

Deep Spatial Autoencoders for Visuomotor Learning

Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, Pieter Abbeel

Abstract—Reinforcement learning provides a powerful and flexible framework for automated acquisition of robotic motion skills. However, applying reinforcement learning requires a sufficiently detailed representation of the state, including the configuration of task-relevant objects. We present an approach that automates state-space construction by learning a state representation directly from camera images. Our method uses a deep spatial autoencoder to acquire a set of feature points that describe the environment for the current task, such as the positions of objects, and then learns a motion skill with these feature points using an efficient reinforcement learning method based on local linear models. The resulting controller reacts continuously to the learned feature points, allowing the robot to dynamically manipulate objects in the world with closed-loop control. We demonstrate our method with a PR2 robot on tasks that include pushing a free-standing toy block, picking up a bag of rice using a spatula, and hanging a loop of rope on a hook at various positions. In each task, our method automatically learns to track task-relevant objects and manipulate their configuration with the robot’s arm.

I. INTRODUCTION

One of the fundamental challenges in applying reinforcement learning to robotic manipulation tasks is the need to define a suitable state space representation. Typically, this is handled manually, by enumerating the objects in the scene, designing perception algorithms to detect them, and feeding high-level information about each object to the algorithm. However, this highly manual process makes it difficult to apply the same reinforcement learning algorithm to a wide range of manipulation tasks in complex, unstructured environments. What if a robot could automatically identify the visual features that might be relevant for a particular task, and then learn a controller that accounts for those features? This would amount to automatically acquiring a vision system that is suitable for the current task, and would allow a range of object interaction skills to be learned with minimal human supervision. The robot would only need to see a glimpse of what task completion looks like, and could then figure out on its own how to change the scene into that configuration.

Directly learning a state space representation from raw sensory signals, such as images from a camera, is an active area of research [1], [2], and while considerable progress has been made in recent years [3], [4], applications to real robotic systems remain exceedingly difficult. The difficulties stem from two challenges. First, learning a good representation with unsupervised learning methods, such as deep learning, often requires a very large amount of data [5]. Second, arbitrary learned visual representations do not always lend themselves well to control. We address these challenges by

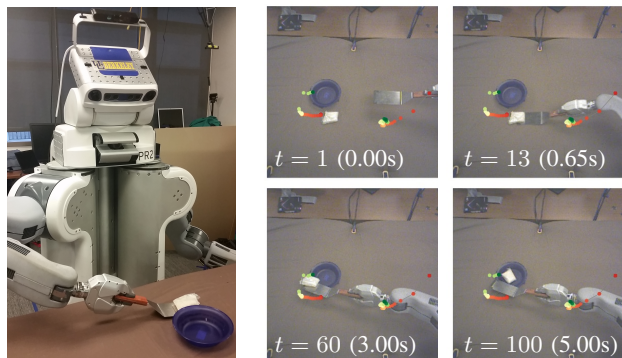


Fig. 1: PR2 learning to scoop a bag of rice into a bowl with a spatula (left) using a learned visual state representation (right).

using a spatial autoencoder architecture to learn a state representation that consists of feature points. Intuitively, these feature points encode the configurations of objects in the scene, and the spatial autoencoder that we describe provides for data-efficient learning by minimizing the number of non-convolutional parameters in the encoder network. Since our learned feature points correspond to object coordinates, this architecture also addresses the second challenge, since real-valued quantities such as positions are more amenable to control than the types of discrete or sparse features that are more commonly learned with deep learning methods [6].

In fact, our experiments show that our learned feature point representation can be used effectively in combination with an efficient trajectory-centric reinforcement learning algorithm. This method iteratively fits time-varying linear models to the states visited at the previous iteration [7], estimating how the robot’s actions (which correspond to joint torques) affect the state. When the state includes visual feature points, this method can learn how the robot’s actions affect the objects in the world, and the trained controllers can perform closed-loop control on the configuration of these objects, allowing the robot to move and manipulate them. Furthermore, this reinforcement learning algorithm can easily be combined with guided policy search to learn complex, nonlinear policies that generalize effectively to various initial states [8], which we demonstrate by training a nonlinear neural network policy on top of our learned representation.

Our main contribution is a method for learning vision-based manipulation skills by combining unsupervised learning using deep spatial autoencoders with simple, sample-efficient trajectory-centric reinforcement learning. We demonstrate that this approach can be used to learn a variety of manipulation skills that require “hand-eye” coordination, including pushing a free-standing toy block, scooping objects into a bowl, using a spatula to lift a bag of

rice from the table (shown in Figure 1), and hanging a loop of rope on hooks at various positions. Each of these tasks are learned using the same algorithm, with no prior knowledge about the objects in the scene, and using only the onboard sensors of the PR2 robot, which consist of joint encoders and an RGB camera. We also compare our approach to a number of baselines that are representative of previously proposed visual state space learning methods.

II. RELATED WORK

Reinforcement learning and policy search methods have been applied to perform a variety of robotic tasks, such as table tennis [9], object manipulation [10], [11], and locomotion [12], [13], [14], [15]. One key challenge to using RL successfully is handling high-dimensional sensory observations needed to perform sensorimotor control. A standard approach is to hand-design low-dimensional features from the observations; however, this requires manual task-specific engineering, which undermines the goal of using an automatic learning-based approach. Other approaches instead use function approximation to replace the value function in standard RL algorithms like TD-learning [16] and approximate Q-learning [17]. These approaches typically require large amounts of time and data to learn the parameter space. In this work, we use the approach of first learning a low-dimensional state representation from observations via unsupervised learning, and then use a data-efficient RL algorithm with the learned state representation to learn robotic manipulation tasks involving perception.

Prior work has proposed to learn representations using general priors from the physical world [2] or the ability to predict future observations [18], [19], [20]. Representation learning has been applied to control from high-dimensional visual input in several recent works. Lange et al. [1] use autoencoders to acquire a state space for Q-learning and evaluate on simple real-world problems with 2 to 3 dimensions. Several methods also train autoencoders that can predict the next image, effectively learning dynamics [3], [21], though such methods have only been demonstrated on toy problems with synthetic images. The quantity of data required to apply deep learning methods to real-world images is known to far exceed that of toy tasks [5], and prior methods generally avoid this issue by using synthetic images. We also use an autoencoder, but we use a spatial architecture that allows us to acquire a representation from real-world images that is particularly well suited for high-dimensional continuous control. Our architecture is also data-efficient and can handle relatively small datasets. We demonstrate our method on a range of real-world robotic tasks.

Other methods have been proposed to train policies from high-dimensional observations, without first acquiring a low-dimensional state space [22], [23]. However, running these methods on a real robotic platform requires either impractical amounts of data [22] or an instrumented training setup that provides knowledge about relevant objects at training time [23]. Our method does not require this knowledge, which enables it to learn tasks such as pushing, tossing, and

scooping without using any additional instrumentation. In fact, our method has no knowledge of which objects make up the scene at all. These tasks cannot be learned using our prior method [23] without mechanisms such as motion capture.

The goal of our method is similar to that of visual servoing, which performs feedback control on image features [24], [25], [26]. However, our controllers and visual features are learned entirely from real-world data, rather than being hand-specified. This gives our method considerable flexibility in how the visual input can be used and, as a result, the tasks that it can learn to perform. Furthermore, our approach does not require any sort of camera calibration, in contrast to many visual servoing methods (though not all – see e.g. [27], [28]).

Most deep learning architectures focus on semantic representation learning, e.g. [29]. Some spatial network architectures have been proposed for tasks such as video prediction and dynamics modeling [30], [21]. The architecture that we use for the encoder in our unsupervised learning method is based on a spatial softmax followed by an expectation operation, which produces a set of spatial features that correspond to points of maximal activation in the last convolutional layer. We previously proposed this type of architecture in the context of policy search for visuomotor policies [23]. However, in our prior work, this type of network was trained with supervised learning, regressing directly to joint torques. Obtaining these joint torque targets requires being able to solve the task, which requires a representation that is already sufficient for the task. In this paper, we show how the representation can be learned from scratch using unsupervised learning, which extends the applicability of this architecture to a much wider range of tasks where a suitable representation is not known in advance.

III. PRELIMINARIES

In this paper, we are primarily concerned with the task of learning a state representation for reinforcement learning (RL) from camera images of robotic manipulation skills. However, we consider this problem in the context of a specific RL algorithm that is particularly well suited to operate on our learned state representations. We describe the RL method we use in this background section, while the Section V describes our representation learning method. The derivation in this section follows prior work [7], but is repeated here for completeness.

A. RL with Local, Time-Varying Linear Models

Let \mathbf{x}_t and \mathbf{u}_t be the state and action at time step t . The actions correspond to motor torques, while states will be learned using the method in the next section. The aim of our RL method is to minimize the expectation $E_{p(\tau)}[\ell(\tau)]$ over trajectories $\tau = \{\mathbf{x}_1, \mathbf{u}_1, \dots, \mathbf{x}_T, \mathbf{u}_T\}$, where $\ell(\tau) = \sum_{t=1}^T \ell(\mathbf{x}_t, \mathbf{u}_t)$ is the total cost, and the expectation is under $p(\tau) = p(\mathbf{x}_1) \prod_{t=1}^T p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)p(\mathbf{u}_t|\mathbf{x}_t)$, where $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$ is the dynamics distribution and $p(\mathbf{u}_t|\mathbf{x}_t)$ is the controller that we would like to learn.

Algorithm 1 RL with linear-Gaussian controllers

-
- 1: initialize $p(\mathbf{u}_t|\mathbf{x}_t)$
 - 2: **for** iteration $k = 1$ to K **do**
 - 3: run $p(\mathbf{u}_t|\mathbf{x}_t)$ to collect trajectory samples $\{\tau_i\}$
 - 4: fit dynamics $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$ to $\{\tau_j\}$ using linear regression with GMM prior
 - 5: fit $p = \arg \min_p E_{p(\tau)}[\ell(\tau)]$ s.t. $D_{\text{KL}}(p(\tau) \|\bar{p}(\tau)) \leq \epsilon$
 - 6: **end for**
-

The controller can be optimized using a variety of model-based and model-free methods [31]. We employ a trajectory-centric algorithm that optimizes time-varying linear-Gaussian controllers, which can be thought of as a trajectory with time-varying linear stabilization [7]. While linear-Gaussian controllers are simple, they admit a very efficient optimization procedure that works well even under unknown dynamics. This method is summarized in Algorithm 1. At each iteration, we run the current controller $p(\mathbf{u}_t|\mathbf{x}_t)$ on the robot to gather N samples ($N = 5$ in all of our experiments), then use these samples to fit time-varying linear-Gaussian dynamics of the form $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t) = \mathcal{N}(f_{\text{xt}}\mathbf{x}_t + f_{\text{ut}}\mathbf{u}_t + f_{\text{ct}}, \mathbf{F}_t)$. This is done by using linear regression with a Gaussian mixture model prior, which makes it feasible to fit the dynamics even when the number of samples is much lower than the dimensionality of the system [7]. We also compute a second order expansion of the cost function around each of the samples, and average the expansions together to obtain a local approximate cost function of the form

$$\ell(\mathbf{x}_t, \mathbf{u}_t) \approx \frac{1}{2}[\mathbf{x}_t; \mathbf{u}_t]^T \ell_{\mathbf{xu}, \mathbf{xu}}[\mathbf{x}_t; \mathbf{u}_t] + [\mathbf{x}_t; \mathbf{u}_t]^T \ell_{\mathbf{xu}} + \text{const.}$$

When the cost function is quadratic and the dynamics are linear-Gaussian, the optimal time-varying linear-Gaussian controller of the form $p(\mathbf{u}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{K}_t\mathbf{x}_t + \mathbf{k}_t, \mathbf{C}_t)$ can be obtained by using the LQR method. This type of iterative approach can be thought of as a variant of iterative LQR [32], where the dynamics are fitted to data. In our implementation, we initialize $p(\mathbf{u}_t|\mathbf{x}_t)$ to be a random controller that is centered around the initial state with low-gain PD feedback, in order to avoid unsafe configurations on the real robot.

One crucial ingredient for making this approach work well is to limit the change in the controller $p(\mathbf{u}_t|\mathbf{x}_t)$ at each iteration, since the standard iterative LQR approach can produce a controller that is arbitrarily far away from the previous one, and visits parts of the state space where the fitted dynamics are no longer valid. To that end, our method solves the following optimization problem at each iteration:

$$\min_{p(\mathbf{u}_t|\mathbf{x}_t)} E_{p(\tau)}[\ell(\tau)] \text{ s.t. } D_{\text{KL}}(p(\tau) \|\bar{p}(\tau)) \leq \epsilon,$$

where $\bar{p}(\tau)$ is the trajectory distribution induced by the previous controller. Using KL-divergence constraints for controller optimization was proposed in a number of prior works [33], [34], [35]. In the case of linear-Gaussian controllers, we can use a modified LQR algorithm to solve this problem. We refer the reader to previous work for details [7].

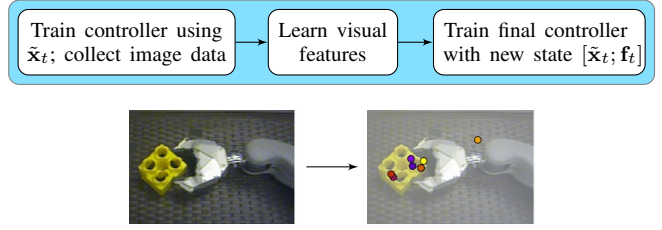


Fig. 2: RL with deep spatial autoencoders. We begin by training a controller without vision to collect images (left) that are used to learn a visual representation (center), which is then used to learn a controller with vision (right). The learned state representation corresponds to spatial feature points (bottom).

B. Learning Nonlinear Policies with Guided Policy Search

Linear-Gaussian controllers are easy to train, but they cannot express all possible control strategies, since they essentially encode a trajectory-following controller. To extend the applicability of linear-Gaussian controllers to more complex tasks, prior work has proposed to combine it with guided policy search [7], [8], [23], which is an algorithm that trains more complex policies, such as deep neural networks, by using supervised learning. The supervision is generated by running a simpler algorithm, such as the one in the previous section, from multiple initial states, generating multiple solutions. In our experimental evaluation, we combine linear-Gaussian controllers with guided policy search to learn a policy for hanging a loop of rope on a hook at different positions. A different linear-Gaussian controller is trained for each position, and then a single neural network policy is trained to unify these controllers into a single policy that can generalize to new hook positions. A full description of guided policy search is outside the scope of this paper, and we refer the reader to previous work for details [7], [8], [23].

IV. ALGORITHM OVERVIEW

Our algorithm consists of three stages. In the first stage, we optimize an initial controller for the task without using vision, using the method described in the previous section. The state space during this stage corresponds to the joint angles and end-effector positions of the robot, as well as their time derivatives. This controller is rarely able to succeed at tasks that require keeping track of objects in the world, but it serves to produce a more focused exploration strategy that can be used to collect visual data, since an entirely random controller is unlikely to explore interesting parts of the state space. In general, this stage can be replaced by any reasonable exploration strategy.

In the second stage, the initial controller is used to collect a dataset of images, and these images are then used to train our deep spatial autoencoder with unsupervised learning, as described in Section V. Once this autoencoder is trained, the encoder portion of the network can be used to produce a vector of feature points for each image that concisely describes the configuration of objects in the scene. The final state space is formed by concatenating the joint angles, end-effector positions, and feature points, and also including their velocities to allow for dynamic tasks to be learned. We must

also define a cost function using this new state space, which we do by presenting an image of the target state to the feature encoder and determining the corresponding state.

In the third stage, a vision-based controller is trained using the new state space that now contains visual features from the encoder, with the new cost function defined in terms of the visual features and the robot’s configuration. This controller is trained using the same trajectory-centric reinforcement learning algorithm as in the first stage, and is able to perform tasks that require controlling objects in the world that are identified and localized using vision.

An overview of this procedure is provided in Figure 2. In the next section, we describe our representation learning method.

V. UNSUPERVISED STATE REPRESENTATION LEARNING FROM VISUAL PERCEPTION

The goal of state representation learning is to find a mapping $\phi(I_t)$ from a high-dimensional observation I_t to a robot state representation for which it is easy to learn a control policy. We will use $\tilde{\mathbf{x}}_t$ to denote the configuration of the robot, \mathbf{f}_t to denote the learned representation, and $\mathbf{x}_t = [\tilde{\mathbf{x}}_t; \mathbf{f}_t]$ to denote the final state space that combines both components. The state of a robotic system should be consistent with properties of the physical world [2]; i.e., it should be temporally coherent and encode only simple, task-relevant information. Moreover, our RL method models the system’s state as a time-varying linear dynamical system, and thus it will perform best if the state representation acts as such. Intuitively, a good state representation should encode the poses of relevant objects in the world using Cartesian coordinates that move with the object over time. Our aim is to learn such a state representation without using human supervision or hand-crafted features. Deep learning offers powerful tools for learning representations [5], but most deep neural network models focus on semantics of *what* is in the scene, rather than *where* the objects are. In this section, we describe our autoencoder architecture which is designed to encode temporally-smooth, spatial information, with particular emphasis on object locations. As we show in our experiments, this kind of representation is well suited to act as a state space for robotic reinforcement learning.

A. Deep Spatial Autoencoders

Autoencoders acquire features by learning to map their input to itself, with some mechanism to prevent trivial solutions, such as the identity function. These mechanisms might include sparsity or a low-dimensional bottleneck. Our autoencoder architecture, shown in Figure 3, maps from full-resolution RGB images to a down-sampled, grayscale version of the input image, and we force all information in the image to pass through a bottleneck of spatial features, which we describe below. Since low-dimensional, dense vectors are generally well-suited for control, a low-dimensional bottleneck is a natural choice for our learned feature representation. A critical distinction in our approach is to modify this bottleneck so that it is forced to encode spatial

feature locations rather than feature values – the “where” rather than the “what.” This architecture makes sense for a robot state representation, as it forces the network to learn object positions; we show in Section VII, that it outperforms more standard architectures that focus on the “what,” both for image reconstruction, and for robotic control. The final state space for RL is then formed by concatenating this learned encoding, as well as its time derivatives (the feature “velocities”), with the robot’s configuration.

The first part of the encoder architecture is a standard 3-layer convolutional neural network with rectified linear units of the form $a_{cij} = \max(0, z_{cij})$ for each channel c and pixel (i, j) . We compute the spatial features from the last convolutional layer by performing a “spatial soft arg-max” operation that determines the image-space point of maximal activation in each channel of conv3. This set of maximal activation points forms our spatial feature representation and forces the autoencoder to focus on object positions. The spatial soft arg-max consists of two operations. The response maps of the third convolutional layer (conv3) are first passed through a spatial softmax $s_{cij} = e^{a_{cij}/\alpha} / \sum_{i',j'} e^{a_{ci'j'}/\alpha}$, where the temperature α is a learned parameter. Then, the expected 2D position of each softmax probability distribution \mathbf{s}_c is computed according to $\mathbf{f}_c = (\sum_i i * s_{cij}, \sum_j j * s_{cij})$, which forms the bottleneck of the autoencoder. This pair of operations typically outputs the image-space positions of the points of maximal activation for each channel of conv3. Finally, the decoder is simply a single linear (fully connected) mapping from the feature points \mathbf{f} to the down-sampled image. We found that this simple decoder was sufficient to acquire a suitable feature representation.

With this architecture, the bottleneck representation \mathbf{f} , which we refer to as learned “feature points,” encodes the positions of the learned features in the image. The autoencoder is forced to compress all of the image information through this spatial feature point representation, which will therefore be capable of directly localizing objects in the image. This means that the state representation used for control will capture the spatial organization of objects in the scene. Example learned feature points are visualized in Figure 4. Note that these feature points pick out task-relevant objects (those which move during the collection of the image data.), and learn to ignore background features.

A feature representation that can reconstruct the image clearly contains the state information captured by the camera. However, the learning algorithm must also be able to predict the dynamics of the state representation, which is not necessarily inherently possible with unconstrained learned features [2]. Thus, we add a smoothness penalty to the feature points as $\|(\mathbf{f}_{t+1} - \mathbf{f}_t) - (\mathbf{f}_t - \mathbf{f}_{t-1})\|_2^2$, which is added to the reconstruction loss term to form the optimization objective.

We optimize the auto encoder using stochastic gradient descent (SGD), and with batch normalization [36] following each of the convolutional operations. The filters of the first convolutional layer are initialized with a network trained on ImageNet [37], [38].

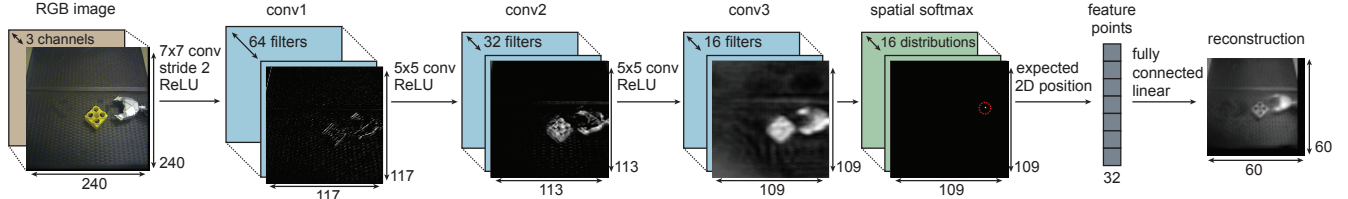


Fig. 3: The architecture for our deep spatial autoencoder. The last convolutional layer is passed through a spatial softmax, followed by an expectation operator that extracts the positions of the points of maximal activation in each channel. A downsampled version of the image is then reconstructed from these feature points.

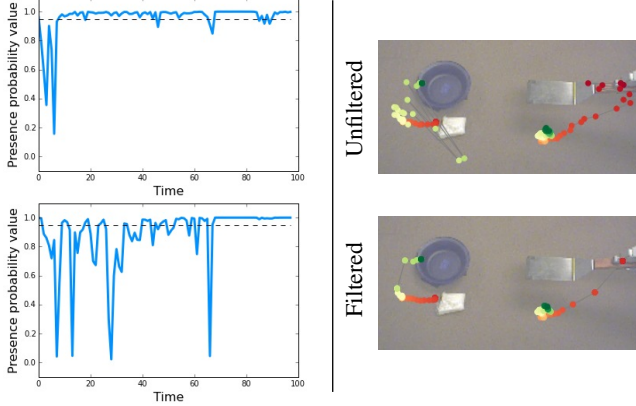


Fig. 4: Left: feature presence probability values plotted for two features, with the threshold β shown in black. Note that a range of thresholds would produce similar results. Right: two sample trajectories of learned feature points, starting from red and ending in green. The features filtered with a Kalman filter using feature presence (bottom) are much smoother than the unfiltered feature points (top)

B. Filtering and Feature Pruning Using Feature Presence

Not all of the learned feature points will encode useful information that can be adequately modeled with a time-varying linear dynamical system by our RL algorithm. Lighting and camera exposure can cause sudden changes in the image that do not reflect motion of relevant objects. However, the autoencoder will assign some subset of its features to model these phenomena to improve its reconstruction, producing noisy feature points that make it difficult to train the controller. Additionally, even useful features can periodically become occluded, resulting in low activations in the last convolutional layer and unpredictable peak locations.

To handle these issues, we introduce the concept of feature presence. The presence of a feature is determined by the softmax probability value s_{cij} at the pixel location of the feature point $\mathbf{f}_c = (i, j)$.¹ This measure reflects the probability mass after the softmax that is actually located at the position of the maximum activation. This measure is usually very high. As shown in Figure 4, learned feature points usually place almost all of their probability mass at \mathbf{f}_c , the location of the maximum. We therefore use a threshold on s_{cij} that marks the feature as present if $s_{cij} \geq \beta$, with $\beta = 0.95$. The positions of the feature points during a trial are filtered by a Kalman filter, which receives observations of

¹Since the feature point position may fall on a boundary between pixels, we actually sum the probability over a 3×3 window of nearby pixels.

\mathbf{f}_c only when $s_{cij} \geq \beta$, producing smooth predictions when the feature is not present. We use a second order Kalman filter and fit the parameters using expectation maximization.

In addition to filtering features that might become occluded, we also automatically determine the features that do not adequately represent task-relevant objects in the scene and prune them from the state. As discussed in Section IV, the goal positions for the feature points are defined by showing an image of the goal state to the robot. To mitigate the effects of noise, we record a short sequence of fifty images over the course of two seconds and average the resulting feature positions. We also measure the feature presence indicators during this period, and prune those points that are not present for the entire sequence, since they are unlikely to be relevant. In our evaluation, we show that this pruning step is important for removing extraneous and noisy features from the state. To validate this heuristic, we conducted a more thorough analysis of the predictability of the feature points to determine its effectiveness.

To analyze the usefulness of the learned feature points for modeling the dynamics of the scene, we define the following measure of feature *predictiveness*:

$$E \left[\sum_{t=1}^T \left(\log p(\mathbf{x}_t^{\mathcal{I}_s} | \mathbf{x}_{t-1}^{\mathcal{I}_s}, \mathbf{u}_{t-1}) + \log p(\mathbf{f}_t^{\mathcal{I} \setminus \mathcal{I}_s} | \mathbf{f}_t^{\mathcal{I}_s}) \right) \right]$$

Here, \mathcal{I} denotes the index set of all features, $\mathcal{I}_s \subset \mathcal{I}$ denotes the indices that are actually selected for inclusion in the state, $\mathbf{f}^{\mathcal{I}_s}$ is shorthand for the corresponding feature vectors, and $\mathbf{x}^{\mathcal{I}_s} = [\tilde{\mathbf{x}}_t; \mathbf{f}_t^{\mathcal{I}_s}]$ is the corresponding state space that includes the robot's configuration. We evaluate the probabilities by fitting a dynamics model to $\mathbf{x}^{\mathcal{I}_s}$ as described in Section III and fitting a linear-Gaussian model to $p(\mathbf{f}_t^{\mathcal{I} \setminus \mathcal{I}_s} | \mathbf{f}_t^{\mathcal{I}_s})$. The expectation is evaluated at samples gathered for autoencoder training. This measure captures the degree to which the state $\mathbf{x}_{t-1}^{\mathcal{I}_s}$ is predictive of the next state $\mathbf{x}_t^{\mathcal{I}_s}$, as well as the degree to which $\mathbf{x}_t^{\mathcal{I}_s}$ can predict $\mathbf{f}_t^{\mathcal{I}}$. Intuitively, a good state space can predict the features at the next step, which are optimized for predicting the entire image. This measure will discourage features with dynamics that are hard to fit, and will also encourage diversity, since multiple features on the same object will be predictable from one another according to $p(\mathbf{f}_t^{\mathcal{I} \setminus \mathcal{I}_s} | \mathbf{f}_t^{\mathcal{I}_s})$. To analyze the relative quality of each feature, we iteratively pruned out the feature that, when removed, produced the least reduction in the likelihood. This yields a predictiveness ranking of the features. We evaluated

Algorithm 2 RL with deep spatial autoencoders

- 1: train exploration controller $\tilde{p}(\mathbf{u}_t|\tilde{\mathbf{x}}_t)$ with simple states $\tilde{\mathbf{x}}_t$ and simple cost function $\ell(\tilde{\mathbf{x}}_t, \mathbf{u}_t)$ (Algorithm 1)
- 2: collect image dataset $\mathcal{D} = \{I_i\}$ by running $\tilde{p}(\mathbf{u}_t|\tilde{\mathbf{x}}_t)$ (or using images collected during step 1)
- 3: train deep spatial autoencoder using \mathcal{D} to obtain feature encoder $\phi(I_t) = \mathbf{f}_t$
- 4: select feature points $\tilde{\mathbf{f}}$ using heuristic described in Section V-B, define full state as $\mathbf{x}_t = [\tilde{\mathbf{x}}_t; \tilde{\mathbf{f}}_t; \dot{\tilde{\mathbf{f}}}_t]$
- 5: define full cost function $\ell(\mathbf{x}_t, \mathbf{u}_t)$ using images of the target state
- 6: train final controller $p(\mathbf{u}_t|\mathbf{x}_t)$ with full cost $\ell(\mathbf{x}_t, \mathbf{u}_t)$ (Algorithm 1)

this metric on the most difficult task, and found that our simple feature presence was a good, quick approximation for selecting highly-predictive features. We report the rankings, along with the chosen features in Table I. We also found that the low-ranked features are always either consistently not present or extremely noisy (see Figure 4).

Task	feature ranking (from best to worst)
Rice scooping	5 15 12 1 4 6 2 14 8 13 7 3 10 9 11 16

TABLE I: Ranking of features according to predictiveness; features chosen based on presence in the target image are shown in bold.

VI. CONTROL OVER VISUAL FEATURES

Putting everything together, our method for training vision-based controllers is outlined in Algorithm 2. We first train a controller to attempt the task without vision, using only the robot’s configuration as the state. The goal for this controller is defined in terms of the robot’s configuration, and the learning is performed using Algorithm 1. While this goal is typically inadequate for completing the task, it serves to explore the state space and produce a sufficiently varied range of images for learning a visual state representation. The representation is then learned on a dataset of images obtained from this “blind” controller using the method described in Section V-A. This produces a feature encoder $\phi(I_t) = \mathbf{f}_t$. We then train Kalman filters for each feature point and automatically prune the features using the heuristic described in Section V-B, yielding the final features $\tilde{\mathbf{f}}$ to be included in the full state $\mathbf{x}_t = [\tilde{\mathbf{x}}_t; \tilde{\mathbf{f}}_t; \dot{\tilde{\mathbf{f}}}_t]$. Since we use torque control, the robot and its environment form a second-order dynamical system, and we must include both the features $\tilde{\mathbf{f}}$ and their velocities $\dot{\tilde{\mathbf{f}}}$ in the state.

In addition to the state representation \mathbf{x}_t , we must also define a cost function $\ell(\mathbf{x}_t, \mathbf{u}_t)$ for the task. As discussed in the preceding section, we set the goal by showing the robot an image of the desired configuration of the scene, denote I_g , in addition to the target end-effector position used for the “blind” controller. The target features are then obtained according to $\mathbf{f}_g = \phi(I_g)$ and the cost depends on the distance to these target features. In practice, we use a short sequence of images to filter out noise. Having defined a state space \mathbf{x}_t

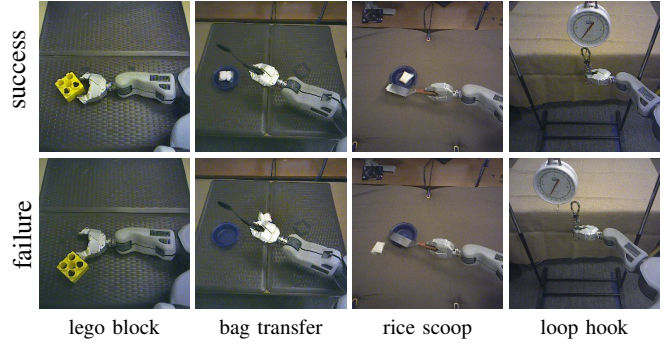


Fig. 5: Illustrations of the tasks in our experiments, as viewed by the robot’s camera, showing examples of successful completion by our controllers and a typical failure by the controller without vision.

and a cost $\ell(\mathbf{x}_t, \mathbf{u}_t)$, we use Algorithm 1 to optimize the final vision-based controllers that perform feedback control on the robot’s configuration and the visual state representation.

VII. EXPERIMENTAL EVALUATION

We evaluated our method on a range of robotic manipulation tasks, ranging from pushing a lego block to scooping a bag of rice into a bowl. The aim of these experiments was to determine whether our method could learn behaviors that required tracking objects in the world that could only be perceived with vision. To that end, we compared controllers learned by our approach to controllers that did not use vision, instead optimizing for a goal end-effector position. We also compared representations learned with our spatial autoencoder to hidden state representations learned by alternative architectures, including ones proposed in previous work.

A. Experimental Tasks

The four experimental tasks in our evaluation are shown in Figure 5. The first and simplest task requires sliding a lego block 30 cm across a table. The robot does not grasp the lego block, but rather pushes it along the table with its gripper, which requires coordinating the motion of the block and balancing it at the end of the fingers. In the second task, the robot is holding a spoon that contains a white bag, and it must use the spoon to drop the bag into a bowl. The robot must balance the bag on the spoon without dropping it, and must angle the spoon such that the bag falls into the bowl. In the third task, the robot must use a spatula to scoop a bag of rice off of the table and into a bowl. In this task, the robot must perform a quick, precise sliding motion under the bag and then lift and angle the spatula such that the rice bag slides or flips into the bowl. This task is illustrated in detail in Figure 1. If the robot does not coordinate the motion of the spatula and rice bag, the bag is pushed aside and cannot be lifted properly.

In the fourth task, we combine our method with guided policy search by training four separate linear-Gaussian controllers for hanging a loop of rope on a hook at different positions. These four controllers are combined into a single neural network policy using guided policy search, and we test the capability of this policy to then hang the rope on the hook in new positions. The robot must use vision to

locate the hook, so making use of the spatial feature points is essential for performing this task.

All of our experiments were performed using one 7 DoF arm of a PR2 robot. The initial state space $\tilde{\mathbf{x}}_t$ contains the robot’s joint angles and end effector pose, as well as their time derivatives. The controls consist of the seven torques at the joints. The images were collected using a consumer RGB camera, and the controller runs at 20 Hz. For each of the tasks, we provided the goal pose of the end-effector and an image of the goal state. For the scooping task, we also provided a sub-goal of the spatula under the bag of rice. Full details of each task and equation for the cost function are both presented in Appendix A.

B. Results and Analysis

We ran each of the learned controller 10 times and reported either the average distance to the goal, or the success rate. For the transfer and scooping task, success required placing the bag inside the bowl, while for the hanging task, success required the rope loop to hang on the hook. For each of the tasks, we compared our method to a controller that used only the robot’s configuration $\tilde{\mathbf{x}}_t$. The results, shown in Table II, show that each of the tasks requires processing visual information, and that our method is able to acquire a state that contains the information necessary for the task.

task	lego block (mean distance)	bag transfer	rice scoop	loop hook
no vision	4.63 cm	1/10	0/10	7/70
ours	0.92 cm	10/10	9/10	60/70

TABLE II: Results of our method, compared to a controller without vision that uses the robot configuration only.

For the generalization experiment with the rope loop and hook task, we trained linear-Gaussian controllers for hook positions that were spaced 8 cm apart, and trained a single neural network using guided policy search that unified these controllers into a single nonlinear policy. This nonlinear policy could succeed at each of the training positions, and was also successful at the three test positions not seen during training, as shown in Table III. The policy trained without vision was unable to localize the hook, and resorted to a random strategy, which was occasionally successful but generally failed at the task. The supplementary video shows the behavior of both the “blind” policy and the policy that used our learned visual state representation.

loop hook	training positions				test positions		
	0 cm	8 cm	16 cm	24 cm	4 cm	12 cm	20 cm
no vision	0/10	1/10	0/10	3/10	2/10	0/10	1/10
ours	10/10	8/10	9/10	8/10	8/10	10/10	7/10

TABLE III: Detailed generalization results for the rope loop task.

We also evaluate two alternative autoencoder architectures that are representative of prior work, trained with the same training and validation datasets as our model. The first comparison closely mirrors the method of Lange et al. [1], but with a bottleneck dimensionality of 10 to account for the greater complexity of our system. The second comparison reflects a more recent architectures, using max-pooling to

reduce the dimensionality of the image maps and batch normalization after the convolutional layers. The bottleneck for this architecture is 32, matching that of our architecture. Details of both baseline architectures are provided in Appendix B. We evaluated both architectures with and without a smoothness penalty. The results, shown in Table IV, show that these methods struggle with our high-dimensional, real-world tasks, despite the larger model achieving a lower reconstruction loss than our autoencoder. For several of the models, we could not obtain a stable controller, while the others did not effectively integrate visual input into the control strategy. Table IV also shows the performance of our method without the smoothness penalty, and without feature pruning, showing that both of these components are critical for obtaining good results.

lego block comparison	reconstruction loss (validation)	network training time	mean distance
AE of [1]	7.23	270 min	8.4 cm
AE of [1], smooth	7.43	420 min	9.7 cm
conv+pool AE	5.29	11 min	*24.3 cm
conv+pool AE, smooth	5.33	32 min	*30 cm
ours, no smooth	6.07	27 min	5.8 cm
ours, no feat pruning	6.01	80 min	*30 cm
ours	6.01	80 min	0.92 cm
No vision	n/a	0 min	4.63 cm

*These controllers became unstable and did not train successfully.

TABLE IV: Comparisons to prior autoencoder architectures and variants of our method.

Another advantage of our approach is its sample efficiency, which is enabled both by the use of simple linear-Gaussian controllers and a data-efficient neural network architecture. The autoencoders used around 50 trials for each task, with each trial consisting of 100 image frames and 5 second of interaction time, for a total of 5000 frames per task. Training the final vision-based controller required another 50-75 trials, which means that each controller was trained with around 10-15 minutes of total robot interaction time.

Videos of each of the learned controllers can be viewed in the supplementary materials and on the project website: <http://rll.berkeley.edu/dsae>.

VIII. DISCUSSION AND FUTURE WORK

We presented a method for learning state representations using deep spatial autoencoders, and we showed how this method could be used to learn a range of manipulation skills that require close coordination between perception and action. Our method uses a spatial feature representation of the environment, which is learned as a bottleneck layer in an autoencoder. This allows us to learn a compact state from high-dimensional real-world images. Furthermore, since this representation corresponds to image-space coordinates of objects in the scene, it is particularly well suited for continuous control. The trajectory-centric reinforcement learning algorithm we employ can learn a variety of manipulation skills with these spatial feature representations using only tens of trials on the real robot. Our evaluation shows that our method can learn skills that require manipulating freely moving objects in the scene, including scooping and pushing.

While we found that controllers trained without vision could adequately explore the space for each task to generate training data for representation learning, there are tasks where even visiting a sufficient range of states can be very difficult without vision. A promising direction for tackling such tasks is to interleave representation learning with controller optimization, in a similar spirit to iterative model learning approaches [39]. Another promising direction for future work is to extend our method to use additional cameras and sensory modalities to learn more advanced sensory state representations, for example depth information and haptics.

The prospect of autonomously learning compact and portable state representations of complex environments entirely from raw sensory inputs, such as images, has widespread implications for autonomous robots. The method presented in this paper takes one step in this direction, by demonstrating that a spatial feature architecture can effectively learn suitable representations for a range of manipulation tasks, and these representations can be used by a trajectory-centric reinforcement learning algorithm to learn those tasks. Further research in this direction has the potential to make it feasible to apply out-of-the-box learning-based methods to acquire complex skills for tasks where a state space is very difficult to design by hand, such as tasks with deformable objects, complex navigation in unknown environments, and numerous other tasks.

REFERENCES

- [1] S. Lange, A. Voigtlaender, and M. Riedmiller, "Autonomous reinforcement learning on raw visual input data in a real world application," in *International Joint Conference on Neural Networks*, 2012.
- [2] R. Jonschkowski and O. Brock, "State representation learning in robotics: Using prior knowledge about physical interaction," in *Robotics: Science and Systems (RSS)*, 2014.
- [3] M. Watter, J. T. 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] T. Lampe and M. Riedmiller, "Acquiring visual servoing reaching and grasping skills using neural reinforcement learning," in *International Joint Conference on Neural Networks (IJCNN)*, 2013.
- [5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [6] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Neural Information Processing Systems (NIPS)*, 2008.
- [7] S. Levine and P. Abbeel, "Learning neural network policies with guided policy search under unknown dynamics," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [8] S. Levine, N. Wagener, and P. Abbeel, "Learning contact-rich manipulation skills with guided policy search," in *International Conference on Robotics and Automation (ICRA)*, 2015.
- [9] J. Kober, E. Oztop, and J. Peters, "Reinforcement learning to adjust robot movements to new situations," in *Robotics: Science and Systems (RSS)*, 2010.
- [10] M. Deisenroth, C. Rasmussen, and D. Fox, "Learning to control a low-cost manipulator using data-efficient reinforcement learning," in *Robotics: Science and Systems (RSS)*, 2011.
- [11] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "Learning and generalization of motor skills by learning from demonstration," in *International Conference on Robotics and Automation (ICRA)*, 2009.
- [12] G. Endo, J. Morimoto, T. Matsubara, J. Nakanishi, and G. Cheng, "Learning CPG-based biped locomotion with a policy gradient method: Application to a humanoid robot," *International Journal of Robotic Research*, vol. 27, no. 2, pp. 213–228, 2008.
- [13] T. Geng, B. Porr, and F. Wörgötter, "Fast biped walking with a reflexive controller and realtime policy searching," in *Advances in Neural Information Processing Systems (NIPS)*, 2006.
- [14] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *International Conference on Robotics and Automation (IROS)*, 2004.
- [15] R. Tedrake, T. Zhang, and H. Seung, "Stochastic policy gradient reinforcement learning on a simple 3d biped," in *International Conference on Intelligent Robots and Systems (IROS)*, 2004.
- [16] B. Sallans and G. E. Hinton, "Reinforcement learning with factored states and actions," *The Journal of Machine Learning Research*, vol. 5, pp. 1063–1088, 2004.
- [17] L. Baird *et al.*, "Residual algorithms: Reinforcement learning with function approximation," in *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 30–37.
- [18] S. Singh, M. L. Littman, N. K. Jong, D. Pardoe, and P. Stone, "Learning predictive state representations," in *ICML*, 2003.
- [19] B. Boots, S. M. Siddiqi, and G. J. Gordon, "Closing the learning-planning loop with predictive state representations," *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 954–966, 2011.
- [20] W. Böhmer, J. T. Springenberg, J. Boedecker, M. Riedmiller, and K. Obermayer, "Autonomous learning of state representations for control," *KI - Künstliche Intelligenz*, pp. 1–10, 2015.
- [21] R. Goroshin, M. Mathieu, and Y. LeCun, "Learning to linearize under uncertainty," *CoRR*, vol. abs/1506.03011, 2015.
- [22] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *NIPS '13 Workshop on Deep Learning*, 2013.
- [23] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *arXiv preprint arXiv:1504.00702*, 2015.
- [24] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 3, 1992.
- [25] K. Mohta, V. Kumar, and K. Daniilidis, "Vision based control of a quadrotor for perching on planes and lines," in *International Conference on Robotics and Automation (ICRA)*, 2014.
- [26] W. J. Wilson, C. W. Hulls, and G. S. Bell, "Relative end-effector control using cartesian position based visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, 1996.
- [27] M. Jägersand, O. Fuentes, and R. C. Nelson, "Experimental evaluation of uncalibrated visual servoing for precision manipulation," in *International Conference on Robotics and Automation (ICRA)*, 1997.
- [28] B. H. Yoshimi and P. K. Allen, "Active, uncalibrated visual servoing," in *International Conference on Robotics and Automation (ICRA)*, 1994.
- [29] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.
- [30] J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun, "Stacked what-where auto-encoders," *CoRR*, vol. abs/1506.02351, 2015.
- [31] M. Deisenroth, G. Neumann, and J. Peters, "A survey on policy search for robotics," *Foundations and Trends in Robotics*, vol. 2, no. 1-2, 2013.
- [32] W. Li and E. Todorov, "Iterative linear quadratic regulator design for nonlinear biological movement systems," in *ICINCO (1)*, 2004.
- [33] J. A. Bagnell and J. Schneider, "Covariant policy search," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.
- [34] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008.
- [35] J. Peters, K. Mülling, and Y. Altun, "Relative entropy policy search," in *AAAI Conference on Artificial Intelligence*, 2010.
- [36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning, ICML 2015*, 2015, pp. 448–456.
- [37] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.
- [39] S. Ross and A. Bagnell, "Agnostic system identification for model-based reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2012.

A. Robotic Experiment Details

All of the robotic experiments were conducted on a PR2 robot. The robot was controlled at 20 Hz via direct effort control,² and camera images were recorded using the RGB camera on a PrimeSense Carmine sensor. The images were downsampled to 240×240 . The learned policies controlled one 7 DoF arm of the robot. The camera was kept fixed in each experiment. Each episode was 5 seconds in length. For each task, the cost function required reaching the goal state, defined both by visual features and gripper pose. Similar to previous work, the cost was given by the following equation:

$$\ell(\mathbf{x}_t, \mathbf{u}_t) = w_{\ell_2} d_t^2 + w_{\log} \log(d_t^2 + \alpha) + w_{\mathbf{u}} \|\mathbf{u}_t\|^2,$$

where d_t is the distance between three points in the space of the end-effector and learned feature points in 2D and their respective target positions³, and the weights are set to $w_{\ell_2} = 10^{-3}$, $w_{\log} = 1.0$, and $w_{\mathbf{u}} = 10^{-2}$. The quadratic term in the cost encourages moving towards the target when it is far, while the logarithm term encourages reaching the target state precisely, as discussed in prior work [8]. The rice scoop task used two target states, in sequence, with half of the episode (2.5 seconds) devoted to each target. For each of the tasks, the objects were reset to their starting positions manually between trials during training. We discuss the particular setup for each experiment below:

a) *Lego block*: The lego block task required the robot to push a lego block 30 cm to the left of its initial position. For this task, we measured and reported the distance between the top corner of the goal block position to the nearest corner of the lego block at the end of the trial. In some of the baseline evaluations, the lego block was flipped over, and the nearest corner was still used to measure distance to the goal.

b) *Bag transfer*: The bag transfer task required the robot to place a white bag into a bowl, using a spoon. At the start of each trial, the robot was grasping the spoon with the bag in the spoon. A trial was considered successful if the bag was inside the bowl and did not extend outside of the bowl. In practice, the bag was very clearly entirely in the bowl, or entirely outside of the bowl during all evaluations.

c) *Rice scoop*: The rice scooping task required the robot to use a spatula to lift a small bag of rice off of a table and place it in a bowl. At the start of each trial, the spatula was in the grasp of the robot gripper, and the bag of rice was on the table, about 3 cm from the bowl. As with the bag transfer task, a trial was considered successful if the bag of rice was inside the bowl and did not extend outside of the bowl. In practice, the bag was very clearly in the bowl, or outside of the bowl during all evaluations.

d) *Loop hook*: The loop hook task required the robot to place a loop of rope onto a metal hook attached to a

scale, for different positions of the hook. At training time, the scale was placed at four different starting positions along a metal pole that were equally spaced across 24 cm of the pole. The test positions were the three midpoints between the four training positions. A trial was considered successful if, upon releasing the rope, the loop of rope would hang from the hook. In practice, the failed trials using our approach were often off by only 1-2 mm, whereas the controller with no vision was typically off by several centimeters.

B. Neural Network Architectures for Prior Work Methods

We compare our network with two common neural network architectures. The first baseline architecture is the one used by Lange et al. [1]. The network is composed of 8 encoder layers and 8 decoder layers. To match the original architecture as closely as possible, we converted our 240×240 RGB images into 60×60 grayscale images before passing them through the network. The encoder starts with 3 convolution layers with filter size 7×7 , where the last convolution layer has stride 2. The last convolution layer is followed by 6 fully connected layers, the size of which are 288, 144, 72, 36, 18 and 10 respectively. The last fully connected layer forms the bottleneck of the autoencoder. We chose 10 as the dimension of the bottleneck, since the system has roughly 10 degrees of freedom. The decoder consists of 6 mirrored fully connected layers followed by 3 deconvolution layers, finally reconstructing the down sampled 60×60 image. We used ReLU nonlinearities between each layer. Following [1], we pre-train each pair of the encoder-decoder layers for 4000 iterations. Then, we perform fine tuning on the entire network until the validation error plateaus.

We also experimented with a more widely adopted convolutional architecture. The $240 \times 240 \times 3$ image is directly passed to the network. This network starts with 3 convolutional layers. As in our network architecture, conv1 consists of $64 \times 7 \times 7$ filters with stride 2, conv2 has $32 \times 5 \times 5$ filters with stride 1, and conv3 has $16 \times 5 \times 5$ filters with stride 1, each followed by batch normalization and ReLU nonlinearities. Unlike our architecture, this baseline architecture performs max-pooling after each convolution layer in order to decrease the dimensionality of the feature maps. The convolution layers are followed by two fully connected layers with 512 and 32 units respectively, the last of which forms the bottleneck of the network. These layers together form the encoder, and a mirroring architecture, consisting of fully connected layers and deconvolution layers, forms the decoder. We initialize the first convolution layer with weights trained on ImageNet, and train the network until validation error plateaus.

²The PR2 robot does not provide for closed loop torque control, but instead supports an effort control interface that directly sets feedforward motor voltages. In practice, these voltages are roughly proportional to feedforward torques, but are also affected by friction and damping.

³Three points fully define the pose of the end-effector.