Few-Shot Goal Inference for Visuomotor Learning and Planning

Annie Xie, Avi Singh, Sergey Levine, Chelsea Finn
University of California, Berkeley

Abstract: Reinforcement learning and planning methods require an objective or reward function that encodes the desired behavior. Yet, in practice, there is a wide range of scenarios where an objective is difficult to provide programmatically, such as tasks with visual observations involving unknown object positions or deformable objects. In these cases, prior methods use engineered problem-specific solutions, e.g., by instrumenting the environment with additional sensors to measure a proxy for the objective. Such solutions require a significant engineering effort on a per-task basis, and make it impractical for robots to continuously learn complex skills outside of laboratory settings. We aim to find a more general and scalable solution for specifying goals for robot learning in unconstrained environments. To that end, we formulate the few-shot objective learning problem, where the goal is to learn a task objective from only a few example images of successful end states for that task. We propose a simple solution to this problem: meta-learn a classifier that can recognize new goals from a few examples. We show how this approach can be used with both model-free reinforcement learning and visual model-based planning and show results in three domains: rope manipulation from images in simulation, visual navigation in a simulated 3D environment, and object arrangement into user-specified configurations on a real robot.

Keywords: goal specification, learning rewards, reinforcement learning, meta-learning

1 Introduction

Reinforcement learning and planning methods assume some form of objective or reward function that encodes the desired outcome or behavior. There is a range of tasks where such an objective is challenging for humans to convey to robots, such as vision-based control with unknown object positions and manipulation of deformable objects, among a number of others. Even for relatively simple skills such as pouring or opening a door, prior works have hand-designed mechanisms to measure a proxy for the objective. This means that even task-agnostic methods such as reinforcement learning still require task-specific engineering to learn a task. In contrast to robots, humans can very quickly infer and understand task goals. This ability to mentally represent an objective and what it means to accomplish a task is a critical aspect of autonomously learning complex skills, as it is the driver of learning progress. If we aim to build robots that can autonomously learn new skills in real-world environments, where external feedback comes rarely, then we must develop robots that can build an internal understanding of its goals and a general mechanism for humans to convey these goals.

One simple approach for specifying tasks is to provide an image of the goal [1, 2, 3, 4, 5, 6], or more generally, provide an observation of one instance of success. There are a number of challenges with this approach, such as measuring the distance between the current and goal observation; but perhaps most saliently, we would like to not only encode a single instance of success, but reason about the entire space of successful behavior and generalize the high-level goal to new scenarios. To encode such goals, we can learn a reward function [7, 8, 9, 10, 11] or success classifier [12, 13, 14] that operates on the robot's observations. Rewards and classifiers can be learned respectively through inverse reinforcement learning from demonstrations and supervised learning from positive and negative examples. Yet, these methods do not solve the entire problem, as they require a considerable amount of data for acquiring an objective for a single task.

2nd Conference on Robot Learning (CoRL 2018), Zurich, Switzerland.

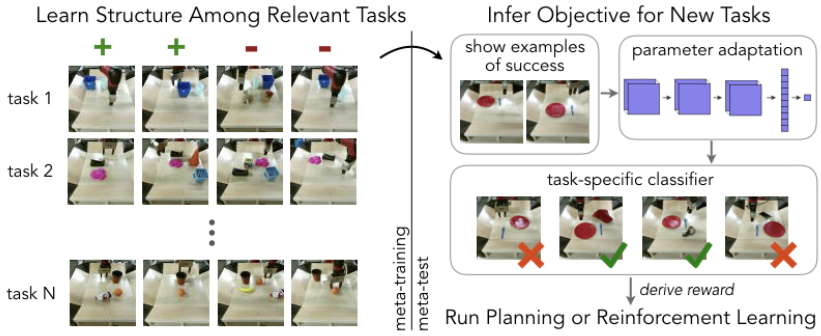


Figure 1: We propose a framework for quickly specifying visual goals. Our goal classifier is meta-trained with positive and negative examples for diverse tasks (left), which allows it to meta-learn that some factors matter for goals (e.g., relative positions of objects), while some do not. At meta-test time, this classifier can learn goals for new tasks from a couple of examples of success (right - the goal is to place the fork to the right of the plate). The reward can be derived from the learned goal classifier for use with planning or reinforcement learning.

If we would like robots to autonomously learn to perform a wide range of tasks, then it is impractical to provide many examples for training a classifier or reward for each and every task. Humans can grasp the goal of a task from just one or two examples of success. Can we allow robots to do the same? If we reuse data from a range of other tasks that we might want a robot to learn, then learning goal metrics for new tasks can be done much more efficiently. In this paper, we consider the problem of learning goal classifiers for new tasks with only a few examples by reusing data from other tasks.

Our main contribution is a framework for specifying goals to robots with minimal human supervision for each individual task. Our approach, illustrated in Figure 1, leverages recent work in meta-learning to enable few-shot acquisition of goal classifiers, and can provide an objective for various control optimizations, including visual model-based planning and model-free reinforcement learning. In our evaluation, we find that our approach provides a means to quickly learn objectives for simulated visual navigation, manipulation of deformable objects in simulation, and multi-stage vision-based manipulation tasks on a real Sawyer robot.

2 Related Work

Specifying goals is a challenge for many real-world robotics and reinforcement learning tasks. Robotic learning methods often sidestep this issue and instead hand-engineer task objectives [15, 16], often using manually instrumented environments (e.g., a thermal camera or scale to evaluate pouring [17, 18, 19], mocap to track a pancake [20], and an accelerometer on a door handle [21]). We propose a more general and scalable framework for specifying a goals that does not require manual instrumentation or shaping, and instead learns to represent goals using the robot's sensors that are used to complete the task.

A number of works have proposed to specify vision-based tasks using an image or visual representation of the goal [1, 2, 3, 4, 22, 5, 23, 6]. Unlike these works, we aim to acquire a classifier that can effectively recognize successful task executions that may not directly match an image of the goal. This enables us to recognize goals that are more abstract than an entire goal visual scene, such as relative positions of objects, approximate shapes of deformable objects, or disjunctions. Other works have sought to learn objectives by training a classifier [12, 13, 14] or reward function [10, 24, 11], including in the framework of inverse RL [25, 7, 8, 26, 27, 9, 28, 29, 30]. However, training a classifier from scratch per task has a number of challenges: modern vision methods require large training sets, and generally require both positive and negative examples. Providing many examples is onerous, and requiring users to provide negative examples is time-consuming and counter-intuitive to the average user. We aim to address both of these issues by considering the fact that we ultimately care about learning objectives of many different skills – we can use meta-learning to share data across tasks, such that only a modest amount of data is needed for any individual task, and learn how to learn a task objective from only a handful of positive examples.

Our work builds upon the ideas of learning-to-learn [31, 32, 33] and few-shot learning [34, 35, 36, 37, 38, 39]. Our general objective of quickly conveying goals to robots is related to that of prior works on one-shot imitation learning [40, 41, 42] and recent work on meta-inverse reinforcement learning [43, 44], both of which strive to learn behavior from one or a few demonstrations either

through direct imitation or by first learning a reward. Unlike these methods, our aim is to acquire goal definitions from example observations without access to a full demonstration. That is, our method only requires examples of *what* successful completion of the task looks like, rather than *how* to do it. In addition to these imitation and inverse RL methods, meta-learning has been used for control in the context of reinforcement learning [45, 46, 38]. Our work focuses on fast objective learning, which could be used in conjunction with meta-reinforcement learning for acquiring behavior quickly from a few examples of success.

3 Generalized Framework for Few-Shot Goal Inference

We aim to develop a framework that makes it easy to specify objectives for new tasks in a way that (1) reduces the manual engineering efforts for specifying a new task to a robot, making it fast and easy to convey goals to robots and (2) is applicable to a wide range of problems and robot learning methods. Since we will be using camera images, we will focus on problems with visually discernible goals, such as object manipulation and visual navigation tasks; but in principle, sensors other than cameras could be used.

Assuming we ultimately want to convey many different tasks to a robot, we consider a multi-task problem setting where we have a modest number of success/failure examples for a large number of tasks. We will use this data for meta-training a classifier such that, at test time, we can learn a goal classifier for a new task from only a few examples of success. By doing this, we minimize the amount of data needed for any particular task and make it possible to easily and quickly convey the goal of any new task. Following the meta-learning literature [47], we will consider the most general notion of a task, which can encapsulate a different objective, a different domain, a different environment, or a combination thereof. For full generality and minimal task-specific or domain-specific engineering, we need to be able to evaluate the objective using the same observation that the robot uses to solve the task, rather than external sensors or privileged information that is not available during deployment. To satisfy this requirement, we will learn success classifiers that operate directly on the robot's observation space. With these two design decisions in mind, we will next formalize the general learning problem and discuss our high-level solution that uses meta-learning.

3.1 Problem Set-up

Formally, we consider a goal classifier $\hat{y} = f(\mathbf{o})$, where $\mathbf{o}$ denotes the robot's observation, such as a camera image, and $\hat{y} \in [0, 1]$ indicates the predicted probability of the observation being of a successful outcome of the task. Such a classifier can be used for specifying the goal to reinforcement learning or planning algorithms, by deriving a reward function from the classifier's predictions; we discuss this in more detail in Section 4.2. Our aim is to learn a goal classifier from a few positive examples of success for a new task $\mathcal{T}_j$, as positive examples of success are intuitive for a human to provide and, in a sense, are the minimal piece of information needed to convey a task goal. Hence, we will be given a dataset $\mathcal{D}_j^+$ of K positive examples of success for a task $\mathcal{T}_j$: $\mathcal{D}_j := \{(\mathbf{o}_k, 1) | k = 1...K\}_j$, and our goal is to infer a classifier for the conveyed task. How might we go about learning to infer goal classifiers for new tasks from only K positive examples? To do this, we will explicitly train a model for the ability to infer goal classifiers for a wide range of previous tasks, $\{\mathcal{T}_i\}$. In particular, we assume a small dataset $\mathcal{D}_i$ for each task $\mathcal{T}_i$, where each dataset consists of both examples of success and not success: $\mathcal{D}_i := \{(\mathbf{o}_n, y_n) | n = 1...N\}_i$.

One natural question at this point is: what can the model learn from the meta-training set that allows it to infer goals for new tasks more effectively than learning each task from scratch? If each task involves a completely distinct visual concept, without any shared structure across tasks, then it seems unlikely that the model will acquire useful knowledge from meta-training. However, practical real-world goals often share many patterns: object rearrangement tasks depend strongly on relative positioning of objects to each other, and are agnostic to the pose of the robot. Tasks that involve placing objects in containers depend on whether or not an object is inside the container, but not the container's position. By extracting such patterns from meta-training tasks, our method should be able to acquire structurally related meta-test tasks efficiently.

3.2 Meta-learning for Few-Shot Goal Inference

To solve the above learning problem, we propose to learn how to learn classifiers for a task from a few positive examples of that task, by using full supervision at the meta-level from both positive

Algorithm 1 Few-Shot Learning of Objectives (FLO)

Require: for each task $\mathcal{T}_i$, a dataset of example successes and failures: $\mathcal{D}_i := \{(\mathbf{o}_n, y_n)\}_i \ \forall\, i$

randomly initialize learner parameters θ
while not done **do**
  Sample training task $\mathcal{T}_i$ (or minibatch)
  Sample examples of success $\mathcal{D}_i^+ \sim \mathcal{D}_i$
  Sample test examples $\mathcal{D}_i^{\text{test}} \sim \mathcal{D}_i \backslash \mathcal{D}_i^+$
  Learn goal classifier from positive examples $g_i(\cdot) = f_{\text{L}}(\mathcal{D}_i^+, \cdot; \theta)$
  Update learner parameters θ according to Eq. 1 using $\mathcal{D}_i^{\text{test}}$

Algorithm 2 FLO test-time

Require: examples of success $\mathcal{D}_j^+$ for new task $\mathcal{T}_j$
Require: learned θ
Infer goal classifier: $g_j(\cdot) = f_{\text{L}}(\mathcal{D}_j^+, \cdot; \theta)$
Run RL or planning, using reward/cost derived from g_j

and negative examples. Then, at test time, we can effectively learn a classifier from only positive examples. Across all of the tasks in the set of training tasks $\{\mathcal{T}_i\}$, we will train a learner f_{L} to learn a goal classifier g_i from a dataset of positive examples $\mathcal{D}_i^+$ and make predictions about new observations $\mathbf{o}$:

$$g_i(\mathbf{o}) = f_{\text{L}}(\mathcal{D}_i^+, \mathbf{o}; \theta)$$

where we use $\mathcal{D}_i^+$ to denote a dataset of K examples sampled uniformly from the positive examples in $\mathcal{D}_i$. To train the meta-learner parameters θ, we will optimize the learned classifier for its ability to accurately classify new examples in $\mathcal{D}_i$, both positive and negative. In particular, we will optimize the following objective:

$$\min_\theta \sum_i \sum_{(\mathbf{o}_n, y_n) \in \mathcal{D}_i^{\text{test}}} \mathcal{L}(y_n, g_i(\mathbf{o}_n)) = \min_\theta \sum_i \sum_{(\mathbf{o}_n, y_n) \in \mathcal{D}_i^{\text{test}}} \mathcal{L}(y_n, f_{\text{L}}(\mathcal{D}_i^+, \mathbf{o}_n; \theta)) \tag{1}$$

where $\mathcal{D}_i^+$ is defined as above, $\mathcal{D}_i^{\text{test}}$ is sampled uniformly from $\mathcal{D}_i \backslash \mathcal{D}_i^+$, and the loss function $\mathcal{L}$ is the standard cross entropy loss that compares the classifier's predictions to the labels. The datapoints $\mathcal{D}^{\text{test}}$ are distinct from $\mathcal{D}^+$ so that we train for good generalization. At test time, after learning f_{L}, we are presented with examples of success $\mathcal{D}_j^+$ for a new task $\mathcal{T}_j$. We can use this data infer a task-specific goal classifier $g_j(\cdot) = f_{\text{L}}(\mathcal{D}_j^+, \cdot; \theta)$. This classifier provides an objective, or part of an objective (as we discuss in the next section), for reinforcement learning or planning. We refer to this approach as few-shot learning of objectives (FLO). The algorithms underlying the meta-training optimization and test-time procedure are outlined in Algorithms 1 and 2.

As discussed by Finn and Levine [48], the view of meta-learning as learning the mapping $f_{\text{L}}(\mathcal{D}, \mathbf{o}; \theta) \rightarrow \hat{y}$ is general to a number of different meta-learning algorithms, including recurrent models [35], learned optimizers [37], and gradient-based methods [38]. Hence, this framework can be combined with any of such meta-learning algorithms for few-shot classifier learning.

4 Few-Shot Goal Inference for Learning and Planning

Having presented the general framework of few-shot goal inference, we will discuss our particular meta-learning implementation, mechanisms needed to mitigate exploitation of the learned objective by the controller, and mechanisms for specifying compound tasks by joining classifiers.

4.1 Concept Acquisition for Goal Classifiers

While the above framework is general to any meta-learning approach, we would use a method that can efficiently learn to learn, to avoid collecting very large amounts data for meta-training. As a result, we choose to build upon model-agnostic meta-learning (MAML) [38], which incorporates the structure of gradient descent for efficient meta-learning. In particular, MAML learns an initial parameter setting θ of the model f_{MAML} such that one or a few steps of gradient descent with a few examples leads to parameters θ' that generalize well to new examples from that task. Grant et al. [49] extended MAML for learning new concepts from only positive examples, referred to as concept acquisition through meta-learning (CAML), akin to how humans learn new concepts. We directly apply CAML to the setting of acquiring binary success classifiers from positive examples. At test time, the learner uses gradient descent to adapt the meta-learned parameters θ to a dataset of positive

examples $\mathcal{D}_j^+$ for task $\mathcal{T}_j$:

$$g_j(\mathbf{o}) = f_{\text{L}}(\mathcal{D}_j^+, \mathbf{o}; \theta) = f_{\text{CAML}}\big(\mathbf{o}; \theta_j'\big) = f_{\text{CAML}}\big(\mathbf{o}; \theta - \alpha \nabla_\theta \sum_{(\mathbf{o}_n, y_n) \in \mathcal{D}_j^+} \mathcal{L}(y_n, f_{\text{CAML}}(\mathbf{o}_n; \theta))\big)$$

where $\mathcal{L}$ is the cross-entropy loss function, α is the step size, and θ_j' denotes the parameters updated through gradient descent on task $\mathcal{T}_j$. We only write out one gradient descent step for convenience of notation; in practice, more may be used. Then, meta-training takes into account this gradient descent adaptation procedure, training for the initial parameters as follows:

$$\min_\theta \sum_i \sum_{(\mathbf{o}_n, y_n) \in \mathcal{D}_i^{\text{test}}} \mathcal{L}(y_n, f_{\text{CAML}}(\mathbf{o}_n; \theta_i'))$$

We optimize this objective using Adam [50] on the initial parameters θ. In our experiments, we consider vision-based tasks where $\mathbf{o}$ is an RGB image. We use a learner f_{CAML} that is represented by a convolutional neural network with RGB image inputs. We provide details on the architecture and other implementation details in Section 5.

4.2 Deriving Rewards from Classifier Predictions

After meta-learning, a goal classifier can be inferred for a new task $\mathcal{T}_j$ by running one gradient descent step from the initial parameters θ with respect to a few examples of success, $\mathcal{D}_j^{\text{test}}$. The inferred goal classifier predicts the probability of an observation $\mathbf{o}$ being successful at performing task $\mathcal{T}_j$. To use this prediction as an objective for planning or reinforcement learning, we need to convert it into a reward function. One simple approach would be to treat the probability of success as the reward for that observation. We find that this works reasonably well, but that the predictions produced by the neural network are not always well-calibrated. To reduce the effect of false positives and mis-calibrated predictions, we use the classifier conservatively by thresholding the predictions so that reward is only given for confident successes. Below this threshold, we give a reward of 0 and above this threshold, we provide the predicted probability as the reward. The threshold is chosen via cross-validation on a set of validation tasks.

4.3 Cascading Classifiers for Compound Tasks

To provide objectives for more complex tasks, we can join multiple classifiers. For example, if we train a classifier to recognize if a particular relative position of two objects is achieved, then we can use multiple classifiers to achieve a particular configuration of multiple objects, like a table setting. To achieve this, we provide a few positive examples $\mathcal{D}_j^+$ of each task $\mathcal{T}_j$ that we would like to achieve, and then infer the classifier for each task. To perform the sequence of tasks, we cascade the inferred classifiers – iteratively setting each objective, and running the planner or policy for each one until the classifier indicates that the subtask has been completed. In our experiments, we illustrate how this cascading technique can be used to maneuver three objects into a desired configuration.

5 Experiments

To study the generality of our approach, we evaluate our learned objectives with both vision-based planning and reinforcement learning methods, with an emphasis on tasks that have visually nuanced goals that are difficult to specify manually. Our experiments focus on object arrangement, rope manipulation, and visual navigation tasks, but our approach is suitable for any domain where a few successful observations can sufficiently communicate the task. Code for training the few-shot success classifier and videos of our results are available on the supplementary website[1].

Since our goal is to provide an easy way to compute an objective for new tasks from a few observations of success for that task, we compare our approach to a few alternative and prior methods for doing so under the same assumptions as our method:

Pixel distance: Given one observation of success, a naive metric is to measure the ℓ_2 distance between the current observation and the successful observation. Will this can sometimes perform well

[1]The supplementary website is at https://sites.google.com/view/few-shot-goals

Figure 3: Object arrangement performance of our method with distractor objects and with two tasks. The left shows a subset of the 5 positive examples that are provided for inferring the goal classifier(s), while the right shows the robot executing the specified task(s) via visual planning.

with low-dimensional observations, this approach cannot capture invariances that a goal classifier can represent, nor does it scale well to visual or other high-dimensional sensory observation spaces.

Latent space distance: An alternative approach is to measure the distance between current and goal observations in a learned latent space [3, 4]. For learning the latent space, we train an autoencoder [51, 4] on the meta-training data used for our approach. We expect this metric to be better grounded than distances in pixel space, but still cannot capture invariances across and within goals.

Oracle: In our rope domain, we derive an objective using the ground truth state of the environment, which is generally not available in the real world. This provides an upper bound on performance.

For all tasks in our evaluation, we consider the 5-shot learning setting, where 5 examples of success are provided to method for conveying the goal of the test task. For both the pixel and latent space distance metrics above, we take the minimum distance across these five examples, allowing the optimization to try to match the closest goal observation. Full experimental details and neural network architecture information can be found in Appendices A and B.

5.1 Visual Planning for Object Arrangement

We study a visual object arrangement task, where different goals correspond to different relative arrangements of a pair of objects. For this setting, we use a Sawyer robot and use the planning method developed by Ebert et al. [52] to optimize actions with respect to the learned objective. This planning approach learns a video prediction model from self-supervised data, and uses a sampling-based optimization with iterative replanning (MPC) to select sequences of actions at each timestep that lead to desirable futures. We evaluate our learned classifier on the predictions made by the video prediction model and derive the cost used for planning using the approach described in Section 4.2.

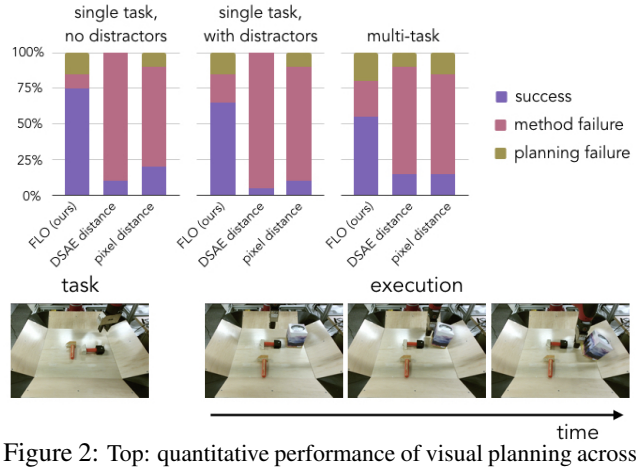


Figure 2: Top: quantitative performance of visual planning across different goal specification methods: ours, DSAE [4], and pixel error. Where possible, we include break down the cause of failures into errors caused by inaccurate prediction or planning and those caused by an inaccurate goal classifier. Bottom: visualization of the most common failure case — the presence of novel distractor objects. More data or data augmentation with a wider range of distractors could help mitigate this source of failure.

We evaluate all approaches in three different experimental settings. In the first setting, the goal is to arrange two objects into a specified relative arrangement. The second setting is the same, but with distractor objects present. In the final, most challenging setting, the goal is to achieve two tasks in sequence. As described in Section 4.3, we provide positive examples for both tasks, infer the classifier for both task, perform MPC for the first task until completion, followed by MPC for the second task. To evaluate the ability to generalize to new goals and settings, we use novel, held-out objects for all of the task and distractor objects in our evaluation.

We qualitatively visualize the evaluation in Figure 3. On the left, we show a subset of the five images provided to illustrate the task(s), and on the left, we show the motions performed by the robot. We

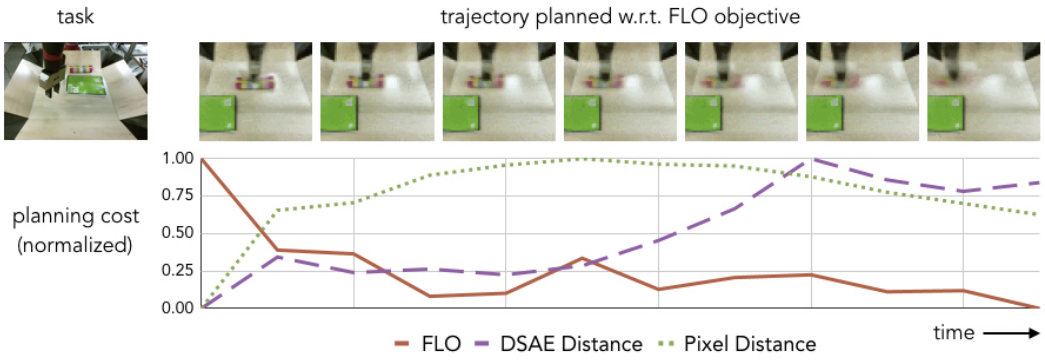


Figure 4: Comparison of the normalized cost determined by each method (bottom). We evaluate each metric on the video prediction corresponding to the trajectory planned with the FLO objective (top).

Figure 5: Rope manipulation policy roll-outs. Left: a subset of the 5 positive examples provided for inferring the goal classifier. Right: the robot executing the policy learned via RL with the learned objective.

see that the robot is able to execute motions which lead to a correct relative positioning of the objects. We quantitatively evaluate each method across 20 tasks, including 10 unique object pairs. The results, shown in Figure 2, indicate that prior methods for learning distance metrics struggle to infer the goal of the task, while our approach leads to substantially more successful behavior on average. Aiming to understand the results further and the fidelity of the inferred goal classifiers, we measure whether the failures are caused by the planner or the classifier. In particular, if the classifier predicted that the final image was a success that was not, or if the classifier classifies a correct plan as not successful, then the fault is on the classifier. Otherwise, the fault is the planner, e.g. if the planner finds an action plan that looks successful according to the video prediction model, but in reality is not. From this analysis, we observe that about half of the single task failures are caused by the planner / predictive model. This suggests at least 85% of the tasks are achievable by the planner with a suitable metric. We further analyze a common failure mode in Figure 2 and provide a direct comparison of each method's cost function in Figure 4 and Appendix C.

5.2 Rope Manipulation with Reinforcement Learning

We next study a setting where the goal is to learn to manipulate a rope of pearls into a particular shape, as specified by a few images. Here, we aim to evaluate how our objective and others can be used with the cross-entropy method [53]. For this setting, we use a parallel-jaw gripper as the manipulator, and our experiments are carried out in the Mujoco simulator [54].

We use 30 tasks for testing, with five success images used per task for inferring the goal. Example tasks are illustrated in Figure 6. For evaluation and the oracle, we need to devise a distance metric. We do so by translating the current and goal rope to have the same center of mass, and then measuring the average distance between corresponding pearls. This metric is used only by the oracle (as reward), and to provide a fast, automatic evaluation. In our results, shown in Figure 6, we first notice that pixel distance provides a reasonably good metric for the tasks. This is not all that surprising, since the simulated images are clean, with a fixed background. As shown in the previous experiments, pixel distance does not provide a good metric when using real images with novel objects, so we do not expect these results to transfer to real world settings. We also observe that FLO performs substantially better than using a distance metric derived from the latent space of an autoencoder, but does not reach the performance of the oracle distance metric. In most tasks where the policy performed poorly using FLO, we observed that RL was able to successfully optimize the objective, indicating that it was exploiting inaccuracies in the classifier. Integrating frameworks for mining negatives and iteratively retraining the objective (such as those presented in Fu et al. [30]) is an interesting direction for future work. We show a qualitative example in Figure 5.

method	median distance
pixel distance	0.54 (0.34, 0.70)
AE distance	0.59 (0.50, 0.69)
FLO (ours)	0.27 (0.18, 0.34)
oracle	0.16 (0.13, 0.23)

Figure 6: Left: Illustration of four different test tasks. Right: Median distance across rope manipulation tasks (lower is better), comparing each method for deriving a reward from images. We compare RL from full state and from visual features. Numbers in parentheses indicate the 25^{th} and 75^{th} percentiles.

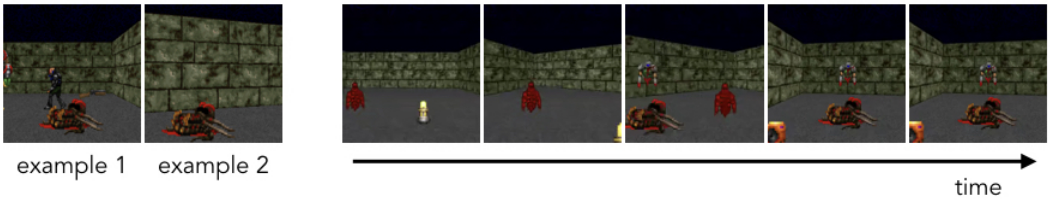
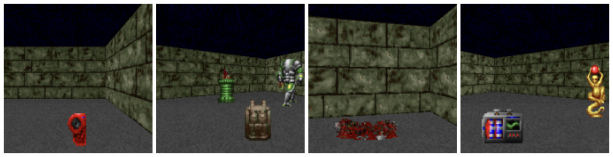


Figure 7: Visual navigation policy roll-out. Left: 2 of the 5 positive examples provided for inferring the goal classifier. Right: the agent executing the policy learned via RL using the learned objective.

5.3 Visual Navigation with Reinforcement Learning

We lastly consider a visual navigation task in VizDoom, a first-person 3D navigation environment. Here, the goal is to learn to navigate to a particular object that is specified by the few observations of success. We use DQN [55] to optimize a policy with respect to the learned reward functions. The observation space for the robot consists of its spatial location and orientation. There are four discrete actions: move forward, turn left, turn right and no action. Each episode is 30 timesteps long, and we run the algorithm for 50K timesteps. Since we found it difficult to engineer a success metric for this task, evaluations are performed by a human. For the same reason, we do not compare to an oracle that receives the true success metric as supervision, and we only compare to using pixel and latent space distance methods described above. For evaluation, a trial is successful if the robot visits a state in which the target object is in the lower third of the agent's field of view.

We evaluate each method across 9 tasks, with 5 trials for each task for a total of 45 trials. We illustrate examples of tasks in Figure 8. The reward is inferred from 5 examples of success provided for each task. Because of the limited number of objects available in VizDoom, we do not evaluate on held-out target objects, but instead evaluate on novel combinations of target and distractor objects in new configurations. As seen in Figure 8, our method is able to learn a success metric and use it to navigate to an object 77.8% of the time, while the alternative distance metrics have much more difficulty encoding the correct task, achieving 33.0% and 44.0% success respectively. We show a qualitivate example in Figure 7.



method	success rate
pixel distance	33.0%
AE distance	44.0%
FLO (ours)	77.8%

Figure 8: Left: Illustration of four different test tasks. Right: Success rate of each method evaluated on the test visual navigation tasks. We compare our method to autoencoder and pixel distance metrics for deriving the reward for reinforcement learning.

6 Conclusion

In this work, we proposed a framework for learning to learn task objectives from a few examples of success, motivated by the challenge of specifying goals in real world settings. We discussed how to derive a reward function from our inferred goal classifier, and how our task goal classifiers can be combined for forming more complex, compound goals. Finally, we showed how our framework for inferring goals can be combined with both reinforcement learning and planning for vision-based skills such as maneuvering a rope into a given shape in simulation, navigating to a specified target in a simulated 3D environment, and rearranging multiple previously-unseen objects in the real world.

Acknowledgments

We thank the anonymous reviewers who gave useful comments. This work was supported by the NSF through IIS-1651843, IIS-1614653, and IIS-1700696, an ONR Young Investigator Program Award, ARL DCIST CRA W911NF-17-2-0181, and Berkeley DeepDrive. We also acknowledge generous equipment and computational resource support from NVIDIA, Google, and Amazon.

References

[1] M. Jagersand and R. Nelson. Visual space task specification, planning and control. In *International Symposium on Computer Vision*, 1995.

[2] K. Deguchi and I. Takahashi. Image-based simultaneous control of robot and target object motions by direct-image-interpretation method. In *International Conference on Intelligent Robots and Systems (IROS)*, 1999.

[3] M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[4] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel. Deep spatial autoencoders for visuomotor learning. In *International Conference on Robotics and Automation (ICRA)*, 2016.

[5] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, 2017.

[6] A. Srinivas, A. Jabri, P. Abbeel, S. Levine, and C. Finn. Universal planning networks. In *International Conference on Machine Learning (ICML)*, 2018.

[7] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

[8] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.

[9] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning (ICML)*, 2016.

[10] P. Sermanet, K. Xu, and S. Levine. Unsupervised perceptual rewards for imitation learning. *Robotics: Science and Systems (RSS)*, 2017.

[11] H.-Y. F. Tung, A. W. Harley, L.-K. Huang, and K. Fragkiadaki. Reward learning from narrated demonstrations. *arXiv:1804.10692*, 2018.

[12] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *International Conference on Robotics and Automation (ICRA)*, 2016.

[13] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[14] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning hand-eye coordination for robotic grasping with large-scale data collection. In *International Symposium on Experimental Robotics (ISER)*, 2016.

[15] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research (JMLR)*, 2016.

[16] A. A. Rusu, M. Vecerik, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot Learning (CoRL)*, 2017.

[17] C. Schenck and D. Fox. Visual closed-loop control for pouring liquids. In *International Conference on Robotics and Automation (ICRA)*, 2017.

[18] C. Schenck and D. Fox. Guided policy search with delayed sensor measurements. *arXiv:1609.03076*, 2016.

[19] A. Yamaguchi, C. G. Atkeson, and T. Ogasawara. Pouring skills with planning and learning modeled from human demonstrations. *International Journal of Humanoid Robotics (IJHR)*, 2015.

[20] P. Kormushev, S. Calinon, and D. G. Caldwell. Robot motor skill coordination with em-based reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2010.

[21] A. Yahya, A. Li, M. Kalakrishnan, Y. Chebotar, and S. Levine. Collective robot reinforcement learning with distributed asynchronous guided policy search. In *International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[22] A. Edwards, C. Isbell, and A. Takanishi. Perceptual reward functions. *arXiv:1608.03824*, 2016.

[23] A. D. Edwards, S. Sood, and C. L. Isbell Jr. Cross-domain perceptual reward functions. *arXiv:1705.09045*, 2017.

[24] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[25] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.

[26] A. Boularias, J. Kober, and J. Peters. Relative entropy inverse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[27] M. Kalakrishnan, P. Pastor, L. Righetti, and S. Schaal. Learning objective functions for manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2013.

[28] M. Wulfmeier, D. Z. Wang, and I. Posner. Watch this: Scalable cost-function learning for path planning in urban environments. In *International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[29] N. Rhinehart and K. M. Kitani. First-person activity forecasting with online inverse reinforcement learning. *International Conference on Computer Vision (ICCV)*, 2017.

[30] J. Fu*, A. Singh*, D. Ghosh, L. Yang, and S. Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[31] S. Thrun and L. Pratt. *Learning to learn*. Springer Science & Business Media, 1998.

[32] J. Schmidhuber. Evolutionary principles in self-referential learning. *Diploma thesis, Institut f. Informatik, Tech. Univ. Munich*, 1987.

[33] S. Hochreiter, A. Younger, and P. Conwell. Learning to learn using gradient descent. *International Conference on Artificial Neural Networks (ICANN)*, 2001.

[34] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015.

[35] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning (ICML)*, 2016.

[36] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[37] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. *International Conference on Learning Representations (ICLR)*, 2017.

[38] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning (ICML)*, 2017.

[39] T. Munkhdalai and H. Yu. Meta networks. *International Conference on Machine Learning (ICML)*, 2017.

[40] Y. Duan, M. Andrychowicz, B. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba. One-shot imitation learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[41] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine. One-shot visual imitation learning via meta-learning. In *Conference on Robot Learning (CoRL)*, 2017.

[42] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, and S. Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *Robotics: Science and Systems (RSS)*, 2018.

[43] K. Xu, E. Ratner, A. Dragan, S. Levine, and C. Finn. Learning a prior over intent via meta-inverse reinforcement learning. *arXiv:1805.12573*, 2018.

[44] A. Gleave and O. Habryka. Multi-task maximum entropy inverse reinforcement learning. *arXiv:1805.08882*, 2018.

[45] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv:1611.02779*, 2016.

[46] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. In *CogSci*, 2017.

[47] C. Finn. *Learning to Learn with Gradients*. PhD thesis, University of California, Berkeley, 2018.

[48] C. Finn and S. Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *International Conference on Learning Representations (ICLR)*, 2018.

[49] E. Grant, C. Finn, J. Peterson, J. Abbott, S. Levine, T. Griffiths, and T. Darrell. Concept acquisition via meta-learning: Few-shot learning from positive examples. In *NIPS Workshop on Cognitively-Informed Artificial Intelligence*, 2017.

[50] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[51] S. Lange, M. Riedmiller, and A. Voigtlander. Autonomous reinforcement learning on raw visual input data in a real world application. In *International Joint Conference on Neural Networks (IJCNN)*, 2012.

[52] F. Ebert, C. Finn, A. X. Lee, and S. Levine. Self-supervised visual planning with temporal skip connections. In *Conference on Robot Learning (CoRL)*, 2017.

[53] R. Rubinstein and D. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Information Science and Statistics. Springer New York, 2013.

[54] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[55] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

Appendix

A Architecture Details

Our model f_{CAML} is represented by a convolutional neural network with RGB image inputs. The network consists of three convolutional layers with 32 3×3 filters and stride 2, each followed by layer normalization and a ReLU non-linearity. A spatial soft-argmax operation extracts spatial feature points from the final convolution layer. These features are passed through one fully-connected layers with 50 units and ReLU non-linearities, followed by a linear layer to the two-dimensional softmax output. The architecture in our simulated experiments is the same, except with 16 filters in each convolution layer, feature flattening instead of a spatial soft-argmax, and three fully-connected layers instead of one. When using real images, the first convolutional layer is initialized with weights from VGG-16.

For the autoencoder, we use a convolutional neural network with three layers and 3x3 filters, where the layers have 64, 32, and 16 filters respectively. The stride is 2 for the first layer, and 1 for subsequent layers. We follow this up with three fully connected layers, that have 200, 100 and 50 units respectively. We train this autoencoder on the entire meta-training dataset, and our target reconstructions have dimension 32x32x3. We run reinforcement learning on top of the features from the last hidden layer of this autoencoder, and keep the autoencoder weights fixed during the policy learning process.

B Experimental Details

B.1 Object Arrangement

To collect data for meta-training the classifier, we randomly select a pair of objects from our set of training objects, and position them into many different relative positions, recording the image for each configuration. One task corresponds to a particular relative positioning of two objects, e.g. the first object to the left of the second, and we construct positive and negative examples for this task by labeling the aforementioned images. We randomly position the arm in each image, as it is not a determiner of task success. A good objective should ignore the position of the arm. We also include randomly-positioned distractor objects in about a third of the collected images. Some of the meta-training data is illustrated in the left of Figure 1. In total, we acquire data for 122 tasks, with roughly 25 positive and 45 negative examples for each task. We use more negatives than positives to account for the larger space of non-goal examples. As some images are shared across tasks, the total number of unique images in the dataset is 5490.

B.2 Rope Manipulation

The tasks (i.e. goal shapes of the rope in this case) are generated through the following automated procedure: the manipulator takes random actions, some of which cause the rope to move. If any single pearl gets displaced by more than 10 cm from its original position, we freeze the rope position and call it a new task. We then displace the rope all over the table, randomly select a color for the rope from a fixed set of 5 colors, and add small perturbations to it to generate multiple examples for this task. This encodes the fact that we only care about the shape of the rope, and not its color nor its position on the table, and small perturbations of the shape are acceptable. We use 360 tasks for training with 30 positives per task. Negative examples for each task are defined using positive examples of other tasks. We also collect a pool of 1000 negative examples by taking random actions and share these examples across all tasks, leading to a total of 11800 unique images being used for training.

B.3 Visual Navigation

We generate the data to meta-train the classifier as follows: we randomly select an object to be used as the target object, initialize the target object in a random position, and place 4 randomly chosen distractor objects in different randomized positions in the room. To collect positive examples, we place the agent between 70 and 90 units away from the object with an orientation between -30 and

30 degrees with respect to the target object and record the resulting image observation as a positive example. Negative examples for each task are collected by initializing the agent 100 to 200 units away from the object with an orientation between -30 and 30 degrees. We repeat this to generate 50 positive and 50 negative examples for each of 100 tasks. We additionally use positive examples from other tasks as negative examples for a particular task, and collect a general pool of 200 negative examples shared across all tasks. For the general pool of negatives, we gather examples by taking random actions and by collecting examples facing walls and corners. In total, there are 10200 unique images in our dataset.

C Comparison of Planning Costs

To further compare FLO to alternative success metrics, we plot the planning cost determined by each metric on trajectories planned with respect to each objective in Figures 9, 10, and 11.

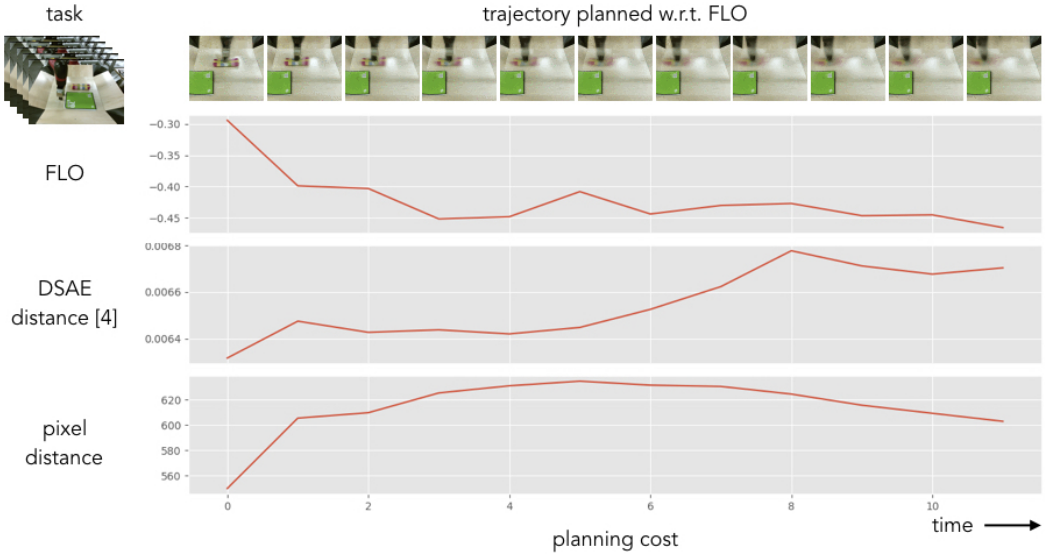


Figure 9: Trajectory proposed with FLO as the planning cost. With our method, the planner is able to discover a trajectory that enables the robot to successfully complete the specified task. For the same trajectory, the costs determined by alternative metrics do not decrease as dramatically despite each frame getting closer to success.

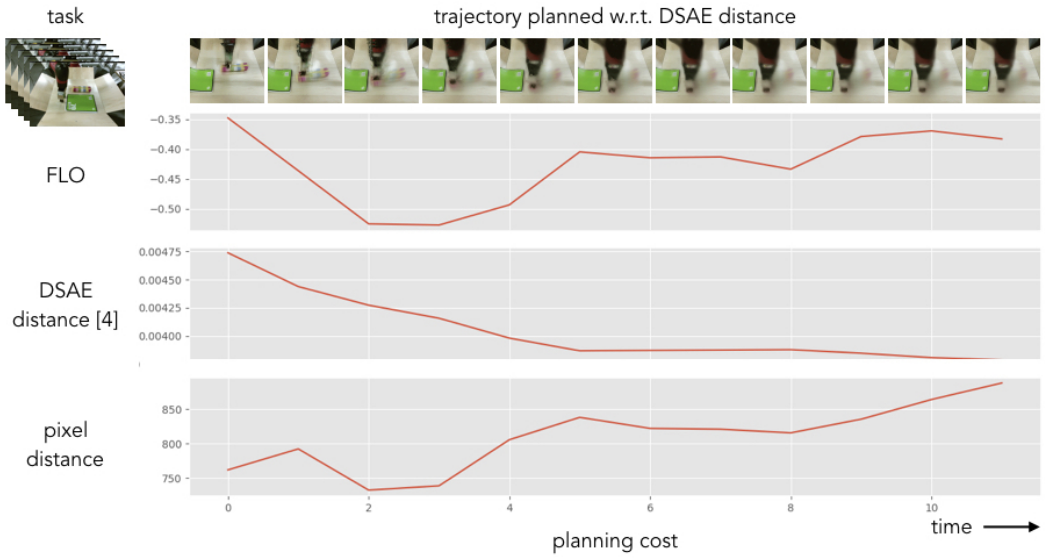


Figure 10: Trajectory proposed with DSAE distance as the planning cost. The planner using the DSAE distance tries to match the arm's position in the given example of success on the left and ignores the task entirely.

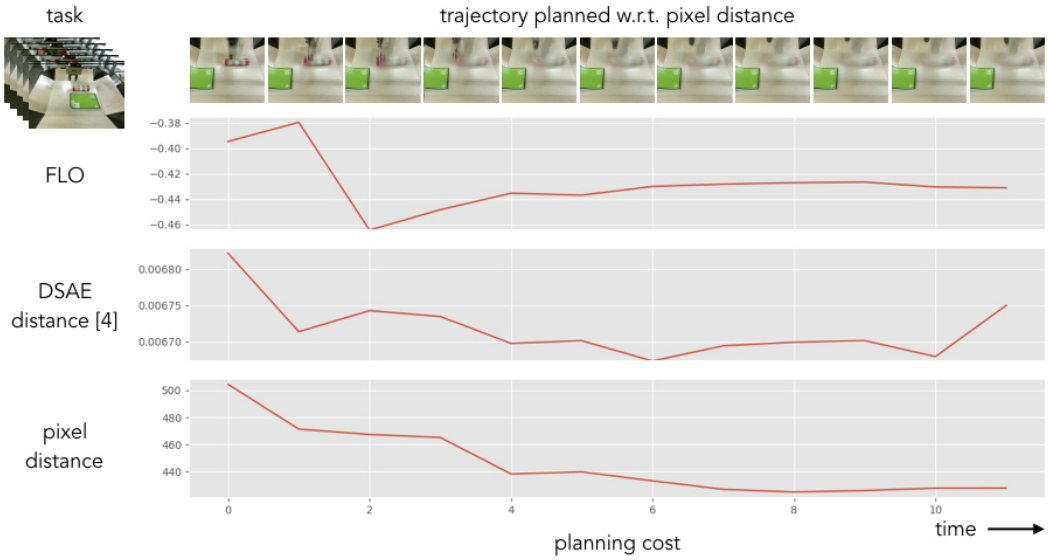


Figure 11: Trajectory proposed with pixel distance as the planning cost. The pixel distance metric also mostly focuses on the position of the arm. Instead of performing the specified task, the planner tries to move the arm to match the given example of success on the left and lower its pixel cost.