Improving State-of-the-Art GAN Video Synthesis

Michael Gentner
TU Munich
85748 Garching
michael.gentner@tum.de

Lars Carius
TU Munich
85748 Garching
lars.carius@tum.de

Abstract

We propose multiple methods for improving state-of-the-art GAN-based video synthesis approaches. We show that GANs using 3D-convolutions for video generation can easily be extended to predicting coherent depth maps alongside RGB frames, but results indicate that RGB accuracy does not improve if depth is available. We further propose critic-correction, a method for improving videos generated by latent space curve fitting. Additionally, we study the effect of Principal Component Analysis as well as different backprojection methods on the quality of generated videos. Our code can be found at `https://github.com/CariusLars/ImprovingVideoGeneration`.

1. Introduction

Synthesizing high-resolution videos is a highly relevant topic for commercial applications of Generative Adversarial Networks (GANs). The technique can be used for creating animations in games and, at some point, photo-realistic video clips. With the possibility to condition the generated videos on input data, GANs have the potential to outperform classic approaches of video editing and special effects.

2. Related Work

Generative Adversarial Networks have been shown to generate high-resolution images that are indistinguishable from real photographs to the human eye [7, 6]. The domain of synthesizing coherent high-resolution videos of considerable length remains an open challenge.
BigGAN [2], a large architecture for image synthesis, was extended to video generation by Clark *et al.* using Recurrent Neural Network (RNN) blocks and separate discriminators for spatial and temporal consistency [4]. Building on Progressively Growing GANs (PGGANs) [6], Aigner and Körner used 3D-convolutions to synthesize video sequences conditioned on input frames [1]. Other approaches include explicit modeling of back- and foreground dynamics to achieve spatio-temporal consistency [10]. The approaches do not consider using depth information.

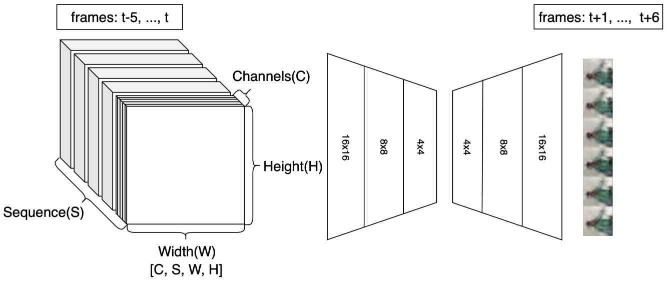


Figure 1: 3D-convolutional video generation pipeline adapted from [1]. 6 RGB/RGB-D frames serve as conditioning input, the network generates the consecutive 6 RGB/RGB-D frames. Every 20 epochs, encoder and decoder are grown to learn features on all resolutions (see [6])

In other research, the low-dimensional latent manifold of GANs has been widely used to interpolate between generated images [6]. Zhu *et al.* use encoded spatial properties to morph images based on user suggestions [11], Chen *et al.* [3] propose additional methods to manipulate specific properties of generated faces.

3. Method

We propose multiple approaches to improve the state-of-the-art of video synthesis:

- Use of depth information to improve the quality of generated videos
- Leveraging curve-fitting on the latent space
- Improving synthesized videos with critic-correction and Principal Component Analysis (PCA)

3.1. Incorporating depth

Due to its promising results, the FutureGAN architecture [1] served as a basis for this approach. We incorporated depth by adding a feature channel containing depth maps normalized to $[-1, 1]$ (see figure 1). From the dense RGB-D dataset generated by Nießner *et al.*, we extracted video snippets of 12 frames each from which the first 6 served as

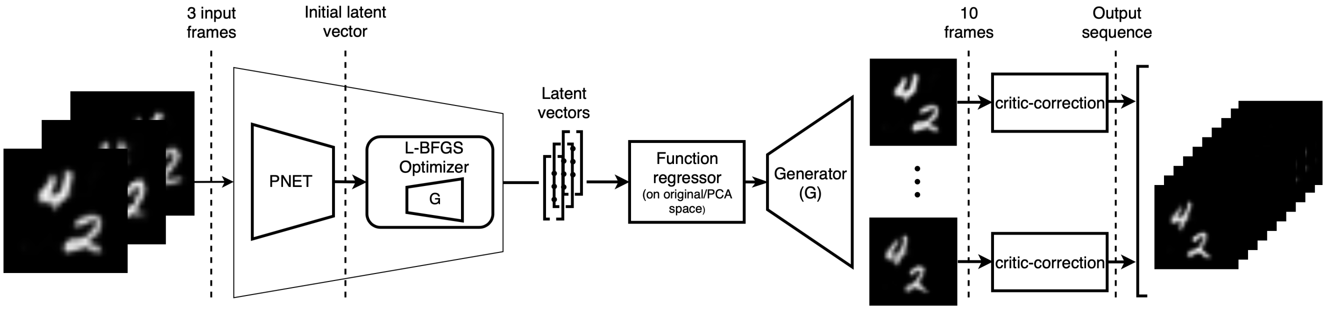


Figure 2: Latent space video generation pipeline. Conditioning frames are projected into latent space using a numerical optimizer initialized with a projection network. A polynomial is then fitted to the latent vectors (if the latent space is very large, PCA before regression and backprojection afterwards proved to be useful) and new points are sampled from the curve to generate the output video in latent space. The generator translates the video into image space, critic-correction improves the video in post-processing

conditioning input frames and the remaining 6 as ground truth for the generated video. We used 40% of frames per sequence as test set. We limited the progressive growing of the network to a maximum resolution of 16×16 px, trained for 20 epochs in each resolution step and adapted the batch size after each resolution change to best utilize the available GPU memory.
We tested the hypothesis of depth information improving the synthesized video quality by allowing to learn spatio-temporal correlations in an ablation study.

3.2. Latent Space Video Generation

We performed video synthesis on MovingMNIST by fitting curves to the latent manifold of a GAN. We propose critic-correction as a postprocessing method for the generated video. Conditioning on input frames is incorporated by backprojecting given images into latent space utilizing optimization techniques and a projection network, the entire pipeline is visualized in figure 2.

3.2.1 GAN-Architecture

We adapted the DCGAN architecture of [8] to incorporate the WGAN-GP loss of [5]. This remedies the diligent training of standard GANs. We used instance normalization [9] to render the network agnostic to high contrast images found in MovingMNIST.

3.2.2 Backprojection

The projection of input frames into latent space is necessary to achieve conditioning of the generated videos. Inspired by Zhu *et al.*, we used a combination of L-BFGS optimization and a projection network [11]. We fed an input image $\mathbf{x}$ into our projection network P (P-net) which was trained to minimize the binary cross-entropy loss between $\mathbf{x}$ and the image generated from the predicted latent vector $\mathbf{z}$. The weights of the generator network $G(P(\mathbf{x}))$ were frozen and P closely resembles the structure of the critic used in the initial GAN training. The latent vector generated by P is used as an initialization for the numerical solver.

3.2.3 Critic-Correction

Due to imperfect manifolds and backprojections, generated frames were often close to the desired image, yet still blurry. We propose the new technique critic-correction to compensate for this. The critic determines the realism of images generated from z, so we used it to optimize our obtained latent vectors. Fixing the parameters of critic C and generator G and using stochastic gradient descent to optimize the latent vector z, the update rule is given by:

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \alpha \nabla_{\mathbf{z}} \left[-\mathbb{E}[C(G(\mathbf{z}_i))] \right] \quad (1)$$

3.2.4 Regression in Latent Space

The projections of input images into the latent space serve as conditioning input. Parametrized by time, polynomials are fitted to the individual dimensions of the samples in latent space. The curves are evaluated at new points in time to synthesize a latent video sample. We assumed that certain properties correspond to certain dimensions in the manifold. Therefore, we applied PCA on the input projections, fitted a curve in the principal component space and applied the inverse projection to the output to reconstruct the original space. We compared these results to approaches without PCA.

4. Results

We conducted separate experiments for both approaches. We summarize our findings below, additional samples can be found in our repository.

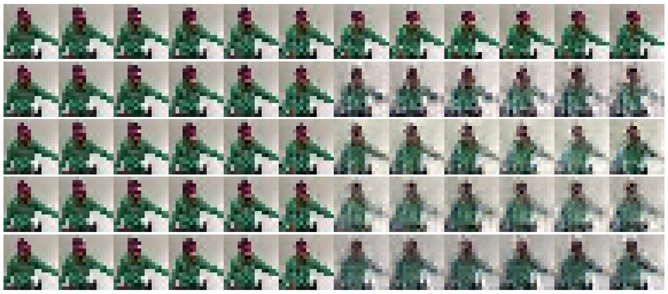
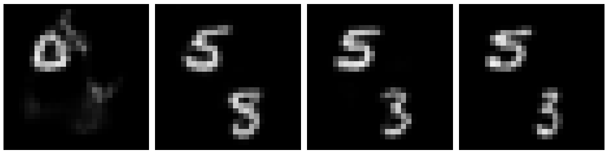


Figure 3: Test set sequence from the ablation study. Each row represents a video sequence with 6 input frames (conditioning) and 6 frames predicted by the network. From top to bottom: ground truth, RGB→RGB, RGB-D→RGB, RGB→RGB-D, RGB-D→RGB-D



(a) L-BFGS (b) P-net (c) Both (d) Ground Truth

Figure 4: Comparison of backprojection methods: a) L-BFGS-optimization b) Projection learned by convolutional neural network c) L-BFGS with initialization by network b d) ground truth. The projections produce latent vectors, for visualization they are fed into the generator to obtain images

4.1. Extending FutureGAN

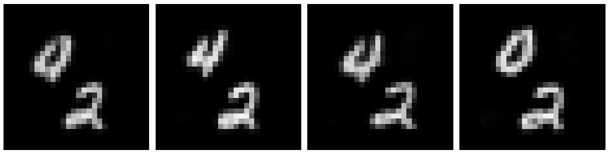
	RGB→RGB	RGB-D→RGB	RGB→RGB-D	RGB-D→RGB-D
MSE	**0.0656**	0.0686	0.0765	0.0749
PSNR	**18.2980**	18.1068	12.7094	17.7737
SSIM	**0.6452**	0.6321	0.6290	0.6188
Depth MSE	-	-	0.0669	**0.0582**
Depth PSNR	-	-	20.2700	**20.9937**

Table 1: Scores of the different methods to incorporate depth information into the FutureGAN architecture. The networks predict 6 consecutive frames based on 6 conditioning frames

Figure 3 shows an example sequence analysed in the ablation study, table 1 summarizes the performance of tested architectures. While the baseline performed best on the RGB channels, low test set errors were achieved on the predicted depth maps in the corresponding approaches. On the one hand, this supports the fact that depth information is not used by the network to infer information about the RGB predictions. On the other hand, the findings show that depth predictions can be learned just as easily as RGB channels. The small increase in RGB error in approaches generating depth shows that additionally learning depth predictions comes at very little cost and can, within given constraints of a suitable training set, even be accomplished without depth input being available at test time.

4.2. Latent Space Video Generation

While the backprojection of test set images into the latent space using L-BFGS-optimization worked well for large latent space dimensions (128 or 256), it performed poorly with reasonable latent space sizes (16). Training a convolutional network for the backprojection task yielded better results, combining both approaches by initializing the L-BFGS-solver with the network outperformed both individual methods significantly (see figure 4). Additional experiments using different numerical solvers did not improve performance.



(a) Before corr. (b) After corr. (c) Before corr. (d) After corr.

Figure 5: Critic correction applied to frames generated by the video synthesis pipeline. Image a) is successfully corrected to image b), the algorithm fails on image c) and results in a digit error displayed in d). The ground truth contains the numbers 4 and 2

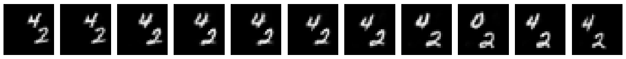


Figure 6: Synthesized video sequence based on 3 conditioning input frames. Regression to a linear curve and critic correction were used for the generation process

Our proposed method of critic-correction produced mixed results. As figure 5 shows, the correction leads to significantly sharper output images and often converges to the correct numbers. In some cases, difficult to interpret numbers converged to wrong digits. In future works, this could be compensated with additional information transfer from surrounding frames (in the time dimension).

On the simple MovingMNIST dataset, videos generated by interpolating the backprojected conditioning frames in latent space led to reasonably coherent output sequences (Figure 6). Regression to a polynomial of rank 1 produced the best results with the fewest digit errors, critic correction reduced blurriness significantly but led to occasional mistakes in the digits. PCA before regression hardly changed the results for a 16-dimensional latent space but did significantly improve the video quality for a 256-dimensional latent space. This indicates that it helps to alleviate the problem of poorly-conditioned latent spaces in video generation tasks.

5. Conclusion

Our experiments indicate that with the given FutureGAN architecture, depth information does not improve RGB video prediction performance. Predicting depth maps, however, comes at reasonable accuracy with very low cost, which is valuable especially for robotics tasks.

Our contributions to video generation by leveraging smooth latent spaces include multiple methods to compensate for instabilities and noise in the regression of latent vectors and confirm the results of previous papers. In a set of constraint applications, both critic-correction and applying PCA before regression improved the synthesized videos.

References

[1] S. Aigner and M. Körner. Futuregan: Anticipating the Future Frames of Video Sequences Using Spatio-Temporal 3D Convolutions in Progressively Growing Gans. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W16:3–11, 2019.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. pages 1–35, 2018.

[3] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic Latent Space Interpolation for Unpaired Image-to-image Translation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2408–2416, 2019.

[4] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial Video Generation on Complex Datasets. pages 1–21, 2019.

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. *Advances in Neural Information Processing Systems*, 2017-Decem:5768–5778, 2017.

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–26, 2018.

[7] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. 2018.

[8] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–16, 2016.

[9] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. (2016), 2016.

[10] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in Neural Information Processing Systems*, (Nips):613–621, 2016.

[11] Jun Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9909 LNCS:597–613, 2016.