Visual Goal-Directed Meta-Imitation Learning

Corban G. Rivera, David A. Handelman, Christopher R. Ratto, David Patrone, Bart L. Paulhamus
Intelligent Systems Center, Applied Physics Lab
Johns Hopkins University, Laurel, MD
corban.rivera@jhuapl.edu

Abstract

The goal of meta-learning is to generalize to new tasks and goals as quickly as possible. Ideally, we would like approaches that generalize to new goals and tasks on the first attempt. Requiring a policy to perform on a new task on the first attempt without even a single example trajectory is a zero-shot problem formulation. When tasks are identified by goal images, the tasks can be considered visually goal-directed. In this work, we explore the problem of visual goal-directed zero-shot meta-imitation learning. Inspired by several popular approaches to Meta-RL, we composed several core ideas related to task-embedding and planning by gradient descent to attempt to explore this problem. To evaluate these approaches, we adapted the Meta-world benchmark tasks to create 24 distinct visual goal-directed manipulation tasks. We found that 7 out of 24 tasks could be successfully completed on the first attempt by at least one of the approaches we tested. We demonstrated that goal-directed zero-shot approaches can translate to a physical robot with a demonstration based on Jenga block manipulation tasks using a Kinova Jaco robotic arm.

1. Introduction

When confronted with a new task, humans are able to generalize from previously learned skills and experience to perform new tasks on the first attempt. Equipping artificial intelligence (AI) agents with similar capabilities would increase their value as robotic teammates by enabling "improvisation" of problem solutions. If an AI could learn to transfer learned skills to new tasks in a way similar to humans, it would make the robotic teammate much more useful and trusted as a partner. It has been thoroughly studied that state-of-the-art AI struggles when presented with a limited number of training examples of new tasks, let alone succeeding in a "zero-shot" task with no previous experience attempting it [29]. One key challenge is learning abstractions and modularity that will allow novel tasks to be completed on the first attempt.

There are several related problems that have received a lot of attention in literature including meta-learning and meta-imitation learning which we will describe briefly.

Meta-learning Meta-learning, or "learning to learn," has a rich history in machine learning literature [3, 10, 19, 26]. Given a family of tasks, training experience for each meta-training task, and performance measures, an algorithm is capable of learning to learn if its performance improves with both additional experience and tasks.

Meta-learning generally refers to the problem of quickly adapting to new tasks given prior experience on dissimilar tasks. These methods frequently include an outer loop which is used to select parameters for an inner loop which is then used to solve the task [3, 26]. Some approaches rely on conditioning networks on a task embedding [17]. Other approaches optimize directly for their ability to fine tune the model quickly to a library of known tasks [2, 7].

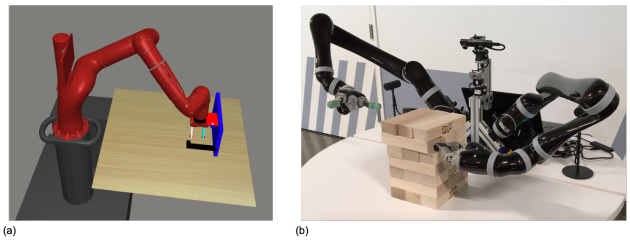


Figure 1. We investigated the problem of zero-shot visual goal-directed meta-learning with experiments making use of the (a) Metaworld ML1 set of manipulation tasks adapted for visual goal-directed task definitions, and a (b) physical demonstration making use of the Kinova Jaco manipulator on JENGA tasks.

Meta-imitation learning Imitation learning allows autonomous agents to learn a skill from demonstrations. Neural network based policies can provide a basic level of generalization to new scenarios, however training these policies may require many demonstrations. Research into few-shot learning and rapid adaptation has focused on decreasing the number of demonstrations or trials of the test-task needed to successfully perform the new task [7, 27]. In the limit, zero demonstrations of the new task are provided. This is a spe-

cial case of *zero-shot learning* in policy domain, which has become an emerging sub-field within meta-learning [12]. *Meta-imitation learning* refers to meta-learning approaches that use imitation learning during the meta-training phase.

Task Objectives given as an Image We focus on approaches that represent their goal as an image. The idea of using an image as a goal has been studied previously [4,28]. An image provides a flexible format that can be easily accommodate diverse tasks. The challenge comes from finding representations of the goal image that allow the policy to accomplish the test task. Using an image as a task goal has unique challenges as many pixel level details of the image may be irrelevant or misleading for the task. Instead, many approaches have explored latent representations of the images that are intended to capture the salient aspects of the task from the image [25,28]. To discover these latent representations, prior work has mostly focused on unsupervised objectives or objectives that are disconnected from learning the policy [8, 28].

1.1. Visual Goal-Directed Zero-shot Meta-Imitation Learning

The problem that we are solving here is distinct from the other formulations of meta-learning and meta-imitation learning in that we aim to complete a meta-test task on the first attempt assuming we are given demonstrations of the meta-training tasks and a single image of the completion of the meta-test task.

Requiring a policy to perform on a new task on the first attempt without even a single example trajectory is considered a zero-shot problem formulation. When those tasks are identified by goal images, the tasks can be considered visual goal-directed. In this work, we explore the problem of visual goal-directed zero-shot meta-imitation learning.

We explored approaches that draw inspiration from several popular approaches to Meta-RL [2, 17, 25] and reinterpret them in the context of visual goal-directed zero-shot meta-learning. We introduce approaches that combine planning by backpropagation [25] with task embedding [17].

Our experiments on the modified Metaworld [30] benchmark tasks show that the approaches we tested were able to successfully complete 7 out of 24 tasks on the first attempt [30]. The primary contributions of this paper include:

1. The introduction of visual goal-directed zero-shot meta-imitation learning as a new meta-learning problem formulation

2. An exploration of several approaches to attempt to solve the problem.

3. Quantitative evaluation of the approaches across the Metaworld series of manipulation tasks adapted for visual goal-directed task definitions.

4. A demonstration that goal-directed zero-shot approaches can translate to a physical robot

2. Related Work

Describing tasks by images has been studied by several authors [8,28]. These approaches do not focus on zero-shot goal-directed task transfer. Other work used an imitation learning objective to optimization a representation for a goal image [21, 22]. In contrast, our work focuses on both imitation learning and planning. Single-shot imitation learning via images has also been explored [13, 23, 29, 31]. Single-shot visual imitation learning approaches are related in the sense that the final image of the single demonstration could be thought of as the goal image.

Meta-learning in the context of reinforcement learning is referred to as Meta-RL [2, 5–7, 17]. These methods are related in that they evaluate the approaches on their ability to learn new tasks with as few steps from the new task as possible. They are different in that they require environments with extrinsic rewards. In contrast, the work here is based on meta-imitation learning. Although, inverse reinforcement learning [1, 14] (IRL) could be used to approximate an objective function which would allow Meta-RL approaches to be applied for learning from demonstrations. Instead of taking the IRL coupled with Meta-RL approach, In our experiments, we adapted concepts from Meta-RL approaches into the visual learning from demonstration context explicitly for the purpose of comparison with a baseline approach that conditions the policy on both a state and task-embedding [17].

2.1. Planning by Backpropagation

Planning through latent space with gradient descent is a core concept shared by several previous works, in which the same embedding is used for both the goal and the current state. The basic idea is that the distance between the latent representation of the current state and the latent representation of the goal could be exploited for planning purposes [12]. The idea can be extended by using a dynamics model to predict the latent state resulting from an action. Prior work [15, 16, 20, 24, 25] has demonstrated the value of using the dynamics model to unroll the policy over a planning horizon with the goal of minimizing the distance between the latent representations of the final predicted state and the goal.

The earliest work on planning by gradient descent was introduced decades ago [11] and was based on known model dynamics. Later work introduced planning with learned dynamics models [18]. Model-based planning was also explored for environments with discrete actions [9]. This approach relies on unsupervised pretraining. Others have explored using planning through latent space with the goal of predicting a value function [15, 20, 24]. These

approaches focus on environments with extrinsic rewards. Other work has looked at approaches for goal-directed imitation learning [16] conditioned on a sequence of images of the test task.

3. Preliminaries

We first introduce some meta-learning preliminaries, then we present the problem statement before describing the approach and implementation.

Meta-learning, or "learning to learn," aims to complete new tasks with very little (if any) training data. To achieve this, a meta-learning algorithm is first pretrained on a repertoire of tasks $\mathcal{T}_i$ in a meta-train phase. The algorithm is then evaluated on a disjoint task $\mathcal{T}_j$. As part of meta-training, the meta-learning algorithm can learn generalizable structure between tasks that allow it to successfully complete the meta-test task.

A task $\mathcal{T}_i$ is a finite-horizon Markov decision process (MDP), $\{S, A, r_i, P_i, H\}$ with state space S given as images, action space A, reward function $r_i : S \times A \rightarrow \mathcal{R}$, dynamics $P_i(s_{t+1}|s_t, a_t)$, and horizon H. It is assumed that the environment dynamics and goals may vary across tasks.

We explore a subset of reward specification $\mathcal{R}$ corresponding to the mean squared error $||f_\phi(s_t) - f_\phi(s_g)||_2$ between state $s_t \in S$ at time t, a goal $s_g \in S$, and image embedding function f_ϕ with parameters ϕ.

3.1. Problem Statement

The goal is to meta-train an agent such that it is more successful at performing a new test task $\mathcal{T}_j$ on the first attempt given a goal g_j for test task j presented as an image.

Meta-train: The agent observes and learns from $m = 1, \ldots, M$ task demonstrations.

Meta-test: The pretrained agent is given a new goal g_j that corresponds to the goal for the test task $\mathcal{T}_j$.

3.2. Universal Planning Networks

The goal of this meta-learning approach is to directly optimize for plannable representations [25]. The approach is inspired by other well known approaches like MAML [7] that optimize a model for an ability to quickly fine tune for a new task. The UPN approach is illustrated in Figure 2. The model produces an action plan $\hat{a}_{t:t+H}$ conditioned on initial and goal observations o_t and o_g, where $\hat{a}_t$ denotes the predicted action at time t over horizon H. Both the encoder f_ϕ and the combined policy and dynamics model g_θ are approximated with neural networks and are fully differentiable. The inner-loop unrolls predictions of intermediate states $x_t : x_{t+H}$ and actions $\hat{a}_{t:t+H}$ over the horizon H. The objective of the inner loop is to minimize the distance between the latent representation of the goal x_g and the final predicted state x_{t+H} using the Huber loss. Updates to the model using this loss result in updated actions $\hat{a}_{t:t+H}$. In the learning from demonstrations scenario, the outer loop uses a behavior cloning objective that minimizes the distance between the predicted action $\hat{a}_t$ for time t and the given action a_t over all timepoints using a mean squared error loss. The planning horizon H and the number of inner-loop updates U are hyperparameters for the approach. In our experiments, hyperparameters are set such that the predicted state embedding at the end of the horizon is as close as possible to the embedding of the goal.

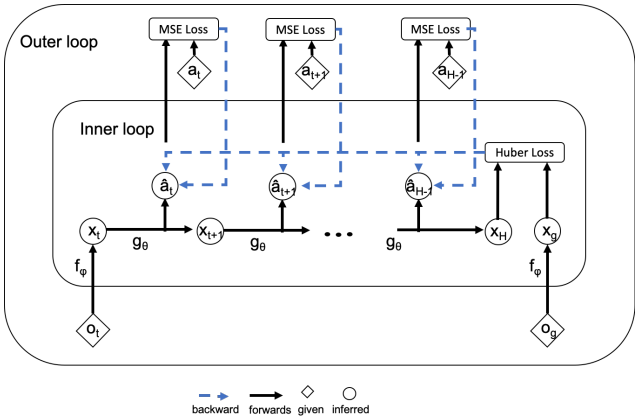


Figure 2. Universal planning networks architecture. The architecture can be described as an inner loop and an outer loop. The objective of the inner loop is to minimize the distance in latent space between the final rolled out state prediction and the latent representation of the goal. For learning from demonstrations, the outer loop is the behavior cloning objective which minimizes the distance between the predicted and observed actions.

3.3. Task Embedding for Meta-learning

Task embedding in reactive networks has been explored for meta-learning [16, 17]. A recent approach called PEARL [17] extends the typical policy formation $p(a|s)$ by additionally conditioning on a latent representation of the task $p(a|s, x_g)$. An additional KL-divergence objective is used to organized the task embedding into a desired distribution. This approach was demonstrated in the context of reinforcement learning. We adapted this approach to learn from demonstration by adopting a behavior-cloning objective. In experiments, we refer to this baseline approach as (TE-BC) for task embedding behavior cloning.

3.4. Approaches

We explored several approaches inspired by popular approaches to Meta-RL and reformulated to perform visual goal-directed zero-shot meta learning. We extended ideas from universal planning networks illustrated in Figure 2 illustrated in Figure 3(b) and visual task-embedding illustrated in Figure 3(a). Task-embedding is integrated by modifying the combined policy and dynamics model by addi-

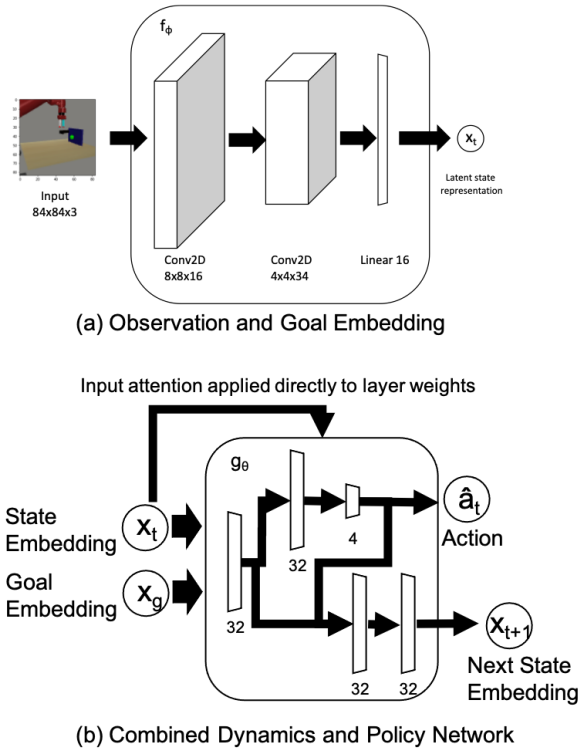


Figure 3. (a) Network architecture of the observation and goal embedding, (b) Structure of the combined policy and dynamics model. A key structural difference between this model and an MLP architecture is the introduction of attention conditioned on the input that attenuates the weights of the model directly.

tionally conditioning on x_g as shown in Figure 3(b). The structure of combined dynamics and policy network has a shared trunk that passes the image embedding to a linear layer. After the shared trunk, the network splits into dynamics and policy branches. The policy branch is composed of linear layers outputs the number of continuous actions. The dynamics branch is composed of two linear layers and is also conditioned on the action prediction. In our experiments with Metaworld, the number of continuous actions is 4 and the internal linear layers are 32 units wide.

4. Results

We designed experiments to answer the following questions: (i) Are approaches inspired by Meta-RL able to be repurposed to explore visual goal-directed meta-imitation learning, and(ii) Can goal-directed zero-shot meta-imitation learning be demonstrated on a physical robot platform?

4.1. Methods for comparison

We compared to two reactive imitation learning approaches and two planning approaches along with a random control. We compared to a behavior cloning approach (BC) that made use of the policy architecture described in Figure 3. We compared to an approach adapted from PEARL [17] to work with visual-goal directed tasks.

Additionally, we compare to two approaches that plan over a finite horizon by gradient decent. The first approach is based on the Universal Planning Network (UPN) [25]. We also compare to an approach that extends UPN with task embedding (TE-UPN) as described earlier. To evaluate both planning approaches on equal terms, the planning horizon was set to 5 steps and a single backpropagation update step was used by each planning approach. Both planning based approaches were evaluated based on the strategy of model predictive control described by Srinivas et al. [25]. We planned actions over a fixed horizon at each time step but only took the first action in the plan.

All methods we tested make use of the same deep network architecture where possible to control for the number of parameters and network structure. Additionally, both the observation and the goal image encoded using the same convolutional network described in Figure 3(a).

4.2. Metaworld tasks

We structured our experiements around 24 distinct visiomotor tasks derived from the Metaworld benchmark [30]. Two example tasks are shown in Figure 4. The tasks contain different objects with different affordances. Additionally, each task contains selectable sub-task variation. The original metaworld benchmarks were designed for multitask and meta-RL with vector representations of state and goals. We adapted the metaworld benchmark tasks to make them suitable for approaches that learn from demonstration with visual observations and goals. The introduction of visual observations add additional complexity that are not part of the existing Metaworld experiments.

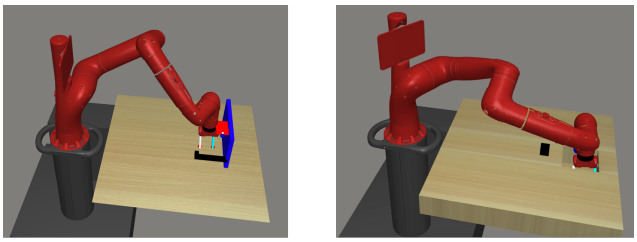


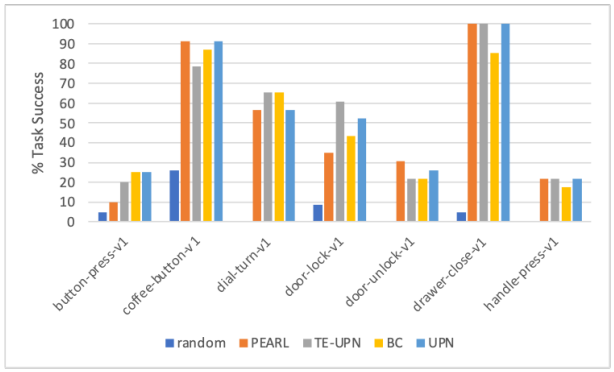
Figure 4. Metaworld is a manipulation benchmark that includes 50 distinct tasks for multi-task learning and meta-RL experiments. The tasks include distinct objects and affordances. We adapted the metaworld benchmark tasks to create 24 zero-shot meta-learning from visual demonstration tasks for evaluation. The use of visual observations and goals is more challenging than the existing metaworld benchmark formulations which are based on coordinate-based observations and goals

4.3. Experiment design

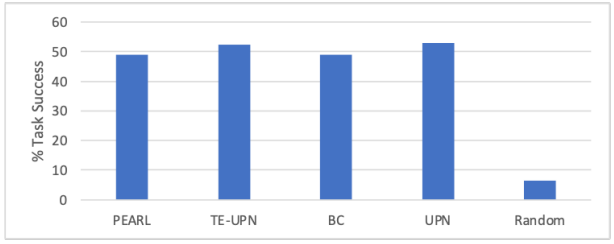
We adopted a cross-validation structure for the experiment. Specifically, to evaluate the first attempt success rate on a meta-test task, the meta-training set consisted of all other tasks.

Training data For each of the 24 tasks, we generated 100 successful demonstrations using heuristic control. For each trial, the position of key objects was randomized within preconfigured bounds. Trajectories in the dataset consisted of 84x84x3 images of each state including the final state which was treated as the goal image. Each approach was trained via Adam optimizer for 50 epochs. We made use of MSE and Huber loss functions as shown in Figure 2.

Evaluation For the task being evaluated, we measured average task success rate over 20 trials. We controlled the random seed to ensure that each approach was evaluated based on the same distribution of task variants.



(a)



(b)

Figure 5. First attempt success rate for several meta-learning and baseline approaches based on reactive policies and planning. (a) Average % task success organized by task. (b) Average % task success averaged over tasks

The results of the experiments separated by task are summarized in Figure 5(a). We found that 7 of the 24 evaluation tasks had at least one approach that was able to complete the task on the first attempt. The approaches were most successful for the coffee button and drawer-close tasks. Nearly all of the approaches we tested achieved 50% success in the dial-turn and door-lock task. For the door-lock task, TE-UPN and UPN achieved over 50% success on the first attempt, while the other reactive approaches performed close to chance. For the drawer-close, dial-turn, coffee-button tasks all of the approaches performed well above chance. In Figure 5(b), we averaged % task success over all tasks by approach. The planning approaches narrowly improved over the behavior cloning baseline and PEARL, and all the approaches that we trained outperformed the random control.

4.4. JENGA domain on a physical robot

We wanted to test the ability of TE-UPN to extrapolate to novel goal targets on physical hardware. Kinova Jaco arms and an oversized Jenga tower were used for the experiment. The task was to demonstrate the ability to poke target blocks in the tower in positions that were were offset along an axis not seen during training.

Heuristic behaviors were used to create a dataset of 21 trajectories of the JACO arm poking a block on the lower row from different starting positions. Figure 6 illustrates the experimental setup. In this experiment, TE-UPN was reconfigured to be conditioned on a block position goal represented as a three dimensional vector. The horizon length and number of planning updates were selected to minimize the deviation between the predicted endpoint of the trajectory and the target block position. In Figure 6 and the supplementary video, we show that the approach was able to plan towards multiple targets that offset along an axis not seen during training.

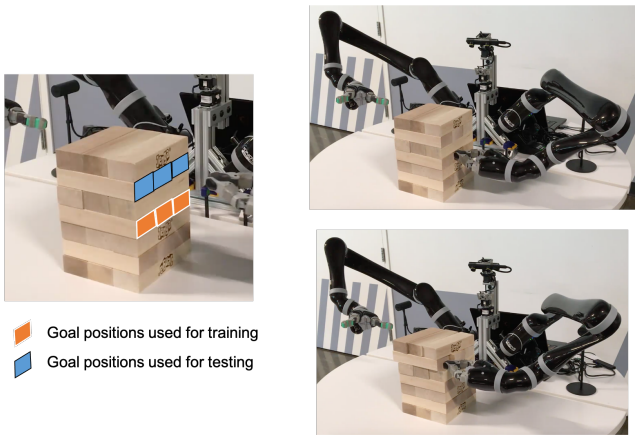


Figure 6. Physical demonstration of planning towards novel goals in Jenga. TE-UPN was trained using trajectories of Jaco arm reaching towards blocks on the lower level. TE-UPN was used to create a trajectory for a target block that were offset along an axis that was not varied during training.

5. Discussion

Of the tasks that were solvable on the first attempt, we were surprised to find that there was no clear superior approach among the approaches that we tested. The best approach frequently varied by task. Occasionally, basic reactive approaches fared as well as more sophisticated planning-based approaches. We found that the TE-UPN approach outperformed the other approaches in the door-lock task.

In the Jenga experiment, we found that TE-UPN was successful at planning towards goals that were not seen during training. This experiment highlights the ability of the TE-UPN approach to extrapolate to new goal targets not seen during training. The demonstration represents extrapolation because the goal targets were offset along an axis that was not in the training data.

We believe that this work takes steps towards understanding the challenges of this difficult zero-shot problem formulation. Still, 17 out of 24 Metaworld-adapted tasks could not be completed with any approach that we tested. This indicates that there is a lot of opportunity for future work.

6. ACKNOWLEDGEMENTS

The authors would like to thank Chace Ashcraft, I-Jeng Wang, and Marie Chau for technical and manuscript feedback.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004. 2

[2] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. *arXiv preprint arXiv:2002.09571*, 2020. 1, 2

[3] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, volume 2. Univ. of Texas, 1992. 1

[4] Koichiro Deguchi and Isao Takahashi. Image-based simultaneous control of robot and target object motions by direct-image-interpretation method. In *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289)*, volume 1, pages 375–380. IEEE, 1999. 2

[5] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016. 2

[6] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning, 2016. 2

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017. 1, 2, 3

[8] Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 512–519. IEEE, 2016. 2

[9] Mikael Henaff, William F Whitney, and Yann LeCun. Model-based planning in discrete action spaces. *arXiv preprint arXiv:1705.07177*, 2017. 2

[10] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001. 1

[11] Henry J Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960. 2

[12] Kara Liu, Thanard Kurutach, Christine Tung, Pieter Abbeel, and Aviv Tamar. Hallucinative topological memory for zero-shot visual planning. *arXiv preprint arXiv:2002.12336*, 2020. 2

[13] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2146–2153. IEEE, 2017. 2

[14] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000. 2

[15] Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. In *Advances in Neural Information Processing Systems*, pages 6118–6128, 2017. 2

[16] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros, and Trevor Darrell. Zero-shot visual imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2050–2053, 2018. 2, 3

[17] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340, 2019. 1, 2, 3, 4

[18] Jürgen Schmidhuber. An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In *1990 IJCNN international joint conference on neural networks*, pages 253–258. IEEE, 1990. 2

[19] Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pages 199–226. Springer, 2007. 1

[20] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019. 2

[21] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 486–487. IEEE, 2017. 2

[22] Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. *arXiv preprint arXiv:1612.06699*, 2016. 2

[23] Kyriacos Shiarlis, Markus Wulfmeier, Sasha Salter, Shimon Whiteson, and Ingmar Posner. Taco: Learning task decomposition via temporal alignment for control. *arXiv preprint arXiv:1803.01840*, 2018. 2

[24] David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, et al. The predictron: End-to-end learning and planning. In *International Conference on Machine Learning*, pages 3191–3199. PMLR, 2017. 2

[25] Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Universal planning networks. *arXiv preprint arXiv:1804.00645*, 2018. 2, 3, 4

[26] Sebastian Thrun and Lorien Pratt. *Learning to Learn: Introduction and Overview*, pages 3–17. Springer US, Boston, MA, 1998. 1

[27] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 1

[28] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pages 2746–2754, 2015. 2

[29] Danfei Xu, Suraj Nair, Yuke Zhu, Julian Gao, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Neural task programming: Learning to generalize across hierarchical tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018. 1, 2

[30] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100, 2020. 2, 4

[31] Allan Zhou, Eric Jang, Daniel Kappler, Alex Herzog, Mohi Khansari, Paul Wohlhart, Yunfei Bai, Mrinal Kalakrishnan, Sergey Levine, and Chelsea Finn. Watch, try, learn: Meta-learning from demonstrations and reward. *arXiv preprint arXiv:1906.03352*, 2019. 2