SEMI: SELF-SUPERVISED EXPLORATION VIA MULTISENSORY INCONGRUITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient exploration is a long-standing problem in reinforcement learning. In this work, we introduce a self-supervised exploration policy by incentivizing the agent to maximize multisensory incongruity, which can be measured in two aspects: perception incongruity and action incongruity. The former represents the uncertainty in multisensory fusion model, while the latter represents the uncertainty in an agent's policy. Specifically, an alignment predictor is trained to detect whether multiple sensory inputs are aligned, the error of which is used to measure perception incongruity. The policy takes the multisensory observations with sensory-wise dropout as input, and outputs actions for exploration. The variance of actions is further used to measure action incongruity. Our formulation allows the agent to learn skills by exploring in a self-supervised manner without any external rewards. Besides, our method enables the agent to learn a compact multimodal representation from hard examples, which further improves the sample efficiency of our policy learning. We demonstrate the efficacy of this formulation 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" (Bellemare et al., 2016; Lopes et al., 2012; Poupart et al., 2006) of less-frequently visited states, "curiosity" (Oudeyer & Kaplan, 2009; Pathak et al., 2017; Schmidhuber, 1990) where prediction error is used as reward signal and "diversity rewards" (Eysenbach et al., 2018; Lehman & Stanley, 2011) 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 continuous space. And for "predictive model" based method, the key challenge is to model and interact with the stochastic world.

As humans, we experience our world through a number of simultaneous sensory streams. The coincidence of senses gives us strong evidence that they were generated by a common, underlying event (Sekuler et al., 1997), since it is unlikely that they co-occurred across multiple modalities merely by chance. Thus, the incongruity between multisensory streams can be used as a strong signal of novelty. Researches in psychology suggest that this incongruity can attract human's attention and trigger further exploration (Berlyne et al., 1963; Dember & Earl, 1957), which has been widely used in product design (Ludden et al., 2012; 2008). Besides, humans are able to integrate multimodal sensory information in a near-optimal manner for decision making (Angelaki et al., 2009; Ma & Pouget, 2008), and are even robust to the loss of some senses (Hoover et al., 2012; Kolarik et al., 2014). Sensory compensation empowers humans to make similar decisions when different senses are used (Cohen et al., 1997; Bavelier & Neville, 2002; Lee et al., 2001). Thus, the incongruity of decisions which are made under different combinations of senses can also be used a signal of novelty.

However, few exploration policies are designed around multimodal feedback, *e.g.*vision, audition and touch. The difficulties are mainly reflected in two aspects: how to leverage multiple modalities with very different dimensions, frequencies and characteristics; and how to measure novelty with multimodal feedback. 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*.

For perceptual incongruity, 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.

For action incongruity, a policy network is trained with multi-modal dropout during multisensory fusion. Proposed by Srivastava *et al.* (Srivastava et al., 2014), *dropout* has been widely used to prevent neural networks from overfitting (LeCun et al., 2015; Huang et al., 2017). Gal *et al.* (Gal & Ghahramani, 2016; Kendall & Gal, 2017) 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. Here we adopt a similar approach by randomly dropping one or several modalities during multisensory fusion to imitate different combinations of senses. The variance of actions suggested by the policy network under different dropout states is used to measure action incongruity, which can also be used as an intrinsic reward.

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

The contributions of this paper can be summarized as follows. Inspired by psychology, we propose SEMI, a novel self-supervised exploration policy through discovering multisensory incongruity; SEMI enables agents to learn compact multimodal representation from hard examples; we demonstrate the efficacy of this formulation across a variety of benchmark environments including object manipulation and audio-visual games; furthermore, we show that SEMI is complementary to other intrinsic and extrinsic rewards.

2 Related Works

Explore with Intrinsic Reward Consider an agent that sees an observation, takes an action and transitions to the next state. We aim to incentivize this agent with a reward relating to how informative the transition was, so that the agent can explore the complicated environment more efficiently. One simple approach to encourage exploration is to use state visitation counts (Bellemare et al., 2016; Fu et al., 2017; Tang et al., 2017), where one maximizes visits on less frequent states. However, counting in the continuous space is usually challenging. Recently a more popular approach is using prediction error (Pathak et al., 2017; Schmidhuber, 1990; Burda et al., 2018), prediction uncertainty (Houthooft et al., 2016; Osband et al., 2016), or improvement (Lopes et al., 2012) of a forward dynamics or value model as intrinsic rewards. As a result, the agent is driven to reach regions of the environment that are difficult to reason with the current model.

A concurrent work from Dean *et al.* (Dean et al., 2020) has also demonstrated the effectiveness of using multisensory signals as intrinsic rewards. Specifically, they focus on the association of audio and visual signals as intrinsic rewards for RL exploration. Different from them, our multisensory incongruity contains both perceptual incongruity and action incongruity.

Multimodal Self-supervised Learning Self-supervised methods learn features by training a model to solve a pretext task derived from the input data itself, without human labeling. A variety of pretext tasks have been proposed to learn representations from different modalities. Several works leverage the natural correspondence (Arandjelovic & Zisserman, 2017; Patrick et al., 2020) and synchronization (Owens & Efros, 2018; Lee et al., 2019) between the audio or tactile and RGB streams to learn representations. Others use a modality distillation framework to learn video (Owens et al., 2016) and sound (Aytar et al., 2016) representations. Recent works have also found that multi-

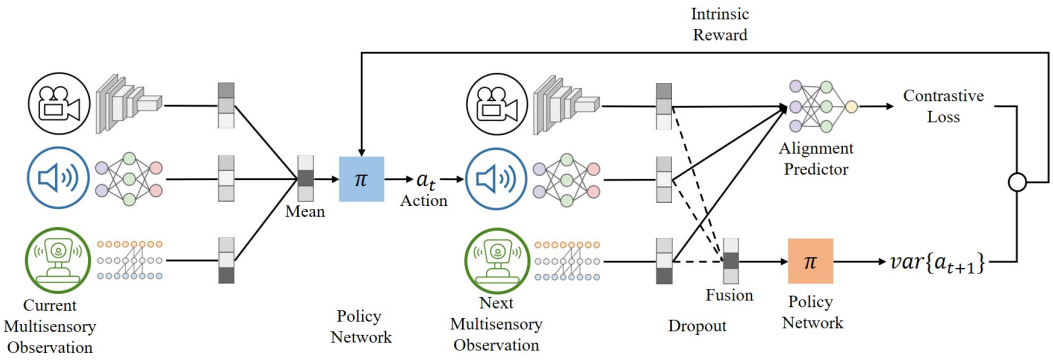

Figure 1: SEMI overview: at time step t, an agent takes action a_t given a multisensory observation O_t as input and ends up in a new state. The multisensory fusion model takes a new observation O_{t+1} as input and predicts whether these sensory inputs are aligned. The prediction loss is used as the 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 π.

modal learning can lead to more robust representations as they can partly account for the different learning speeds of the different modalities (Alwassel et al., 2019).

Noise-contrastive Estimation Noise-contrastive estimation (Hadsell et al., 2006; Gutmann & Hyvärinen, 2010; Mnih & Kavukcuoglu, 2013) measures the similarity/compatibility between sample pairs in a representational space and is at the core of several recent works on unsupervised feature learning (He et al., 2019; Hadsell et al., 2006; Patrick et al., 2020; Chen et al., 2020). It reduces a density estimation problem into a simpler probabilistic classification problem, circumventing the need to design handcrafted tasks in the raw signal space. Contrastive learning has recently been shown to yield good performance for image and video representation learning (Oord et al., 2018; He et al., 2019; Han et al., 2019). Prominently, Chen *et al.* (Chen et al., 2020) demonstrated that proper combination of data augmentation strategies and noise-contrastive re-identification achieves superior unsupervised learning results.

3 METHOD

3.1 FORMULATION

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 π 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 $O_{t+1} = \{o_{t+1}^1, o_{t+1}^2, ..., o_{t+1}^M\}$.

SEMI is composed of two sub-modules: an alignment predictor that generates a perceptual incongruity reward r_t^p, and a policy network that generates an action together with an action incongruity reward r_t^a. The alignment predictor observes multiple raw sensory streams, and detects the misalignment between them. We use the prediction error as a measure of perceptual incongruity. The policy network also observes multisensory inputs. The variance of actions suggested by the policy network given different modalities is used to measure action incongruity. Both incongruities are fed to the agent as intrinsic rewards to encourage its exploration. Figure 1 gives an overview of the formulation of SEMI, and we will detail each sub-module in the following.

3.2 MULTISENSORY PERCEPTUAL INCONGRUITY

The synchrony of multiple senses is a fundamental property of natural event perception, and we humans are extremely sensitive to the incongruity between them, which is a strong signal of novelty.

Therefore, to guide an agent to explore novel states, we propose an alignment predictor to discover this 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, and suffer from overfitting to trivial details or noises (Pathak et al., 2017).

Along the idea of contrastive learning (Oord et al., 2018; Chen et al., 2020), our design of alignment predictor directly maximizes agreement between different modalities of the same example 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 minibatch of N observations, there are $M \times (N-1)$ sensory streams from other examples, they are all treated as negative 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.*

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

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_k^i, o_k^j) is defined as

$$\mathcal{L}(o_k^i, o_k^j) = -\log \frac{\exp(\text{sim}(o_k^i, o_k^j)/\tau)}{\sum_{n=1}^{N}\sum_{m=1}^{M}\exp(\text{sim}(o_k^i, o_n^m)/\tau)}, \tag{2}$$

where τ denotes a temperature parameter.

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

3.3 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. Therefore in the setting of robot exploration, we use the incongruity with drop of senses as an indicator of novelty.

We use a modal-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 π_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,

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

3.4 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 π_r used to calculate intrinsic reward r_t^a is different from that used for exploration π. Inspired by Double Q-learning (Van Hasselt et al., 2016) and Dual Policy Iteration (Sun et al., 2018), π_r, with parameters θ being the same as π except that its parameters are copied every τ steps from the π. 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 $r_t = r_t^p + \gamma \times r_t^a$, where γ is a weight factor. The agent is optimized using PPO (Schulman et al., 2017) to maximize the expected reward according to:

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

4 Experiments

We evaluate the performance of SEMI in two environments, *OpenAI Robotics* and *Atari*. Three settings are considered and discussed: exploration with multisensory incongruity only (Section 4.1), combining multisensory incongruity with extrinsic reward (Section 4.2), and combining multisensory incongruity with other intrinsic rewards (Section 4.3).

4.1 Exploration via Multisensory Incongruity

4.1.1 Environment and Setting

OpenAI Robotics The first environment we evaluate on is the OpenAI Robotics (Plappert et al., 2018), where we consider the object manipulation task with MuJoCo. Our setup consists of a 7-DOF Fetch robotic arm that could be tasked to interact with the objects kept on the table in front of it. The objects are kept randomly in the workspace of the robot on the table. Robot's action space is 5-dimensional: a) 3 dimensions specify the desired gripper movement in Cartesian coordinates, b) angle of approach θ (1-dim), and c) gripper status (1-dim), a binary value indicating whether to grasp (open the gripper fingers) or push (keep fingers close). In this environment, we consider RGB input and Depth input as our multi-modal observation. To achieve this, we mount an RGBD camera at a fixed location from the robot to acquire $224 \times 224 \times 3$ RGB images and $224 \times 224 \times 1$ Depth images. To accurately grasp or push objects, the agent needs to figure out an accurate combination of location, orientation and gripper status.

Atari Our second environment is Atari, where vision and audio are considered as multi-modal inputs. Instead of using the Arcade Learning Environment (ALE) (Bellemare et al., 2013), we use Gym Retro (Nichol et al., 2018) in order to access game audio. We exclude some games due to lack of audio (*e.g.* Pong) and the presence of background music (*e.g.* RoadRunner), with more details in Appendix. In this environment, actions are encoded as a 12-dimensional vector representing 12 different buttons. We render gray-scale images with size to $64 \times 64 \times 1$, while audio is sample at 44.1kHz. At each time step, the observation contains 5 consecutive frames and their corresponding audio clip.

4.1.2 Training Details

OpenAI Robotics We use 5 convolutional layers to extract RGB features, and a similar network to extract depth features. Our policy network is a multi-layer perceptron (MLP) with 4 hidden layers.

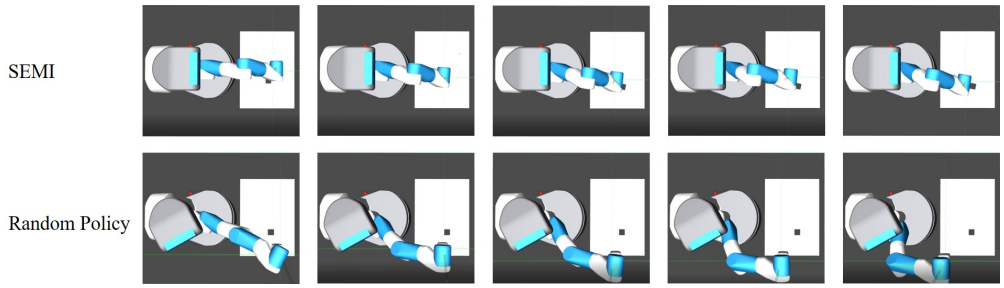

Figure 2: Object manipulation in MuJoCo. *Top*: Self-supervised exploration policy trained with multisensory incongruity (SEMI). *Bottom*: Random policy. In this example, agent trained with SEMI is able to interact with the object, whereas random policy fails.

We used PPO (Schulman et al., 2017) to maximize the intrinsic reward with an Adam optimizer. The learning rate is set to 10^{-3}. We empirically found that applying multisensory action incongruity in the later training stage could help the model training to better converge. So we use multisensory action incongruity after the policy is trained with 20 epochs, where the policy is trained 50 epochs in total. The policy used for calculating multisensory action incongruity is updated every 5 epochs.

Atari We stack the 5 consecutive frames channel-wisely and use 5 convolutional layers to extract the image features. The observed audio waveforms are first transformed into normalized log-mel spectrograms with 512 frequency bins. Three 1D convolutions are then used to extract the audio features. Our policy network is a multi-layer perceptron (MLP) with 4 hidden layers to predict the action distribution. Actions are sampled from a normalized categorical distribution given the 12 action parameters. During training, we replace the reward in the sampled trajectory by multisensory perceptual incongruity, which will guide the algorithm as intrinsic reward. We also use PPO (Schulman et al., 2017) and store the advantage value into the replay buffer. Each time we collect a replay buffer, we iterate throughout the entire buffer 3 times and then move to next rollout. Adam optimizer with a learning rate at 10^{-4} is adopted to train both intrinsic model and policy model.

Note that we do not apply multisensory action incongruity on Atari games since agents cannot always make reasonable decisions given only audio signals.

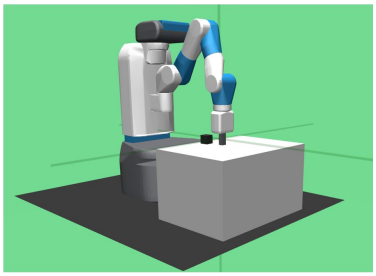

Figure 3: Experiment Setup.

Exploration Strategy		Interaction Rate
Uni-IR	Curiosity	2.7%
	Random	8.4%
	Disagreement	26.3%
	SEMI (P)	30.5%
	SEMI (PA)	34.4%
Multi-IR	Curiosity + SEMI (PA)	35.8%
	Disagreement + SEMI (PA)	37.1%

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).

4.1.3 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).

We evaluate two different versions of our method:

- *SEMI (P)*: We first use only the multisensory perceptual incongruity as our intrinsic reward, as described in Section 3.2.
- *SEMI (PA)*: Second, we use both multisensory perceptual incongruity and multisensory action incongruity as our intrinsic reward.

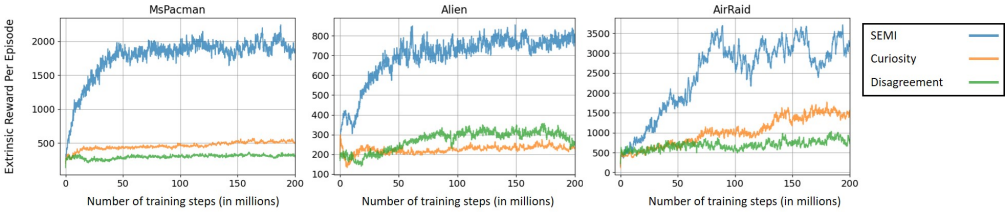

Figure 4: Average extrinsic reward of the agent during training with different intrinsic rewards.

We compare SEMI to the following baselines:

- *Curiosity*: Proposed by (Pathak et al., 2017; Burda et al., 2018), the exploration policy is trained jointly with a predictive model. The predictive model is trained to predict the next state from current state and action, and the error of which is used as intrinsic reward. The intuition is unpredictable situations are more likely novel and therefore ones the agent should explore.
- *Random*: As a sanity check, we propose to use a random policy, which moves randomly.
- *Disagreement*: Proposed by (Pathak et al., 2019), they first train an ensemble of dynamics models and incentivize the agent to explore such that the disagreement of those ensembles is maximized.

As shown in Table 1, our method, outperforms all of these baselines. The method of Disagreement (Pathak et al., 2019) has a performance close to that of our method. In Figure 2, we show examples where our method interacts with objects, whereas the baseline Random policy fails.

We perform an ablation analysis to quantify the performance of each component of our system (4th and 5th row in Table 1). We see that both multisensory perceptual incongruity and multisensory action incongruity contribute to the robot exploration.

Atari We also test out method in Atari Games. Figure 4 shows the extrinisic reward of Atari during exploration with SEMI in comparison of intrinsic reward via Curiosity and Disagreement. It should be pointed that during training the agent only has access to the intrinsic reward. As illustrated in Figure 4, our method obtains the best results comparing with all baseline methods. This results depict that SEMI can explore the environment more efficiently and more thoroughly.

4.2 Combining with Extrinsic Reward

Considering our proposed intrinsic rewards are designed to guide the agent to explore the environment, are they compatible with external reward? To verify this, we conduct additional experiments in OpenAI Robotics and Atari Games. While the network architecture and training schema are exactly the same as Section 4.1, we use the sum of SEMI and extrinsic rewards as training signal,

$$R_t = r_t + \beta \times r_t^{(e)} \tag{6}$$

where $r_t^{(e)}$ is the external reward provided by the environment. We set the reward weighting β to 1 in all the experiments.

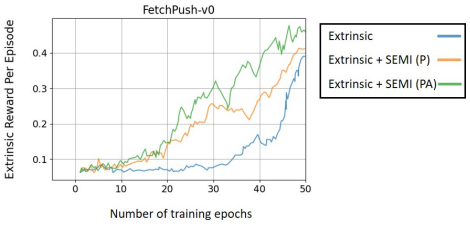

Figure 5: Average extrinsic reward of the agent trained with multisensory incongruity against without using intrinsic rewards.

Figure 5 shows the average extrinsic reward of the FetchPushing task training with multisensory incongruity against without using intrinsic rewards. The extrinsic rewards are sparse and binary: The agent obtains a reward of 0 if the goal has been achieved (within some task-specific tolerance) and -1 otherwise. Training with SEMI significantly improves the learning efficiency compared with training with only sparse extrinsic reward.

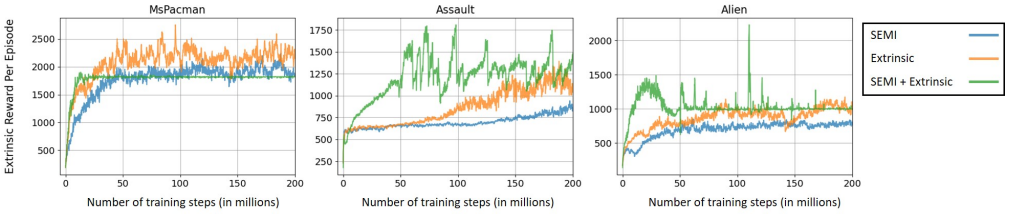

Figure 6: Average extrinsic reward of the agent trained with multisensory incongruity against without using intrinsic rewards.

By adding action variance in SEMI paradigm (PA), the performance improves further when policy model learns meaningful mapping from observation to action.

Figure 6 shows the effectiveness of our method for efficient exploration in Atari Games. Training with SEMI always leads to a faster convergence, which indicates that it is able to speed up exploring the environments. Besides, the final performance of the agent does not deteriorate with faster convergence, showing the compatibility of SEMI with any extrinsic rewards.

4.3 Combining with Other Intrinsic Rewards

We further show that exploration via multisensory incongruity is complementary to other self-supervised exploration methods, *e.g.* prediction-based curiosity. To demonstrate this, we simply sum the multisensory incongruity with other intrinsic rewards, and use it to train the agents. We evaluate this setup in both OpenAI Robotics and Atari Games. The network architecture and training schema are exactly the same as mentioned in Section 4.1.

Table 1 (5th and 6th row) shows the interaction rate of object manipulation during exploration with a combination of intrinsic rewards:

- *SEMI + Curiosity*: The agent receives a combination of intrinsic rewards: we sum the losses from multisensory incongruity and visual prediction error.
- *SEMI + Disagreement*: The agent receives a combination of intrinsic rewards: we sum the losses from multisensory incongruity and the disagreement of dynamics model ensembles.

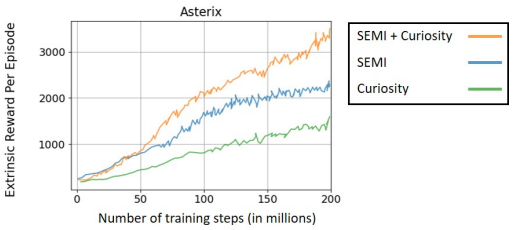

Figure 7: Average extrinsic reward of the agent training with multisensory incongruity joint with other intrinsic rewards.

The agent maximizing the sum of multiple intrinsic rewards explores better than an agent maximizing single intrinsic rewards, which shows that SEMI is complementary to many existing intrinsic rewards.

Similarly, Figure 7 shows the extrinsic reward of Atari during exploration with a combination of intrinsic rewards. The performance of an agent trained with a combination of SEMI and Curiosity performs better than trained with SEMI or Curiosity alone.

5 Conclusion

In conclusion, we propose a self-supervised exploration strategy by incentivizing the agent to maximize multisensory incongruity. We show that through the use of multisensory perceptual incongruity and multisensory action incongruity, our learned policy can explore the environment efficiently. We also show the compatibility of our proposed method with extrinsic rewards and other intrinsic rewards. We hope that our work paves the way towards to a direction for intelligent agents to continually develop knowledge and acquire new skills from multisensory observations without human supervision.

REFERENCES

Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

Dora E Angelaki, Yong Gu, and Gregory C DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–458, 2009.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pp. 892–900, 2016.

Daphne Bavelier and Helen J Neville. Cross-modal plasticity: where and how? *Nature Reviews Neuroscience*, 3(6):443–452, 2002.

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*, pp. 1471–1479, 2016.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Daniel E Berlyne, Margaret A Craw, Philip H Salapatek, and Judith L Lewis. Novelty, complexity, incongruity, extrinsic motivation, and the gsr. *Journal of Experimental Psychology*, 66(6):560, 1963.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Leonardo G Cohen, Pablo Celnik, Alvaro Pascual-Leone, Brian Corwell, Lala Faiz, James Dambrosia, Manabu Honda, Norihiro Sadato, Christian Gerloff, M Dolores Catala, et al. Functional relevance of cross-modal plasticity in blind humans. *Nature*, 389(6647):180–183, 1997.

Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. *arXiv preprint arXiv:2007.03669*, 2020.

William N Dember and Robert W Earl. Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological review*, 64(2):91, 1957.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Justin Fu, John Co-Reyes, and Sergey Levine. Ex2: Exploration with exemplar models for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2577–2587, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Adria EN Hoover, Laurence R Harris, and Jennifer KE Steeves. Sensory compensation in sound localization in people with one eye. *Experimental brain research*, 216(4):565–574, 2012.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pp. 1109–1117, 2016.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Alex Kendall and Yarin 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 (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5574–5584. Curran Associates, Inc., 2017.

Andrew J Kolarik, Silvia Cirstea, Shahina Pardhan, and Brian CJ Moore. A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing research*, 310:60–68, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Dong Soo Lee, Jae Sung Lee, Seung Ha Oh, Seok-Ki Kim, Jeung-Whoon Kim, June-Key Chung, Myung Chul Lee, and Chong Sun Kim. Cross-modal plasticity and cochlear implants. *Nature*, 409(6817):149–150, 2001.

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. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8943–8950. IEEE, 2019.

Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems*, pp. 206–214, 2012.

Geke DS Ludden, Hendrik NJ Schifferstein, and Paul Hekkert. Surprise as a design strategy. *Design Issues*, 24(2):28–38, 2008.

Geke DS Ludden, Hendrik NJ Schifferstein, and Paul Hekkert. Beyond surprise: A longitudinal study on the experience of visual-tactual incongruities in products. *International journal of design*, 6(1), 2012.

Wei Ji Ma and Alexandre Pouget. Linking neurons to behavior in multisensory perception: A computational review. *Brain research*, 1242:4–12, 2008.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pp. 2265–2273, 2013.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pp. 4026–4034, 2016.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

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)*, pp. 631–648, 2018.

Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pp. 801–816. Springer, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pp. 5062–5071, 2019.

Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 697–704, 2006.

Jürgen 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*, pp. 222–227, Cambridge, MA, USA, 1990. MIT Press. ISBN 0-262-63138-5. URL http://dl.acm.org/citation.cfm?id=116517.116542.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

R. Sekuler, A. Sekuler, and R Lau. Sound alters visual motion perception. *Nature*, pp. 308, 1997.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Wen Sun, Geoffrey J Gordon, Byron Boots, and J Bagnell. Dual policy iteration. In *Advances in Neural Information Processing Systems*, pp. 7059–7069, 2018.

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*, pp. 2753–2762, 2017.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

A APPENDIX

A.1 EXPLORATION VIA MULTISENSORY INCONGRUITY

We test SEMI on Atari Breakout, Assault, MsPacman, AirRaid, Alien, Qbert, Space Invaders, and Beam Rider. The average extrinsic reward of the agent training with multisensory incongruity outperforms curiosity on 5 out of 8 environments. On Atari Alien, MsPacman, *etc.*, our method converges faster and achieves better performances compared with baselines. The reason is that audio signals are always triggered by significant events (*e.g.* eating pellets). Thus, the multisensory incongruity is more indicative compared with curiosity and disagreement baselines, which are influenced by the stochasticity of the environments.

On environments such as Breakout, Space Invader, and Beam Rider, SEMI does not outperform the curiosity baseline. Environment like Breakout shows a trivial audio-visual association that does not contain enough information for the multisensory incongruity method. In the meantime, the Breakout environment is deterministic, which makes it easier to learn a predictive model. And in Beam Rider, the noisy background music keeps appearing whenever the agent is making a move. These also provide no information for the multisensory incongruity module to neither learn the association nor guide the RL algorithm to learn better policy.

A.2 COMBINING WITH EXTRINSIC REWARD

We combined the multisensory incongruity reward and extrinsic reward with equal weight. We tested the experiments on Atari MsPacman, AirRaid, Assault, and Alien. The average extrinsic reward of the agent significantly improved on 3 out of 4 environments compared with trained with only extrinsic reward or intrinsic reward. With the guidance of the extrinsic reward, the agent can learn to explore the environment in the early stage and also learn from the extrinsic reward signal.

A.3 COMBINING WITH OTHER INTRINSIC REWARDS

While experiments on our method lead to efficient self-exploration behavior, we test the compatibility of our method with other contemporary self-supervised learning methods. We test SEMI combined with curiosity method. On environments such as Atari Asterix, Breakout, Assault, MsPacman, AirRaid, Alien, Qbert, Space Invaders, and Beam Rider, the average extrinsic reward improved in 4 of 9 environments. On environment such as Asterix, the extrinsic reward converge significantly faster than trained with only visual prediction or SEMI method. But on environments like Alien and Space Invaders, the performance does not improve compared to the visual prediction baselines no matter whether multisensory incongruity outperforms curiosity. Since it is unclear how the intrinsic rewards will affect each other when trained jointly, it is possible that optimizing some rewards can bring negative impacts on the others, which will lead to a worse exploration efficiency.

A.4 FAILURE CASES

While SEMI generally shows improvement in exploring the environment and is compatible with training with extrinsic reward, there are still some Atari environments where it does not improve exploration efficiency 8. We dig into the games and analyze the feature of them to explain why these environments lead to failures.

1) The game presents constant sound patterns. For example in Beam Rider, there is a fixed background sound whenever the agent makes a move. Thus, the multisensory incongruity method will not learn useful patterns to distinguish the incongruity even in the basic situations, therefore the agent cannot learn from any meaningful intrinsic reward signal.

2) The game shows trivial multisensory association. In environments like Breakout, the audio is almost the same when the agent is interacting with the environment, *i.e.* the sound in Breakout only indicates the ball is making contact with objects in the scene. The multisensory incongruity module could easily distinguish the incongruity in almost all cases in the game. A newly reached game situation will not lead to high intrinsic rewards. Therefore, multisensory incongruity method cannot motivate the agent to explore unseen situations.

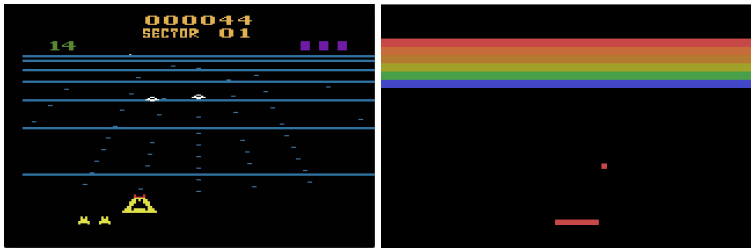

Figure 8: The failure cases where SEMI does not improve exploration efficiency. Beam Rider (Left), Breakout (Right)