Vision Research 174 (2020) 79–93

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Lessons from reinforcement learning for biological representations of space

Alex Muryy[a], N. Siddharth[b], Nantas Nardelli[b], Andrew Glennerster[a,*], Philip H.S. Torr[b]

[a] School of Psychology and Clinical Language Sciences, University of Reading, UK
[b] Department of Engineering Science, University of Oxford, UK

ARTICLE INFO

Keywords:
Deep Reinforcement Learning
3D spatial representation
Moving observer
Navigation
View-based
Parallax

ABSTRACT

Neuroscientists postulate 3D representations in the brain in a variety of different coordinate frames (e.g. 'head-centred', 'hand-centred' and 'world-based'). Recent advances in reinforcement learning demonstrate a quite different approach that may provide a more promising model for biological representations underlying spatial perception and navigation. In this paper, we focus on reinforcement learning methods that reward an agent for arriving at a target image without any attempt to build up a 3D 'map'. We test the ability of this type of representation to support geometrically consistent spatial tasks such as interpolating between learned locations using decoding of feature vectors. We introduce a hand-crafted representation that has, by design, a high degree of geometric consistency and demonstrate that, in this case, information about the persistence of features as the camera translates (e.g. distant features persist) can improve performance on the geometric tasks. These examples avoid Cartesian (in this case, 2D) representations of space. Non-Cartesian, learned representations provide an important stimulus in neuroscience to the search for alternatives to a 'cognitive map'.

1. Introduction

The discovery of place cells, grid cells, heading direction cells, boundary vector cells and similar neurons in the mammalian hippocampus and surrounding cortex has been interpreted as evidence that the brain builds up an allocentric, world-based representation or 'map' of the environment and indicates the animals movement within it (Hafting, Fyhn, Molden, Moser, & Moser, 2005; Lever, Burton, Jeewajee, O'Keefe, & Burgess, 2009; O'Keefe & Dostrovsky, 1971; Taube, Muller, & Ranck, 1990). However, this interpretation is increasingly questioned and alternative models are proposed that do not involve a 'cognitive map' (Acharya, Aghajan, Vuong, Moore, & Mehta, 2016; Warren, 2019; Glennerster et al., 2016). Computer vision and robotics provide a useful source of inspiration for models of spatial representation in animals because their performance can be tested. Until recently, the predominant computer vision model for 3D representation has been simultaneous localisation and mapping (SLAM) where a 3D reconstruction of the scene and the agent's location within it are continually updated as new sensory information is received (Davison, 2003; Fuentes-Pacheco, Ruiz-Ascencio, & Rendón-Mancha, 2015) (and non-visual precursors of SLAM such as Chatila & Laumond (1985)). Although there are many variations on this theme, the essence of SLAM is that a set of corresponding features in images taken from different vantage points are used to recover (i) the 3D structure of those points in the scene and (ii) the rotation and translation of the camera per frame, where scene structure and camera pose are all described in the same 3D coordinate frame.

However, since the advent of deep neural networks, there has been a move to try out a quite different approach to representing a 3D environment and controlling movement within it. The agent is tasked with matching the input resulting from a particular camera pose (i.e. an image, not a 3D location) and rewarding, however sparsely, the actions that lead it on a path to that goal. Eventually, after many trials, the agent learns to take a sequence of actions ('turn left', 'turn right', 'go forward') that take it from the current image to the goal although it never builds an explicit 'map' in the sense of a representation of the scene layout with an origin and coordinate frame. These networks are different from earlier attempts to model mammalian navigation that used information about the location of the agent gained from model place cells (Foster, Morris, & Dayan, 2000) or using idiothetic information from proprioceptive and vestibular inputs (Arleo et al., 2000). Instead, they rely on visual information alone to build up a representation of space and are, in that sense, directly comparable with SLAM models. The recent RL models also differ from early attempts to represent space using very simple visual inputs such as Franz, Schölkopf, Mallot, and Bülthoff (1998) where the input was a set of 1-D omnidirectional measurements of luminance values and the robot laid down a new 'snapshot' whenever the view differed significantly from its

* Corresponding author.
E-mail address: a.glennerster@reading.ac.uk (A. Glennerster).

https://doi.org/10.1016/j.visres.2020.05.009
Received 31 October 2019; Received in revised form 26 April 2020; Accepted 26 May 2020

current stored snapshots, generating a topological graph of space as it went (Chatila & Laumond, 1985; Barrera & Weitzenfeld, 2008). For one thing, the rules for storing the feature vectors were quite different in these approaches, although in some ways they were forerunners of the modern RL approach. Also radically different are the inverse RL approaches that have been used to predict human movement in relation to obstacles and goals (Rothkopf & Ballard, 2013). These fit human navigation data in a low dimensional space of control parameters that, while successful in explaining obstacle avoidance, do not relate to an allocentric space representation.

1.1. A classic reinforcement learning approach to navigation: Zhu et al.

In 2017, Zhu et al. (2017) showed that reinforcement learning could be applied successfully to a navigation task in which the agent was rewarded for arriving at a particular image (i.e. a given location and pose of the camera, although these 3D variables were not explicitly encoded in the input the agent received, only the current image and the goal image). It is one of the key papers in this emerging field of reinforcement learning (RL)-based *perceptual-goal-driven* navigation (Zhu et al., 2017; Anderson et al., 2018; Dhiman, Banerjee, Griffin, Siskind, & Corso, 2018; Edwards, 2017; Sermanet, Xu, & Levine, 2016; Singh, Yang, Hartikainen, Finn, & Levine, 2019; Yang, Wang, Farhadi, Gupta, & Mottaghi, 2018). Zhu et al. (2017) in particular was one of the first to show it is possible to construct an end-to-end architecture for visual-goal-driven navigation using a modern deep learning stack trained with RL. This was in contrast to more typical RL work that treats the task of navigation to particular positions of the world just as part of general, global, state-based reward function (e.g. all the work on taxi-world (Mnih et al., 2016) and most other tasks based on minigrids, or even the more recent BabyAI (Chevalier-Boisvert et al., 2018)). We illustrate what the system has learned by relating the stored vectors in the representation to the agent's location and orientation in space. We show, in particular, that the contexts that the representation recognises are heavily dependent on the agent's current goal. The fact that the agent's task is integrated into the representation of current and stored states is reminiscent of many results in biological representation of shape and space (Bradshaw, Parton, & Glennerster,2000; Bradshaw et al., 2000; Smeets et al., 2008; Warren, 2019; Glennerster et al., 2016). Since Zhu et al. (2017), there have been a number of important developments in this type of approach. Mirowski et al. (2018) adapted the method to cover much larger spatial regions using images from *Google StreetView*; Eslami et al. (2018) have shown that behaviour one might have thought would require a 3D model (e.g. predicting a novel view from a novel location in a novel scene) can be learned by carrying out the same task in many similar scenes; and others have included an explicit coordinate frame in the stored memory (Gupta, Davidson, Levine, Sukthankar, & Malik, 2017; Chen, Gupta, & Gupta, 2019; Kumar, Gupta, & Malik, 2019; Mirowski et al., 2018; Mirowski et al., 2016) while Kanitscheider and Fiete (2017) have built on the biologically-inspired (but allocentric, coordinate-based) RatSLAM model of Milford et al. (2010). In contrast to these coordinate-based advances, progress since Zhu et al. (2017) on pure image-based approaches to large-scale spatial representation for navigation has slowed, as the community has been primarily focused on improving the visual navigation testbeds (Savva et al., 2019). Another paper that incorporates an explicit biological perspective in relation to navigation is Wayne et al. (2018) who have shown the importance of storing 'predictions that are consistent with the probabilities of observed sensory sequences from the environment'. They use a Memory Based Predictor to do this and draw attention to the similarities between the MBP and some of the proposed functions of the hippocampus.

In this paper, we examine the feature vectors in the stored representation after learning in the Zhu et al. (2017) study to explore the extent to which they reflect the spatial layout of the scene. We show that, although spatial information *is* present in the representation, sufficient to be decoded, the organisation of the feature vectors is dominated by other factors such as the goal and the orientation of the camera (as one might expect, given the inputs to the network) and that it is possible to use these feature vectors to carry out simple spatial tasks such as interpolating between two learned locations.

1.2. A hand-crafted alternative representation using relative visual directions

We compare the performance of this RL network to a representation of location that (i) avoids any explicit 3D coordinate frame (like the RL approach), (ii) represents the current sensory state as a high dimensional vector (like the RL approach) but (iii) unlike the RL approach, is built on information that is known to be important in biological vision. The visual system is much more sensitive to the spatial separation (relative visual direction) of points than it is to their absolute visual direction (Kinchla et al., 1971; Westheimer, 1979; Erkelens & Collewijn, 1985; Thomas, Cumming, & Parker, 2002; Regan, Erkelens, & Collewijn, 1986; Glennerster, Hansard, & Fitzgibbon, 2001) and it has been suggested on the basis of psychophysical evidence (Watt, 1987) that the visual system uses a reference frame for egocentric visual direction that is built from the relative visual direction of points and hence has no single 2D coordinate frame encompassing the sphere of visual directions (Glennerster, Hansard, & Fitzgibbon, 2009; Watt, 1987, 1988; Glennerster & Read, 2018; Glennerster et al., 2001). This representation is very similar to a list of the saccades (magnitude and direction) that would take the eye from one point to another in the scene. Information about the 3D structure of the scene can be added to this representation by incorporating information about the *change* in the relative visual direction of points when the camera translates (motion parallax or binocular disparity). Glennerster et al. (2001) showed how the pattern of eye movements that animals generally adopt, which is to fixate on a point as they move, is a distinct advantage for interpreting retinal flow if one assumes that the goal of the visual system is to update a representation of this sort. If animals fixate a point as they move, retinal motion provides information straightforwardly about changes in the relative visual direction of points with respect to the fixation point and the information can be used to build up a representation like the one we describe below. We call this a 'relative visual direction' representation (RVD) (Glennerster, 2016; Glennerster, Hansard, & Fitzgibbon, 2009; Glennerster & Read, 2018; Glennerster et al., 2001). In the simplistic implementation we describe here, the input is 1-dimensional and spans the entire 360° field of view, whereas in practice the input would be 2-dimensional and the field of view would be limited so information would have to be gathered over successive fixations. The skeletal version used here nevertheless illustrates some key points about the information that is available in a representation that stores information in a relatively raw form, without building a 3D coordinate frame. In particular, we show how motion parallax can be useful in separating out information in the representation that is likely to persist as the observer translates while other information is likely to go rapidly 'out of date'.

1.3. Comparison of feature vector models

We report on the performance of both types of model when faced with tasks that require basic spatial knowledge. The tasks we chose were interpolating between two locations or interpolating between two visual directions because these test whether the network contains information about novel locations or directions that it has not learned about during training. Bisection tasks have been carried out in humans (Purdy & Gibson, 1955; Rieser, Ashmead, Talor, & Youngquist, 1990; Bodenheimer et al., 2007) and are simpler to imitate than other tests of 'map-like' properties of spatial representation in humans such as a triangle completion task (Klatzky, Beall, Loomis, Golledge, & Philbeck, 1999; Foo, Warren, Duchon, & Tarr, 2005).

The input to the two algorithms is utterly different (2D images of a

naturalistic scene or a 1D image of synthetic points and the fields of view are quite different) and so it is not possible to make a fair comparison of their performance in these tasks. Instead, our aim is to show how, in principle, a representation that does not include a 3D coordinate frame (which is true of both models) could, nevertheless, contain useful information relating to the distance of features, rather like Marr's idea of a $2\frac{1}{2}$-D sketch (Marr, 1982) and to demonstrate how this information could be useful in the tasks we examine. The way forward for these non-3D representations is clearly to build on the success of RL demonstrations such as Zhu et al. (2017), not simplistic handcrafted models, but, we argue, this development may be helped by considering ways to incorporate motion parallax information.

2. Methods

Our goal is to compare performance of two algorithms, one based on a learned representation, developed by Zhu et al. (2017), and one based on a hand-crafted representation. To analyse these methods, we use two different tasks: the first is to find the mid-point in space between two locations that have been learned (or are 'known') already; the second is to do the same in the orientation domain, i.e. to find the mid-bearing between two 'known' bearings. These tasks test for geometric consistency within a representation i.e., in this case, whether there is any implicit knowledge in the representation about locations or orientations other than the ones that have been learned about during training. We also probe the representations more directly, looking for systematic spatial organisation in the arrangement of the learned feature vectors when related to corresponding locations in space.

We begin with an account of the contribution Zhu et al. (2017) make in the context of reinforcement learning and describe how decoding can be used to query the information stored in the network. We then describe our hand-crafted representation which records information about the angles between pairs of visible points and about the extent to which these change as the optic centre translates. It is hardly a surprise that this representation performs well on geometric tasks, and we are not making a claim that this representation is in any sense 'better' than the learnt one - the representations are, after all, utterly different. Nevertheless, it is informative to compare the performance of the representations side-by-side in order to inform the debate about improving learned representations in future in a way that incorporates information that is particularly important to animals.

2.1. Reinforcement learning for visual navigation

Reinforcement learning (RL) (Sutton & Barto, 2018) is a framework for optimizing and reasoning about sequential decision-making. Tasks are modelled as Markov Decision Processes (MDPs), $\langle S, A, T, R, \gamma \rangle$ tuples where S represents the state space, A the set of actions, T: $S \times A \times S \rightarrow [0, 1]$, R the reward function R: $S \times A \times S \rightarrow \mathbb{R}$, and $\gamma \in [0, 1)$ a discount factor. Solving an MDP is defined as finding a policy $\pi(a|s) = p(A = a|S = s)$ that maximizes the expected discounted cumulative return $\sum_{k=0}^{\infty} \gamma^k r_{k+1}$.

Deep reinforcement learning (DRL) is an extension of standard RL in which the policy is approximated by a Deep Neural Network, and where RL algorithms are combined with stochastic gradient descent to optimise the parameters of the policy. Popular instances of DRL methods include: Deep Q-Network (DQN) (Mnih et al., 2015) and its variants (Hessel et al., 2018), which regress a state-action value function; policy gradient methods, which directly approximate the policy (Sutton, McAllester, Singh, & Mansour, 2000), and actor-critic methods (Silver et al., 2014; Mnih et al., 2016), which combine value-based methods with policy gradient algorithms to stabilize the training of these policies. DRL methods have been successful in solving complex tasks such as Go and other popular board games (Silver et al., 2017, 2018), and have proved to be necessary to tackle decision-making tasks with high-dimensional or visual state representation (Mnih et al., 2015; Levine, Finn, Darrell, & Abbeel, 2016). These breakthroughs in visual learning and control have also created a surge in work on *active vision* (Ruiz-del Solar, Loncomilla, & Soto, 2015), and several *visual-based navigation* (Savva et al., 2019) frameworks have recently been proposed to formalize and tackle many 3D navigation tasks.

We focus here on the task of goal-driven visual navigation, where the agent is asked to navigate to an entity in a high-fidelity 3D environment, given either an image of the entity, a natural language description, some coordinates, or other relevant information. As we set out in the Introduction, the case we have chosen to analyse is the one proposed by Zhu et al. (2017), which aims to solve the problem of learning a policy conditioned on both the target image and the current observation. The architecture is composed as follows: the observation and target images are generated using an agent in a virtual environment, AI2-THOR (Kolve et al., 2017). First, these images are passed separately through a set of siamese layers (which means that the parameters in the twinned networks are identical, despite the input to the two networks being different) (Chopra, Hadsell, & LeCun, 2005). These are based on a pretrained ResNet-50 network and have a feedforward layer, embedding these images into the same embedding space. These embeddings are then concatenated and further passed through a fusion layer, which outputs a joint representation of the state. The joint representation is finally sent to *scene-specific* feedforward layers, which produce a policy output and a value as required by a standard actor-critic model (see Fig. 1). This split architecture allows for the embedding layers to focus on providing a consistent representation of the MDP instance based on the goal and the agent's observation, while providing capacity to the network to create separate feature filters that can condition on specific scene features such as map layouts, object arrangement, lighting, and visual textures, thus obtaining the capability to arbitrarily generalize across many different scenes.

2.2. Knowledge decoder

It is not possible to tell from the architecture described in the previous section whether any of the environment properties that are available in a 'cognitive map' (e.g., location/orientation of the target, agent position, angles to the target) are present in the transformations encoded in the network's weights. To test whether information about location and orientation is encoded, we trained a decoder which takes the agent's internal representation as input and outputs one of the desired properties, such as (x, y) coordinates of a chosen observation. More specifically, to build the dataset we used Zhu et al. (2017)'s architecture as described above. This generates an embedding up to the final feedforward layer (before it gets sent into the policy and value heads) for each target-observation pair of the training set, while also recording the agent's (x, y) coordinates and angle θ. We primarily employ multi-layer perceptrons (MLPs) to perform this decoding. MLPs characterise flexible non-linear functions, and are constructed by interleaving linear transformations with non-linear activations/transformations (e.g. ReLU or TanH). The decoder is a 2-layer MLP in the case of a single value regressor (i.e., the angle), or a 3-layer MLP with multiple "heads"—additional MLPs to split common computation—when regressing to (x, y) coordinates or to the orientation, θ, of the agent. We use an MSE loss trained with Adam (Kingma & Ba, 2014), together with dropout (see Table A1 for hyperparameters).

2.3. Relative visual direction (RVD) representation

This section describes a simple representation of the angles between pairs of points around the observer. It is not learned, like the Zhu et al. (2017) representation; it is hand-crafted and it contains all the information that would be required to reconstruct the 3D structure of the scene. However, it does not do that. Instead, it keeps the information in a relatively raw state so that current and stored states can be compared in a high dimensional space, just as they are in the Zhu et al. (2017)

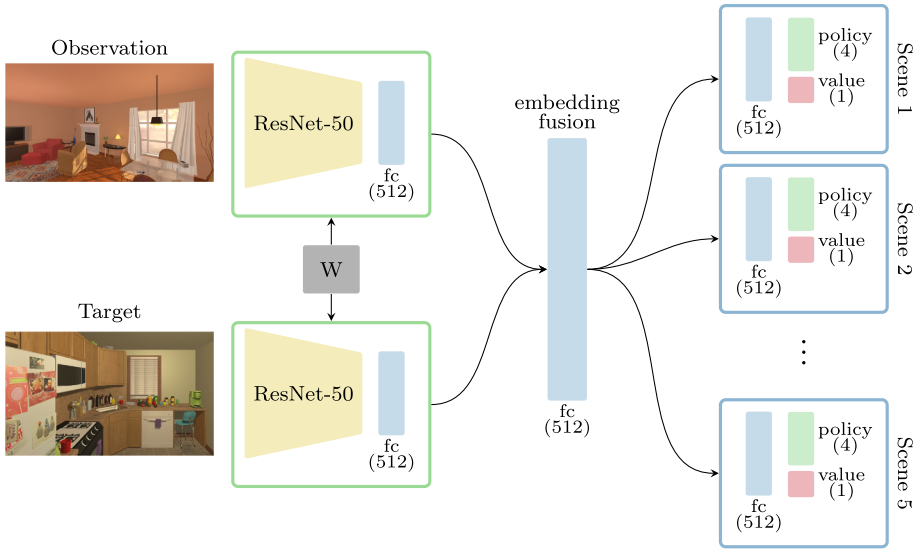


Fig. 1. Architecture of the Zhu et al. (2017) siamese network. See text for details.

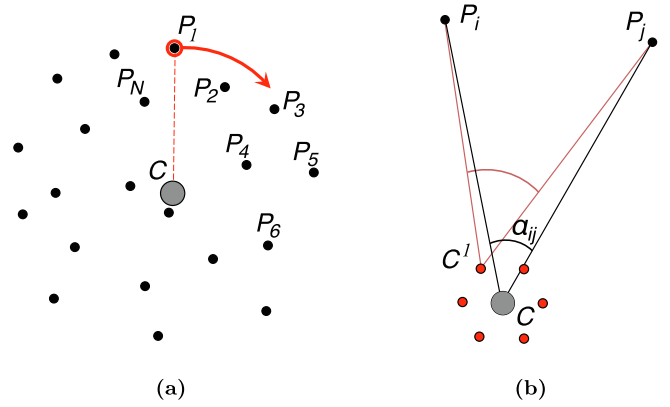


Fig. 2. (A) 2D scene containing N random points and camera C in the centre. The points are ordered clockwise in angular sense with respect to the reference point P_1, which is marked red. (B) Angular and parallax features. P_i and P_j are scene points, (C) is camera location, C^1–C^6 are sub-cameras. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

representation. As discussed in the Introduction, information about relative visual directions and changes in relative visual direction are important in biological vision and are key to this representation. Fig. 2a shows a 2D scene containing an optic centre, C, surrounded by N random points $P_1, ..., P_N$. The points in the scene are numbered and ordered clockwise with respect to the first point, P_1, marked in red (this is relevant for the mid-bearing task described later). The angle subtended by a pair of points (P_i, P_j) at the optic center, C, indicating the relative visual direction, is denoted $\alpha_{ij} = \angle P_i C P_j$ (such that $\alpha_{ji} = \angle P_j C P_i = 2\pi - \alpha_{ij}$). The vector of all such angles, between every possible pair of points P_i and P_j viewed from the optic center C is denoted ε (Fig. 2b). We assume an omnidirectional view with no occlusions. The dimensionality of ε is thus $M = N^2 - N$, since we exclude angles between a point and itself. The elements of ε are ordered in a particular way, following

$$\varepsilon = \{\alpha_{ij}: i = 1, ..., N, j = (i+1), ..., N, 1, ..., (i-1)\}. \quad (1)$$

The reason ε is ordered in such a manner is to assist in extracting subsets of elements when the task relates to visual direction (2.5). However, elements of ε can be indexed in other ways, as the next section shows.

2.4. Mid-point for translation of the camera

Although ε contains all possible angular features, for certain tasks such as interpolating between locations some angular features are more informative than others. In particular, angular features that arise from pairs of distant points are more stable (i.e. vary less) during translation of the optic centre and thus are more useful for the interpolation task than are the angles between nearby points since these vary rapidly with optic centre translation. First, we extract a subset of the elements of ε using a criterion based on parallax information. We define a measure of parallax that assumes we have access to more views of the scene, as if the camera has moved by a small amount as shown in Fig. 2b. For such individual 'sub-cameras', C^k, $k = 1, ..., n_C$, where n_C is the number of sub-cameras, we can construct angular feature vectors ε_{C^k} similar to that constructed at the optic centre, ε, with exactly the same ordering of elements. A 'mean parallax vector', ψ, can then be computed from the difference between these sub-camera views, C^k, and the original view at C.

$$\psi = \{\psi_n\}_{n=1}^{N} = \frac{1}{n_C} \sum_{k=1}^{n_C} \frac{\varepsilon - \varepsilon_{C^k}}{\varepsilon} \quad (2)$$

Since ψ has the same ordering of elements as ε, each element of ψ contains a parallax-related measure referring to that particular pair of points.

It will prove useful to identify the pairs of points that are more distant, using the observation that the parallax values recorded in ψ are small in these cases. For a particular threshold value T_ψ on parallax, we define ρ as the *mask* on ψ, such that $\rho_i = 1$, if $\psi_i \leqslant T_\psi$, to identify the subset of ε with relatively small parallax values as $\varepsilon \odot \rho$. These elements of ε are, by design, those that are likely to change relatively slowly as the camera moves over larger distances.

2.5. Mid-bearing for rotation of the camera

We now consider a task of interpolating between camera *bearing* (viewing direction), rather than location. The goal is to estimate a bearing that is half way between two given views of the camera. A view, $\vartheta^{\theta,\omega}$, in this context, involves both a bearing, θ, and an angular range, ω, specifying the field of view for that camera (here, taken to be a fixed value of 90°) and is defined as a list of all the elements of ε that appear within that field of view. Note that the goal here is closely related, but not identical, to the task in the previous section of picking out an entire view that is half-way between two given views captured from different

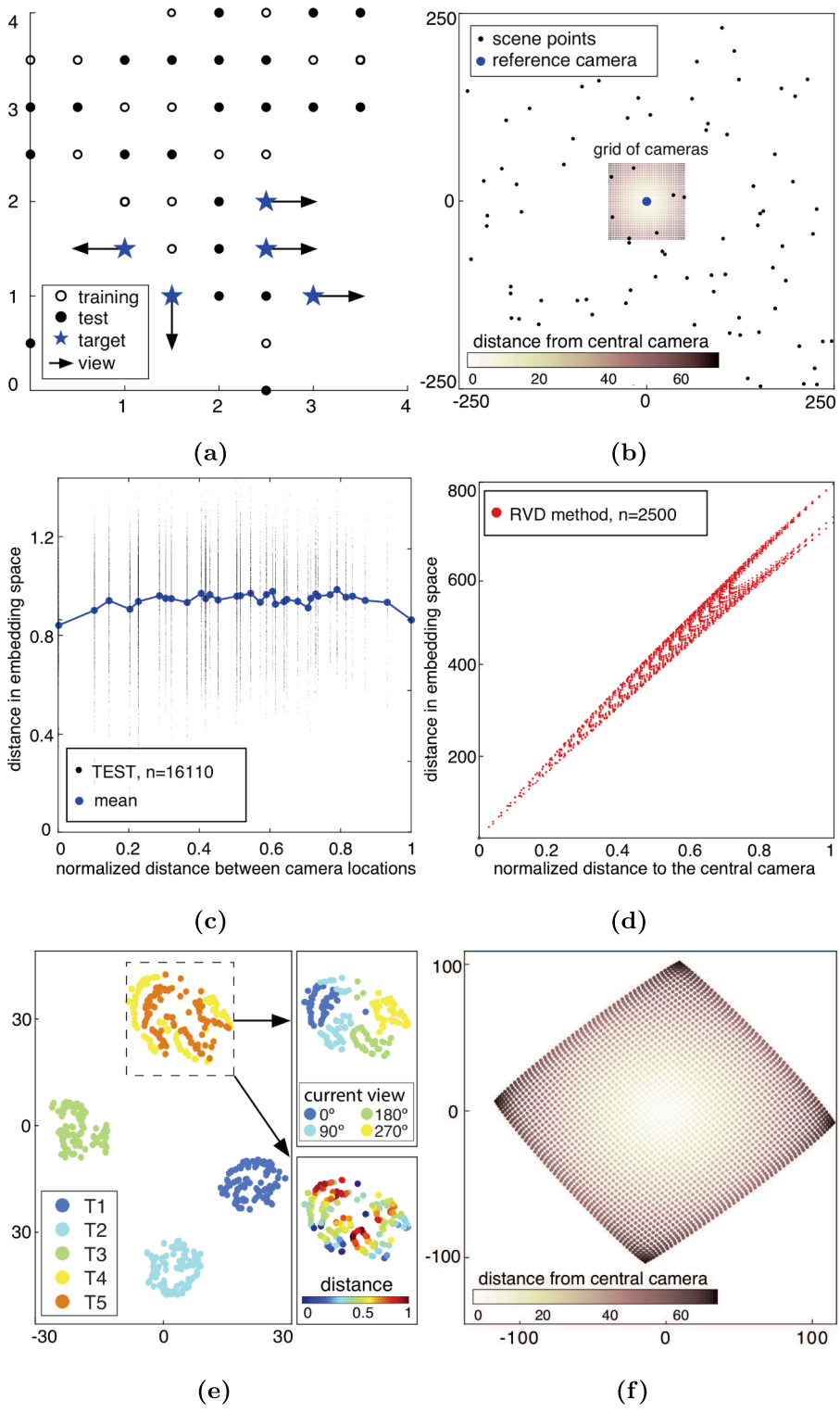


Fig. 3. Relationship between scene location and feature vectors for Zhu et al. (2017) and the relative visual direction (RVD) method. (a) shows a plan view of the Bathroom scene in Zhu et al. (2017). Open circles show the camera locations for images used in the training set, closed circles show the locations used in the test set. Blue stars and black arrows show the location and viewing direction of the camera for the target images. (b) An example of a random 2D scene with N = 100 points used in the RVD model. Cameras are placed in the middle of the scene as a 50 × 50 grid, which is 1/5 of the scene. The colour of each camera location indicates the distance of the camera from the central camera, *C*. For each of the 2500 camera locations we calculated a vector, ε, describing the angle between pairs of scene points as viewed from that camera (see Methods). (c) For the Zhu et al. (2017) method, the Euclidean distance in $\mathbb{R}^{512}$ between pairs of embedded feature vectors is plotted against the separation between the corresponding pairs of camera locations in the scene. (d) For the 'relative visual direction' (RVD) method, the Euclidean distance between the feature vectors for each camera and the feature vector for the central camera, *C*, is plotted against the separation between the corresponding pairs of camera locations in the scene. (e) A t-SNE plot that projects the stored feature vectors in the Zhu et al. (2017) network into 2D (see text for details). (f) Same as (e) but now for the RVD model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

locations. The way ε is organised, such that elements are ordered by reference point (see Eq. (1)), means that there is a consistent (albeit approximate) relationship between the index of the element and the bearing of the reference point for that element.

To consider all the elements of ε that appear in a given view we construct a mask, κ, similar to ρ above, but now the mask is based on

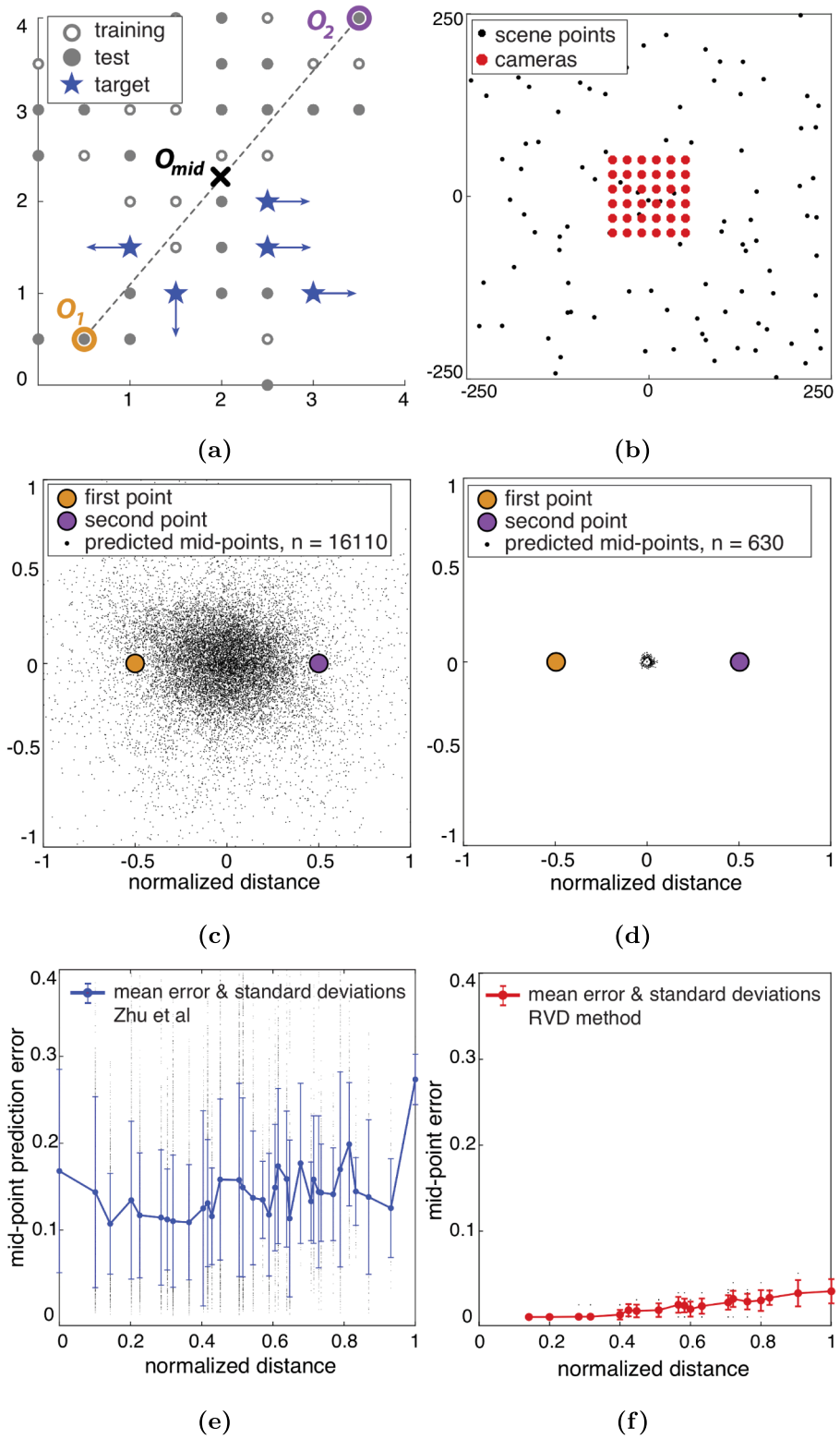
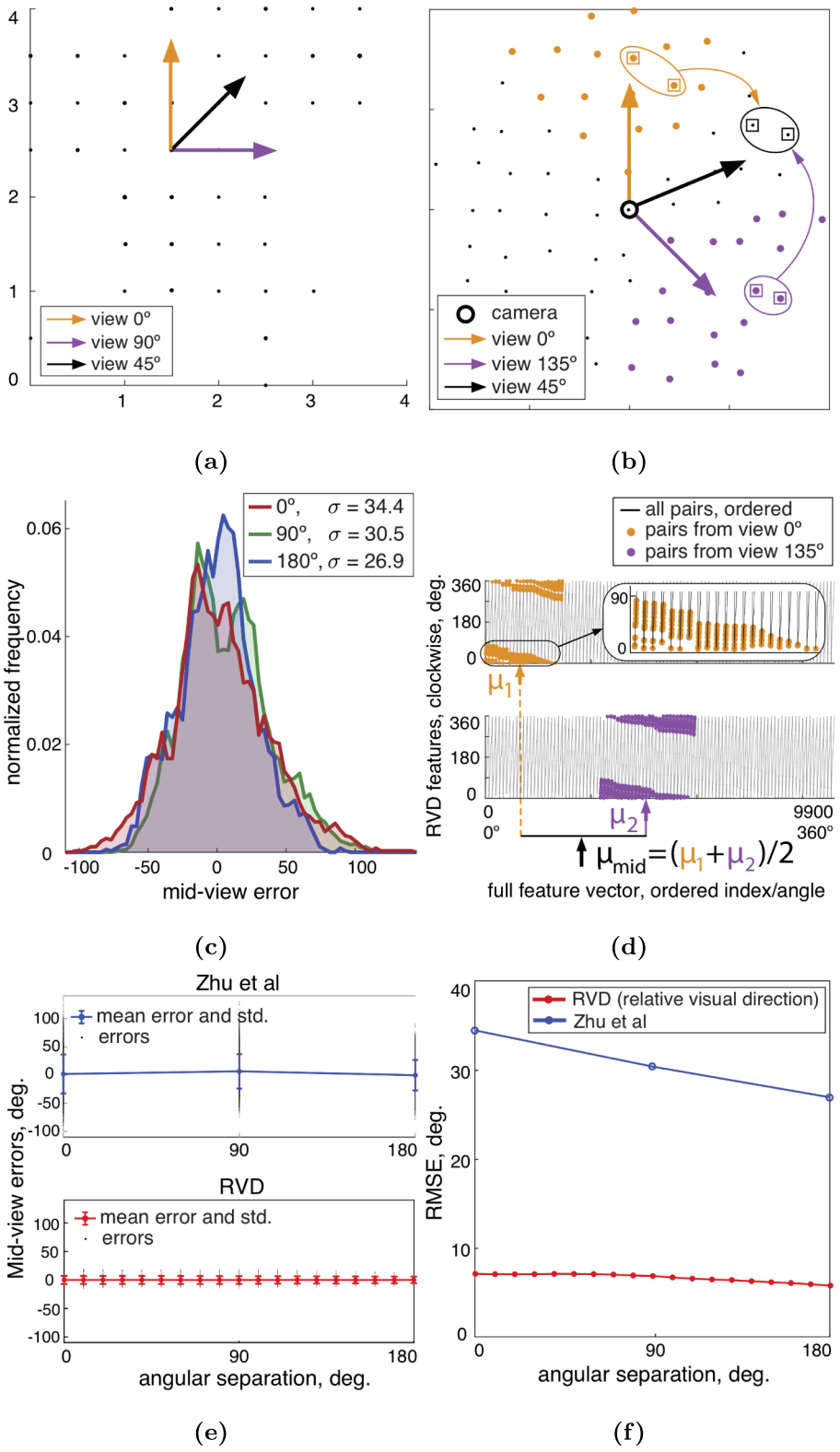


Fig. 4. Estimate of midpoints between pairs of observation locations. (a) shows the Bathroom scene with two observation locations, O_1 and O_2, and a midpoint, O_{mid}. (b) shows a random 2D scene with a 6 × 6 grid of cameras in the middle. For each camera, we calculated a feature vector $\varepsilon \odot \rho$ (see Section 2.4). (c) shows the estimated midpoints for all possible pairs of observations (where an observation is defined as a location, orientation and target), using Zhu et al. (2017) feature vectors and decoding (see Section 2.2). Orange and purple circles show the normalised location of the two observation locations and the black dots show, in this normalised coordinate frame, the location of the estimated midpoints. (d) shows the same as (c) but for the feature vectors in the RVD model. The black dots show midpoints for all possible pairs of camera locations. (e) shows the midpoint prediction error from (c) (absolute errors) plotted against the separation of the observation locations (O_1 and O_2). The separation between observation locations is normalised by the maximum possible location of two observation locations in the room. Error bars show one standard deviation. (f) shows the same for the RVD method. We considered all possible pairs of cameras (n = 630). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(caption on next page)

Fig. 5. Estimate of new views at an orientation half way between learned views. (a) shows a plan view of a bathroom scene in Zhu et al. (2017) and the 45 locations the camera could occupy. Orange and purple arrows indicate two camera orientations and the black arrow indicates an orientation halfway between these (not used in Zhu et al. (2017)). (b) Similar to (a) but for the RVD method. Points visible in views 0° (north) and 135° (south-east) are marked as orange and purple circles, where the field of view (ω) is limited to 90°. The ground-truth mid-view is indicated by the black arrow (see text). (c) Distribution of errors in computing the mid-view orientation from a decoding of orientation in the Zhu et al. (2017) trained network. Red, green and blue distributions are for camera orientations separated by 0, 90 and 180 degrees respectively. (d) Full vector of angular features, ε, (black saw-tooth plot). The y-axis shows the magnitude the elements in ε, i.e. the angle between pairs of points. The x-axis represents indices of the vector's elements (9900 in this case) see Eq. (1). The x-axis also provides an approximate indication of visual direction, from 0° to 360°, see text. The elements that correspond to pairs of points visible in the north and south-east views are marked with orange and purple circles respectively (see inset). Mid-indices μ_1 and μ_2 are marked as orange and purple arrows, while the index of the predicted mid-view μ_{mid} is marked as a black arrow. (e) All the mid-view errors for the Zhu et al. (2017) method for camera orientations separated by 0, 90 and 180 degrees. Mean and standard error shown in blue. Plot below shows the same for the RVD method. Mean and standard error shown in red. (f) Shows the RMSE error of predicted mid-view with respect to the ground truth as a function of angular separation between the views. For the RVD method we considered views separated by many different angles (in increments of 10°), while for Zhu et al. (2017) the data limited analysis to only three separations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 6. Visual servo-ing to maintain postural stability. Looking straight out on the mountains, almost all motion parallax is removed because the scene is distant and so cannot drive postural reflexes. In a normal scene, there are objects visible at a range of distances, giving rise to both large and small magnitudes of motion parallax. Removal of close objects in this scene has the same effect as setting T_ψ to mask out all but the lowest parallax elements of ε in the RVD representation. This is one example, in addition to the two examined in Figs. 4 and 5, where indexing different elements of ε and monitoring changes in those elements is helpful for accomplishing a task. License to use Creative Commons Zero – CC0.

whether both scene points P_i and P_k that define an element in ε are visible in a particular view: $\kappa^{\theta,\omega} = \{\kappa_j\}_{j=1}^{N}$, where $\kappa_j = 1$, if $P_i, P_k \in \vartheta^{\theta,\omega}$, where $\varepsilon^j = \alpha_{ik} = \angle P_i C P_k$. The relevant elements are denoted $\varepsilon \odot \kappa^{\theta,\omega}$. Given two such views $\vartheta_i^{\theta_i,\omega}$ and $\vartheta_j^{\theta_j,\omega}$, we can use the indices of the elements in each view to estimate the indices of the view that is mid-way between the two (Section 3.3).

3. Results

Figs. 3–5, show the results for three comparisons between the models. Fig. 3 relates physical distance between locations to the separation of corresponding feature vectors in the representation. Fig. 4 illustrates the ability of both models to interpolate correctly between the representation of two learned/known locations while Fig. 5 does the same for interpolation between two learned/known visual directions.

3.1. Correlation between physical separation and feature separation in the representation

Fig. 3 compares the representation of a scene in the two models we have discussed, based on Zhu et al. (2017) (left hand column) or relative visual direction (RVD, right hand column). Fig. 3a shows a plan view of the scenes used by Zhu et al. (2017) (where filled and closed symbols show the camera locations at test and training) and Fig. 3b shows the 2D layout of scene points (black dots) and camera locations (coloured points) in a synthetic 2D scene that was used as input for the RVD method. In Zhu et al. (2017), the environment was a highly realistic 3D scene in which the agent was allowed to make 0.5 m steps and turn by 0, +90 or −90 degrees (figures are from the Bathroom scene, see Appendix for others). Target views are marked by blue stars and arrows. For the RVD method, we generated a random 2D scene with 100 points. Cameras were placed in the middle of the scene as a regular 50 × 50 grid, which occupied 1/5 of the scene (Fig. 3b). The colour indicates the distance of a camera from the central reference camera.

For each learned context in Zhu et al. (2017) (where a learned context is defined by an observation location, a camera orientation and a target), there is a corresponding feature vector (i.e. 20 feature vectors per location). These observation locations are the 'trained' locations illustrated by open circles in Fig. 3. Fig. 3c shows the Euclidean distance between pairs of feature vectors ($\mathbb{R}^{512}$) from the test set, for all possible pairings, and plots this distance against the distance between the corresponding observation locations ($\mathbb{R}^2$). Fig. 3c shows that there is only a weak correlation between distance in the embedding space and physical distance between observation locations for this scene in the Zhu et al. (2017) paper (Pearson correlation coefficient, R, is 0.09, see Fig. A3 for other scenes) whereas Fig. 3d shows that, for the RVD method, there is a clear positive correlation ($R = 0.99$). Zhu et al. (2017) quoted a correlation of 0.62 between feature vector separation and separation in room space, but we are only able to reproduce a similarly high correlation by considering the distance between pairs of feature vectors when the agent had the *same* goal and the *same* viewing direction ($R = 0.67$ for all such pairings in the Bathroom scene). By contrast, Fig. 3c refers to all possible pairings in the test phase.

The right hand column of Fig. 3 shows results of the 'relative visual direction' (RVD) model. At each camera location ($N = 2500$), we generated a truncated angular feature vector $\varepsilon \odot \rho$ (see Section 2.4) as a representation of the scene as viewed from that location. We used the 30^{th} percentile of the parallax values as a threshold for inclusion of elements (T_ψ), i.e. the truncated feature vectors contained only the elements of ε that corresponded to pairs of points with the smallest parallax values, where 'small' in this case means the bottom 30% when

ordered by parallax magnitude. The exact choice of threshold is not important; in the Appendix, Fig. A1, we show the same result for different values of this threshold. Using the *top* 30% of ε when ordered by parallax, or using the entire ε vector, gives rise to worse performance on the interpolation task. Note that we have used the same ordering of elements in $\varepsilon \odot \rho$ for all cameras. Specifically, the ordering of ε and ρ were established for the central reference camera and applied to all other cameras (see Eq. (1)).

Fig. 3e and f visualise the embedding space for the Zhu et al. (2017) and RVD representations respectively using a t-SNE projection (Maaten & Hinton, 2008). This projection attempts to preserve ordinal information about the Euclidean distance between high dimensional vectors when they are projected into 2-D. In Zhu et al. (2017) (Fig. 3e), feature vectors are clumped together in the t-SNE plot according to the agent's target image. Targets 4 and 5 were very similar images, so it is understandable that the feature vectors for locations with these targets are mixed (yellow and orange points). Although target is the dominant determinant of feature vector clustering, information about camera orientation and camera location is still evident in the t-SNE plot. The top-right sub-panel colour-codes the same T4/T5 cluster but now according to the orientation of the camera: this shows that orientation also separates out very clearly. Finally, there is also information in the t-SNE plot about camera location. Colours in the bottom right subplot indicate distance of the camera location from a reference point, (0,0); there is a gradation of colours along strips of a common camera orientation and this systematic pattern helps to explain why camera location can be decoded (see Section 3.2). For the RVD method, the configuration of feature vectors preserves the structural regularity of the camera positions, as can be seen from the t-SNE projection in Fig. 3f. We now explore how these differences affect the ability of each representation to support interpolation between learned/stored locations.

3.2. Interpolation between stored locations in the representation

Fig. 4 shows the results of the location interpolation task which was to estimate the mid-point between two locations (e.g. in Fig. 4a O_{mid} is halfway between O_1 and O_2) based on the midpoint between two feature vectors. For the Zhu et al. (2017) model, this requires a decoder for 2-D position learned from the stored feature vectors (see Section 2.2). The results are shown in Fig. 4c using a normalized scale to illustrate the errors relative to the two input locations.

For the RVD model, decoding is much more direct, as one would expect from the t-SNE plot (Fig. 3f). The details are as follows. Fig. 4b shows a random 2D scene with a 6 × 6 grid of cameras in the middle. For each camera C_j, $j = 1, \ldots, 36$, we calculated a feature vector ε_{C_j} and a parallax mask ρ_{C_j} as described in Section 2.4. The feature vector for the mid-point between two cameras C_i and C_j was computed as $\varepsilon_{C_i,C_j} = \frac{1}{2}((\varepsilon_{C_i} \odot \rho_{C_i}) + (\varepsilon_{C_j} \odot \rho_{C_j}))$. Then, to find the midpoint, we searched over a fine regular grid (step = 1) of camera locations to find the camera $C_k^{\star}$ that was best matched with the estimated feature ε_{C_i,C_j}, that is,

$$C_k^{\star} = \underset{c_k}{\operatorname{argmin}} \|\varepsilon_{C_i,C_j} - \varepsilon_{C_k}\| \tag{3}$$

This is equivalent to, but simpler than, the decoding stage using a MLP for the Zhu et al. (2017) model. Fig. 4d shows estimated mid-points calculated this way for all possible pairs of the 36 cameras (n = 630). For the Zhu et al. (2017) method, Fig. 4e shows the absolute errors relative to the true mid-point between O_1 and O_2 as a function of the separation between O_1 and O_2. Fig. 4f shows the same for the RVD method. As discussed in the Introduction, it is not fair to make a direct comparison between the magnitude of the errors for the two models given how different their inputs are but one can compare the way that the errors change with separation between O_1 and O_2. This shows a monotonic rise for the RVD model, as one would expect from a geometric representation, whereas this is not true for the Zhu et al. (2017) method (Fig. 4e).

3.3. Interpolation between stored viewing orientations in the representation

Fig. 5a shows the scene layout from Zhu et al. (2017) and two views from a single location. The goal in this case is to find an intermediate *bearing* (as shown by the black arrow) half way between the bearing of the two reference images (orange and purple arrows). Fig. 5c shows the error in the decoded mid-bearing when the input images are taken from views that are 0, 90 or 180° apart. Note that the two input images need not necessarily be taken from the same location in the room (either in training the decoder or in recovering a mid-bearing). Fig. 5c and e show that there is no systematic bias to the mid-bearing errors but the spread of errors is large compared to that for the 'relative visual direction' (RVD) method (Fig. 5f). The RVD method uses a very simple algorithm to estimate the mean bearing. It assumes that the ordering of elements in ε has a linear relationship to the bearing of a view, i.e. that as the bearing changes (going from orange view to purple view in Fig. 5d) the index of the corresponding elements in ε will change systematically and hence the mean index of the elements within a view is useful in determining the bearing of that view. This is not strictly true, but the fact that it is a useful approximation is because of the way that the vector, ε, was set up in the first place (Eq. (1)). In more detail, Fig. 5b and d shows how the bearing of a mid-view (θ_{mid}) is estimated using the over-simplified assumption that the bearing of the reference point in a pair of views varies linearly with index in ε. In fact, of course, the relationship between bearing and element index depends on the layout of the scene. The mean index of a view, $\vartheta^{\theta,\omega}$, is computed from its corresponding mask, $\kappa^{\theta,\omega}$, as the middle index, μ, of all 'on' mask elements, $\kappa_j = 1$, for that view. Given two views $\vartheta_i^{\theta_i,\omega}$ and $\vartheta_j^{\theta_j,\omega}$, we estimate a nominal bearing of the mid-view image, μ_{mid}, from the average of their mean indices:

$$\mu_{mid} = (\mu_i + \mu_j)/2. \tag{4}$$

and $\theta_{mid} \propto \mu_{mid}$.

This heuristic is illustrated in Fig. 5d. For the purposes of illustration only, this shows the i^{th} element in the orange image (pair of dots outlined in orange) and the i^{th} pair in the purple image (outlined in purple). Considering the indices of these two elements in ε, the rounded mean of these two indices gives an index to an element of , i.e. it corresponds to a pair of points. For the purposes of illustration, these are shown by the black squares in Fig. 5d which, in this case, happen to lie close to the mid-bearing direction. However, the heuristic simply reports the estimated orientation of the mid-view as described above (Eq. (4)). The bias and variability of the estimates of the mid-view in both models are shown in Fig. 5e and f respectively. Again, given the very different nature of the inputs to the two models, it is not fair to comment on the relative magnitude of errors in the two models. Neither model shows the Weber's law increase in errors with angular separation between μ_1 and μ_2 that we saw in Fig. 4f.

4. Discussion

There has been a long-standing assumption that the brain generates spatial representations from visual input and does so in a variety of 3D coordinate frames including eye-centred (V1), ego-centred (parietal cortex) or world-centred (hippocampus and parahippocampal gyrus). Computer vision and robotics research has also concentrated on algorithms that generate representations in a 3D frame (a world-based one). Biological models have not tried to recapitulate the complexities of photogrammetry (computing 3D structure from images) but instead have generally assumed that the generation of a 'cognitve map' relies on other inputs such as proprioceptive signals or pre-existing place cell or grid cell input, to provide spatial structure to the representation (Foster et al., 2000; Bush, Barry, Manson, & Burgess, 2015; Banino et al., 2018; Behrens et al., 2018).

We have chosen to examine in detail the RL method described by Zhu et al. (2017) for learning to navigate to an image using visual inputs alone, because this has now become a general method on which several more recent and complex algorithms have been based (Chen, DeAngelis, & Angelaki, 2011; Gupta, Davidson, Levine, Sukthankar, & Malik, 2017; Chen, Gupta, & Gupta, 2019; Kumar, Gupta, & Malik, 2019; Mirowski et al., 2018; Mirowski et al., 2016). We have compared the Zhu et al. (2017) representation to a hand-crafted representation (based on relative visual directions and using highly simplistic input) in order to illustrate two points. First, in Zhu et al. (2017), the relationship between stored feature vectors and the locations of the camera in the scene (Fig. 3a) is quite a complex one, while for the RVD model the relationship is simple and transparent. In the case of Zhu et al. (2017), it is possible to build a decoder to describe the mapping between feature vectors and location (as illustrated by the systematic distance information visible in Fig. 3e) but this is quite different from the smooth, one-to-one relationship between stored feature vectors and space illustrated in Fig. 3f, at least over the range of camera locations illustrated here (Fig. 3b). The decoding required to extract location from the Zhu et al. (2017) representation is reminiscent of the decoding that has been described as a way to use the aliased grid cell activity as a signal for location in rats (Bush et al., 2015), i.e. substantially more complex than the interpolation of the feature vectors of the RVD model which generates a sensible result directly (e.g. Fig. 4d). Like the decoding of location in the Zhu et al. (2017) model, interpreting the output of grid cells would need a sophisticated decoding mechanism if they were to be used on their own for navigation (Bush et al., 2015) and neural network implementations have been proposed to solve this problem. For example, it is possible to decode the distance and direction of a goal given high dimensional vectors ($\mathbb{R}^{512}$) of grid cell activity at the current and goal locations (Banino et al., 2018) but grid cell firing rates are not the only high dimensional vectors encoding spatial location that could be used. The vector ε that we have described in this paper would be likely to do equally well and potentially even better since the aliased nature of grid cell firing is a disadvantage rather than an advantage in this context.

Answering the question 'where am I?' does not necessarily imply a coordinate frame (Gillner & Mallot, 1998; Glennerster, 2016; Glennerster, Hansard, & Fitzgibbon, 2009; Rosenbaum, Besse, Viola, Rezende, & Eslami, 2018; Warren, 2019). Instead, one can offer a restricted set of alternative hypotheses. These potential answers to the question may correspond to widely separated locations in space, in which case the catchment area of each hypothesis is large, but the answer can be refined by adding more alternatives (i.e. more specific hypotheses about where the agent is). This makes the representation of space hierarchical (Hirtle & Jonides, 1985; Wiener & Mallot, 2003; Milford & Wyeth, 2010) and compositional in the following sense. Consider the RVD representation of a scene that includes distant objects such as the stars or the mountains in Fig. 6. The angles between these (which are elements of ε) do not change, however much the observer moves. If the objects are stars, then the catchment area of the hypothesis covers the whole Earth. Adding in objects that are nearer than the mountains refines the catchment area and this can be done progressively, providing a more and more accurate estimate of the location of the observer (hence, the representation is compositional) as elements with higher parallax are added to ε. This provides a hierarchy of hypotheses about location, from coarse to fine, without generating a 3D coordinate frame.

5. Conclusion

Biological models of spatial representation have often assumed that the brain builds a reconstruction of the world using an allocentric (world-based) or ego-centric 3-dimensional coordinate frame. The representations we have examined here are different in that they store high dimensional vectors describing the sensory information (and, in the case of Zhu et al. (2017), also the agent's goal) at each location. Given that this type of representation is being used increasingly in deep reinforcement learning implementations of agents that are capable of predicting novel views of scene, route-following and taking short-cuts (Banino et al., 2018; Eslami et al., 2018; Mirowski et al., 2016), this type of model is an important existence proof that there are alternatives to 3-dimensional coordinate frame hypotheses of spatial representation. We have shown here how, in developing high dimensional features to represent images, it can be advantageous to introduce information about the distance of features and, especially, to identify elements of the input that are likely to be long- or short-lived in the scene as the camera translates.

CRediT authorship contribution statement

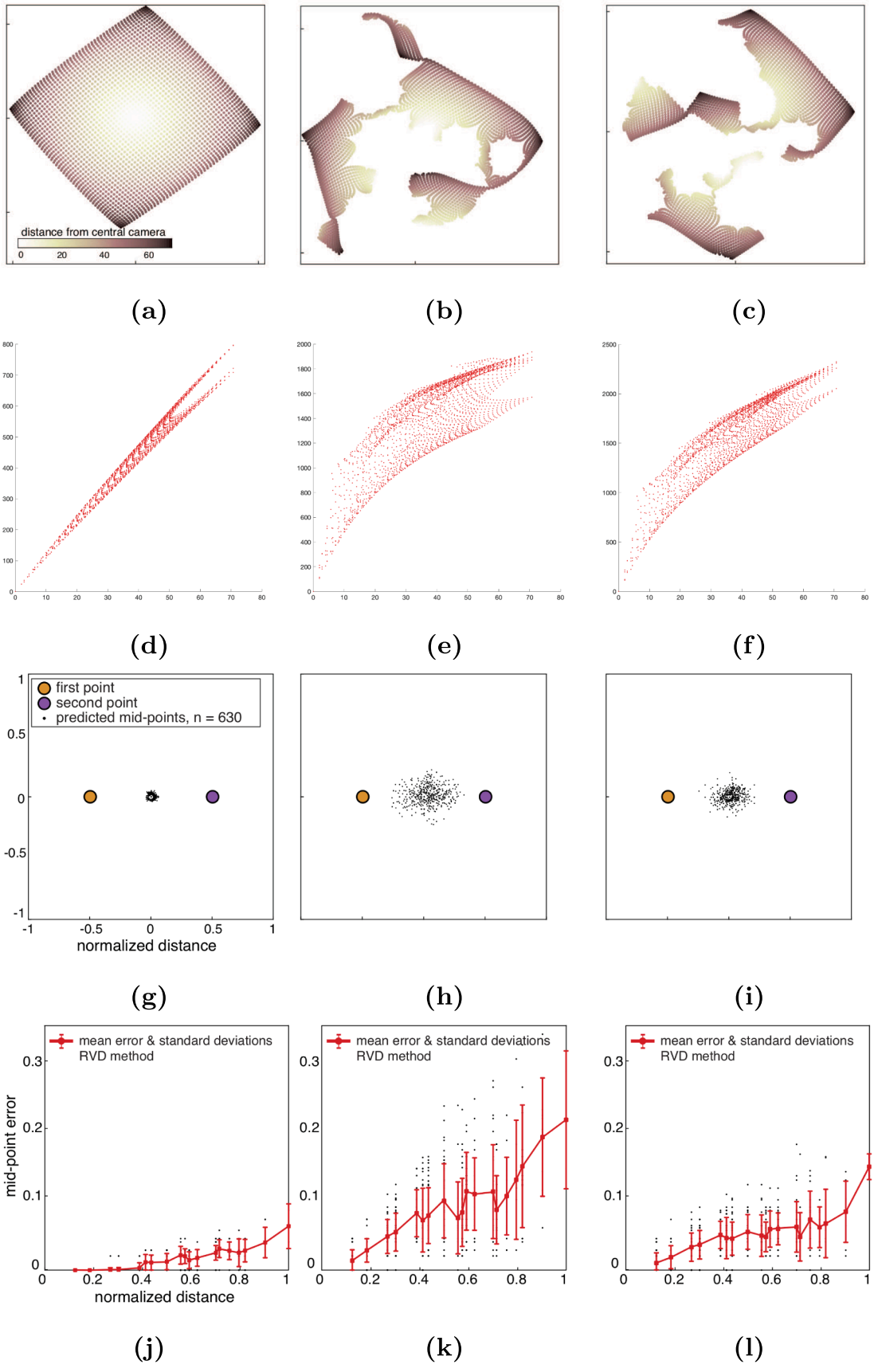
Alex Muryy: Conceptualization, Formal analysis, Software, Writing - original draft, Writing - review & editing. **N. Siddharth:** Conceptualization, Formal analysis, Software, Writing - original draft, Writing - review & editing. **Nantas Nardelli:** Conceptualization, Formal analysis, Software, Writing - original draft, Writing - review & editing. **Andrew Glennerster:** Conceptualization, Formal analysis, Software, Writing - original draft, Writing - review & editing. **Philip H. S. Torr:** Funding acquisition, Conceptualization.

Acknowledgements

We are grateful to Abhinav Gupta for providing code and advice and to Aidas Kilda and Andrew Gambardella for their help. This research was supported by EPSRC/Dstl grant EP/N019423/1 (AG). PHST, NS, & NN were supported by EPSRC/MURI grant EP/N019474/1. PHST was additionally supported by ERC grant ERC-2012-AdG 321162-HELIOS and EPSRC grant Seebibyte EP/M013774/1, and would also like to acknowledge the Royal Academy of Engineering and FiveAI.

Appendix A

See Figs. A1, A2, A3 and Table A1.



(caption on next page)

Fig. A1. Consequences of using large-parallax elements in the RVD model. (a) re-plots the t-SNE projection of the RVD feature vectors from Fig. 3f. (b) shows the disruption in the representation caused by using a different subspace of ε, namely picking out 30% of the elements of ε that have the *greatest* magnitude of motion parallax (Eq. (2)) rather than the smallest parallax, as we have used in all the previous figures. (c) shows the effect of using *all* of ε rather than a subspace. (d)–(f) show the distance between feature vectors plotted against distance to the central camera (see Fig. 3d) using the feature vectors illustrated in (a)–(c) respectively. (g)–(i) show the consequence of using the vectors illustrated in (a)–(c) for the mid-point task (so (i) is a repeat of Fig. 4d). (j)–(l) show the magnitude of the midpoint errors, following the format of Fig. 4f.

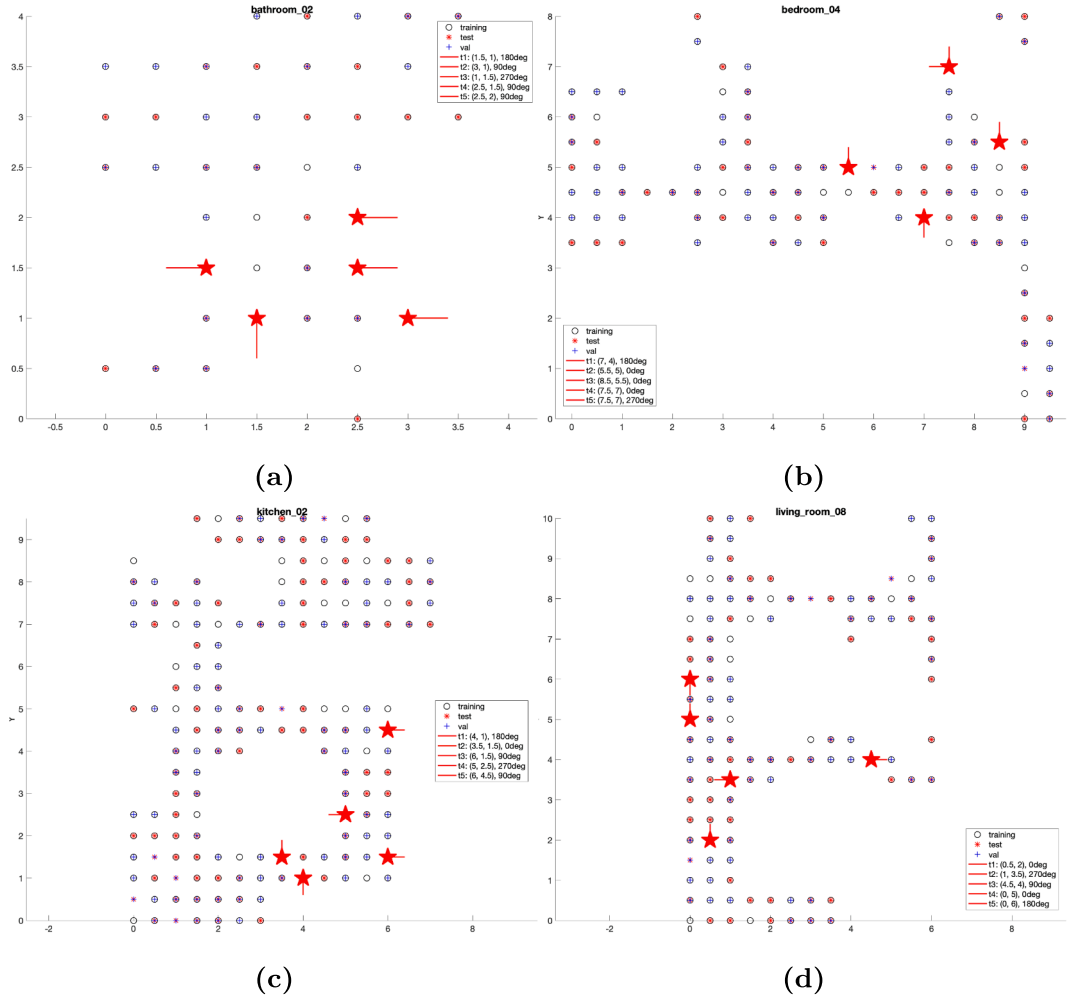


Fig. A2. Plan views of all 4 scenes used by Zhu et al. (2017). (a) bathroom, (b) bedroom, (c) kitchen, (d) living room. Symbols are as for Fig. 3a.

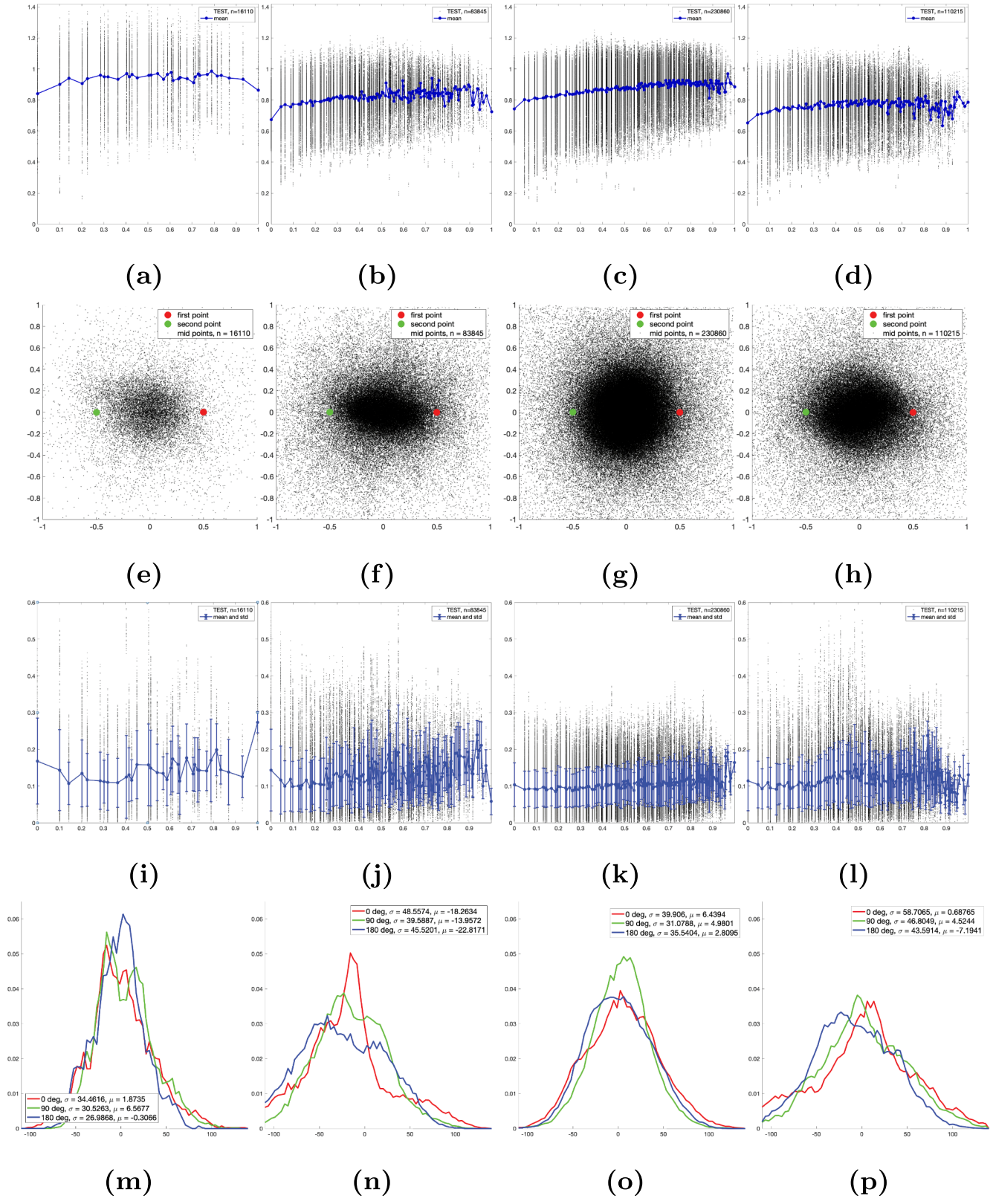


Fig. A3. Results for the bathroom scene were shown in Figs. 3–5 and are re-plotted here (left hand column). Results for the bedroom, kitchen and living room are shown in columns 2 to 4 respectively. In the top row, (a–d), the correlations, R, are 0.088, 0.22, 0.24 and 0.14 respectively. For details of what is plotted in (e–h) see Fig. 4c, for (i–l) see Fig. 4e, and for (m–p) see Fig. 5c.

Table A1
Hyperparameters for the original trained network and the two decoder networks. The original trained network from Zhu et al. (2017) was used throughout the paper, eg the t-SNE plot in Fig. 3e. The position decoder was used for the results shown in Fig. 4. The angle decoder was used for Fig. 5.

Default parameters for Adam	
β_1	0.9
β_2	0.999
ϵ	10^{-8}
use-locking	False
Position decoder	
learning rate	0.00001
λ_{L2}	0
Viewing angle decoder	
learning rate	0.0005
λ_{L2}	0.04

References

Acharya, L., Aghajan, Z. M., Vuong, C., Moore, J. J., & Mehta, M. R. (2016). Causal influence of visual cues on hippocampal directional selectivity. *Cell, 164*(1–2), 197–207.

Arleo, A., & Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biological Cybernetics, 83*(3), 287–299.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature, 557*(7705), 429–433.

Barrera, A., & Weitzenfeld, A. (2008). Biologically-inspired robot spatial cognition based on rat neurophysiological studies. *Autonomous Robots, 25*(1–2), 147–169.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron, 100*(2), 490–509.

Anderson, P., Chang, A. X., Chaplot, D. S., Dosovitskiy, A., Gupta, S., Koltun, V., et al. (2018). On evaluation of embodied navigation agents. CoRR abs/1807.06757, URL: http://arxiv.org/abs/1807.06757.

Bodenheimer, B., Meng, J., Wu, H., Narasimham, G., Rump, B., McNamara, T. P., et al. (2007). Distance estimation in virtual and real environments using bisection. In *Proceedings of the 4th symposium on applied perception in graphics and visualization* (pp. 35–40).

Bradshaw, M. F., Parton, A. D., & Glennerster, A. (2000). *The task-dependent use of binocular disparity and motion parallax information* 40, 3725–3734.

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using grid cells for navigation. *Neuron, 87*(3), 507–520.

Chatila, R., & Laumond, J.-P. (1985). Position referencing and consistent world modeling for mobile robots. In *Proceedings. 1985 IEEE international conference on robotics and automation* (pp. 138–145). Vol. 2, IEEE.

Chen, A., DeAngelis, G. C., & Angelaki, D. E. (2011). Convergence of vestibular and visual self-motion signals in an area of the posterior sylvian fissure. *Journal of Neuroscience, 31*(32), 11617–11627.

Chopra, S., Hadsell, R., LeCun, Y., et al. (2005). Learning a similarity metric discriminatively, with application to face verification. *CVPR, 1*, 539–546.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T.H., & Bengio, Y. (2018). BabyAI: First steps towards grounded language learning with a human in the loop, arXiv preprint arXiv:1810.08272.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., et al. (2019). Habitat: A platform for embodied ai research, arXiv preprint arXiv:1904.01201.

Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., et al. (2016). Learning to navigate in complex environments, arXiv preprint arXiv:1611.03673.

Chen, T., Gupta, S., & Gupta, A. (2019). Learning exploration policies for navigation, arXiv preprint arXiv:1903.01959.

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *ICCV* (pp. 1403–1410).

Dhiman, V., Banerjee, S., Griffin, B., Siskind, J. M., & Corso, J. J. (2018). A critical investigation of deep reinforcement learning for navigation. CoRR abs/1802.02274, URL: http://arxiv.org/abs/1802.02274.

Edwards, A. D. (2017). *Perceptual goal specifications for reinforcement learning.* Ph.D. thesisGeorgia Institute of Technology.

Erkelens, C. J., & Collewijn, H. (1985). Motion perception during dichoptic viewing of moving random-dot stereograms. *Vision Research, 25*, 583–588.

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., et al. (2018). Neural scene representation and rendering. *Science, 360*(6394), 1204–1210.

Foo, P., Warren, W. H., Duchon, A., & Tarr, M. J. (2005). Do humans integrate routes into a cognitive map? Map-versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning, Memory, and Cognition* ***31***(2), 195.

Foster, D., Morris, R., & Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus, 10*(1), 1–16.

Franz, M. O., Schölkopf, B., Mallot, H. A., & Bülthoff, H. H. (1998). Learning view graphs for robot navigation. *Autonomous Robots, 5*, 111–125.

Fuentes-Pacheco, J., Ruiz-Ascencio, J., & Rendón-Mancha, J. M. (2015). Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review, 43*(1), 55–81.

Gillner, S., & Mallot, H. A. (1998). Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience, 10*(4), 445–463.

Glennerster, A. (2016). A moving observer in a three-dimensional world. *Philosophical Transactions of the Royal Society B, 371*(1697) 20150265.

Glennerster, A., & Read, J. C. (2018). A single coordinate framework for optic flow and binocular disparity, arXiv preprint arXiv:1808.03875.

Glennerster, A., Hansard, M. E., & Fitzgibbon, A. W. (2001). Fixation could simplify, not complicate, the interpretation of retinal flow. *Vision Research, 41*, 815–834.

Glennerster, A., Hansard, M. E., & Fitzgibbon, A. W. (2009). View-based approaches to spatial representation in human vision. *Lecture Notes in Computer Science, 5064*, 193–208.

Gupta, S., Davidson, J., Levine, S., Sukthankar, R., & Malik, J. (2017). Cognitive mapping and planning for visual navigation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2616–2625). .

Hafting, T., Fyhn, M., Molden, S., Moser, M-B, & Moser, E. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature, 436*(7052), 801–806 http://www.ncbi.nlm.nih.gov/pubmed/15965463.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., et al. (2018). Rainbow: combining improvements in deep reinforcement learning. In *Thirty-Second AAAI conference on artificial intelligence* (pp. 3215–3222).

Hirtle, S. C., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition, 13*(3), 208–217.

Kanitscheider, I., & Fiete, I. (2017). Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems. In *Advances in neural information processing systems* (pp. 4529–4538).

Kinchla, R. (1971). Visual movement perception: a comparison of absolute and relative movement discrimination. *Perception & Psychophysics, 9*(2), 165–171.

Klatzky, R. L., Beall, A. C., Loomis, J. M., Golledge, R. G., & Philbeck, J. W. (1999). Human navigation ability: tests of the encoding-error model of path integration. *Spatial Cognition and Computation, 1*(1), 31–65.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., & Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience, 29*(31), 9771–9777.

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research, 17*(1), 1334–1373.

Kumar, A., Gupta, S., & Malik, J. (2019). Learning navigation subroutines by watching videos, arXiv preprint arXiv:1905.12612.

Kolve, E., Mottaghi, R., Gordon, D., Zhu, Y., Gupta, A., & Farhadi, A. (2017). Ai2-thor: An interactive 3d environment for visual ai, arXiv preprint arXiv:1712.05474.

Singh, A., Yang, L., Hartikainen, K., Finn, C., & Levine, S. (2019). End-to-end robotic reinforcement learning without reward engineering. CoRR abs/1904.07854, URL: http://arxiv.org/abs/1904.07854.

Rosenbaum, D., Besse, F., Viola, F., Rezende, D.J., & Eslami, S. (2018). Learning models for visual 3D localization with implicit mapping, arXiv preprint arXiv:1807.03149.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* ***9***(Nov), 2579–2605.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freemann and Company. ISBN 0262514621.

Milford, M., & Wyeth, G. (2010). Persistent navigation and mapping using a biologically inspired SLAM system. *The International Journal of Robotics Research, 29*(9), 1131–1153.

Milford, M., & Wyeth, G. (2010). Persistent navigation and mapping using a biologically inspired SLAM system. *The International Journal of Robotics Research, 29*(9), 1131–1153.

Mirowski, P., Grimes, M., Malinowski, M., Hermann, K.M., Anderson, K., Teplyashin, D., et al. (2018). Learning to navigate in cities without a map. In *Advances in neural information processing systems* (pp. 2419–2430).

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning* (pp. 1928–1937). .

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529.

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.

Purdy, J., & Gibson, E. J. (1955). Distance judgment by the method of fractionation. *Journal of Experimental Psychology, 50*(6), 374.

Regan, D., Erkelens, C. J., & Collewijn, H. (1986). Necessary conditions for the perception of motion in depth. *Investigative Ophthalmology & Visual Science, 27*(4), 584–597.

Rieser, J. J., Ashmead, D. H., Talor, C. R., & Youngquist, G. A. (1990). Visual perception and the guidance of locomotion without vision to previously seen targets. *Perception, 19*(5), 675–689.

Rothkopf, C. A., & Ballard, D. H. (2013). Modular inverse reinforcement learning for visuomotor behavior. *Biological Cybernetics, 107*(4), 477–490.

Ruiz-del Solar, J., Loncomilla, P., & Soto, N. (2015). A survey on deep learning methods for robot vision, arXiv preprint arXiv:1803.10862.

Sermanet, P., Xu, K., & Levine, S. (2016). Unsupervised perceptual rewards for imitation learning. CoRR abs/1612.06699, URL: http://arxiv.org/abs/1612.06699.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning* (pp. 387–395).

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science, 362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354.

Smeets, J. B., & Brenner, E. (2008). Grasping Weber's law. *Current Biology, 18*(23), R1089–R1090.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction.* MIT press.

Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems* (pp. 1057–1063).

Taube, S., Muller, U., & Ranck, B. (1990). Head-direction cells recorded from the post-subiculum in freely moving rats. I. Description and quantitative analysis. *The Journal of Neuroscience, 10*(2), 420–435.

Thomas, O. M., Cumming, B. G., & Parker, A. J. (2002). A specialization for relative disparity in V2. *Nature Neuroscience, 5*(5), 472–478.

Warren, W. H. (2019). Non-Euclidean navigation, *Journal of Experimental Biology* ***222***(Suppl 1), jeb187971.

Watt, R. J. (1987). Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *Journal of the Optical Society of America A, 4*, 2006–2021.

Watt, R. J. (1988). *Visual Processing: Computational, Psychophysical and Cognitive Research.* Hove: Lawrence Erlbaum Associates.

Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., et al. (2018). Unsupervised predictive memory in a goal-directed agent, arXiv preprint arXiv:1803.10760.

Westheimer, G. (1979). *Cooperative neural processes involved in stereoscopic acuity* 36, 585–597.

Wiener, J. M., & Mallot, H. A. (2003). 'Fine-to-coarse'route planning and navigation in regionalized environments. *Spatial Cognition and Computation, 3*(4), 331–358.

Yang, W., Wang, X., Farhadi, A., Gupta, A., & Mottaghi, R. (2018). Visual semantic navigation using scene priors. CoRR abs/1810.06543, URL: http://arxiv.org/abs/1810.06543.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., & Fei-Fei, L. (2017). Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3357–3364). IEEE.

Zhu, Y., Gordon, D., Kolve, E., Fox, D., Fei-Fei, L., Gupta, A., Mottaghi, R., & Farhadi, A. (2017). Visual semantic planning using deep successor representations. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 483–492). .