Disentangling Policy from Offline Task Representation Learning via Adversarial Data Augmentation

Chengxing Jia*
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
Polixir Technologies, Nanjing, China
jiacx@lamda.nju.edu.cn

Fuxiang Zhang*
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
zhangfx@lamda.nju.edu.cn

Yi-Chen Li
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
Polixir Technologies, Nanjing, China
liyc@lamda.nju.edu.cn

Chen-Xiao Gao
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
gaocx@lamda.nju.edu.cn

Xu-Hui Liu
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
Polixir Technologies, Nanjing, China
liuxh@lamda.nju.edu.cn

Lei Yuan
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
Polixir Technologies, Nanjing, China
yuanl@lamda.nju.edu.cn

Zongzhang Zhang
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
zzzhang@nju.edu.cn

Yang Yu†
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, Nanjing, China
Polixir Technologies, Nanjing, China
yuy@nju.edu.cn

ABSTRACT

Offline meta-reinforcement learning (OMRL) proficiently allows an agent to tackle novel tasks while solely relying on a static dataset. For precise and efficient task identification, existing OMRL research suggests learning separate task representations that be incorporated with policy input, thus forming a context-based meta-policy. A major approach to train task representations is to adopt contrastive learning using multi-task offline data. The dataset typically encompasses interactions from various policies (i.e., the behavior policies), thus providing a plethora of contextual information regarding different tasks. Nonetheless, amassing data from a substantial number of policies is not only impractical but also often unattainable in realistic settings. Instead, we resort to a more constrained yet practical scenario, where multi-task data collection occurs with a limited number of policies. We observed that learned task representations from previous OMRL methods tend to correlate spuriously with the behavior policy instead of reflecting the essential characteristics of the task, resulting in unfavorable out-of-distribution generalization. To alleviate this issue, we introduce a novel algorithm to disentangle the impact of behavior policy from task representation learning through a process called adversarial data augmentation. Specifically, the objective of adversarial data augmentation is not merely to generate data analogous to offline data distribution; instead, it aims to create adversarial examples designed to confound learned task representations and lead to incorrect task identification. Our experiments show that learning from such adversarial samples significantly enhances the robustness and effectiveness of the task identification process and realizes satisfactory out-of-distribution generalization. The results in MuJoCo locomotion tasks demonstrate that our approach surpasses other OMRL baselines across various meta-learning task sets.

KEYWORDS

Offline reinforcement learning; Meta-reinforcement learning

ACM Reference Format:
Chengxing Jia*, Fuxiang Zhang*, Yi-Chen Li, Chen-Xiao Gao, Xu-Hui Liu, Lei Yuan, Zongzhang Zhang, and Yang Yu†. 2024. Disentangling Policy from Offline Task Representation Learning via Adversarial Data Augmentation. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 16 pages.

*These authors contributed equally.
†Corresponding author.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

In recent years, reinforcement learning (RL) has seen substantial progress, primarily in situations where abundant interactive data

is accessible for the learning of policies [42, 44, 49]. However, the practical implementation of RL in fields such as robotics [21] and healthcare [16] remains constrained by persistent challenges relating to data efficiency and generalization [9, 23], especially when environmental interaction becomes costly or poses potential risks. Offline RL [11, 24–27, 41, 51], proposing to learn policy contexts from a pre-existing environmental dataset, offers a viable solution to address the safety concerns and sample efficiency of RL. In parallel, meta-RL [7, 8, 40, 63] aims to train generalizable policies, adaptable to various tasks sharing common structural aspects in terms of environmental dynamics and rewards.

As the intersection of offline RL and meta-RL, Offline Meta-RL (OMRL) [6, 35, 50, 62] garners growing recognition due to its potential to enhance both data efficiency and generalization. OMRL usually relies on a multi-task offline dataset collected by different behavior policies to train a meta-policy that can effectively adapt to various tasks within the range. Prior OMRL works [28, 29, 38] focus on context-based methods using a separate context encoder to extract task representations from context data. The context encoder learning is crucial for task identification and meta task generalization. Although current OMRL methods show tremendous potential on generalization to unseen tasks without the requirement of real-time interactions, they usually posit a comprehensive coverage of the offline dataset [29, 58], i.e., data from each task is collected by diverse behavior policies. Consequently, the context encoder, employed for task identity identification, can naturally develop a robust manner from diverse data. Nevertheless, in instances where interactions entail substantial risk or cost, such a presumption or condition lacks realism. We, therefore, contemplate a more feasible scenario where the data emerges from a limited number of policies, resulting in poor data coverage for task representation extraction. When deployed in tasks, the meta-policy may encounter contexts that deviate from the trained dataset, resulting in out-of-distribution scenarios, and subsequently, a failure to generalize effectively for the given tasks.

Recent studies integrate contrastive learning to optimize task representation learning, such as FOCAL [29], proposing a more stable contrastive learning objective, and CORRO [58], adopting a bi-level structure to enhance robustness. When conducting experiments in such a poor-coverage dataset, we find that task representations through naive contrastive learning correlate spuriously with behavior policies from the dataset rather than recognizing task characteristics. The reason comes from the fact that the state-action sequence, which composes the context data, is formed not only by the environment nature but also by the decision from behavior policies. In this case, the context encoder is prone to fit the properties of behavior policies to output task representations. Such a manner severely degrades policy performance as the context encoder may treat data collected by different policies as different tasks and vice versa. While existing methods develop various contrastive learning strategies to comprehend task representations [29, 58] that better outline task properties, the spurious relationship still remains since these methods do not account for the effect of behavior policies. Therefore, employing a task representation learning methodology that can disentangle behavior policies from learned representations would significantly enhance out-of-distribution generalization, subsequently facilitating adaptation to a spectrum of unseen tasks.

To disentangle the effect of behavior policy from task representations, we hope the context encoder to solely capture task-relevant data from the given context. Specifically, we find that this spurious relationship is amplified primarily when the task variation involves different environment transitions rather than reward functions. Since rewards are independently computed and will not affect subsequent trajectories, prior research can effectively address it by sharing data and relabeling rewards across tasks [28, 55]. Conversely, the variation of transition functions will result in different context data coupled with behavior policies, which cannot be simply relabeled. Consequently, our focus shifts to establishing a mapping process from transition information into acquired task representations.

To eliminate the impact of behavior policies in original offline data, we introduce adversarial data augmentation to assist task representation learning. Different from traditional data augmentation approaches that usually generate in-distribution data, the proposed adversarial data augmentation is intended to reshape data distribution from the original offline dataset so that the bias induced by behavior policies can be removed. To be more specific, the data augmentation process tries to generate the most indistinguishable interaction data to confound the context encoder within an adversarial learning objective. We adopt a model-based RL approach to generate such data without environment interactions via multiple pretrained dynamics models and a particular adversarial policy for data collection. Using augmented training data, the context encoder learns to marginalize the effect of behavior policies and correctly identify tasks based on the environment characteristics. Our results on several meta-RL task sets from locomotion benchmarks demonstrate that our OMRL method with adversarial data augmentation exhibits definite abilities to distinguish tasks and significantly outperforms other OMRL baselines even when encountering out-of-distribution context data.

2 PRELIMINARIES

2.1 Offline Meta-RL

Reinforcement learning (RL) tasks can be formulated as Markov decision processes (MDPs) [46] $M = (\mathcal{S}, \mathcal{A}, T, r, d_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T(s'|s,a)$ is the transition function, $r(s,a)$ is the reward function, $d_0(s)$ is the initial state distribution, and $\gamma \in [0,1)$ is the discount factor. In state s, an agent takes action $a \sim \pi(\cdot \mid s)$ according to its policy $\pi(a \mid s)$ and results in the next state $s' \sim T(\cdot \mid s,a)$ and the reward $r(s,a)$. The objective of RL is to maximize the cumulative discounted reward (a.k.a. return) as follows:

$$\max_{\pi} R_M(\pi) = \mathbb{E}_{s_0 \sim d_0, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim T(\cdot|s_t,a_t)} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t).$$

Meta-RL considers deploying a meta-policy that can perform well over a task distribution $P(M)$, where the meta-policy is trained in a task set and is deployed for tasks that may be unseen during the meta-train period. Offline meta-RL (OMRL) then extends this concept to offline RL settings, where the meta-policy should be learned from a static dataset $\{D_i\}_{i=1}^n$ containing interactions from different tasks $M_1, M_2, \ldots, M_n$. Every single-task data D_i is collected by some unknown behavior policies, which may be trained with modern single-task RL algorithms, for example, TD3 [12] and SAC [17] for

MuJoCo locomotion tasks [47]. The goal of OMRL is to learn a meta-policy π that maximizes the cumulative discounted reward of each task under the task distribution $P(M)$ as follows:

$$\max_{\pi} \mathbb{E}_{M\sim P(\cdot)} \left[R_M(\pi)\right].$$

It is worth noting that OMRL does not posit the number of behavior policies for single-task data. Traditionally, previous approaches [29, 58] adopt a variety of policy checkpoints to collect interaction data for each task, resulting in a diverse data distribution for task identification. Such comprehensive data coverage may be plausible to alleviate the data bias induced by specific behavior policies. However, using exhaustive policies to collect data is not practical for realistic scenarios, which, in most cases, is what OMRL methods should be applied to.

2.2 Task Representation Learning

Task identification is vital for meta-RL to realize fast adaptation to unseen tasks [3]. Here we focus on context-based meta-RL methods, which learn a meta-policy $\pi(a \mid s, z)$ that is explicitly conditioned on a task representation z. It additionally adopts a context encoder $\phi(z \mid x)$ to extract task representations z from context data x. The context data x, which is often a transition tuple or a sub-trajectory of multiple interaction steps, can be collected from an arbitrary policy. This feature enables zero-shot generalization of the meta-policy when providing pre-collected or bootstrapped context data without extra updates during the adaptation phase.

Although online meta-RL often trains the context encoder and the meta-policy jointly [40], in OMRL settings they are typically learned separately with different objectives to stabilize the training process. From the perspective of information theory, an ideal context encoder should maximize the mutual information $I(z; M)$ between the task representation z and the corresponding MDP M. As $I(z; M)$ is not directly tractable, we proceed by optimizing an InfoNCE objective [36], which is a lower bound of the mutual information metric as proved by [58]:

$$\begin{aligned} I(z;M) &\geq \mathbb{E}_{M\sim P(\cdot), z, z^*\sim\phi(\cdot|x_M)} \log \frac{\exp(S(z,z^*))}{\Phi(z)} + C, \\ \text{where } \Phi(z) &= \sum\nolimits_{M'\sim P(\cdot), z'\sim\phi(\cdot|x_{M'})} \exp(S(z,z')). \end{aligned} \tag{1}$$

We denote x_M and $x_{M'}$ as context data from tasks M and M' respectively, and $S(\cdot,\cdot)$ as a score function which measures the similarity between two representations, e.g., their inner product or cosine similarity. C is a constant that relates to the task distribution $P(M)$.

An intuitive understanding of maximizing the lower bound in Equation (1) is to make representations of the same task similar while keeping those of different tasks apart. Nevertheless, naively applying this objective may fail to establish the correct relationship between the characteristics and representation of the tasks in the offline setting, because x_M is sampled from the offline dataset and influenced by the behavior policy and task M jointly. When each task data D_i is collected by a few ad-hoc behavior policies and comparatively limited in coverage, the context encoder is prone to overfitting the behavior data and fails to generalize to unseen data distributions during evaluation. In the following section, we propose a method to disentangle the effect of behavior policy and base the representation only on the task nature.

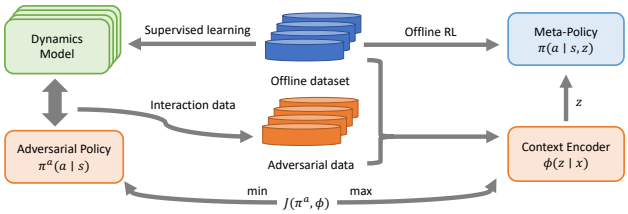


Figure 1: The overall process of using adversarial data augmentation for offline meta-RL.

3 ADVERSARIAL DATA AUGMENTATION

In this section, we present a novel OMRL algorithm that can disentangle behavior policies during task representation learning via an effective approach called adversarial data augmentation, as shown in Figure 1. This augmentation process, different from generating in-distribution auxiliary data, aims to provide a more robust data distribution that can minimize the influence of specific behavior policies. Therefore, consequent task representations can break down the spurious relationship between behavior policies and task representations. In Section 3.1, we will introduce the adversarial training objective for our data augmentation phase. In Section 3.2, we demonstrate how we utilize a model-based RL approach to acquire additional data for task representation learning in the offline setting. In Section 3.3, we develop the algorithmic framework to use adversarial data argumentation in OMRL and present its implementation details.

3.1 Adversarial Training Objective

As mentioned in Section 2.2, we can derive a lower bound in the form of InfoNCE [36] from the mutual information between task representations and the task MDP, which can be described as the following objective:

$$\begin{aligned} J(\phi) &= \mathbb{E}_{M\sim P(\cdot), z, z^*\sim\phi(\cdot|x_M)} \log \frac{\exp(z^\top z^*)}{\Phi(z)}, \\ \text{where } \Phi(z) &= \sum\nolimits_{M'\sim P(\cdot), z'\sim\phi(\cdot|x_{M'})} \exp(z^\top z'). \end{aligned} \tag{2}$$

Here we specify the score function $S(\cdot,\cdot)$ in Equation (1) to be the inner product, which aligns to prevalent contrastive learning works [13, 39]. Though these works suggest that InfoNCE can be an effective objective to learn powerful representations according to given labels, the objective leads to building a connection between task representations z and task context data x_M, rather than the task M itself. Since the context data x_M is sampled from the offline dataset, collected by specific behavior policies, this objective establishes a spurious relationship between behavior policies and learned task representations. In Section 5.1 we will show an instantiation of this spurious relationship.

To eliminate such a spurious relationship, we envision that the context encoder can derive correct task representations from not only in-distribution data but also context data collected by any policy. To realize such generalization, the context encoder is expected to learn from a wider range of context data, which may help marginalize the effect of behavior policies. However, it is intractable to involve diverse context data collected by disparate policies in

offline training. Inspired by previous progress in adversarial training [15, 18], we find that an intuitive approach is to make the context encoder capable of discerning the most indistinguishable samples. Therefore, the context encoder can possess adequate robustness to identify the task no matter what behavior policies the data comes from.

To generate additional context data for adversarial training, we adopt a model-based RL approach [31, 33] to learn environment models and thus provide interaction data in the offline setting. For a particular meta-RL task distribution $P(M)$, previous methods usually consider task variants on environment dynamics T and reward functions r. As discussed in the introduction section, we only consider task variations where environment dynamics differ since it is the key factor in the spurious relationship. Given a multi-task offline dataset $\{D_i\}_{i=1}^N$, we learn several dynamics models via supervised learning with transition tuples sampled from the dataset. These learned dynamics models formulate a new task distribution $\hat{P}(\hat{M})$ that is quite similar to the real task distribution. Denoting $x_{\hat{M}}$ as context data collected from the simulating task $\hat{M}$ with an auxiliary policy π^a, we summarize our adversarial optimization problem in the max-min form below:

$$\max_{\phi} \min_{\pi^a} J(\pi^a, \phi) = \mathbb{E}_{\hat{M}\sim\hat{P}(\cdot), z, z^*\sim\phi(\cdot|x_{\hat{M}}(\pi^a))} \log \frac{\exp(z^\top z^*)}{\Phi(z)}, \quad \text{where } \Phi(z) = \sum_{\hat{M}'\sim\hat{P}(\cdot), z'\sim\phi(\cdot|x_{\hat{M}'}(\pi^a))} \exp(z^\top z'). \tag{3}$$

Note that we mark the context data $x_{\hat{M}}(\pi^a)$ for its sources from the simulating MDP $\hat{M}$ and the auxiliary policy π^a. Here we name the auxiliary policy π^a as *adversarial policy* since it is trained against the context encoder's goal. Using the InfoNCE objective $J(\pi^a, \phi)$ as the basic form, the minimization part guarantees that the adversarial policy can learn to search the most indistinguishable data for a specific task. Then the maximization part helps the task identification of the context encoder in the worst-case data. Therefore, the context encoder is expected to recognize tasks well for any data distribution, as it has learned from the weakest data the adversarial policy chooses to generate.

3.2 Adversarial Data Generation

Starting from Equation (3), we consider how to generate data from a simulating task distribution. To compensate for the absence of the task MDP in offline settings, we first train several dynamics models for each task data D_i like previous model-based RL approaches [31, 33], following a supervised learning manner:

$$\max_{\hat{T}_i} \mathbb{E}_{(s,a,s')\sim D_i}[\log \hat{T}_i(s'|s,a)]. \tag{4}$$

In addition, the reward function is learned in a similar supervised learning way with all of the task data, as the reward functions are shared across different tasks. According to Equation (3), we require context data from simulating task distribution, which is determined by both the simulating environment model and an adversarial policy. As the environment model should precisely reflect the original task distribution, the functionality of producing adversarial data to confound the context encoder, i.e., minimizing $J(\pi^a, \phi)$, falls to the adversarial policy π^a. Since context data forms a sequence-style format, we adopt reinforcement learning to optimize the adversarial policy, which helps make the goal of obscuring task identification consistent along the whole trajectory. Specifically, we define the reward as how much ambiguity the task representations are measured by InfoNCE compared to the previous step. Letting $R(z_t) = \mathbb{E}_{z^*}[\log \frac{\exp(z_t^\top z^*)}{\Phi(z_t)}]$ denote the value of $J(\pi^a, \phi)$ at step t, the reward for training π^a towards the adversarial objective is:

$$r_t^{\text{adv}} = R(z_{t+1}) - R(z_t) = \mathbb{E}_{z^*}\left[\log \frac{\exp(z_{t+1}^\top z^*)}{\Phi(z_{t+1})} - \log \frac{\exp(z_t^\top z^*)}{\Phi(z_t)}\right]. \tag{5}$$

Intuitively, we reward the adversarial policy for taking actions that will decrease the task representation similarity within the same task. Therefore, the adversarial policy π^a is helpful to mine the most indistinguishable data to enhance the robustness of the context encoder ϕ. Nevertheless, we notice that adopting the reward function of Equation (5) is not sufficient to train a well-behaved adversarial policy. The learned policy may explore states that are far beyond the environment model's support, which leads to useless data. On the other hand, when only following the adversarial reward, the policy will learn to make decisions that are irrelevant to the task goals. To this point, we highlight two additional terms that should be incorporated with the adversarial reward:

- *Uncertainty penalty.* Like previous offline model-based RL works [22, 57], we quantify the uncertainty $u_t(s, a)$ of the current state-action pair by the aleatoric uncertainty of the learned dynamics models on the state transition. Specifically, the aleatoric uncertainty is the maximum standard deviation of the Gaussian distribution of the next states among a group of dynamics models. This uncertainty metric serves as a penalty for rollouts in the models to prevent the policy from collecting imprecise data.
- *Task completeness.* We also append the task reward $\hat{r}(s, a)$ provided by the learned reward model to the total reward used for training the adversarial policy. Adopting the task reward should be a natural approach since we focus more on contexts that are related to the task goal rather than other meaningless data.

We also draw theoretical inspirations for the design of these two reward components. We illustrate these reward terms are beneficial to generate useful data in Appendix A. Combining the two additional reward terms above, we can acquire the total reward to train the adversarial policy:

$$J(\pi^a) = \mathbb{E}_{\hat{M}\sim\hat{P}(\cdot), s_t\sim\hat{M}, a_t\sim\pi^a(\cdot|s_t)}\Big[\sum_{t=0}^{\infty}\gamma^t\big(r_t^{\text{adv}} - \lambda_1 u_t(s,a) + \lambda_2\hat{r}(s,a)\big)\Big], \tag{6}$$

where we get states s_t from the learned environment model and actions a_t from the adversarial policy π^a for each task. λ_1 and λ_2 are two coefficients to balance the weights of different reward terms. We adopt the branch rollout approach [19] to mitigate the influence of model error and avoid the requirement of initial state distribution. Instead of using a single dataset, we draw initial states from the joint dataset $D = \bigcup_i D_i$ to enhance the coverage of rollout.

Algorithm 1 Meta-training with adversarial data augmentation

Input: The offline dataset $\{D_i\}_{i=1}^n$ from n tasks, the context encoder $\phi(z|x)$, the meta-policy $\pi(a|s,z)$, and the adversarial policy $\pi^a(a|s)$

▷ *Dynamics model training*
for $i = 1, 2, \cdots, n$ **do**
  Learn m dynamics models with supervised learning on D_i to represent the task dynamics
end for
▷ *Task representation learning*
for $K = 1, \ldots, K_c$ **do**
  ▷ *Collect adversarial data*
  for $i = 1, 2, \cdots, n$ **do**
    Sample adversarial data $\hat{D}_i$ from learned dynamics models $\hat{M}_i$ using adversarial policy π^a
  end for
  ▷ *Train the adversarial policy*
  Compute rewards from sampled adversarial data $\hat{D}_1, \ldots, \hat{D}_n$ according to Equation (6)
  Update adversarial policy $\pi^a(a|s)$ using SAC
  ▷ *Train the context encoder*
  Update context encoder $\phi(z|x)$ with sampled adversarial data $\hat{D}_1, \ldots, \hat{D}_n$
end for
▷ *Meta-policy learning*
Update meta-policy $\pi(a|s,z)$ with dataset $\{D_i\}_{i=1}^n$ with fixed context encoder ϕ

As for the maximization part of Equation (3), the context encoder ϕ can learn from the adversarial data generated by the adversarial policy in the environment model by fixing the adversarial policy. The learning process is through simple contrastive learning with the InfoNCE objective:

$$J(\phi) = \mathbb{E}_{\hat{M}\sim\hat{P}(\cdot),z,z^*\sim\phi(\cdot|x_{\hat{M}}(\pi^a))} \log \frac{\exp(z^\top z^*)}{\Phi(z)}. \tag{7}$$

3.3 Overall Framework

Though most of our work focuses on the stage of task representation learning, we further propose an algorithm that can meta-train a policy using adversarial data augmentation. As shown in Algorithm 1, the meta-training process consists of three stages: task dynamics model training, task representation learning, and meta-policy learning. Utilizing provided offline datasets, we first train several dynamics models for each task, which is useful for data generation. During task representation learning, we train the adversarial policy π^a and the context encoder ϕ alternatively like common generative adversarial learning approaches. Finally, we fix the context encoder ϕ and train the meta-policy with a standard offline RL algorithm. To be more specific, we adopt the classic SAC algorithm [17] and add a standard behavior-cloning-style regularization term [10] for its offline adaptation. The context encoder $\phi(z \mid x)$ has a Transformer-based structure [48] for better capturing context in the form of sequence information. For detailed attention computation, we concatenate the current observation and the last action as the source of query, and the context is in the form of state-action pair and serves as the source of key and value. We detach the gradient of output task representation and feed it into the meta-policy $\pi(a \mid s, z)$ along with the observed state s. The network structure of the meta-policy is a simple multi-layer perceptron (MLP), which is the same as the adversarial policy $\pi^a(a \mid s)$ despite the removal of input task representation z.

4 RELATED WORK

4.1 Offline Reinforcement Learning

Offline reinforcement learning proposes to learn policies from a fixed offline dataset and is considered helpful for many realistic domains such as healthcare and robotics [16, 21] to avoid exhaustive online interactions that is usually impractical for real-world scenarios. The main issue with offline learning is the distribution shift caused by the mismatch between the behavior and exploitation policies [27], inducing a huge challenge for online evaluation. When adopting RL algorithms from online RL literature, offline model-free RL methods usually adopt a conservative way to learn policies. For example, they explicitly constrain the policy to be similar to the behavior policy with different regularization manners [4, 10, 11, 25, 34, 34, 37, 54, 59, 61], imposing conservative penalty on out-of-distribution actions [1, 2, 26], or re-weighting offline samples during policy evaluation and improvement [24, 60].

Another line of research is offline model-based RL that incorporates dynamics models [5, 20, 22, 43, 45, 56, 57] to synthesize data for better generalization. However, the dynamics models are typically trained from offline dataset via supervised learning and still face model errors on unseen state-action pairs. One approach to account for this problem is to modify the learned dynamics [22] or add reward penalties to highly uncertain areas [57], which is quantified by disagreement or maximum standard deviation over an ensemble of dynamics. COMBO [56] bypasses uncertainty quantification by leveraging the model-free method for optimization. Recently, RAMBO [43] introduces the idea of robust RL and proposes to learn adversarial dynamics models to incorporate conservatism.

Our paper generally adopts offline model-free RL approaches to train meta-policies by introducing SAC [17] with adding a behavior-cloning-style regularization term [10]. When tackling the auxiliary adversarial policy, we also utilize a model-based RL approach. We train several dynamics models to augment the original dataset and penalize the policy by the model uncertainty, which builds a connection with the model-based offline RL literature [22, 57].

4.2 Offline Meta-Reinforcement Learning

Meta-reinforcement learning [3] is a popular research area that learns a policy that is capable of adapting to any new task from the task distribution with as little data as possible. When more and more meta-RL approaches are shown to learn a good meta-policy with extremely comprehensive interactions [8, 53], offline meta-RL [6] extends the boundary of meta-RL by learning the meta-policy directly from offline data.

Different from a major research direction on applying model-agnostic meta-learning (MAML) methods [8] to offline meta-RL [32], we here concentrate more on adopting context-based meta-RL approaches [40]. As MAML-style approaches often require few-shot updates for task adaptation, context-based meta-RL methods can

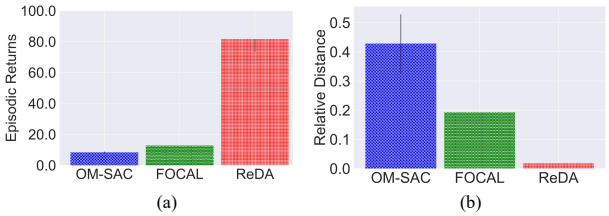


Figure 2: (a) Performance on `InvertedPendulum-v2` with 1.0x gravity coefficient. (b) The relative representation metric of different methods.

utilize pre-collected context data from other policies or its bootstrapped interactions, realizing a promising capability of zero-shot generalization. When extending context-based meta-RL methods to the offline setting, previous works typically employ contrastive-style learning objectives to learn discriminative representations for each task, such as triplet loss [28], InfoNCE [36, 58], and a negative-power variant contrastive loss [29]. Similar to our motivation, MBML [28] and CORRO [58] also point out the impact of behavior policy on task identification, and try to alleviate this issue via relabeling data with known reward functions or learned generative models like conditional variational autoencoders. Nonetheless, we note that these works do not indeed eliminate the correlation between the behavior policy and the task transition. CSRO [14] aims to solve this problem by introducing an information bottleneck between the policy and task features. However, we notice that previous works assume that the behavior policies used to collect the dataset are diversified enough (e.g. CORRO uses each saved checkpoint during the training process to collect data), which alleviates the issue of building the spurious relationship with behavior policies. To the best of our knowledge, we are the first to exhaustively consider the spurious relationship in task representation learning and eliminate the effect of behavior policies.

5 EXPERIMENTS

In this section, we conduct experiments to evaluate the proposed adversarial data augmentation. We investigate whether the task **Re**presentation learning via adversarial **D**ata **A**ugmentation (ReDA) can disentangle the effect of behavior policies and thus helps train generalizable meta-policies. In Section 5.1, we use a didactic example to show the issue of spurious relationship and how ReDA can help address it via a classic control task. In Section 5.2, we establish tasks and datasets to benchmark ReDA and other OMRL baselines. We particularly design two evaluation manners called on-policy and off-policy protocols to verify the effectiveness of learned task representation under different circumstances. Finally in Section 5.3, we ablate multiple design choices in the adversarial data augmentation process, for example, the use of models and different reward terms on training adversarial policy. We also provide illustrative explanations on task representations in out-of-distribution data.

5.1 A Didactic Example

As a straightforward example to show the spurious relationship between behavior policies and task representations, we conduct experiments on the `InvertedPendulum-v2` environment from Gym's classic control benchmarks. To make task variants, we modify the gravity of the environment by multiplying it by 1.0 and 2.0, creating two distinct tasks. For offline dataset collection, we independently train SAC policies on each task. We select the checkpoint that achieves an expected episodic return of 100 for gravity 1.0 (task 1) and that of 50 of gravity 2.0 (task 2) to collect training data. We denote these two datasets as D_1^{train} and D_2^{train}. After training, we choose the test dataset from task 1 with returns of 50, which is denoted by D_1^{test}. To highlight the spurious relationship issue, we evaluate the meta-policy on task 1 with context data from D_1^{test}. We speculate that the existence of spurious relationship may encourage the context encoder to predict task representations more similar to task 2. We test the performance with several baselines. FOCAL is an OMRL baseline [29] that designs an ad-hoc contrastive learning objective for task representation learning. OM-SAC is a baseline that adopts the same context encoder structure as ReDA but only adopts the same offline RL manner as ReDA without the data augmentation and context encoder learning.

We depict the performance in task 1 when using context data from D_1^{test} in Figure 2(a). The majority of baselines exhibit poor performance, indicating a failure in task identification and generalization. In contrast, our method ReDA can better capture task features and thus performs exceptionally well when facing out-of-distribution contexts. To get deeper insights on how well the task representations are, we define a relative representation metric $d(\phi)$ to delineate whether the task representation is close to task 1 given its out-of-distribution context data, as follows:

$$d(\phi) = \frac{|D_2^{\text{train}}| \sum_{z_1 \sim \phi(\cdot|x_1), x_1 \sim D_1^{\text{train}}} \sum_{z \sim \phi(\cdot|x), x \sim D_1^{\text{test}}} \|z_1 - z\|^2}{|D_1^{\text{train}}| \sum_{z_2 \sim \phi(\cdot|x_2), x_2 \sim D_2^{\text{train}}} \sum_{z \sim \phi(\cdot|x), x \sim D_1^{\text{test}}} \|z_2 - z\|^2}.$$

We acquire this metric by feeding context data from each training dataset and computing the mean square errors with task representations from test data. It is obvious that when $d(\phi)$ is close to 0, the context encoder correctly recognizes most context data as task 1. A large value of $d(\phi)$ means the context encoder cannot perform task identification well. As illustrated in Figure 2(b), we find that only the context encoder of ReDA reaches a $d(\phi)$ near 0, indicating its effectiveness. In contrast, OM-SAC and FOCAL both have a larger $d(\phi)$ compared to ReDA, while the particular contrastive objective of FOCAL makes it slightly better than OM-SAC. We include a detailed description about this example in Appendix B.

5.2 Performance on MuJoCo Benchmarks

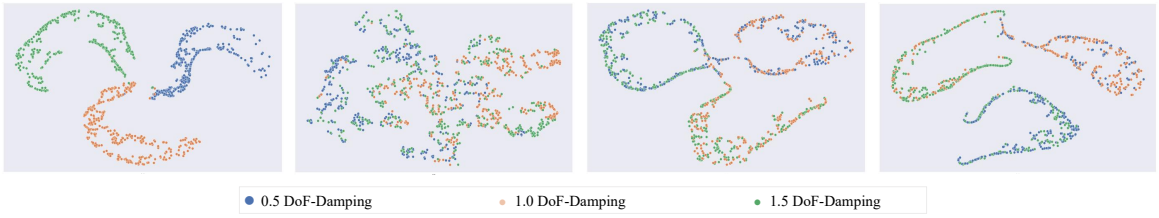
Following the experiment setups in previous meta-RL works [29, 30, 58], we construct multiple tasks by varying certain hyperparameters of the simulator in MuJoCo locomotion tasks [47]. Specifically, we choose a set of Gym's MuJoCo environments, including `HalfCheetah-v2`, `Hopper-v2`, `Walker2d-v2`, and `Ant-v2`. For each environment, we perturb the environment parameters, namely `gravity` and `dof-damping`, of the original simulator to create a wide range of tasks. During offline training, the meta-policy is trained when the environment parameters are set to 0.5, 1.0, and 1.5 times of their original values for each environment. We also choose 0.8 and 1.2 times of the parameters as unseen tasks to evaluate the generalization ability. For each task, we train behavior policies of multiple qualities via SAC [17]. Similar to that in CORRO [58], we save the checkpoint during the whole learning process but only

Table 1: Average score of the meta-policy under the on-policy protocol. The performance is averaged over 5 random seeds. We abbreviate task names for simplicity. For example, `Walker2d Gravity-1` denotes the `Walker2d-v2` task with training datasets collected from 1 checkpoint.

Task Type	OM-BC	OM-SAC	CORRO	FOCAL	PromptDT	ReDA (ours)
Walker2d Gravity-1	835.7 ± 158.9	1227.4 ± 176.8	965.7 ± 283.5	1260.2 ± 292.1	1019.7 ± 216.4	**1720.8 ± 321.3**
Walker2d Gravity-3	1110.4 ± 170.6	1180.1 ± 175.7	1269.8 ± 309.9	1584.6 ± 361.4	1125.0 ± 226.0	**1611.6 ± 288.0**
Walker2d Gravity-5	657.4 ± 265.8	1216.5 ± 388.4	1412.5 ± 387.7	2181.7 ± 355.7	1082.8 ± 277.7	**2560.7 ± 362.1**
Walker2d Dof-Damping-1	2692.3 ± 488.5	2995.4 ± 287.9	2756.5 ± 379.5	3409.2 ± 343.9	2635.0 ± 298.6	**3769.2 ± 364.2**
Walker2d Dof-Damping-3	1734.5 ± 309.0	1941.4 ± 342.5	2232.0 ± 332.4	1940.7 ± 321.9	1935.1 ± 342.3	**3536.7 ± 258.7**
Walker2d Dof-Damping-5	998.8 ± 231.3	1478.9 ± 267.6	1647.5 ± 423.3	2222.0 ± 311.0	1136.0 ± 281.2	**2497.6 ± 339.4**
HalfCheetah Gravity-1	4204.0 ± 840.3	4641.7 ± 745.6	5353.7 ± 869.6	3651.9 ± 505.9	5034.1 ± 644.3	**6772.0 ± 649.5**
HalfCheetah Gravity-3	3719.9 ± 614.6	5419.6 ± 682.0	5054.6 ± 754.7	2950.0 ± 391.7	3900.9 ± 944.0	**6130.8 ± 856.4**
HalfCheetah Gravity-5	2684.8 ± 584.5	4398.2 ± 778.3	4165.7 ± 654.7	1863.9 ± 633.4	3494.8 ± 711.6	**4549.0 ± 753.9**

Table 2: Average score of the meta-policy under the off-policy protocol. The performance is averaged over 5 random seeds.

Task Type	OM-BC	OM-SAC	CORRO	FOCAL	PromptDT	ReDA (ours)
Walker2d Gravity-1	102.9 ± 16.5	773.9 ± 115.4	354.9 ± 113.3	972.4 ± 314.9	516.5 ± 165.3	**1349.8 ± 374.2**
Walker2d Gravity-3	887.5 ± 197.7	1096.0 ± 182.6	755.7 ± 387.5	1331.1 ± 384.5	996.6 ± 204.1	**1336.3 ± 363.3**
Walker2d Gravity-5	655.4 ± 253.2	983.0 ± 394.0	899.1 ± 298.5	2318.6 ± 312.3	1254.8 ± 255.0	**2421.0 ± 398.7**
Walker2d Dof-Damping-1	2566.4 ± 358.6	2567.3 ± 381.1	2138.2 ± 432.6	2141.7 ± 299.3	2594.3 ± 315.6	**3274.5 ± 429.5**
Walker2d Dof-Damping-3	1599.2 ± 340.2	1611.1 ± 415.6	1765.8 ± 498.6	1619.0 ± 362.0	1883.9 ± 331.2	**3236.0 ± 322.7**
Walker2d Dof-Damping-5	823.6 ± 340.1	1798.6 ± 345.6	1667.2 ± 438.8	1943.4 ± 353.4	1446.5 ± 326.5	**2243.7 ± 352.2**
HalfCheetah Gravity-1	6024.0 ± 768.2	3985.1 ± 676.9	5244.7 ± 766.6	3960.4 ± 517.0	5755.3 ± 582.9	**6243.8 ± 604.4**
HalfCheetah Gravity-3	2112.9 ± 601.1	4792.5 ± 596.7	4865.2 ± 710.8	2293.1 ± 532.1	3211.5 ± 660.6	**5749.2 ± 599.0**
HalfCheetah Gravity-5	3006.2 ± 684.3	4164.7 ± 602.3	3654.6 ± 677.0	2181.2 ± 414.6	3243.8 ± 601.1	**4268.1 ± 562.2**



(a) ReDA (b) OM-SAC (c) FOCAL (d) CORRO

Figure 3: Visualization on task representations with t-SNE dimensionality reduction for (a) ReDA, (b) OM-SAC, (c) FOCAL, and (d) CORRO on the task set Walker2d Dof-Damping-1. Task representations from different tasks are shown in distinct colors.

select 1, 3, and 5 checkpoints from all to collect training data and use the rest checkpoints to collect test data. Our data collection is distinguished from that in CORRO, which utilizes all checkpoints to collect training data. We name the task type by the number of checkpoints used for training. For example, `Walker2d Gravity-1` means that the offline dataset of Walker2d is collected by 1 checkpoint in 0.5x, 1.0x, and 1.5x gravity. We put more detailed descriptions on how we select checkpoints and collect data to Appendix B. Besides above baselines, we also include OM-BC, which trains a meta-policy with behavior cloning, an OMRL baseline CORRO [58], and PromptDT [52]. To better show the generalization ability of different OMRL algorithms, we apply two evaluation protocols to shape different context data: *on-policy protocol* and *off-policy protocol*.

On-policy protocol. In the on-policy protocol, the context data comes from interactions the meta-policy encounters when deployed to the task. Compared to the behavior policies that collect the offline data, the meta-policy is often an improved one, and thus there still exists a mismatch between training-time and test-time distribution. We report partial performance in Table 1 and due to the limitation of space, the full results are provided in Appendix B.10. The results demonstrate that ReDA significantly outperforms other baseline methods when using bootstrapped context data, indicating the superior ability of task recognition and generalization. Previous

Table 3: Ablations on different variants of ReDA. The variant w/o model has no data augmentation and the variant w/o adv. augments data with rollouts from random policies. "Dof-Damp" is short for Dof-Damping.

Task Param	w/o model	w/o adv.	ReDA
Walker2d Gravity-1	834.0	1197.5	**1349.8 ± 321.3**
Walker2d Gravity-3	1135.5	1268.9	**1336.3 ± 363.3**
Walker2d Gravity-5	1348.6	1220.2	**2421.0 ± 398.7**
Walker2d DofDamp-1	2606.5	3186.2	**3274.5 ± 429.5**
Walker2d DofDamp-3	1773.3	2845.8	**3236.0 ± 322.7**
Walker2d DofDamp-5	1641.7	1566.0	**2243.7 ± 352.2**

Table 4: Ablations on adversarial policy learning with different reward terms. The variant w/o UP learns the policy is ReDA with $\lambda_1 = 0$. The variant w/o TC is ReDA with $\lambda_2 = 0$.

Task Param	w/o UP	w/o TC	ReDA
Walker2d Gravity-1	1185.4	1281.3	**1349.8 ± 321.3**
Walker2d Gravity-3	1249.0	1292.3	**1336.3 ± 363.3**
Walker2d Gravity-5	1572.1	2212.5	**2421.0 ± 398.7**
Walker2d DofDamp-1	2893.2	3002.4	**3274.5 ± 429.5**
Walker2d DofDamp-3	2985.2	3145.8	**3236.0 ± 322.7**
Walker2d DofDamp-5	1787.8	2209.2	**2243.7 ± 352.2**

works like FOCAL and CORRO exhibit similar performance to OM-SAC, indicating that their task representation learning approaches cannot work for this extreme setting of poor data coverage.

Off-policy protocol. To further investigate the generalization ability of the context encoders, we introduce the off-policy protocol where we replace the context with interactions collected by unseen checkpoints. This evaluation type typically raises a higher requirement to the generalization capability of the context encoder since it should output correct task representations collected from arbitrary policies. Due to the limitation of space, we provide the partial results of performance in Table 2 and full results in Appendix B.10. We notably find that almost all methods have performance drops in the off-policy protocol, while our algorithm still achieves significant performance advantages compared to other baselines.

5.3 Ablation Studies

We design a few ablation studies to investigate why ReDA works in the experiments above. To compare how well the method identifies different tasks, we visualize the task representations trained in the `Walker2d Dof-Damping-1` setting in Figure 3. The task representations are obtained by feeding the context encoders with interaction data from tasks including different gravity coefficients. We find that ReDA can generate generally distinct task representations for different tasks even though some data patterns are never encountered during training, indicating that the proposed adversarial data augmentation helps improve out-of-distribution generalization. When we remove the adversarial data augmentation of ReDA and keep other procedures the same (shown in the results of OM-SAC), the consequent task representations become indistinguishable, as well as those from FOCAL and CORRO. To ablate each component in our adversarial data augmentation process, we also design two variants of ReDA and compare their performance on `Walker2d Gravity` and `Walker2d Dof-Damping` dataset. One variant is to implement ReDA but does not introduce pre-trained dynamics models for data augmentation (w/o model). The other variant is to utilize the model for data augmentation but with a random policy rather than an additional adversarial policy (w/o adv.). As shown in Table 3, the results on the Walker2d task sets demonstrate that both two variants degrade severely compared to ReDA, which verifies the effectiveness of the proposed adversarial data augmentation. Besides, The performance of the ReDA variant without data augmentation is not as high as the ReDA variant using dynamics models, indicating that introducing additional context data from pre-trained dynamics models is helpful for generalization of task representations. Then we turn to ablate the design of reward for learning the adversarial policy. As we adopt two additional reward terms in our methodology, we propose two variants of ReDA. One variant (w/o UP) is ReDA without uncertainty penalty ($\lambda_1 = 0$) and the other (w/o TC) is ReDA without task completeness reward ($\lambda_2 = 0$). The results in Table 4 show that our reward design could promote task representation generalization while these two variants show less promising performance. The uncertainty penalty, which is adopted from offline model-based literature [57], is shown to be more critical to policy performance. We also provide more results of task representation visualization on mentioned ablation variants in Appendix B, which shows that ReDA outputs more distinct representations than those variants.

6 CONCLUSION

In this paper, we shift our attention to OMRL in a low-data regime, where the offline dataset is limited in terms of coverage by collective policy. We underscore the criticality of disentangling behavior policies from task representation learning and offer a practical solution named adversarial data augmentation. We train an adversarial policy to prevent the context encoder from identifying the task and augment the offline dataset. The data is collected in learned dynamics models, which is irrelevant to the original behavior policies and thus can assist in learning robust task representations. Experimental results reveal that when the dataset is encompassed by limited behavior policies, previous OMRL methods encounter difficulties with task identification. Nevertheless, task representation learning with adversarial data augmentation resolves this problem, effectively generalizing to out-of-distribution context data at test time. Our aspiration for this research is to stimulate further exploration into the impact of behavior policies and the standardization of OMRL benchmarks. One identified limitation of our work is the rudimentary designation of the context encoder, with us merely inheriting a simple Transformer architecture, tasked with extracting and aggregating context information. Due to its departure from this study's primary focus, we reserve this issue for future work.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation of China (61921006) and Tencent AI Lab (RBFR2023011). We thank the anonymous reviewers for their support and helpful discussions.

REFERENCES

[1] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. 2021. Uncertainty-based Offline Reinforcement Learning with Diversified Q-Ensemble. In *Advances in Neural Information Processing Systems*. 7436–7447.

[2] Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhi-Hong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. 2022. Pessimistic Bootstrapping for Uncertainty-Driven Offline Reinforcement Learning. In *International Conference on Learning Representations*.

[3] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa M. Zintgraf, Chelsea Finn, and Shimon Whiteson. 2023. A Survey of Meta-Reinforcement Learning. *arXiv preprint arXiv 2301.08028* (2023).

[4] Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, and Jun Zhu. 2023. Offline Reinforcement Learning via High-Fidelity Generative Behavior Modeling. In *International Conference on Learning Representations*.

[5] Xiong-Hui Chen, Yang Yu, Qingyang Li, Fan-Ming Luo, Zhiwei (Tony) Qin, Wenjie Shang, and Jieping Ye. 2021. Offline Model-Based Adaptable Policy Learning. In *Advances in Neural Information Processing Systems*. 8432–8443.

[6] Ron Dorfman, Idan Shenfeld, and Aviv Tamar. 2021. Offline Meta Reinforcement Learning–Identifiability Challenges and Effective Data Collection Strategies. In *Advances in Neural Information Processing Systems*. 4607–4618.

[7] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv preprint arXiv:1611.02779* (2016).

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *International conference on machine learning*. PMLR, 1126–1135.

[9] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. 2018. An Introduction to Deep Reinforcement Learning. *Foundations and Trends in Machine Learning* 11, 3-4 (2018), 219–354.

[10] Scott Fujimoto and Shixiang Shane Gu. 2021. A Minimalist Approach to Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 20132–20145.

[11] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In *International Conference on Machine Learning*. 2052–2062.

[12] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*. 1582–1591.

[13] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Conference on Empirical Methods in Natural Language Processing*. 6894–6910.

[14] Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, Qi Guo, Ling Li, and Yunji Chen. 2023. Context Shift Reduction for Offline Meta-Reinforcement Learning. *arXiv preprint arXiv:2311.03695* (2023).

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[16] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. 2019. Guidelines for Reinforcement Learning in Healthcare. *Nature Medicine* 25, 1 (2019), 16–18.

[17] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*. 1856–1865.

[18] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*. 4565–4573.

[19] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. 2019. When to Trust Your Model: Model-Based Policy Optimization. In *Advances in Neural Information Processing Systems*. 12498–12509.

[20] Xue-Kun Jin, Xu-Hui Liu, Shengyi Jiang, and Yang Yu. 2022. Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning. *arXiv preprint arXiv:2206.02000* (2022).

[21] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. 2019. Residual Reinforcement Learning for Robot Control. In *International Conference on Robotics and Automation*. 6023–6029.

[22] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. MOReL: Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 21810–21823.

[23] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2021. A Survey of Generalisation in Deep Reinforcement Learning. *arXiv preprint arXiv:2111.09794* (2021).

[24] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*.

[25] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems*. 11761–11771.

[26] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 1179–1191.

[27] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643* (2020).

[28] Jiachen Li, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Henrik Christensen, and Hao Su. 2020. Multi-Task Batch Reinforcement Learning with Metric Learning. In *Advances in Neural Information Processing Systems*. 6197–6210.

[29] Lanqing Li, Rui Yang, and Dijun Luo. 2021. FOCAL: Efficient Fully-Offline Meta-Reinforcement Learning via Distance Metric Learning and Behavior Regularization. In *International Conference on Learning Representations*.

[30] Fan-Ming Luo, Shengyi Jiang, Yang Yu, Zongzhang Zhang, and Yi-Feng Zhang. 2022. Adapt to Environment Sudden Changes by Learning a Context Sensitive Policy. In *AAAI Conference on Artificial Intelligence*. 7637–7646.

[31] Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. 2022. A Survey on Model-Based Reinforcement Learning. *arXiv preprint arXiv 2206.09328* (2022).

[32] Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. 2021. Offline Meta-Reinforcement Learning with Advantage Weighting. In *International Conference on Machine Learning*, Vol. 139. 7780–7791.

[33] Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. 2023. Model-Based Reinforcement Learning: A Survey. *Foundations and Trends in Machine Learning* 16, 1 (2023), 1–118.

[34] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. AWAC: Accelerating Online Reinforcement Learning with Offline Datasets. *arXiv preprint arXiv:2006.09359* (2020).

[35] Fei Ni, Jianye Hao, Yao Mu, Yifu Yuan, Yan Zheng, Bin Wang, and Zhixuan Liang. 2023. MetaDiffuser: Diffusion Model as Conditional Planner for Offline Meta-RL. In *International Conference on Machine Learning*. 26087–26105.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

[37] Zhiyong Peng, Changlin Han, Yadong Liu, and Zongtan Zhou. 2023. Weighted Policy Constraints for Offline Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*. 9435–9443.

[38] Vitchyr H. Pong, Ashvin V. Nair, Laura M. Smith, Catherine Huang, and Sergey Levine. 2022. Offline Meta-Reinforcement Learning with Online Self-Supervision. In *International Conference on Machine Learning*. 17811–17829.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. 8748–8763.

[40] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. 2019. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *International Conference on Machine Learning*. 5331–5340.

[41] Yuhang Ran, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, and Yang Yu. 2023. Policy Regularization with Dataset Constraint for Offline Reinforcement Learning. In *International Conference on Machine Learning*. 28701–28717.

[42] Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. 2020. Rl-cyclegan: Reinforcement Learning Aware Simulation-to-real. In *Conference on Computer Vision and Pattern Recognition*. 11157–11166.

[43] Marc Rigter, Bruno Lacerda, and Nick Hawes. 2022. RAMBO-RL: Robust Adversarial Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 16082–16097.

[44] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering The Game of Go without Human Knowledge. *Nature* 550, 7676 (2017), 354–359.

[45] Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. 2023. Model-Bellman Inconsistency for Model-Based Offline Reinforcement Learning. In *International Conference on Machine Learning*. 33177–33194.

[46] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction (Second Edition)*. MIT press.

[47] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A Physics Engine for Model-Based Control. In *International Conference on Intelligent Robots and Systems*. 5026–5033.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[49] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Çaglar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney,

Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster Level in StarCraft II using Multi-agent Reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[50] Jianhao Wang, Jin Zhang, Haozhe Jiang, Junyu Zhang, Liwei Wang, and Chongjie Zhang. 2023. Offline Meta Reinforcement Learning with In-Distribution Online Adaptation. In *International Conference on Machine Learning*. 36626–36669.

[51] Yifan Wu, George Tucker, and Ofir Nachum. 2019. Behavior Regularized Offline Reinforcement Learning. *arXiv preprint arXiv:1911.11361* (2019).

[52] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua B. Tenenbaum, and Chuang Gan. 2022. Prompting Decision Transformer for Few-Shot Policy Generalization. In *International Conference on Machine Learning*. 24631–24645.

[53] Zhongwen Xu, Hado van Hasselt, and David Silver. 2018. Meta-Gradient Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 2402–2413.

[54] Yiqin Yang, Hao Hu, Wenzhe Li, Siyuan Li, Jun Yang, Qianchuan Zhao, and Chongjie Zhang. 2023. Flow to Control: Offline Reinforcement Learning with Lossless Primitive Discovery. In *AAAI Conference on Artificial Intelligence*. 10843–10851.

[55] Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Sergey Levine, and Chelsea Finn. 2021. Conservative Data Sharing for Multi-Task Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 11501–11516.

[56] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. 2021. COMBO: Conservative Offline Model-Based Policy Optimization. In *Advances in Neural Information Processing Systems*. 28954–28967.

[57] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. MOPO: Model-Based Offline Policy Optimization. In *Advances in Neural Information Processing Systems*. 14129–14142.

[58] Haoqi Yuan and Zongqing Lu. 2022. Robust Task Representations for Offline Meta-Reinforcement Learning via Contrastive Learning. In *International Conference on Machine Learning*. 25747–25759.

[59] Hongyu Zang, Xin Li, Jie Yu, Chen Liu, Riashat Islam, Remi Tachet des Combes, and Romain Laroche. 2023. Behavior Prior Representation learning for Offline Reinforcement Learning. In *International Conference on Learning Representations*.

[60] Hongchang Zhang, Yixiu Mao, Boyuan Wang, Shuncheng He, Yi Xu, and Xiangyang Ji. [n.d.]. In-sample Actor Critic for Offline Reinforcement Learning. In *International Conference on Learning Representations*.

[61] Hongchang Zhang, Jianzhun Shao, Shuncheng He, Yuhang Jiang, and Xiangyang Ji. 2023. DARL: Distance-Aware Uncertainty Estimation for Offline Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*. 11210–11218.

[62] Ying Zhou, Shan Cong, and Chao Yu. 2023. Task Inference for Offline Meta Reinforcement Learning via Latent Shared Knowledge. In *International Conference on Knowledge Science, Engineering and Management*. 356–365.

[63] L Zintgraf, K Shiarlis, M Igl, S Schulze, Y Gal, K Hofmann, and S Whiteson. 2020. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. In *International Conference on Learning Representations*.