Learning Predictive Models from Observation and Interaction

Karl Schmeckpeper[1][0000−0003−4989−2022], Annie Xie[3][0000−0003−1736−3775], Oleh Rybkin[1][0000−0002−5898−006X], Stephen Tian[2][0000−0003−3148−5541], Kostas Daniilidis[1][0000−0003−0498−0758], Sergey Levine[3][0000−0001−6764−2743], and Chelsea Finn[2][0000−0001−6298−0874]

[1] University of Pennsylvania, Philadelphia PA
[2] Stanford University, Stanford CA
[3] University of California, Berkeley, Berkeley CA
karls@seas.upenn.edu

Abstract. Learning predictive models from interaction with the world allows an agent, such as a robot, to learn about how the world works, and then use this learned model to plan coordinated sequences of actions to bring about desired outcomes. However, learning a model that captures the dynamics of complex skills represents a major challenge: if the agent needs a good model to perform these skills, it might never be able to collect the experience on its own that is required to learn these delicate and complex behaviors. Instead, we can imagine augmenting the training set with observational data of other agents, such as humans. Such data is likely more plentiful, but cannot always be combined with data from the original agent. For example, videos of humans might show a robot how to use a tool, but (i) are not annotated with suitable robot actions, and (ii) contain a systematic distributional shift due to the embodiment differences between humans and robots. We address the first challenge by formulating the corresponding graphical model and treating the action as an observed variable for the interaction data and an unobserved variable for the observation data, and the second challenge by using a domain-dependent prior. In addition to interaction data, our method is able to leverage videos of passive observations in a driving dataset and a dataset of robotic manipulation videos to improve video prediction performance. In a real-world tabletop robotic manipulation setting, our method is able to significantly improve control performance by learning a model from both robot data and observations of humans.
always

Keywords: video prediction, visual planning, action representations, robotic manipulation

1 Introduction

Humans have the ability to learn skills not just from their own interaction with the world but also by observing others. Consider an infant learning to use tools.

[1] Correspondence to: Karl Schmeckpeper <karls@seas.upenn.edu>.

In order to use a tool successfully, it needs to learn how the tool can interact with other objects, as well as how to move the tool to trigger this interaction. Such intuitive notion of physics can be learned by observing how adults use tools. More generally, observation is a powerful source of information about the world and how actions lead to outcomes. However, in the presence of physical differences (such as between an adult body and infant body), leveraging observation is challenging, as there is no direct correspondence between the demonstrator's and observer's actions. Evidence from neuroscience suggests that humans can effectively infer such correspondences and use them to learn from observation [45,44]. In this paper, we consider this problem: can agents learn to solve tasks using both their own interaction and the passive observation of other agents?

In model-based reinforcement learning, solving tasks is commonly addressed via learning action-conditioned predictive models. However, prior works have learned such predictive models from interaction data alone [24,23,28,16,68]. When using both interaction and observation data, the setup differs in two important ways. First, the actions of the observed agent are not known, and therefore directly learning an action-conditioned predictive model is not possible. Second, the observation data might suffer from a domain shift if the observed agent has a different embodiment, operates at a different skill level, or exists in a different environment. Yet, if we can overcome these differences and effectively leverage observational data, we may be able to unlock a substantial source of broad data containing diverse behaviors and interactions with the world.

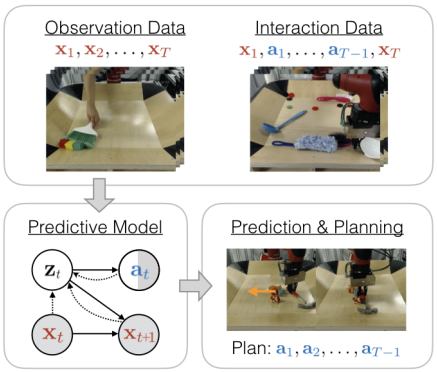

Fig. 1: Our system learns from action-observation sequences collected through interaction, such as robotic manipulation or autonomous vehicle data, as well as action-free observations of another demonstrator agent, such as data from a human or a dashboard camera. By combining interaction and observation data, our model is able to learn to generate predictions for complex tasks and new environments without costly expert demonstrations.

Our main contribution is an approach for learning predictive models that can leverage both videos of an agent annotated with actions and observational data for which actions are not available. We formulate a latent variable model for prediction, in which the actions are observed variables in the first case and unobserved variables in the second case. We further address the domain shift between the observation and interaction data by learning a domain-specific prior over the latent variables.

We instantiate the model with deep neural networks and train it with amortized variational inference. In two problem settings – driving and object manipulation – we find that our method is able to effectively leverage observational data from dashboard cameras and humans, respectively, to improve the performance of action-conditioned prediction. Furthermore, we find that the resulting model enables a robot to solve pushing and sweeping tool-use tasks, and achieves significantly greater success than a model that does not use observational data of a human using tools. Finally, we release our dataset of human demonstrations of pushing and sweeping with tools to allow others to study this problem. To the best of our knowledge, this is the first work to demonstrate a method for learning predictive models from both observation and interaction data.

2 Related Work

Predictive models Video prediction can be used to learn useful representations and models in a fully unsupervised manner. These representations can be used for tasks such as action recognition [50], action prediction [62], classification [14], and planning [18,19,16,29,6,23,24,30,20]. Many different approaches have been applied to video prediction, including patch-centric methods [43], compositional models of content and motion [61,14,58], pixel autoregressive models [31], hierarchical models [8,40,39], transformation-based methods [37,42,18,63,34,36,33,3,9,1], and other techniques [12,67,5,38]. We choose to leverage transformation-based models, as they have demonstrated good results on robotic control domains [18,16]. Recent work has also developed stochastic video prediction models for better handling of uncertainty [13,33,3,69,9,64]. We also use a stochastic latent variable, and unlike these prior works, use it to model actions.

Learning action-conditioned visual dynamics models was proposed in [41,18,10]. Using model predictive control techniques, flow based action-conditioned prediction models have been applied to robotic manipulation [18,19,16,29,6,73]. Other works address video games or physical simulation domains [23,24,30,20,66].

The models have been shown to generalize to unseen tasks and objects while allowing for challenging manipulation of deformable objects, such as rope or clothing [65,11,70]. Unfortunately, large amounts of robotic interaction data containing complex behavior are required to train these models. These models are unable to learn from cheap and abundantly available natural videos of humans as they are trained in *action-conditioned* way, requiring corresponding control data for every video. In contrast, our method can learn from videos without actions, allowing it to leverage videos of agents for which the actions are unknown.

Learning to control without actions Recent work in imitation learning allows the agent to learn without access to the ground-truth expert actions. One set of approaches learn to translate the states of the expert into actions the agent can execute [55,72]. Action-free data can also be used to learn a set of sub-goals for hierarchical RL [32,48]. Another common approach is to learn a policy in the agent's domain that matches the expert trajectories under some similarity

metric. Adversarial training or other metrics have been used to minimize the difference between the states generated by the demonstrated policy and the states generated by the learned policy [56,57,51,53]. Liu et al. transform images from the expert demonstrations into the robot's domain to make calculating the similarity between states generated by different policies in different environments more tractable [35]. Edwards et al. learn a latent policy on action-free data and use action-conditioned data to map the latent policy to real actions [17]. Several works learn state representations that can be used to transfer policies from humans to robots [47,15,2]. Shon et al. learn a mapping between human and robot degrees of freedom to allow the robot to match the human's pose [49]. Sun et al. use partially action-conditioned data to train a generative adversarial network to synthesize the missing action sequences [52]. Unlike these works, which aim to specify a specific task to be solved through expert demonstrations, we aim to learn predictive models that can be used for multiple tasks, as we learn general properties of the real world through model-building.

Recent prior work has considered learning predictive models from an initial dataset that is entirely action-free [46], learning a mapping from actions to latent variables post-hoc. However, this approach has been limited to simple simulated settings with no domain shift. Unlike this prior work, we explicitly handle domain shift between the interaction and observational data, and consider challenging real video datasets. Furthermore, our experiments indicate that our approach substantially outperforms the approach of Rybkin et al. [46] on multiple domains.

Domain adaptation In order to handle both observational and interaction data, our method must handle the missing actions and bridge the gap between the two domains (e.g., human arms vs. robot arms). Related domain adaptation methods have sought to map samples in one domain into equivalent samples in another domain [75,4,54,26], or learn feature embeddings with domain invariance losses [60,76,21,22,59]. In our setting, regularizing for invariance across domains is insufficient. For example, if the observational data of humans involves complex manipulation (e.g., tool use), while the interaction data involves only simple manipulation, we do not want the model to be invariant to these differences. Therefore, instead of regularizing for invariance across domains, we explicitly model the distributions over (latent) action variables in each of the domains.

Related to our method, DIVA [27] aims to avoid losing this information by proposing a generative model with a partitioned latent space. The latent space is composed of both components that are domain invariant and components that are conditioned on the domain. This allows the model to use domain-specific information while still remaining robust to domain shifts. We find that using an approach similar to DIVA in our model for learning from observation and interaction makes it more robust to the domain shift between interaction and observation data. However, in contrast to DIVA, our method explicitly handles sequence data with missing actions in one of the domains.

3 Learning Predictive Models from Observation and Interaction

In our problem setting, we assume access to observation data of the form $[\mathbf{x}_1, \ldots, \mathbf{x}_T]$ and interaction data of the form $[\mathbf{x}_1, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1}, \mathbf{x}_T]$, where $\mathbf{x}_i$ denotes the i^{th} frame of a video and $\mathbf{a}_i$ denotes the action taken at the i^{th} time step. Domain shift may exist between the two datasets: for example, when learning object manipulation from videos of humans and robotic interaction, as considered in our experiments, there is a shift in the embodiment of the agent. Within this problem setting, our goal is to learn an action-conditioned video prediction model, $p(\mathbf{x}_{c+1:T}|\mathbf{x}_{1:c}, \mathbf{a}_{1:T})$, that predicts future frames conditioned on a set of c context frames and sequence of actions.

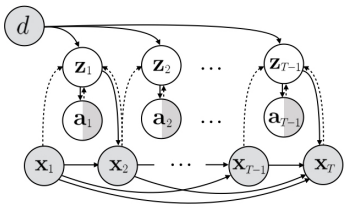

Fig. 2: We learn a predictive model of visual dynamics (in solid lines) that predicts the next frame x_{t+1} conditioned on the current frame x_t and action representation z_t. We optimize the likelihood of the interaction data, for which the actions are available, and observation data, for which the actions are missing. Our model is able to leverage joint training on the two kinds of data by learning a latent representation z that corresponds to the true action.

To approach this problem, we formulate a probabilistic graphical model underlying the problem setting where actions are only observed in a subset of the data. In particular, in Subsection 3.1, we introduce a latent variable that explains the transition from the current frame to the next and, in the case of interaction data, encodes the action taken by the agent. We further detail how the latent variable model is learned from both observation and interaction data by amortized variational inference. In Subsection 3.2, we discuss how we handle domain shift by allowing the latent variables from different datasets to have different prior distributions. Finally, we discuss implementation details in Subsection 3.3.

3.1 Graphical Model

To leverage both passive observations and active interactions, we formulate the probabilistic graphical model depicted in Figure 2. To model the action of the agent $\mathbf{a}_t$, we introduce a latent variable $\mathbf{z}_t$, distributed according to a domain-dependent distribution. The latent $\mathbf{z}_t$ generates the action $\mathbf{a}_t$. We further introduce a forward dynamic model that, at each time step t, generates the frame $\mathbf{x}_t$ given the previous frames $\mathbf{x}_{1:t-1}$ and latent variables $\mathbf{z}_{1:t-1}$. The generative

model can be summarized as:

$$\mathbf{z}_t \sim p(\mathbf{z}_t|d) \tag{1}$$
$$\mathbf{a}_t \sim p(\mathbf{a}_t|\mathbf{z}_t) \tag{2}$$
$$\mathbf{x}_{t+1} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \mathbf{z}_{1:t}). \tag{3}$$

The domain-dependent distribution over $\mathbf{z}_t$ is Gaussian with learned mean and variance, described in more detail in Subsection 3.2, while the *action decoder* $p(\mathbf{a}_t|\mathbf{z}_t)$ and *transition model* $p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \mathbf{z}_{1:t})$ are neural networks with Gaussian distribution outputs, described in Subsection 3.3.

The transition model takes $\mathbf{z}_t$ as input and thus necessitates the posterior distributions $p(\mathbf{z}_t|\mathbf{a}_t)$ and $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$. We require $p(\mathbf{z}_t|\mathbf{a}_t)$ to generate latent variables for action-conditioned video prediction, i.e. sampling from

$$p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \mathbf{a}_{1:t}) = \mathbb{E}_{p(\mathbf{z}_{1:t}|\mathbf{a}_{1:t})}\left[p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \mathbf{z}_{1:t})\right].$$

We also require $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$ since the actions are not available in some trajectories to obtain the first distribution.

The computation of these two posterior distributions is intractable, since the model is highly complex and non-linear, so we introduce the variational distributions $q_{\text{act}}(\mathbf{z}_t|\mathbf{a}_t)$ and $q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$ to approximate $p(\mathbf{z}_t|\mathbf{a}_t)$ and $p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$. The distributions are modeled as Gaussian and the variational parameters are learned by optimizing the evidence lower bound (ELBO), which is constructed by considering two separate cases. In the first, the actions are observed, and we optimize an ELBO on the joint probability of the frames and the actions:

$$\begin{aligned}\log p(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}) \geq &\sum_t \mathbb{E}_{q_{\text{act}}(\mathbf{z}_{1:t}|\mathbf{a}_{1:t})}\left[\log p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \mathbf{z}_{1:t}) + \log p(\mathbf{a}_t|\mathbf{z}_t)\right] \\ &- \sum_t D_{KL}(q_{\text{act}}(\mathbf{z}_t|\mathbf{a}_t)||p(\mathbf{z}_t)) = -\mathcal{L}_i(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}).\end{aligned} \tag{4}$$

In the second case, the actions are not observed, and we optimize an ELBO on only the probability of the frames:

$$\begin{aligned}\log p(\mathbf{x}_{1:T}) \geq &\sum_t \mathbb{E}_{q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})}\left[\log p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{z}_t)\right] \\ &- \sum_t D_{KL}(q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})||p(\mathbf{z}_t)) = -\mathcal{L}_o(\mathbf{x}_{1:T}).\end{aligned} \tag{5}$$

The full ELBO is the combination of the lower bounds for the interaction data with actions, D^i, and the observation data without actions, D^o:

$$\mathcal{J} = \sum_{(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}) \sim D^i} \mathcal{L}_i(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}) + \sum_{\mathbf{x}_{1:T} \sim D^o} \mathcal{L}_o(\mathbf{x}_{1:T}). \tag{6}$$

We also add an auxiliary loss to align the distributions of $\mathbf{z}$ generated from the encoders $q_{\text{act}}(\mathbf{z}_t|\mathbf{a}_t)$ and $q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$, since the encoding $\mathbf{z}$ should be

independent of the distribution it was sampled from. We encourage the two distributions to be similar through the Jensen-Shannon divergence:

$$\mathcal{L}_{JS} = \sum_{(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}) \sim D^i} D_{JS}(q_{\text{act}}(\mathbf{z}_t|\mathbf{a}_t) \| q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})). \tag{7}$$

Our final objective combines the evidence lower bound for the entire dataset and the Jensen-Shannon divergence, computed for the interaction data:

$$\mathcal{F} = \mathcal{J} + \alpha \mathcal{L}_{JS}. \tag{8}$$

We refer to our method as prediction from observation and interaction (POI).

3.2 Domain Shift

When learning from both observation and interaction, domain shift may exist between the two datasets. For instance, in the case of a robot learning by observing people, the two agents differ both in their physical appearance, as well as their action spaces. To address these domain shifts, we take inspiration from the domain-invariant approach described in [27]. We divide our latent variable $\mathbf{z}$ into $\mathbf{z}^{\text{shared}}$, which captures the parts of the latent action that are shared between domains, and $\mathbf{z}^{\text{domain}}$, which captures the parts of the latent action that are unique to each domain.

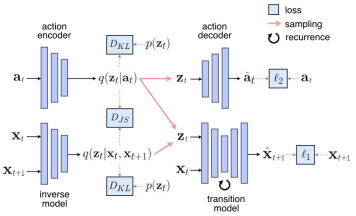

Fig. 3: Network architecture. To optimize the ELBO, we predict the latent action $\mathbf{z}_t$ from $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ using the inverse model q_{inv}. When the true actions are available, we additionally predict the latent action from the true action $\mathbf{a}_t$ using the action encoder q_{act}, and encourage the predictions from q_{act} and q_{inv} to be similar with a Jensen-Shannon divergence loss. The next frame is predicted from $\mathbf{z}_t$ and $\mathbf{x}_t$.

We allow the network to learn the difference between the $\mathbf{z}^{\text{domain}}$ for each dataset by using different prior distributions. The prior $p(\mathbf{z}_t^{\text{shared}})$ is the same for both domains, however, the prior for $\mathbf{z}_t^{\text{domain}}$ is different for the interaction dataset, $p_i(\mathbf{z}_t^{\text{domain}})$, and the observational dataset, $p_o(\mathbf{z}_t^{\text{domain}})$. $p(\mathbf{z}_t^{\text{shared}})$ and $p_a(\mathbf{z}_t^{\text{domain}})$ are both multivariate Gaussian distributions with a learned mean and variance for each dimension. The prior is the same for all timesteps t.

Unlike the actions for the robot data, which are sampled from the same distribution at each time step, the actions of the human are correlated across time. For the human observation data, the prior $p_o(\mathbf{z}_{1:T}^{\text{domain}}|\mathbf{x}_1)$ models a joint distribution over timesteps, and is parameterized as a long short-term memory (LSTM) network [25]. The input to the LSTM at the first timestep is an encoding of the initial observation, and the LSTM cell produces the parameters of the multivariate Gaussian distribution for each time step.

3.3 Deep Neural Network Implementation

A high-level diagram of our network architecture is shown in Figure 3. Our action encoder $q_{\text{act}}(\mathbf{z}_t|\mathbf{a}_t)$ is a multi-layer perceptron (MLP) with 3 layers of 64 units to encode the given action $\mathbf{a}_t$ to the means and variances for each dimension of the encoding. Our action decoder predicts the mean of the distribution $p(\mathbf{a}_t|\mathbf{z}_t)$ using an MLP with 3 layers of 64 units each, while using a fixed unit variance.

Our inverse model $q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$ is a convolutional network that predicts the distribution over the action encoding. The network is made up of three convolutional layers with {32, 64, 128} features with a kernel size of 4 and a stride of 2. Each convolutional layer is followed by instance normalization and a leaky-ReLU. The output of the final convolutional layer is fed in a fully connected layer, which predicts the means and variances of the action encoding.

We encourage the action encodings generated by the action encoder q_{act} and the inverse model q_{inv} to be similar using the Jensen-Shannon divergence in Equation 7. Since the Jensen-Shannon divergence does not have a closed form solution, we approximate it by using a mean of the Gaussians instead of a mixture. Our model uses a modified version of the SAVP architecture [33] as the transition model which predicts $\mathbf{x}_{t+1}$ from $\mathbf{x}_t$ and an action encoding $\mathbf{z}$, either sampled from $q_{\text{act}}(\mathbf{z}_t|\mathbf{a}_t)$ or from $q_{\text{inv}}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t+1})$. In the case where the actions are observed, we generate two predictions, one from each of q_{inv} and q_{act}, and in the case where actions are not observed, we only generate a prediction from the inverse model, q_{inv}. This architecture has been shown to be a useful transition model for robotic planning in [16,11].

4 Experiments

We aim to answer the following in our experiments:

1. Do passive observations, when utilized effectively, improve an action-conditioned visual predictive model despite large domain shifts?
2. How does our approach compare to alternative methods for combining passive and interaction data?
3. Do improvements in the model transfer to downstream tasks, such as robotic control?

To answer question 1, we compare our method to a strong action-conditioned prediction method, SAVP [33], which is trained only on interaction data as it is not able to leverage the observation data. To answer question 2, we further compare to CLASP [46], a prior method that infers actions in a post-hoc manner and does not model domain shift. We study questions 1 and 2 in both the driving domain in Subsection 4.1 and the robotic manipulation domain in Subsection 4.2 and evaluate the methods on action-conditioned prediction. We evaluate question 3 by controlling the robotic manipulator using our learned model.

Method	PSNR (↑)	SSIM (↑)	LPIPS [74] (↓)
SAVP [33] (Boston w/ actions)	19.74 ± 0.41	0.5121 ± 0.0164	0.1951 ± 0.0075
CLASP [46] (Boston w/ actions, Singapore w/o actions)	20.57 ± 0.48	0.5431 ± 0.0161	0.1964 ± 0.0076
POI (ours) (Boston w/ actions, BDD100K w/o actions)	$\mathbf{20.88 \pm 0.24}$	$\mathbf{0.5508 \pm 0.0076}$	0.2106 ± 0.0089
POI (ours) (Boston w/ actions, Singapore w/o actions)	20.81 ± 0.49	0.5486 ± 0.0164	$\mathbf{0.1933 \pm 0.0074}$
Oracle: SAVP [33] (Boston w/ actions. Singapore w/ actions)	21.17 ± 0.47	0.5752 ± 0.0156	0.1738 ± 0.0076

Table 1: Action-conditioned prediction results on the Singapore portion of the nuScenes dataset, reporting the mean and standard error of each metric. By leveraging observational driving data from Singapore or from BDD dashboard cameras, our method is able to outperform prior models that cannot leverage such data (i.e. SAVP) and slightly outperform alternative approaches to using such data.

4.1 Visual Prediction for Driving

We first evaluate our model on video prediction for driving. Imagine that a self-driving car company has data from a fleet of cars with sensors that record both video and the driver's actions in one city, and a second fleet of cars that only record dashboard video, without actions, in a second city. If the goal is to train an action-conditioned model that can be utilized to predict the outcomes of steering actions, our method allows us to train such a model using data from both cities, even though only one of them has actions.

We use the nuScenes [7] and BDD100K [71] datasets for our experiments. The nuScenes dataset consists of 1000 driving sequences collected in either Boston or Singapore, while the BDD100K dataset contains only video from dashboard cameras. In nuScenes, we discard all action and state information for the data collected in Singapore, simulating data that could have been collected by a car equipped with only a camera. We train our model with action-conditioned video from Boston and action-free video either from the nuScenes Singapore data or the BDD100K data, and evaluate on action-conditioned prediction on held-out data from Singapore (from nuScenes). Since the action distribution for all datasets is likely very similar as they all contain human driving, we use the same learned means and variances for the Gaussian prior over z for both por-

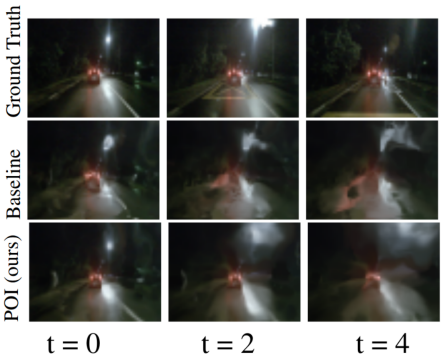

Fig. 4: Example predictions on the Singapore portion of the Nuscenes dataset. This sequence was selected for large MSE difference between the models. More examples are available in the supplementary material. We compare our model to the baseline of the SAVP model trained on the Boston data with actions. Our model is able to maintain the shape of the car in front.

tions of the dataset. We additionally train a our model with the action-conditioned video from Boston and action-free video taken from the BDD100K dataset [71].

We compare our predictions to those generated by the SAVP [33] model trained with only the action-conditioned data from Boston, since SAVP cannot leverage action-free data for action-conditioned prediction. We additionally compare our predictions to those generated by CLASP [46] trained with action-conditioned video from Boston, and action-free video from Singapore. As an upper-bound, we train the SAVP [33] model with action-conditioned data from Boston and action-conditioned data from Singapore.

Comparisons between these methods are shown in Table 1. Qualitative results are shown in Figure 4. With either form of observational data, BDD2K or nuScenes Singapore, our method significantly outperforms the SAVP model trained with only action-conditioned data from Boston, demonstrating that our model can leverage observation data to improve the quality of its predictions. Further, our method slightly outperforms alternative approaches to learning from observation and interaction.

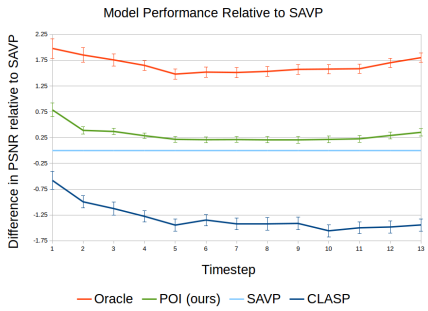

Fig. 5: Frame-by-frame differences in PSNR relative to SAVP, on the robotic domain. Our method consistently outperforms both SAVP and CLASP.

4.2 Robotic Manipulation: Prediction

We evaluate our model on the robotic manipulation domain, which presents a large distributional shift challenge between robot and human videos. In particular, we study a tool-use task and evaluate whether human videos of tool-use can improve predictions of robotic tool-use interactions.

Learning predictive models from interaction with the world allows an agent, such as a robot, to learn about how the world works, and then use this learned model to plan coordinated sequences of actions to bring about desired outcomes.

For our interaction data, we acquired 20,000 random trajectories of a Sawyer robot from the open-source datasets from [16] and [68], which consist of both video and corresponding actions. We then collected 1,000 videos of a human using different tools to push objects as the observation data. By including the human videos, we provide the model with examples of tool-use interactions, which are not available in the random robot data. Our test set is composed of 1,200 kinesthetic demonstrations from [68], in which a human guides the robot

Fig. 6: Example images from the robot (left) and human (right) datasets.

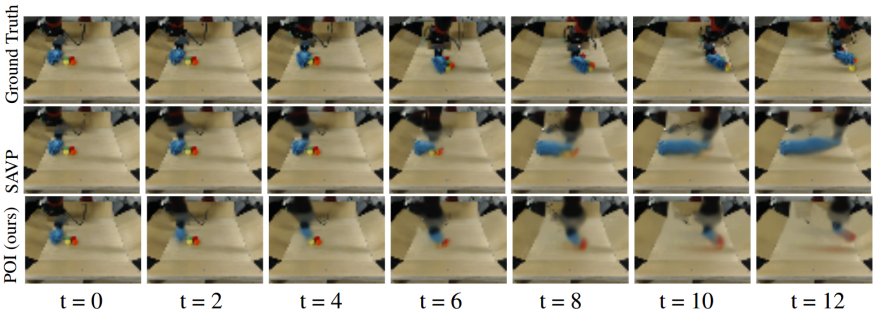

Fig. 7: Example predictions on the robotic dataset. We compare our model to the SAVP model trained with random robot data. This sequence was selected to maximize the MSE difference between the models. More examples are available in the supplementary material. Our model more accurately predicts both the tool and the object it pushes.

to use tools to complete pushing tasks similar to those in the human videos. Kinesthetic demonstrations are time-consuming to collect, encouraging us to build a system that can be trained without them, but they serve as a good proxy for evaluating robot tool-use behavior. Example images from the datasets are shown in Figure 6.[4] This dataset is especially challenging because of the large domain shift between the robot and human data. The human arm has a different appearance from the robot and moves in a different action space.

Method	PSNR (↑)	SSIM (↑)	LPIPS [74] (↓)
CLASP [46] (random robot, expert human)	22.14 ± 0.11	0.763 ± 0.004	0.0998 ± 0.0023
SAVP [33] (random robot)	23.31 ± 0.10	0.803 ± 0.004	0.0757 ± 0.0022
POI (ours) (random robot, expert human)	$\mathbf{23.79 \pm 0.12}$	$\mathbf{0.813 \pm 0.005}$	$\mathbf{0.0722 \pm 0.0024}$
Oracle: SAVP [33] (random robot, expert kinesthetic)	24.99 ± 0.11	0.858 ± 0.003	0.0486 ± 0.0017

Table 2: Means and standard errors for action-conditioned prediction on the manipulation dataset. By leveraging observational data of human tool use, our model was able to outperform prior models that cannot leverage such data (i.e. SAVP) and slightly outperform alternative approaches to using such data.

[4] Data will be made available at `https://sites.google.com/view/lpmfoai`

We compare to the CLASP model [46] trained with the same data as our model. We also evaluate the SAVP model [33], trained the same robot data, but without the human data, since the SAVP model is unable to leverage action-free data for action-conditioned prediction. For an oracle, we trained the SAVP model [33] on both the random robot trajectories and the kinesthetic demonstrations.

As shown in Table 2, our model is able to leverage information from the human videos to outperform the other models. Our model outperforms the SAVP model trained on only the random robot data, showing that it is possible to leverage passive observation data to improve action-conditioned prediction, even in the presence of the large domain shift between human and robot arms. Figure 5 shows the frame-by-frame differences in PSNR relative to SAVP.

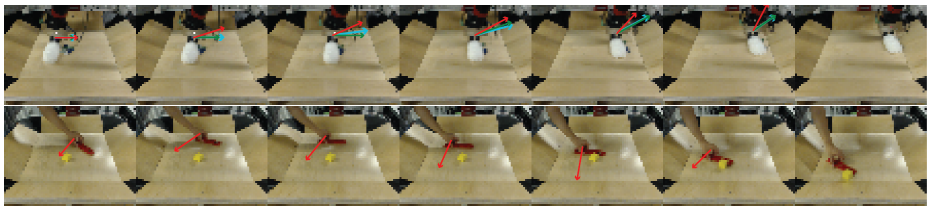

Fig. 8: Action predictions on human and robot data. The sequences of images show the ground truth observations, while the arrows show the action in the (x, y) plane between each pair of frames. The blue arrow is the ground truth action, the green arrow is the action generated from decoding the output of the action encoder, and the red is the action generated by decoding the output of the inverse model. The human data only has actions generated by the inverse model. Our model is able to infer plausible actions for both domains, despite never seeing ground truth human actions.

Qualitative results are shown in Figure 7. Our model is able to generate more accurate predictions than the baseline SAVP model that was trained with only the robotic interaction data. In addition to predicting future states, our model is able to predict the action that occurred between two states. Examples for both robot and human demonstrations are shown in Figure 8. Our inverse model is able to generate reasonable actions for both the robot and the human data despite having never been trained on human data with actions. Our model can reconstruct the actions with an average percent error of 14.3, while CLASP reconstructs the actions with an average percent error of 70.4. Our model maps human and robot actions to a similar space, allowing it to exploit their similarities to improve prediction performance on robotic tasks.

4.3 Robotic Manipulation: Planning and Control

To study the third and final research question, we evaluate the efficacy of our visual dynamics model in a set of robotic control experiments. We evaluate

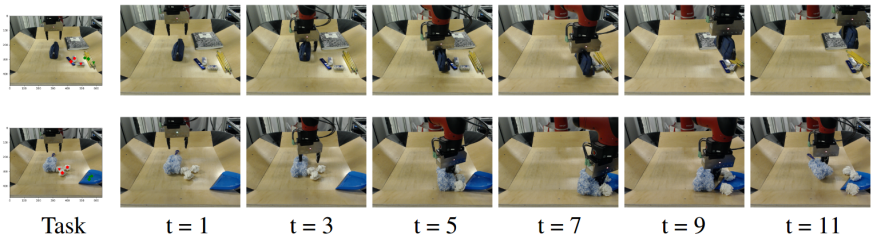

Fig. 9: Examples of a robot using our model to successfully complete tool use tasks. The robot must move the objects specified by the red symbols to the locations of the corresponding green symbols. The robot uses a tool to simultaneously move several objects to their goal locations.

each predictive model's ability to control the robot on a baseline task [68] by integrating the model with an existing visual model predictive control pipeline, which optimizes actions with respect to a user-provided task [19,16].

To evaluate the importance of the human data, we focus on control tasks that involve moving multiple objects, which would be difficult to complete without using a tool. While [19,16] only evaluated on simple one-step planning tasks, we want to see whether our model can be used to successfully solve more complex tasks by incorporating observational data of humans. Therefore, for testing our model, we position the objects such that it is not possible to solve the task greedily by moving directly towards the goal, following the evaluation setup for pushing and sweeping tool use tasks from Xie et al. [68]. In each task setting, several objects, as well as a tool that the robot could potentially use to complete the task, are placed in the scene. Tasks are specified by designating a pixel corresponding to an object and the goal position for the object, following [19,16,68]. We specify moving multiple objects by selecting multiple pairs of pixels. We quantitatively evaluate each model on 15 tasks with tools seen during training and 15 tasks with previously unseen tools. In Figure 9, we show qualitative examples of the robot completing tool-use tasks.

Method	Success Rate
SAVP [33] (random)	$23.3 \pm 7.7\%$
POI (ours) (random, human)	$\mathbf{40.0 \pm 8.9}\%$
Oracle (random, kinesthetic)	$36.7 \pm 8.8\%$

Table 3: Robotic control results, measuring the success rate and standard error for three object manipulation tasks. "random" denotes random robot data, "human" denotes human interaction data, and "kinesthetic" is an oracle dataset of expert robot trajectories. POI performs comparably to the oracle, and successfully leverages the observational videos to improve over SAVP.

The quantitative results, in Table 3, indicate that the planner can leverage our model to execute more successful plans relative to the baseline SAVP model, which was trained only using random robot trajectories. In our evaluation, a trial is successful if the average distance between the objects and their respective goal positions at the final time step is less than or equal to 10 centimeters. Using our model, the robot achieves similar performance to the oracle model trained on kinesthetic demonstrations with action labels. This suggests that our model's improvements on prediction leads to a corresponding improvements on control.

5 Conclusion

We present a method for learning predictive models from both passive observation and active interactions. Active interactions are usually more expensive and less readily-available than passive observation: for example, consider the amount of observational data of human activities on the internet. Active interaction, on the other hand, is especially difficult when the agent is trying to collect information about regions of the state-space which are difficult to reach. Without an existing policy that can guide the agent to those regions, time consuming on-policy exploration, expert teleoperated or kinesthetic demonstrations are often required, bringing additional costs.

By learning a latent variable over the semi-observed actions, our approach is able to leverage passive observational data to improve action-conditioned predictive models, even in the presence of domain shift between observation and interaction data. Our experiments illustrate these benefits in two problem settings: driving and object manipulation, and find improvements both in prediction quality and in control performance when using these models for planning.

Overall, we hope that this work represents a first step towards enabling the use of broad, large-scale observational data when learning about the world. However, limitations and open questions remain. Our experiments studied a limited aspect of this broader problem where the observational data was either a different embodiment in the same environment (i.e. humans manipulating objects) or a different environment within the same underlying dataset (i.e. driving in Boston and Singapore). In practice, many source of passive observations will exhibit more substantial domain shift than those considered in this work. Hence, an important consideration for future work is to increase robustness to domain shift to realize greater benefits from using more large and diverse observational datasets. Finally, we focused our study on learning predictive models; an exciting direction for future work is to study how to incorporate similar forms of observational data in representation learning and reinforcement learning.

Acknowledgements

We thank Karl Pertsch, Drew Jaegle, Marvin Zhang, and Kenneth Chaney. This work was supported by the NSF GRFP, ARL RCTA W911NF-10-2-0016, ARL DCIST CRA W911NF-17-2-0181, and by Honda Research Institute.

References

1. van Amersfoort, J., Kannan, A., Ranzato, M., Szlam, A., Tran, D., Chintala, S.: Transformation-Based Models of Video Sequences. arXiv preprint (jan 2017), http://arxiv.org/abs/1701.08435
2. Aytar, Y., Pfaff, T., Budden, D., Paine, T., Wang, Z., de Freitas, N.: Playing hard exploration games by watching YouTube. Advances in Neural Information Processing Systems 31 (2018), http://papers.nips.cc/paper/7557-playing-hard-exploration-games-by-watching-youtube.pdf
3. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. In: International Conference on Learning Representations (2018)
4. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
5. Byeon, W., Wang, Q., Kumar Srivastava, R., Koumoutsakos, P.: Contextvp: Fully context-aware video prediction. In: The European Conference on Computer Vision (ECCV) (September 2018)
6. Byravan, A., Leeb, F., Meier, F., Fox, D.: Se3-pose-nets: Structured deep dynamics models for visuomotor planning and control. Proceedings of International Conference in Robotics and Automation (ICRA) (2017)
7. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
8. Castrejon, L., Ballas, N., Courville, A.: Improved Conditional VRNNs for Video Prediction. arXiv preprint (apr 2019), http://arxiv.org/abs/1904.12165
9. Chen, B., Wang, W., Wang, J., Chen, X.: Video Imagination from a Single Image with Transformation Generation. arXiv preprint (jun 2017), http://arxiv.org/abs/1706.04124
10. Chiappa, S., Racanière, S., Wierstra, D., Mohamed, S.: Recurrent environment simulators. In: International Conference on Learning Representations (2017)
11. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C.: RoboNet: Large-Scale Multi-Robot Learning. Conference on Robot Learning (oct 2019), http://arxiv.org/abs/1910.11215
12. De Brabandere, B., Jia, X., Tuytelaars, T., Van Gool, L.: Dynamic Filter Networks. Neural Information Processing Systems (may 2016), http://arxiv.org/abs/1605.09673
13. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: International Conference on Machine Learning (ICML) (2018)
14. Denton, E., Birodkar, V.: Unsupervised learning of disentangled representations from video. In: Neural Information Processing Systems. pp. 4417–4426 (2017)
15. Dwibedi, D., Tompson, J., Lynch, C., Sermanet, P.: Learning actionable representations from visual observations. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1577–1584. IEEE (2018), https://arxiv.org/abs/1808.00928
16. Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., Levine, S.: Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. arXiv:1812.00568 (2018)
17. Edwards, A.D., Sahni, H., Schroecker, Y., Isbell, C.L.: Imitating Latent Policies from Observation. International Conference on Machine Learning (may 2019), http://arxiv.org/abs/1805.07914

18. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Neural Information Processing Systems (2016)
19. Finn, C., Levine, S.: Deep visual foresight for planning robot motion. In: Proceedings of International Conference in Robotics and Automation (ICRA) (2017)
20. Fragkiadaki, K., Agrawal, P., Levine, S., Malik, J.: Learning Visual Predictive Models of Physics for Playing Billiards. International Conference on Learning Representations (nov 2016), http://arxiv.org/abs/1511.07404
21. Ganin, Y., Lempitsky, V.: Unsupervised Domain Adaptation by Backpropagation. International Conference on Machine Learning (ICML) (2015)
22. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural net-works. The Journal of Machine Learning Research (2016)
23. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Neural Information Processing Systems (2018)
24. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. International Conference on Machine Learning (ICML) (2019)
25. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
26. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: CyCADA: Cycle-Consistent Adversarial Domain Adaptation. International Conference on Machine Learning (ICML) (nov 2018), http://arxiv.org/abs/1711.03213
27. Ilse, M., Tomczak, J.M., Louizos, C., Welling, M.: DIVA: Domain Invariant Variational Autoencoders. arXiv preprint (may 2019), http://arxiv.org/abs/1905.10427
28. Janner, M., Fu, J., Zhang, M., Levine, S.: When to trust your model: Model-based policy optimization. NeurIPS (2019)
29. Janner, M., Levine, S., Freeman, W.T., Tenenbaum, J.B., Finn, C., Wu, J.: Reasoning About Physical Interactions with Object-Oriented Prediction and Planning. International Conference on Learning Representations (dec 2019), http://arxiv.org/abs/1812.10972
30. Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R.H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., Sepassi, R., Tucker, G., Michalewski, H.: Model-based reinforcement learning for atari (2019)
31. Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., Kavukcuoglu, K.: Video Pixel Networks. arXiv preprint (oct 2016), http://arxiv.org/abs/1610.00527
32. Kumar, A., Gupta, S., Malik, J.: Learning navigation subroutines by watching videos. CoRR **abs/1905.12612** (2019), http://arxiv.org/abs/1905.12612
33. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. arXiv:1804.01523 **abs/1804.01523** (2018)
34. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual Motion GAN for Future-Flow Embedded Video Prediction. International Conference on Computer Vision (aug 2017), http://arxiv.org/abs/1708.00284
35. Liu, Y., Gupta, A., Abbeel, P., Levine, S.: Imitation from Observation: Learning to Imitate Behaviors from Raw Video via Context Translation. Ph.D. thesis, University of California, Berkeley (jul 2018), http://arxiv.org/abs/1707.03374
36. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4463–4471 (2017)

37. Lotter, W., Kreiman, G., Cox, D.: Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. arXiv preprint (may 2016), http://arxiv.org/abs/1605.08104
38. Lu, C., Hirsch, M., Scholkoph, B.: Flexible spatio-temporal networks for video prediction. Computer Vision and Pattern Recognition (2017)
39. Luc, P., Neverova, N., Couprie, C., Verbeek, J., LeCun, Y.: Predicting Deeper into the Future of Semantic Segmentation. International Conference on Computer Vision (mar 2017), http://arxiv.org/abs/1703.07684
40. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: International Conference on Learning Representations (2016)
41. Oh, J., Guo, X., Lee, H., Lewis, R., Singh, S.: Action-conditional video prediction using deep networks in atari games. In: Neural Information Processing Systems (2015)
42. Patraucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory. arXiv preprint (nov 2015), http://arxiv.org/abs/1511.06309
43. Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. arXiv preprint arXiv:1412.6604 (2014)
44. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annu. Rev. Neurosci. **27**, 169–192 (2004)
45. Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L.: Premotor cortex and the recognition of motor actions. Cognitive Brain Research **3**(2) (1996)
46. Rybkin, O., Pertsch, K., Derpanis, K.G., Daniilidis, K., Jaegle, A.: Learning what you can do before doing anything. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=SylPMnR9Ym
47. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S.: Time-contrastive networks: Self-supervised learning from video. Proceedings of International Conference in Robotics and Automation (ICRA) (2018), http://arxiv.org/abs/1704.06888
48. Sharma, P., Pathak, D., Gupta, A.: Third-Person Visual Imitation Learning via Decoupled Hierarchical Controller. Neural Information Processing Systems (2019)
49. Shon, A.P., Grochow, K., Hertzmann, A., Rao, R.P.: Learning shared latent structure for image synthesis and robotic imitation. Advances in Neural Information Processing Systems pp. 1233–1240 (2005)
50. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: International Conference on Machine Learning (ICML) (2015)
51. Stadie, B.C., Abbeel, P., Sutskever, I.: Third-person imitation learning. arXiv preprint arXiv:1703.01703 (2017)
52. Sun, M., Ma, X.: Adversarial Imitation Learning from Incomplete Demonstrations. International Joint Conference on Artificial Intelligence (may 2019), http://arxiv.org/abs/1905.12310
53. Sun, W., Vemula, A., Boots, B., Bagnell, J.A.: Provably Efficient Imitation Learning from Observation Alone. International Conference on Machine Learning (may 2019), http://arxiv.org/abs/1905.10948
54. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised Cross-Domain Image Generation. International Conference on Learning Representations (nov 2017), https://arxiv.org/abs/1611.02200

55. Torabi, F., Warnell, G., Stone, P.: Behavioral Cloning from Observation. International Joint Conference on Artificial Intelligence (may 2018), http://arxiv.org/abs/1805.01954
56. Torabi, F., Warnell, G., Stone, P.: Generative Adversarial Imitation from Observation. arXiv preprint (jul 2018), http://arxiv.org/abs/1807.06158
57. Torabi, F., Warnell, G., Stone, P.: Imitation Learning from Video by Leveraging Proprioception. International Joint Conference on Artificial Intelligence (may 2019), http://arxiv.org/abs/1905.09335
58. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: MoCoGAN: Decomposing motion and content for video generation. In: Computer Vision and Pattern Recognition (2018)
59. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial Discriminative Domain Adaptation. Computer Vision and Pattern Recognition (feb 2017), http://arxiv.org/abs/1702.05464
60. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep Domain Confusion: Maximizing for Domain Invariance. arXiv preprint (dec 2014), http://arxiv.org/abs/1412.3474
61. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. In: International Conference on Learning Representations (2017)
62. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Computer Vision and Pattern Recognition (2016)
63. Vondrick, C., Torralba, A.: Generating the future with adversarial transformers. Conference on Vision and Pattern Recognition (2017)
64. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An Uncertain Future: Forecasting from Static Images using Variational Autoencoders. European Conference on Computer Vision (jun 2016), http://arxiv.org/abs/1606.07873
65. Wang, A., Kurutach, T., Tamar, A., Abbeel, P.: Learning Robotic Manipulation through Visual Planning and Acting. Robotics: Science and Systems (2019)
66. Watter, M., Springenberg, J.T., Boedecker, J., Riedmiller, M.: Embed to control: A locally linear latent dynamics model for control from raw images. In: Neural Information Processing Systems (2015)
67. Wichers, N., Villegas, R., Erhan, D., Lee, H.: Hierarchical long-term video prediction without supervision. ICML (2018)
68. Xie, A., Ebert, F., Levine, S., Finn, C.: Improvisation through Physical Understanding: Using Novel Objects as Tools with Visual Foresight. Robotics: Science and Systems (apr 2019), http://arxiv.org/abs/1904.05538
69. Xue, T., Wu, J., Bouman, K.L., Freeman, W.T.: Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (jul 2016), http://arxiv.org/abs/1607.02586
70. Yen-Chen, L., Bauza, M., Isola, P.: Experience-Embedded Visual Foresight. Conference on Robot Learning (nov 2019), http://arxiv.org/abs/1911.05071
71. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. arXiv preprint (may 2018), http://arxiv.org/abs/1805.04687
72. Yu, T., Finn, C., Xie, A., Dasari, S., Zhang, T., Abbeel, P., Levine, S.: One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning. Robotics: Science and Systems (feb 2018), http://arxiv.org/abs/1802.01557

73. Zhang, M., Vikram, S., Smith, L., Abbeel, P., Johnson, M.J., Levine, S.: SOLAR: Deep Structured Representations for Model-Based Reinforcement Learning. International Conference on Machine Learning (aug 2018), http://arxiv.org/abs/1808.09105
74. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Computer Vision and Pattern Recognition (2018)
75. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)
76. Zhuang, F., Cheng, X., Luo, P., Pan, S.J., He, Q.: Supervised representation learning with double encoding-layer autoencoder for transfer learning. International Joint Conference on Artifical Intelligence (2015). https://doi.org/10.1145/3108257