
See, Hear, Explore: Curiosity via Audio-Visual Association

Anonymous Author(s)

Affiliation

Address

email

Abstract

Exploration is one of the core challenges in reinforcement learning. A common formulation of curiosity-driven exploration uses the difference between the real future and the future predicted by a learned model [1]. However, predicting the future is an inherently difficult task which can be ill-posed in the face of stochasticity. In this paper, we introduce an alternative form of curiosity that rewards novel associations between different senses. Our approach exploits multiple modalities to provide a stronger signal for more efficient exploration. Our method is inspired by the fact that, for humans, both sight and sound play a critical role in exploration. We present results on several Atari environments and Habitat (a photorealistic navigation simulator), showing the benefits of using an audio-visual association model for intrinsically guiding learning agents in the absence of external rewards.

1 Introduction

Many successes in reinforcement learning (RL) have come from agents maximizing a provided extrinsic reward such as a game score. However, in real-world settings, reward functions are hard to formulate and require significant human engineering. On the other hand, humans explore the world driven by intrinsic motivation, such as curiosity, often in the absence of rewards. But what is curiosity and how would one formulate it?

Recent work in RL has focused on curiosity using future prediction [1–3]. In this formulation, an exploration policy receives rewards for actions that lead to differences between the real future and the future predicted by a forward dynamics model. In turn, the dynamics model improves as it learns from novel states. While the core idea behind this curiosity formulation is simple, putting it into practice is quite challenging. Learning and modeling forward dynamics is still an open research problem; it is unclear how to handle multiple possible futures, whether to explicitly incorporate physics, or even what the right prediction space is (pixel space or some latent space).

The use of multiple modalities in *human* learning has a long history. Research in psychology has suggested that humans look for incongruity [4]. A baby might hit an object to hear what it sounds

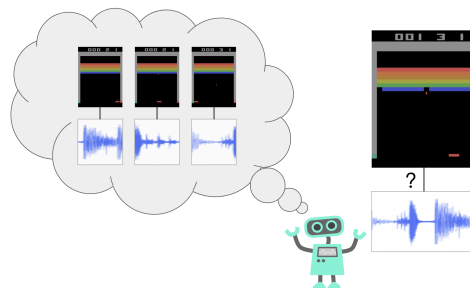


Figure 1: **See, Hear, Explore:** We propose a formulation of curiosity that encourages the agent to explore novel associations between modalities, such as audio and vision. On Atari games, like the one shown above, our method allows for more efficient exploration than baselines.

37 like. Have you ever found yourself curious to touch a material that looks different from anything
38 you have seen before? In these cases, humans are drawn towards discovering and exploring novel
39 associations between different modalities. Dember and Earl [5] argued that intrinsically motivated
40 behaviors tend to create a discrepancy between expected sensory perception and the actual stimulus.
41 In cognitive development, both sight and sound guide exploration: babies are drawn towards colorful
42 toys that create squeaking and rattling sounds [6].

43 Inspired by human exploration, we introduce See Hear Explore (SHE): a curiosity for novel associa-
44 tions between sensory modalities (Figure 1). SHE rewards actions that generate novel associations
45 between different sensory modalities (in our case, pixels and sounds). We first demonstrate that
46 our formulation is useful in several Atari games: SHE allows for more exploration, is more sample-
47 efficient, and is more robust to noise compared to existing curiosity baselines on these environments.
48 Finally, we show experiments on area exploration in the realistic Habitat simulator [7]. Our results
49 demonstrate that in this setting our approach outperforms baselines by a significant margin.

50 To summarize, our contributions in this paper include: 1) SHE, a curiosity formulation that searches
51 for novel associations in the world. To the best of our knowledge, multimodal associations have not
52 been investigated in self-supervised exploration; 2) we show our approach outperforms the commonly-
53 used curiosity approaches on standard Atari benchmark tasks; 3) most importantly, multimodality
54 is one of the most basic facets of our rich physical world (audio and vision are generated by the
55 same physical processes [8]). We show experiments on realistic area exploration in which SHE
56 significantly outperforms baselines. This work builds on efficient exploration, which will be crucial
57 as we want agents to explore more complex unknown environments.

58 2 Related Work

59 Our work uses audio as an additional modality for self-supervised exploration. We divide the prior
60 work into two categories: exploration (Section 2.1) and multimodal learning (Section 2.2).

61 2.1 Exploration

62 Prior work on exploration has used error [9, 10, 1, 11], uncertainty [12, 13, 3], and potential improve-
63 ment [14] of a prediction model as intrinsic motivation. Some approaches have used count-based or
64 pseudo-count-based exploration [15, 16]. Others use auxiliary losses to supplement reward functions
65 and improve sample efficiency [17, 18].

66 One popular approach to self-supervised exploration is curiosity by self-supervised prediction [1, 2].
67 In this form of curiosity, an intrinsic reward encourages an agent to explore situations with high
68 error under a jointly-trained predictive model. The predictive model’s error is a proxy for novelty:
69 unpredictable situations are more likely novel and therefore ones the agent should explore. Predictive
70 models, especially in visual space, can be difficult to train. Our method also looks at self-supervised
71 exploration, but our intrinsic reward does not rely on visual prediction. We circumvent the need for
72 predictive models by leveraging multimodal input. Our key insight is that *associative* models across
73 modalities are simpler to learn, and their accuracy is also indicative of novelty.

74 2.2 Multimodal Learning

75 Multimodal settings are especially amenable to self-supervision, as information from one modality
76 can be used to supervise learning for another modality. One prior work learned a joint visual and
77 language representation using Flickr images and associated descriptors [19]. In computer vision,
78 audio can provide additional information that complements images [20–22]. Recent work [23] has
79 looked at audio-visual embodied navigation, in which audio is emitted from a goal point to aid in
80 supervised learning of navigation. In the same environment, Gao et al. [24] used audio and visual
81 information for learning visual feature representations. We test on this audio-visual navigation
82 environment, but for unsupervised exploration in RL; we have no goal states.

83 Audio and visual information are closely linked, and since we commonly have access to both in the
84 form of video, this is a rich area for self-supervision. Aytar et al. [25] used audio from Atari in the
85 form of YouTube videos of people playing the games. This work uses audio-visual demonstrations
86 from YouTube to learn a visual embedding. The setup here is learning from demonstrations from

humans. In our case, on the other hand, the associations help drive intrinsically motivated exploration. We learn multimodal alignment from active data: the agent collects and uses its own data.

In robotics settings, the use of additional modalities such as tactile sensing [26, 27] or audio [28] is increasingly popular for grasping and manipulation tasks. Lee et al. [29] showed the effectiveness of self-supervised training of tactile and visual representations by demonstrating its use on a peg insertion task. While these previous approaches have demonstrated the benefits of using multiple sensory modalities for learning better representations or accurately solving tasks, in this work we demonstrate its utility for allowing agents to explore. To the best of our knowledge, using audio to learn actions for exploration is unique to our work.

3 See, Hear, Explore

We now describe SHE, our exploration method based on associating audio and visual information. Our goal is to develop a form of curiosity that exploits the multimodal nature of the input data. Our core idea is that the agent learns a model that captures associations between two modalities. We use this model to reward actions that lead to unseen associations between the modalities. By rewarding such actions, we guide the exploration policy towards discovering new combinations of sight and sound.

More formally, we consider an agent interacting with an environment that contains visual and sound features, which we call $x_t = (v_t, s_t)$ for time t where v_t is the visual feature vector and s_t is the sound feature vector. The agent explores using a policy $a_t \sim \pi(v_t; \theta)$ where a_t corresponds to an action taken by the agent at time t . To make for easier comparison to visual-only baselines, our agent is only given access to the visual features v_t and not the audio features s_t . To enable this agent to explore, we train a discriminator D that tries to determine whether an observed multimodal pair (v_t, s_t) is novel, and we reward the agent in states where the discriminator is surprised by the observed multimodal association.

3.1 Why Novel Associations?

The goal of an exploration policy is to perform actions that uncover states that lead to a better understanding of the world. One commonly used exploration strategy involves rewarding actions that lead to unseen or novel states [15]. While this strategy seems intuitive, it does not handle the fact that while some states might not have been seen, we still understand them and hence they do not need to be explored. In light of this, recent approaches have used a prediction-based formulation. If a model cannot predict the future, it needs more data points to learn. However, sometimes we may have seen enough examples, and prediction is still challenging, leading a prediction-based exploration policy to get stuck. For example, the random TV in the Unity environment (as described in Burda et al. [2]) yields high error for prediction models, so prediction-based curious agents receive high rewards for staring at the TV, though this is not a desirable type of exploration.

Our underlying hypothesis is that discovering new sight and sound associations will help mitigate the shortcomings of the previously described count-based and prediction-based exploration strategies. By using an association model, we ask a simpler question: can this image co-occur with this sound? Such a model can allow generalization to unseen states, and it also does not need to predict the future to provide an informative signal for exploration.

3.2 Association Novelty via Alignment

The core of our method is the ability to determine whether a given pair (v_t, s_t) represents a novel association. To tackle this problem, we learn a model in an online manner. Given past trajectories, a model learns whether a certain audio-visual input comes from a seen or new phenomenon. One way to model this would be to use a generative model such as a VAE [30] or GAN [31], which could determine if the image-audio combination is within the distribution or out of distribution. However, generative models are also difficult to train, so we instead propose using a discriminator to predict if the image-audio pair is novel, which has a much smaller, binary output space.

We train this discriminator to distinguish real audio-visual pairs from ‘fake’ pairs from another distribution, with the insight that the learned model is more likely to classify novel pairs as fake. Here,

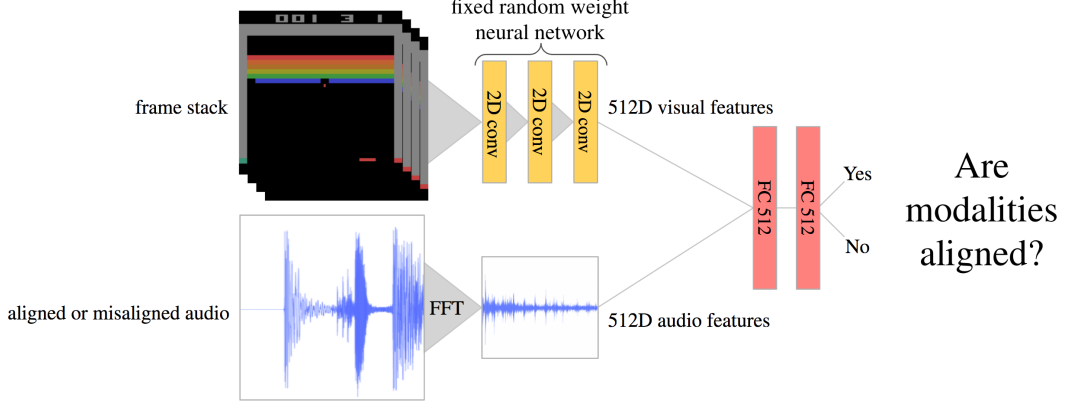


Figure 2: **Our audio-visual association model:** The frames (top left) and potentially misaligned audio waveform (bottom left) are preprocessed into 512-dimensional feature vectors using a random feature network and FFT, respectively. The discriminator network (right) takes these features as inputs and is trained to output whether or not they are aligned.

the observed image-audio pairs during exploration act as positive training examples, but a critical question is how to obtain negative image-audio pairs. To this end, we reformulate the problem as whether image-audio pairs are aligned or not: we obtain ‘fake’ samples by randomly misaligning the audio and visual modalities, similar to Owens and Efros [21]. The positive data is then the aligned image-audio pairs, and the negative data is comprised of misaligned ones. The discriminator model, as shown in Figure 2, outputs values between 0 and 1, with 1 representing high probability of audio-visual alignment and 0 representing misalignment. We can then leverage the misalignment likelihood as an indicator of novelty since the discriminator would be uncertain in such instances.

3.3 Training

Having introduced association novelty via alignment, we now describe how we implement this idea using function approximators. During training, the agent policy is rolled out in parallel environments. These yield trajectories which are each chunked into 128 time steps. A trajectory consists of pairs of preprocessed visual and sound features: $(v_1, s_1), (v_2, s_2) \dots (v_{128}, s_{128})$. These trajectories are used for two purposes: 1) updating the discriminator D as described below and 2) updating the exploration policy based on the intrinsic reward r_t^i (computed using the discriminator), also described below.

Training the Alignment Discriminator The discriminator D is a neural network that takes as input a visual and sound feature pair and outputs an alignment probability. To train D, we start with positive examples from the visual and sound feature pairs (v_t, s_t) . With 0.5 probability we use the true aligned pair, and with 0.5 probability we create a false pair consisting of the true visual feature vector v_t and a sound feature vector uniformly sampled from the current trajectory. We call this false sound s'_t . We also define a binary variable z_t to indicate whether the true audio was used, i.e. when we give the discriminator the true audio s_t , we set $z_t = 1$, and when we give the discriminator the false audio s'_t , $z_t = 0$. We use a cross-entropy loss to train the discriminator:

$$\mathcal{L}_t(v_t, s_t, z_t) = \begin{cases} -\log(D(v_t, s_t)), & \text{if } z_t = 1 \\ -\frac{\|s_t - s'_t\|_2}{\mathbb{E}_{\text{batch}}\|s_t - s'_t\|_2} \log(1 - D(v_t, s'_t)), & \text{if } z_t = 0 \end{cases}$$

In the $z_t = 0$ case above, we weight the cross-entropy loss to prevent the discriminator from being penalized in cases where the true and false audio are similar. We weight by the L2 difference between the true and false audio feature vectors and normalize by dividing by the mean difference across samples in the batch of 128 trajectories. This loss is used for updating the discriminator and is not used in computing the agent’s intrinsic reward.

Training the Agent via Intrinsic Reward We want to reward actions that lead to unseen image-audio pairs. For a given image-audio pair, if the discriminator predicts 0 (unseen or unaligned), we



Figure 3: **Habitat visualization:** Left: an example agent view. Right: the top-down map for apartment 0 (not seen by agent). The agent is the blue arrow and the audio source is the green square. Gray areas are open space, while white areas are obstacles, which make exploration challenging.

want to reward the agent. On the other hand, if the discriminator correctly outputs 1 on a true pair, the agent receives no reward. Mathematically, the agent’s intrinsic reward is the negative log-likelihood of the discriminator evaluated on the true pairs: $r_t^i := -\log(D(v_t, s_t))$, where the output of D is between 0 and 1. Audio-visual pairs that the discriminator knows to be aligned get a reward of 0, but if the discriminator is uncertain (the association surprised the discriminator) the agent receives a positive reward. The agent takes an action and receives a new observation v_t and intrinsic reward r_t^i (note that the agent does not have access to the sound s_t). The agent is trained using PPO [32] to maximize the expected reward: $\max_{\theta} \mathbb{E}_{\pi(v_t; \theta)} [\sum_t \gamma^t r_t^i]$. *The agent does not have access to the extrinsic reward. Extrinsic reward is used only for evaluation.* This will enable the use of our method on future tasks for which we cannot easily obtain a reward function.

4 Experiments

In this section, we will test our method in two exploration settings (Atari and Habitat) and compare it against commonly-used curiosity formulations.

4.1 Environments

Atari Similar to prior work, we demonstrate the effectiveness of our approach on 12 Atari games. We chose a subset of the Atari games to represent environments used in prior work and a range of difficulty levels. We excluded some games due to lack of audio (e.g. Amidar, Pong) or the presence of background music (e.g. RoadRunner, Super Mario Bros). The action space is different from the one used in large-scale curiosity [2], as we use Gym Retro [33] in order to access game audio, and Retro environments use a larger action space. The original work reported results using the minimal action space, Discrete(4), whereas we use Discrete(6). We note that the larger action space does slow exploration, but it is used for both our method and the baselines for fair comparisons. To compute audio features, we take an audio clip spanning 4 time steps and apply a Fast Fourier Transform (FFT). The FFT output is downsampled using max pooling to a 512-dimensional feature vector, which is used as input to the discriminator along with a 512-dimensional visual feature vector.

Habitat Navigation We also test our method in a navigation setting using Habitat [7] (Figure 3). In this environment, the agent moves around a photorealistic Replica scene [34]. We use the largest Replica scene, Apartment 0, which has 211 discrete locations. In each location, the agent can face in 4 directions. At each timestep, the agent takes one of 3 discrete actions: turn left, turn right, or move forward. As in our Atari experiments, the agent is not given any extrinsic reward; we simply want to see how well it can explore the area without supervision. We use the audio-visual navigation extension from Chen et al. [23], which emits a fixed audio clip from a fixed location and allows our agent to hear the sound after simulating room acoustics. The perceived sound at each time step is less than 1 second long, and we zero pad this audio to 1 second. We apply FFT and downsample to a 512-dimensional feature vector, the same as done in Atari, described above.

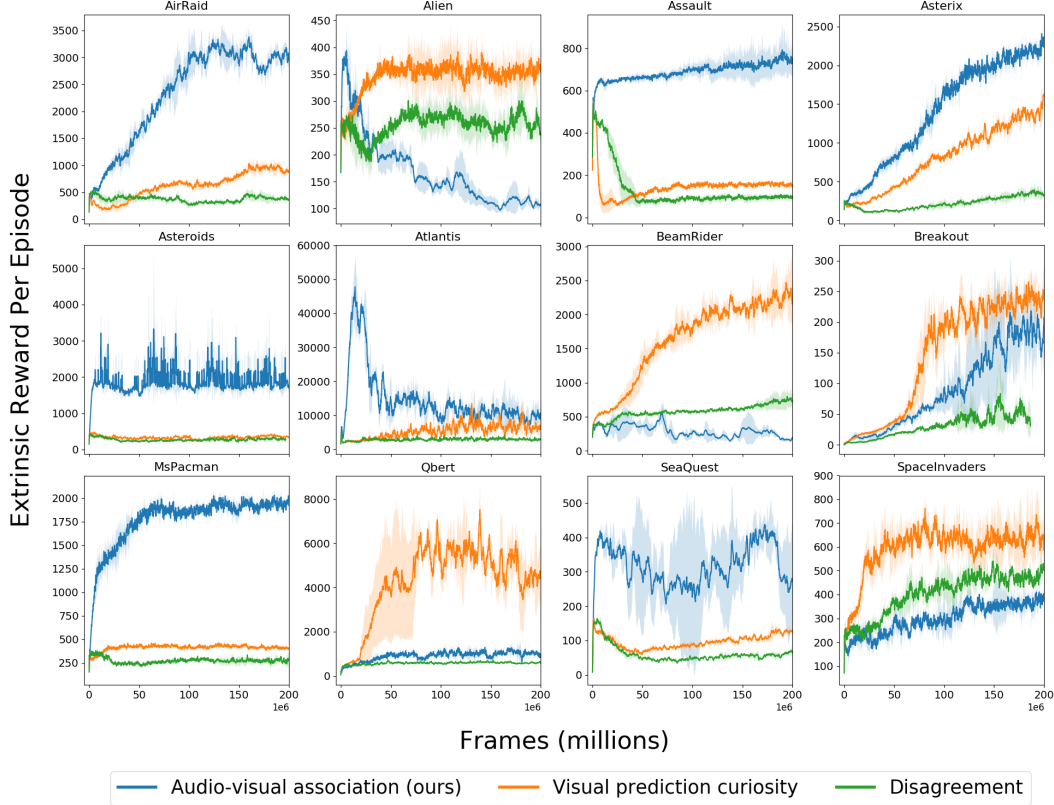


Figure 4: **Atari training curves**: Average extrinsic reward (never seen by the agent) throughout training for our method, visual prediction [2], and exploration via disagreement [3]. Our method outperforms the baselines in 8 of 12 environments, supporting our hypothesis that audio-visual association is a useful signal for accelerating exploration.

4.2 Baselines

We compare to large-scale curiosity [2], which as previously described performs visual prediction. We use the open-source code from the authors (see the supplementary material for more details). We also compare to exploration via disagreement [3] and Random Network Distillation (RND) [35]. We use the same hyperparameters (which were optimized for baselines) for policy learning across all approaches. We use random features [2, 35] for the visual feature representation for our method and the baselines in all experiments.

4.3 Atari Experimental Results

We trained our approach and baselines for 200 million frames using the *intrinsic* reward and measure performance by the *extrinsic* reward throughout learning. Figure 4 shows these results. Each method was run with three random seeds, and the plots show the mean and standard error for each method. Please see the supplementary material for more experimental details. Across many environments, our method enables better exploration (as judged by the extrinsic reward) and is more sample efficient than the baselines. Of the 12 environments, SHE outperforms disagreement in 9 and visual prediction in 8. We hypothesize that states leading to novel audio-visual associations, such as a new sound when killing an enemy, are more indicative of a significant event than ones inducing high prediction error (which can happen due to inaccurate modeling or stochasticity) and this is why our approach is more efficient across these environments.

Understanding Failure Cases While our approach generally exceeds the performance of or is comparable to the curiosity baselines, there are some environments where SHE underperforms. We have analyzed these games and found common failure cases: 1) Audio-visual association is trivial.

For example in Qbert, the discriminator easily learns the associations: every time the Qbert agent jumps to any cube the same sound is made, thus making the discriminator’s job easy, leading to a low agent reward. Visiting states with already learned audio-visual pairs is necessary for achieving a high score, even though they may not be crucial for exploration. The game Atlantis had similarly high discriminator performance and low agent rewards. 2) The game has repetitive background sounds. Games like SpaceInvaders and BeamRider have background sounds at a fixed time interval, but it is hard to visually associate these sounds. Here the discriminator has trouble learning basic cases, so the agent is unmotivated to further explore.

Hard Exploration Environment According to Taïga et al. [36], Gravitar is a hard exploration environment. Such environments are particularly difficult to solve without learning from demonstrations [25], using extrinsic reward [36], or exploiting structure in the game [37]. Even for humans, it can be unclear how to play Gravitar upon first introduction, in contrast with other Atari games that are intuitively simple. Despite Gravitar’s difficulty, audio-visual curiosity enables the agent to explore well, while the baselines perform poorly (Figure 5). After examining the game, we hypothesize that the game’s visual dynamics are not that interesting on their own, but the audio-visual associations are. We also applied our method and the baselines to other hard exploration games, but in these cases, no method was successful in the training time allotted.

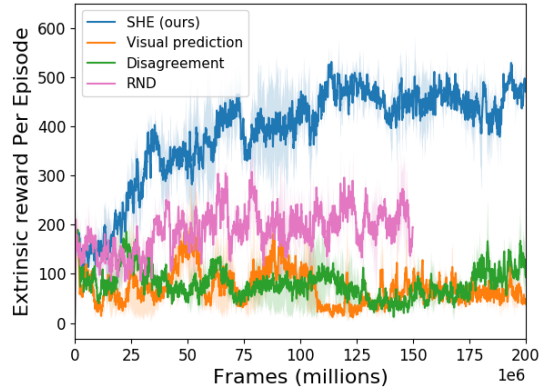
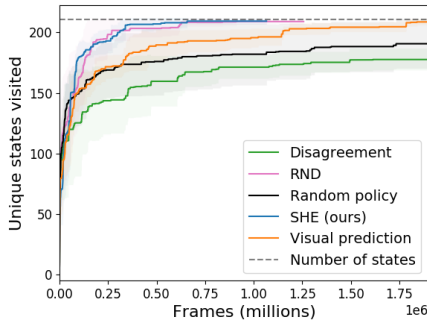
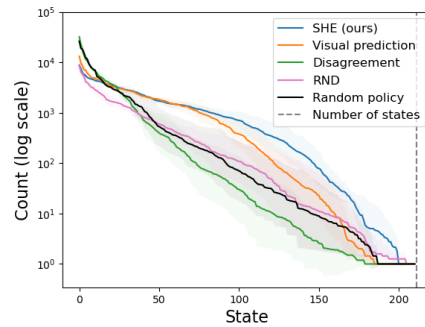


Figure 5: **Case study on Gravitar:** Our method is able to explore this hard environment, while baselines have negligible increase in extrinsic rewards. Due to compute constraints, this figure uses two random seeds.

4.4 Habitat Experimental Results



(a) State coverage, as measured by unique states visited throughout training. Our method achieves full state coverage about 3 faster than visual prediction curiosity.



(b) State counts: the number of times each state is visited in the first 2000 episodes, sorted by frequency and shown on a log scale. Our method has a wider tail, visiting rare states one to two orders of magnitude more frequently than the baselines.

Figure 6: Habitat exploration results for SHE and baselines. Each method is run with three different seeds and each seed uses a different start location.

Here we present results from unsupervised area exploration in the biggest scene in Replica [34] with realistic acoustic responses [23]. Figure 6 shows the quantitative results. SHE (blue) has similar coverage to RND and reaches full state coverage 3 times faster than visual prediction curiosity (Figure 6(a)). We can also look at how much each state is visited (Figure 6(b)). A good exploration method will have higher counts in the rare states. Our method visits these rare states (Figure 6(b) right) about

8 times more frequently than the next-best baseline. It does so by visiting common states (Figure 6(b) left) less frequently. SHE’s strong performance on this more realistic task holds promise for future work exploring the real world.

4.5 Ablations

Audio in baseline One hypothesis for why our method outperforms baselines is that SHE has access to additional information in the form of audio. To test the benefit of including audio without the use of our association method, we created an audio-visual prediction baseline. In this baseline, the prediction space is concatenated audio and visual features: the intrinsic model takes an audio-visual feature vector as input and predicts an audio-visual feature vector. As the results in supplementary material indicate, this does not lead to significant improvement over the visual-only baseline.

Robustness to noise Predicting the future can be especially difficult in the face of inherent uncertainty. To analyze our approach in such a setting, we created a noisy version of the environments, where Gaussian noise is added to the intrinsic model inputs. Our approach can be affected by noise in both audio and visual observations, whereas the baseline is only affected by the visual noise. For these experiments, we chose three environments: one where our method was better (MsPacman), one where the baseline was better (SpaceInvaders), and one where both methods performed well (Asterix). Figure 7 depicts results across these three environments both with and without noise. We observe that visual prediction curiosity is not robust to such noise: the performance degrades significantly in both Asterix and SpaceInvaders. In contrast, as our approach only relies on associations, it is more robust to such noise; thus only resulting in a smaller performance drop in one environment.

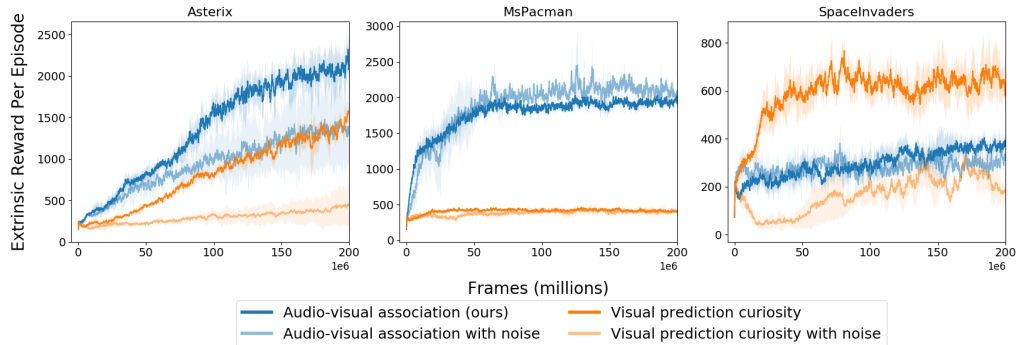


Figure 7: **Effect of input noise on performance:** Our method (blue) maintains similar performance with the introduction of noisy observations, while the baseline performance (orange) degrades.

Multiple Curiosity Modules Curiosity can have multiple forms, e.g. prediction-based and multimodal, and these are complementary to each other. To demonstrate this, we ran a joint method combining intrinsic rewards: we sum the losses from visual prediction and the audio discriminator. The resulting method is better than the visual-only baseline in 10 of 12 games, sometimes surpassing both (see supplementary for the detailed results).

5 Conclusion

Multimodality is one of the most basic facets of our rich physical world. Our formulation of curiosity enables an autonomous agent to efficiently explore a new environment by exploiting relationships between sensory modalities. With results on Atari games, we demonstrated the benefit of using audio-visual association to compute the intrinsic reward. Our method showed improved exploration over baselines in several environments. The most promise lies in our approach’s significant gains when used on a more realistic task, exploration in the Habitat environment, where audio and visual are governed by the same physical processes. We anticipate multimodal agents exploring in the real world and discovering even more interesting associations. Instead of building robots that perform like adults, we should build robots that can learn the way babies do. These robots will be able to explore autonomously in real-world, unstructured environments.

Broader Impact

The lasting impact of RL will be from these algorithms working in the real world. As such, our work is centered around increasing sample efficiency and adaptability. By leveraging self-supervision, we can avoid cumbersome reward shaping, which becomes exponentially more difficult as tasks grow more complex. Although our work here uses simulated agents, our longer-term goal is to deploy multimodal curiosity on physical robots, enabling them to explore in a more sample-efficient manner. Multimodal learning could have a near-immediate impact in autonomous driving, where different sensory modalities are used for perception of near, far, small, and large entities.

Autonomous RL agents have many potential positive outcomes, such as home robots aiding elderly people or those with disabilities. They will save time and money in many sectors of industry. However, they also have the potential to displace significant parts of the workforce [38].

There could be privacy concerns if merged multimodal data is hard to anonymize. There could also be privacy concerns with respect to recording audio data in the wild [39]. With unsupervised RL, it can be hard to predict what behaviors will be learned. For example, a robot using our algorithm might learn to damage sensors to create novel associations. The inability to predict agent behavior can make ensuring safety difficult, which would have consequences in safety-critical settings like autonomous driving or healthcare. Some work has been done on safety in RL [40], and there is more to be done, especially on analyzing the safety of RL exploration policies during training.

References

- [1] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *CoRR*, abs/1705.05363, 2017. URL <http://arxiv.org/abs/1705.05363>.
- [2] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.
- [3] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. *arXiv preprint arXiv:1906.04161*, 2019.
- [4] JMcVLevine Hunt. Intrinsic motivation and its role in psychological development. In *Nebraska symposium on motivation*, volume 13, pages 189–282. University of Nebraska Press, 1965.
- [5] William N Dember and Robert W Earl. Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological review*, 64(2):91, 1957.
- [6] Barbara A Morrongiello, Kimberley D Fenwick, and Graham Chance. Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior and Development*, 21(4): 543–553, 1998.
- [7] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [8] Katsumi Watanabe and Shinsuke Shimojo. When sound affects vision: effects of auditory grouping on visual motion perception. *Psychological Science*, 12(2):109–116, 2001.
- [9] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *CoRR*, abs/1703.01732, 2017. URL <http://arxiv.org/abs/1703.01732>.
- [10] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.
- [11] Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *CoRR*, abs/1507.00814, 2015. URL <http://arxiv.org/abs/1507.00814>.
- [12] Rein Houthoofd, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Curiosity-driven exploration in deep reinforcement learning via bayesian neural networks. *CoRR*, abs/1605.09674, 2016. URL <http://arxiv.org/abs/1605.09674>.
- [13] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

- [14] Jürgen Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991.
- [15] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*, pages 1471–1479, 2016.
- [16] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pages 2753–2762, 2017.
- [17] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.
- [18] Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*, 2016.
- [19] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.
- [20] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016.
- [21] Andrew Owens and Alexei 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.
- [22] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018.
- [23] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Audio-visual embodied navigation. *arXiv preprint arXiv:1912.11474*, 2019.
- [24] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. *arXiv preprint arXiv:2005.01616*, 2020.
- [25] Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. In *Advances in Neural Information Processing Systems*, pages 2930–2941, 2018.
- [26] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *CoRR*, abs/1805.11085, 2018. URL <http://arxiv.org/abs/1805.11085>.
- [27] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. *arXiv preprint arXiv:1805.04201*, 2018.
- [28] Samuel Clarke, Travers Rhodes, Christopher G Atkeson, and Oliver Kroemer. Learning audio feedback for estimating amount and flow of granular material. In *Conference on Robot Learning*, pages 529–550, 2018.
- [29] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *CoRR*, abs/1810.10191, 2018. URL <http://arxiv.org/abs/1810.10191>.
- [30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [33] Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, and John Schulman. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018.

- 394 [34] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul
395 Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan,
396 June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler,
397 Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele,
398 Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv*
399 *preprint arXiv:1906.05797*, 2019.
- 400 [35] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network
401 distillation. *arXiv preprint arXiv:1810.12894*, 2018.
- 402 [36] Adrien Ali Taïga, William Fedus, Marlos C Machado, Aaron Courville, and Marc G Bellemare. Benchmark-
403 ing bonus-based exploration methods on the arcade learning environment. *arXiv preprint arXiv:1908.02388*,
404 2019.
- 405 [37] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new
406 approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.
- 407 [38] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*,
408 358(6370):1530–1534, 2017.
- 409 [39] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions,
410 concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer*
411 *Interaction*, 2(CSCW):1–31, 2018.
- 412 [40] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal*
413 *of Machine Learning Research*, 16(1):1437–1480, 2015.