. the place cell community, there is debate about these cells' mechanisms and function (eg memory, navigation, etc), or if they can even be said to serve only a single function. However, they are still referred to as place cells, not as a statement of their function but as a history-dependent label that refers to their observed correlates with experimental variables. Thus, the observation that spatially tuned cells are "inevitable derivatives of any complex system" is itself an interesting finding which *contributes to*, rather than contradicts, the study of these cells. It seems that the authors have a specific definition in mind when they say that a cell is "truly" "spatial" or that a biological or artificial neural network is a "spatial system", but this definition is not stated, and it is not clear that the terminology used in the field presupposes their definition.

In sum, the authors have demonstrated the existence of a control/null hypothesis for observations of spatially-tuned cells. However, 1) It is not enough to show that a control (null

hypothesis) exists, one needs to test if experimental observations are no better than control, in order to make strong claims about experimental data, 2) the authors do not acknowledge the work that has been done in many cases specifically to control for this null hypothesis in experimental work or to test the sensory influences on these cells, and 3) the authors do not rigorously test the degree or source of spatial tuning of their units.

https://doi.org/10.7554/eLife.99047.1.sa1

Author response:

We thank the reviewers for their engagement and constructive comments. This provisional response aims to clarify key misconceptions, address major criticisms, and outline our revision plans.

A primary concern of the reviewers appears to be our model's limitations in addressing a broad range of empirical findings. This, however, misinterprets our core contribution. Our work centers on a cautionary tale that before advocating for newly discovered cell types and their purported special roles in spatial cognition—an approach prevalent in the field—such claims must be tested against alternative (null) hypotheses that may contradict intuitive expectations. We present such an alternative hypothesis regarding spatial cells and their assumed privileged roles. We show that key findings in the field - spatial “cell types”, arise in a set of null models without spatial grounding (including untrained variants) despite the models not being a model for spatial processing, and we also found that they had no privileged role for representing spatial information.

Our proposal is not a new model attempting to explain the brain, and therefore we do not aim to capture every empirical finding. Indeed, we would not expect an object recognition model (and its untrained variant) with no explicit spatial grounding to account for all phenomena in spatial cognition. This underscores our key point: if there exists a basic, spatially agnostic model that can explain certain degrees of empirical findings using criteria from the literature (i.e. place, head-direction and border cells), what implications does this have for the more complex theories and models proposed as underlying mechanisms of special cell types?

Regarding concerns about the limited scope and generalizability of our setting, we will clarify that we considered multiple DNN architectures, both trained and untrained, on multiple decoding tasks (position, head direction, and nearest-wall distance). We plan to extend our experiments further as detailed in the revision plan below.

Further, there was a methodological concern about using a linear decoder on a fixed DNN for spatial decoding tasks being a form of "hacking". However, linear readout is standard practice in neuroscience to characterize information available in a neural population. Moreover, our tests on untrained networks also showed spatial decoding capabilities, suggesting it's not solely due to the linear readout.

For our full revision plan:

(1) We will revise the manuscript to better reflect these above points, clarifying our paper's stance and improving the writing to reduce misconceptions.

(2) We will address individual public reviews in more detail.

(3) We intend to address key reviewer recommendations, focusing on better situating our work within the broader context of the existing literature whilst emphasizing the null hypothesis perspective.

(4) In general, we will consider additional aspects of the literature and conduct new experiments to strengthen the relevance of our work to existing work. We highlight a number of potential experiments which we believe can address reviewer concerns:

a. Blurring the visual inputs to DNNs to match rodent perception.

b. Vary environmental settings to verify whether our findings are more

generalizable (which we predict to be the case).

c. Vary the environment to assess remapping effects, which will strengthen the

connection of our work to the literature.

https://doi.org/10.7554/eLife.99047.1.sa0