Video Synthesis from a Single Image

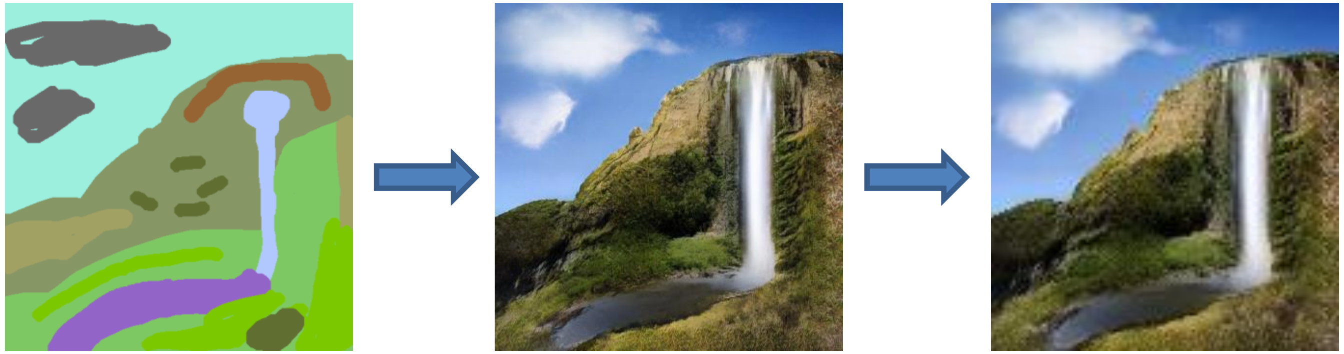
Patrick Radner

Problem Statement

Generate a short, realistic video given a single image.

Motivation

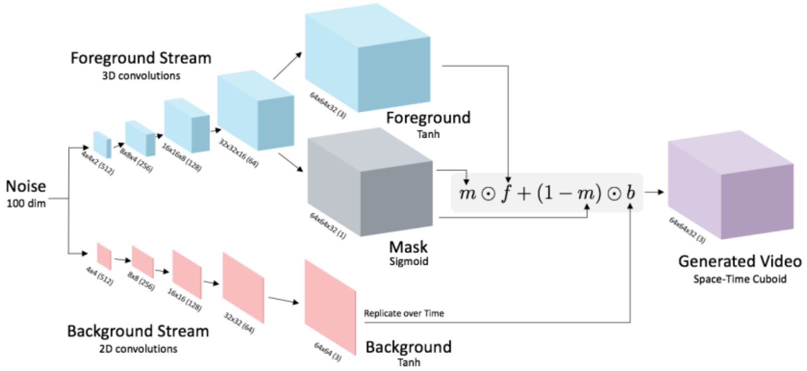
- Image GANs
- “Bringing Landscape Images to Life”



Some results: mirjang.github.io/mt_videosynthesis

Related Work: VGAN, TGAN, MoCoGAN

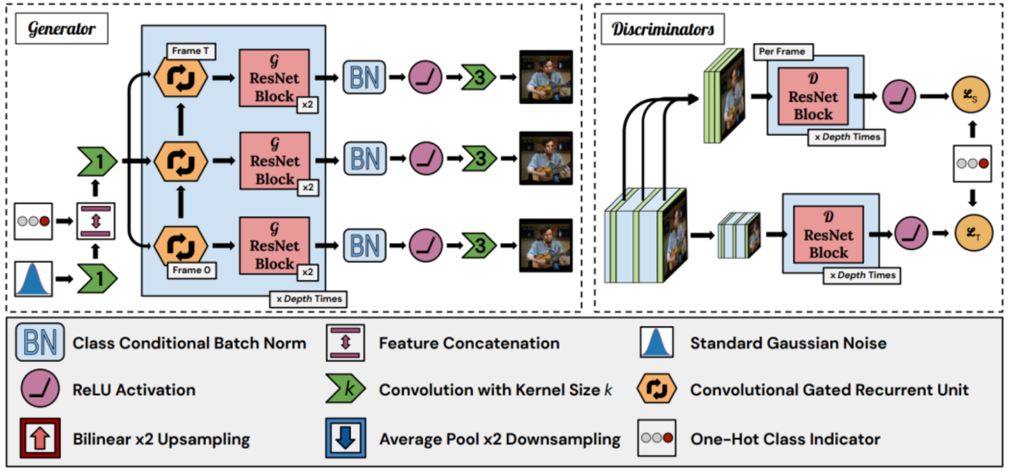
- VGAN (2016)

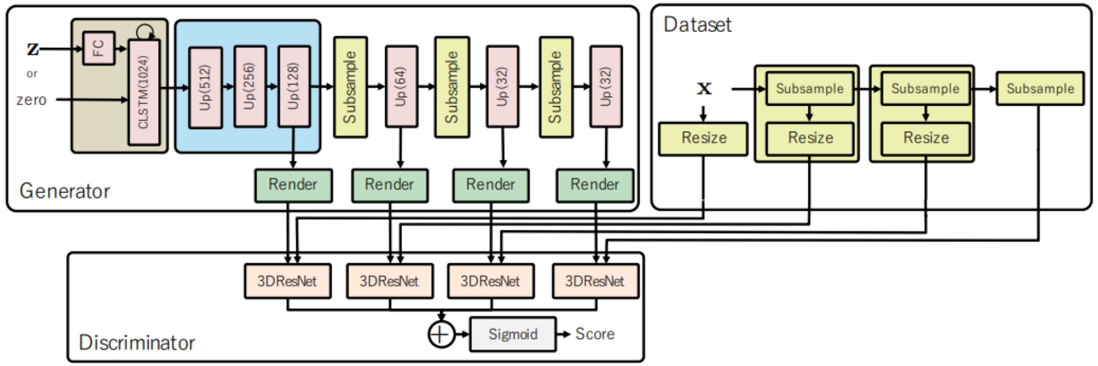


- TGAN (2017)
  - Temporal generator
- MoCoGAN (2017)
  - Discriminator decomposition

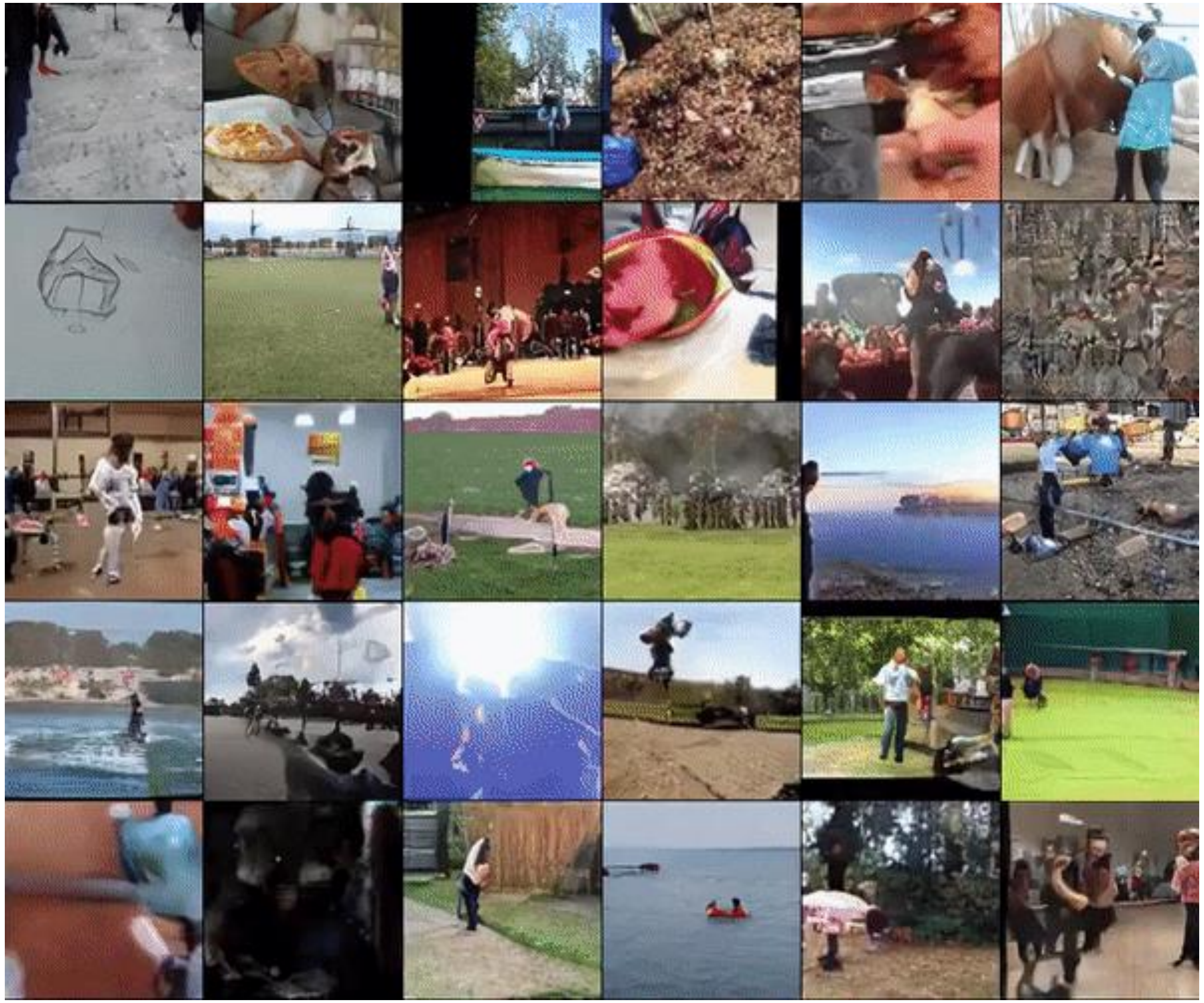
Related Work: State of the Art

- DVDGAN (2019)
  - Feature pyramid
  - “BigGAN for videos”
- TriVDGAN (2020)
  - TSRU
- TGANv2 (2020)
  - Subsampling for efficiency
- Latent Video Transformer





Samples DVDGAN and TGANv2



Method

- “DVDGAN w/ Style blocks”
- No BatchNorm

  →Small batch size
- 128x128 resolution
- 65M parameters
  - 20M for Generator

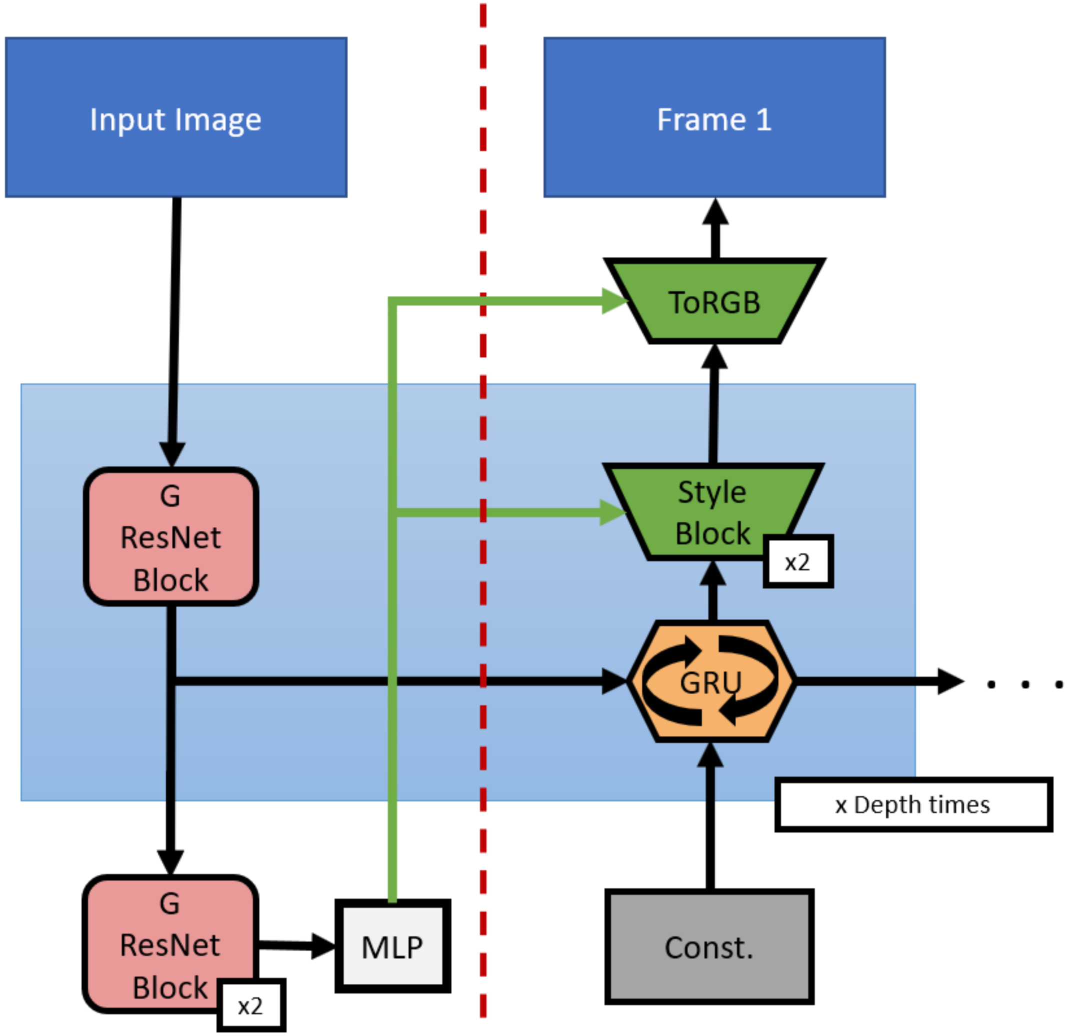


Image Generation - BigGAN

- ResNet-like blocks
- Large batch size
- Best practices for GANs

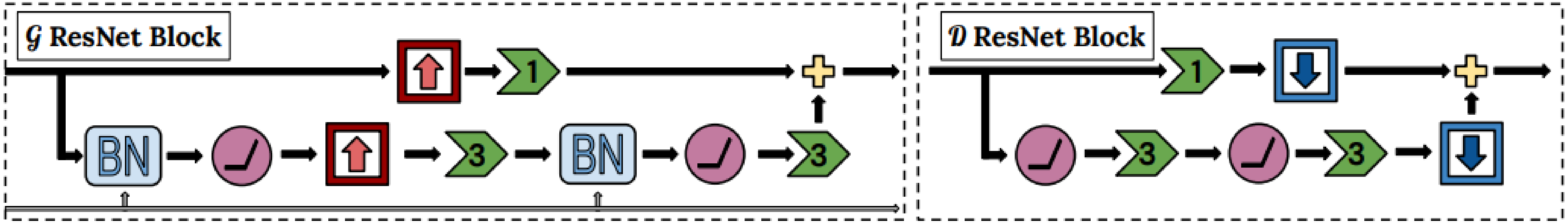
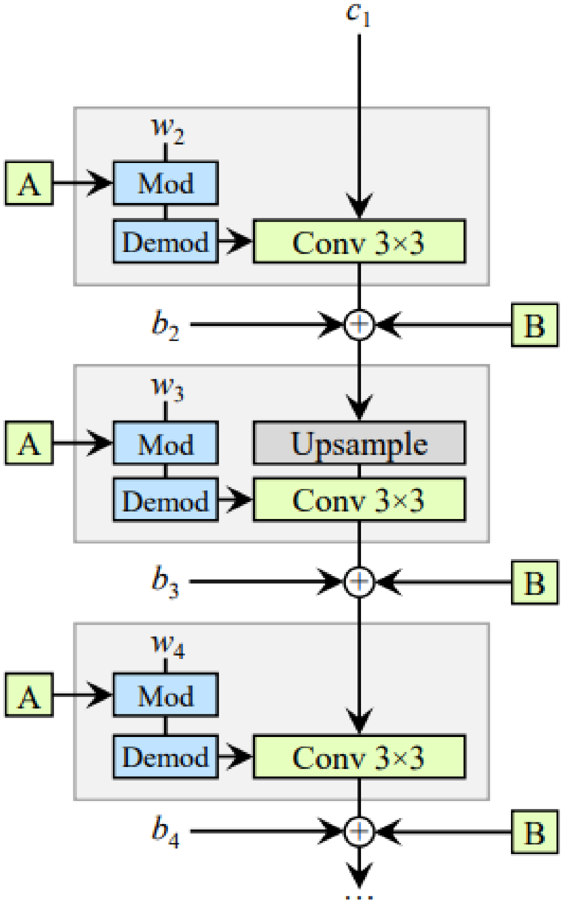


Image Generation – StyleGAN2

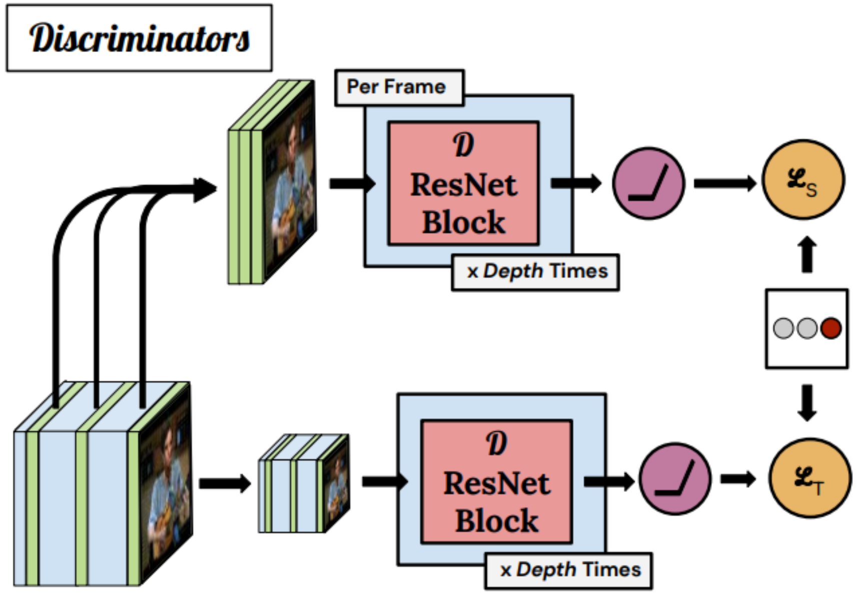
- Style vector
- No BatchNorm!

$$w'_{ijk} = s_i \cdot w_{ijk}$$

$$w''_{ijk} = w'_{ijk} \Big/ \sqrt{\sum_{i,k} {w'_{ijk}}^2 + \epsilon}$$



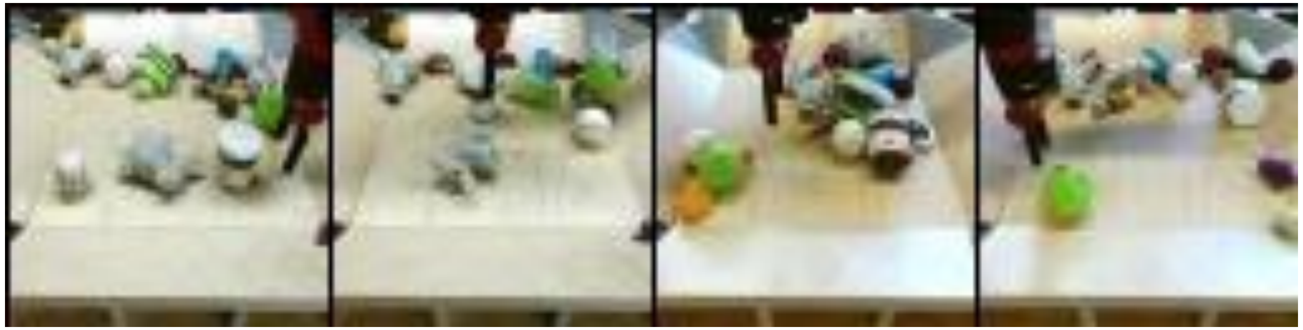
Discriminators



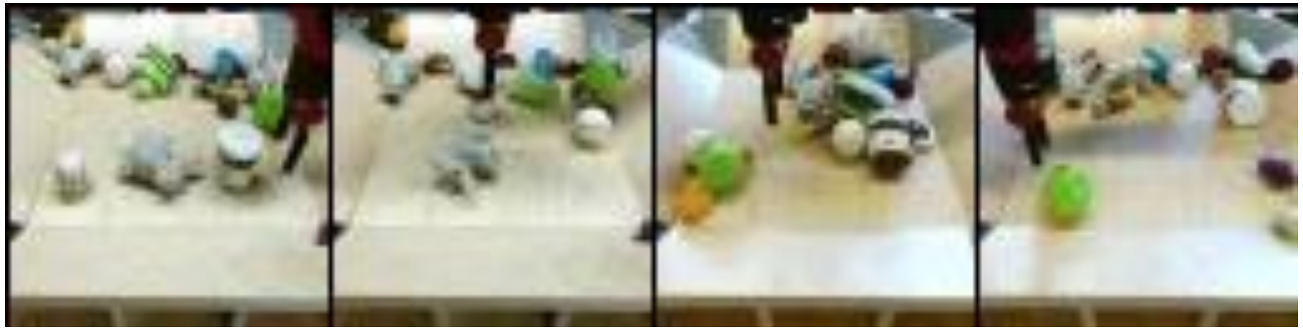
- Decomposition into $\mathcal{D}_S$ and $\mathcal{D}_T$
- $\#pixels = K \times H \times W + T \times \frac{H}{\phi_H} \times \frac{W}{\phi_W}$
- WGAN-GP

Results: BAIR dataset

- 64x64 videos
- Static camera
- Little diversity



ours



DVDGAN-S

Bringing Landscape Images to Life

- Sky Time-lapse dataset
  - Ca. 1000 long videos
  - 2400 clips

- Custom Dataset:
  - ca. 500 YouTube videos
  - ca. 7500 clips
  - Duplicates

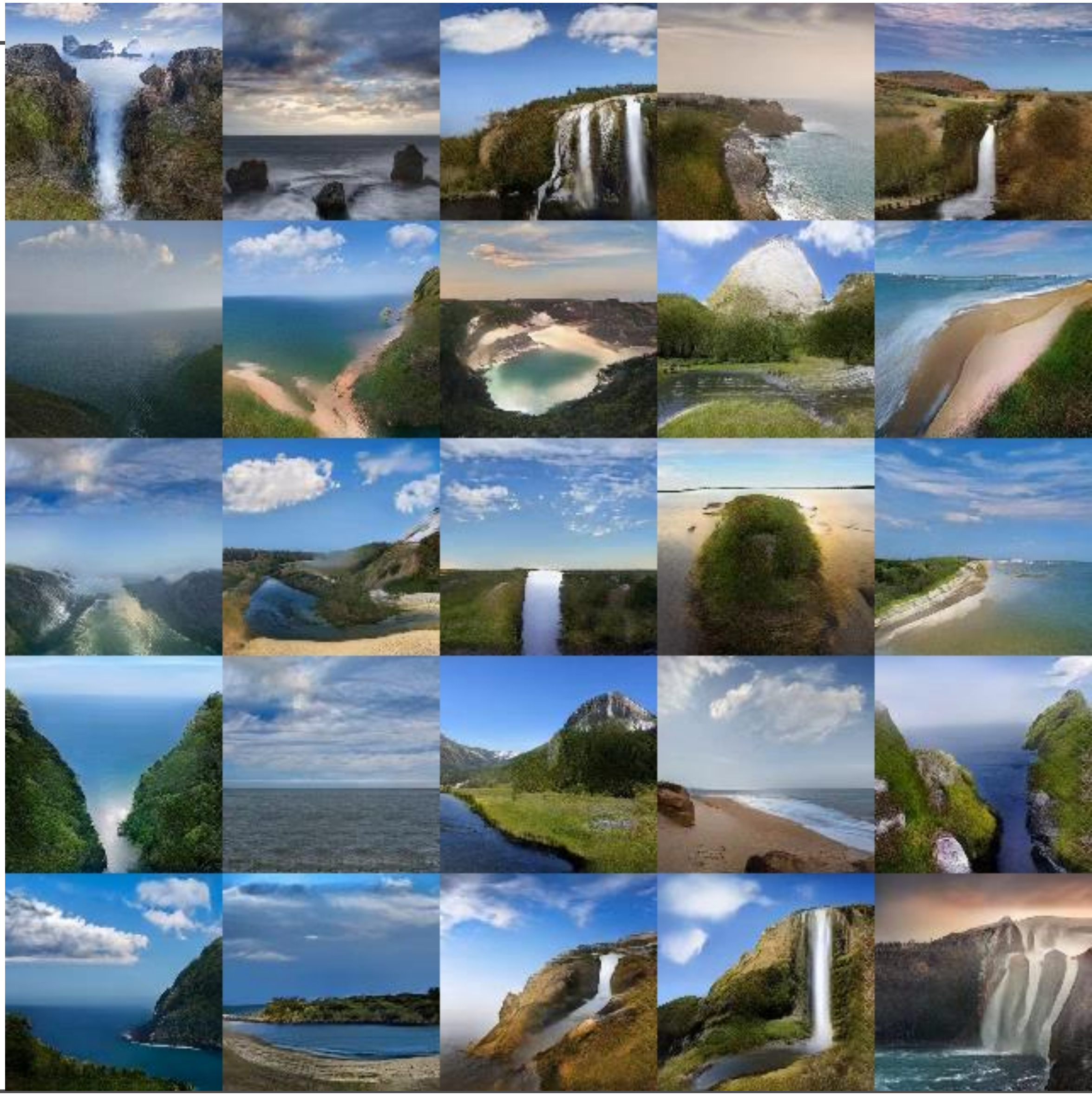
Quantitative Evaluation

	BAIR		Custom Dataset	
	IS (↑)	FID(↓)	IS(↑)	FID(↓)
DVDGAN-S	10.68	81.02	**29.27**	194.30
Ours	**14.68**	**41.47**	13.07	**108.96**

- DVDGAN-S:
  - approx. same # of params as ours
  - Batch size 128

GauGAN videos

- GauGAN
  - Image2Image translation
  - Easy to use demo
- Domain gap
  - Dataset mostly close-up



Discussion

- Autoregressive Models
- TrajGRU
- Stochasticity
- Global effects

- General model for video prediction
- Landscape videos
  - Better specialized solutions

General Problems

- Resource intensive
- No unified dataset
- No unified metrics
- Lots of hyperparameters
  - Input: k-Frames, 1 Frame, unconditional, class-cond.
  - Output: length, resolution

Questions

Sources

- BigGAN
  - A. Brock, J. Donahue, and K. Simonyan.Large Scale GAN Training for High FidelityNatural Image Synthesis. 2019. arXiv:1809.11096 [cs.LG]
- StyleGAN
  - T. Karras, S. Laine, and T. Aila.A Style-Based Generator Architecture for GenerativeAdversarial Networks. 2019. arXiv:1812.04948 [cs.NE]
  - T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila.Analyzing andImproving the Image Quality of StyleGAN. 2020. arXiv:1912.04958 [cs.CV]
- DVDGAN
  - A. Clark, J. Donahue, and K. Simonyan.Adversarial Video Generation on ComplexDatasets. 2019. arXiv:1907.06571 [cs.CV]
- GAUGAN
  - T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu.Semantic Image Synthesis with Spatially-Adaptive Normalization. 2019. arXiv:1903.07291 [cs.CV].
  - nvidia.com/en-us/research/ai-playground