
Action-Conditional Video Prediction using Deep Networks in Atari Games

Junhyuk Oh

Computer Science & Engineering
University of Michigan
junhyuk@umich.edu

Xiaoxiao Guo

Computer Science & Engineering
University of Michigan
guoxiao@umich.edu

Honglak Lee

Computer Science & Engineering
University of Michigan
honglak@umich.edu

Richard Lewis

Psychology
University of Michigan
rickl@umich.edu

Satinder Singh

Computer Science & Engineering
University of Michigan
baveja@umich.edu

Abstract

Motivated by vision-based reinforcement learning (RL) problems, in particular Atari games from the recent benchmark Arcade Learning Environment (ALE), we consider spatio-temporal prediction problems where future (image-)frames are dependent on control variables or actions as well as previous frames. While not composed of natural scenes, frames in Atari games are high-dimensional in size, can involve tens of objects with one or more objects being controlled by the actions directly and many other objects being influenced indirectly, can involve entry and departure of objects, and can involve deep partial observability. We propose and evaluate two deep neural network architectures that consist of encoding, action-conditional transformation, and decoding layers based on convolutional neural networks and recurrent neural networks. Experimental results show that the proposed architectures are able to generate visually-realistic frames that are also useful for control over approximately 100-step action-conditional futures in some games. To the best of our knowledge, this paper is the first to make and evaluate long-term predictions on high-dimensional video conditioned by control inputs.

1 Introduction

Over the years, deep learning approaches (see [6, 21] for survey) have shown great success in many visual perception problems (e.g., [14, 7, 26, 8]). However, modeling videos (i.e., building a generative model) is still a very challenging problem because it usually involves high-dimensional natural-scene data with complex temporal dynamics. Thus, recent studies have mostly focused on modeling simple video data, such as bouncing balls or small video patches, where the next frame is highly-predictable based on the previous frames [23, 17, 16]. In many applications, however, future frames are not only dependent on previous frames but also on additional control or action variables. For example, the first-person-view in a vehicle is affected by wheel-steering and acceleration actions. The camera observation of a robot is similarly dependent on its movement and changes of its camera angle. More generally, in vision-based reinforcement learning (RL) problems, learning to predict future images conditioned on future actions amounts to learning a model of the dynamics of the

agent-environment interaction; such transition-models are an essential component of model-based learning approaches to RL. In this paper, we focus on Atari games from the Arcade Learning Environment (ALE) [4] as a source of challenging action-conditional video modeling problems. While not composed of natural scenes, frames in Atari games are high-dimensional in size, can involve tens of objects with one or more objects being controlled by the actions directly and many other objects being influenced indirectly, can involve entry and departure of objects, and can involve deep partial observability. To the best of our knowledge, this paper is the first to make and evaluate long-term predictions on high-dimensional images conditioned by control inputs.

This paper proposes, evaluates, and contrasts two spatio-temporal prediction architectures based on deep networks that incorporate action variables (See Figure 1). The architectures divide the prediction problem into three parts: encoding, action-conditional transformation, and decoding. One architecture is based on convolutional neural networks (CNNs) and the other on recurrent neural networks (RNNs). The encoding part computes high-level abstractions of input frames, the action-conditional transformation part predicts the abstraction of the next frame conditioned on the action, and finally the decoding part maps the predicted high-level abstraction to a detailed frame. The feed-forward architecture takes the last 4 frames as input while the recurrent architecture takes just the last frame but has recurrent connections as shown in Figure 1b. Our experimental results on predicting images in Atari games show that our architectures are able to generate realistic frames over 100-step action-conditional future frames without diverging. We show that the representations learned by our architectures 1) approximately capture natural similarity among actions, and 2) discover which objects are directly controlled by the agent’s actions and which are only indirectly influenced or not controlled at all. We evaluated the usefulness of our architectures for control in two ways: 1) by replacing emulator frames with predicted frames in a previously-learned model-free controller (DQN; DeepMind’s state of the art Deep-Q-Network for Atari Games [18, 19]), and 2) by using the predicted frames to drive a more informed than random exploration strategy to improve a model-free controller (also DQN).

2 Related Work

Uncontrolled Video Prediction using Deep Networks. The problem of video prediction has led to a variety of architectural proposals in the deep learning literature. A recurrent temporal restricted Boltzmann machine (RTRBM) [23] was proposed to learn temporal correlations from sequential data by introducing recurrent connections in RBM. A structured RTRBM (sRTRBM) [17] scaled up RTRBM by learning dependency structures between observations and hidden variables from data. More recently, Michalski et al. [16] proposed a higher-order gated autoencoder (HGAE) that defines multiplicative interactions between consecutive frames and mapping units. By assuming temporal consistencies in high-order mapping units, they suggest that temporal prediction problem can be viewed as learning and inferring transformations between consecutive images. Srivastava et al. [22] applied sequence to sequence learning framework [25] to a video domain, and showed that deep LSTM networks are capable of generating video of bouncing handwritten digits. In contrast to these previous studies that consider only previous frames to predict future frames, this paper tackles problems where control variables affect temporal dynamics, motivated by vision-based RL. In addition, we scale up spatio-temporal prediction to larger-size images than previous work.

ALE: Combining Deep Learning and RL. Atari 2600 games provide very challenging environments for RL because of high-dimensional visual observations, partial observability, and delayed rewards. Approaches that combine deep learning and RL have made significant advances [18, 19, 9]. Specifically, DQN [18] combined Q-learning [29] with a CNN and achieved state-of-the-art performance on many Atari games. Guo et al. [9] used the ALE-emulator as a means for making action-conditional predictions and slow UCT [13], a Monte-Carlo tree search method, to generate training data for a fast-acting CNN, which outperformed DQN on several domains. Throughout this paper we will use DQN to refer to the architecture used in [18] (a more recent work [19] used a deeper CNN with more data to produce the currently best-performing Atari game players). There have been few attempts to learn from ALE data a transition-model that makes action-conditional predictions of future frames. One line of work [2, 3] divides game images into patches and applies a Bayesian model averaging framework to predict patch-based observations. However, this approach assumes that neighboring patches are enough to predict the center patch, which is not true in Atari games because of many complex interactions (e.g., shooting an enemy changes the score pixels).

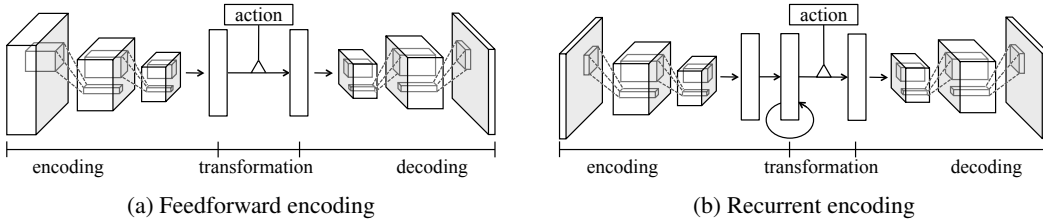


Figure 1: Two Encoding-Transformation-Decoding network architectures.

They present average log loss of 1-step predictions. In this paper we make and evaluate long-term predictions both for quality of pixels generated and for usefulness to control.

3 Proposed Architectures and Training Method

The goal of our architectures is to learn a function $f : \mathbf{x}_{t-k+1:t}, \mathbf{z}_t \rightarrow \mathbf{x}_{t+1}$, where \mathbf{x}_t and \mathbf{z}_t are the frame and action variables at time t , and $\mathbf{x}_{t-k+1:t}$ are the frames from time $t - k + 1$ to time t , i.e., the last k frames. Figure 1 shows our two architectures that are each composed of encoding layers that extract spatio-temporal features from the input frames (Section 3.1), action-conditional transformation layers that transform the encoded features into a prediction of the next frame in high-level feature space by introducing action variables as additional input (Section 3.2) and finally decoding layers that map the predicted high-level features into pixels (Section 3.3). Our architectural contributions are in the novel action-conditional transformation component as well as in the novel use of the overall resulting architecture in vision-based RL domains.

3.1 Feedforward encoding and Recurrent encoding

We present two different types of encoding architecture: *feedforward encoding* and *recurrent encoding* as shown in Figure 1.

Feedforward encoding takes a fixed history of previous frames as an input, which is concatenated through channels (see Figure 1a), and stacked convolution layers extract spatio-temporal features directly from the concatenated frames. The encoded feature vector $\mathbf{h}_t^{enc} \in \mathbb{R}^h$ at time t can be formulated as:

$$\mathbf{h}_t^{enc} = \text{CNN}(\mathbf{x}_{t-k+1:t}), \quad (1)$$

where $\mathbf{x}_{t-k+1:t} \in \mathbb{R}^{k \times n \times m}$ denotes k frames of $n \times m$ pixel images. CNN is a mapping from raw concatenated pixels to a high-level feature vector using multiple convolution layers, each of which is followed by a rectifier nonlinearity [20], and a fully-connected layer at the end. This encoding can be viewed as *early-fusion* [12] (other types of fusions, e.g., *late-fusion* or 3D convolution [28] can also be applied to this architecture).

Recurrent encoding takes one frame as an input for each time-step and extracts high-level spatio-temporal features using an RNN in which the temporal dynamics is modeled by the recurrent layer on top of the high-level feature vector extracted by convolution layers (see Figure 1b). In this paper, long short-term memory (LSTM) [10] without peephole connection is used for the recurrent layer as follows:

$$\mathbf{h}_t^{enc} = \text{LSTM}(\text{CNN}(\mathbf{x}_t)). \quad (2)$$

Intuitively, LSTM units retain information from deep history of inputs using *memory cells*, and $\text{CNN}(\mathbf{x}_t)$ is given as input to the LSTM so that the LSTM captures temporal correlations from high-level spatial features.

3.2 Multiplicative Action-Conditional Transformation

The transformation layer should be able to predict different frames for different agent-actions. A straightforward approach would be to simply concatenate the action into the encoded feature vector and use a fully-connected layer to map to the predicted feature vector. In this approach, however, the encoded feature vector and the action-vector contribute to the prediction independently as an additive interaction, which arguably makes it difficult for the actions to “condition” the transformation. To allow the action to condition the transformation, we propose multiplicative interactions between the

encoded feature vector and the control variables as follows:

$$h_{t,i}^{dec} = \sum_{j,k} W_{ikj} z_{t,j} h_{t,k}^{enc} + b_i \quad (3)$$

where $\mathbf{h}_t^{enc} \in \mathbb{R}^h$ is encoded feature, $\mathbf{h}_t^{dec} \in \mathbb{R}^h$ is transformed feature, $\mathbf{z}_t \in \mathbb{R}^z$ is action-vector at time t , $\mathbf{W} \in \mathbb{R}^{h \times h \times z}$ is 3-way tensor weight, and $\mathbf{b} \in \mathbb{R}^h$ is bias. When the action \mathbf{z} is represented using one-of-k vectors, using a 3-way tensor is equivalent to using different weight matrices for each action. This enables the architecture to model different transformations for different actions. The advantages of multiplicative interactions have been discussed in image and text modelling problems [27, 24, 15]. In practice, however, the 3-way tensor is not scalable because of its large number of parameters. Thus, we approximate the tensor by factorizing into three matrices as follows [27],

$$\mathbf{h}_t^{dec} = \mathbf{W}^{dec} (\mathbf{W}^{enc} \mathbf{h}_t^{enc} \odot \mathbf{W}^z \mathbf{z}_t) + \mathbf{b} \quad (4)$$

where $\mathbf{W}^{dec} \in \mathbb{R}^{h \times f}$, $\mathbf{W}^{enc} \in \mathbb{R}^{f \times h}$, $\mathbf{W}^z \in \mathbb{R}^{f \times z}$, $\mathbf{b} \in \mathbb{R}^h$, and f is the number of factors. Unlike the 3-way tensor, the above factorization shares the weights between different actions by mapping them to the size- f factors. This sharing may be desirable relative to the full 3-way tensor when there are common temporal dynamics in the data across different actions (e.g., ‘left+fire’ and ‘left’ in Atari games); this is discussed further in Section 4.3.

3.3 Convolutional Decoding

It has been recently shown that a CNN is capable of generating an image given fully-specified attributes of the image [1]. Inspired by this, we apply this idea to our end-to-end deep architecture for video prediction. In our method, convolutional filters are used to decode high-level features encoded and transformed by CNN instead of attributes. More specifically, the transformed feature vector \mathbf{h}^{dec} is decoded into pixels by repeatedly applying 2×2 upsampling and a convolution layer.

3.4 Incremental Training

Given the training data $D = \{((\mathbf{x}_1^n, \mathbf{z}_1^n), \dots, (\mathbf{x}_{T_n}^n, \mathbf{z}_{T_n}^n))\}_{n=1}^N$, the model is trained to minimize the sum of the squared loss of K -step predictions as follows:

$$\mathcal{L}_K(\theta) = \frac{1}{2} \sum_n \sum_t \sum_{k=1}^K \|\hat{\mathbf{x}}_{t+k}^n - \mathbf{x}_{t+k}^n\|^2 \quad (5)$$

where $\hat{\mathbf{x}}_{t+k}^n$ is a k -step future prediction. Intuitively, the network is repeatedly *unrolled* through K time steps by using its prediction as an input for the next time-step. The model is trained in multiple phases based on increasing K , as suggested by Michalski et al. [16]. In other words, the model is trained to predict short-term future frames and fine-tuned to predict longer-term future frames after the previous phase converges. A mini-batch gradient descent with backpropagation through time (BPTT) is used to optimize the parameters of the network by sampling training sequences from the training data.

4 Experiments

In the experiments that follow, we have the following goals for our two architectures. 1) To evaluate the predicted frames in two ways: qualitatively evaluating the generated video, and quantitatively evaluating the pixel-based squared loss, 2) To evaluate the usefulness of predicted frames for control in two ways: by replacing the emulator’s frames with predicted frames for use by DQN, and by using the predictions to improve exploration in DQN, and 3) To qualitatively evaluate the representations learned by our architectures.

We begin by describing the details of the data, and model architecture parameters, and baselines.

Data and Preprocessing. We replicated DQN and used our replication to generate game-play video datasets using ϵ -greedy policy with $\epsilon = 0.2$, which means DQN is forced to choose a random action with 20% probability. For each game, the dataset consists of about 600,000 training frames and 60,000 test frames with actions chosen by DQN. Following DQN, actions are chosen once every 4 frames which reduces the 60fps video to 15fps video. The number of actions available in games

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
271						↑
272						↑
273						↑

Figure 2: Example of predictions over 270 steps in Freeway. ‘Step’ column shows the number of prediction steps, while the ‘Action’ column shows the actions taken. The white boxes indicate the object controlled by the agent. From prediction step 272 to 273 the controlled object crosses the top boundary and reappears at the bottom; this nonlinear shift is predicted by our architectures and is not predicted (indeed, not even represented) by naLinear and naFf. The horizontal movement of the uncontrolled objects (most clearly seen near the left boundary of the frame) are predicted by our architectures and naFf but not by naLinear.

varies from 3 to 18, and they are represented as 1-of-k vectors. As in DQN, every image (210×160) is down-sampled to 84×84 pixels and gray-scaled. We preprocessed the images by subtracting mean pixel values and dividing each pixel value by 128.

Network Architecture and Training. The same network architecture is used for all game domains. The encoding layers consist of 3 convolution layers and one fully-connected layer. The first convolution layer has $64, 6 \times 6$ filters with stride of 2. The second and third convolution layers have $64, 6 \times 6$ filters with stride of 2 and padding of 2. The fully-connected layer consists of 1024 hidden units. Every layer is followed by a rectified linear function [20]. For the feedforward encoding network, last 4 frames are used as an input for each time-step. In the recurrent encoding network, a LSTM layer with 1024 hidden units is added on top of the fully connected layer. The recurrent encoding network takes one frame for each time-step, but it is unrolled through last 11 frames to initialize the LSTM hidden units before making a prediction. The number of factors in the transformation layer is 2048. The architecture of decoding layers is symmetric with respect to the encoding layers. We trained the model by increasing the number of prediction steps: 1, 3, and 5. Our implementation is based on Caffe toolbox [11].

Baselines. In the following experiments, the proposed models are compared with two baselines that do not incorporate actions. The first, patch-wise no-action linear regression (or *naLinear*) takes randomly sampled 16×16 patches concatenated through last four frames ($4 \times 16 \times 16$), and predicts a 16×16 patch of the next frame. During testing, the linear regression model predicts every patch by sliding windows with overlap, and the overlapped pixels are averaged. A second baseline, no-action feedforward (or *naFf*) is the same as the feedforward encoding architecture (Figure 1a) except that the transformation layer consists of one fully-connected layer that does not get the action as input.

4.1 Evaluation of Predicted Frames

Qualitative Evaluation: Prediction video. The prediction videos of our models and baselines are available at the following website: <https://sites.google.com/a/umich.edu/junhyuk-oh/action-conditional-video-prediction>.¹ As seen in the videos, our proposed models make qualitatively reasonable predictions over 30 ~ 500 steps depending on the game. In all games, the naLinear baseline quickly diverges, and the naFf baseline fails to predict the

¹In addition to the main results on down-sampled images, we have also trained larger networks on the original images (210×160 RGB pixels). These preliminary results are also demonstrated in the website.

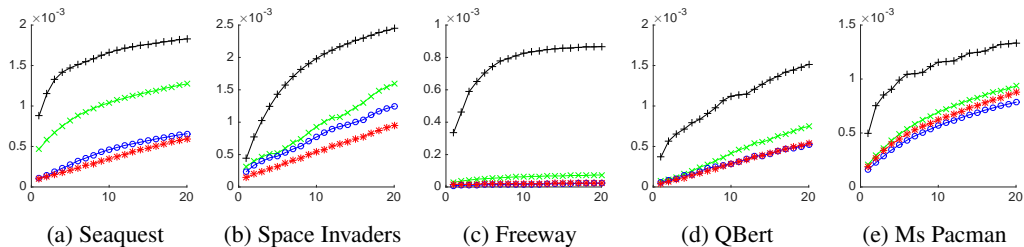


Figure 3: Mean squared error over 20-step predictions. The blue curves ('o') and red curves ('*') correspond to feedforward encoding and recurrent encoding respectively. The black ('+') and green ('x') curves represent naLinear and naFf respectively. The x-axis is the number of steps of prediction. The y-axis is the per-pixel mean squared error.

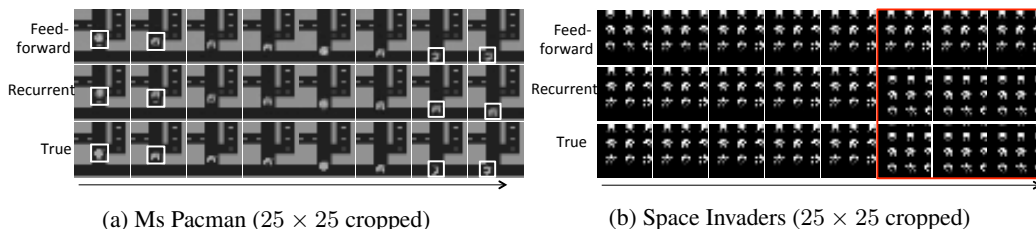


Figure 4: Comparison between two encoding models. (a) Controlled object is seen in white square. As the recurrent encoding model makes a small translation error at the beginning, this leads to a situation in which the true position of the controlled object is in the bottom corridor while the predicted position is still above the bottom corridor. The agent moves leftwards from then on which is not possible in the predicted position and so the controlled object in recurrent encoding gets stuck. This is less likely to happen in feedforward encoding because its position prediction is more accurate. (b) The objects move down after staying at the same location for the first five steps. The feedforward encoding model fails to predict this movement (see last three frames) because it only gets the last four frames as input while the recurrent encoding model predicts the downwards movement correctly.

object controlled by the actions. An example of predictions over 270 steps in Freeway is illustrated in Figure 2. We observed that both models predict complex local translations very well such as the movement of vehicles and the object controlled by the agent. They can predict interactions between objects such as collision of two objects. Since our architectures effectively extract hierarchical features using CNN, they are able to make a prediction that requires a global context. For example, in Figure 2, the model predicts the sudden change of the controlled-object location (from the top lane to the bottom lane) at 273-step.

However, both of our models have difficulty in accurately predicting small objects, such as bullets in Space Invaders. This is because squared loss gives only small error signals when the model fails to predict small objects during training. Another difficulty is in handling stochasticity. In Seaquest, new objects appear from the left side or right side randomly, and these are hard to predict. Although our models do generate new objects with reasonable shapes and movements (e.g., after appearing they move as in the true frames), the generated frames do not necessarily match the ground-truth.

Quantitative Evaluation: Squared Prediction Error. Mean squared error over 20-step predictions is reported in Figure 3 (see caption for details). Our predictive models outperform the two baselines for all domains. However, the gap between our predictive models and no-action-feedforward baseline is not very large except for Seaquest. This is due to the fact that the object directly controlled by the action occupies only a small part of the image.

Qualitative Analysis of Relative Strengths and Weaknesses of Feedforward encoding and Recurrent encoding. We hypothesize that feedforward encoding can model precise spatial transformations more easily because its convolutional filters can learn temporal correlations directly from pixels in the multiple concatenated frames. In contrast, convolutional filters in recurrent encoding can learn only spatial features from the one-frame input, and the temporal context has to be captured by the recurrent layer on top of the high-level CNN feature vector, which does not have localized

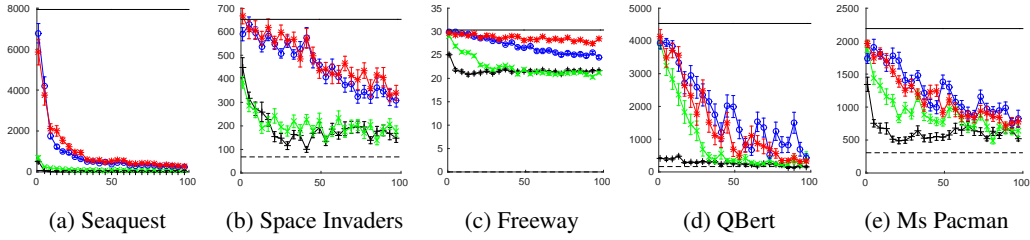


Figure 5: Game play performance using the predictive model as an emulator. The models corresponding to each curve are the same as in Figure 3. The solid horizontal lines are the performance of the DQN controller when given the true frames, and the dashed horizontal lines correspond to scores achieved by purely random play. The x-axis is the number of steps of prediction before re-initialization with real frames. The y-axis is the average game score measured from 30 plays.

information. On the other hand, recurrent encoding is more flexible and potentially better for modelling long-term dependencies, because the recurrent layer can account for an arbitrary number of frames, whereas feedforward encoding is not suitable for modelling long-term dependencies because it takes more memory and parameters as more frames are concatenated.

As evidence, in Figure 4a we show a case that feedforward encoding is better at predicting the precise movement of the action-controlled object, while recurrent encoding makes a 1-2 pixel translation error. This small spatial error leads to entirely different predicted frames after a few steps. Since the architecture of feedforward encoding and recurrent encoding is identical except for the encoding part, we conjecture that this result is mainly due to the failure of precise spatio-temporal encoding in recurrent encoding. On the other hand, recurrent encoding is better at predicting when the enemies move in Space Invaders as illustrated in Figure 4b. This is due to the fact that the enemies move after 9 steps, which is hard for feedforward encoding to predict because it takes only the last four frames as input. We observed similar results showing that feedforward encoding cannot handle long-term dependencies in other games. In Freeway, for example, the controlled-object cannot move for 9 steps when it starts a new stage. Feedforward encoding sometimes moves the controlled-object in this situation.

4.2 Evaluating Usefulness of Predictions for Control

Replacing Real Frames with Predicted Frames as Input to DQN. Since squared loss does not measure how meaningful the predictions are for playing the games, we implement an alternative evaluation method that uses the predictive model to replace the game emulator as follows. A DQN controller that takes the last four frames is first pre-trained using real frames and then used to play the games based on $\epsilon = 0.05$ —greedy policy where the input frames are generated by our predictive model instead of the game emulator. To evaluate how the depth of predictions influence the quality of control, we re-initialize the predictions using the true last frames after every n -steps of prediction for $1 \leq n \leq 100$. Note that the DQN controller never sees a true frame, just the outputs of our predictive models.

The results are shown in Figure 5. Unsurprisingly, replacing real frames with predicted frames reduces the score. However, in all the games using the model to repeatedly predict only a few time steps yields a score very close to that of using real frames. Our two architectures produce much better scores than the two baselines for deep predictions than would be suggested based on the much smaller differences in squared loss. The likely cause of this is that our models are better able to predict the movement of the object directly controlled by the actions relative to the baselines even though such an ability may not always lead to better squared loss error. In three out of the five games the score remains much better than the score of random play even when using a 100 steps of prediction (repeatedly) before re-initialization. In Freeway and Space Invaders the predictions seem particularly good at replacing the emulator. In Freeway, where recurrent encoding clearly outperforms feedforward encoding, we observed that the feedforward encoding network sometimes fails to keep track of the directly-controlled-object when the agent starts a new stage.

Model	Seaquest	S. Invaders	Freeway	QBert	Ms Pacman
DQN - Random exploration	13119 (538)	698 (20)	30.9 (0.2)	3876 (106)	2281 (53)
DQN - Informed exploration	13265 (577)	681 (23)	32.2 (0.2)	8238 (498)	2522 (57)

Table 1: Average game score of DQN over 100 plays. The first row and the second row show the performance of our DQN replication with random exploration and informed exploration respectively.

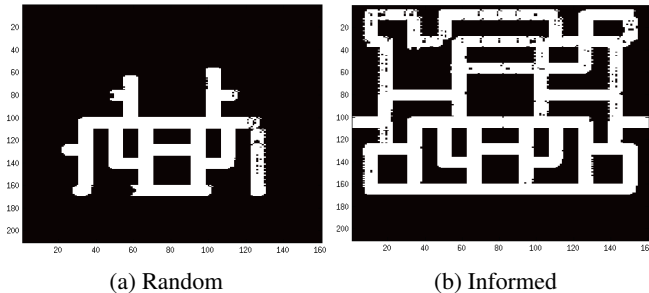


Figure 6: Comparison between two exploration methods on Ms Pacman. Each heat map shows the trajectories of the controlled object measured over 2500 steps for the corresponding exploration strategy.

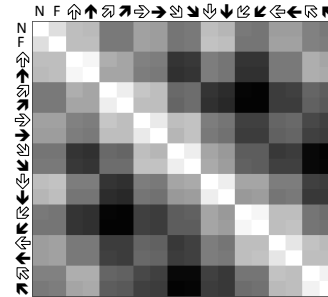


Figure 7: Cosine similarity between every pair of action factor representations.

Improving DQN via Informed Exploration. To learn control in an RL domain, exploration of actions and states is necessary because without it the agent can get stuck in a badly sub-optimal policy. In the DQN paper, the CNN-based agent was trained using an ϵ -greedy policy in which the agent picks the greedy action $1 - \epsilon$ percent of the time and a random action ϵ percent of the time. Such random exploration is a basic strategy that produces sufficient exploration, but can be slower than more informed exploration strategies. Our *informed exploration* strategy is to again take the greedy action $1 - \epsilon$ percent of the time but for the ϵ -percent exploratory actions to pick one that takes the agent to a frame that has been visited least often (say in the last d time steps). Implementing such a exploration strategy requires a predictive model because the next frame for each possible action has to be considered.

Specifically, we store the most recent d frames in a *trajectory memory*, denoted $D = \{\mathbf{x}^i\}_{i=1}^d$. The predictive model is used to get the next frame \mathbf{x}^a for every action a . We estimate the visit-frequency for every predicted frame by summing the similarity between the predicted frame and the most d recent frames stored in the trajectory memory using an exponential kernel over the pixels as follows:

$$n_D(\mathbf{x}^a) = \sum_{i=1}^d k(\mathbf{x}^a, \mathbf{x}^i) \quad (6)$$

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\sum_j \min(\max((x_j - y_j)^2 - \delta, 0), 1)/\sigma\right) \quad (7)$$

where $\mathbf{x}^i \in D$ is the i -th frame in the trajectory memory, δ is a threshold, and σ is a kernel bandwidth.²

Table 1 summarizes our results. The informed exploration improves DQN’s performance using our predictive model in three of five games (with the most significant improvement in QBert); see Figure 6 for a heatmap that shows the informed exploration strategy improving the initial experience of DQN. This preliminary result shows one way our predictive models can be used to improve DQN’s game play; recall that DQN is the state of the art architecture for playing Atari games.

4.3 Analysis of Learned Representations

Similarity among Action Representations. In the factored multiplicative interactions, every action is linearly transformed to f factors ($\mathbf{W}^z \mathbf{z}$ in Equation 4). In Figure 7 we present the cosine

²The size of trajectory memory is 200 for QBert and 20 for the other games, δ is 0 for Freeway and 50 for the others, σ is 100 for all games. We use our feedforward encoding architecture to predict \mathbf{x}^a .

similarity between every pair of action-factor representations after training in Seaquest. ‘N’ and ‘F’ corresponds to ‘no-operation’ and ‘fire’. Black arrows and white arrows correspond to movements with or without ‘fire’. It turns out that there are strong positive correlation between actions that have the same movement directions such as ‘up’ and ‘up+fire’. There are also negative correlations between actions that have opposite moving directions such as ‘up+right’ and ‘down+left’. Both of these effects are reasonable and discovered automatically in learning good predictions.

Distinguishing Controlled and Uncontrolled Objects. This is a hard and interesting problem in it’s own right. Bellemare et al. [5] proposed a framework to learn *contingent regions*, the parts of an image affected by the agent’s action, and suggested that contingency awareness is useful for model-free RL agents. As we illustrate below, our architectures implicitly learn contingent regions as they learn to predict the entire image.

In our architectures, a factor dimension ($f_i = (w_i^z)^T \mathbf{z}$) with higher variance measured over all possible actions, $\text{Var}(f_i) = \mathbb{E}_{\mathbf{z}} \left[(f_i - \mathbb{E}_{\mathbf{z}}[f_i])^2 \right]$, is more likely to transform an image differently depending on actions. Thus, we assume that such factor dimensions are responsible for transforming the part of the image related to the actions. Based on this assumption, we collected the high variance (>0.001) dimensions from the model trained on Space Invaders into a “highvar” subset (around 40% of factors), and collected the remaining dimensions into a “lowvar” subset. Given an image and an action, we did two controlled forward propagations: giving only highvar factor dimensions (by setting the other factors to zeros) and vice versa. The results are visualized as ‘Action’ and ‘Non-Action’ image in Figure 8. Interestingly, given only highvar-dimensions (Action case), the model predicts sharply the movement of the object controlled by actions, while the other objects are much more blurry. In contrast, given only lowvar-dimensions (Non-Action case), the model predicts the movement of the enemies and the background more sharply, and the controlled object stays at the previous location. This result implies that our model learns to distinguish between controlled objects and uncontrolled objects and transform them using disentangled representations.

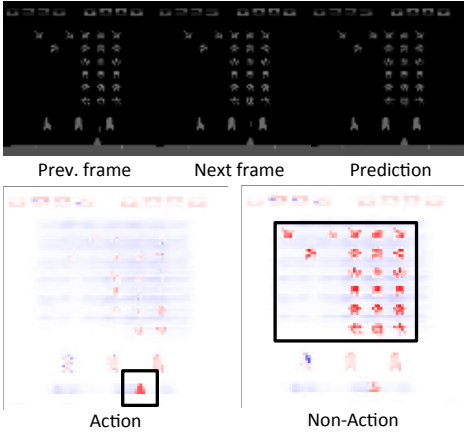


Figure 8: Distinguishing controlled and uncontrolled objects. The red, blue, and white colors represent positive, negative, and zero value from the network output. See text for details.

5 Conclusion

This paper introduced two different novel deep architectures that predict future frames that are dependent on actions and showed qualitatively and quantitatively that they are able to predict visually-realistic and useful-for-control frames over 100-step futures on several Atari game domains. To our knowledge, this is the first paper to show good deep predictions in Atari games. Since our architectures were domain independent we expect that they will generalize to many vision-based RL problems. In future work we will learn models that predict future reward in addition to predicting future frames and evaluate the performance of our architectures in model-based RL.

References

- [1] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015.
- [2] M. Bellemare, J. Veness, and M. Bowling. Bayesian learning of recursively factored environments. In *ICML*, 2013.
- [3] M. Bellemare, J. Veness, and E. Talvitie. Skip context tree switching. In *ICML*, 2014.
- [4] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *JAIR*, 47:253–279, 06 2013.

- [5] M. G Bellemare, J. Veness, and M. Bowling. Investigating contingency awareness using atari 2600 games. In *AAAI*, 2012.
- [6] Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009.
- [7] Dan Ciresan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR*, 2012.
- [8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. IEEE, 2014.
- [9] X. Guo, S. Singh, H. Lee, R.L Lewis, and X. Wang. Deep learning for real-time atari game play using offline monte-carlo tree search planning. In *NIPS*, 2014.
- [10] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural Computation.*, 9(8):1735–1780, 1997.
- [11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.
- [13] L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *ECML*. Springer, 2006.
- [14] A. Krizhevsky, I Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *NIPS*, pages 1097–1105. Curran Associates, Inc., 2012.
- [15] R. Memisevic. Learning to relate images. *PAMI*, 35(8):1829–1846, 2013.
- [16] V. Michalski, R. Memisevic, and K. Konda. Modeling deep temporal dependencies with recurrent “grammar cells”. In *NIPS*. 2014.
- [17] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee. Structured Recurrent Temporal Restricted Boltzmann Machines. In *ICML*, 2014.
- [18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.
- [19] V. Mnih, K. Kavukcuoglu, D. Silver, A. a. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.
- [20] V. Nair and G. E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.
- [21] J. Schmidhuber. Deep learning in neural networks: An overview. *arXiv preprint arXiv:1404.7828*, 2014.
- [22] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *arXiv preprint arXiv:1502.04681*, 2015.
- [23] I. Sutskever, G.E. Hinton, and G. Taylor. The Recurrent Temporal Restricted Boltzmann Machine. *NIPS*, 21(1), 2008.
- [24] I. Sutskever, J. Martens, and G.E Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.
- [25] I. Sutskever, O Vinyals, and QVV Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.
- [26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.
- [27] G. W Taylor and G. E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *ICML*, 2009.
- [28] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.
- [29] C. JCH Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

A Informed Exploration

The entire DQN algorithm combined with informed exploration is described in Algorithm 1. The bold texts show the modifications on the original DQN algorithm.

Algorithm 1 Deep Q-learning with informed exploration

```
Allocate capacity of replay memory  $R$ 
Allocate capacity of trajectory memory  $D$ 
Initialize parameters  $\theta$  of DQN
while  $steps < M$  do
  Reset game and observe image  $x_1$ 
  for  $t=1$  to  $T$  do
    Sample  $c$  from Bernoulli distribution with probability  $\epsilon$ 
    Set  $a_t = \begin{cases} \mathbf{argmin}_a n_D(x_t^a) \text{ in Eq 6} & \text{if } c = 1 \\ \mathbf{argmax}_a Q(\phi(s_t), a; \theta) & \text{otherwise} \end{cases}$ 
    Choose action  $a_t$ , observe reward  $r_t$  and image  $x_{t+1}$ 
    Set  $s_{t+1} = x_{t-2:t+1}$  and preprocess images  $\phi_{t+1} = \phi(s_{t+1})$ 
    Store image  $x_{t+1}$  in  $D$ 
    Store transition  $(\phi_t, a_t, r_t, \phi_{t+1})$  in  $R$ 
    Sample a mini-batch of transitions  $(\phi_j, a_j, r_j, \phi_{j+1})$  from  $R$ 
    Update  $\theta$  based on the mini-batch and Bellman equation.
     $steps = steps + 1$ 
  end for
end while
```

Comparison to random exploration. Figure 9 visualizes the difference between random exploration and informed exploration in two games. In Freeway, where the agent gets rewards by reaching the top lane, the agent move only around the bottom area in the random exploration so that it takes about 4.6×10^5 steps to get the first reward. On the other hand, the agent moves around all locations in the informed exploration and receives the first reward in 86 steps. The similar result is found in Ms Pacman.

Application to Deep Q-learning. The results of the informed exploration using the game emulator and our predictive model are reported in Figure 10 and Table 2 with the baselines. Two published results of DQN [18, 19] and our DQN replication. Our replication follows [18], which uses a smaller CNN than [19].

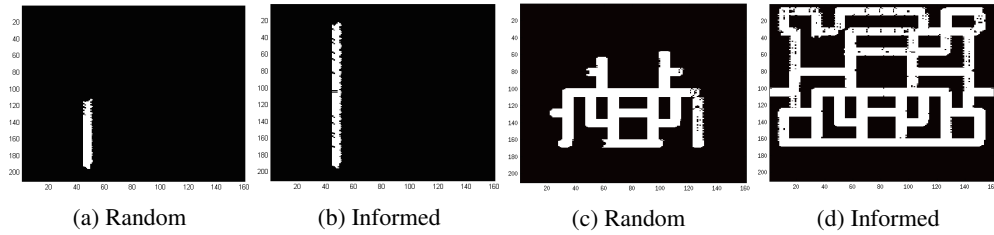


Figure 9: Comparison between two exploration methods on Freeway (Left) and Ms Pacman (Right). Each heat map shows the trajectories of the agent measured from 2500 steps from each exploration strategy.

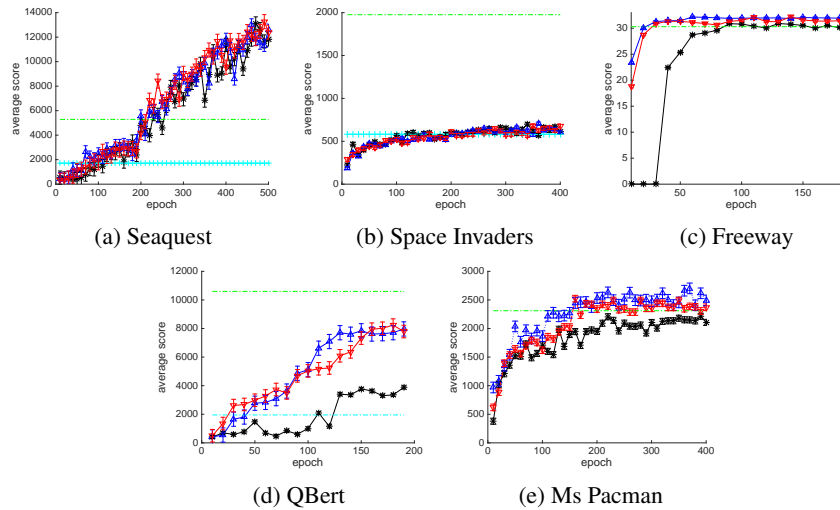


Figure 10: Learning curves of DQNs. The red and blue curves are informed exploration using our predictive model and the emulator respectively. The black curves use random exploration (original DQN). The average game score measured from 100 game plays with 0.05-greedy policy.

Model	Seaquest	S. Invaders	Freeway	QBert	Ms Pacman
DQN (Nature) [19]	5286	1976	30.3	10596	2311
DQN (NIPS) [18]	1705	581	-	1952	-
Our replication of [18]	13119 (538)	698 (20)	30.9 (0.2)	3876 (106)	2281 (53)
I.E (Prediction)	13265 (577)	681 (23)	32.2 (0.2)	8238 (498)	2522 (57)
I.E (Emulator)	13002 (498)	708 (17)	32.2 (0.2)	7969 (496)	2702 (92)

Table 2: Game score. The average game score is measured from 100 plays for every epoch (50,000 weight updates), and the maximum average score is reported. I.E indicates DQN combined with the informed exploration method. ‘Emulator’ and ‘Prediction’ correspond to the emulator and our predictive model.

B Squared Loss

Data	Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Seaquest	Linear	8.81	11.5	13.2	14.2	14.7	15.1	15.4	15.8	16.2	16.5	16.8	17.1	17.2	17.4	17.6	17.7	17.9	18.0	18.1	18.2
	naFf	4.67	5.81	6.74	7.53	8.21	8.84	9.29	9.72	10.0	10.4	10.7	10.9	11.2	11.4	11.7	11.9	12.1	12.3	12.5	12.7
	Feedforward	1.10	1.41	1.89	2.32	2.74	3.23	3.57	3.93	4.30	4.62	4.87	5.13	5.32	5.52	5.72	5.92	6.09	6.26	6.43	6.57
	Recurrent	1.04	1.27	1.55	1.75	2.09	2.35	2.66	2.89	3.20	3.45	3.76	4.02	4.30	4.56	4.81	5.04	5.30	5.52	5.75	5.95
S.Invaders	Linear	4.45	7.68	10.2	12.4	14.2	15.7	17.0	18.1	19.0	19.8	20.5	21.1	21.6	22.1	22.5	22.9	23.4	23.8	24.1	24.4
	naFf	3.17	3.99	4.64	5.13	5.18	5.98	6.83	7.39	8.39	9.29	10.0	10.6	10.7	11.5	12.2	12.9	13.9	14.8	15.4	15.9
	Feedforward	2.38	3.37	3.96	4.50	4.76	5.30	5.86	6.29	7.05	7.70	8.36	8.98	9.32	9.63	10.0	10.3	11.0	11.6	12.1	12.4
	Recurrent	1.47	1.98	2.38	2.83	3.17	3.65	4.08	4.47	4.96	5.41	5.78	6.26	6.66	7.00	7.36	7.73	8.27	8.75	9.15	9.49
Freeway	Linear	3.37	4.62	5.89	6.51	7.05	7.45	7.76	7.93	8.10	8.24	8.33	8.41	8.47	8.53	8.56	8.58	8.61	8.64	8.64	8.67
	naFf	0.28	0.36	0.42	0.48	0.51	0.53	0.56	0.59	0.62	0.62	0.65	0.65	0.68	0.68	0.68	0.70	0.70	0.70	0.73	0.73
	Feedforward	0.08	0.11	0.11	0.11	0.14	0.14	0.14	0.14	0.17	0.17	0.17	0.17	0.17	0.19	0.19	0.19	0.19	0.22	0.22	0.22
	Recurrent	0.17	0.17	0.17	0.17	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.22	0.22	0.22	0.22	0.22	0.22	0.22
QBert	Linear	3.71	5.69	6.57	7.17	7.90	8.36	9.04	9.77	10.6	11.2	11.3	11.4	12.1	12.6	13.2	13.6	14.2	14.4	14.7	15.1
	naFf	0.76	0.96	1.21	1.58	1.98	2.43	2.77	3.23	3.65	4.13	4.56	4.90	5.41	5.52	5.86	6.29	6.54	6.88	7.25	7.48
	Feedforward	0.59	0.79	1.04	1.41	1.55	1.89	2.32	2.49	2.63	2.89	3.11	3.42	3.79	3.85	4.10	4.42	4.70	4.81	4.98	5.18
	Recurrent	0.42	0.65	0.87	1.13	1.44	1.70	1.98	2.26	2.55	2.83	3.08	3.34	3.68	3.88	4.16	4.42	4.67	4.93	5.13	5.41
Ms.Pacman	Linear	4.96	7.51	8.50	9.04	9.89	10.4	10.4	10.6	11.1	11.5	11.6	11.6	12.0	12.3	12.4	12.5	12.9	13.2	13.2	13.3
	naFf	2.09	2.91	3.59	4.30	4.87	5.38	5.83	6.23	6.57	6.91	7.22	7.51	7.79	8.04	8.27	8.53	8.75	8.92	9.12	9.35
	Feedforward	1.61	2.26	2.86	3.42	3.91	4.33	4.73	5.04	5.35	5.66	5.95	6.17	6.43	6.68	6.88	7.11	7.31	7.51	7.68	7.85
	Recurrent	1.89	2.69	3.37	3.96	4.45	4.90	5.27	5.61	5.89	6.20	6.49	6.74	7.02	7.31	7.53	7.82	8.04	8.27	8.50	8.75

all values are multiplied by 10^4

Table 3: Mean squared error over 20-step predictions. The numbers represent $\frac{1}{n} \|\hat{\mathbf{x}}_{t+k} - \mathbf{x}_{t+k}\|^2$ measured from randomly sampled 30,000 sequences from the test data. Pixel values are scaled to $[0, 1]$.

C Correlation between actions

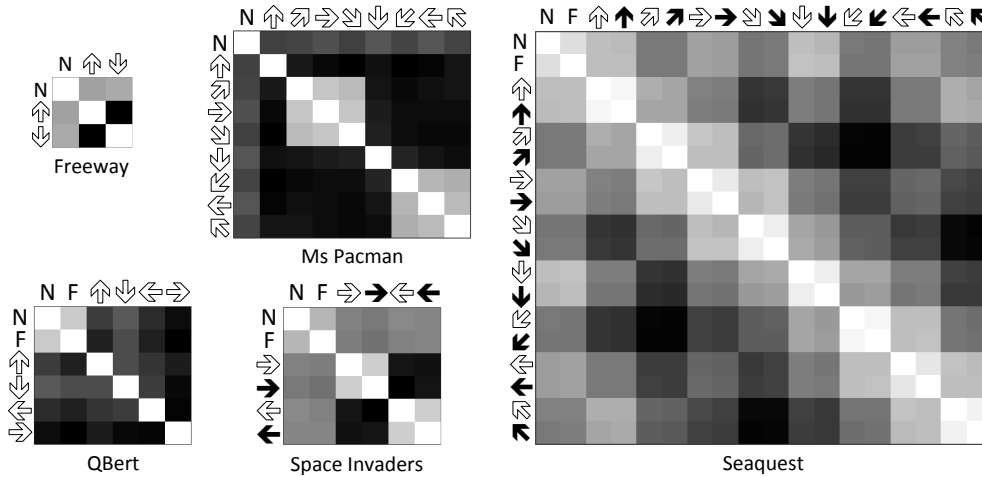
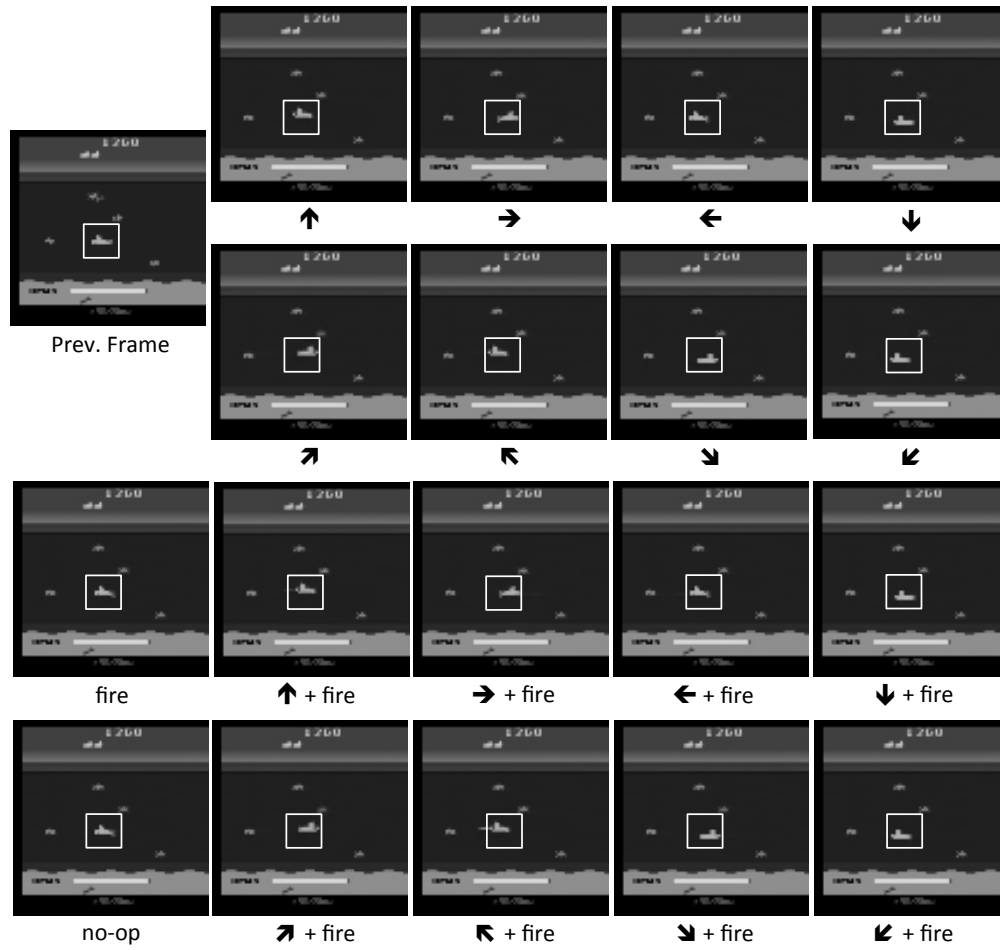
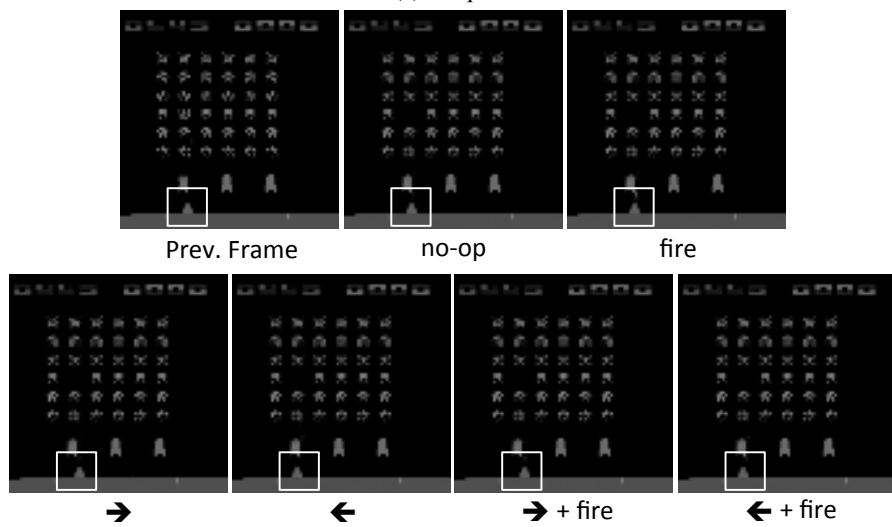


Figure 11: Correlations between actions. The brightness represents cosine similarity between every pair of factors.

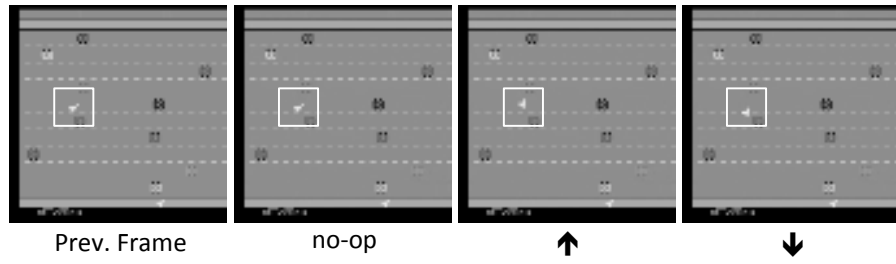
D Handling different actions



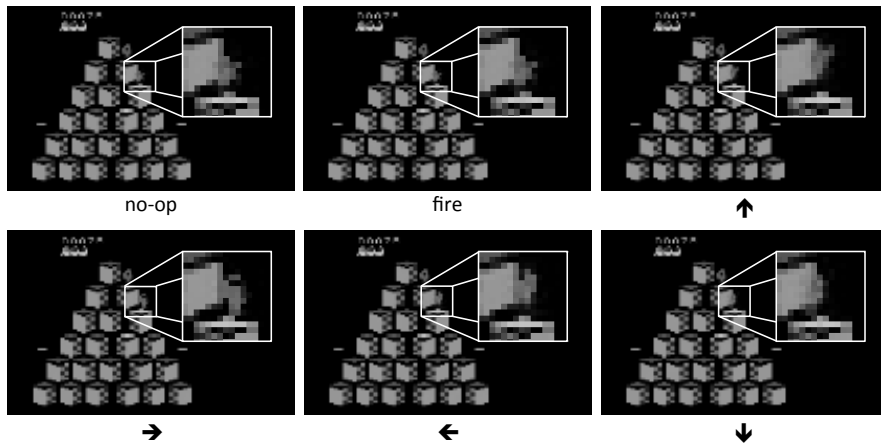
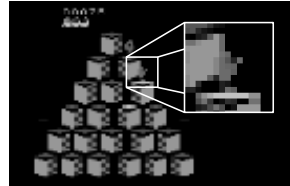
(a) Seaquest



(b) Space Invaders

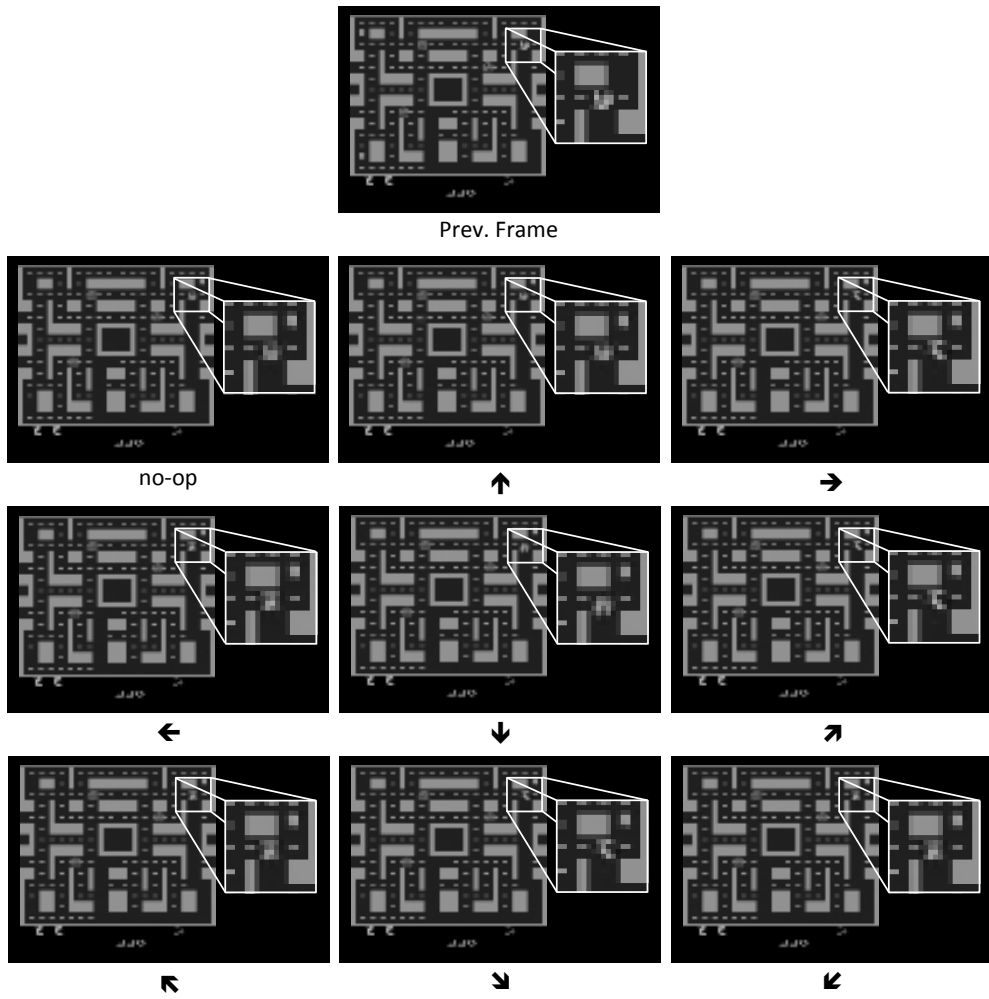


(a) Freeway



(b) QBert

Figure 11: Predictions given different actions



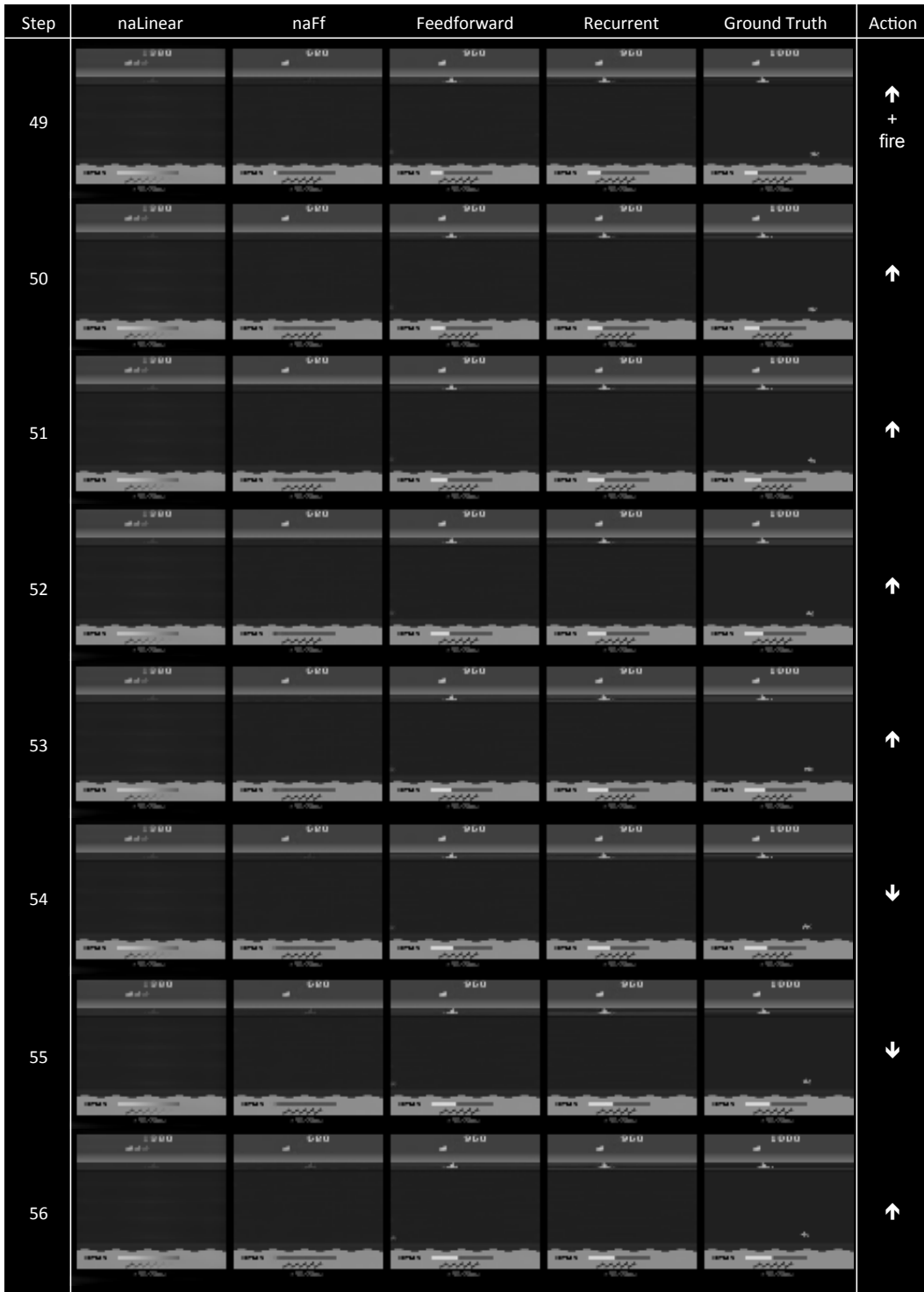
(c) Ms Pacman

Figure 11: Predictions given different actions

E Prediction video on down-sampled images

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
1						←
2						↗
3						↑
4						→
5						→
6						→
7						↑
8						↑

(a) Seaquest (1 ~ 8 steps). The proposed models (feedforward and recurrent) predict the movement of the submarine as well as the enemies correctly. It also predicts the blinking oxygen level which is running out.



(a) Seaquest (49 ~ 56 steps). The submarine is filling the oxygen tank at the surface of the sea (the oxygen level is increasing). It stays at the same location regardless of the actions until the oxygen tank is filled up.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
141						↘ + fire
142						↖
143						↖
144						↓
145						↗ + fire
146						↖ + fire
147						↖ + fire
148						↖ + fire

(a) Seaquest (141 ~ 148 steps). The recurrent model predicts new enemies coming from the right side. Although they do not match the ground-truth images, the generated objects are realistic in terms of shapes and movements.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
1						fire
2						fire
3						fire
4						fire
5						→
6						fire
7						fire
8						fire

(a) Space Invaders (1 ~ 8 steps). The controlled object is located in the bottom, while the enemies are arranged in a grid. Since there are not many variations in the early part of the game, the models predict the dynamics of the game well.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
169						→ + fire
170						→ + fire
171						→ + fire
172						→ + fire
173						no-op
174						→ + fire
175						→ + fire
176						→ + fire

(a) Space Invaders (169 ~ 176 steps). The recurrent model predicts enemies moving down in 171-step, while the other models fail to predict it.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
193						← + fire
194						← + fire
195						fire
196						fire
197						←
198						← + fire
199						→ + fire
200						→ + fire

(a) Space Invaders (193 ~ 200 steps). In 200-step, our models keep track of the object. The recurrent model predicts the positions of the enemies more accurately than other models.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
1						↑
2						↑
3						no-op
4						↑
5						no-op
6						↓
7						↑
8						↓

(a) Freeway (1 ~ 8 steps). The predictions made by our models are almost same as ground-truth images. The small object (chicken), which is controlled by the agent, is diffused in the predictions of naFf.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
132						↑
133						no-op
134						↑
135						↓
136						↑
137						no-op
138						↓
139						no-op

(a) Freeway (132 ~ 139 steps). The chicken crashes with one of the cars in 133-step, so it is forced to move down regardless of the actions.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
337						↑
338						↑
339						no-op
340						no-op
341						no-op
342						no-op
343						no-op
344						↑











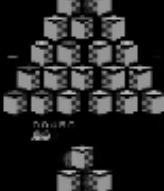
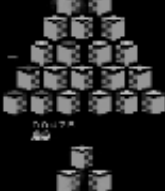
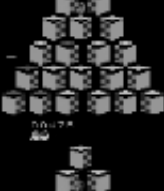
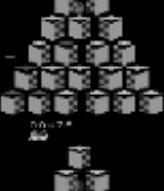
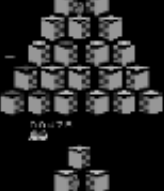


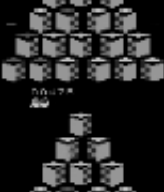
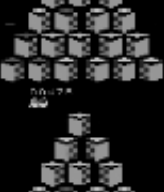
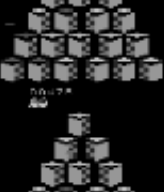















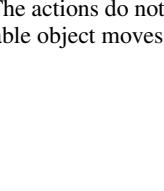
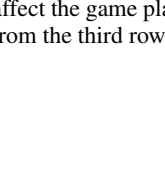
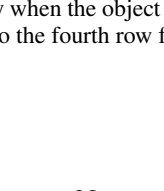
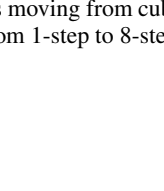
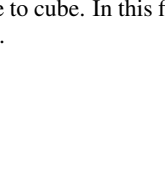
(a) Freeway (337 ~ 344 steps). When the chicken reaches the top lane in 338-step, it is relocated to the bottom lane in the next step. The feedforward model has a small translation.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
351						↑
352						↑
353						↑
354						↑
355						↑
356						↑
357						↑
358						↑

(a) Freeway (351 ~ 358 steps). The chicken disappears in the predictions made by the feedforward model. This is due to the fact that the chicken cannot move for 8-steps regardless of the action whenever it starts a new stage from the bottom lane. We conjecture that the feedforward model cannot handle this type of dependencies very well.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
493						no-op
494						no-op
495						↓
496						↑
497						↓
498						↓
499						↑
500						↑

(a) Freeway (493 ~ 500 steps). The recurrent network successfully predicts every object up to 500-step.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
1						no-op
2						↓
3						↓
4						fire
5						no-op
6						no-op
7						no-op
8						no-op

(a) QBert (1 ~ 8 steps). In this game, the player has to visit every cube location in order to change their colors. The actions do not affect the game play when the object is moving from cube to cube. In this figure, the controllable object moves from the third row to the fourth row from 1-step to 8-step.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
37						→
38						←
39						←
40						↓
41						no-op
42						no-op
43						↑
44						↓

(a) QBert (37 ~ 44 steps). When the agent visits every cube (40-step), the entire cubes flash for a few seconds (41 ~ 52 steps). Our models successfully handle this type of global changes in the screen.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
45						←
46						↑
47						←
48						←
49						↑
50						fire
51						↑
52						fire

(a) QBert (45 ~ 52 steps). The entire cubes are flashing because the player changed the colors of all the cubes in 40-step.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
110						→
111						→
112						→
113						←
114						→
115						→
116						↓
117						↓

(a) QBert (110 ~ 117 steps). After 114-step, the predicted object from the recurrent model disappears, while the feedforward model keeps track of it.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
1						→
2						→
3						→
4						→
5						→
6						↗
7						↗
8						↗

(a) Ms Pacman (1 ~ 8 steps). The pacman, which is controlled by the player, moves from left to right in the bottom corridor. The challenge in this game is that the shape of the pacman is very similar to that of the enemies, and the enemies are randomly blinking.

Step	naLinear	naFf	Feedforward	Recurrent	Ground Truth	Action
45						↑
46						→
47						↑
48						↑
49						no-op
50						no-op
51						no-op
52						no-op

(a) Ms Pacman (45 ~ 52 steps).

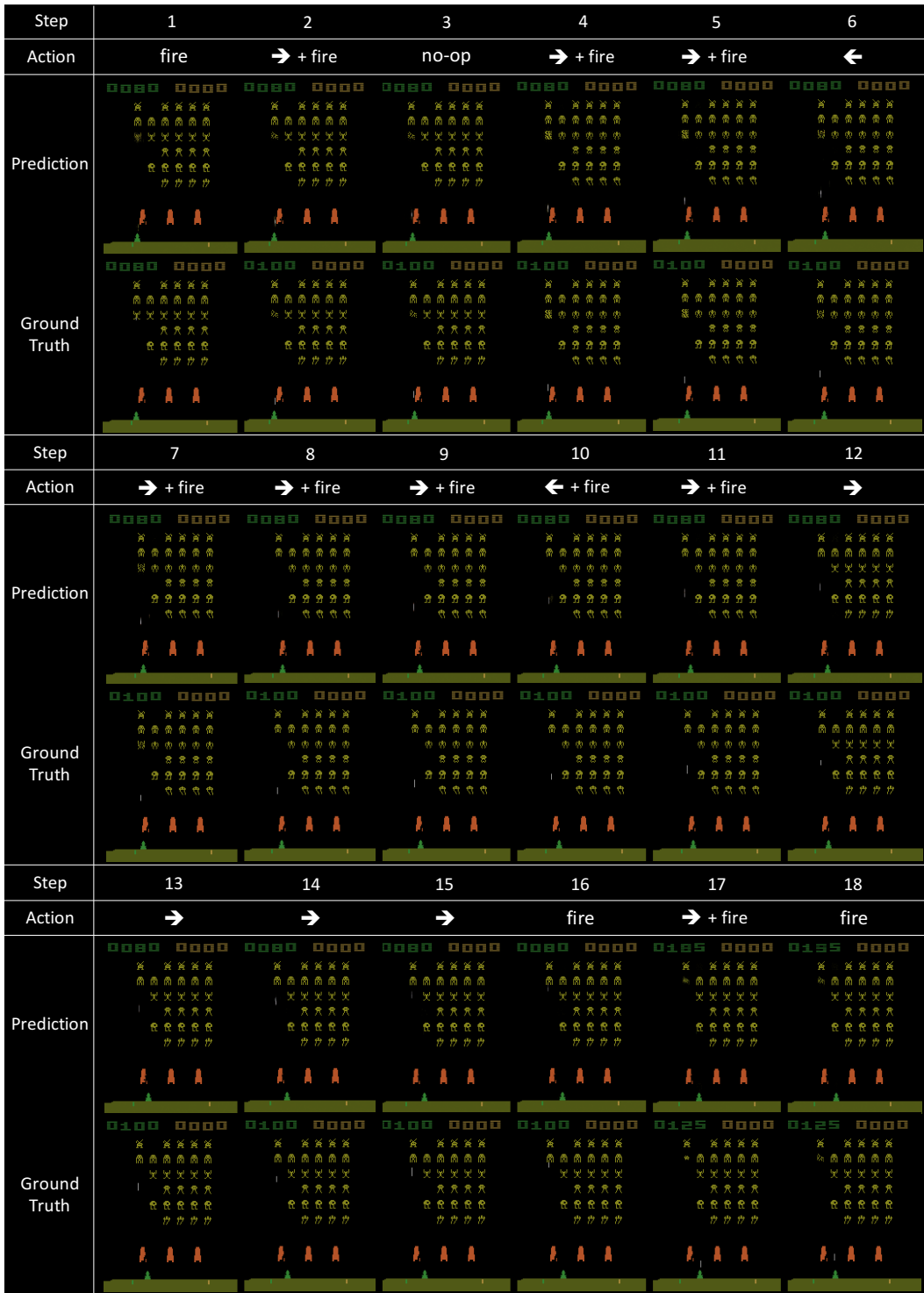
F Prediction video on original images

Step	1	2	3	4	5	6
Action	←	← + fire	↖	↘ + fire	↘ + fire	↘ + fire
Prediction						
Ground Truth						
Step	7	8	9	10	11	12
Action	↓	↓	↖ + fire	↖	← + fire	↓
Prediction						
Ground Truth						
Step	13	14	15	16	17	18
Action	↓	↘ + fire	↓	↘ + fire	↘ + fire	↖
Prediction						
Ground Truth						

(b) Seaquest (1 ~ 15 steps). Our model predicts the movement of the submarine and enemies correctly. It also predicts disappearing objects when the submarine is heading for them.

Step	153	154	155	156	157	158
Action	↑	← + fire	↖ + fire	↖	↓	↑
Prediction						
Ground Truth						
Step	159	160	161	162	163	164
Action	↖ + fire	↑ + fire	↓ + fire	no-op	↑ + fire	↗
Prediction						
Ground Truth						
Step	165	166	167	168	169	170
Action	↗	↖ + fire	↓ + fire	↖	↗	↗
Prediction						
Ground Truth						

(c) Seaquest (153 ~ 170 steps). The model generates new objects. Although the generated objects are not correct, their shape and color are realistic. In this figure, the model predicts that the submarine dies when it crashes with a 'virtual' enemy.



(d) Space Invaders (1 ~ 15 steps). The enemies move and change their shapes after 8 steps (see 3 ~ 4-step and 11 ~ 12 step). This requires the model to capture the temporal dependencies from the images in order to make an accurate prediction.

Step	154	155	156	157	158	159
Action	fire	← + fire	fire	fire	← + fire	→
Prediction						
Ground Truth						
Step	160	161	162	163	164	165
Action	← + fire	→	← + fire	←	fire	fire
Prediction						
Ground Truth						
Step	166	167	168	169	170	171
Action	fire	fire	fire	→ + fire	←	← + fire
Prediction						
Ground Truth						

(e) Space Invaders (154 ~ 171 steps). Although the model makes predictions errors in this long-term predictions, they are still realistic in the sense that the objects are reasonably arranged in a 2D-grid and moving in the right directions.

Step	1	2	3	4	5	6
Action	↑	↑	↑	↓	↑	↑
Prediction						
Ground Truth						
Step	7	8	9	10	11	12
Action	no-op	↑	↑	↑	↓	↑
Prediction						
Ground Truth						
Step	13	14	15	16	17	18
Action	↑	↑	no-op	↓	↓	↑
Prediction						
Ground Truth						

(f) Freeway (1 ~ 15 steps). The chicken controlled by the agent reaches the top lane in 3-step and starts a new stage from the bottom lane. Whenever it starts a new stage, the agent cannot move the chicken for a while (from 4-step to 12-step). Our model successfully handles this constraint.

Step	483	484	485	486	487	488
Action	↑	↑	↑	↑	no-op	↑
Prediction						
Ground Truth						
Step	489	490	491	492	493	494
Action	no-op	↓	↑	↑	↑	↑
Prediction						
Ground Truth						
Step	495	496	497	498	499	500
Action	↑	↑	↓	↑	↓	↓
Prediction						
Ground Truth						

(g) Freeway (483 ~ 500 steps). The predictions over 500 steps are very accurate in this game domain.

Step	1	2	3	4	5	6
Action	no-op	no-op	no-op	no-op	←	no-op
Prediction						
Ground Truth						
Step	7	8	9	10	11	12
Action	no-op	no-op	→	→	no-op	→
Prediction						
Ground Truth						
Step	13	14	15	16	17	18
Action	→	no-op	↓	→	←	→
Prediction						
Ground Truth						

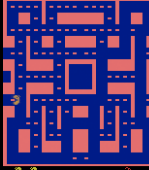
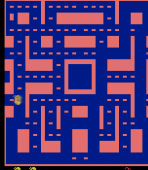
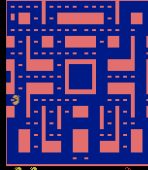
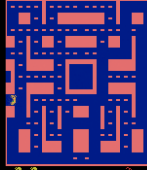
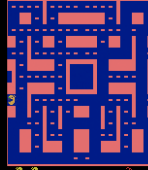
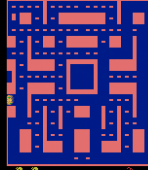
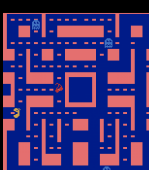
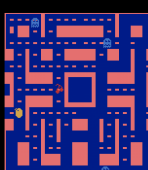

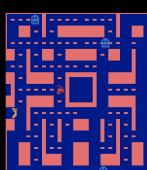
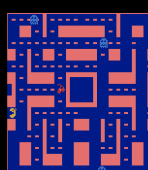
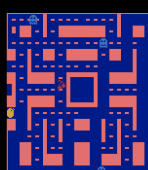
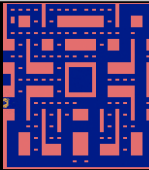
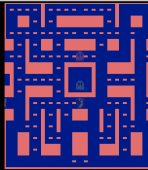
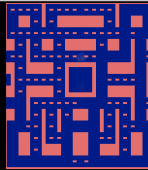
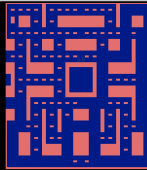
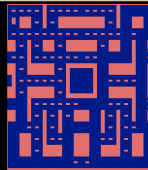
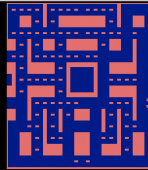
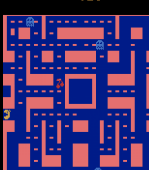
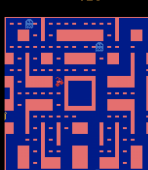
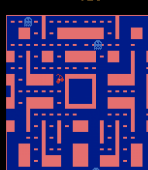
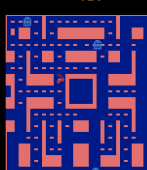
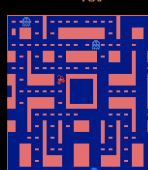
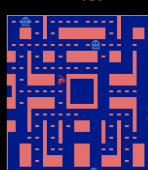
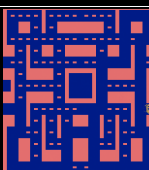
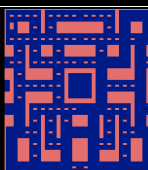
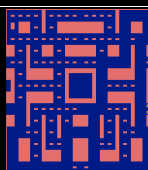
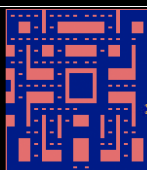
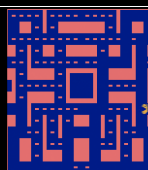
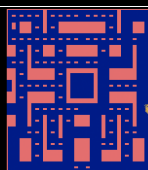
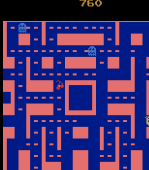
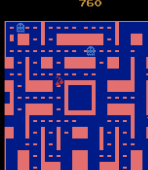
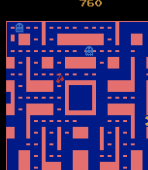
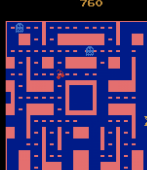
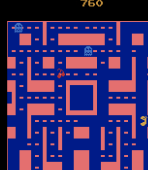
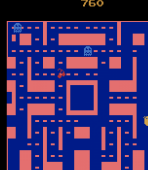
(h) QBert (1 ~ 15 steps). The controlled object is orange-colored, while the enemy is green-colored. The object moves between the third row and the fourth row and changes the color of the cube from blue to yellow, while the enemy moves to the third row and changes the color back to blue. Our model not only predicts the dynamic of the controllable object but also often predicts the movement of the enemy.

Step	72	73	74	75	76	77
Action	no-op	←	no-op	→	→	→
Prediction						
Ground Truth						
Step	78	79	80	81	82	83
Action	→	no-op	↑	no-op	no-op	no-op
Prediction						
Ground Truth						
Step	84	85	86	87	88	89
Action	no-op	↓	no-op	↑	↓	fire
Prediction						
Ground Truth						

(i) QBert (72 ~ 89 steps). The model predicts the movement of the object controlled by the agent well, while it is unable to predict the new object (the purple-colored ball).

Step	1	2	3	4	5	6
Action	↙	←	→	↘	↙	→
Prediction						
Ground Truth						
Step	7	8	9	10	11	12
Action	↙	↑	↙	↙	↘	↘
Prediction						
Ground Truth						
Step	13	14	15	16	17	18
Action	↘	↘	←	↙	↙	↙
Prediction						
Ground Truth						

(j) Ms Pacman (1 ~ 15 steps). Pacman moves and eats the blocks in the corridors. The model predicts the changing score as Pacman eats blocks.

Step	95	96	97	98	99	100
Action	↗	←	↙	↙	↙	↙
Prediction						
	740	740	740	750	750	750
Ground Truth						
	740	740	740	750	750	750
Step	101	102	103	104	105	106
Action	↙	↙	↙	↙	↓	→
Prediction						
	750	750	750	750	760	760
Ground Truth						
	750	750	750	750	750	760
Step	107	108	109	110	111	112
Action	↑	↙	↓	no-op	→	→
Prediction						
	760	760	760	760	760	760
Ground Truth						
	760	760	760	760	760	760

(k) Ms Pacman (95 ~ 112 steps). As pacman goes to the left end of the screen in 102-step, it appears in the right end of the screen in 104-step. This is an example of highly non-linear transformation, and requires the model to consider the global spatial context to predict the pixels. However, our model does not predict the movement of the enemies very well.