

---

# FutureGAN: Anticipating the Future Frames of Video Sequences using Spatio-Temporal 3d Convolutions in Progressively Growing GANs

---

Sandra Aigner and Marco Körner

Technical University of Munich

Munich, Germany

{sandra.aigner, marco.koerner}@tum.de

## Abstract

We introduce a new *encoder-decoder GAN* model, *FutureGAN*, that predicts future frames of a video sequence conditioned on a sequence of past frames. During training, the networks solely receive the raw pixel values as an input, without relying on additional constraints or dataset specific conditions. To capture both the spatial and temporal components of a video sequence, spatio-temporal 3d convolutions are used in all encoder and decoder modules. Further, we utilize concepts of the existing *progressively growing GAN (PGGAN)* that achieves high-quality results on generating high-resolution single images. The FutureGAN model extends this concept to the complex task of video prediction. We conducted experiments on three different datasets, *MovingMNIST*, *KTH Action*, and *Cityscapes*. Our results show that the model learned representations to transform the information of an input sequence into a plausible future sequence effectively for all three datasets. The main advantage of the FutureGAN framework is that it is applicable to various different datasets without additional changes, whilst achieving stable results that are competitive to the state-of-the-art in video prediction. Our code is available at <https://github.com/TUM-LMF/FutureGAN>

## 1 Introduction

Anticipating a possible future based on experience is an important part of the human decision-making process. Simulating this process in machines by teaching them to anticipate future events based on internal representations of the environment could be of relevance for many tasks. Automatically predicting the future frames of a video is one approach to tackle this problem. Video predictions can be of use for planning in robotics, as well as in autonomous driving, especially in reinforcement learning settings. They can lead to better decisions, or at least to faster executions, when used as an additional input to the agent. As shown by Mathieu et al. [29], other tasks, such as object recognition, detection, and tracking, can benefit from the representations that are implicitly learned by such a model.

There are several different approaches that address the pixel-level prediction of video frames. Early, often purely deterministic, approaches tended to insufficiently model the uncertainty of the output, which led to blurry predictions. Using *generative adversarial networks (GANs)* [12] is one way to appropriately model the uncertainty of the multi-modal output. We build on this idea of training a generative model in an adversarial setting. GANs learn to model the underlying data distribution implicitly by utilizing a critic, the *discriminator* network, during training time. While being trained, the critic constantly provides feedback to the *generator*, whether the generated samples appear real or not. This forces the generator to output samples of a similar data distribution as those of the real samples.

Although GAN based video prediction methods usually manage to better preserve the sharpness in the predicted frames, there are two major drawbacks. First, GANs are hard to train because the training process is highly unstable. Secondly, GANs often suffer from mode collapse effects [35], where the generator learns to fool the discriminator by producing samples of a limited set of modes. This means, the resulting generative model will not be able to fully capture the underlying data distribution.

In our *FutureGAN* model, we utilize the training strategy of *progressively growing GANs (PGGANs)* [18] that effectively managed to overcome these problems. The PGGAN of Karras et al. [18] was originally designed for generating high-resolution images from a set of random latent variables. On this task, it achieved high-quality results. The basic principle is to gradually increase the image resolution as the training proceeds by progressively adding layers in both networks. For further stabilization of the training, the authors introduced normalization techniques to constrain the signal magnitudes and the competition in both the discriminator and the generator. We extend this architecture for the complex task of video prediction to benefit from the positive effects on the GAN training.

The primary contribution of this paper is to provide a simple GAN-based model for video prediction that is directly applicable to different datasets using, in general, the same setting. Our FutureGAN predicts multiple future frames at once when conditioned on a set of past frames. Contrary to other approaches, both networks solely use the raw pixel value information as an input, without relying on additional conditions. To evaluate the FutureGAN framework, we conducted experiments on three datasets of increasing complexity, *i.e.* the *MovingMNIST* dataset [37], the *KTH Action* dataset [36], and the *Cityscapes* dataset [6]. Figure 1 provides example predictions. We show that our model is able to generate plausible futures for all three datasets, while avoiding the problems that typically arise when training GANs. The predicted frames indicate that the model effectively learned representations of spatial and temporal transformations.

## 2 Related Work

Since 2014, predicting the future frames of a video, from either a single input frame or a sequence of input frames, has become a widely researched topic. Ranzato et al. [33] were the first to provide a baseline model for video prediction with deep neural networks. Since then, various other approaches were introduced. Most of these combine the raw pixel values of the input frame(s) with learned temporal components [13, 17, 24, 26, 31, 37, 43, 45], dynamically learned filters [7], latent variables [13], or by explicitly incorporating time dependency [43]. Oliu et al. [31], for example, generate future video frames with a *folded recurrent neural network (fRNN)*. This network uses *bijective gated recurrent units (bGRUs)* that learn shared video representations between the encoder and decoder. Others learn separate representations for the static and dynamic components of a video by adding action or geometry-based conditions, such as pose, optical flow, or depth information [4, 11, 15, 28, 30, 32, 47].

The most promising results, especially for long-term predictions, have been achieved just recently by approaches that explicitly include stochasticity in their models [2, 8, 9, 22, 44, 47]. Those methods directly address the uncertainty in predicting future frames. Generating a set of possible predictions, rather than a single prediction that averages over all modes, prevents from the effect of blurred predictions for an increasing number of time steps.

Another attempt to address the uncertainty in predicting future frames, thus reducing the blurring effect, is to train the generative models in an adversarial setting. Our approach follows this research branch. Mathieu et al. [29] showed first, networks trained with an adversarial loss term tend to produce sharper results compared to networks only trained on pixel error-based loss metrics, such as the L2 loss. The idea of using GANs for making video predictions further evolved when traditional image generation GANs were extended for video generation [34, 41]. Vondrick et al. [41] use a two-

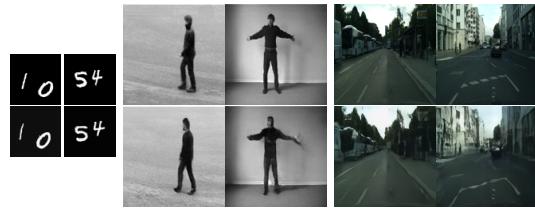


Figure 1: Example predictions for the MovingMNIST, KTH Action, and Cityscapes dataset. Top row: Last observed input frame. Bottom row: Next frames predicted by our FutureGAN.

stream network, where foreground and background streams are separated. This network generates a sequence of 32 frames using layer-wise spatial and temporal up-sampling with *3d convolutions* [38]. When exchanging the generator’s input from random latent variables to the pixel values of an input image, the network then learns to predict future frames. Kratzwald et al. [20] build on the approach of Vondrick et al. [41]. They jointly predict the dynamic and static patterns with an extended Wasserstein GAN (WGAN) [1]. For video generation and prediction, they combine the application-specific L2 loss and the adversarial loss term.

In contrast to our approach, many GAN-based video prediction methods add additional information, such as motion, content, geometry or action-based conditions, or learn those components separately [3, 5, 9, 23, 27, 39, 40, 40, 42, 46, 48]. Villegas et al. [40], for instance, use an encoder-decoder convolutional neural network (CNN) with *convolutional long short-term memory (ConvLSTM)* units to make pixel-level predictions. Their network learns to model the motion and content component of the input sequence independently, using separate encoders. GAN approaches often use deterministic *autoencoder (AE)*-based networks with LSTM units where the networks are trained in an adversarial setting. In many cases, the adversarial term is then added to the loss function [25].

Mostly related to our approach are [3, 20, 29, 41], but the applied losses and the training strategies differ. We follow the idea of using a multi-scale GAN setting for video prediction. The idea of a multi-scale or multi-stage GAN for this task has previously been addressed, by either having separate networks, or layer-wise up-sampling operations [3, 20, 29, 41, 42]. It is, however, new in this context to add the layers progressively for increasing the image resolution during training.

### 3 FutureGAN Model

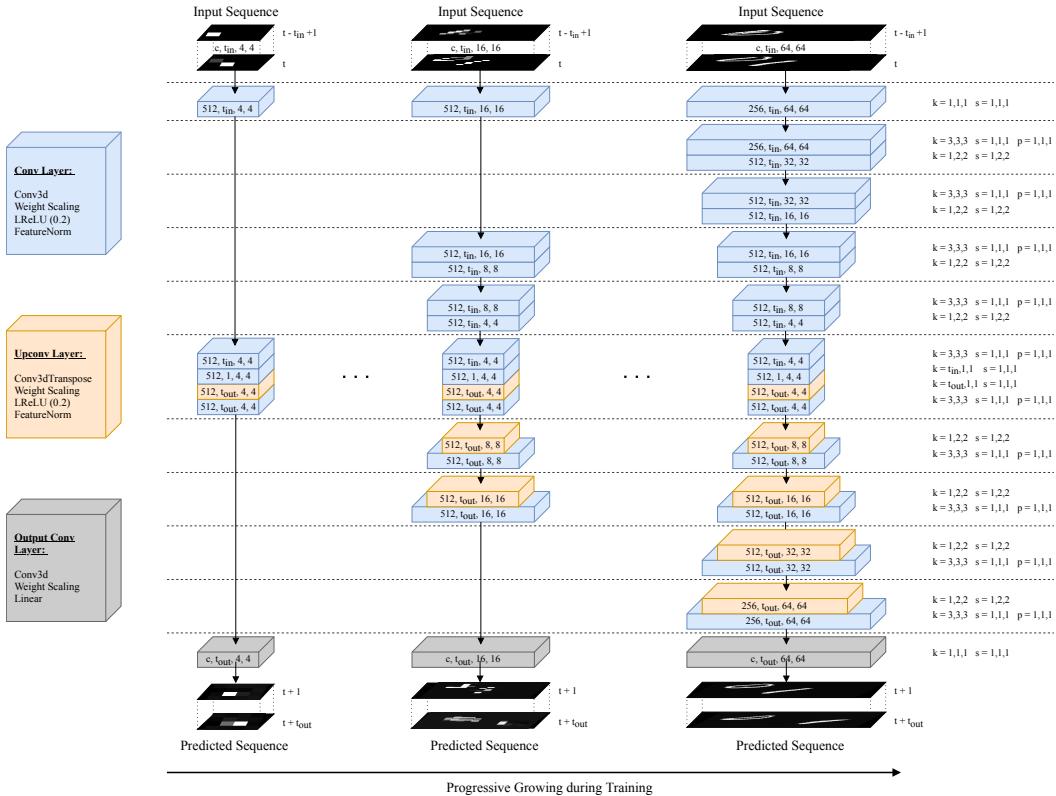


Figure 2: **FutureGAN generator during training.** We initialize our model to take a set of  $4 \times 4$  px resolution frames and output frames of the same resolution. During training, layers are added progressively to double the resolution after a specified number of iterations. The resolution of the input frames always matches the resolution of the current state of the network. This figure illustrates the growth progress of the generator exemplary for the MovingMNIST dataset with a final resolution of  $64 \times 64$  px. Intermediate  $8 \times 8$  px and  $32 \times 32$  px resolution steps are left out for visual clarity.

The FutureGAN framework is based on the idea of training a generative model in an adversarial setting and therefore consists of two separate networks. Our generator network is trained to predict a sequence of future video frames given a sequence of past frames. The second network, the discriminator, is trained to distinguish between the generated sequence and a real sequence from the training dataset. The discriminator alternately receives real and fake sequences as an input and calculates a score whether the sequence appears real or not. An output score close to 0 indicates the discriminator rates a given sequence as probably fake. The higher the output score of the discriminator for a given sequence, the more realistic it appears to the network. The generator network updates its weight parameters according to the feedback it receives from the discriminator, trying to generate sequences that will fool the discriminator. Because the training of GANs tended to be highly unstable, we build on the recently proposed PGGAN approach by Karras et al. [18] that effectively managed to overcome these problems. We describe the architecture and training strategy of our proposed FutureGAN model in the following.

### 3.1 Generator Network

Our generator network  $G$  processes the frames of an input sequence and transforms them into future frames of this sequence. The output of the generator can be described as the sequence of future frames  $\tilde{\mathbf{x}} = G(\mathbf{z}) = (\tilde{\mathbf{x}}_{t+1}, \dots, \tilde{\mathbf{x}}_{t+t_{\text{out}}})$ , and the input as the sequence of past frames  $\mathbf{z} = (\mathbf{x}_{t-t_{\text{in}}+1}, \dots, \mathbf{x}_t)$ . The parameter  $t_{\text{in}}$  corresponds to the temporal depth of the input sequence,  $t_{\text{out}}$  corresponds to the temporal depth of the output sequence. To enable predictions of video frames based on an input sequence, the FutureGAN generator consists of an encoder and a decoder part. We extend the PGGAN generator of Karras et al. [18] to include an encoder that learns a latent representation of the input. This latent representation is used by a decoder to generate the predictions.

For the decoder part of our generator, we modify the basic architecture of the PGGAN generator. Figure 2 illustrates the detailed structure and main components of the FutureGAN generator. Instead of 2d convolutions, we use 3d convolutions in all convolutional layers. This enables the generator to properly encode and decode both the spatial and temporal components of the input sequence. Additionally, we realize the spatial upsampling between layers operating on different frame resolutions within a single convolutional layer. To perform spatial upsampling only, we use transposed 3d convolutions with asymmetric kernel sizes and strides. Originally, Karras et al. [18] use a nearest neighbor upsampling layer and a convolutional layer separately. The encoder part of our generator mirrors the structure of the decoder part, except that the spatial upsampling layers are replaced by spatial downsampling layers. We use 3d convolutions with asymmetric kernel sizes and strides to perform spatial downsampling only. The bottleneck layers of our generator perform temporal downsampling and upsampling operations to match the temporal depth to the number of input frames and desired output frames, respectively. Following the basic design of the PGGAN generator, we add two convolutional layers in the encoder and in the decoder part to increase the network resolution. To introduce non-linearity in the networks, leaky rectified linear units (*LReLU*) follow each convolution in the hidden layers. After each *LReLU* activation function, a pixel-wise feature vector normalization layer is inserted.

### 3.2 Discriminator Network

The discriminator of our FutureGAN model is designed to distinguish between real and fake sequences. As an input, the discriminator network  $D$  alternately receives  $\mathbf{x} = (\mathbf{x}_{t-t_{\text{in}}+1}, \dots, \mathbf{x}_{t+t_{\text{out}}})$  frames from the training set, representing the ground truth sequence, and  $\tilde{\mathbf{x}} = (\mathbf{z}, G(\mathbf{z})) = (\mathbf{x}_{t-t_{\text{in}}+1}, \dots, \tilde{\mathbf{x}}_{t+t_{\text{out}}})$ . The latter sequence consists of the input and output frames of the generator. The output of the discriminator network is a score  $s = D(\mathbf{x})$  or  $\tilde{s} = D(\tilde{\mathbf{x}})$ , respectively. This score ranks the given input as either being real or fake. We set the labels for the real sequence to  $l_{\text{real}} = 1$  and the labels for fake sequences to  $l_{\text{real}} = 0$ .

Apart from the bottleneck layers, the FutureGAN discriminator closely resembles the encoder part of our generator network. One important difference is that there are no pixel-wise feature vector normalization layers in the discriminator. Additionally, a mini-batch standard deviation layer is added to one of the last layers. Karras et al. [18] inserted this layer to increase the variation in the generator's outputs, thus to prevent mode collapse. This layer computes the standard deviation for each feature in each spatial location over the mini-batch. Averaging these values over all features and spatial locations produces a scalar value. This value is replicated for every spatial location in the

mini-batch, which generates an additional feature map. We modify the original layer to calculate this constant feature map for temporal depth as well as spatial locations, in order to increase variation, especially in the temporal domain. To reduce the output of the discriminator to a single scalar, the last layer consists of a fully connected layer, followed by a linear activation function. A figure showing the detailed structure and main components of the discriminator is included in appendix B.

### 3.3 Training Procedure

We initialize our networks to start the training process with a frame resolution of  $4 \times 4$  px. This resolution is gradually increased by a factor of two after the networks have trained for a specified number epochs. The number of feature maps in each layer initially is 512. Starting from a frame resolution of  $64 \times 64$  px, the number of feature maps is halved for all newly added layers. Figure 2 illustrates the progressive growing for the FutureGAN generator. Our FutureGAN training closely follows the training procedure described in [18]. The next paragraphs briefly introduce the main concepts, for further details we refer to the original paper.

**Adding layers for increased resolutions** Adding new layers to the networks is completed in two steps to ensure a smooth transition between two resolutions. The first step is the *transition phase*, where the layers operating on the frames of the next resolution are treated as a residual block whose weights  $\alpha$  increase linearly from 0 to 1. While the model is in the transition phase, interpolated inputs are fed into both of the networks, making the input frames match the resolution of the current state of the networks. The second step is the *stabilization phase*, where the networks are trained for a specified number of iterations before the resolution is doubled again. Growing the networks progressively both speeds up and stabilizes the training, as the networks only need to learn small transformations between the existing and the newly added layers.

**Weight scaling** To further stabilize the training, a weight-scaling layer is added on top of all the layers. This layer estimates the element-wise standard deviation of the weights and normalizes them to  $\hat{w}_i = w_i/c$ , where  $w_i$  are the layer weights and  $c$  is the normalization constant from *He's initializer* [16]. Using this layer in a network equalizes the dynamic range, and thus the learning speed, for all weights.

**Feature normalization in the generator** Another element for stabilizing the training process is the pixel-wise feature vector normalization in the generator. This element follows the activations of each convolutional layer. Based on a variant of the *local response normalization* [21], the feature vector is normalized to unit length in each pixel. To make this layer applicable to the FutureGAN generator, we modified it to operate on both the spatial and temporal elements of the feature maps.

The procedure can be described as  $b_{x,y,z} = a_{x,y,z} / \sqrt{\frac{1}{n_f} a_{x,y,z}^\top a_{x,y,z} + \epsilon}$ , where  $\epsilon = 10^{-8}$ ,  $n_f$  is the number of feature maps,  $a_{x,y,z}$  is the original, and  $b_{x,y,z}$  the normalized feature vector of the pixel  $(x, y, z)$ . Using this technique prevents the escalation of signal magnitudes in the generator and discriminator that result from an unhealthy competition between the two networks.

**WGAN-GP loss with epsilon penalty** Our loss function consists of the *Wasserstein GAN with gradient penalty* (WGAN-GP) loss [14] and an additional term to prevent the loss from drifting, the epsilon-penalty term. Using the WGAN-GP loss effectively increases the quality of the generated frames.

The WGAN-GP loss with epsilon penalty for optimizing the discriminator is defined as

$$L_D(\mathbf{x}, \tilde{\mathbf{x}}, \hat{\mathbf{x}}) = \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_g} [D(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r} [D(\mathbf{x})]}_{\text{WGAN loss}} + \underbrace{\lambda \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} [(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2]}_{\text{gradient-penalty}} + \underbrace{\varepsilon \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r} D(\mathbf{x})^2}_{\text{epsilon-penalty}}, \quad (1)$$

where  $\mathbb{P}_r$  is the data distribution,  $\mathbb{P}_g$  is the model distribution implicitly defined by  $\tilde{\mathbf{x}} = G(\mathbf{z})$ ,  $\tilde{\mathbf{x}} \sim p(\tilde{\mathbf{x}})$ ,  $\varepsilon$  is the epsilon-penalty coefficient, and  $\lambda$  is the gradient-penalty coefficient.  $\mathbb{P}_{\hat{\mathbf{x}}}$  is implicitly defined, sampling uniformly along straight lines between pairs of points sampled from the data distribution  $\mathbb{P}_r$  and the generator distribution  $\mathbb{P}_g$ .

The WGAN(-GP) loss for optimizing the generator is defined as

$$L_G(\tilde{\mathbf{x}}) = - \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_g} [D(\tilde{\mathbf{x}})]. \quad (2)$$

## 4 Experiments and Evaluation

We conducted experiments on three datasets of increasing complexity, *i.e.* the MovingMNIST dataset [37], the KTH Action dataset [36], and the Cityscapes dataset [6]. The experiments on the MovingMNIST and the KTH Action dataset were carried out on an NVIDIA Tesla P100 GPU with 16 GB of RAM. For the experiments on the Cityscapes dataset, we used an NVIDIA Titan X Pascal GPU with 12 GB RAM. The FutureGAN model is implemented in PyTorch. For the optimization, we used the *ADAM optimizer* [19] with  $\beta_1 = 0.0$  and  $\beta_2 = 0.99$ . Our initial learning rate was heuristically set to  $l = 0.001$ . Every resolution step, we adjusted the batchsize dynamically during training, according to available GPU RAM. Therefore, we decay our learning rate by a factor of 0.87 in each resolution step. The penalty coefficients of the WGAN-GP loss with epsilon-penalty were set to  $\lambda = 10$  and  $\varepsilon = 0.001$ , as proposed in [18].

On the MovingMNIST dataset, we trained our network until a resolution of  $64 \times 64$  px. The resolution of the MovingMNIST data already matched the final network resolution. For the KTH Action dataset, and the Cityscapes dataset, we used a final resolution of  $128 \times 128$  px. The original size of the KTH Action videos is  $160 \times 120$  px, the Cityscapes frames have an original size of  $2048 \times 1024$  px. These resolutions did not match the size of our final network resolution. Therefore, we resized all frames of both datasets bicubically to a resolution of  $128 \times 128$  px, beforehand. During training, the frames were downsampled to match the current resolution of the networks using nearest neighbor interpolation. All networks were trained for 10 epochs each in transition and stabilization phase of every resolution step, and another 20 epochs in the final phase. This results in a total number of training epochs of 120 for MovingMNIST, and 140 for KTH Action and Cityscapes.

To evaluate the networks quantitatively, we provide values for the *mean squared error (MSE)*, the *peak signal-to-noise ratio (PSNR)*, and *structural similarity index (SSIM)* between the ground truth and the predicted frame sequence. We compare our FutureGAN models to a naive baseline of simply copying the last frame of the input sequence, as well as to state-of-the-art approaches. Additionally, we provide comparisons between the optical flow maps of the ground truth and the predicted frames in appendix A.

Table 1: Average results over 6 frames for the full test splits (best results are bold)

	MovingMNIST		KTH Action	
	MSE	SSIM	MSE	SSIM
CopyLast	0.2580	0.6791	0.0116	0.8532
FutureGAN (ours)	<b>0.1603</b>	<b>0.7780</b>	0.0120	0.6180
fRNN [31]	0.1854	0.7408	0.0181	0.7053
MCNet [40]	—	—	<b>0.0048</b>	<b>0.8692</b>

### 4.1 MovingMNIST

To verify the effectiveness of our model architecture in general, we utilized the MovingMNIST dataset as a toy example. Following the procedure described in [37], we generated a set of 4500 videos for training, each of length 36 frames. Every MovingMNIST frame displays two white bouncing digits of distinct classes on a black background. Our generator network was trained to predict six future frames while being conditioned on six input frames, thus a total of 13499 sequences was used for training. For testing, we generated another set of 2250 videos of length 36 frames, resulting in a test set containing 6750 sequences.

In Figure 4, we show a qualitative comparison of our FutureGAN model to the fRNN model of Oliu et al. [31]. The average quantitative results are listed in table 1. We provide the per frame values of the quantitative measures in figure 3. Note that we used the pre-trained models provided by the original authors to generate the results. This means, the fRNN was trained to predict 10 frames based on 10 input frames.

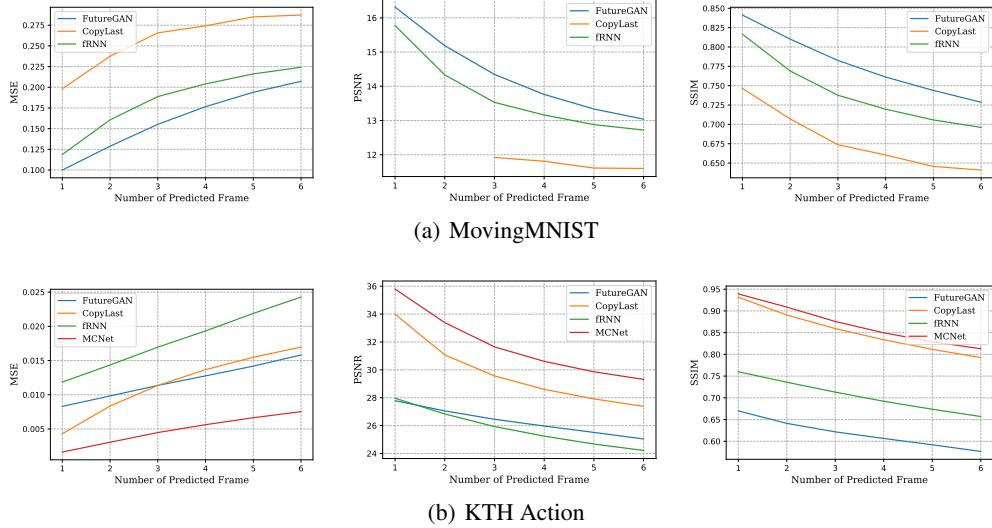


Figure 3: Quantitative results per predicted frame for the full test splits.

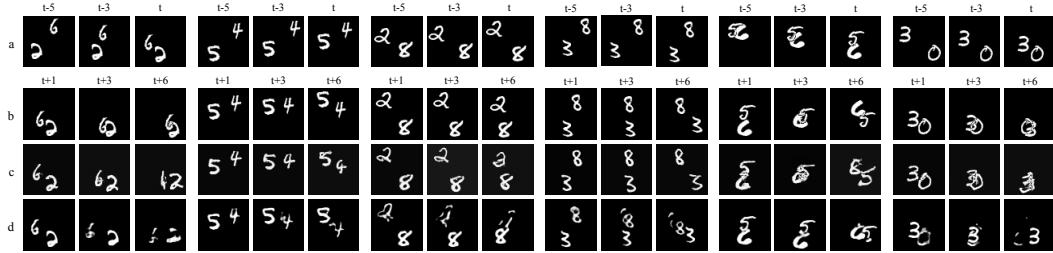


Figure 4: Prediction results for the MovingMNIST test sequences. a: Input, b: Ground Truth, c: FutureGAN (ours), d: fRNN [31]

## 4.2 KTH Action

In the second set of experiments, we used the KTH Action dataset. This dataset consists of 600 videos that display 25 different persons, each performing six actions in four different scenarios. The grayscale videos were recorded with a frame rate of 25 fps and have varying length. We split the dataset into person 1 to 16 for training, and 17 to 25 for testing. The FutureGAN model was trained to predict six future frames conditioned on six past frames. In total, our training set consists of 15156 sequences for predicting. Our test set had 8722 sequences.

In Figure 5, we show a qualitative comparison of our FutureGAN model to the fRNN and the *MCNet* of Villegas et al. [40]. The average quantitative results are listed in table 1. We provide the per frame values of the quantitative measures in figure 3. For testing the MCNet, we used the pre-trained models provided by the original authors to generate the results. The fRNN was originally trained for frame resolutions of  $80 \times 64$  px. We re-trained the fRNN model on sequences with a  $128 \times 128$  px frame resolution, following the procedure of [31]. This means, both the fRNN and the MCNet model were trained to predict 10 frames based on 10 input frames.

## 4.3 Cityscapes

To further investigate whether our model is able to scale to more complex real-world scenes, we trained it on the Cityscapes dataset. This dataset contains 2975 training videos and 1525 test videos, each of 30 frames in length. The 16 bit color videos were recorded with a frame rate of 17 fps in 50 different cities of Germany. We took the training and testing set as split by Cordts et al. [6]. Each split contains the videos from a different set of cities. The FutureGAN was trained to predict five

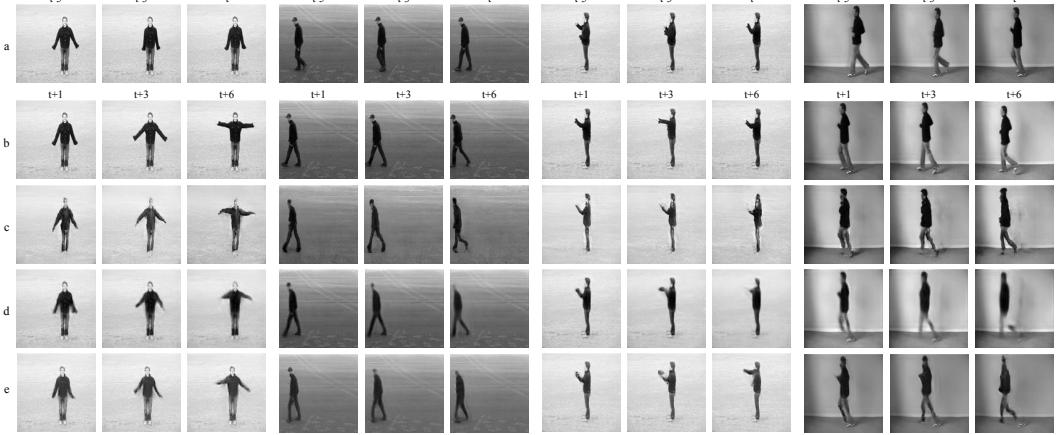


Figure 5: Prediction results for the KTH Action test split. a: Input, b: Ground Truth, c: FutureGAN (ours), d: fRNN [31], e: MCNet [40].

future frames based on five input frames. In total, our training set consisted of 8924 frame sequences. The test set had 4574 sequences.

In Figure 6, we display the qualitative results of our FutureGAN model trained on Cityscapes. For brevity, we provide the quantitative values in appendix A.4 figure A.6.

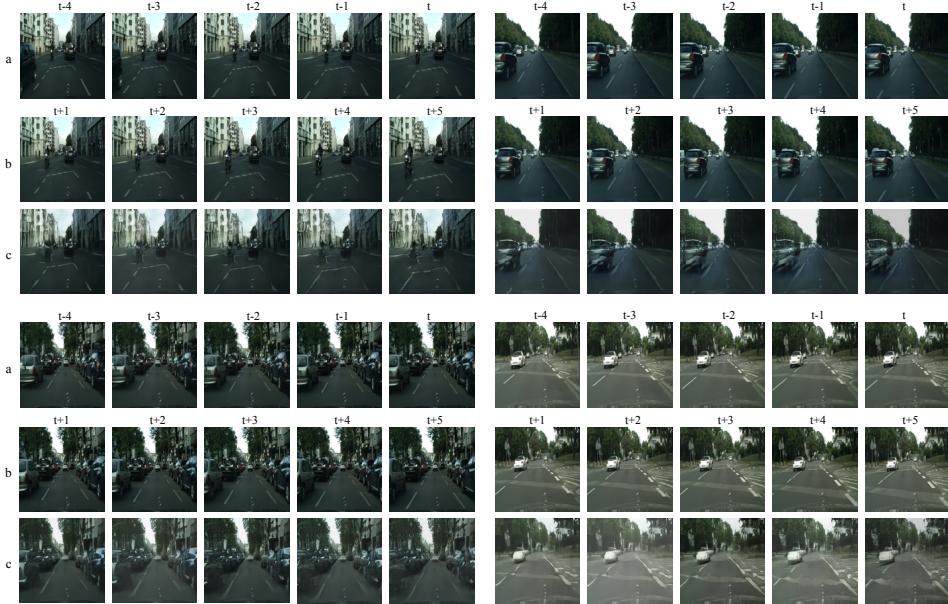


Figure 6: Prediction results for the Cityscapes test sequences. a: Input, b: Ground Truth, c: FutureGAN (ours)

#### 4.4 Long-term Predictions

To test the generalization abilities of our network, we generated long-term predictions for all three datasets. This was achieved by feeding the predictions recursively back in as inputs.

Figure 7 shows the qualitative results of this experiment. For MovingMNIST, we generated predictions for 30 frames ahead, letting the network observe only one real sequence of 6 input frames. On the KTH Action dataset, also predictions up to 30 frames ahead were made, while only one real sequence of 6 input frames was observed. Additionally, we provide results for predictions up to 120 steps

ahead for the KTH Action dataset in appendix A.3. For Cityscapes, we generated predictions 25 frames ahead, when the network only observed one real sequence of 5 input frames.

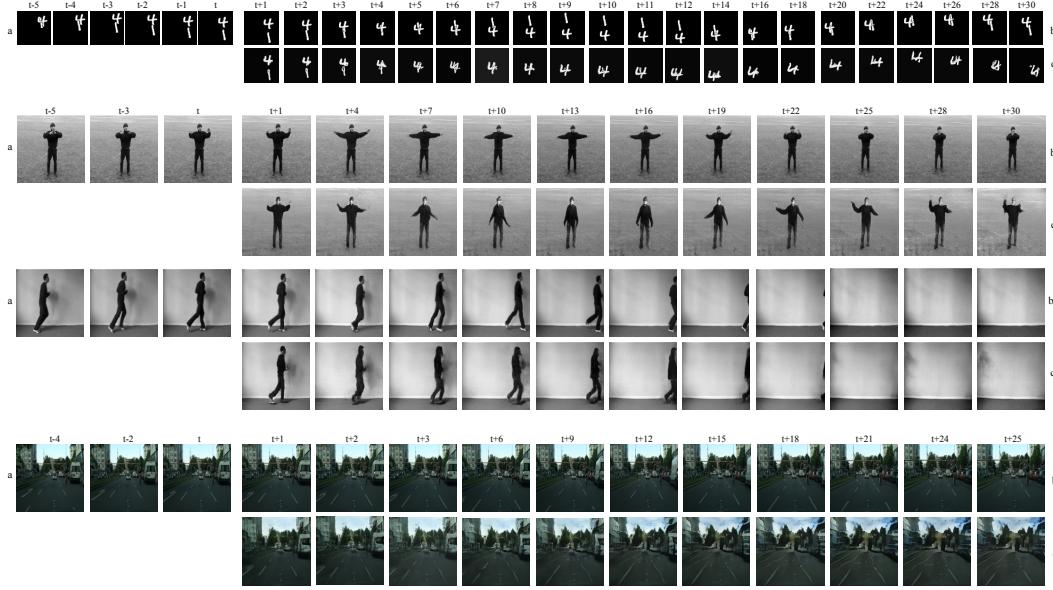


Figure 7: Prediction results of the long-term predictions for all three datasets. a: Input, b: Ground Truth, c: FutureGAN (ours)

## 5 Conclusion and Discussion

In this paper, we have proposed FutureGAN, a new model that predicts future video frames conditioned on an input sequence. By extending the existing PGGAN architecture to video prediction, we are able to predict future frames that appear realistic, while the problems that typically arise when training GANs are avoided. Our proposed model is trained to predict multiple future frames at once, using a similar setting for different datasets. This makes FutureGAN directly applicable to a variety of datasets without utilizing dataset specific domain knowledge. Contrary to other approaches, our networks solely use the raw pixel values as an input, without relying on additional priors, or conditional information.

To evaluate our model, we trained and tested it on three datasets of increasing complexity. For MovingMNIST and KTH Action, we used an identical training setting, except for the dataset size and final frame resolution. The predicted frames show that the network effectively learned representations of spatial and temporal transformations for the two datasets. Our network identifies moving pixels in the input frames and transforms them based on its learned internal representations. For both datasets, the results are competitive to the state-of-the-art. The qualitative results of the Cityscapes dataset suggest that our model scales to complex natural traffic scenes as well. We observed that FutureGAN applies separate motion patterns to the background and foreground pixels. Furthermore, it seems as if the network was able to learn scene-specific representations of ego-motion. In most cases, it applies the correct motion patterns for either a static or dynamic background based on the input sequence. Even though the networks were trained on fewer frames, they generalize reasonably well to predict deeper into future for the KTH Action and Cityscapes dataset. The predictions still appear plausible, although the frames tend to get blurrier for increasing numbers of time steps.

Our experiments verify that the progressive growth strategy of Karras et al. [18] scales effectively to the more complex video prediction task. FutureGAN is a highly flexible model that can easily be trained on various datasets of different resolutions without prior knowledge about the data.

## Acknowledgements

The authors are very grateful for the computing resources provided by the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities (BAdW) and their excellent team, with special thanks to Yu Wang. Further, we gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal GPU, used for this research. We thank Marc Oliu Simón for kindly providing the pre-trained fRNN models. Finally, we want to appreciate the valuable feedback of our colleague Lloyd Hughes and our colleagues of the Computer Vision Research Group.

## References

- [1] M. Arjovsky, S. Chitala, and L. Bottou. Wasserstein GAN. In *ICML*, 2018.
- [2] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic Variational Video Prediction. In *ICLR*, 2018.
- [3] P. Bhattacharjee and S. Das. Temporal Coherency based Criteria for Predicting Video Frames using Deep Multi-stage Generative Adversarial Networks. In *NIPS*. Curran Associates, Inc., 2017.
- [4] W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos. Fully Context-Aware Video Prediction. In *ECCV*, 2018.
- [5] B. Chen, W. Wang, J. Wang, and X. Chen. Video Imagination from a Single Image with Transformation Generation. In *MM*, 2017.
- [6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.
- [7] B. De Brabandere, X. Jia, T. Tuytelaars, and L. Van Gool. Dynamic Filter Networks. In *NIPS*. Curran Associates, Inc., 2016.
- [8] E. Denton and R. Fergus. Stochastic Video Generation with a Learned Prior. In *ICML*, 2018.
- [9] E. L. Denton and V. Birodkar. Unsupervised Learning of Disentangled Representations from Video. In *NIPS*. Curran Associates, Inc., 2017.
- [10] G. Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In *SCIA*, 2003.
- [11] C. Finn, I. Goodfellow, and S. Levine. Unsupervised Learning for Physical Interaction through Video Prediction. In *NIPS*. Curran Associates, Inc., 2016.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *NIPS*. Curran Associates, Inc., 2014.
- [13] R. Goroshin, M. Mathieu, and Y. LeCun. Learning to Linearize Under Uncertainty. In *NIPS*. Curran Associates, Inc., 2015.
- [14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. In *NIPS*. Curran Associates, Inc., 2017.
- [15] Z. Hao, X. Huang, and S. Belongie. Controllable Video Generation with Sparse Trajectories. In *CVPR*, 2018.
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *ICCV*, 2015.
- [17] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video Pixel Networks. In *ICML*, 2017.
- [18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018.
- [19] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.
- [20] B. Kratzwald, Z. Huang, D. P. Paudel, , A. Dinesh, and L. Van Gool. Improving Video Generation for Multi-functional Applications. *CoRR*, abs/1711.11453, 2017.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*. Curran Associates, Inc., 2012.
- [22] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic Adversarial Video Prediction. *CoRR*, abs/1804.01523, 2018.
- [23] X. Liang, L. Lee, W. Dai, and E. P. Xing. Dual Motion GAN for Future-Flow Embedded Video Prediction. In *ICCV*, 2017.
- [24] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video Frame Synthesis Using Deep Voxel Flow. In *ICCV*, 2017.
- [25] W. Lotter, G. Kreiman, and D. Cox. Unsupervised Learning of Visual Structure using Predictive Generative Networks. In *ICLR*, 2016.
- [26] W. Lotter, G. Kreiman, and D. Cox. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. In *ICLR*, 2017.
- [27] C. Lu, M. Hirsch, and B. Schlkopf. Flexible Spatio-Temporal Networks for Video Prediction. In *CVPR*, 2017.
- [28] R. Mahjourian, M. Wicke, and A. Angelova. Geometry-Based Next Frame Prediction from Monocular Video. In *IV*, 2017.

- [29] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.
- [30] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-Conditional Video Prediction using Deep Networks in Atari Games. In *NIPS*. Curran Associates, Inc., 2015.
- [31] M. Oliu, J. Selva, and S. Escalera. Folded Recurrent Neural Networks for Future Video Prediction. In *ECCV*, 2018.
- [32] V. Patraucean, A. Handa, and R. Cipolla. Spatio-Temporal Video Autoencoder with Differentiable Memory. In *ICLR*, 2016.
- [33] M.'A. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (Language) Modeling: A Baseline for generative Models of natural Videos. *CoRR*, abs/1412.6604, 2014.
- [34] M. Saito, E. Matsumoto, and S. Saito. Temporal Generative Adversarial Nets With Singular Value Clipping. In *ICCV*, 2017.
- [35] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *NIPS*. Curran Associates, Inc., 2016.
- [36] C. Schüldt, I. Laptev, and B. Caputo. Recognizing Human AAction: A Local SVM Approach. In *ICPR*, 2004.
- [37] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised Learning of Video Representation using LSTMs. In *ICML*, pages 843–852, 2015.
- [38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, 2015.
- [39] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *CVPR*, 2018.
- [40] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing Motion and Content for Natural Video Sequence Prediction. In *ICLR*, 2017.
- [41] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating Videos with Scene Dynamics. In *NIPS*. Curran Associates, Inc., 2016.
- [42] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating the Future with Adversarial Transformers. In *CVPR*, 2017.
- [43] V. Vukoti, A.-L. Pintea, C. Raymond, G. Gravier, and J. Van Gemert. One-Step Time-Dependent Future Video Frame Prediction with a Convolutional Encoder-Decoder Neural Network. In *NCCV*, 2016.
- [44] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An Uncertain Future: Forecasting from Static Images using Variational Autoencoders. In *ECCV*, 2016.
- [45] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu. PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs. In *NIPS*. Curran Associates, Inc., 2017.
- [46] W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo. Learning to Generate Time-Lapse Videos Using Multi-Stage Dynamic Generative Adversarial Networks. In *CVPR*, 2018.
- [47] T. Xue, J. Wu, K. L. Bouman, and W. T. Freeman. Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks. In *NIPS*. Curran Associates, Inc., 2016.
- [48] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. C. Niebles. Visual Forecasting by Imitating Dynamics in Natural Sequences. In *ICCV*, 2017.

# Appendix

## A Optical Flow Evaluation

We computed the dense optical flow maps using the Gunnar Farneback's algorithm [10].

### A.1 MovingMNIST Predictions

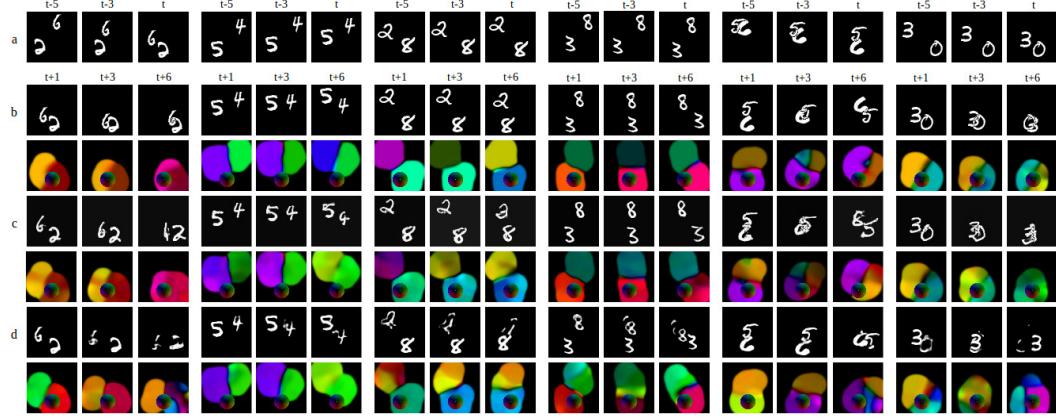


Figure A.1: Prediction results for the MovingMNIST test sequences. a: Input, b: Ground Truth, c: FutureGAN (ours), d: fRNN [31]

### A.2 KTHAction Predictions

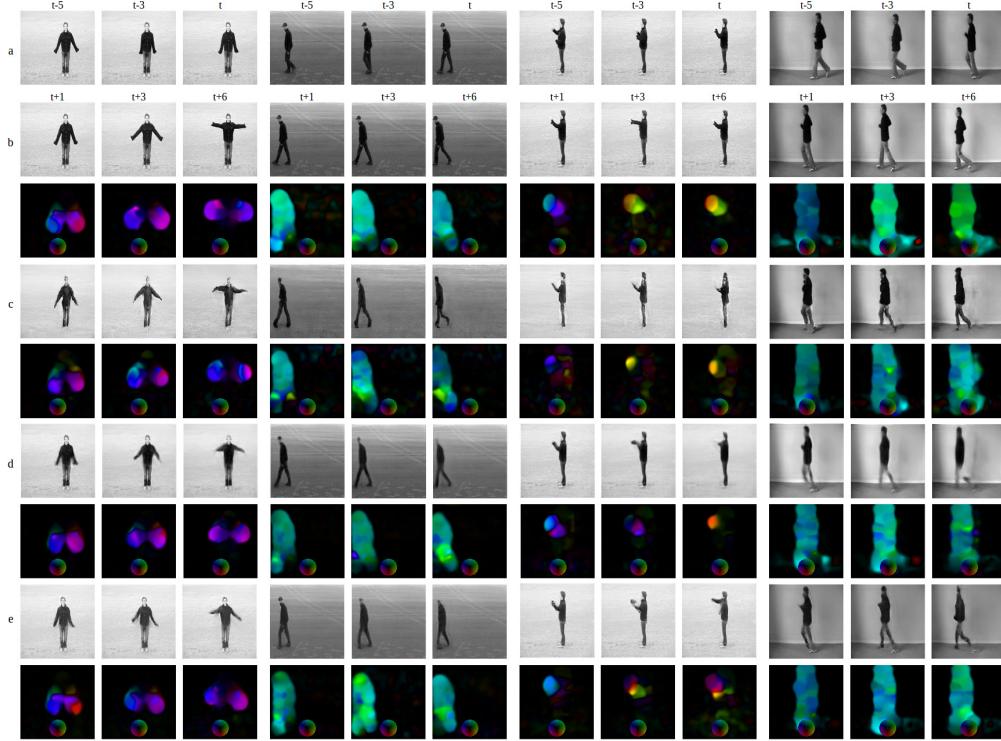


Figure A.2: Prediction results for the KTH Action test split. a: Input, b: Ground Truth, c: FutureGAN (ours), d: fRNN [31], e: MCNet [40].

### A.3 KTHAction 120 step Long-Term Predictions

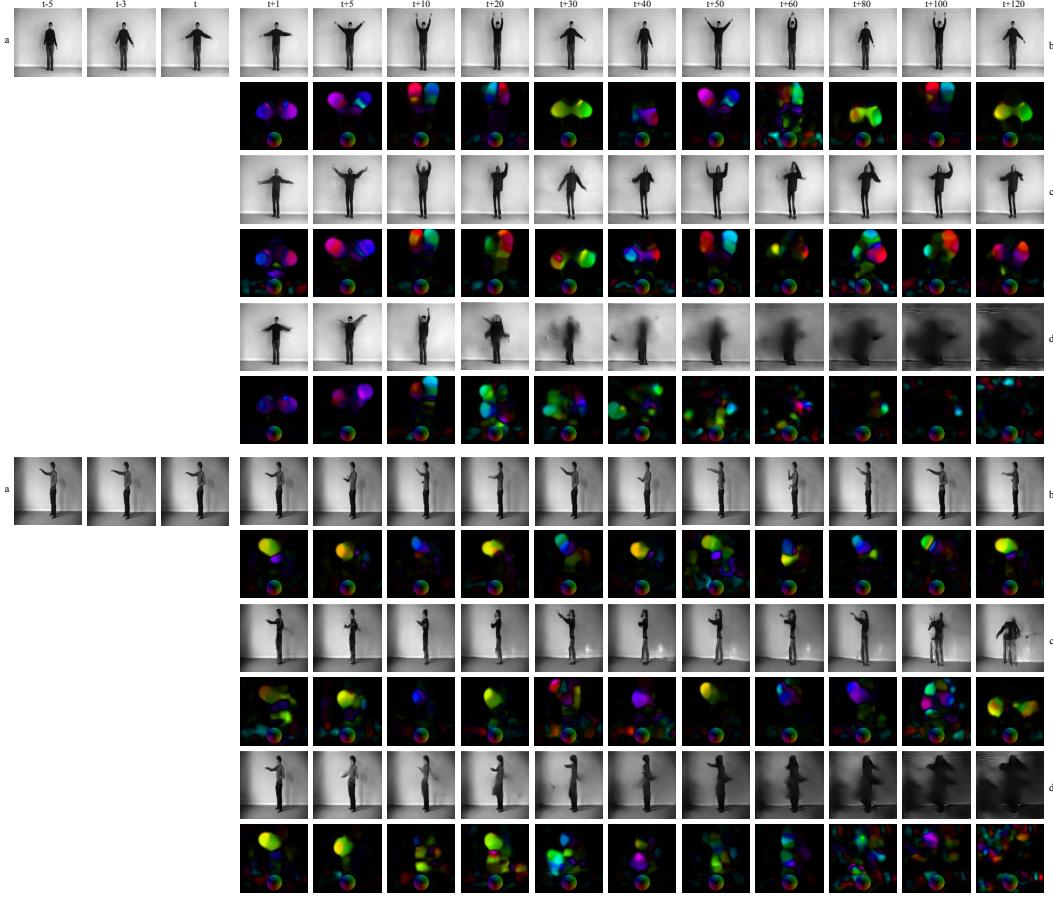


Figure A.3: Prediction results of the 120 step long-term predictions for all three datasets. a: Input, b: Ground Truth, c: FutureGAN (ours), d: MCNet [40].

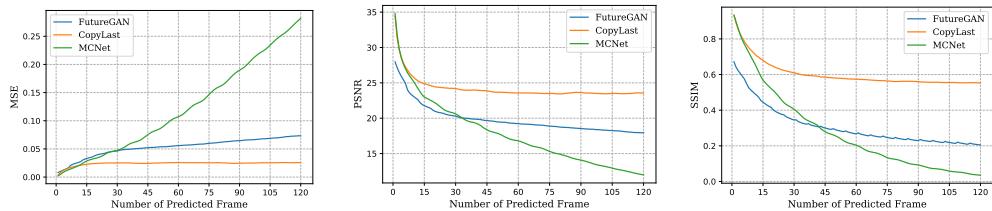


Figure A.4: Quantitative results per predicted frame for the full KTH Action test split.

#### A.4 Cityscapes

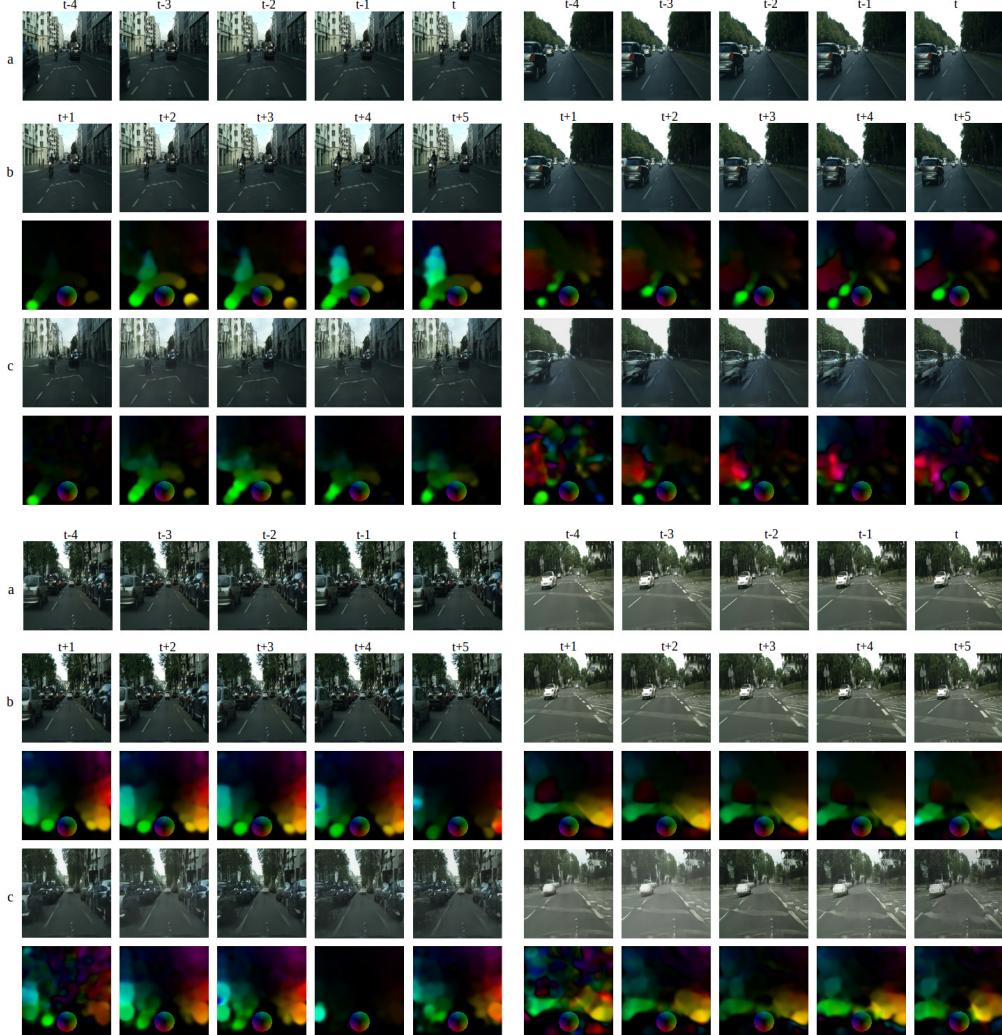


Figure A.5: Prediction results for the Cityscapes test sequences. a: Input, b: Ground Truth, c: FutureGAN (ours)

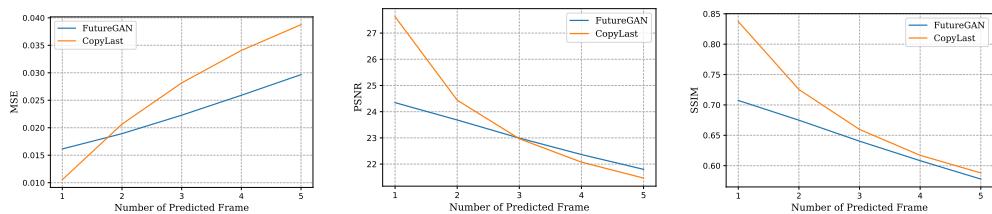
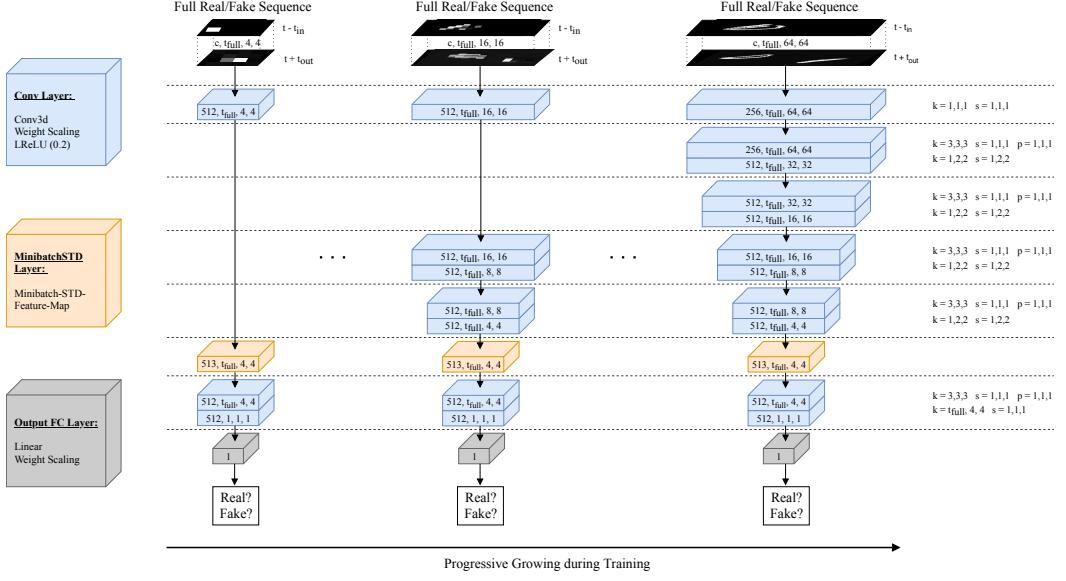


Figure A.6: Quantitative results per predicted frame for the full Cityscapes test split.

## B Discriminator Structure



**Figure A.7: FutureGAN discriminator during training.** We initialize our model to take a set of  $4 \times 4$  px resolution frames and output frames of the same resolution. During training, layers are added progressively to double the resolution after a specified number of iterations. The resolution of the input frames always matches the resolution of the current state of the network. This figure illustrates the growth progress of the discriminator exemplary for the MovingMNIST dataset with a final resolution of  $64 \times 64$  px. Intermediate  $8 \times 8$  px and  $32 \times 32$  px resolution steps are left out for visual clarity.