Continual Predictive Learning from Videos

Geng Chen[1*] Wendong Zhang[1*] Han Lu[1] Siyu Gao[1] Yunbo Wang[1†]
Mingsheng Long[2] Xiaokang Yang[1]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]School of Software, BNRist, Tsinghua University
{chengeng, diergent, yunbow}@sjtu.edu.cn

Abstract

Predictive learning ideally builds the world model of physical processes in one or more given environments. Typical setups assume that we can collect data from all environments at all times. In practice, however, different prediction tasks may arrive sequentially so that the environments may change persistently throughout the training procedure. Can we develop predictive learning algorithms that can deal with more realistic, non-stationary physical environments? In this paper, we study a new continual learning problem in the context of video prediction, and observe that most existing methods suffer from severe catastrophic forgetting in this setup. To tackle this problem, we propose the continual predictive learning (CPL) approach, which learns a mixture world model via predictive experience replay and performs test-time adaptation with non-parametric task inference. We construct two new benchmarks based on RoboNet and KTH, in which different tasks correspond to different physical robotic environments or human actions. Our approach is shown to effectively mitigate forgetting and remarkably outperform the naïve combinations of previous art in video prediction and continual learning.

1. Introduction

Predictive learning is an unsupervised learning technique to build a world model of the environment by learning the consequences from historical observations, sequences of actions, and corresponding future observation frames. The standard predictive learning setup is assumed to operate the model in a stationary environment with relatively fixed physical dynamics [9, 15, 38, 41]. However, the assumption of stationarity does not always hold in more realistic scenarios, such as in the settings of continual learning (CL), where the model is learned through tasks that arrive sequentially. For example, in robotics (see Fig. 1), world models often serve as the representation learners of model-based control systems [11, 17–19], while the agent may be subjected to non-stationary environments in different training periods. Under these circumstances, it is not practical to maintain a single model for each environment or each task, nor is it practical to collect data from all environments at all times. A primary finding of this paper is that most existing predictive networks [9, 15, 38, 41] cannot perform well when trained in non-stationary environments, suffering from a phenomenon known as catastrophic forgetting [13].

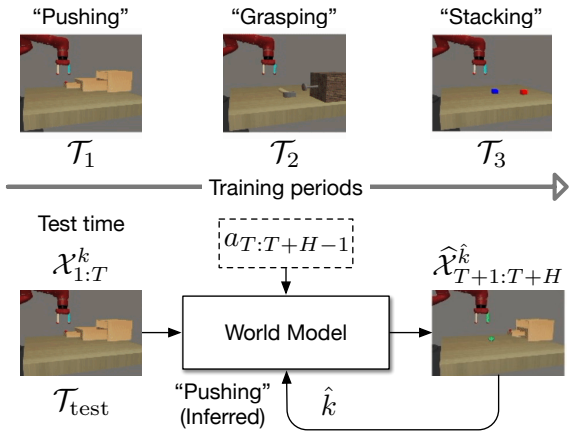

Figure 1. The new problem of continual predictive learning and the general framework of our approach at test time.

We formalize this problem setup as *continual predictive learning*, in which the world model is trained in time-varying environments (*i.e.*, "tasks" in the context of continual learning) with non-stationary physical dynamics. The model is expected to handle both newer tasks and older ones after the entire training phase (see Section 2 for detailed setups). There are two major challenges.

1.1. Covariate–Dynamics–Target Shift

Unlike in the settings of *domain-incremental* or *class-incremental* CL for deterministic models, the world model,

* Equal contribution.
† Corresponding author: Yunbo Wang.

which can be viewed as a conditioned generative model, cannot assume a stationary distribution of training targets or fixed target space. Therefore, different from all previous CL problems, the unique challenge of continual predictive learning is due to *the co-existence of three types of distribution shift, including the covariate shift in P_X, the target shift in P_Y, and the dynamics shift $P_{Y|X}$*[1]. Notably, the covariate shift [14, 28–31, 39, 43] and target shift [2, 16, 21, 27, 50] have been widely considered by existing methods, whereas the conditional distribution is typically assumed to be invariant. In our setup, however, the conditional distribution $P_{Y|X}$ corresponding to the spatiotemporal dynamics also changes over training periods. It significantly increases the probability of catastrophic forgetting in the world model.

To combat the dynamics shift, we first present a new world model that learns multi-modal visual dynamics of different tasks on top of task-specific latent variables. Future frames are generated by drawing samples from learned mixture-of-Gaussian priors conditioned on a set of categorical task variables, and combining them with a deterministic component of future prediction (see Section 3.1).

Second, we specifically design a novel training scheme named *predictive experience replay*. Like deep generative replay (DGR) [40], the proposed training method leverages a learned generative model to produce samples of previous tasks. Yet, in our approach, these samples are fed into the world model as the first frames to generate entire sequences, which can be reused as model inputs for rehearsal. The world model alternates between (i) generating rehearsal data without backpropagating the gradients, (ii) regressing the facilitate future frames of previous tasks produced by the world model itself, and (iii) generating future frames from real data of the current task. Another advantage of this training scheme is about the memory efficiency, as it only retains parts of low-dimensional action vectors in the buffer for action-conditioned predictive replay (see Section 3.2).

1.2. Task Inference: Coupled Forgetting Issues

The second challenge in continual predictive learning is the task ambiguity at test time, which can greatly affect the prediction results. Unlike existing CL methods for fully generative models [33, 40], in our setup, the models are required not only to solve each task seen so far, but also to infer which task they are presented with. A naïve solution is to infer the task using another neural network. However, due to the inevitable forgetting issue of the task inference model itself, coupled with that of the world model, this method is unlikely to perform well. In Section 3.3, we propose the non-parametric task inference strategy, which overcomes the intrinsic nature of forgetting of a deterministic model. We also present a self-supervised, test-time training process that recalls the pre-learned knowledge of the inferred task through one or several online adaptation steps.

We construct two new benchmarks for continual predictive learning based on real-world datasets, RoboNet [6] and KTH [37], in which different tasks correspond to different physical robotic environments or human actions. Our CPL approach is shown to effectively avoid forgetting and remarkably outperform the straightforward combinations of previous art in video prediction and continual learning.

[1] In predictive learning settings, the input X is in forms of sequential observation frames $\mathcal{X}_{1:T}$ and the training target Y corresponds to future frames $\mathcal{X}_{T+1:T+H}$. We here skip the input action signals for simplicity.

2. Problem Setup

Unlike existing predictive learning approaches, we consider to learn a world model ($\mathcal{M}$) in non-stationary environments (*i.e.*, the evolution of tasks), such that

$$\widehat{\mathcal{X}}_{T+1:T+H} \sim \mathcal{M}(\mathcal{X}_{1:T}, a_{T:T+H-1}, \hat{k}), \tag{1}$$

where $\mathcal{X}_{1:T}$ and $\mathcal{X}_{T+1:T+H}$ are respectively the observed frames and future frames to be predicted. The task index k is known at training, but not observed at test. It requires our approach not only to solve each task seen so far, but also to infer which task it is presented with, denoted as $\mathcal{T}_{\hat{k}}$. Here, $a_{T:T+H-1}$ is the optional inputs of action signals when $\mathcal{M}$ is learned for vision-based robot control, as in the action-conditioned video prediction experiments in this paper. Formally, continual predictive learning assumes that:

$$\begin{aligned}
&\text{Covariate shift: } P(\mathcal{X}^k_{1:T}) \neq P(\mathcal{X}^{k+1}_{1:T}) \\
&\text{Dynamics shift: } P(\mathcal{X}^k_{T+1:T+H}|\mathcal{X}^k_{1:T}) \neq P(\mathcal{X}^{k+1}_{T+1:T+H}|\mathcal{X}^{k+1}_{1:T}) \\
&\text{Target shift: } P(\mathcal{X}^k_{T+1:T+H}) \neq P(\mathcal{X}^{k+1}_{T+1:T+H}),
\end{aligned} \tag{2}$$

where we leave out $a_{T:T+H-1}$ for simplicity in the conditional distribution of visual dynamics. The setup is in part similar to class-incremental CL for supervised tasks that assumes $P(\mathcal{X}^k) \neq P(\mathcal{X}^{k+1})$, $\{\mathcal{Y}^k\} = \{\mathcal{Y}^{k+1}\}$, $P(\mathcal{Y}^k) \neq P(\mathcal{Y}^{k+1})$. $\{\mathcal{Y}^k\}$ denotes a constant label set for discriminative models. In contrast, continual predictive learning does not assume a fixed target space, and therefore may have more severe catastrophic forgetting issues.

3. Approach

In this section, we present the new continual predictive learning (CPL) approach, which first mitigates catastrophic forgetting within the world model from two aspects:

- **Mixture world model:** A recurrent network that captures multi-modal visual dynamics. Unlike existing models [9, 18], the learned task-specific priors are in forms of mixture-of-Gaussians to overcome dynamics shift.
- **Predictive experience replay:** A new rehearsal-based training scheme that combats the forgetting within the world model and is efficient in memory usage.

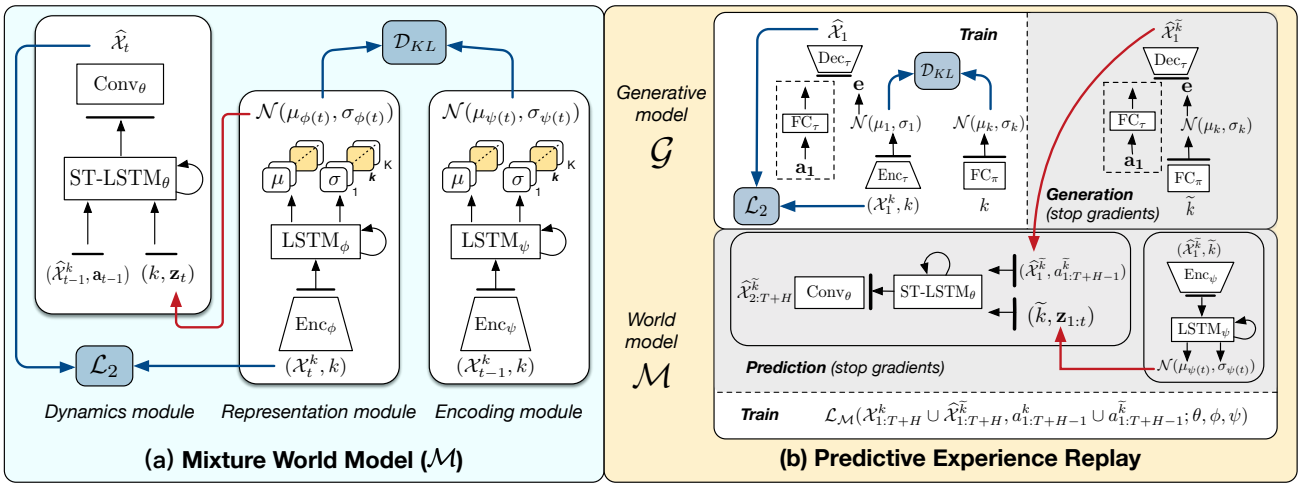

Figure 2. The overall network architecture of the *mixture world model* and the *predictive experience replay* training scheme in the proposed CPL method. (a) The world model learns representations in the forms of mixture-of-Gaussians based on categorical task variables. (b) As for the predictive experience replay, the world model ($\mathcal{M}$) interacts with the initial-frame generative model ($\mathcal{G}$). In this replay stage, we first use $\mathcal{G}$ to generate the first frames of previous tasks without backpropagating the gradients, then use $\mathcal{M}$ to predict the corresponding future frames, and finally combine the rehearsal data and real data to jointly train $\mathcal{M}$ and $\mathcal{G}$.

To cope with the challenge of task ambiguity when testing the world model in an unknown task, we propose:

- **Non-parametric task inference:** Instead of using any parametric task inference model that may introduce extra forgetting issues, we use a trial-and-error strategy over the task label set to determine the present task.

3.1. Mixture World Model

The world model $\mathcal{M}$ considers a new remedy to catastrophic forgetting from the perspective of spatiotemporal representation. As mentioned above, the forgetting problem within existing world models [9, 18] is mainly caused by the covariate-dynamics-target shift in time-varying environments. Therefore, the key idea of the proposed world model in CPL is to use mixture-of-Gaussian variables to capture the multi-modal distribution of visual dynamics in the latent space, as well as that of spatial appearance in the input/output observation space. Accordingly, as shown in Fig. 2(a), the world model consists of three components:

$$\begin{aligned} \text{Representation module:} \quad & \mathbf{z}_t \sim q_\phi(\mathcal{X}^k_{1:t}, k) \\ \text{Encoding module:} \quad & \hat{\mathbf{z}}_t \sim p_\psi(\mathcal{X}^k_{1:t-1}, k) \\ \text{Dynamics module:} \quad & \widehat{\mathcal{X}}_t = p_\theta(\mathcal{X}^k_{1:t-1}, a_{1:t-1}, \mathbf{z}_{1:t}, k). \end{aligned} \tag{3}$$

The representation module infers the latent state $\mathbf{z}_t$ from the target frames. It takes as input the categorical task variable $k \in \{1, \dots, K\}$ to cope with the target shift in continual predictive learning scenarios. The encoding module corresponds to the covariate shift and dynamically maps the input frames to $\hat{\mathbf{z}}_t$ in the same latent subspace as $\mathbf{z}_t$. The dynamics module learns the deterministic transition component from inputs to prediction targets. It responds to multi-modal spatiotemporal dynamics by taking as input the task-specific latent variable $\mathbf{z}_t$. All components are implemented as neural networks, in which the dynamics module is particularly composed of stacked ST-LSTM layers [47].

The task-specific latent representation $\mathbf{z}_t$ is drawn from a mixture-of-Gaussian distribution, inspired by existing unsupervised learning methods that use Gaussian Mixture priors for variational autoencoders [10, 22, 33]. Our mixture world model is an early work that uses this representation form to model the multi-modal priors in spatiotemporal dynamics. Specifically, for each task, the representation module and the encoding module are both conditioned on the present task label. They are jointly trained to learn the posterior and prior distribution of $\mathbf{z}_t$ by optimizing the Kullback-Leibler divergence. At task $\mathcal{T}_k$, the objective function $\mathcal{L}^k_{\mathcal{M}}(\mathcal{X}^k_{1:T+H}, a^k_{1:T+H-1})$ combines the KL loss with the reconstruction loss:

$$\begin{aligned} \sum_{t=2}^{T+H} \big[& \mathbb{E}_{q(\mathbf{z}_{1:t} \,|\, \mathcal{X}^k_{1:t}, k)} \log p(\mathcal{X}^k_t \,|\, \mathcal{X}^k_{1:t-1}, a^k_{1:t-1}, \mathbf{z}_{1:t}, k) \\ & - \alpha D_{KL}(q(\mathbf{z}_t \,|\, \mathcal{X}^k_{1:t}, k) \,||\, p(\hat{\mathbf{z}}_t \,|\, \mathcal{X}^k_{1:t-1}, k)) \big], \end{aligned} \tag{4}$$

where α is set to 10^{-4} in our experiments. In the test phase, we discard the representation module q_ϕ and only use the encoding module p_ψ to sample task-specific latent variables for frames generation.

3.2. Predictive Experience Replay

The two main challenges in typical CL setups are catastrophic forgetting and memory limitation. Due to the co-

Algorithm 1 Predictive experience replay

Input: Training data $\{\mathcal{X}^k_{1:T+H}\}_{k=1}^K$, $\{a^k_{1:T+H-1}\}_{k=1}^K$
Output: World model $\mathcal{M}$, generative model $\mathcal{G}$

Train $\mathcal{M}$ at $\mathcal{T}_1$ according to Eq. (4)
Train $\mathcal{G}$ at $\mathcal{T}_1$ according to Eq. (6) with $k=1$
for $k = 2, \ldots, K$ **do**
  # Replay video sequences (skip the batch size)
  for $\tilde{k} = 1, \ldots, k-1$ **do**
    $\widehat{\mathcal{X}}^{\tilde{k}}_1 \leftarrow \mathcal{G}(a^{\tilde{k}}_1, \tilde{k})$
    $\widehat{\mathcal{X}}^{\tilde{k}}_{2:T+H} \leftarrow \mathcal{M}(\widehat{\mathcal{X}}^{\tilde{k}}_1, a^{\tilde{k}}_{2:T+H-1}, \tilde{k})$
  end for
  # Mix replayed data at $\mathcal{T}_{1:k-1}$ and real data at $\mathcal{T}_k$
  $(\widehat{\mathcal{X}}^{1:k-1}_{1:T+H}, a^{1:k-1}_{1:T+H-1}) \cup (\mathcal{X}^k_{1:T+H}, a^k_{1:T+H-1})$
  Train $\mathcal{M}$ according to Eq. (5)
  Train $\mathcal{G}$ according to Eq. (6)
end for

existence of covariate shift, target shift, and dynamics shift, these challenges become even more urgent in the context of continual predictive learning based on video data. One common way to tackle these challenges is generative replay [33, 40], which considers using a generative model to produce samples of previous tasks. However, the generative replay method cannot be used directly in our setup, as it is extremely difficult to generate a valid video sequence using a generative model alone.

Therefore, we propose the predictive experience replay, which firmly combines an initial-frame generative model ($\mathcal{G}$), which learns to generate the first frame of videos at previous tasks given the task labels, with the world model ($\mathcal{M}$). To counter the coviariate shift of image appearance in non-stationary environments, $\mathcal{G}$ also uses learnable mixture-of-Gaussian latent priors, denoted by $\boldsymbol{e}$. As shown in Fig. 2(b), for each previous task $\mathcal{T}_{\tilde{k}}$, we first use $\mathcal{G}$ to generate the first frames of the rehearsal video sequences, and then use $\mathcal{M}$ to predict the corresponding future frames. Finally, we mix the rehearsal sequences at previous tasks and real sequences at the present task $\mathcal{T}_k$ to train $\mathcal{G}$ and $\mathcal{M}$ in turn. We summarize the training procedure in Alg. 1. The predictive experience replay is different from all existing generative replay methods because the world model plays a key role in the rehearsal process.

In particular, for action-conditioned predictive learning scenarios, we maintain a buffer to keep parts ($\sim$7%) of the low-dimensional action sequences from previous tasks. During predictive experience replay, we first sample an action sequence from the buffer $a^{\tilde{k}}_{1:T+H-1}$ at a previous task $\mathcal{T}_{\tilde{k}}$. We feed the initial action $a^{\tilde{k}}_1$ and the task label $\tilde{k}$ into $\mathcal{G}$ to ensure that the generated first frame $\widehat{\mathcal{X}}^{\tilde{k}}_1$ is valid for robot control, and perform $\mathcal{M}$ to produce predictive replay results $\widehat{\mathcal{X}}^{\tilde{k}}_{2:T+H}$ given $\widehat{\mathcal{X}}^{\tilde{k}}_1$ and $a^{\tilde{k}}_{2:T+H-1}$. In predictive experience replay, we train the world model $\mathcal{M}$ at $\mathcal{T}_k$ by minimizing

$$\begin{aligned}\mathcal{L}_{\mathcal{M}} = &\sum_{\tilde{k}=1}^{k-1} \mathcal{L}^{\tilde{k}}_{\mathcal{M}}(\widehat{\mathcal{X}}^{\tilde{k}}_{1:T+H}, a^{\tilde{k}}_{1:T+H-1}) \\ &+ \mathcal{L}^{k}_{\mathcal{M}}(\mathcal{X}^{k}_{1:T+H}, a^{k}_{1:T+H-1}).\end{aligned} \tag{5}$$

The objective function of the initial-frame generative model $\mathcal{G}$ can be written as

$$\begin{aligned}\mathcal{L}_{\mathcal{G}} = &\ \mathbb{E}_{q(\boldsymbol{e}\,|\,\mathcal{X}^k_1,k)} \log p(\mathcal{X}^k_1 \,|\, \boldsymbol{e}, a^k_1, k) \\ &- \beta D_{KL}(q(\boldsymbol{e}\,|\,\mathcal{X}^k_1, k)\,||\,p(\hat{\boldsymbol{e}}\,|\,k)) \\ &+ \sum_{\tilde{k}=1}^{k-1} \big[\mathbb{E}_{q(\boldsymbol{e}\,|\,\widehat{\mathcal{X}}^{\tilde{k}}_1,\tilde{k})} \log p(\widehat{\mathcal{X}}^{\tilde{k}}_1 \,|\, \boldsymbol{e}, a^{\tilde{k}}_1, \tilde{k}) \\ &- \beta D_{KL}(q(\boldsymbol{e}\,|\,\widehat{\mathcal{X}}^{\tilde{k}}_1, \tilde{k})\,||\,p(\hat{\boldsymbol{e}}\,|\,\tilde{k}))\big],\end{aligned} \tag{6}$$

where the reconstruction loss is in an ℓ_2 form and β is set to 10^{-4} through empirical grid search.

3.3. Non-Parametric Task Inference

In the mixture world model, the task label has a significant impact on the learned priors and corresponding prediction results. Since it is unknown at test time, it can only be inferred from the input observation sequences, *i.e.*, video classification. However, existing video classification models tend to underperform in the domain-incremental CL setting, which will magnify the catastrophic forgetting problem jointly trained with the world model. To avoid the inherent forgetting issue of model-based task inference, we propose a new non-parametric method that only exploits the learned mixture world model to make task inference.

More precisely, as shown in Alg. 2, we feed the first half of each input sequence into the world model $\mathcal{X}_{1:T/2}$, along with a hypothetic task label k. We then enumerate each task label $k \in \{1, \ldots, K\}$ and evaluate the outputs of the world model on the remaining frames of the input sequence $\mathcal{X}_{T/2+1:T}$. Finally, we choose the task label $\hat{k}$ that leads to the best prediction quality.

Algorithm 2 Testing procedure

Input: Observation frames $\mathcal{X}_{1:T}$, optional actions $a_{1:T+H}$
Output: Predicted future frames $\widehat{\mathcal{X}}_{T+1:T+H}$

 # Non-parametric task inference
for $k = 1, \ldots, K$ **do**
  $\widehat{\mathcal{X}}^k_{T/2+1:T} \leftarrow \mathcal{M}(\mathcal{X}_{1:T/2}, a_{1:T-1}, k)$
end for
$\hat{k} = \arg\min_{k\in\{1,\ldots,K\}} \sum_{t=T/2+1}^{T} (\mathcal{X}_t - \widehat{\mathcal{X}}^k_t)^2$
 # Test-time adaptation (optional)
Optimize $\mathcal{M}$ with $\mathcal{L}^{\hat{k}}_{\mathcal{M}}(\mathcal{X}_{1:T}, a_{1:T-1})$
 # Model deployment
$\widehat{\mathcal{X}}_{T+1:T+H} \leftarrow \mathcal{M}(\mathcal{X}_{1:T}, a_{1:T+H-1}, \hat{k})$

Method	Action-conditioned		Action-free	
	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$ ($\times 10^{-2}$)	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$ ($\times 10^{-2}$)
SVG [9]	18.72 $\pm$ 0.61	68.59 $\pm$ 2.22	18.92 $\pm$0.51	68.08 $\pm$ 2.20
PredRNN [47]	19.45	66.38	19.56	69.92
PhyDNet [15]	19.60	68.68	21.00	75.47
PredRNN + LwF [26]	19.10	64.73	19.79	71.43
PredRNN + EWC [24]	21.15	74.72	21.15	78.02
CPL-base + EWC [24]	21.29 $\pm$ 0.30	75.16 $\pm$ 0.98	21.38 $\pm$ 0.18	76.68 $\pm$ 0.69
CPL-base	19.36 $\pm$ 0.00	63.57 $\pm$ 0.00	20.15 $\pm$ 0.02	71.15 $\pm$ 0.08
CPL-full	**23.26 $\pm$ 0.10**	**80.72 $\pm$ 0.23**	**22.48 $\pm$ 0.03**	**78.84 $\pm$ 0.07**
CPL-base (Joint training)	24.64 $\pm$ 0.01	83.73 $\pm$ 0.00	22.56 $\pm$ 0.01	79.57 $\pm$ 0.02

Table 1. Quantitative results of continual predictive learning on the RoboNet benchmark in both action-conditioned and action-free setups. (**Lines 1-3**) Existing video prediction models with i.i.d. assumption. (**Lines 4-6**) Combinations of predictive models and continual learning approaches. (**Lines 7-8**) Our predictive model based on learned mixture-of-Gaussian priors, and the the entire CPL with predictive experience replay and non-parametric task inference. (**Line 9**) A baseline model jointly trained on all tasks throughout the training procedure, whose results can be roughly viewed as the upper bound of our approach.

In addition to using $P(\mathcal{X}_{T/2+1:T}|\mathcal{X}_{1:T/2})$ to perform task inference, we also use this self-supervision for test-time adaptation, which allows the model to continue training after deployment. Test-time adaptation effectively recalls the pre-learned knowledge in the inferred task $\mathcal{T}_{\hat{k}}$ through one-step (or few-steps) online optimization, thus further alleviating the forgetting problem.

4. Experiment

4.1. Experimental Setup

Benchmarks. We quantitatively and qualitatively evaluate CPL on the following two real-world datasets:

- **RoboNet** [6]. The RoboNet dataset contains action-conditioned videos of robotic arms interacting with a variety of objects in various environments. We divide the whole dataset into four continual learning tasks according to the environments (*i.e.*, *Berkeley* $\rightarrow$ *Google* $\rightarrow$ *Penn* $\rightarrow$ *Stanford*). For each task, we collect about 3,840 training sequences and 960 testing sequences.
- **KTH action** [37]. This dataset contains gray-scale videos which include 6 types of human actions. We directly use the action labels to divide the dataset into 6 tasks (*i.e.*, *Boxing* $\rightarrow$ *Clapping* $\rightarrow$ *Waving* $\rightarrow$ *Walking* $\rightarrow$ *Jogging* $\rightarrow$ *Running*). For each task, we collect about 1,500 training sequences and 800 testing sequences in average.

We define the task orders by random sampling, and without loss of generality, our approach is effective to any task orders (see Section 4.4). More experimental configurations and the implementation details can be found in the Supplementary Material.

Evaluation criteria. We adopt SSIM and PSNR from previous literature [9, 47] to evaluate the prediction results. We run the continual learning procedure 10 times and report the mean results and standard deviations in the two metrics.

Compared methods. We compare CPL with the following baselines and existing approaches:

- **CPL-base**: A baseline model that excludes the new components of Gaussian mixtures, predictive replay, and task inference.
- **PredRNN** [47], **SVG** [9], **PhyDNet** [15]: Video prediction models focused on stochastic, deterministic, and disentangled dynamics modeling respectively.
- **LwF** [26]: It is a distillation-based CL method built on the memory state of PredRNN [47].
- **EWC** [24]: It constrains the parameters of PredRNN and CPL-base on new tasks with additional loss terms.

4.2. RoboNet Benchmark

We first evaluate CPL on the real-world RoboNet benchmark, in which different continual learning tasks are divided by laboratory environments. We conduct both action-conditioned and action-free video prediction on RoboNet. The former follows the common practice [3, 48] to train the world model to predict 10 frames into the future from 2 observations and corresponding action sequence at the 11 time steps. For the action-free setup, we use the first 5 frames as input to predict the next 10 frames.

Quantitative comparison. Table 1 gives the quantitative results on RoboNet, in which the models are evaluated on the test sets of all 4 tasks after the training period on the last task. We have the following findings here. **First**, CPL outperforms existing video prediction models by a large margin. For instance, in the action-conditioned setup, it improves SVG in PSNR by 24.3%, PredRNN by 19.6%, and PhyDNet by 18.7%. **Second**, CPL generally performs better than previous continual learning methods (*i.e.*, LwF and EWC) combined with video prediction backbones. Note that a naïve implementation of LwF on top of PredRNN even leads to a negative effect on the final results. **Third**,

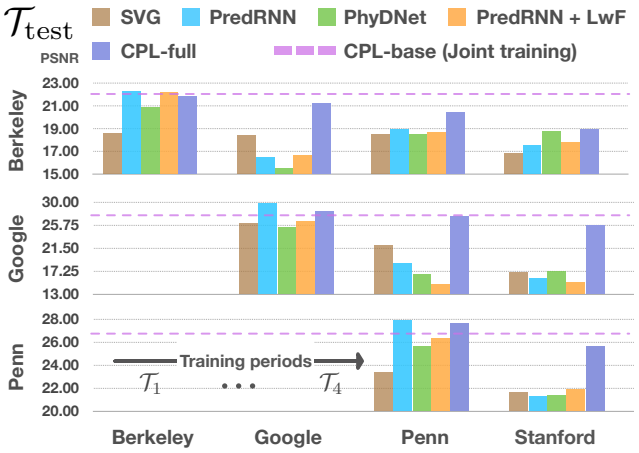

Figure 3. Results on the action-conditioned RoboNet benchmark. The horizontal axis represents the sequential training process, and the vertical axes represent test results on particular tasks after each training period. The purple dashed line indicates the results of the baseline model jointly trained on all tasks.

by comparing CPL-full (our final approach) with CPL-base (w/o Gaussian mixture latents, predictive experience replay, or non-parametric task inference), we can see that the new technical contributions have a great impact on the performance gain. We provide more detailed ablation studies in Section 4.4. **Finally**, CPL is shown to effectively ease catastrophic forgetting by approaching the results of jointly training the world model on all tasks in the i.i.d. setting (23.26 vs. 24.64 in PSNR). Apart from the average scores for all tasks, in Fig. 3, we provide the test results on particular tasks after individual training periods. As shown in the bar charts right to the main diagonal, CPL performs particularly well on previous tasks, effectively alleviating the forgetting issue. Please refer to the Supplementary Material for detailed comparison results.

Qualitative comparison. Fig. 4 provides the qualitative comparisons on the action-conditioned RoboNet benchmark. Specifically, we use the final models after the training period of the last task to make predictions on the first task. We can see from these demonstrations that our approach is more accurate in predicting both future dynamics of the objects as well as the static information of the scene. In contrast, the predicted frames by PredRNN+LwF and CPL-base+EWC suffer from severe blur effect in the moving object or the static (but complex) background, indicating that directly combining existing CL algorithms with the world models cannot effectively cope with the dynamics shift in highly non-stationary environments.

4.3. KTH Benchmark

Quantitative comparison. Table 2 shows the quantitative results on the test sets of all 6 tasks after the last training

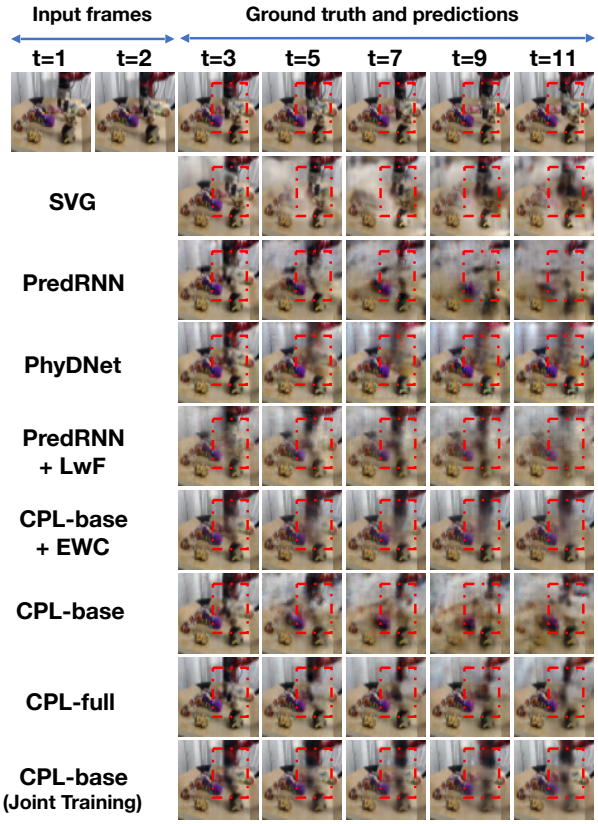

Figure 4. Showcases of action-conditioned video prediction in the first environment of RoboNet (*i.e.*, *Berkeley*) after training the models in the last environment (*i.e.*, *Stanford*).

Method	PSNR	SSIM ($\times 10^{-2}$)
SVG [9]	22.20±0.02	69.23±0.01
PredRNN [47]	23.27	70.47
PhyDNet [15]	23.68	72.97
PredRNN + LwF [26]	24.25	70.93
CPL-base + EWC [24]	24.32 ± 0.15	69.02 ± 0.48
CPL-base	22.96 ± 0.05	68.98 ± 0.02
CPL-full	**29.12 ± 0.03**	**84.50 ± 0.04**
CPL-base (Joint train)	28.12 ± 0.01	82.16 ± 0.00

Table 2. Quantitative results on the KTH benchmark.

period of the models on the last task. We can observe that CPL significantly outperforms the compared video prediction methods and continual learning methods in both PSNR and SSIM. Furthermore, an interesting result is that our approach even outperforms the joint training model, as shown in the bottom line in Table 2. While we do not know the exact reasons, we state two hypotheses that can be investigated in future work. First, the Gaussian mixture priors enable the world model to better disentangle the representations of visual dynamics learned in different continual learning tasks. Second, the predictive experience replay allows the pre-learned knowledge on previous tasks to facilitate the

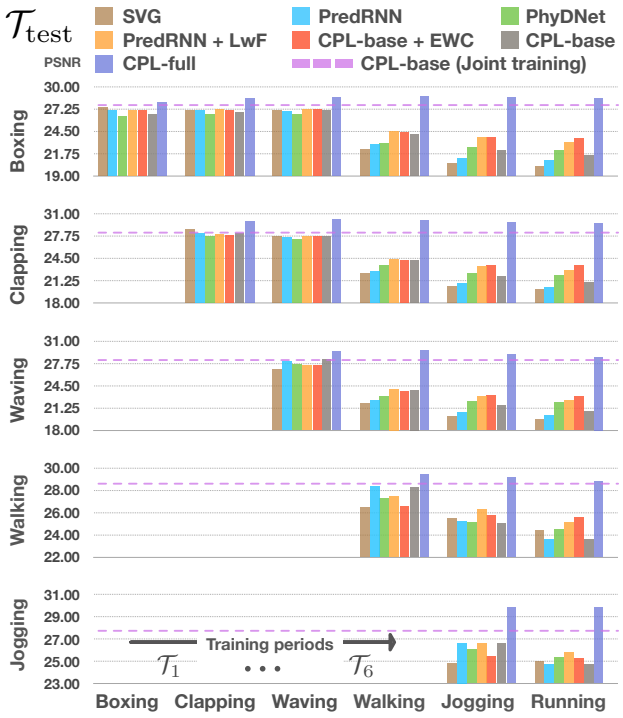

Figure 5. Results on the KTH benchmark. The horizontal axis represents the sequential training process, and the vertical axes represent test results on particular tasks after each training period.

Replay	Infer k	Random k	Adapt	PSNR	SSIM
✗	✗	✗	✗	22.96	68.98
✓	✗	✗	✗	27.21	79.99
✓	✓	✗	✗	27.82	81.51
✓	✗	✓	✗	26.56	78.64
✓	✓	✗	✓	**29.12**	**84.50**

Table 3. Ablation study for each component of CPL on the KTH benchmark. "Replay" denotes the use of predictive experience replay. "Infer k" indicates the use of non-parametric task inference. "Random k" means that the world model takes as input a random task label at test time. "Adapt" means test-time adaptation.

learning process on new tasks. Fig. 5 provides the intermediate test results on particular tasks after each training period, which confirm the above conclusions.

Qualitative comparison. We visualize a sequence of predicted frames on the first task of KTH in Fig. 6. As shown, all existing video prediction models and even the one with LwF generate future frames with the dynamics learned in the last task (*i.e.*, *Running*), which clearly demonstrates the influence of the dynamics shift. Images generated by CPL-base+EWC suffer from a severe blur effect, indicating that the model cannot learn disentangled representations for different dynamics in the non-stationary training environments. In comparison, CPL produces more reasonable results. To testify the necessity of task inference, we also

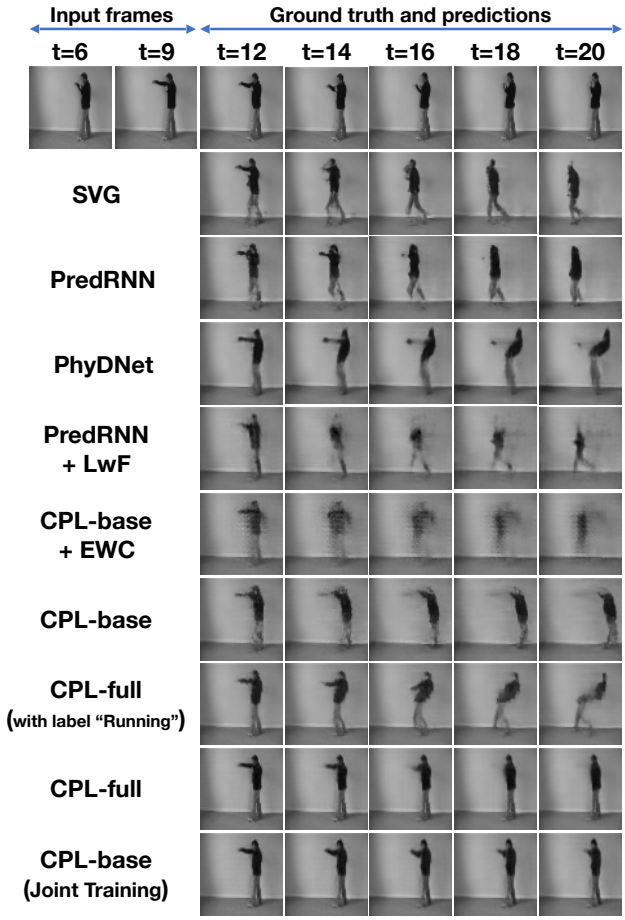

Figure 6. Showcases of predicted frames of the first task (*i.e.*, *Boxing*) after the training period of the last task (*i.e.*, *Running*).

provide incorrect task labels for CPL. As shown in the third line from the bottom, the model takes as input the *Boxing* frames along with an erroneous task label of *Running*. Interestingly, CPL combines the inherent dynamics of input frames (reflected in motion of arms) with the dynamics priors from the input task label (reflected in motion of legs).

4.4. Ablation Study

Effectiveness of each component in CPL. We conduct ablation studies on the KTH benchmark step by step. In Table 3, the first line shows the results of the CPL-base model and the bottom line corresponds to our final approach. In the second line, we train CPL-base with predictive experience replay and observe a significant improvement from 22.96 to 27.21 in PSNR. In the third line, we improve the world model with mixture-of-Gaussian priors and accordingly perform non-parametric task inference at test time. We observe consistent improvements in both PSNR and SSIM upon the previous version of the model. In the fourth line, we skip the non-parametric task inference during testing and use a random task label instead. We observe that

Dataset	PSNR	SSIM ($\times 10^{-2}$)
RoboNet	23.58 $\pm$0.28	79.67 $\pm$3.75
KTH	28.93$\pm$0.14	83.99 $\pm$0.40

Table 4. Robustness of CPL on random task orders.

the performance drops from 27.82 to 26.56 in PSNR, indicating the importance of task inference to predictive experience replay. Finally, in the bottom line, we introduce the self-supervised test-time adaptation. It shows a remarkable performance boost compared with all the above variants.

Is CPL robust to the task order? As shown in Table 4, we further conduct experiments to analyze that whether CPL can effectively alleviate catastrophic forgetting regardless of the task order. We additionally train the CPL model in 3-4 random task orders. From the results, we find that the proposed techniques including mixture world model, predictive experience replay, and non-parametric task inference are still effective despite the change of training order.

5. Related Work

Continual learning of supervised tasks. Continual learning is designed to cope with the continuous information flow, retaining or even optimizing old knowledge while absorbing new knowledge. Mainstream paradigms include regularization, replay, and parameter isolation [8]. The regularization approaches typically tackle catastrophic forgetting [13] by constraining the learned parameters on new tasks with additional loss terms, *e.g.*, EWC [24], or distilling knowledge from old tasks, *e.g.*, LwF [26]. For replay-based approaches, a typical solution is to retain a buffer on earlier tasks of representative data or feature exemplars [1, 34, 35]. Some approaches also use generative networks to encode the previous data distribution and synthesize fictitious data for experience replay, *e.g.*, DGR [40] and CURL [33]. The parameter isolation approaches allow the neural networks to dynamically expand when new tasks arrive [36] or encourage the new tasks to use previously "unused" parameter subspaces [20].

Continual learning of unsupervised tasks. Most existing approaches are mainly focused on supervised tasks of image data. Despite the previous literature that discussed unsupervised CL [5, 23, 33], our approach is significantly different from these methods as it explores the specific challenges of continual predictive learning for video data, especially the covariate-dynamics-target shift. The most related method to CPL is CURL [33], which introduces a mixture-of-Gaussian latent space for class-incremental CL and combat forgetting via generative replay. There are three major differences between CPL and CURL. First, CURL cannot be directly used in our setup as it does not handle the dynamics shift in non-stationary spacetime, while CPL tackles it through a new world model. Second, CPL greatly benefits from the carefully-designed predictive replay algorithm, while it is extremely difficult for CURL to replay valid video frames using a fully generative model alone. Third, CPL provides a non-parametric task inference method as opposed to the model-based inference method in CURL.

Video prediction. RNN-based models have been widely used for deterministic video prediction [7, 32, 38, 41, 42, 46, 47, 49]. Shi *et al*. [38] proposed ConvLSTM to improve the learning ability of spatial information by combining convolutions with LSTM transitions. Following this line, Wang *et al*. [47] proposed PredRNN, modeling memory cells in a unified spatial and temporal representation. Stochastic video prediction models assume that different plausible outcomes would be equally probable for the same input, and thus incorporate uncertainty in the models using GANs [44, 45] or VAEs [3, 4, 9, 12, 25]. Particularly, Yao *et al*. proposed to adapt video prediction models from multiple source domains to a target domain via distillation [49]. However, it cannot be easily used as a solution to continual predictive learning, as the number of retained model parameters increases linearly with the number of tasks.

6. Discussion

In this paper, we explored a new research problem of continual predictive learning, which is challenging due to the co-existence of the covariate, dynamics, and target shift. We proposed an approach named CPL, whose major contributions of CPL can be viewed in three aspects. First, it presents a new world model to capture task-specific visual dynamics in a Gaussian mixture latent space. Second, it introduces the predictive experience replay method to overcome the forgetting issue in the world model. Third. it leverages a non-parametric task inference strategy to avoid coupling the forgetting issues caused by the introduction of a task inference model. Our approach has shown competitive results on RoboNet and KTH benchmarks, achieving remarkable improvements over the naïve combinations of existing world models and CL algorithms.

Although CPL can be easily extended to more complex tasks, the potential limitation is that it has not been evaluated in the entire pipeline of vision-based robot control, which includes the processes of predictive learning and decision making. In future work, we plan to integrate CPL in a model-based reinforcement learning framework to further validate its effectiveness for downstream tasks.

7. Acknowledgement

This work was supported by NSFC grants (U19B2035, 62106144, 62021002, 62022050), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and Shanghai Sailing Program (21Z510202133).

References

[1] Ali Ayub and Alan R Wagner. EEC: Learning to encode and regenerate images for continual learning. In *ICLR*, 2021. 8

[2] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019. 2

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *ICLR*, 2018. 5, 8

[4] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional VRNNs for video prediction. In *CVPR*, pages 7608–7617, 2019. 8

[5] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *ICCV*, pages 9516–9525, 2021. 8

[6] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *CoRL*, pages 885–897, 2019. 2, 5

[7] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NeurIPS*, pages 667–675, 2016. 8

[8] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 8

[9] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, pages 1182–1191, 2018. 1, 2, 3, 5, 6, 8

[10] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016. 3

[11] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, pages 2786–2793, 2017. 1

[12] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *ICML*, pages 3233–3246, 2020. 8

[13] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1, 8

[14] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009. 2

[15] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *CVPR*, pages 11474–11484, 2020. 1, 5, 6

[16] Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, and Dacheng Tao. LTF: A label transformation framework for correcting label shift. In *ICML*, pages 3843–3853, 2020. 2

[17] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018. 1

[18] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020. 1, 2, 3

[19] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, pages 2555–2565, 2019. 1

[20] Xu He and Herbert Jaeger. Overcoming catastrophic interference using conceptor-aided backpropagation. In *ICLR*, 2018. 8

[21] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *ICML*, pages 530–538, 2014. 2

[22] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *IJCAI*, pages 1965—-1972, 2017. 3

[23] Zixuan Ke, Bing Liu, Hu Xu, and Lei Shu. Classic: Continual and contrastive learning of aspect sentiment classification tasks. In *EMNLP*, pages 6871–6883, 2021. 8

[24] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 5, 6, 8

[25] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 8

[26] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 5, 6, 8

[27] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, pages 3122–3130, 2018. 2

[28] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *ICML*, pages 4013–4022, 2019. 2

[29] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 2

[30] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2017. 2

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017. 2

[32] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 8

[33] Dushyant Rao, Francesco Visin, Andrei A Rusu, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Continual unsupervised representation learning. In *NeurIPS*, 2019. 2, 3, 4, 8

[34] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 8

[35] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR*, 2019. 8

[36] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 8

[37] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, pages 32–36, 2004. 2, 5

[38] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, pages 802–810, 2015. 1, 8

[39] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 2

[40] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, pages 2990–2999, 2017. 2, 4, 8

[41] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *ICML*, pages 843–852, 2015. 1, 8

[42] Jiahao Su, Wonmin Byeon, Furong Huang, Jan Kautz, and Animashree Anandkumar. Convolutional tensor-train lstm for spatio-temporal learning. In *NeurIPS*, 2020. 8

[43] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Von Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NeurIPS*, volume 7, pages 1433–1440, 2007. 2

[44] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018. 8

[45] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, pages 613–621, 2016. 8

[46] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li. Fei-Fei. Eidetic 3D LSTM: A model for video prediction and beyond. In *ICLR*, 2019. 8

[47] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *NeurIPS*, pages 879–888, 2017. 3, 5, 6, 8

[48] Bohan Wu, Suraj Nair, Roberto Martín-Martín, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *CVPR*, pages 2318–2328, 2021. 5

[49] Zhiyu Yao, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Unsupervised transfer learning for spatiotemporal predictive networks. In *ICML*, pages 10778–10788, 2020. 8

[50] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *ICML*, pages 819–827, 2013. 2