Follow the Object: Curriculum Learning for Manipulation Tasks with Imagined Goals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning robot manipulation through deep reinforcement learning in environments with sparse rewards is a challenging task. In this paper we address this problem by introducing a notion of imaginary object goals. For a given manipulation task, the object of interest is first trained to reach a desired target position on its own, without being manipulated, through physically realistic simulations. The object policy is then leveraged to build a predictive model of plausible object trajectories providing the robot with a curriculum of incrementally more difficult object goals to reach during training. The proposed algorithm, Follow the Object (FO), has been evaluated on 7 MuJoCo environments requiring increasing degree of exploration, and has achieved higher success rates compared to alternative algorithms. In particularly challenging learning scenarios, e.g. where the object's initial and target positions are far apart, our approach can still learn a policy whereas competing methods currently fail.

1 Introduction

Reinforcement Learning (RL) aims to solve sequential decision making problems where the decision maker seeks to find the optimal actions that maximise a feedback signal [1]. Robotic problems are characterised by the need to make sequential decisions, and recent advances in this domain have been made through deep RL. Typical applications involve locomotion [2–4] and manipulation tasks [5, 6], such as grasping [7, 8], stacking [9] and dexterous hand manipulation [10, 11]. Since the large majority of today's RL algorithms are still sample inefficient, learning from trial and error in the real world is often unfeasible, and it is commonplace to train RL policies for continuous control using a simulated environment, e.g. through a physics engine such as MuJoCo [12].

Traditional RL approaches for robotics have required manually designed and problem-specific reward functions providing a smoothly varying feedback signal for every visited state [13, 14, 8, 6]. Recently, there has been increasing evidence that similar or even superior policies for robotics tasks can be obtained using sparse rewards, i.e. when non-zero rewards are given to the agent only when the task is successfully completed [15]. Although exploration becomes much harder in this setting, one particular approach - Hindsight Experience Replay (HER) [16] - has been proven to enable learning.

Despite such progress, contradicting claims have been reported in the literature about the empirical performance of HER even for simple robotics tasks. For instance, in *PickAndPlace*, HER has been found to fail in some cases [9] or to succeed in other cases, provided that some of the training episodes are started from 'easy' states, e.g. when the object has already been grasped by the robot [16]. Other studies have indicated that initial grasping is not a prerequisite to success as long as the object's target position is 'easy', e.g. the target position is on the table in at least some training episodes [17]. Overall, the currently available experimental evidence confirms what intuition alone would suggest:

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

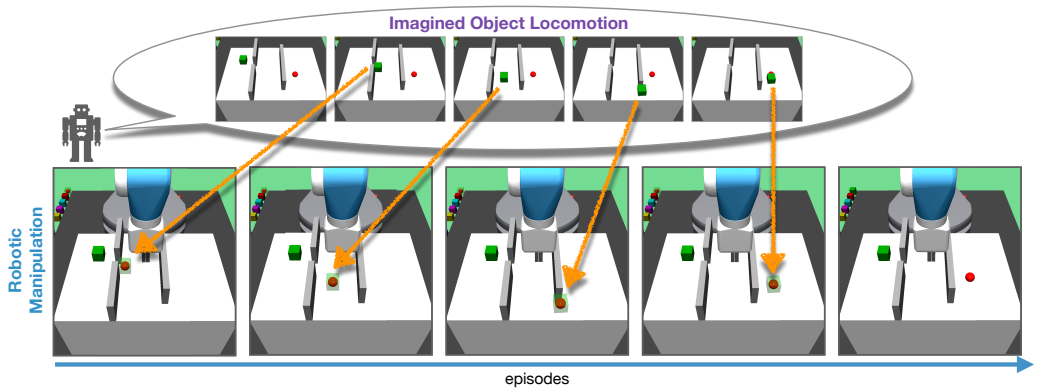

Figure 1: An *object locomotion policy* allows an object (green block) to reach its final target position (red dot). This policy is used to generate trajectories of intermediary goals: imagined object positions closer to the object's initial position are easier to achieve and those closer to the original target are more difficult. The robot's learning procedure follows a curriculum: it starts with the simplest imagined goal, and progressively proceeds to a more difficult one once the robot has mastered the current difficulty level. Ultimately, the robot is able to solve the original task with only sparse rewards.

starting from 'easy' configurations generally leads to successful roll-outs whereas more 'difficult' initial configuration may significantly hinder the learning ability [9].

Inspired by these findings, we set out to investigate mechanisms to further improve upon the performance of HER on manipulation tasks that are currently difficult to learn, e.g. when the object's target position is far apart from its starting position. These situations require a significantly higher degree of exploration which makes learning with only sparse rewards particularly challenging. The approach we propose stems from recognising the importance of learning over a curriculum of progressively more difficult goals. Curriculum learning is based upon the principle that mastering simple skills first can help master harder ones later [18]. However, how to implement it on manipulation tasks with only sparse reward signals is not obvious.

The approach proposed here relies on a mechanism to imagine object positions that are progressively more difficult for the object to reach during training. These intermediary goals guide the progression of mastering simple skills first (e.g. pushing an object to a nearby target) followed by more difficult ones. Our developments in this direction are inspired by the notion of *object locomotion policies* [19], where the objects involved in a manipulation task are modelled as independent agents that must learn a locomotion policy (i.e. how to move from any initial position to a final one). The authors used the object policies to define auxiliary rewards for the main robot manipulation task at hand.

We initially follow a similar approach as in [19], and obtain an object locomotion policy for the given task. Learning such policies successfully with only sparse rewards can be accomplished using a physics engine such as MuJoCo where objects are allowed to move freely in space, on their own, without being manipulated. Our key contribution here is to leverage the object policy to implement a curriculum learning strategy for the manipulation task. As the robot learns the task, imagined object positions are produced, each one conditional on the object's initial and target position. These imagined positions act as intermediary goals of increasing complexity. Since this complexity can be controlled precisely, the resulting learning strategy facilitates mastering simple tasks first followed by more difficult ones, and results in higher success rate overall. The proposed approach, Follow the Object (FO), is illustrated in Figure 1 and described in Section 4. Its performance has been tested on 7 MuJoCo environments using a Fetch robotic arm with different degrees of task complexity and different starting configurations. Our empirical results and comparisons to related algorithms are presented in Section 5, and indicate that FO is able to learn in particularly challenging scenarios whereas competing methods currently fail.

2 Related Work

Imitation learning. Imitation Learning (IL) approaches have been used to facilitate reinforcement learning of robotic manipulation tasks with sparse rewards [15, 20, 9]. IL casts the sequential decision

making problem as a supervised learning problem requiring either pre-collected observational data [21, 22] or an interactive demonstrator [23, 24]. Demonstrations can also be used to learn a reward function [25–27]. In general, collecting demonstrations for robotics tasks can be time consuming and requires a dedicated data collection setup, e.g. virtual reality or tele-operation facility.

Goal conditional policies. Recently, the advantages of goal-conditioned policies for robotics problems have been discussed [28]. Different approaches have investigated how to automatically generate such goals. For instance, [29] follow a self-play approach on reversible or resettable environments, [30] employ adversarial training for robotic locomotion tasks, [31] use variational autoencoders for visual robotics tasks and [16] propose an experience sampling approach, Hindsight Experience Replay (HER), to randomly draw synthetic goals from previously experienced trajectories. HER in itself may be seen as a form of implicit curriculum learning. The effectiveness of this approach has been demonstrated on a suite of challenging continuous control environments [17].

Curiosity. In these approaches, the notion of *novelty* of a state provides the RL agent with an additional reward and motivation for exploring less novel states [32, 33]. Curiosity has been defined in many ways, e.g. as the error in predicting the RL agent's actions by an inverse dynamics model [34], as the error in predicting the output of a fixed randomly initialised neural network [35], or by evaluating the novelty of a recently visited state according to its reachability from those stored in a memory [36]. The large majority of these works describe gaming or maze environments with discrete action spaces, but a locomotion task with continuous actions has also been considered [36].

Curriculum learning. Curriculum learning has been widely adopted into the RL framework and many researchers have sought to find a solution to the problem of how to automatically generate a way to order the goals according to difficulty. For example, [37] and [38] propose to automatically adjust the difficulty of the gaming environment, [39] use the policies of the intermediate agents obtained during the training, [40] evaluate the occurrence frequencies of pre-defined events. In robotics, [41] propose to schedule the initial positions, [42] create easier sub-goals spanning over one dimension of the target position, and [43] select experiences adaptively according to diversity-based curiosity.

Object motion prediction. A body of works exists on predicting how objects behave under manipulative actions. For example, [44] learn a probabilistic model using a real system where a robotic arm applies random pushes to various objects, and [45] propose a simulation-based framework. A combination of these methods has also been proposed [46]. Some efforts have also been made to model the dynamics of complex scenes involving non-rigid bodies, e.g. using a particle-based simulator [47]. Similarly, the object locomotion policies we use here can predict the most likely trajectory that an object has to travel to reach a given target. Unlike these works, our policies are learned irrespective of the manipulative actions exercised by the robot. Advantages and disadvantages of this approach are discussed in Section 5.3.

3 Problem Definition and Background

3.1 Multi-goal RL for Robotic Manipulation

We are concerned with solving a manipulation task: an object is presented to the robot, and has to be manipulated so as to reach a target position. In the tasks we consider, the target goal is specified by the object location and orientation, and the robot is rewarded only when it reaches its goal. We model the robot's sequential decision process as a Markov Decision Process (MDP) defined by a tuple, $M = \langle \mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{G}$ is the set of goals, $\mathcal{A}$ is the set of actions, $\mathcal{T}$ is the state transition function, $\mathcal{R}$ is the reward function and $\gamma \in [0, 1)$ is the discounting factor. At the beginning of an episode, the environment samples a goal $g \in \mathcal{G}$. The position of the object at time t is denoted by o_t and the state of the environment is $s_t \in \mathcal{S}$, which includes o_t. We assume that, given s_t, we can recover o_t through a known mapping, i.e. $o_t = m(s_t)$. A robot's action is controlled by a deterministic policy, i.e. $a_t = \mu_\theta(s_t, g) : \mathcal{S} \times \mathcal{G} \rightarrow \mathcal{A}$, parameterised by θ. The environment moves to its next state through its state transition function, i.e. $s_{t+1} = \mathcal{T}(s_t, a_t) : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, and provides an immediate and sparse reward r_t, defined as

$$r_t = \mathcal{R}(o_{t+1}, g) = \begin{cases} 0, & \text{if } ||o_{t+1} - g||_2 \leq \epsilon \\ -1, & \text{otherwise} \end{cases} \tag{1}$$

where ϵ is a pre-defined threshold. Following its policy, the robot interacts with the environment until the episode terminates after T steps. The interaction between the robot and the environment

generates a trajectory, $\tau = (g, s_1, a_1, r_1, \ldots, s_T, a_T, r_T, s_{T+1})$. The ultimate learning objective is to find the optimal policy that maximises the expected sum of the discounted rewards over the time horizon T, i.e.

$$J(\mu_\theta) = \mathbb{E}_{\tau \sim \mathbb{P}(\tau|\mu_\theta)}[R(\tau) = \sum_{i=1}^{T} \gamma^{i-1} r_i] \tag{2}$$

where γ is the discount factor.

3.2 Deep Deterministic Policy Gradient Algorithm

Deep Deterministic Policy Gradient (DDPG) [2] is adopted here as our main learning algorithm, however any other off-policy algorithm that operates on continuous action domains could be equally used. DDPG integrates deterministic policy functions [48] with non-linear function approximators such as deep neural networks, and maintains a policy (actor) network $\mu_\theta(s_t, g)$ and an action-value (critic) network $Q^\mu(s_t, a_t, g)$. Further details about our implementation are provided in the Appendix.

3.3 Hindsight Experience Replay

Suppose we are given an observed trajectory, $\tau = (g, s_1, a_1, r_1, \ldots, s_T, a_T, r_T, s_{T+1})$. Since o_t can be obtained from s_t using a fixed and known mapping, the path that was followed by the object during the trajectory, i.e. $o_1, \ldots, o_{T+1}$, can be easily extracted. HER samples a new goal from this path, i.e. $\tilde{g} \sim \{o_1, \ldots, o_T\}$, and the rewards are recomputed with respect to $\tilde{g}$, i.e. $\tilde{r}_t = \mathcal{R}(o_{t+1}, \tilde{g})$. Using these rewards and $\tilde{g}$, a new trajectory is created implicitly, i.e. $\tilde{\tau} = (\tilde{g}, s_1, a_1, \tilde{r}_1, \ldots, s_T, a_T, \tilde{r}_T, s_{T+1})$. These HER trajectories $\tilde{\tau}$ are used to train the policy parameters together with the original trajectories.

4 Methodology

In this section, we explain the steps involved in our proposed procedure, i.e. (a) how object locomotion policies are learned, (b) how the novel curriculum learning model is defined upon them, and (c) how the curriculum is implemented for robotic manipulation.

4.1 Learning Object Locomotion Policies

The object involved in the manipulation task is initially modelled as an agent capable of independent decision making abilities, and its decision process is modelled by a separate MDP defined by a tuple $L = \langle \mathcal{Z}, \mathcal{G}, \mathcal{U}, \mathcal{Y}, \mathcal{R}, \gamma \rangle$. Here, $\mathcal{Z}$ is the set of states, $\mathcal{G}$ is the set of goals, $\mathcal{U}$ is the set of actions, $\mathcal{Y}$ is the state transition function, $\mathcal{R}$ is the reward function and $\gamma \in [0, 1)$ is the discounting factor. The same goal space, $\mathcal{G}$, is used as in the robotic manipulation. $z_t \in \mathcal{Z}$ is a reduced version of s_t that only involves object-related features including the position of the object, i.e. $o_t \subset z_t$. The object's action space explicitly controls the pose of the object, and these actions are controlled by a deterministic policy, i.e. $u_t = \nu_\theta(z_t, g) : \mathcal{Z} \times \mathcal{G} \rightarrow \mathcal{U}$. The state transition is now defined on different spaces than robotic manipulation, i.e. $\mathcal{Y} : \mathcal{Z} \times \mathcal{U} \rightarrow \mathcal{Z}$; however, the same sparse reward function is used here as before. Figure 2a illustrates the training procedure used in this context and based on DDPG with HER. The optimal object policy ν_θ maximises the expected return $J(\nu_\theta) = \mathbb{E}_{g, z_t \sim \mathcal{D}}[\sum_{i=1}^{T} \gamma^{i-1} r_i]$ where $\mathcal{D}$ denotes the replay buffer containing the trajectories, indicated by η, obtained by ν_θ throughout training.

4.2 Learning to Imagine Goals

Upon training, the object locomotion policy ν_θ can be used to generate trajectories, i.e. $(g, z_1, u_1, r_1, \ldots, z_T, u_T, r_T, z_{T+1}) \sim \mathbb{P}(\eta|\nu_\theta)$. Since $o_t \subset z_t$, an object path originating at o_1 and moving all the way to the final target g, i.e. $(o_1, \ldots, o_{T+1})$, can be obtained from $(z_1, \ldots, z_{T+1})$. We use these trajectories to train a model, $\mathcal{F}_\phi$, to predict the object's position k-steps ahead within a trajectory. The model takes o_1 and g together with an integer $k \in [1, T]$ as input, and outputs the prediction for the object position at time-step $k+1$, i.e. $\hat{o}_{k+1} = \mathcal{F}_\phi(o_1, g, k)$. The model parameters

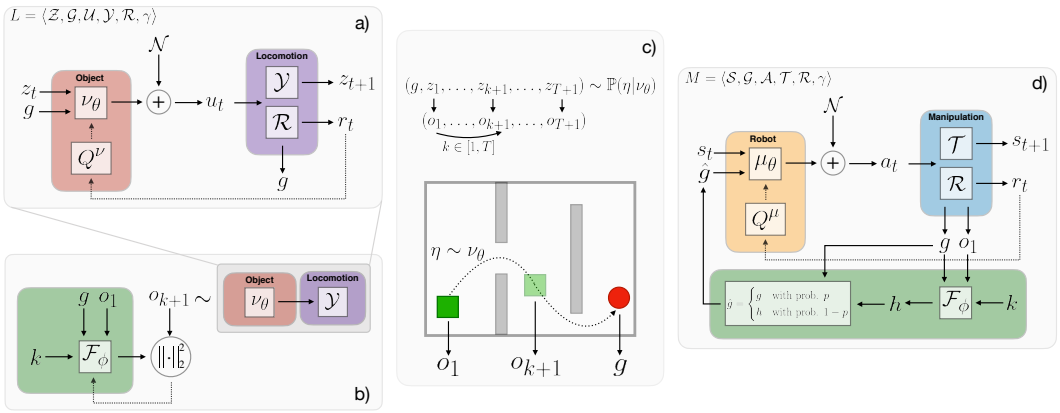

Figure 2: a) Learning the object policy on locomotion. b) Adopting object policy to generate trajectories for training of $\mathcal{F}_\phi$. c) A visual example for training of $\mathcal{F}_\phi$: the green coloured object whose initial position denoted by o_1 needs to pass the obstacles and eventually reach the target position g in red. ν_θ is used to obtain the trajectory moving the object from o_1 to g. Using these trajectories, we train $\mathcal{F}_\phi$ to predict the object position at time-step $k+1$ for given o_1 and g. d) $\mathcal{F}_\phi$ is used to generate a curriculum over object positions on robotic manipulation. The difficulty of h is adjusted through k starting from 1 and reaching T as the robot performs better on imagined goals.

ϕ are found such that (see also Figure 2b):

$$\arg\min_{\phi} \mathbb{E}_{\substack{(g,o_1,\ldots,o_T)\sim\mathbb{P}(\eta|\nu_\theta)\\ \forall k\in[1,T]}} \left[\left|\left| o_{k+1} - \mathcal{F}_\phi(o_1, g, k) \right|\right|_2^2 \right] \tag{3}$$

4.3 Learning the Robot Policy for Manipulation with Imagined Goals

In this final step, the predictive model $\mathcal{F}_\phi$ is leveraged to generate a sequence of imagined goals that can facilitate the learning of the original manipulation policy. For every new episode characterised by o_1 and g, a predicted object position at time-step $k+1$, i.e. $h = \mathcal{F}_\phi(o_1, g, k)$, is used as an intermediate goal to train the robot. By adjusting k throughout the training, a curriculum is formed.

Adjusting the curriculum difficulty: The value of k controls the difficulty of an imagined goal h. Starting from $k = 1$ and all the way up to $k = T$, one could simply increase k by one every time the robot masters the current level. However, we have observed that this simple strategy can underperform due to some form of forgetting; this happens when the robot fails to master the current level, and the replay buffer starts to be filled with failing trajectories. Instead, k is sampled from a uniform distribution, $U(1, k_{max})$, and k_{max} is increased upon reaching fluency at the current difficulty level. This remedy resolves the forgetting issue since training at a given difficulty level now allows for goals at lower levels to be randomly introduced; this forces the robot to occasionally 'practice' on simpler problems rather than always being confronted with more difficult ones. In order to decide whether the robot is fluent enough with the current difficulty level, we evaluate its performance for $k = k_{max}$, i.e. $h_{max} = \mathcal{F}_\phi(o_1, g, k_{max})$. When the robot achieves a success rate higher than some threshold (e.g. 0.2) for a given h_{max}, we increase k_{max} by one, and keep increasing it until T is reached. This scheme is somewhat inspired by the *boundary-sampling* idea used in the Automatic Domain Randomization (ADR) algorithm [49].

Learning the Robot Policy: The policy μ_θ is learned by DDPG as in Section 3.2. We ensure that both original goals g and imagined goals h are used during training. That is, goal $\hat{g}$ where $\hat{g} = g$ with probability p, and otherwise $\hat{g} = h$. Introducing g early on in the training allows for faster learning as the policy may become capable to accomplish the original goals before the curriculum provided through h is completed at $k = T$. Using a mixture of g and h, the robot interacts with the environment and obtains trajectories $\tau = (\hat{g}, s_1, a_1, \hat{r}_1, \ldots, s_T, a_T, \hat{r}_T, s_{T+1})$, where rewards are calculated with respect to $\hat{g}$, i.e. $\hat{r}_t = \mathcal{R}(o_t, \hat{g})$. Experienced trajectories are stored into the replay buffer $\mathcal{D}$. The parameters of μ_θ are updated to maximise Eq. (2) using the trajectories sampled from $\mathcal{D}$. Figure 2d illustrates this procedure, and Algorithm 1 provides the pseudo code.

Algorithm 1: Learning manipulation policy with imagined goals

Given : Robotic Manipulation MDP $M = \langle \mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, Imagined goal generator $\mathcal{F}_\phi(o_1, g, k)$, Random process $\mathcal{N}$ for exploration, Fixed and known mapping function $m : \mathcal{S} \rightarrow \mathcal{O}$

Initialise : Parameters θ for robot policy μ_θ, Experience replay buffer $\mathcal{D}$

$k_{max} = 2$

for *episode* $= 1$ ***to*** N **do**

  Receive initial state s_1 and g, $o_1 = m(s_1)$

  Sample $k \sim U(1, k_{max})$ where U is uniform distribution

  $h = \mathcal{F}_\phi(o_1, g, k)$

  $\hat{g} = \begin{cases} g & \text{with prob. } p \\ h & \text{with prob. } 1-p \end{cases}$

  for $t = 1, T$ **do**

    Sample an action: $a_t = \mu_\theta(s_t, \hat{g}) + \mathcal{N}$

    Execute the action: $s_{t+1} = \mathcal{T}(s_t, a_t)$ and $r_t = \mathcal{R}(o_{t+1}, \hat{g})$

  Store $\tau = (\hat{g}, s_1, a_1, \hat{r}_1, \ldots, s_T, a_T, \hat{r}_T, s_{T+1})$ in memory buffer $\mathcal{D}$

  Update μ_θ with trajectories drawn from $\mathcal{D}$

  if $k_{max} \leq T$ **then**

    Test μ_θ for $h_{max} = \mathcal{F}_\phi(o_1, g, k_{max})$

    if $success \geq threshold$ **then**

      $k_{max} = k_{max} + 1$

5 Experiments

5.1 Environments

We evaluate our method on 7 simulated MuJoCo environments that use a 7-DoF Fetch robotics arm. We build these environments upon the *Push* and *PickAndPlace* environments from [17]. As visualised in Figure 3, the environments differ in terms of initial and target object positions, as well as the obstacles between these two. In all the environments, the target indicates the desired 3D position of the object. A reward of 0 is provided to the agent if the object is within 5-cm range to the target, and -1 otherwise. Robot actions are 4-dimensional: 3D to specify the desired arm movement in Cartesian coordinates and 1D to control the opening of the gripper. In *Push* tasks, the robot is not allowed to control the opening of the gripper to prevent grasping. The observations include the positions and linear velocities of the robot arm and the gripper, as well as the object's position, rotation and angular velocity, and its relative position and linear velocity to the gripper. An episode terminates after 50 time-steps, except for *Push-DoubleObstacles* which terminates after 80 time-steps. In the auxiliary object locomotion tasks, on the other hand, observations include the object's position, rotation and angular velocity. The object actions are 7 dimensional (3D for translation and 4D for rotation). In practice, we define the object as a *mocap* MuJoCo entity which enables us to control its 7D pose. In all cases, the object actions correspond to the desired relative change in the object pose between two consecutive time-steps; however, their realisation depends on the dynamics executed by the simulator. The same reward function and termination criteria are used as in manipulation.

5.2 Implementation and Training Process

We use three-layer neural networks with ReLU activations for all models, and optimise them using Adam [50]. We adopt the optimised hyperparameter values for HER (the baseline approach) from

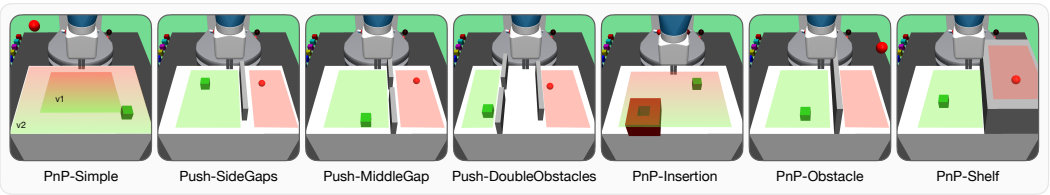

Figure 3: Illustrations of the environments. The object is coloured in green and the target position is visualised as a red sphere. The green and red shaded areas indicate possible initial and target object positions.

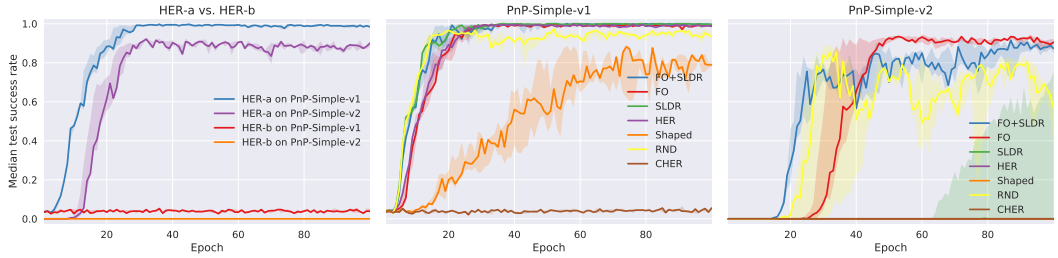

Figure 4: a) The performances of HER-a and HER-b on both *PnP-Simple* versions. b) The performances of all baselines on PnP-Simple-v1. c) The performances of all baselines on PnP-Simple-v2.

[17], and use them for all models. We train the models on a GPU enabled machine using its CPU cores to generate experiences and GPU to optimise the parameters. At each epoch of the training, we generate $38 * 50 = 1900$ full episodes and update the parameters 40 times using batches of size 4864. We observed that the performance of the proposed approach is robust with respect to the model specific hyperparameters and report the results by choosing $p = 0.2$ and a success threshold of 0.25. An ablation study on the effects of these two hyperparameters can be found in Appendix, where we also describe all the remaining hyperparameters in greater detail.

5.3 Comparison and Performance Evaluation

We consider 6 alternative methods for comparison. All methods are built upon DDPG [2] and implement HER [16] as standard. Differences are as follows: *HER* is the baseline as described in [16] using sparse rewards. *Shaped* uses distance-based shaped rewards instead of sparse rewards. *SLDR* uses auxiliary rewards introduced by [19] together with sparse rewards. In *SLDR*, the optimal object locomotion actions are compared with those caused by the robot in terms of the action-values, and the difference is introduced as an auxiliary reward. *RND* uses curiosity-based exploration bonuses together with sparse rewards. Exploration bonuses are obtained by adopting Random Network Distillation (RND) proposed in [35]. Curriculum-guided HER (*CHER*) [43] brings adaptive experience selection upon baseline HER. *FO* refers to the approach introduced in this paper. We also combine our approach with SLDR and refer this combination as *FO+SLDR*. Following [17], we evaluate the performances after each training epoch by computing the test success rate averaged across 380 deterministic rollouts with original goals g. In all cases, we repeat an experiment with 5 different random seeds and report results by computing the median test success rate with the interquartile range.

Demonstrating the Problem on Simple *PickAndPlace*: To investigate how initial and target object positions affect learning, we define 2 different versions of *PickAndPlace* (*PnP*). *PnP-Simple-v1* corresponds to the version mentioned in [17]. In *PnP-Simple-v2*, we simply remove two constraints that condition the initial and target object positions, i.e. unlike [17] we do not require the positions to be sampled from within a smaller square centred around the gripper, and we do not enforce the targets to be on the table level in some of the training episodes. We adopt the baseline HER from [16] to train two policies, i.e. *HER-a* on *PnP-Simple-v1* and *HER-b* on *PnP-Simple-v2*, following identical learning procedures. Throughout the training, we test their performances on both versions as presented in Figure 4. *HER-a* achieves *PnP-Simple-v1* with 1.0 success rate, and *PnP-Simple-v2* with 0.9. This result indicates that the basic characteristics of the required robotic behaviours are similar in both versions, and *PnP-Simple-v2* is not drastically harder to perform than *PnP-Simple-v1*. However, *HER-b* fails on both versions suggesting that *PnP-Simple-v2* is harder to learn for HER. This is explained by the fact that easy-to-solve episodes are seen less frequently in *PnP-Simple-v2* than in *PnP-Simple-v1*. Figure 4 also demonstrates the performances of all 7 approaches on both versions. It can be noted that our approach is robust to such differences in the environments at a small cost of a slightly slower learning rate. We observe that this trade-off is generally compensated when combining *FO* with *SLDR*, as reported in the following section.

Performance Evaluation on Other Environments: Before evaluating the performances, it is worth noting a difference between *Push* and *PnP* tasks in terms of exploration requirements. *Push* variations only require horizontal exploration in the space of object positions. Even a random robot policy may change the object position in the x- and y- axes, and contribute to horizontal exploration. *PnP*

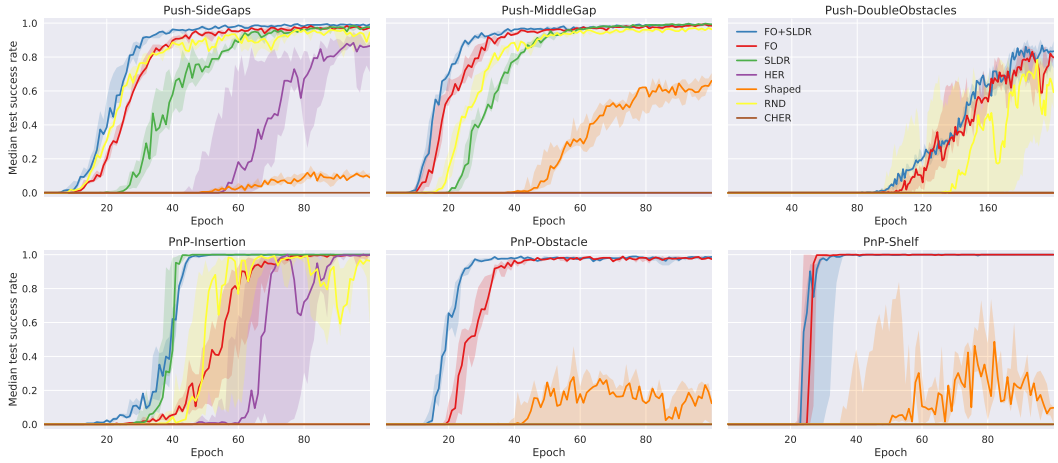

Figure 5: The learning curves of 7 approaches on 6 environments. *FO* can learn all 6 manipulation tasks with greater exploration demands. When combined with SLDR, *FO+SLDR* results in faster learning compared to *FO* alone.

variations, on the other hand, additionally require vertical exploration, which becomes possible only after the robot has learned to grasp the object. In general, the *PnP* variations are associated with greater exploration demands. However, *PnP-Insertion* proves to be easier than *PnP-Obstacle* and *PnP-Shelf*; unlike the other 5 environments, in this task the initial and target object positions are sampled from intersecting sets (see Figure 3 for a visualisation.)

The *HER* baseline fails in most cases as the exploration becomes harder. Interestingly, in some cases, it is outperformed by *Shaped*, which uses distance-based rewards as the only difference. *RND* demonstrates that curiosity-based exploration bonuses can improve the performance beyond *HER* as the exploration demand increases. However, it fails in more difficult cases such as *PnP-Obstacle* and *PnP-Shelf*. One can also notice the fluctuations in the learning curves of *RND*. In terms of curiosity, the robot here would be rewarded when, for example, it throws the object out of the table for exploration. However, these rewards can jeopardise the learning process towards the main task. Although *SLDR*'s manipulation-specific auxiliary rewards may seem as providing an advantage over *RND*, *SLDR* can also fail in more difficult environments, and in general its performance is poorer than *RND*. We also observe that *CHER* fails to learn our manipulation tasks. Overall, *FO* outperforms these 5 approaches and achieves encouraging learning rates in all 6 environments. Moreover, the combined approach *FO+SLDR* results in faster learning and has been proved to be the best approach overall on these environments. This is explained by the fact that both approaches, individually, motivate the robot to follow the object policy.

A potential limitation of the proposed approach is that not all the trajectories generated by the object locomotion policy are always guaranteed to be attainable by the robot. For example, in *Push* tasks, occasionally the simulated object trajectories may pass over the obstacles, which clearly are not feasible for the robot. Our investigations have indicated that including unattainable imaginary goals do not hinder the performance of the proposed methodology in any significant way, and further discussions of these findings can be found in the Appendix.

6 Conclusions

We have proposed *Follow the Object* (*FO*), a new curriculum generation framework for learning robotic manipulation tasks with only sparse rewards. Using an object locomotion policy, a predictive model provides a curriculum of intermediary imagined goals resulting in high learning success rate. The object policies can be learned easily and realistically with a physics simulation engine. We have demonstrated that *FO* is particularly beneficial in difficult manipulation problems requiring substantial exploration and can improve upon the performance obtained by existing algorithms. In future work, this framework can be extended to include more complex objects (e.g. flexible objects) and manipulation tasks, and account for uncertainty in the imagined goals, e.g. by using a distribution of predicted goals.

References

[1] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press, 1998.

[2] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

[3] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[4] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

[5] Justin Fu, Sergey Levine, and Pieter Abbeel. One-shot learning of manipulation skills with online dynamics adaptation and neural network priors. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4019–4026. IEEE, 2016.

[6] Shixiang Gu, Ethan Holly, Timothy P. Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *International Conference on Robotics and Automation*, pages 3389–3396, 2017.

[7] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[8] Ivaylo Popov, Nicolas Heess, Timothy P. Lillicrap, Roland Hafner, Gabriel Barth-Maron, Matej Vecerik, Thomas Lampe, Yuval Tassa, Tom Erez, and Martin A. Riedmiller. Data-efficient deep reinforcement learning for dexterous manipulation. *CoRR*, abs/1704.03073, 2017.

[9] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *International Conference on Robotics and Automation*, pages 6292–6299, 2018.

[10] Henry Zhu, Abhishek Gupta, Aravind Rajeswaran, Sergey Levine, and Vikash Kumar. Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost. *arXiv preprint arXiv:1810.06045*, 2018.

[11] Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*, 2018.

[12] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[13] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Sixteenth International Conference on Machine Learning*, pages 278–287, 1999.

[14] Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *International Conference on Machine Learning*, pages 463–471, 1998.

[15] Matej Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin A. Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *CoRR*, abs/1707.08817, 2017.

[16] Marcin Andrychowicz, Dwight Crow, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5055–5065, 2017.

[17] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *CoRR*, abs/1802.09464, 2018.

[18] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48, 2009.

[19] Ozsel Kilinc and Giovanni Montana. Reinforcement learning for robotic manipulation using simulated locomotion demonstrations. *arXiv preprint arXiv:1910.07294*, 2019.

[20] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John Agapiou, Joel Z. Leibo, and Audrunas Gruslys. Deep q-learning from demonstrations. In *AAAI Conference on Artificial Intelligence*, pages 3223–3230, 2018.

[21] Dean Pomerleau. ALVINN: an autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, pages 305–313, 1988.

[22] Yan Duan, Marcin Andrychowicz, Bradly C. Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. In *Advances in Neural Information Processing Systems*, pages 1087–1098, 2017.

[23] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.

[24] Nathan D. Ratliff, James A. Bagnell, and Siddhartha S. Srinivasa. Imitation learning for locomotion and manipulation. In *International Conference on Humanoid Robots*, pages 392–397, 2007.

[25] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pages 663–670, 2000.

[26] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pages 49–58, 2016.

[27] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.

[28] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1312–1320, 2015.

[29] Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. In *International Conference on Learning Representations*, 2018.

[30] Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International Conference on Machine Learning*, pages 1514–1523, 2018.

[31] Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, pages 9209–9220, 2018.

[32] Jürgen Schmidhuber. Curious model-building control systems. In *IEEE International Joint Conference on Neural Networks*, pages 1458–1463. IEEE, 1991.

[33] 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.

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

[35] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[36] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018.

[37] Yuandong Tian, Qucheng Gong, Wenling Shang, Yuxin Wu, and C. Lawrence Zitnick. ELF: an extensive, lightweight and flexible research platform for real-time strategy games. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2659–2669, 2017.

[38] Yuxin Wu and Yuandong Tian. Training agent for first-person shooter game with actor-critic curriculum learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[39] Wojciech Marian Czarnecki, Siddhant M. Jayakumar, Max Jaderberg, Leonard Hasenclever, Yee Whye Teh, Nicolas Heess, Simon Osindero, and Razvan Pascanu. Mix & match agent curricula for reinforcement learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1095–1103. PMLR, 2018.

[40] Niels Justesen and Sebastian Risi. Automated curriculum learning by rewarding temporally rare events. In *2018 IEEE Conference on Computational Intelligence and Games, CIG 2018, Maastricht, The Netherlands, August 14-17, 2018*, pages 1–8. IEEE, 2018.

[41] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pages 482–495. PMLR, 2017.

[42] Manfred Eppe, Sven Magg, and Stefan Wermter. Curriculum goal masking for continuous deep reinforcement learning. In *Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2019, Oslo, Norway, August 19-22, 2019*, pages 183–188. IEEE, 2019. doi: 10.1109/DEVLRN.2019.8850721. URL `https://doi.org/10.1109/DEVLRN.2019.8850721`.

[43] Meng Fang, Tianyi Zhou, Yali Du, Lei Han, and Zhengyou Zhang. Curriculum-guided hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 12602–12613, 2019.

[44] Marek Kopicki, Sebastian Zurek, Rustam Stolkin, Thomas Mörwald, and Jeremy Wyatt. Learning to predict how rigid objects behave under simple manipulation. In *2011 IEEE International Conference on Robotics and Automation*, pages 5722–5729. IEEE, 2011.

[45] Lars Kunze, Mihai Emanuel Dolha, Emitza Guzman, and Michael Beetz. Simulation-based temporal projection of everyday robot object manipulation. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 107–114. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[46] Dominik Belter, Marek Kopicki, Sebastian Zurek, and Jeremy Wyatt. Kinematically optimised predictions of object motion. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4422–4427. IEEE, 2014.

[47] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*, 2018.

[48] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pages 387–395, 2014.

[49] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[50] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.