Improving State-of-the-Art GAN Video Synthesis

Michael Gentner Lars Carius

Advanced Deep Learning for Computer Vision (IN2364)

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 this does not improve RGB accuracy 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.

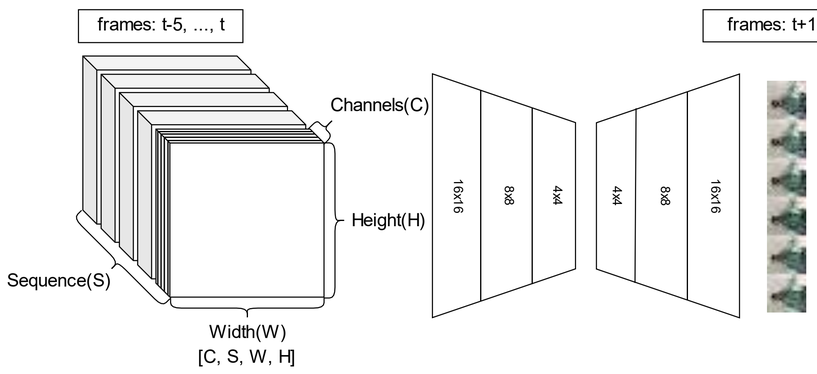
GitHub Repo

Related Work

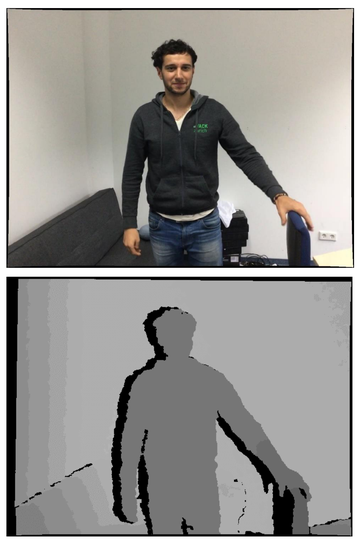
BigGAN was extended to video generation by Clark *et al.* using Recurrent Neural Network (RNN) blocks and separate discriminators for spatial and temporal consistency [1]. Building on PGGAN [2], FutureGAN uses 3D-convolutions to synthesize video sequences conditioned on input frames [3].

The low-dimensional latent manifold of GANs was used to interpolate between generated images [4]. Zhu *et al.* use encoded spatial properties to morph images based on user suggestions [5], Chen *et al.* [6] propose additional methods to manipulate specific properties of generated faces.

Incorporating Depth



Progressively Growing 3D-convolution architecture for video synthesis



Deform Dataset

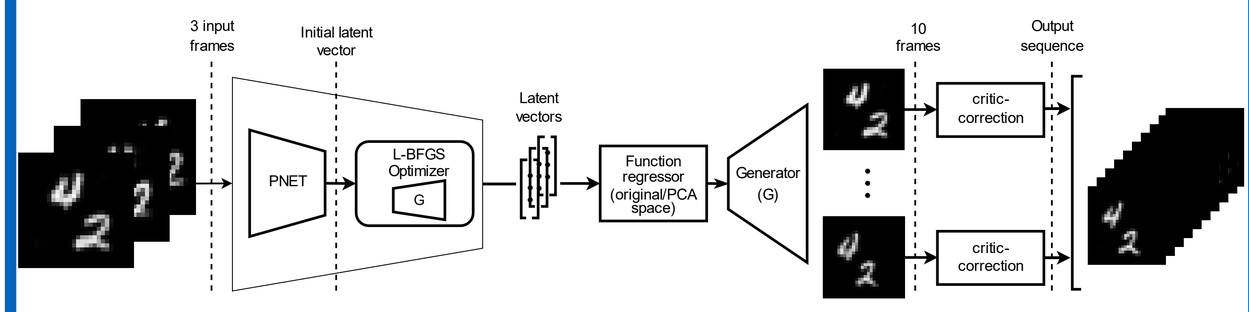
Key Facts

- 6 input frames, 6 output frames
- 3D-convolutions, WGAN-GP loss
- Progressive growing every 20 episodes (transition & stabilization phases)
- Trained on DEFORM dataset

Ablation Study

- Dense depth maps as an additional channel
- All combinations of depth input & output

Latent Space Video Generation (LSVG)

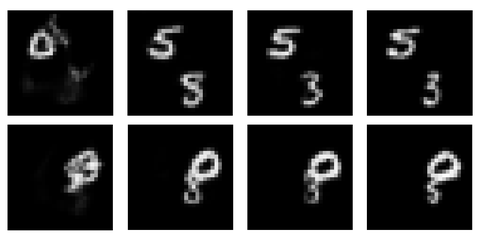


Pipeline for generating videos via backprojection of conditioning frames, curve-fitting in latent space and generation of new samples including post-processing with critic-correction

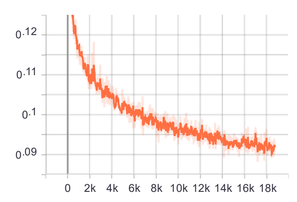
Key Facts

- Arbitrary number of input frames
- Backprojection using a hybrid approach (Projection-net + L-BFGS optimization)
- Polynomial fitting in latent space (16 dim.)
- GAN generator to synthesize an arbitrary number of new frames
- Critic-correction as a postprocessing step

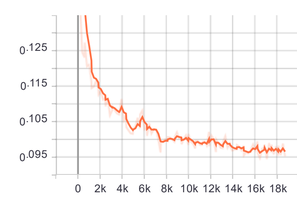
Backprojection



L-BFGS | **P-Net** | **L-BFGS+ P-Net** | **Ground truth**



P-Net train loss



P-Net validation loss

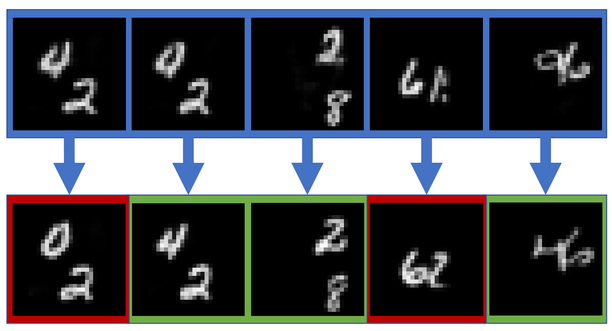
Numerical Solver

- L-BFGS worked best
- Minimizes reconstruction error with fixed generator weights

Projection-Net

- Similar structure as critic, BCE loss
- Trained on MovingMNIST
- Used to initialize numerical solver

Critic-Correction



Backprojected video frames before and after critic-correction. The frames become considerably sharper and more readable, but not all digits converge to the correct ground truth

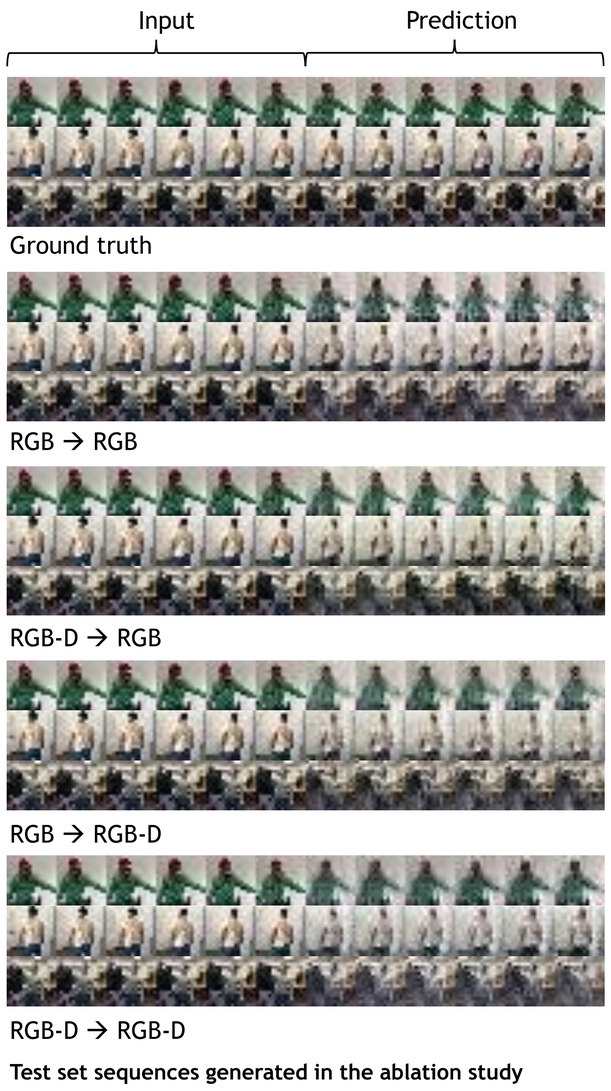
$$\mathbf{z}_{i+1} = \mathbf{z}_i - \alpha \nabla_{\mathbf{z}} \left[-\mathbb{E}\left[C\left(G\left(\mathbf{z}_i \right) \right) \right] \right]$$

Update formula for critic-correction. The latent vector is optimized to achieve a high critic score and thus to produce a realistic image

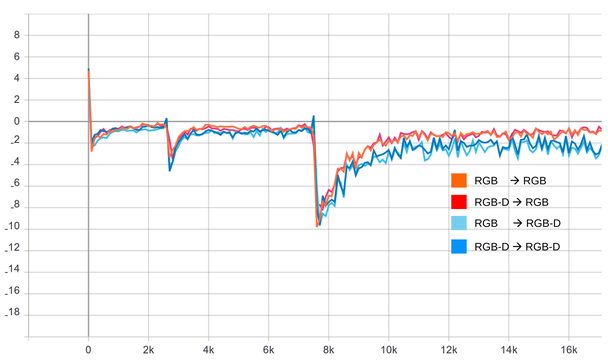
Key Insights

- Critic rates realism of frame
- Method recovers blurry images
- 100 update steps worked best

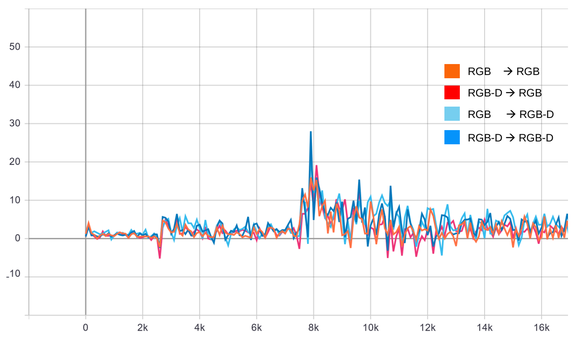
Incorporating Depth - Results



Test set sequences generated in the ablation study



Discriminator loss



Generator loss

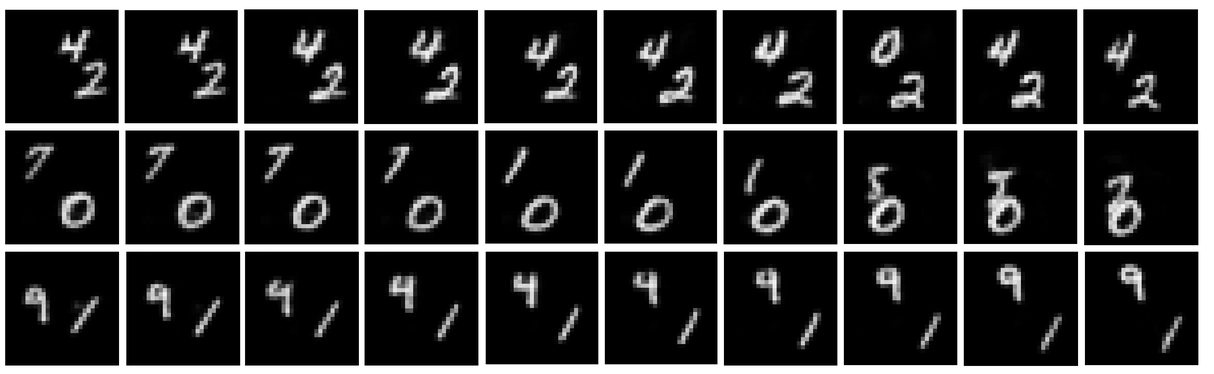
	RGB→RGB	RGB-D→RGB	RGB→RGB-D	RGB-D→RGB-D
MSE	**0.0656**	0.0686	0.0765	0.0749
PSNR	**18.2980**	18.1068	12.7094	17.7737
SSIM	**0.6452**	0.6321	0.6290	0.6188
Depth MSE	-	-	0.0669	**0.0582**
Depth PSNR	-	-	20.2700	**20.9937**

Scores from the ablation study

Key Insights

- Depth information does not improve RGB prediction
- Network does not draw a connection between RGB and D-channel
- Learning depth works well with little extra cost
- Predicting depth works well without depth availability at test time (in given constraints)

LSVG - Results



Video sequences generated by our pipeline. 3 frames are used as input, backprojected into latent space using P-net and L-BFGS. A linear curve is fit via regression, 10 points are sampled and fed into the generator network and critic-correction is applied as a postprocessing step

Key Insights

- Video generation by latent space curve-fitting is possible but requires very good latent space
- Hybrid backprojection performs best
- Low-rank polynomials perform better for curve-fitting
- PCA can help overcome ill-conditioned latent spaces
- Critic-correction can improve image quality but might need additional information

References

[1] Adversarial Video Generation on Complex Datasets. *Clark et al., 2019.*
[2] Progressive growing of GANs for improved quality, stability, and variation. *Karras et al., ICLR 2018.*
[3] Futuregan: Anticipating the Future Frames of Video Sequences Using Spatio-Temporal 3D Convolutions in Progressively Growing Gans. *Aigner et al., ISPRS 2019.*
[4] Progressive growing of GANs for improved quality, stability, and variation. *Karras et al., ICLR 2018.*
[5] Generative visual manipulation on the natural image manifold. Zhu *et al., Lecture Notes in Computer Science 2016.*
[6] Homomorphic Latent Space Interpolation for Unpaired Image-to-image Translation. *Chen et al., CVPR 2019.*