The Role of Embodiment in Learning Representations: Discovering Space Topology via Sensory-Motor Prediction

Oksana Hagen[1], **Swen Gaudl**[2]

[1]University of Plymouth, Drake Circus, Pl4 8AA, Plymouth, UK
[2]University of Gothenburg, 41296 Goteborg, Sweden
oksana.hagen@plymouth.ac.uk, swen.gaudl@gu.se

Abstract

This paper explores the crucial role of embodiment in learning representations for space topology in robotics. Embodiment, the ability of an agent to interact with its environment and receive sensory feedback, is fundamental to developing accurate and efficient representations. In this work, we investigate this by applying an action-conditional prediction algorithm to data collected from a simulated environment, aiming to learn the topology of the environment through sequences of random interactions. Using a simple mobile-robot-like scenario, by leveraging sensory-motor interactions we demonstrate how the agent can discover the topology of its environment. Our results demonstrate the importance of embodiment in the development of representations and potential applicability in robotic tasks, and a simple but effective method of integrating actions into a learning loop. We suggest that building abstract representations through the use of action-conditional prediction is a step towards unification of the representations used in robotics.

Introduction

The main distinction between learning representations for robotics and the majority of other machine learning domains is the presence of embodiment. In most machine learning domains, representation learning is performed on static data, neglecting the dynamic nature of real-world interactions. The presence of embodiment, the ability to actively engage with the environment and receive feedback, presents a unique challenge and an opportunity to develop useful and efficient representations for robotics. We aim to highlight the role of action information as a strong informational signal that should be leveraged when learning representations.

This paper explores the importance of embodiment for developing representations that can accurately describe the environment and be subsequently used for robotic tasks. We illustrate this idea by applying a simple action-conditional prediction representation learning algorithm to a set of data collected in a simulated environment to learn the environment topology from sequences of random interactions with the environment. When discussing the topology of an environment, we are refer to a property that describes, how different parts of the environment are connected or related to each other. This includes the overall structure of the environment in terms of connectivity and adjacency and is especially relevant to building unified representations of abstract state spaces, where the exact measures could be difficult to establish.

In robotics, a wide variety of definitions of embodiment exist, ranging from basic ones (such as the ability to move) to the requirement of real-time sensing in all the body elements (Wilson 2002). In this paper, we intentionally define embodiment in a very loose form: the agent can move in a continuous space, where a few simple principles of physical space hold, and it has a first-person view. While the actual environments used in the experiments are based on a simulation, we consider a mobile-robot-like scenario. Our results demonstrate how the accurate topology of the environment can be discovered by leveraging sensory-motor interaction patterns between the agent and the environment, even when the observation spaces are identical.

Embodied Representations

Embodiment refers to the concept that intelligence arises not just from the brain or an abstract mind, but through the intricate interactions between a physical body and its environment. While the debate about the role of embodiment, as a building block, or even requirement for the emergence of intelligence (for example, as argued by (Scheier and Pfeifer 1999)) is still ongoing, growing evidence supports the significant role of embodiment in cognition and intelligence. This perspective on embodiment highlights that a body (natural or mechanical) equipped with sensors and actuators, and its engagement with the surrounding context, play a key role in shaping the representations, internal models, and cognitive strategies that are learned.

The relationship between the agent's ability to act and observe the consequences of its action is considered a key building block for emerging intelligence within the predictive coding paradigm (Huang and Rao 2011), and related theories, such as free energy principle (Friston 2010; Schro¨dinger 1944) and sensory-motor contingency theory (Degenaar and O'Regan 2015). These theories emphasize that the physical embodiment of an agent —its ability to take actions in an environment and receive sensory feedback as a consequence of those actions — is foundational to its cognitive processes. This perspective asserts that an agent's intel-

ligence emerges through the continuous cycle of action and perception, highlighting the indispensable role of the embodiment in the emergence of intelligent behaviour.

Even in this weak sense of using actions to build representations, current state-of-the-art AI models are, at large, not embodied. Large language models (Brown et al. 2020) and diffusion models (Ho, Jain, and Abbeel 2020; Yang et al. 2023) are trained on large-scale datasets that are static in nature. This removes a crucial signal that could improve learning and leads to, among other things, problems with reasoning, that become apparent in, for example, video generation tasks, and further highlights the importance of integrating motor information into learning representations (Paolo, Gonzalez-Billandon, and Ke´gl 2024).

In the intersection of robotics and machine learning communities, efforts are made towards integrating embodied information into the learning loop (Firoozi et al. 2023). In general, however, there is no consensus on the integration of actions into representation learning for robotics. When learning geometric physical spaces surrounding the robot, the state-of-the-art methods include simultaneous localisation and mapping (SLAM), where the agent builds a geometric map of the environment by integrating its sensor measurements and motor information (Placed et al. 2023). This is a good example of using embodied information for building representations. Similarly, robotic priors (Jonschkowski and Brock 2015) enable the agent to learn the structure of the environment by using prior knowledge about the basic physical properties of the environment. In both cases, reliance on the prior information naturally limits the adaptability of these methods to different domains and sensory modalities, as the prior knowledge depends heavily on the context.

In contrast, the learning algorithm used in this work is designed to be as generic as possible, incorporating minimal prior information about the environment. This ensures its potential applicability beyond the simple environment proposed. This approach shares similarities with commonly used methods for learning state spaces in reinforcement learning, such as Ha and Schmidhuber (2018) or Barreto et al. (2017), which are designed to be general and independent of specific sensory inputs or assumptions about the environment. The system presented here is meant to facilitate a further discussion on how embodiment can be effectively used for building unified representations.

Experimental Environments

To demonstrate the importance of embodiment for learning representations, we propose three environments, see fig.1. They are based on the Flatland simulation (Caselles-Dupre´ et al. 2018) that features continuous space and first-person view and was developed for quick and efficient prototyping of learning algorithms with the view of further applying them to robotic tasks, while simplifying the observation space and reducing computational overhead, by using one-dimensional observations. (inspired by Jonschkowski and Brock (2015)). The second one (fig.1b) has a *wall* that cuts the environment into two rooms, connected by a passage in the centre. The last environment (fig.1c) has the same shape as the second, but the wall is transparent and not visible to

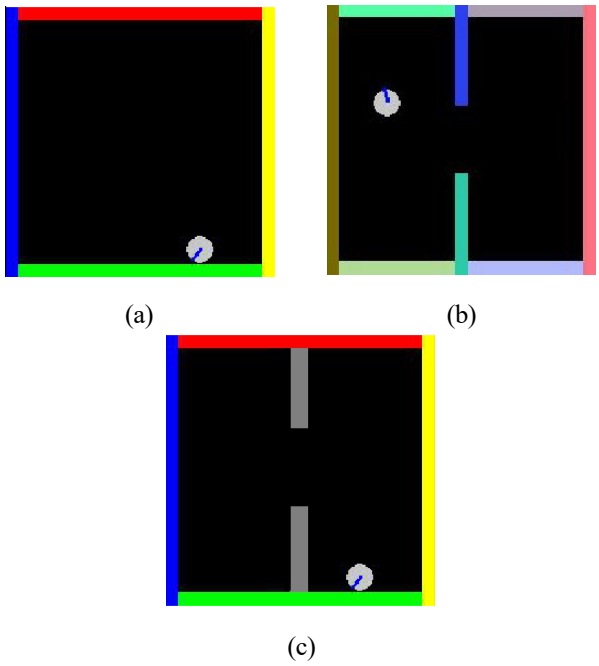


Figure 1: The environments used in the demonstration: a - *simple* room, b - room with a *wall* and a passage, c - room with an *invisible wall*. The agent can move around by either moving forward or rotating left or right at every time step. The rotation speed is within $(-\pi/2,\ \pi/2)$ and the speed is bounded by $(0,\ 10)$ pixel units.

the agent, so the observation space $O(t)$ is identical to the environment in fig.1a. We will refer to it as an *invisible wall* environment.

There are two distinct topologies: an open square space, represented by the *simple* environment, and a more complex two-room space, represented by the *wall* and *invisible wall* environments. Additionally, there are two different observation spaces: one with four walls in four primary colours, in the *simple* and *invisible wall* environments, and another with a randomly chosen set of eight colours, present in the *wall* environment. The *invisible wall* shares the observation space with the *simple* environment and topology with the *wall* environment.

We included the environment with an invisible wall to illustrate the importance of including motor information in the learning loop. The *invisible wall* environment's sensory space is identical to the *simple* environment, but its shape is the same as the wall environment. Hence, the only way the agent may discover the environment topology is by observing different sensory-motor dynamics. As the observations of both *simple* and *invisible wall* environments are identical, the difference in the resulting learned state space between the two can be attributed to the agent discovering the invisible wall through interaction. Similarly, despite different observations of the *wall* and *invisible wall* environments, similarities between the resulting observation spaces would indicate that the agent has managed to accurately extract the

structure of these environments.

On each discrete time-step, t the agent can either move forward for up to 10 pixel units or rotate within $(-\pi/2, \pi/2)$, thus producing a 2-dimensional vector a_t. The dimension of the space is 200×200 pixels. The field of view of the agent is set to $\pi/2$, meaning that it operates under the conditions of partial observability. Fig.3 shows examples of observations that the agent receives in all three environments.

Representation Learning Architecture and Training Procedure

In line with the concept of interaction between motor and sensory information, the proposed architecture employs a predictive learning objective, essentially building a version of a forward model. Forward model is a system used by the agent to predict the future state of its environment based on its current state and actions (Wolpert, Ghahramani, and Jordan 1995; Lesort et al. 2018). implementing a version of a forward model by setting up a predictive objective for self-supervised learning of representations is an established method. For example, such prominent concepts as successor representations (Dayan 1993; Kulkarni et al. 2016) and World Models (Ha and Schmidhuber 2018) use a predictive objective for learning. Moreover, in (Watter et al. 2015; Kulak and Garcia Ortiz 2018; Recanatesi et al. 2021) prediction is used to build latent spaces, similarly to our experiment.

In our case, the model receives a sequence of observations and actions (processed sequentially using an RNN) and is trained to predict the final observation in the sequence. The difference between the predicted and actual observation generates an error signal, which is used to update the model. By learning to predict future sensory inputs based on the agent's actions, the model develops representations that are closely aligned with the environment's dynamics. This predictive learning approach is particularly effective in robotics, where an agent's actions directly impact its sensory feedback. While our experiments focus on visual data, the same principle could, in principle, be extended to other sensory modalities, making the approach broadly applicable.

We use a simple version of predictive architecture, based on end-to-end training of a convolutional encoder-decoder pair and an LSTM, as outlined in fig.2. More formally, let's consider a sequence of observations $O = (o_1, o_2, ..o_{n-1}, o_n)$ and the corresponding sequence of motor commands $A = (a_{1,2}, x_{2,3}, .., a_{n-1,l})$, where n is the length of the sequence, and each step represents a discrete time interval. Then we assign the last observation o_l to be the *target* real observation, and the rest of the sequence $(o_1, o_2, .., o_{n-1})$ is used to derive the *estimation* $\hat{o}_n$.

As shown in fig.2, each observation o_i is processed using a convolutional encoder and compressed to a 3-dimensional vector, which is then concatenated with the 2-dimensional a_i. This concatenated vector is fed into LSTM cell (Hochreiter and Schmidhuber 1997). The hidden 3-dimensional state of the LSTM cell is then propagated to the next time step. The final 3-dimensional state z is obtained after processing $(n-1)$ observation-action pairs. This state is then fed into a

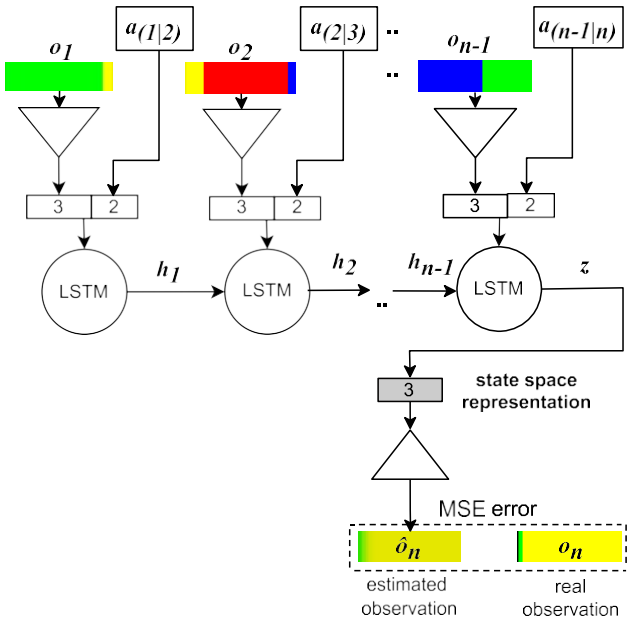
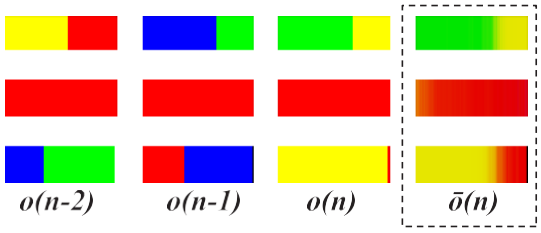


Figure 2: The architecture of the state representation learning model. The input consists of pairs of (o_t, a_t). The visual RGB input (3 × 256) is compressed into 3-dimensional vector using an encoder, implemented using Conv1D based decoder. The resulting compressed representations is concatenated with a_t and used as an input into LSTM cell. After repeating this process n = 15 times, the final output of LSTM z is decoded using an equivalent Deconv1D-decoder back into RGB vector (3 × 256) $\hat{o}_n$ and compared to the true value of the last observation o_n. The resulting MSE is back-propagated thought the network.

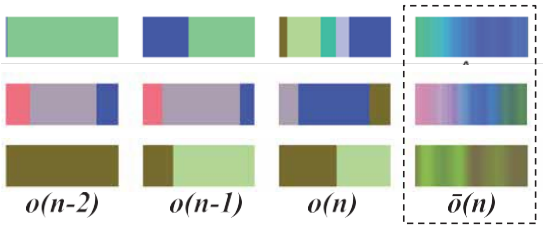
deconvolutional decoder that recovers the visual information from the 3-dimensional state z into an estimated observation $\hat{o}_n$. A 3-dimensional information bottleneck is deliberately chosen for its easy visual representation and interpretability, and as it is the minimal number of necessary dimensions to describe the agent's position in the space (x, y, θ). This allows us to inspect the representation space directly without any additional processing. Such limitation of course is not appropriate for most practical applications, and the bottleneck has to be significantly relaxed if the representation is to include any other information aside from the (x, y, θ).

During training the mean-square error (MSE) of $\hat{o}_n$ and o_n is calculated and propagated through the gradients of the weights through the system. During testing, only the state estimation is used, up to where the z state value is estimated.

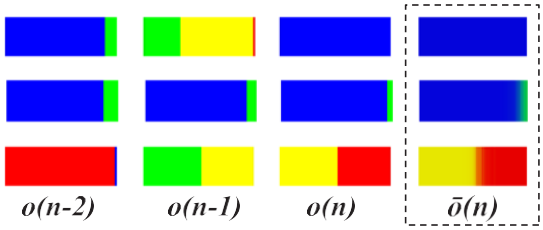
The data for the training and validation were collected separately on each of the three environments as a series of observation sequences obtained by random actions over 10,000 time steps. The agent can move around by either moving forward or rotating left or right at every time step. The rotation speed is within $(-\pi/2, \pi/2)$ and the speed is



(a)



(b)



(c)

Figure 3: The last three observations in the three example testing sequences (o_{n-2}, o_{n-1}, o_n) together with the predicted $\hat{o}_n$. In all cases, the estimated $\hat{o}_n$ resembles the real observation o_n closely, just as a more blurred version, which demonstrates that the model is capable of predicting the next observation.

bounded by (0, 10) pixel units. The data was split into two parts, with (2:1) split between training and validation data. A separate model was trained for each environment using the same procedure.

Fig.3 demonstrates the last observations in a few random validation n-step sequences, the last o_n used as a training signal, and the observation $\hat{o}_n$ estimated by the network. In all cases, the estimated $\hat{o}_n$ resembles the real observation o_n, as a more blurred version, which demonstrates that the model is capable of predicting the next observation. In the case of the *wall* environment (fig.3b), the prediction is notably worse than its counterparts from the other two environments. Most likely, this is the result of a more complex observation space compared to the other two environments, resulting in the small neural network used for both the encoder and the decoder reaching its representational capacity.

Results and Discussion

The visualization of state space Z is shown in fig.4: (a) *simple*, (b) *wall* and (c) *invisible wall*. To obtain these point clouds, random sequences of samples from the validation set were processed with the representation learning architecture to obtain the 3-dimensional representation z, which describes the state of the agent at the time. For illustration purposes, the point clouds in the fig.4 are shown in the colours that correspond to the ground truth position of the agent: (x, y, θ) (x - right, θ - left and y - bottom graph). State space point clouds show a high level of structure with respect to the ground truth position of the agent. Units of the graphs are abstract representations of space units and do not have a semantic meaning.

Notice the similarity between the *wall* environment and the *invisible wall* environment point clouds. We can see that clouds generated by the representation spaces of *invisible wall* and *wall* are similar. On the other hand, while the observation spaces are similar in both *simple* and *invisible wall* cases, the resulting topologies are different; the *simple* environment generates an almost flat rectangular point cloud, but both *wall* and *invisible wall* environments' representations result in a crescent-shaped point cloud.

Hence, by integrating sequences of sensory-motor data, the system is able to discover accurate topology of the space, even despite similarities in the observations. This highlights the important role of embodiment for learning representations, since even such a weak form of embodiment as integrating actions into the learning loop was integral to identifying the topology of the environment.

We would like to highlight that in our case, there are two different sources of embodied signal present in the data: (1) the interaction between the observations and actions and (2) the observation sequences pattern itself, as none of the trajectories cross the wall. Most likely, both of those two aspects contribute to the outcome of the learning process. However, we argue, that actions take a prominent role as they enable the system to estimate predictions and, consequently, uncover the dynamics of the environment.

It is important to note that our example shows a basic form of embodiment and one could argue that the data was still "static" in nature - obtained separately from building the representation itself - thus lacking true interactivity. Our goal, however, is to show that even in this very redacted form it is an invaluable source of information, that should be considered very closely when building robotic representations. The proposed predictive architecture is one way, in which motor information could be integrated, along with all the sensory information.

Since the proposed architecture is rather generic, it could be, in principle, applied to different domains and sensory modalities, although confirming its efficacy would be a subject for future work. In general, the usage of a sensory-motor stream of data and predictive objectives is applicable to any sensory domain — be it auditory, tactile, or even multimodal.

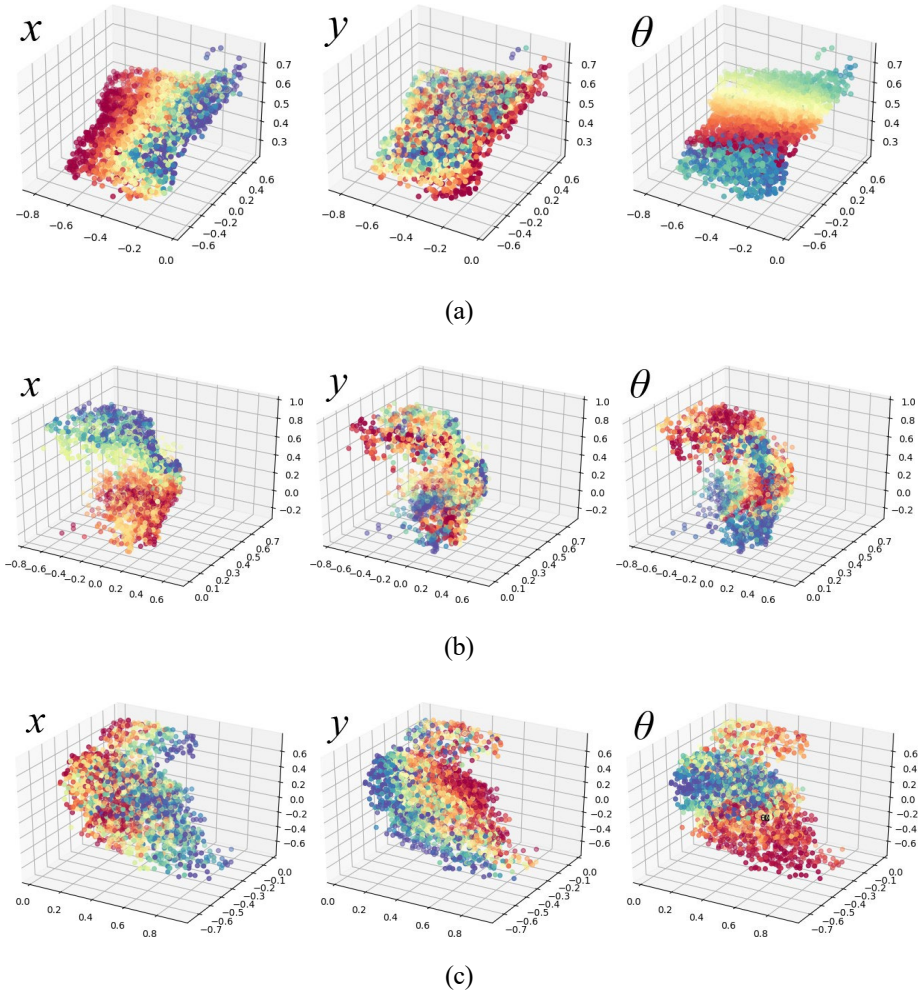


Figure 4: The visualization of the state space Z for the the three environments: (a) *simple*, (b) *wall* and (c) *invisible wall*. Colours correspond to the ground truth of the position of the agent (x, y, θ). Both of the environments with the wall feature some curvature but also have the same gradient distribution. Units of the graphs are abstract representations of space units and do not have a semantic meaning, and cannot be compared. Note the similarities between (b) and (c) spaces.

Conclusions

The results from our simulated environments reveal that the agent's ability to interact with and perceive the consequences of its actions is crucial for developing representations that align with the actual environment structure. This highlights the importance of considering embodiment in the design of representation learning algorithms for robotic learning tasks.

This work serves as a reminder that actions should be a central component of the loop when learning representations, particularly in robotics. The strong informational signal provided by actions is too valuable to be overlooked, as it allows for the development of more accurate and useful representations. Our results demonstrate the efficacy of this approach using a predictive objective.

From the perspective of unification of representations for the use of robotics, the use of action-conditional prediction in an abstract space allows for integrating signals from different sensory modalities. We hope this work encourages further exploration into different ways of including the unique and valuable property of embodiment we have in robotics in the development of intelligent systems.

Acknowledgments

This work originally began during the first author's tenure at SoftBank Robotics Europe AI Lab (Paris, France), supported by Horizon 2020 APRIL ITN funding. We thank Dr. M. G. Ortiz, Dr. P. Loviken and other colleagues at the AI Lab for the discussions on the topic of this paper and the provided support.

References

Barreto, A.; Dabney, W.; Munos, R.; Hunt, J. J.; Schaul, T.; van Hasselt, H. P.; and Silver, D. 2017. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.

Caselles-Dupre´, H.; Annabi, L.; Hagen, O.; Garcia-Ortiz, M.; and Filliat, D. 2018. Flatland: a lightweight first-person 2-d environment for reinforcement learning. *arXiv preprint arXiv:1809.00510*.

Dayan, P. 1993. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, 5(4): 613–624.

Degenaar, J.; and O'Regan, J. K. 2015. Sensorimotor theory of consciousness. *Scholarpedia*, 4952.

Firoozi, R.; Tucker, J.; Tian, S.; Majumdar, A.; Sun, J.; Liu, W.; Zhu, Y.; Song, S.; Kapoor, A.; Hausman, K.; Ichter, B.; Driess, D.; Wu, J.; Lu, C.; and Schwager, M. 2023. Foundation Models in Robotics: Applications, Challenges, and the Future. *ArXiv*, abs/2312.07843.

Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127.

Ha, D.; and Schmidhuber, J. 2018. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems 31*, 2451–2463. Curran Associates, Inc. https://worldmodels.github.io.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Huang, Y.; and Rao, R. P. 2011. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5): 580–593.

Jonschkowski, R.; and Brock, O. 2015. Learning state representations with robotic priors. *Autonomous Robots*, 39(3): 407–428.

Kulak, T.; and Garcia Ortiz, M. 2018. Emergence of Sensory Representations Using Prediction in Partially Observable Environments. In *27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part II*, 489–498. ISBN 978-3-030-01420-9.

Kulkarni, T. D.; Saeedi, A.; Gautam, S.; and Gershman, S. J. 2016. Deep Successor Reinforcement Learning. *ArXiv*, abs/1606.02396.

Lesort, T.; Rodr´ıguez, N. D.; Goudou, J.; and Filliat, D. 2018. State Representation Learning for Control: An Overview. *CoRR*, abs/1802.04181.

Paolo, G.; Gonzalez-Billandon, J.; and Ke´gl, B. 2024. Position: A Call for Embodied AI. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 39493–39508. PMLR.

Placed, J. A.; Strader, J.; Carrillo, H.; Atanasov, N.; Indelman, V.; Carlone, L.; and Castellanos, J. A. 2023. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Transactions on Robotics*, 39(3): 1686–1705.

Recanatesi, S.; Farrell, M.; Lajoie, G.; et al. 2021. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, 12: 1417.

Scheier, C.; and Pfeifer, R. 1999. The embodied cognitive science approach. In *Dynamics, synergetics, autonomous agents: Nonlinear systems approaches to cognitive psychology and cognitive science*, 159–179. World Scientific.

Schro¨dinger, E. 1944. *What is life? The physical aspect of the living cell and mind*. Cambridge university press Cambridge.

Watter, M.; Springenberg, J.; Boedecker, J.; and Riedmiller, M. 2015. Embed to control: A locally linear latent dynamics model for control from raw images. *Advances in neural information processing systems*, 28.

Wilson, M. 2002. Six views of embodied cognition. *Psychonomic bulletin & review*, 9: 625–636.

Wolpert, D. M.; Ghahramani, Z.; and Jordan, M. I. 1995. An Internal Model for Sensorimotor Integration. *Science*, 269(5232): 1880–1882.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.