In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23 Oct-27 Oct 2022.

Impact Makes a Sound and Sound Makes an Impact: Sound Guides Representations and Explorations

Xufeng Zhao[1,*], Cornelius Weber[1], Muhammad Burhan Hafez[1], and Stefan Wermter[1]

***Abstract*— Sound is one of the most informative and abundant modalities in the real world while being robust to sense without contacts by small and cheap sensors that can be placed on mobile devices. Although deep learning is capable of extracting information from multiple sensory inputs, there has been little use of sound for the control and learning of robotic actions. For unsupervised reinforcement learning, an agent is expected to actively collect experiences and jointly learn representations and policies in a self-supervised way. We build realistic robotic manipulation scenarios with physics-based sound simulation and propose the Intrinsic Sound Curiosity Module (ISCM). The ISCM provides feedback to a reinforcement learner to learn robust representations and to reward a more efficient exploration behavior. We perform experiments with sound enabled during pre-training and disabled during adaptation, and show that representations learned by ISCM outperform the ones by vision-only baselines and pre-trained policies can accelerate the learning process when applied to downstream tasks.**

I. Introduction

Research in the field of neuroscience shows that with multiple cues from a diverse range of sensory modalities comes enhanced behavioral performance towards faster response, more accurate movement, and a better sense of stimulus [1]. When presented with multiple modalities, e.g., a combination of auditory, haptic, and visual perception, an observer will make the *assumption of unity* that decides whether the multimodal information originates from a common source or from some separated objects and events [2]. The perception of unity arises when the perceiver assumes that a physical event is redundantly expressed and sensed across diverse modalities, and decisions are commonly made based on the temporal and spatial consistency of information [3], or on semantic congruence factors [1].

Undoubtedly, vision is extremely information-rich and is one of the most important senses for humans to perceive the world, but is nevertheless hard for a robot to directly extract knowledge from. Though the issue is dramatically alleviated when combined with deep neural networks, visual representations usually are hard to interpret and somehow constrain the tasks they are trained on. For many vision tasks, a common behavior begins by constructing neural networks based on pre-trained models, or by training neural networks in a self-supervised way, e.g., an intra-modal design of simple but diverse sub-tasks [4], or crossmodal prediction of information consistency [5], [6]. However, only the later design can, at least partially, persist the assumption of unity.

In most scenarios, a vision-based reinforcement learner requires to learn representations and policy jointly [7]. Both are highly coupled: sufficient and stable representations are essential for policy learning [8]; a diverse and near-optimal policy is needed to collect samples to learn unbiased representations. Humans can benefit from multiple sensing cues in terms of both perception and behavior. Intuitively, an active agent who is allowed to explore freely can benefit from multimodal cues in two aspects: 1) learning meaningful representations by crossmodal self-supervision [9], [10], [11], and 2) being intrinsically motivated to explore the environment under the unity assumption reflected by the uncertainty of crossmodal predictions.

Sounds are generally much more distinctive compared with visual events. For some specific tasks related to physical properties estimation, the sound alone is reliable to guide a robot and measure its performance [12]. For others, it may be informative but not sufficient, e.g., a classification of objects that share common auditory properties [13], or precise control of a water-pouring robot [14]. In this case, sounds are supposed to fuse with other sensory inputs to present a much more robust description of states, or to scaffold the agent's exploration.

There are more chances that sound is abundantly distributed while hardly considered for general manipulations due to the facts that 1) vision is content-rich and is thus sufficient for traditional planning-based robots so the sound is often ignored; 2) the correlation of sound events with a task goal could be implicit to program or to discover automatically by traditional methods, which further limits its exploitation. However, things go the other way when a deep reinforcement learner is deployed to control. 1) Learning exclusively with vision can be exhausting. Though deep neural networks are capable of extracting features from high dimensional inputs, there is no guarantee of information sufficiency as samples are collected gradually. Representations can possibly overfit the trajectories of a non-optimal agent, especially when transferred to new scenes where a biased policy could lead to a worse learning process. Moreover, exploration time for robots is often desired to be minimal for natural wear and safety concerns, which calls for the necessity of efficient and robust pixel interpretation. 2) Fortunately, latent associations among modalities [15], [16] and behavior consequences [17] can be discovered

This research was funded by the German Research Foundation (DFG) in the project Crossmodal Learning (TRR-169) and the China Scholarship Council (CSC).

[1]The authors are with the Knowledge Technology Group, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany. E-mail: {xufeng.zhao, cornelius.weber, burhan.hafez, stefan.wermter}@uni-hamburg.de.

*Corresponding author, Email: xufeng.zhao@uni-hamburg.de.

automatically by deep learning, which prompts the potential of crossmodal control.

Therefore, our approach contains two phases: first, to train the image encoder of a Reinforcement Learning (RL) agent with visual-auditory correlations, and second, to use the crossmodal error as an intrinsic reward to encourage meaningful exploration. Contributions in this paper include: 1) the ManipulateSound[†] environment built upon the ThreeDWorld simulator [18] that comprises robotic control with physically generated sound (see Fig.1); 2) a general architecture to utilize sound feedback for unsupervised RL exploration, resulting in more robust representation and active exploration.

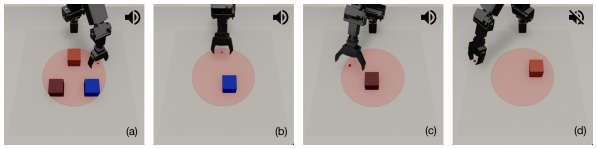

Fig. 1. *ManipulateSound* environments with different objects that have different physical properties: (a) a task with three different cubes to push out; (b) a fine-tuning task with a single blue ceramic cube to push out; (c) a task with a single brown wooden cube to push out; (d) a task with a single red metal cube to push out; sound intentionally turned off during evaluation.

II. Related Work

We introduce sound to boost self-supervised representation learning as well as active exploration of unsupervised RL agents.

A. Self-supervised Representation Learning

Self-supervised learning is a collection of methods to learn representations from data that has automatically created pseudo-labels according to certain objectives. Based on the sensory inputs, they can be roughly classified into two categories: intra-modal and crossmodal self-supervised learning.

A common intra-modal way to create pseudo-labels of images is to perform multiple parameterized augmentations. Then, neural networks are trained to predict which transformation has been carried out on each sample [4], [19]. We argue that it makes more sense to a robot when the transformation owns a realistic meaning. For instance, to obtain representations with ego-motion equivariance addressed, images are collected with a camera on a moving car and grouped into neighbor pairs by driving commands [20]. The forward model in the Intrinsic Curiosity Module (ICM) [21] predicts s_{t+1} with (s_t, a_t) so that the agent can learn to represent the environmental dynamics.

Self-supervised representation learning is naturally applicable for scenarios with multiple modalities involved. Representations emerge concurrently with different focuses and biases, but often have strong relations from one to another. To jointly model multiple modalities, such as audio and visual components of videos [22], a binary classification model to discriminate whether the visual and auditory input are aligned [23], [24], or a regression model to predict corresponding audio statistics given vision [25] can be established. Although these settings are simple enough, they reveal the unity assumption of events, such that extraordinary abilities can be acquired, e.g., sound localization, audio-visual retrieval [6] and speech separation [26]. In our case, we train a discriminating model which is easy to implement and applicable for general usage.

When applied to robotic control, the available sensory perception is much more diverse [27], [28], [29], [30]. A work by [31] shows that a fused state of visual input, force-torque sensing, and proprioception trained by self-supervision is beneficial for sample efficiency. However, it can be difficult to handcraft such sub-tasks and properly assign weights among modalities. We keep the complexity low by focusing on the impact of sound.

B. Active Exploration

A reinforcement learning agent can gain remarkable abilities by purely maximizing the reward of experiences [17]. However, for a task with sparse rewards [32], [33], which is a common case, the learning process can be quite slow due to the inefficiency of sampling. Reward-shaping [34] is a commonly used method to alleviate this problem, but it requires expert knowledge and human effort to tune and is vulnerable to environmental disturbance. Many active exploration strategies have been investigated to encourage the agent to seek novel states [35], [36], [21], [37] among which ICM proves to be robust on many tasks [8], [7]. So we construct our auditory-curiosity module on top of ICM, building on an existing visual processing pathway.

As an alternative to sound, haptic sense [38] achieves good performance and active exploration in terms of frequent contacts, supporting sample-efficient learning. Similar to our work, [39] use vision and action to predict next clustered auditory events, and the classification error will thus be used as the overall intrinsic reward. However, the transferability of learned representations is not as well studied as in our work. Work in [24] trains a discriminator to exploit information consistency of aligned image sequences and audio, and intrinsic reward is computed according to the uncertainty of the classifier. Despite the extra efforts required to construct offline data sets, they are restricted to Atari games or audio-dense scenarios. When applied to robotic control, an object will only produce sound when there is a contact. Silence or background noise dominates most of the time. It is even harder to construct misaligned pairs because a random shuffle strategy fails in cases where silence is capable of being aligned with most of the visual scenes. Moreover, a cold-starting problem will arise, particularly when the policy is not sufficiently rewarded to produce collisions. Therefore, we use intrinsic motivations extracted from both visual and auditory cues.

III. Intrinsic Sound Curiosity Module

Typical reinforcement learning problems are formulated as Markov Decision Processes (MDPs), comprised by states $\mathcal{S} = \{s_t\}$, actions $\mathcal{A} = \{a_t\}$, transition probability $\mathcal{P}^a_{ss'}$,

[†]https://github.com/xf-zhao/ManipulateSound

and rewards $\mathcal{R} = \{r_t\}$. The goal of the agent is to find the optimal policy $\pi^*(s_t, a_t)$ that maximizes the expected discounted sum of rewards $\mathbb{E}_\pi \sum_{n=0}^{\infty} \gamma^n r_{t+n}$. Usually, out of realistic constraints and generality consideration, we do not have full access to internal states $\mathcal{S}$ but a series of sensors attached to the workspace, resulting in partial observations $\mathcal{O} = \{o_t\}$. Before being fed into the policy module, high-dimensional sensory inputs must be compressed to latent states that can efficiently represent the environment [40], [8].

A. Visual Representation Learning

Visual exploration is a fundamental task for embodied AI agents, where the agent is allowed to actively gather visual information about the environment and then distill knowledge into models such as a topological map or a dynamics model [41]. Generally, the agent is supposed to explore as many novel states as possible with an internal encouragement aligned to certain targets, e.g. a measure of the *coverage* such as the amount of visited unique states in a navigation scenario [24], a *prediction error* of a learned dynamics model [21], [42] or of a reconstruction model when an agent tries to generate other views of an object than the observed ones [41].

With a combination of multiple sensory inputs for internal states, the agent is allowed to have a more comprehensive view of the environment. However, it will require either a lot of domain-specific assumptions or an increase in model complexity [15], [43] to derive efficient representations from fused inputs. In order to make a fair comparison with vision-only baselines, we use sound in a supplementary way. Only in the pre-training stage has the agent access to sound. The baseline encoder is trained by dynamically modeling the environment with visual states, while the one of ISCM (Intrinsic Sound Curiosity Module) additionally fits a visual-auditory sub-task (see Fig.2). Before adaptation to downstream tasks, visual encoders of the DDPG [44] learner are initialized with weights from the ISCM and ICM baseline.

Let the visual and auditory observation at time step t be denoted as o_t^V and o_t^A, respectively. A visual encoding function $\varphi(\cdot)$ comprised of convolutional neural networks is thus applied on o_t^V to compute the state $s_t = \varphi(o_t^V)$, which is later used for both policy learning and dynamic environment modeling. Evidence shows that a well-pre-trained encoder is essential for the generalization of supervised learning models [45], [4] and RL agents [7], [8]. Hence, the sound-free visual encoder $\varphi(\cdot)$ and the sound-guided counterpart $\tilde{\varphi}(\cdot)$ are trained separately for comparison.

There are two jointly-trained dynamics models in ICM: a forward model $\mathbb{D}^F$ and an inverse model $\mathbb{D}^I$. The forward model tries to predict the forward n-step transition s_{t+n} given the current state s_t and action a_t, i.e. $\hat{s}_{t+n} = \mathbb{D}^F(s_t, a_t)$, while the inverse one tries to predict the action taken between aligned states $\hat{a}_t = \mathbb{D}^I(s_t, s_{t+n})$, which encourages noise-robust representations [21]. These two dynamics models are optimized concurrently with respect to L_2 constraints, defined as $L_t^F = \|\hat{s}_{t+n} - s_{t+n}\|_2^2$ and $L_t^I = \|\hat{a}_t - a_t\|_2^2$. Note that here we use L_2 loss also for action predictions since we control the continuous actions of the robot arm, otherwise a cross-entropy loss can be considered for discrete actions.

To benefit from sound, a crossmodal prediction model $\mathbb{C}(\cdot)$, which can be either a discriminator $\mathbb{C}^D(\cdot)$ or a regressor $\mathbb{C}^R(\cdot)$, is then trained to learn the associations of concurrent vision and sound. It is optimized by minimizing the error between the visual-auditory projection $\hat{s}_t^A = \mathbb{C}(s_t)$ and the targeted latent auditory feature $s_t^A = \phi(o_t^A)$, where $\phi(\cdot)$ is a fixed auditory encoder with output suitable for either discrimination or regression. Typically, to construct auditory features for regression, $\phi(\cdot)$ consists of randomly initialized neural networks, with no requirements of any further training. These representations are compact, stable, and generally reliable [8], [37], especially when dealing with impact sound whose information density could be low compared to information in speech. Alternatively, $\phi(\cdot)$ can be chosen as a threshold to distinguish valid event sound from background noise, considering the simplicity and the aforementioned knowledge that even with a simple discriminating task, surprisingly good abilities can be acquired through cross-modal learning [6], [24], [5]. Much of the time in a manipulation scenario, there is just silence before any valid collision or friction happens. To avoid the model eagerly collapsing to zero prediction and causing dying neurons [46], we use weighted cross entropy loss by ω to amplify the importance of positive samples, i.e.

$$L_t^{\mathbb{C}^D} = -\omega \cdot s_t^A \log \hat{s}_t^A - \left(1 - s_t^A\right) \log(1 - \hat{s}_t^A). \quad (1)$$

For regression, the optimization is similar except for an unweighted L_2 loss $L_t^{\mathbb{C}^R} = \|\hat{s}_t^A - s_t^A\|_2^2$.

To summarize, the optimal encoders for visual representations in vanilla ICM and the proposed ISCM are separately written as

$$\varphi^* = \arg\min_{\varphi} \mathbb{E}_t \left[L_t^{\mathbb{D}}\right] \quad (2)$$

and

$$\tilde{\varphi}^* = \arg\min_{\tilde{\varphi}} \mathbb{E}_t \left[(1-\alpha) L_t^{\mathbb{D}} + \alpha L_t^{\mathbb{C}}\right], \quad (3)$$

where $L_t^{\mathbb{D}} = \beta L_t^F + (1-\beta) L_t^I$ is the overall dynamics loss and α, β are hyper-parameters to mediate the relative importance between modules. Note that the objective is expected to be minimized over samples with time stamp t. Therefore, it is reasonable to encourage the agent to collect informative samples by injecting the model's prediction error, as a form of intrinsic reward, into the agent's exploration objective.

B. Intrinsic Visual-Auditory Reward

Unlike typical supervised learning in which the data is drawn from a stationary distribution, RL agents actively seek samples according to the policy that updates towards reward-weighted maximum likelihood estimation [47]. So when dealing with the sparse-reward case, the intrinsic reward mechanism helps prevent representations to focus too much on non-interesting areas.

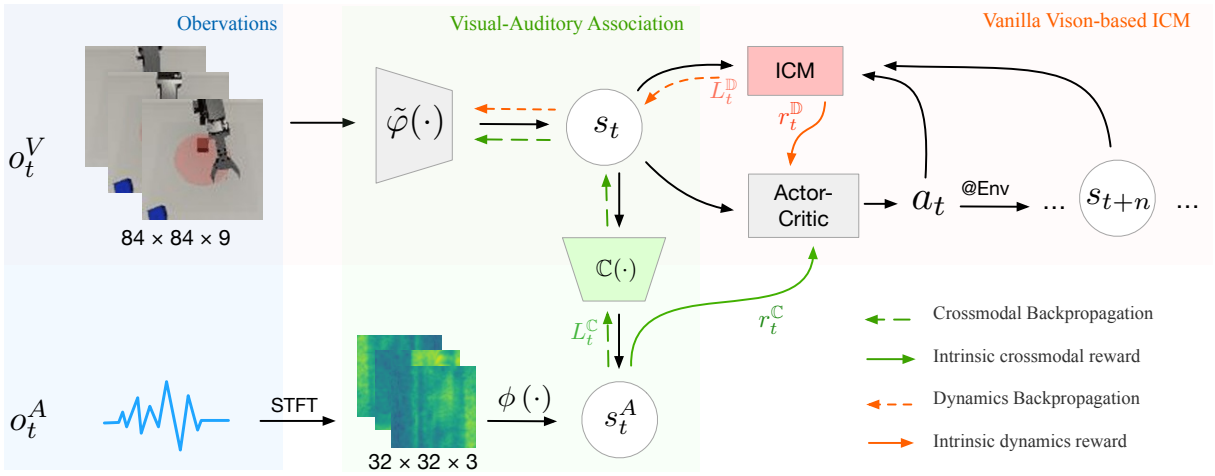

Fig. 2. An overview of the Intrinsic Sound Curiosity Module (ISCM) comprised of: 1) visual-auditory observations available in exploration (blue-shaded square), crossmodal learning (green-shaded square) and vanilla vision-based ICM architecture (red-shaded square).

The visual-auditory reward in our case is defined as $r_t^{\mathbb{C}} = \log(L_t^{\mathbb{C}} + \epsilon)$ — if the agent's assumption violates its perception, it will be encouraged to experience more, and vice versa. ϵ is a constant added to maintain numerical stability, particularly for values near zero. With $r_t^{\mathbb{D}} = \log(L_t^{\mathbb{D}} + \epsilon)$ as the ICM reward when modeling the environment dynamics, the overall intrinsic reward of ISCM is computed as

$$r_t = \lambda r_t^{\mathbb{C}} + (1 - \lambda) r_t^{\mathbb{D}}, \tag{4}$$

where λ controls the relative importance of crossmodal prediction and dynamics modeling for exploration.

C. Learning

The learning process is separated into 1) fully unsupervised pre-training and 2) task-specific fine-tuning stages with the curiosity mechanism omitted. It begins with an agent freely exploring an environment, trajectories of $\{o_t^V\}$ and $\{o_t^A\}$ are accumulated for representation learning; intrinsic rewards are computed for policy learning. When the freedom limit is reached or when the agent is believed to have enough knowledge, the pre-trained visual encoder will be fixed, and the actor-critic networks will be fine-tuned on downstream tasks with only vision and extrinsic sparse rewards accessible. Refer to Algorithm 1 for pseudo code.

IV. Experiments

A. Environments

The experiments are carried out in simulation because unsupervised exploration in the real world is costly which we leave for future work. One way to manipulate objects with authentic sound is to use a fixed data set with a physics computation interface [48]. For generality, we build our manipulation scenarios based on ThreeDWorld (TDW) [18], a novel embodied AI simulator [41] which is built upon the Unity game engine with multimodal capacities. To the best of our knowledge, it is the only one so far that supports physically simulated impact and scrape sounds [49], [50] at run time. The tabletop robot is composed of a 6-DoF OpenManipulator-Pro robotic arm and a 2-DoF gripper[‡]. It is allowed to manipulate cubes with diverse physical properties that are essential for both dynamics and sound characteristics, e.g., masses, materials, and bounciness.

Observations A camera and a single-channel microphone are placed above the table to capture observations. We focus more on vision and sound, so the robot's proprioception is not included, and the robot has no knowledge of the objects' coordinates.

Rewards One or several cubes are randomly placed inside a red circular area, and the goal is to push them out of the circle within a limited number of steps. Specifically, each step will have a penalty of -1/50, and an immediate reward of 1 will be delivered once the task is completed, otherwise the episode ends at 50 steps.

B. Implementations

We use the ICM implementation of URLB [7] as the baseline, and further extend it to our ISCM architecture[§].

1) Visual Observation: a) Raw RGB image observations (o_{t-2}^V, o_{t-1}^V, o_t^V) are stacked to the size of $84 \times 84 \times 9$ pixels. b) Four layers of CNN with ReLU activation are applied subsequently to encode vision to a latent state s_t. c) A model with two layers of fully connected neural networks with ReLU activation is constructed for sound prediction. d) Visual inputs are available in both pre-training and fine-tuning.

2) Auditory Observation: a) An auditory observation o_t^A is generated at run-time by a physical engine; it is then converted to the spectrogram o_t^S using Short-Time Fourier Transform (STFT). This is a consideration that complex sounds that come from objects with distinct materials are more distinguishable in the frequency domain with the help of the Fourier transform. Since the agent is updated with samples from a replay buffer and actions are chosen solely based on the visual input, there is no wait for the computation

[‡]https://github.com/ROBOTIS-GIT/open_manipulator_p
[§]https://github.com/xf-zhao/ISCM

Algorithm 1: Pseudo Code for ISCM Learning

Initialize: Replay buffer $\mathcal{D} \leftarrow \emptyset$, policy neural networks π, visual encoder $\tilde{\varphi}$, auditory encoder ϕ;
for $n = 1$ **to** $N_{pre-train}$ **do** ▷ Exploration
  Observe $o_t = \{o_t^V, o_t^A\}$;
  $s_t \leftarrow \tilde{\varphi}(o_t^V)$, $s_t^A \leftarrow \phi(o_t^A)$;
  $a_t \leftarrow \pi(s_t)$;
  Observe $o_{t+1} \sim \mathcal{P}_{ss'}^a$;
  $\mathcal{D} \leftarrow \mathcal{D} \cup (o_t, a_t, o_{t+1})$;
  Sample $\mathcal{D}_{batch}$ from $\mathcal{D}$;
  Update $\tilde{\varphi}$, π using samples in $\mathcal{D}_{batch}$ with Eq. 3 and Eq. 4;
end
Fix visual encoder $\tilde{\varphi}^* \leftarrow \tilde{\varphi}$ for evaluations;
Chose task T;
$\mathcal{D} \leftarrow \emptyset$;
for $n = 1$ **to** $N_{fine-tune}$ **do** ▷ Adaptation
  Observe o_t^V;
  $s_t \leftarrow \tilde{\varphi}^*(o_t^V)$;
  $a_t \leftarrow \pi(s_t)$;
  Observe $o_{t+1}, r \sim \mathcal{P}_{ss'}^a$;
  $\mathcal{D} \leftarrow \mathcal{D} \cup (o_t, a_t, r, o_{t+1})$;
  Sample $\mathcal{D}_{batch}$ from $\mathcal{D}$;
  Update π using samples in $\mathcal{D}_{batch}$ with extrinsic rewards;
end
Evaluate π with the accumulated rewards on task T for performance;

of STFT in real-time control. b) Spectrograms (o_{t-2}^S, o_{t-1}^S, o_t^S) are then stacked as the auditory input of $32 \times 32 \times 3$ size. c) Finally, s_t^A is obtained by applying a certain threshold for silence discrimination; and by passing through a fixed auditory encoder with 36-dimensional output for regression. Auditory inputs are available only in pre-training.

3) ICM Modeling: a) Trajectories of (s_t, a_t, s_{t+n}) are fed into the ICM dynamics models for both encoder training (Eq.2 with $\beta = 0.5$) and intrinsic reward $r_t^{\mathbb{D}}$ computation with $\epsilon = 1$. b) The sample with $r_t^{\mathbb{D}}$ is thus used to train a DDPG base learner. c) After enough explorations, the DDPG model will have to adapt to tasks with supervised rewards.

4) ISCM Modeling: a) Paired multimodal observations (o_t^V, o_t^A) are used to train the visual encoder (Eq.3 and Eq.1 with $\omega, \alpha, \beta = 100, 0.2, 0.5$) and to compute intrinsic crossmodal rewards $r_t^{\mathbb{C}}$. b) Overall intrinsic reward (Eq.4 with $\lambda = 0.8, \epsilon = 1$) is thus computed to train a DDPG-based learner.

All the mentioned neural networks are optimized by RAdam [51] with a learning rate equal to 0.001. For many unsupervised RL approaches, the performance decays with an excessive number of environment interactions [7]. There is so far no general strategy to determine when to early-stop explorations for better generality. We empirically choose 200K environment steps to pre-train and 30K steps to fine-tune, considering the convergence of learning curves. The result is averaged over 4 runs with different seeds.

C. Evaluation

The performance of unsupervised agents can be evaluated by means of measuring the adaptation process on downstream tasks or by statistically analyzing data diversity, e.g., counting of collisions [39], variance in the introduced sensory vector [38], or transformations (distance of movement, orientation changes) of objects. However, the latter method varies from task to task and is not always applicable.

Whereas the main focus of this work is to demonstrate the effectiveness of learned representations, the tasks are chosen to be simple to master for an agent. In this case, accumulated reward rather than success rate is more appropriate to compare the learning efficiency because the former can reflect the consumed steps, under the setting that the agent is punished for every unfinished step. Following Michael et al. [7], task rewards are solely used as the evaluation metrics: 1) as a measurement of active interactions, extrinsic rewards are accumulated but never leaked during pre-training; 2) the extrinsic rewards accumulated in the adaptation stage.

D. Results and Discussion

We observe that with sound involved, the agent is more interested in interacting with objects, resulting in more occasions of accidental completions (see Fig.3).

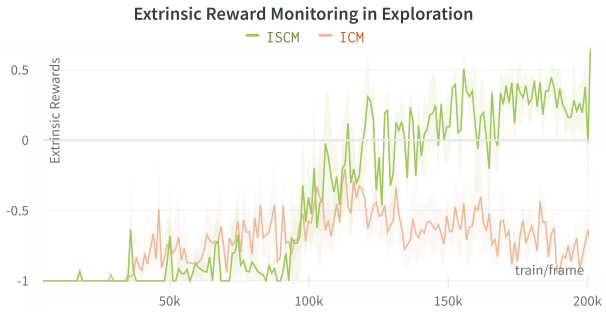

Fig. 3. Monitoring of extrinsic rewards (recorded but never used) in exploration. The ISCM agent has more chances of accidentally accumulating extrinsic rewards as a result of sound contributing to additional rewards.

Observations in the Unsupervised Reinforcement Learning Benchmark (URLB) [7] indicate that the learned representations are universally generalizable while the behavior policy maybe not, especially the policy learned with perfect states (full observable MDPs). As is shown in Fig.4, we reiterate that representations learned in unsupervised exploration are essential, and add further findings:

- There is a big performance gap between the DDPG learned from scratch (DDPG, dashed gray curve) and the other four with pre-trained weights (colored curves), which suggests that unsupervised exploration is helpful for faster adaptation to new tasks.
- The full pre-trained module (representations and behavior policy) with sound (ISCM, solid green curve) outperforms the baseline that solely depends on vision (ICM, solid orange curve).

- Without considering pre-trained policies, representations learned with a visual-auditory prediction (ISCM-PR, dashed green curve) outperform the ones learned with only vision (ICM-PR, dashed orange curve).
- Moreover, by comparing all solid with dashed curves, we find pre-trained policies to have positive effects on task adaptation, which reveals that skills acquired in unsupervised exploration are also reusable. However, more studies on policy analysis, e.g. decomposition of the learned policy for abstract behaviors are required for a clear view.

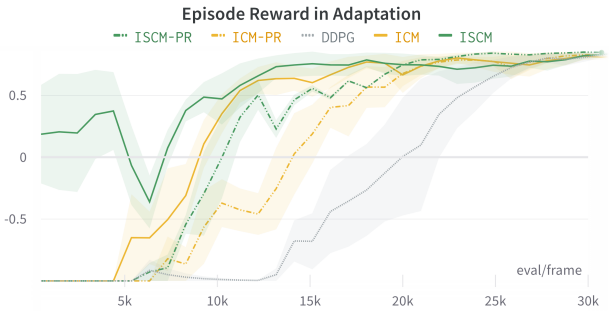

Fig. 4. Episode rewards in fine-tuning stage accumulated by DDPG learners with all hyper-parameters configured the same except for the initialization of models: 1) ICM: models with representations and policy pre-trained by ICM. 2) ICM-PR: models with ICM pre-trained representations but a re-initialized policy. 3) ISCM: models with representations and policy pre-trained by ISCM. 4) ISCM-PR: models with ISCM pre-trained representations but a re-initialized policy. 5) DDPG: models without pre-training.

A vision-to-sound regression model is also trained with all other hyper-parameters configured the same (see Fig.5 for a clear comparison). Though a vector (for regression) rather than a scalar (for discrimination) is believed to have a higher capacity of information, we find the discriminator setup (green curves) achieves a comparative performance with a regressor (red curves), while being simple to implement. Similar findings can be also found in recent works [39] where clustered auditory events are being predicted instead of regressing sound features. It may be a result of the following reasons: 1) impact sound presents not much more information than a deduction of event occurrence; 2) simulated sound is still far away from perfect such that vision, sound, and dynamics are not matched well as in reality. Future work will include constructions of more complex environments and sim-to-real adaptations to investigate more on these research questions.

V. Conclusions

Sound is one of the most common and efficient modalities, but is yet less considered to learn either simulated or real-world robotic manipulations. Unlike many of the curiosity-driven RL variants, especially the ones combined with audio that pay attention to non-robotics applications such as playing Atari games, we are focusing on investigating how robots can benefit from exploring multimodal environments. In this paper, the importance of unsupervised representation

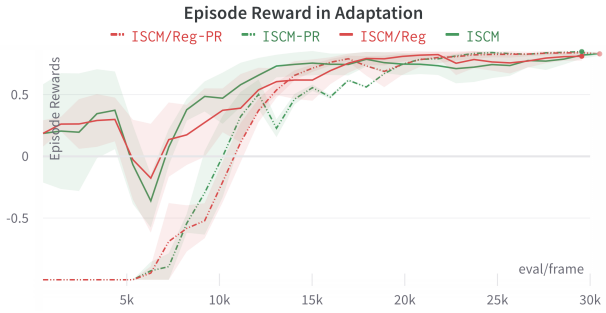

Fig. 5. Episode rewards in fine-tuning stage are accumulated by base DDPG learners that are initialized differently. 1) ISCM: fully pre-trained module with a discrimination auditory encoder. 2) ISCM/PR: pre-trained representations (but policy re-initialized) with a discrimination auditory encoder. 3) ISCM/Reg: fully pre-trained module with a regression auditory encoder. 4) ISCM/Reg-PR: pre-trained representations (but policy re-initialized) with a regression auditory encoder.

learning and of active exploration is addressed. We further propose the ISCM architecture to use physics-based sound as guidance regarding both aspects. Our experiments demonstrate that a sound-guided reinforcement learner is more active and has a great superiority to form sufficient as well as stable representations over vision-only baselines. In future work towards more applicable scenarios, we anticipate novel and interesting robot behaviors emerging in multimodal environments.

References

[1] P. J. Laurienti, R. A. Kraft, J. A. Maldjian, J. H. Burdette, and M. T. Wallace, "Semantic congruence is a critical factor in multisensory behavioral performance," *Experimental brain research*, vol. 158, no. 4, pp. 405–414, 2004.

[2] R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy." *Psychological bulletin*, vol. 88, no. 3, p. 638, 1980.

[3] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli," *Perception & Psychophysics*, vol. 69, no. 5, pp. 744–756, July 2007.

[4] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060.

[5] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 570–586.

[6] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 435–451.

[7] M. Laskin, D. Yarats, H. Liu, K. Lee, A. Zhan, K. Lu, C. Cang, L. Pinto, and P. Abbeel, "URLB: Unsupervised reinforcement learning benchmark," in *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[8] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, "Large-scale study of curiosity-driven learning," in *7th International Conference on Learning Representations*, May 2019, pp. 1–17.

[9] A. Eisermann, J. H. Lee, C. Weber, and S. Wermter, "Generalization in multimodal language learning from simulation," in *2021 International Joint Conference on Neural Networks*. IEEE, 2021, pp. 1–8.

[10] F. L. Higgen, P. Ruppel, M. Görner, M. Kerzel, N. Hendrich, J. Feldheim, S. Wermter, J. Zhang, and C. Gerloff, "Crossmodal pattern discrimination in humans and robots: A visuo-tactile case study," *Frontiers in Robotics and AI*, vol. 7, p. 540565, 2020.

[11] G. I. Parisi, P. Barros, D. Fu, S. Magg, H. Wu, X. Liu, and S. Wermter, "A neurorobotic experiment for crossmodal conflict resolution in complex environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 2330–2335.

[12] S. Clarke, T. Rhodes, C. G. Atkeson, and O. Kroemer, "Learning audio feedback for estimating amount and flow of granular material," in *Proceedings of the 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, Oct. 2018, pp. 529–550.

[13] G. Mir, M. Kerzel, E. Strahl, and S. Wermter, "A humanoid robot learning audiovisual classification by active exploration," in *2021 IEEE International Conference on Development and Learning*, 2021, pp. 1–6.

[14] H. Liang, C. Zhou, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang, "Robust robotic pouring using audition and haptics," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 10 880–10 887, 2020.

[15] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4651–4664.

[16] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artificial Intelligence*, vol. 299, p. 103535, 2021.

[18] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. D. Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, M. Sano, K. Kim, E. Wang, M. Lingelbach, A. Curtis, K. T. Feigelis, D. Bear, D. Gutfreund, D. D. Cox, A. Torralba, J. J. DiCarlo, J. B. Tenenbaum, J. Mcdermott, and D. L. Yamins, "ThreeDWorld: A platform for interactive multi-modal physical simulation," in *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[19] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *2015 IEEE International Conference on Computer Vision*, 2015, pp. 37–45.

[20] D. Jayaraman and K. Grauman, "Learning image representations tied to ego-motion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1413–1421.

[21] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2778–2787.

[22] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *European Conference on Computer Vision*. Springer, 2020, pp. 658–676.

[23] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[24] V. Dean, S. Tulsiani, and A. Gupta, "See, hear, explore: Curiosity via audio-visual association," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 14 961–14 972.

[25] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 801–816.

[26] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 631–648.

[27] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[28] A. Murali, Y. Li, D. Gandhi, and A. Gupta, "Learning to grasp without seeing," in *International Symposium on Experimental Robotics*. Springer, 2018, pp. 375–386.

[29] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *2020 IEEE International Conference on Robotics and Automation*. IEEE, 2020, pp. 9701–9707.

[30] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European Conference on Computer Vision*. Springer, 2020, pp. 17–36.

[31] 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*, 2019, pp. 8943–8950.

[32] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," *2018 IEEE International Conference on Robotics and Automation*, pp. 6292–6299, 2018.

[33] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, "Planning to explore via self-supervised world models," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, July 2020, pp. 8583–8592.

[34] Y. Hu, W. Wang, H. Jia, Y. Wang, Y. Chen, J. Hao, F. Wu, and C. Fan, "Learning to utilize shaping rewards: A new approach of reward shaping," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 931–15 941, 2020.

[35] D. Pathak, D. Gandhi, and A. Gupta, "Self-supervised exploration via disagreement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5062–5071.

[36] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *International Conference on Learning Representations*, 2019.

[37] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *International Conference on Learning Representations*, 2019.

[38] S. Rajeswar, C. Ibrahim, N. Surya, F. Golemo, D. Vazquez, A. Courville, and P. O. Pinheiro, "Haptics-based curiosity for sparse-reward tasks," in *5th Annual Conference on Robot Learning*, 2021.

[39] C. Gan, X. Chen, P. Isola, A. Torralba, and J. B. Tenenbaum, "Noisy agents: Self-supervised exploration by predicting auditory events," *CoRR*, vol. abs/2007.13729, 2020.

[40] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[41] J. Duan, S. Yu, T. Li, H. Zhu, and C. Tan, "A survey of embodied AI: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, pp. 230–244, 2022.

[42] Y. Du, C. Gan, and P. Isola, "Curious representation learning for embodied intelligence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 408–10 417.

[43] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2022.

[44] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning." in *International Conference on Learning Representations (Poster)*, 2016.

[45] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[46] L. Lu, Y. Shin, Y. Su, and G. Em Karniadakis, "Dying ReLU and initialization: Theory and numerical examples," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1671–1706, 2020.

[47] J. Peters, K. Mulling, and Y. Altun, "Relative entropy policy search," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[48] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations," in *5th Annual Conference on Robot Learning*, 2021.

[49] J. Traer, M. Cusimano, and J. H. McDermott, "A perceptually inspired generative model of rigid-body contact sounds," *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Sept. 2019.

[50] V. Agarwal, M. Cusimano, J. Traer, and J. H. McDermott, "Object-based synthesis of scraping and rolling sounds based on non-linear physical constraints," in *The 24th International Conference on Digital Audio Effects (DAFx-21)*, Sept. 2021.

[51] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proceedings of the Eighth International Conference on Learning Representations*, Apr. 2020.