# SEMI: Self-supervised Exploration via Multisensory Incongruity

## Abstract

*Efficient exploration is a long-standing problem in reinforcement learning since extrinsic rewards are usually sparse or missing. A popular solution to this issue is to feed an agent with novelty signals as intrinsic rewards. In this work, we introduce SEMI, a self-supervised exploration policy by incentivizing the agent to maximize a new novelty signal: multisensory incongruity, which can be measured in two aspects, perception incongruity and action incongruity. The former represents the misalignment of the multisensory inputs, while the latter represents the variance of an agent's policies under different sensory inputs. Using both incongruities as intrinsic rewards, SEMI allows an agent to learn skills by exploring in a self-supervised manner without any external rewards. The effectiveness of SEMI is demonstrated across a variety of benchmark environments including object manipulation and audio-visual games.*

## 1. Introduction

Efficient exploration is a major bottleneck in reinforcement learning problems. In many real-world scenarios, rewards extrinsic to an agent are extremely sparse or completely missing, leading to nearly random exploration of states. A common remedy to exploration is adding intrinsic rewards, *i.e.*, rewards automatically computed based on the agent's model of the environment. Existing formulations of intrinsic rewards include maximizing "visitation count" [4, 23, 36] of less-frequently visited states, "curiosity" [30, 32, 37] where future prediction error is used as reward signal and "diversity rewards" [12, 22] which incentivizes diversity in the visited states. These rewards provide continuous feedback to the agent when extrinsic rewards are sparse, or even absent. However, it is challenging to deploy these methods in practice. For "visitation count" based method, it is hard to count in a continuous space. And for "predictive model" based method, the key challenge is to model and interact with a stochastic world where multiple futures are available.

Recently a popular line of works in using intrinsic reward to train an RL agent are using prediction error [32, 37, 6, 7], prediction uncertainty [15, 29], or improvement [23] of a forward dynamics or value model as intrinsic rewards. A concurrent work from Dean *et al*. [10] has also demonstrated the effectiveness of using multisensory signals as intrinsic rewards. In using multimodal signal for self-supervision, several works leverage the natural correspondence [2, 34] and synchronization [31, 21] between the audio or tactile and RGB streams to learn representations. A recent work from Dean *et al*. [10] has also used the temporal incongruity from multisensory perception as intrinsic rewards.

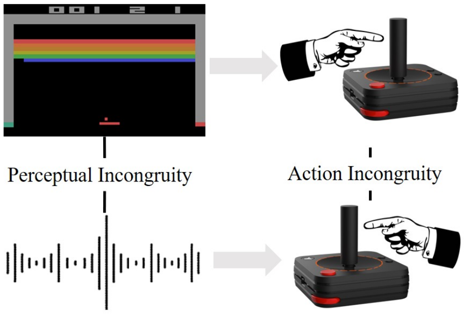

**Figure 1:** SEMI: a self-supervised exploration policy by incentivizing the agent to maximize multisensory incongruity, including *perceptual incongruity* and *action incongruity*. *Perceptual incongruity* indicates the misalignment between the multisensory perceptual inputs, and *action incongruity* refers to the discrepancy of actions under different perceptual inputs.

In this work, we introduce **SEMI**, a self-supervised exploration method by incentivizing the agent to maximize multisensory incongruity, including *perceptual incongruity* and *action incongruity*, as shown in Figure 1.

*Perceptual incongruity* is defined as the misalignment between multisensory inputs. As humans, the coincidence of senses gives us strong evidence that they were generated by a common, underlying event [39], since it is unlikely that they co-occurred across multiple modalities merely by chance. Thus, the misalignment or incongruity between multisensory streams can be used as a strong signal of novelty. Researches in psychology suggested that this incongruity can attract human's attention and trigger further exploration [5, 11], which has been widely used in product design [25, 24]. In SEMI, we use such novelty to guide robot exploration. Specifically, an alignment predictor is trained to detect misalignment between multisensory inputs. The model observes raw sensory streams — some of which are paired, and some have been shuffled — and we task it with distinguishing between the two. This challenging task forces the model to fuse information from multiple modalities and meanwhile learn a useful feature representation. The prediction error of the sensor fusion model serves as a metric of perceptual incongruity, which is further used as an intrinsic reward to guide the agent's exploration.

*Action incongruity* is defined as the discrepancy of an agent's decisions when it perceives different senses of the same underlying event. This is inspired from the fact that humans are able to integrate multimodal sensory information in a near-optimal manner for decision making [1, 26],

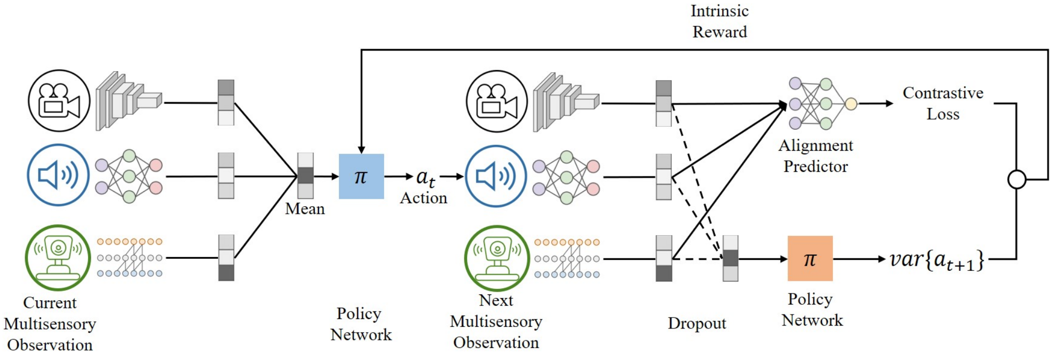

**Figure 2:** SEMI pipeline overview: at time step $t$, an agent takes action $a_t$ given a multisensory observation $O_t$ as input and ends up in a new state. The multisensory fusion model takes a new observation $O_{t+1}$ as input and predicts whether these sensory inputs are aligned. The prediction loss is used as the measure of perceptual incongruity. The variance of actions suggested by the policy network given different combination of multisensory inputs is used to measure action incongruity. Both incongruities are used as intrinsic rewards to train the policy $\pi$.

and are even robust to the loss of some senses [14, 18]. Sensory compensation empowers humans to make similar decisions when different senses are used [9, 3, 20]. In SEMI, a policy network is learned with multi-modal dropout during multisensory fusion. Concretely, we randomly drop one or several modalities during multisensory fusion to imitate loss of senses. The variance of actions suggested by the policy network under different dropout states is used to measure action incongruity, which is also used as an intrinsic reward for better exploration.

SEMI is evaluated in two challenging scenarios: object manipulation (vision and depth) and audio-visual games (Gym Retro). We show that SEMI outperforms “predictive model” based exploration policy by a large margin in both scenarios.

## 2. Method

SEMI is a self-supervised exploration policy that incentivizes agents to maximize multisensory incongruities, which we formulate as two aspects: perceptual incongruity (Section 2.1) and action incongruity (Section 2.2). Both incongruities are fed to the agent as intrinsic rewards to encourage its exploration. Figure 2 gives an overview of the pipeline of SEMI, and we will detail each sub-module in the following.

**Notation.** Given an agent's current observation $O_t$ at time $t$, our goal is to generate intrinsic curiosity reward $r_t$ so that the agent learns a policy $\pi$ to explore unknown and difficult environment. In this paper, we focus on the multisensory setting, where the agent observes a set of perceptual inputs $O_t = \{o_t^1, o_t^2, ..., o_t^M\}$, where $M$ is the number of modalities, which could represent vision, audio, touch, *etc*. By executing an action $a_t$ produced by the policy, the agent further observes the next state, which we denote as $O_{t+1} = \{o_{t+1}^1, o_{t+1}^2, ..., o_{t+1}^M\}$.

### 2.1. Multisensory Perceptual Incongruity

The synchrony of multiple senses is a fundamental property of natural event perception. We humans are extremely sensitive to the incongruity between these senses, which is a strong signal of novelty. For example, if a common object makes an uncommon sound, we are motivated to further interact with this object to gain better knowledge about it. Inspired by this observation, we aim to use such novel association signals as curiosity to drive an RL agent to explore unfamiliar states.

To guide an agent to explore novel states, we propose an alignment predictor to discover the perceptual incongruity. Alignment prediction can take various forms, one possible design is to predict one sensory stream from other streams. For example, we could generate sounds from a corresponding visual input, or generate images from its sounds. However, generating data in the raw signal space is proved to be challenging, since (1) it does not handle the cases of multiple possible targets, (2) it suffers from overfitting to trivial details or noises [32].

A better idea is to predict the compatibility of multisensory streams in the latent space. Along the idea of contrastive learning [28, 8], our design of alignment predictor directly maximizes the agreement between different modalities of the same event. This is achieved by predicting positive (aligned) modality streams from negative ones via a contrastive loss penalty in the latent space. The predicted alignment score can then be used as an indicator of perceptual incongruity.

Concretely, the alignment predictor comprises the following two major components.

- A set of neural network base encoders $(f_1(\cdot), ..., f_M(\cdot))$ that extracts representation vectors from each modality. Our framework is agnostic to the choices of neural network architectures. In the following experiments, we use a 2D ConvNet to extract RGB visual features, another 2D ConvNet to obtain depth features, and a Short Time Fourier Transform (STFT) followed by a 1D ConvNet to extract the audio features.
- A contrastive loss function defined for a contrastive learning. Given one sensory stream $o^j$ from a multisensory observation $O = \{o^i\}|_{i=1,...,M}$ (we omit time $t$ in the following for brevity), we define the other $M-1$ simultaneous sensation streams $\{o^i\}|_{i\neq j}$ as positive examples. In a mini-batch of $N$ observations, there are $M \times (N-1)$ sensory streams from other modalities, which can be used to build misaligned examples. The contrastive prediction task aims to identify aligned sensory streams from these misaligned examples.

The similarity of a pair of multimodal observation $(o^i, o^j)$ are measured by the cosine distance, *i.e*.

$$\mathrm{sim}(o^i, o^j) = \cos(\mathbf{f_i}, \mathbf{f_j}) = \frac{\mathbf{f_i^T} \cdot \mathbf{f_j}}{||\mathbf{f_i}|| \cdot ||\mathbf{f_j}||}, \tag{1}$$

where $\mathbf{f_i} = f_i(o^i), \mathbf{f_j} = f_j(o^j)$ are features from different modalities. Then the contrastive loss function for a pair of positive observation $(o^i_k, o^j_k)$ is defined as

$$\mathcal{L}(o^i_k, o^j_k) = -\log \frac{\exp(\mathrm{sim}(o^i_k, o^j_k)/\tau)}{\sum_{n=1}^{N} \sum_{m=1}^{M} \exp(\mathrm{sim}(o^i_k, o^m_n)/\tau)}, \tag{2}$$

where $\tau$ denotes a temperature parameter.

The *multisensory perceptual incongruity* of an observation $O_k$ is then defined numerically as the sum of losses of all possible multisensory pairs from the same timestep, which can be used as an intrinsic reward $r^p = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathcal{L}(o^i_k, o^j_k)$.

### 2.2. Multisensory Action Incongruity

Congruity in actions is inspired from the fact that human perception is robust to the partly loss of senses, and humans have an exceptional ability to compensate for the loss with other senses. If we make different decisions with different sensory inputs, it suggests we have low confidence of the event we experienced, *e.g*. an inexperienced driver might change lane recklessly without a good understanding of the distance of cars from the sound noise. Inspired by the above observation, we further aim to use the action incongruity as an indicator of novelty in RL exploration.

Here we implement the action incongruity via drop of senses. Proposed by Srivastava *et al*. [40], *dropout* has been widely used to prevent neural networks from overfitting [19, 16]. Gal *et al*. [13, 17] further cast dropout training in deep neural networks as approximate Bayesian inference in deep Gaussian processes, which offers a mathematically grounded framework to reason about model uncertainty.

We adopt a similar approach by taking a sensory-wise dropout strategy during sensor fusion for the policy network. Then multisensory action incongruity is defined as the divergence of actions suggested by the policy network given different combinations of multisensory observations.

Specifically, we combine features of different modalities with dropout to obtain a fused perceptual feature $z$,

$$z = \frac{1}{\sum_{i=1}^{M} \mathbb{1}^i} (\sum_{i=1}^{M} \mathbb{1}^i \mathbf{f_i}) \tag{3}$$

where $\mathbb{1}^i \in \{0, 1\}$ indicates the existence of $\mathbf{f_i}$. Apparently, different combinations of $\mathbb{1}^i$ will lead to different $z$. We collect the action outputs from the policy network $\pi_r$ given all possible inputs $z$'s ($2^M - 1$ possible inputs in total), and define the variance of these actions as the *multisensory action incongruity*. The action incongruity is further used as an intrinsic reward $r^a$ for exploration,

$$r^a = \frac{1}{2^M - 1} \sum_{k=1}^{2^M-1} ||\pi_r(z^k) - \frac{1}{2^M - 1} \sum_{k=1}^{2^M-1} \pi_r(z^k)||_2^2. \tag{4}$$

### 2.3. Multisensory Incongruities as Intrinsic Rewards

To summarize, we use both multisensory perceptual incongruity and multisensory action incongruity as intrinsic rewards. It is worth noting that the policy network $\pi_r$ used to calculate intrinsic reward $r^a_t$ is different from that used for exploration $\pi$. Inspired by Double Q-learning [42] and Dual Policy Iteration [41], $\pi_r$, with parameters $\theta$ being the same as $\pi$ except that its parameters are copied every $\tau$ steps from the $\pi$. This simple strategy not only reduces the observed overestimations, but also leads to better convergence.

At time step $t$, the agent takes action $a_t$ given multisensory observation $O_t$ with modality dropout as input and receives a new observation $O_{t+1}$ and intrinsic reward in calculated as $r_t = r^p_t + \gamma \times r^a_t$, where $\gamma$ is a weight factor. The agent is optimized using PPO [38] to maximize the expected reward according to

$$\max_{\theta} \mathbf{E}_{\pi(O_t;\theta)}(\sum_t r_t). \tag{5}$$

## 3. Experiments

We evaluate the performance of SEMI in two environments, *OpenAI Robotics* and *Atari*.

| Exploration Strategy | | Interaction Rate (1 objects) | Convergence Iteration | Interaction Rate (1 of 3 objects) |
|---|---|---|---|---|
| Uni-IR | Curiosity | 2.7% | 25 | 8.3% |
| | Random | 8.4% | 0 | 22.6% |
| | Disagreement | 26.3% | 23 | 64.3% |
| | SEMI (P) | 30.5% | 20 | 81.4% |
| | SEMI (PA) | 34.4% | 33 | 82.1% |

**Table 1:** We measure the exploration quality by evaluating the object interaction frequency of the agent trained with different intrinsic rewards (Row 1-5) and a combination of intrinsic rewards (Row 6-7).

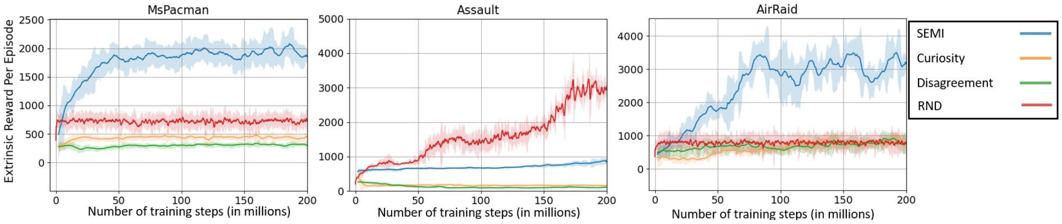

**Figure 3:** We compare different intrinsic reward formulations across different Atari games. We run three independent runs of each algorithm and show the mean extrinsic reward during training. SEMI far outperforms curiosity-based baseline and disagreement-based baseline, and also learns more efficiently.

## 3.1. Exploration via Multisensory Incongruity

### 3.1.1 Environment and Setting

**OpenAI Robotics** We evaluate our method on OpenAI Robotics [35], where robot receives RGB image and Depth image as two modalities, and controls the gripper Cartesian movement, gripper rotation as well as gripper open or close.

**Atari** We also evaluate our method on Atari games, where vision and audio are considered as multi-modal inputs. We use Gym Retro [27] in order to access game audio. Further details for the two evaluation environments are described in the supplementary materials.

### 3.1.2 Training Details

In general, we used 5 convolutional layers to extract RGB features, a similar network to extract depth features or 5 consecutive frames channel-wise spectrum to represent audio feature. We used a 4-layer multi-layer perceptron (MLP) as our policy network and used PPO to maximized the intrinsic reward with an Adam Optimizer. During training, all rewards that are collected in trajectories will be replaced or added by intrinsic reward.

### 3.1.3 Results

**OpenAI Robotics** Table 1 shows the exploration performance of object manipulation using the multisensory incongruity, which are measured by the frequency at which our agent interacts (*i.e.*, touches) with the object (*i.e.* interaction rate). The interaction rate is defined as *#trials robot interact with object/#total trials*.

We evaluate two different versions of our method. We first use only the multisensory perceptual incongruity as our intrinsic reward, as described in Section 2.1. Second, we use both multisensory perceptual incongruity and multisensory action incongruity as our intrinsic reward.

We compare SEMI to Curiosity [32, 6] and Disagreement [33] as our baselines. Also, we compared with a random policy as a sanity check, which samples its action uniformly from the action space.

As shown in Table 1, our method outperforms all of these baselines. The method of Disagreement [33] has a performance close to that of our method.

**Atari** We also test out method in Atari MsPacman, Assault, AirRaid, Alien, Space Invaders, Breakout, and Beam Rider. Figure 3 shows the extrinsic reward of some Atari games during exploration with SEMI in comparison of intrinsic reward via RND, Curiosity and Disagreement. It should be pointed that during training the agent only has access to the intrinsic reward. As illustrated in Figure 3, our method converges faster and achieves better performances comparing with most of the baseline methods. The reason is that audio signals are always triggered by significant events (*e.g.* eating pellets) in these games. Thus, the multisensory incongruity is more indicative compared with curiosity and disagreement baselines, which are influenced by the stochasticity of the environments.

## References

[1] D. E. Angelaki, Y. Gu, and G. C. DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–458, 2009. 1

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 1

[3] D. Bavelier and H. J. Neville. Cross-modal plasticity: where and how? *Nature Reviews Neuroscience*, 3(6):443–452, 2002. 2

[4] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016. 1

[5] D. E. Berlyne, M. A. Craw, P. H. Salapatek, and J. L. Lewis. Novelty, complexity, incongruity, extrinsic motivation, and the gsr. *Journal of Experimental Psychology*, 66(6):560, 1963. 1

[6] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018. 1, 4

[7] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018. 1

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[9] L. G. Cohen, P. Celnik, A. Pascual-Leone, B. Corwell, L. Faiz, J. Dambrosia, M. Honda, N. Sadato, C. Gerloff, M. D. Catala, et al. Functional relevance of cross-modal plasticity in blind humans. *Nature*, 389(6647):180–183, 1997. 2

[10] V. Dean, S. Tulsiani, and A. Gupta. See, hear, explore: Curiosity via audio-visual association. *arXiv preprint arXiv:2007.03669*, 2020. 1

[11] W. N. Dember and R. W. Earl. Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological review*, 64(2):91, 1957. 1

[12] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018. 1

[13] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 3

[14] A. E. Hoover, L. R. Harris, and J. K. Steeves. Sensory compensation in sound localization in people with one eye. *Experimental brain research*, 216(4):565–574, 2012. 2

[15] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016. 1

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[17] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017. 3

[18] A. J. Kolarik, S. Cirstea, S. Pardhan, and B. C. Moore. A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing research*, 310:60–68, 2014. 2

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 3

[20] D. S. Lee, J. S. Lee, S. H. Oh, S.-K. Kim, J.-W. Kim, J.-K. Chung, M. C. Lee, and C. S. Kim. Cross-modal plasticity and cochlear implants. *Nature*, 409(6817):149–150, 2001. 2

[21] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019. 1

[22] J. Lehman and K. O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011. 1

[23] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems*, pages 206–214, 2012. 1

[24] G. D. Ludden, H. N. Schifferstein, and P. Hekkert. Surprise as a design strategy. *Design Issues*, 24(2):28–38, 2008. 1

[25] G. D. Ludden, H. N. Schifferstein, and P. Hekkert. Beyond surprise: A longitudinal study on the experience of visual-tactual incongruities in products. *International journal of design*, 6(1), 2012. 1

[26] W. J. Ma and A. Pouget. Linking neurons to behavior in multisensory perception: A computational review. *Brain research*, 1242:4–12, 2008. 1

[27] A. Nichol, V. Pfau, C. Hesse, O. Klimov, and J. Schulman. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018. 4

[28] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[29] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034, 2016. 1

[30] P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009. 1

[31] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1

[32] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017. 1, 2, 4

[33] D. Pathak, D. Gandhi, and A. Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pages 5062–5071, 2019. 4

[34] M. Patrick, Y. M. Asano, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 1

[35] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018. 4

[36] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704, 2006. 1

[37] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, pages 222–227, Cambridge, MA, USA, 1990. MIT Press. 1

[38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[39] R. Sekuler, A. Sekuler, and R. Lau. Sound alters visual motion perception. *Nature*, page 308, 1997. 1

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3

[41] W. Sun, G. J. Gordon, B. Boots, and J. Bagnell. Dual policy iteration. In *Advances in Neural Information Processing Systems*, pages 7059–7069, 2018. 3

[42] H. Van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 3