Visual Forecasting by Imitating Dynamics in Natural Sequences

Kuo-Hao Zeng†‡   William B. Shen†   De-An Huang†   Min Sun‡   Juan Carlos Niebles†
†Stanford University   ‡National Tsing Hua University
{khzeng, bshen88, dahuang, jniebles}@cs.stanford.edu   sunmin@ee.nthu.edu.tw

Abstract

We introduce a general framework for visual forecasting, which directly imitates visual sequences without additional supervision. As a result, our model can be applied at several semantic levels and does not require any domain knowledge or handcrafted features. We achieve this by formulating visual forecasting as an inverse reinforcement learning (IRL) problem, and directly imitate the dynamics in natural sequences from their raw pixel values. The key challenge is the high-dimensional and continuous state-action space that prohibits the application of previous IRL algorithms. We address this computational bottleneck by extending recent progress in model-free imitation with trainable deep feature representations, which (1) bypasses the exhaustive state-action pair visits in dynamic programming by using a dual formulation and (2) avoids explicit state sampling at gradient computation using a deep feature reparametrization. This allows us to apply IRL at scale and directly imitate the dynamics in high-dimensional continuous visual sequences from the raw pixel values. We evaluate our approach at three different level-of-abstraction, from low level pixels to higher level semantics: future frame generation, action anticipation, visual story forecasting. At all levels, our approach outperforms existing methods.

1. Introduction

Our goal is to expand the boundaries of video analysis towards visual forecasting by building intelligent systems that are capable of imitating the dynamics in natural visual sequences. The idea of visual forecasting has been studied in several contexts, such as trajectory forecasting [1, 19, 46], activity prediction [23, 40], and future frame synthesis [18, 50]. However, most previous work either assumes short forecasting horizons [42, 47], or focuses on a specific application, such as walking trajectories [1, 19, 49] or human actions [14, 23], where handcrafted features play an important role.

In this paper, we propose a general framework for visual forecasting, where it can be applied to learn a model at any
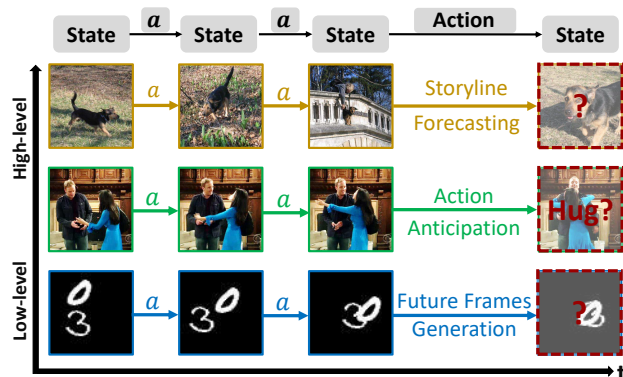

Figure 1. We propose a general framework for visual forecasting. We formulate this as an imitation learning problem, where the natural visual sequences are served as expert behaviors, and visual forecasting is reformulated as learning a policy to reproduce the expert's behavior. Our framework imitate dynamics in visual sequences directly from its raw pixel value and is applicable to tasks with different levels of abstraction.

given semantic-level (see Figure 1) and does not require any domain knowledge or handcrafted features. Such framework for forecasting can be widely useful in real world scenarios: robot motion planning [17, 20] or warning systems for autonomous vehicles [4]. We achieve this by formulating visual forecasting as an inverse reinforcement learning (IRL) [31] problem, where the natural sequences serve as expert demonstrations, and the goal becomes recovering the underlying cost function that prioritizes the entire expert trajectories over others. This eliminates the compounding error problem of single-timestep decisions [11], and has been proven effective in long-term prediction in both robotics [6, 21, 54] and vision [19, 27].

Our key insight is to extend previous IRL work in vision to directly imitate the dynamics in visual sequences from their raw pixel values, where no domain knowledge or handcrafted feature would be required. This is in contrast to previous work that applies IRL on handcrafted features, such as the semantic scene labeling for trajectory forecasting [19, 26] and interaction statistics for dual agent interaction [14]. The main challenge is that both the state and action space would become *high-dimensional* and *continu-*

ous in such setting, which prevent scaling up existing IRL algorithms used in previous work [11, 54].

We address this challenge by extending recent progress in model-free imitation learning [11, 12] with the capacity of jointly trainable deep feature representations. More specifically, we extend the Generative Adversarial Imitation Learning framework of Ho and Ermon [11] with deep feature reparametrization to forecast the dynamics in natural visual sequences. The deep representation is fully differentiable and we optimize it jointly using policy gradient, which results in learning an intermediate representation that is best-suited for visual forecasting. The resulting model (1) bypasses the dynamic programming based subroutine of [54] that exhaustively visits every state-action pair by using a dual formulation, and (2) avoids explicit state sampling (image synthesis in this case) of gradient computation in [11] with reparametrization using deep feature representations. This allows us to operate beyond low-dimensional state-action spaces and directly forecast the natural sequence from the pixel-level.

We demonstrate the effectiveness of our approach by applying our model to three tasks that require different levels of understanding of visual dynamics. First, we examine the ability of our method to perform generation of future frames on the moving MNIST data sequences [41]. Results from future frame generation demonstrate that our framework can capture the low-level (pixel-level) visual dynamics between consecutive frames. Second, we examine the ability of our method to perform action anticipation by unsupervised training on the THUMOS15 dataset [8], and evaluating the resulting model on the TV Human Interactions dataset [35]. Results from action anticipation show that our framework is effective in modeling mid-level (motion-level) visual dynamics in natural videos. Finally, we apply our method on the Visual Storytelling dataset [15] and try to forecast the next photo in a story composed of images. Results on this task demonstrate that our framework can generalize to visual dynamics on a highly abstract level (semantic-level). In each task, our method outperforms all existing methods.

In summary, the main contributions of our work are: (1) We propose a general IRL based formulation for visual sequence forecasting that requires no domain knowledge or handcrafted features. (2) We address the computational bottlenecks of IRL when operating at high-dimensional continuous spaces by using a dual formulation and reparametrization with deep feature representations. (3) We evaluate our method on three tasks at different levels of abstraction and show that our framework outperforms all existing methods.

2. Related Work

Visual Prediction. There has been growing interest in developing computational models of human activities that can extrapolate unseen information and predict future unobserved activities [5, 28, 33, 41, 43, 45, 50, 53]. Some of the existing approaches [5, 28, 43, 44, 45, 50] tried to generate realistic future frames using generative adversarial networks [7]. Unlike these methods, we emphasize longer-term sequential dynamics in videos using inverse reinforcement learning. Other line of work attempted to infer the action or human trajectories that will occur in the subsequent time-step based on previous observation [1, 14, 19, 23, 42]. Our model directly imitates the natural sequence from the pixel-level and assumes no domain knowledge.

Reinforcement Learning in Computer Vision. Reinforcement learning (RL) achieves remarkable success in multiple domains ranging from robotics [9], computer vision [3, 30, 32] and natural language processing [25, 52]. In the RL setting, the reward function that the agent aims to maximize is given as signal for training, where the goal is to learn a behavior that maximizes the expected reward. On the other hand, we work on the inverse reinforcement learning (IRL) problem, where the reward function must be discovered from demonstrated behavior [11, 31, 54]. This is inspired by recent progress of IRL in computer vision [6, 14, 19, 26, 27, 46]. Nonetheless, these frameworks require heavy use of domain knowledge to construct the handcrafted features that are important to the task. Unlike these approaches, we aim to generalize IRL to natural sequential data without annotations.

Unsupervised Representation Learning in Video. Our combination of deep neural networks with IRL to directly imitate natural videos is related to recent progress in unsupervised representation learning in videos. The temporal information in the video has been utilized to learn useful features based on tracking [48], flow prediction [47], future frame synthesis [41], frame order [29], and object movement [34]. Our work utilizes not only the frame-to-frame dynamics, but tries to imitate the entire sequence directly to utilize long-range temporal information.

Generative Adversarial Learning. Our extension of generative adversarial imitation learning [11] is related to recent progress in generative adversarial networks (GAN) [7]. While there has been multiple works on applying GAN to image and video [16, 37, 43, 51], we extend it to long-term prediction of natural visual sequences and directly imitate the high-dimensional continuous sequence.

3. Forecasting Natural Sequences

Our goal is to build a system that can forecast the dynamics of natural videos. We formulate this as an inverse reinforcement learning problem, where the goal is to recover the underlying cost function that induces the natural visual sequence. This is equivalent to learning a policy $\pi(v'|v)$ that parameterizes the transition from a current frame v to the next frame v'. As shown in Figure 2, at training time,
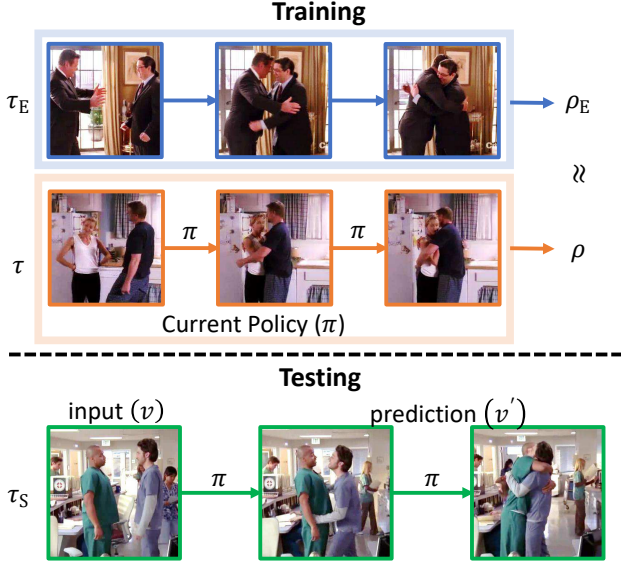
Training

τ_E

ρ_E

π π

τ

ρ

Current Policy (π)

Testing

input (v)   prediction (v')

τ_S

π π

Figure 2. We formulate visual prediction as an imitation learning problem, where the goal becomes learning a policy π that can generate trajectory τ following the behavior of natural visual sequence, or expert demonstration τ_E. This is reformulated as finding a policy π that has a state-action pair visit distribution ρ that is similar to that of the expert (ρ_E). At test time, we thus have a learned policy π that can mimic the behavior of natural visual sequence and perform visual prediction.

our model observes and imitates a large amount of natural visual sequences to learn the policy. At testing time, the learned policy $\pi(v'|v)$ can then be directly applied to an image for long-term visual prediction. Despite its many potential applications, learning $\pi(v'|v)$ is challenging as v is an image that is high-dimensional and continuous. This makes both the exhaustive state-action pairs visit and the explicit state sampling in previous work [11, 54] intractable at scale. Our solution to this extends and reparametrizes the model from Ho and Ermon [11] with deep representations. Our framework bypasses exhaustive state-action pair visit of [54] by the dual formulation of [11], and avoids explicit state sampling (which is equivalent to frame synthesis) during the gradient computation in [11].

In the following, we discuss how we formulate natural visual sequences as Markov Decision Processes so that our inverse reinforcement learning framework can be applied.

3.1. Visual Sequences as Markov Decision Process

In this work, we model natural visual sequences as Markov Decision Processes (MDP) [36]. At each time step, the sequence is at some frame/image v, which is defined as the *state* of our MDP. Given a chosen action a from the state, the process responds by moving to a new state, or frame v' based on the transition model $p(v'|v,a)$ and receives a cost $c(v,a)$. In this work, the transition model is *deterministic* as we are directly imitating the natural visual

sequences. Therefore, $p(v'|v,a) = 1$ for the single visual state $v' = v_a$ that we can visit after taking action a, and is zero otherwise. In this case, the policy $\pi(v_a|v)$ of moving to a state v_a is equivalent to choosing the corresponding action $\pi(a|v)$. As result, we are able to use $\pi(v'|v)$ and $\pi(a|v)$ interchangeably depending on the context (same for $c(v,v')$ and $c(v,a)$). Our goal of inverse reinforcement learning (IRL) is in contrast to reinforcement learning (RL), where the cost function is observed and the goal is to learn the optimal policy $\pi(a|v)$ that minimizes the expected cost. In IRL, the cost function is *not* observed, and the goal is to imitate the behaviors of the expert by recovering the underlying cost function $c(v,a)$ that induces the observed optimal behavior by the expert.

3.2. Imitating Natural Visual Sequences

We formulate the visual prediction as an inverse reinforcement learning problem, where the goal is to imitate the behavior of natural sequences that are treated as expert demonstrations. In this section, we discuss the model-free imitation learning framework of Ho and Ermon [11] that allows us to bypass the exhaustive state-action pair visit in dynamic programming of Maximum Entropy (MaxEnt) IOC [54] used in previous works [14, 19]. In the next section, we discuss how we apply it efficiently to natural visual sequences from the pixel level.

In contrast to reinforcement learning, the cost function in IRL is not given and has to be recovered from expert demonstrations τ_E. We follow previous work and adopt the maximum causal entropy framework [54] for the optimization of our cost function:

$$\max_{c \in C} \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(v,a)] \right) - \mathbb{E}_{\pi_E}[c(v,a)], \quad (1)$$

where Π is a policy space and $H(\pi) \triangleq \mathbb{E}_\pi[-log\,\pi(a|v)]$ is the γ-discounted causal entropy [2] of the policy π.

The standard approach for this IRL problem used in previous work [19, 54] involves a RL inner loop that utilizes dynamic programming to exhaustively visit every state and the available actions for value function computation. However, such procedure is intractable in our problem, as both our state and action spaces are high-dimensional and continuous. The key observation to resolve this problem is that IRL Eq. (1) is actually a dual of an occupancy measure matching problem [12]. Here, the occupancy measure $\rho_\pi(v,a)$ of a policy π is defined as the distribution of the visit of state-action pairs while following the policy π. This reformulates Eq. (1) as:

$$\min_\pi d_\psi(\rho_\pi, \rho_{\pi_E}) - H(\pi), \quad (2)$$

where $d_\psi(\cdot, \cdot)$ measures the difference between two occupancy measures depending on the regularizer ψ. Intuitively,
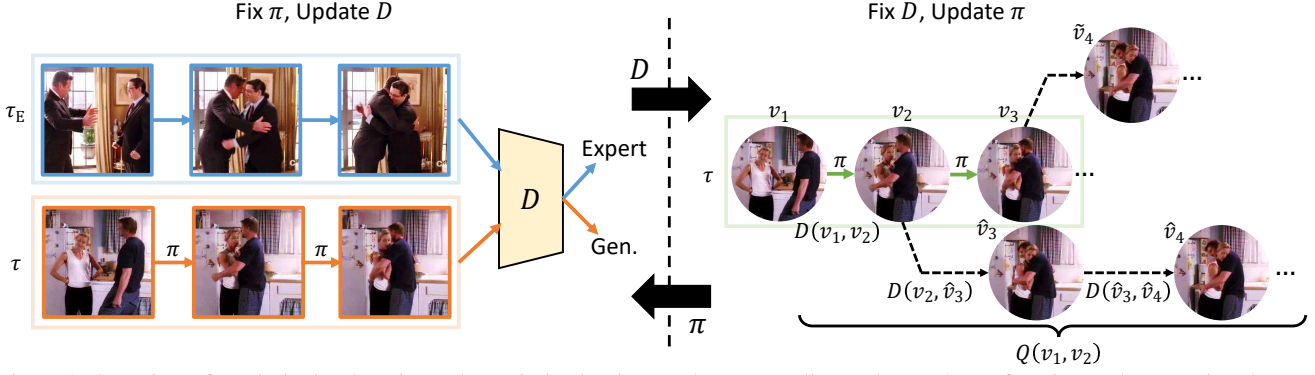
Figure 3. Overview of our imitation learning. The optimization iterates between policy update and cost function update. During the cost (D) update, the goal is to find a cost that best differentiate expert's trajectory τ_E from our generated trajectory τ. During the policy (π) update, we estimate the expected cost Q and use policy gradient with roll out to minimize it (see more detail in supplementary). This framework also has a generative adversarial learning interpretation, where the cost is operating as the discriminator and the policy is functioning as the generator.

this aims to induce a policy that visits state-action pairs with similar frequency to that of the expert policy. This allows us to use the demonstrated state-action pairs of expert directly as learning signal to encourage the policy toward expert-like region in state-action space and bypass the exhaustive visit of dynamic programming.

As presented in [11], under certain selection regularizer ψ, $d_\psi(\cdot, \cdot)$ becomes:

$$\max_D \mathbb{E}_{\pi_\theta}[\log(D(v, a))] + \mathbb{E}_{\pi_E}[\log(1 - D(v, a))], \quad (3)$$

where $D(v, a)$ can be interpreted as a discriminator that tries to distinguish the state-action pairs from the expert or the induced policy. This interpretation connects imitation learning with generative adversarial network (GAN) [7]. Essentially, our imitation learning of natural visual sequences can be seen as jointly training (1) a discriminator that tries to distinguish our generated sequence from expert's and (2) a policy that forms as a generator that uses the learning signal provided by the discriminator to move toward expert-like regions. We refer the readers to [11] for more details of the algorithm. The important note is that no specific restriction would be put on the form of approximation for both D and π and this allows us to directly imitate high-dimensional continuous sequence.

Optimization. Let D_w be D parameterized by weights w, π_θ be π parameterized by θ, $\tau = \{v_t\}$ be trajectory of states v^t from current policy, and $\tau_E = \{v_t^E\}$ be expert trajectory given as demonstration, the optimization of Eq. (2) with Eq. (3) alternates between the maximizing the discriminative power of D_w through (the left panel in Figure 3):

$$\mathbb{E}_{\tau \sim \pi_\theta}[\nabla_w \log(D_w(v_t, v_{t+1}))] \quad (4)$$

$$+ \mathbb{E}_{\tau_E \sim \pi_E}[\nabla_w \log(1 - D_w(v_t^E, v_{t+1}^E))], \quad (5)$$

and a policy gradient step (the right panel in Figure 3):

$$\mathbb{E}_{\tau \sim \pi_\theta}[\nabla_\theta \log(\pi_\theta(v_{t+1}|v_t))Q(v_t, v_{t+1})] \quad (6)$$

that aims to maximize the expected return $Q(v_t, v_{t+1})$, which is in turn defined as

$$Q(v, v') = \mathbb{E}_{\tau \sim \pi_\theta}[\log(D_w(v_t, v_{t+1}))|v_0 = v, v_1 = v'], \quad (7)$$

the expectation of confusing the discriminator that the generated sequence τ is a expert's sequence.

3.3. Efficient Sampling with Deep Representation

Now we have formulated visual prediction as an IRL problem and presented a generative adversarial imitation learning framework [11] that is applicable to our problem, because it bypasses the dynamic programming based subroutine that iterates through all the states and actions. However, we would like to point out there is still a remaining computational bottleneck that prevents the direct application of this framework to the image space.

Notice in the gradient computation for both the discriminator and the policy (Eq. (5) and Eq. (6)) that we are required to sample trajectories τ from the current policy π_θ. This is equivalent to sampling a sequence of images/frames at each gradient computation. The problem of image synthesis itself is challenging and computational intensive [18], and prevents the direct application of their framework to large scale visual sequence datasets.

Our key insight to address this is to parameterize both the discriminator D_w and the policy π_θ with a intermediate representation $\phi(\cdot)$ that is feasible to sample at the optimization, but still captures sufficient information from the image:

$$\pi(v_{t+1}|v_t) = \hat{\pi}(\phi(v_{t+1})|\phi(v_t)), \quad (8)$$

$$D(v_t, v_{t+1}) = \hat{D}(\phi(v_t), \phi(v_{t+1})). \quad (9)$$

In this case, the sampling of v_{t+1} from v_t is a three step process (all steps in Figure 4):

$$v_{t+1} = \phi^{-1}(h_{t+1}), \quad h_{t+1} \sim \hat{\pi}_\theta(\cdot|h_t), \quad h_t = \phi(v_t). \quad (10)$$
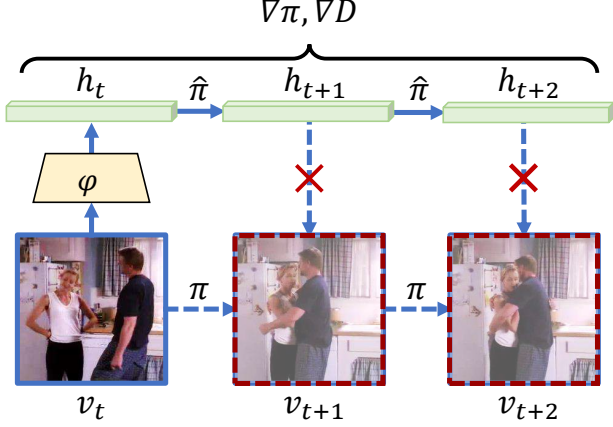
Figure 4. We reparametrize the policy with deep feature representation to avoid explicit state sampling during the gradient computation. The steps of sampling from π is decomposed in to (1) applying $\phi(v_t)$ to get the hidden representation h_t, (2) sampling from reparametrized policy $\hat{\pi}$ to get h_{t+1}, (3) applying $\phi^{-1}(h_{t+1})$ for the next state. However, the third step is not required as we subsequently apply $\phi(\phi^{-1}(h_{t+1}))$ for the sampling in the next time step and cancel out the inverse operation.

The main advantage of this parameterization is that the challenging ϕ^{-1} (image synthesis from hidden representation) is actually not required in the gradient computation, since:

$$\pi(v_{t+1}|v_t) = \hat{\pi}(\phi(v_{t+1})|\phi(v_t)) \qquad (11)$$

$$= \hat{\pi}(\phi(\phi^{-1}(h_{t+1}))|h_t) = \hat{\pi}(h_{t+1}|h_t), \qquad (12)$$

and similar for the discriminator. In this case, only the intermediate representation is required to be sampled for the gradient computation (the top path in Figure 4). In this work, we use the deep convolutional neural networks as our $\phi(\cdot)$ because of its success in representing the visual information [22]. It is important to note that this $\phi(\cdot)$ is also fully differentiable and we approximate it by optimizing it jointly in the policy gradient step of Eq. (6). This reparametrization allows us to directly imitate natural visual sequences from its pixel value and goes beyond the computational limitation in previous works.

4. Experiments

We evaluate our model on three challenging tasks at three different levels of abstraction. The first is future frame generation in the moving MNIST dataset [41], where the image synthesis (ϕ^{-1}) is feasible to verify the effectiveness of our imitation learning formulation. The second task is action anticipation on the TV Human Interactions dataset [35]. We verify the predictive capacity of our model directly imitating natural video [8] from the pixel-level with our deep reparametrization. Finally, we apply our method on the Visual Storytelling dataset [15] to predict the next photo in a storyline that consists of five photos. This is extremely challenging as there can be no clue from direct visual appear-

ance matching. We show that our method can still mimic human behavior at the *semantic*-level from the pixel values.

4.1. Future Frame Generation

In this experiment, we examine our method's ability to perform future frame generation by training on the moving MNIST data sequences [41]. Unlike the two following experiments, frame generation (ϕ^{-1}) is feasible at this scale and we thus also optimize it in our framework and explicitly sample images at the gradient computation.

Dataset. The moving MNIST dataset generates moving digits by randomly selecting 2 digits from MNIST dataset [24] and randomly moving the digits according to the positions in the previous frame. The size of frames are 64×64.

Implementation Details. We note that the state of this dataset is not fully specified by a single image, because the lack of motion information. We thus define the state as three consecutive images. We extend the architecture of DCGAN [37] to have multi-frame input for frame synthesis, which consists of a 5-layer $3D$ convolutional neural network encoder capturing the information of the image, and a 5-layer $2D$ deconvolutional neural network with upsampling as decoder for image generation (See more detail in supplementary). The encoder is treated as our $\phi(\cdot)$ and decoder is treated as our $\phi(\cdot)^{-1}$. We train on sequences of length 10. We use curriculum learning to sequentially increase the prediction horizon. At test time, the goal is to generate four consecutive frames given only the initial frame.

Baselines. We compare to the following models:
- *L1 loss.* We use the same network architecture as ours based on DCGAN [37], but only use the $L1$ reconstruction error of future frame as loss.
- *LSTM.* We compare to the LSTM prediction model of [41].
- *Generative Adversarial Network (GAN).* We perform the ablation study of our model by removing the sequence modeling of imitation learning to demonstrate its importance. In this case, the policy directly aims to minimize the local cost function D from the discriminator without sampling it to be its long term expectation Q. On the other hand, the discriminator only aims to differentiate a single step generated by our policy from the expert's. This reduces our model to be equivalent to GAN [7].

Metric. We follow previous works [43] and evaluate all methods by human evaluation (See more detail for the setting of human evaluation in supplementary). We show on average how many times the output sequence of a method can fool the annotator that it is an actual sequence from the moving MNIST dataset. In this case, the upper bound would be 50% where the generated sequence is the same as real, and the differentiation of real and generated is just by chance.

Results. The result is shown in Table 1. It verifies that

Accuracy (%)	Real sequence
Real sequence	50.0
Our method	15.1
GAN	10.2
LSTM [13]	1.4
L1	0.3

Table 1. Human evaluation results on the moving MNIST dataset. Our imitation learning framework is able to predict and synthesize more realistic consecutive future frames compared to the baselines because our IRL's minimization of long-term expected cost at training.
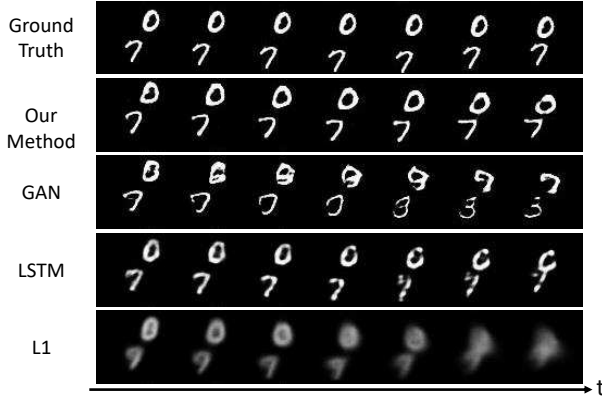

Figure 5. Qualitative results on the moving MNIST dataset. Our model trained with IRL aims to minimize long-term expected cost over the entire sequence. This results in having consistent high-quality frame prediction across time step, and avoids the compounding error problem of baselines.

our method can better generate realistic future frames compared to all baselines. The qualitative results are shown in Figure 5 (See more qualitative results in supplementary). It is important to note the difference of our method compared to GAN. As shown in Figure 5, the result of GAN becomes much worse and drifts away from the original digits as the time step increases. This is due to the fact that it only considers the local cost at training time in contrast to our IRL framework that handles the compounding error problem in long-term prediction. On the other hand, our method is trained with IRL objective and effectively show high-quality synthesis even for long-term prediction.

4.2. Action Prediction on Videos in the Wild

We examined our method's ability to perform action prediction on the videos in the wild following the setup of [42], where the goal is to forecast basic human actions one second before they start. In this experiment, frames synthesis (ϕ^{-1}) is no longer computationally feasible at scale. Therefore we need to use our deep reparametrization of the framework in Section 3.3. We demonstrate the model's ability to capture the dynamics of natural videos directly from raw pixels.

Dataset. We evaluate our results on TV Human Interac-

tions dataset [35], which has 300 videos of four different interactive actions performed by people. We follow previous work [42] and use the THUMOS15 dataset [8] as our unsupervised imitation learning dataset. THUMOS15 consists of trimmed UCF-101 dataset [39] and untrimmed background videos. It has 101 action categories covering a wide range of actions in real world.

Implementation Details. We follow the training procedure and hyperparameters in [42] and reimplemented their methods for a fair comparison. Our reported number is similar to that reported in [42]. We train our model with the same setting. We first do imitation learning on the THUMOS dataset with the states/frames being sampled at the same rate as test time (one second in this case). The predictive methods use the "adapted" setting in [42] during training. As it is infeasible to perform frame synthesis in this task, we use our reparametrization of policy. Our $\phi(\cdot)$ uses the architecture in [10]. This architecture is also used baselines. Our $\hat{\pi}$ for sampling the learned deep representation uses autoencoder with 3-layer encoder and decoder (See more detail in supplementary).

Baselines. We compare to the following methods:

- *SVM*. Two setups on SVM training are considered [42]. The first setup is to train SVM one second before the action starts in the training time. In this case, the distribution of visual features are the same at both training and testing time. The second is to train SVM during the action at the training time. In this case, the input at testing (one second before the action) has different distribution from training, but training frames are given the correct feature for the action. We refer to the first one as SVM and the second and SVM static.

- *LSTM*. We compare to the LSTM prediction model of [41] as sequential training baseline that addresses the compounding error problem.

- *Deep Regression Network* [42]. We compare to the direct deep feature regression approach of Vondrick *et al.* [42]. Their $K = 1$ model can be treated as a proxy of ablation study of our approach by replacing IRL with $L2$ regression loss. We also reimplemented their best performing approach with explicit mixture model ($K = 3$) for predicting multiple possible futures as reference.

-*Pixel Synthesis.* As mentioned in Section 3.3, we resolve the challenge of computational bottleneck from pixel synthesis by introducing a reparametrization that jointly learns a deep representation. It is important to point out that this not only makes our framework efficient, but also makes it effective for semantic prediction. We develop a baseline to demonstrate that our reparametrization is more effective than synthesizing pixels of the future frames. We follow our approach in Section 4.1 and apply the model to downsampled version (64 x 64) of the videos. We anticipate the action with the same setting used by other baselines. The only difference is that now the SVM is applied to deep features
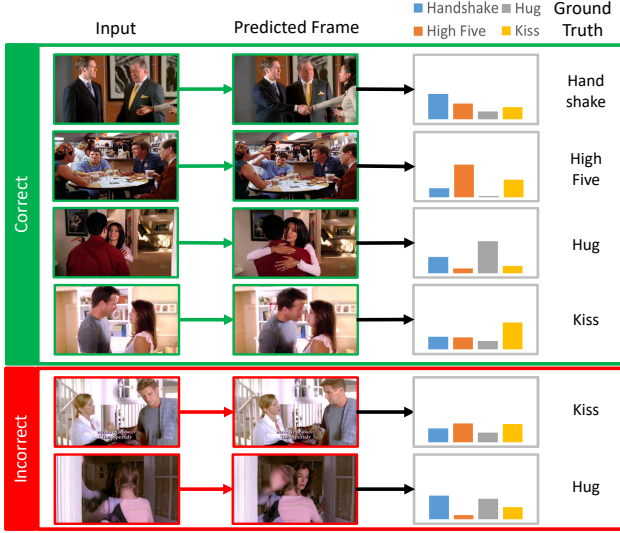
Figure 6. Single step action prediction results. Given an input frame one second before the the interaction takes place, our model is able to robustly predict the future frame and improve significantly over the baselines for action prediction. The frame visualization is retrieved by nearest neighbor image of the predicted deep representation.

Model	Accuracy (%)
Random	25.0
SVM static	36.2
SVM	35.8
Deep Regression ($K = 1$) [42]	37.5
LSTM [41]	40.5
Pixel Synthesis	26.1
Ours	**45.8**
Deep Regression ($K = 3$) [42]	44.0

Table 2. Accuracy of action prediction on the TV Human Interaction dataset. Our learned policy significantly outperforms reconstruction loss based baselines. Our single mode policy also outperforms deep mixture regression model which is designed specifically for this task.

extracted from the synthesized future frames rather than the deep representation directly predicted by our model.

Results. The results for action anticipation are shown in Table 2. Our IRL formulation of action prediction significantly outperforms the corresponding version using just $L2$ loss (Deep Regression $K = 1$). In addition, we are able to outperform the sequential baseline using just the reconstruction loss (LSTM). It is also important to note that our single mode policy without mixture model is already able to outperform the best performing Deep Regression with explicit mixture model for handling multiple possible futures. Note that our reported reimplementation of $K = 1$ is slightly worse than their reporting number because we perform unsupervised learning for all methods only on the THUMOS dataset for a fair comparison. This indicates
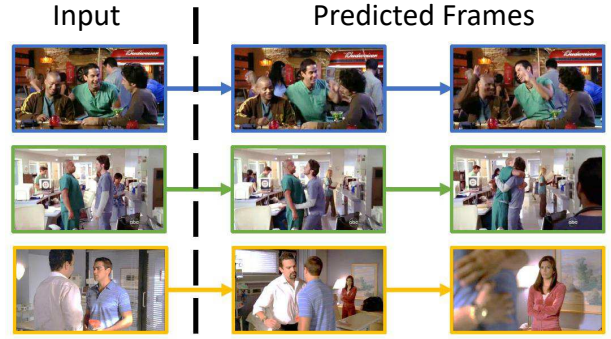

Figure 7. Multi-step video prediction results. Given only a single frame as input, our learned policy is able to accurately forecast multiple future frames in natural video. This verifies the effectiveness of our IRL formulation and reparametrization that allows direct imitation from the pixel level. Images are retrieved by nearest neighbor of predicted deep feature.

that single mode regression is more prone to be affected by dataset bias and demonstrate the robustness of our method. On the other hand, the accuracy of Pixel Synthesis is 26.1%, which is lower than our framework by a large margin. This verifies the importance of bypassing pixel synthesis in our framework for anticipating action in natural videos. Qualitative results of action prediction is shown in Figure 6 (See more qualitative results in supplementary). It can be seen that our learned policy robustly predicts the action after one second. We further demonstrate that our model is able to predict more than one step into future frames. The qualitative results are shown in Figure 7. This shows the effectiveness of our imitation learning framework for visual prediction, and verifies that our model can imitate natural video from the pixel value using our reparametrization of deep representation.

4.3. Storyline Forecasting on Semantic Dynamics

In the final experiment, we evaluate our method on the challenging semantic forecasting in visual storylines [15, 38]. Given a storyline consisting of images, our goal is to predict the next image. This is challenging as the change of semantic meaning might not be capture directly through visual feature matching. Our setup is similar to the prediction task of [38].

Dataset. We use the Visual Storytelling Dataset (VIST) [15] for our storyline forecasting task. The dataset consists of photos from more than 100,000 albums from Flickr. We randomly select one photo sequence of storyline from each album. Using the original dataset split in the dataset, this results in total 8024 training, 1011 testing, and 998 validation storylines, where each consists of five photos illustrating a story. Note that while we follow the evaluation setup of [38], their dataset is not applicable to our task as they only have 10 storylines per concept for evaluation only and does not have the large scale human demonstrated
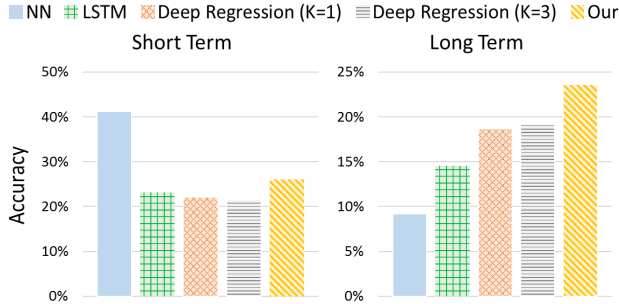
Figure 8. Accuracy of storyline prediction on the Visual Story-telling Dataset. We follow previous work and evaluate short-term and long-term prediction separately. It can be seen that nearest neighbor on the deep feature representation is a strong baseline is this case because the goal is similar to retrieve visually similar frames. On the other hand, nearest neighbor does not work for long-term prediction. Our general visual prediction framework performs the best for the long-term prediction. This verifies that our model can also be applied to the challenging semantic prediction, while learning directly from the pixel.

storyline available for imitation learning.

Experimental Setup and Metrics. We follow [38] and split the storyline evaluation into two goals: the first is short-term, where the image is visually similar consecutive image in the album. The second goal is long-term prediction, where prediction is for the next representative event involving large visual appearance change. We follow [38] and posed this as a classification task, where the goal is to select the true image along with other four other images selected randomly from the same album. We use the training set storylines as expert demonstrations for both our imitation learning and unsupervised learning of baselines. We report resulting accuracy for the test split on both the short-term and long-term prediction (See more detail for experimental setup in supplementary).

Baselines. We again compare to LSTM [41] and Deep Regression Network [42] from the previous task. In addition, we compare to *nearest neighbor* of the ResNet-101 feature representation [10], which is proven to be a strong baseline for short-term prediction [38].

Results. The results of short-term and long-term prediction are shown in Figure 8. For short term prediction, we observe the similar results as with in [38], where nearest neighbor is the strongest baseline as the goal is similar to finding the photo with the most similar visual feature. It is important to note that our method still outperforms reconstruction loss based baselines (LSTM and Deep Regression Network) in this case. For the long-term prediction, it can be seen that the nearest neighbor of visual appearance matching no longer works and performs poorly. On the other hand, our direct imitation of storyline photos from the pixel values is proven effective compared to both LSTM and Deep Regression Networks. This verifies that our ap-

Figure 9. Qualitative results of storyline forecasting. Green boxes indicate correct prediction of the next image in the storyline, and red boxes indicate incorrect predictions. Our imitation learning based framework can best capture the semantics by imitating storylines generated by human. On the other hand, baselines focusing on appearance matching or reconstruction loss is harder to successfully predict semantics in the storylines.

proach is general and can even applied to imitate human behavior at the challenging semantic-level. Qualitative results are shown in Figure 9 (See more qualitative results in supplementary). It can be seen from the figure that our method can better predict semantically meaning next photo in the storyline. First three rows show our correct prediction, and the final row shows incorrect prediction. It can be seen that our method is able to capture the semantic meaning (event or location) of the input photo and continue to forecast photos with similar semantic meaning.

5. Conclusion

We have presented a general framework for long-term visual prediction. We formulate this as an IRL problem, and extend previous works to directly imitate natural visual sequences from the pixel level. The key challenge is the high-dimensional and continuous natural of images. We address this by (1) introducing dual formulation of IRL to natural visual sequences that bypasses exhaustive state-action pair visit, and (2) reparametrization using deep representation to avoid explicit state synthesis during gradient computation. We demonstrated that our framework is general and effective using three tasks at different level of abstractions: (1) future frame generation, (2) action prediction, (3) storyline forecasting. We verified that our unified framework is able to outperform existing methods on all three tasks.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 1, 2

[2] M. Bloem and N. Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *CDC*, 2014. 3

[3] N. J. Butko and J. R. Movellan. Optimal scanning for faster object detection. In *CVPR*, 2009. 2

[4] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun. Anticipating accidents in dashcam videos. In *ACCV*, 2016. 1

[5] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016. 2

[6] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *ICML*, 2016. 1, 2

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 4, 5

[8] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2015. 2, 5, 6

[9] S. Gu, E. Holly, T. Lillicrap, and S. Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *ICRA*, 2017. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[11] J. Ho and S. Ermon. Generative adversarial imitation learning. In *NIPS*, 2016. 1, 2, 3, 4

[12] J. Ho, J. K. Gupta, and S. Ermon. Model-free imitation learning with policy optimization. In *ICML*, 2016. 2, 3

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997. 6

[14] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *ECCV*, 2014. 1, 2, 3

[15] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, J. Devlin, A. Agrawal, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *NAACL*, 2016. 2, 5, 7

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[17] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *ICRA*, 2016. 1

[18] N. Kalchbrenner, A. v. d. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. *arXiv preprint arXiv:1610.00527*, 2016. 1, 4

[19] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 1, 2, 3

[20] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *TPAMI*, 2015. 1

[21] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *IJRR*, 2016. 1

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[23] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. In *ECCV*, 2014. 1, 2

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *IEEE*, 1998. 5

[25] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, 2017. 2

[26] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *CVPR*, 2017. 1, 2

[27] S. Mathe and C. Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *NIPS*, 2013. 1, 2

[28] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016. 2

[29] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016. 2

[30] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 2

[31] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *ICML*, 2000. 1, 2

[32] L. Paletta, G. Fritz, and C. Seifert. Q-learning of sequential attention for visual object recognition from informative local descriptors. In *ICML*, 2005. 2

[33] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi. Egocentric future localization. In *CVPR*, 2016. 2

[34] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *CVPR*, 2017. 2

[35] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in tv shows. In *BMVC*, 2010. 2, 5, 6

[36] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 3

[37] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2, 5

[38] G. A. Sigurdsson, X. Chen, and A. Gupta. Learning visual storylines with skipping recurrent neural networks. In *ECCV*, 2016. 7, 8

[39] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[40] B. Soran, A. Farhadi, and L. Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *ICCV*, 2015. 1

[41] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 2, 5, 6, 7, 8

[42] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016. 1, 2, 6, 7, 8

[43] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, 2016. 2, 5

[44] C. Vondrick and A. Torralba. Generating the future with adversarial transformers. In *CVPR*, 2017. 2

[45] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from variational autoencoders. In *ECCV*, 2016. 2

[46] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2014. 1, 2

[47] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *ICCV*, 2015. 1, 2

[48] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 2

[49] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring "dark matter" and "dark energy" from videos. In *ICCV*, 2013. 1

[50] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016. 1, 2

[51] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016. 2

[52] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: sequence generative adversarial nets with policy gradient. In *AAAI*, 2017. 2

[53] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun. Agent-centric risk assessment: Accident anticipation and risky region localization. In *CVPR*, 2017. 2

[54] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008. 1, 2, 3