ICLR

A Good Image Generator Is What You Need For High-Resolution Video Synthesis

Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng,
Dimitris N. Metaxas, Sergey Tulyakov

Snap Inc.

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

UNIVERSITY OF DELAWARE

Video Synthesis: Background

- Video synthesis: Create *unseen* video clips from random noises

Random noise

Video Synthesis: Difficulties

- Lack of training data.
- Large Models, hard to train.
- High cost for data collection / training.

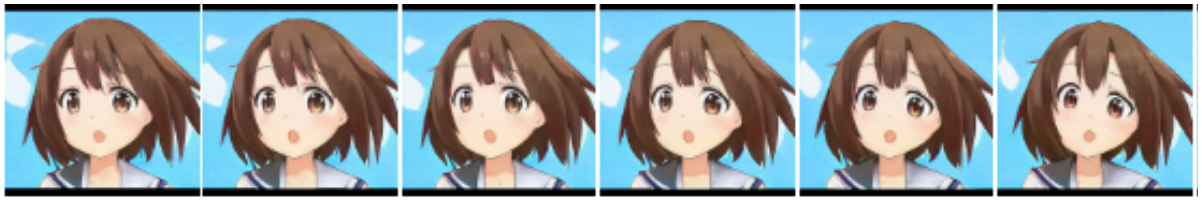
StyleGAN2[1]: 1024 resolution

TGANv2[2]: 256 resolution

[1] Analyzing and Improving the Image Quality of StyleGAN, CVPR 2020
[2]Train Sparsely, Generate Densely: Memory-efficient Unsupervised Training of High-resolution Temporal GAN, IJCV 2020

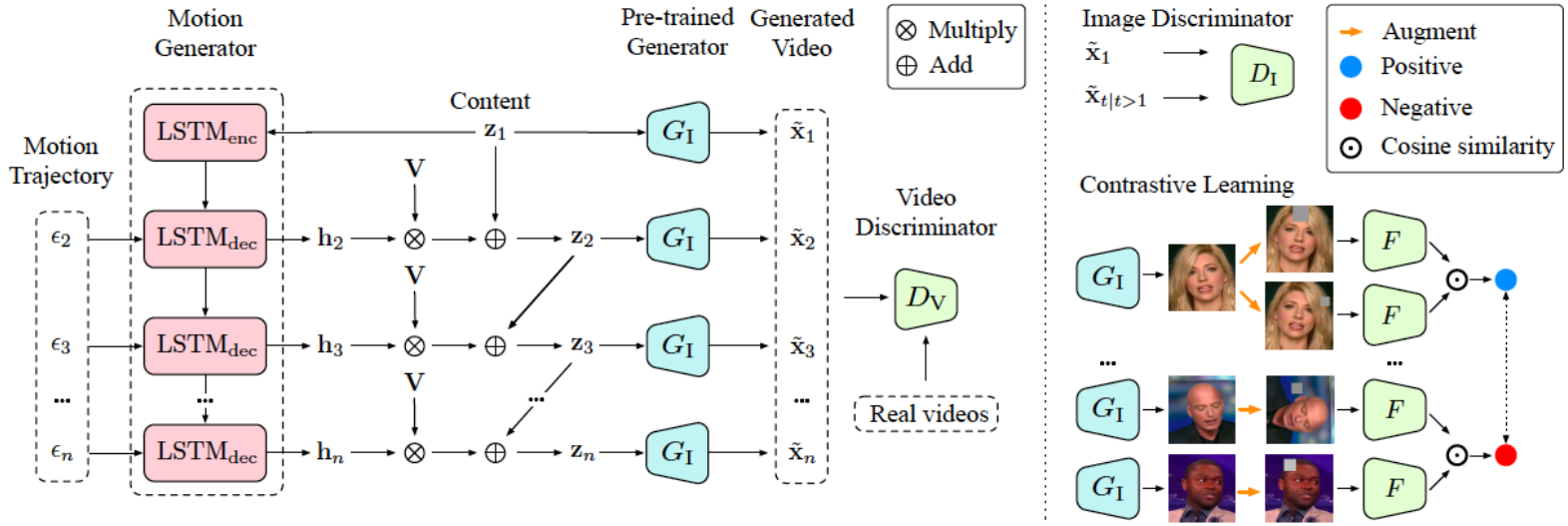
Methods: Intuition

- Reuse <u>pre-trained</u> image generator G_I in video synthesize training.
- Given image G_I: represent a video with a trajectory of random noises.



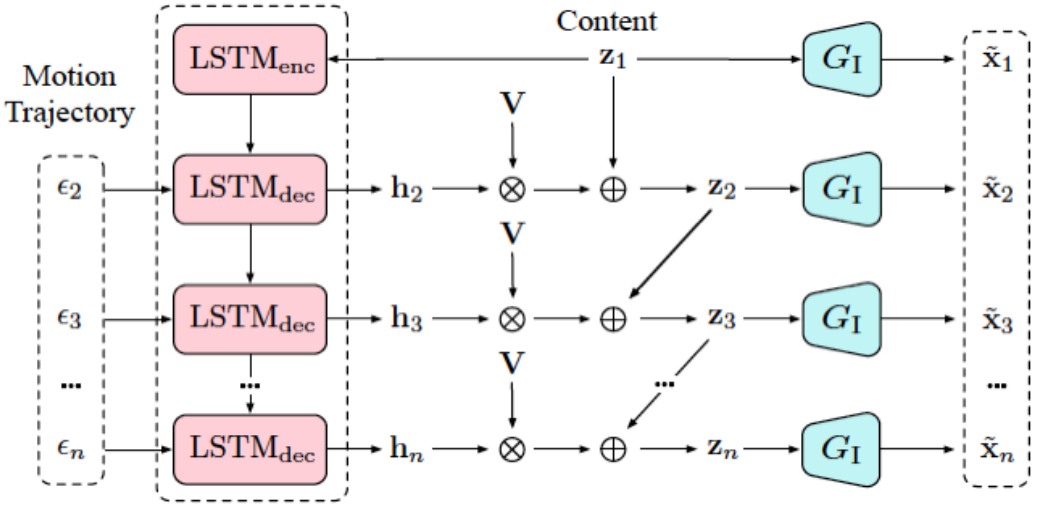
Methods: Framework

Motion generator (find trajectory) + Image generator (pre-trained, fixed) + 2D discriminator (content consistency) + 3D discriminator (motion)



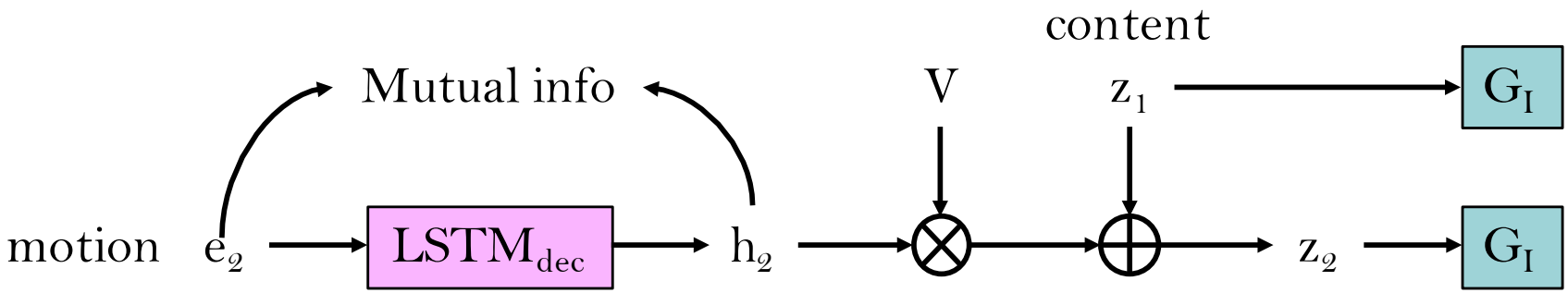
Methods: Motion Generator

- Motion generator: find the trajectory in the latent space of image generator
- LSTM *encoder* + LSTM *decoder*
- Estimate residual of previous frame → content/motion disentanglement



Motion Generator: Improve Diversity

- Mutual information loss: Maximize mutual info between e and h
- e ~ Gaussian: motion randomness
- h: output of LSTM decoder

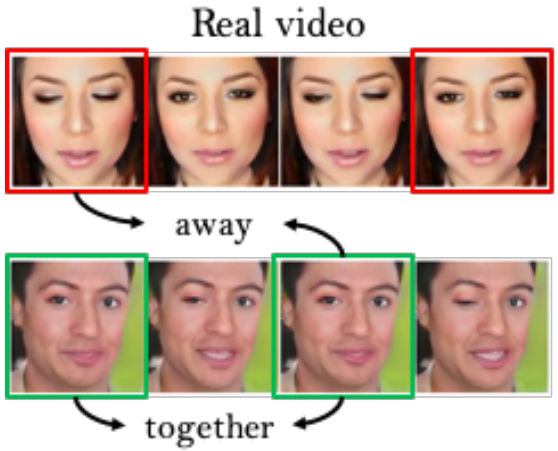


Methods: Image Generator

- Train image generator with video frames: in-domain video synthesis
- Use off-the-shelf image generator: cross-domain video synthesis
  - Cross-domain video synthesis: content from image dataset, motion from video dataset.
  - Save lots of **costs** in data collection: synthesis dog videos with dog images & human videos.

Methods: Contrastive 2D Discriminator

In-domain



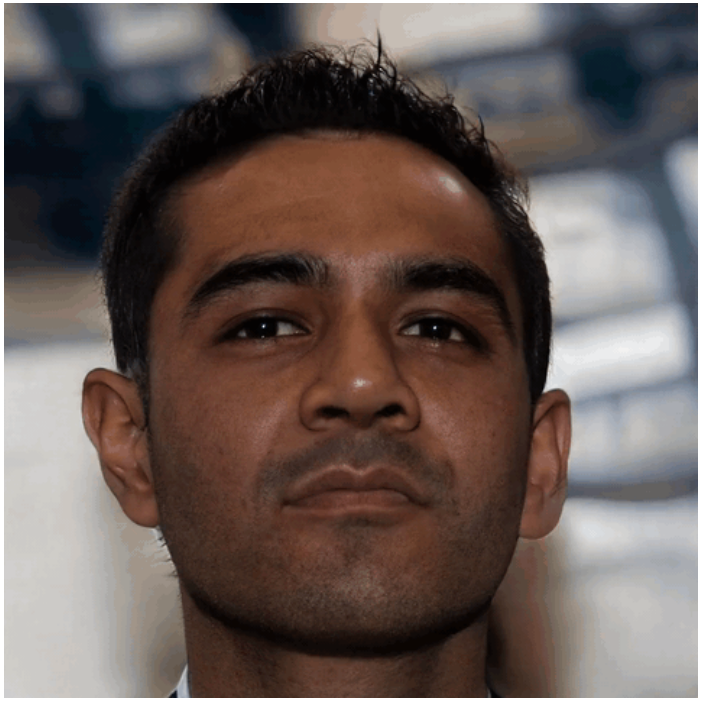
Cross-domain

Framework: Properties

- **High-resolution** video synthesis
- Motion/content **disentanglement**
- **Cross-domain** video synthesis
- **Long** sequence generation
- **Low** computation cost

Properties: High-resolution

Pre-trained image generator: 1024x1024

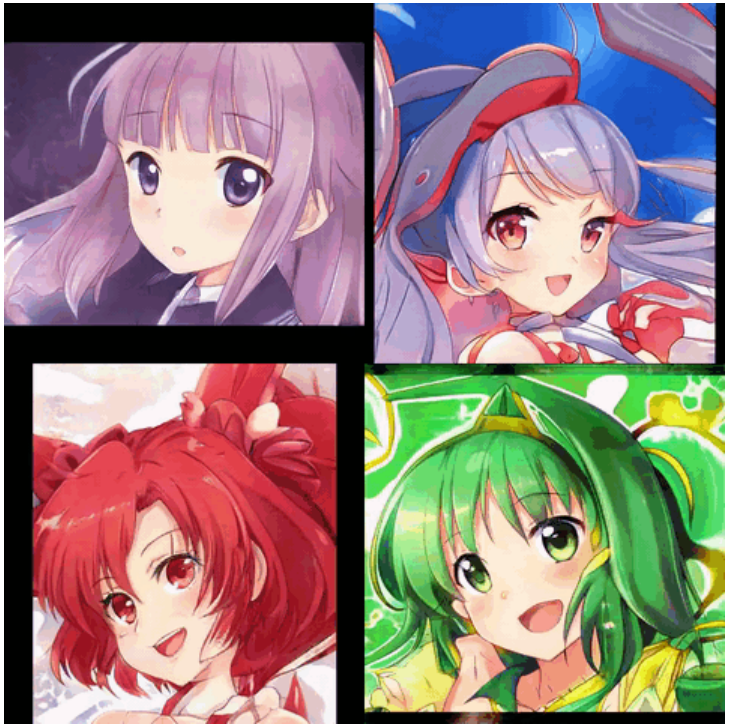


Properties: Motion/content Disentanglement

Residual design in motion generator

Properties: Cross-domain Video Synthesis

512 resolution, human face videos as training data



Properties: Cross-domain Video Synthesis

256 resolution, time-lapse videos as training data

Properties: Long Sequence Generation

(AFHQ, Vox): Interpolation

FaceForensics: LSTM unrolling

Properties: Low Computation Cost

- Pre-trained image generator:
  - **Small** batch-size for video training (as low as 8)
  - **Fixed** image generator: No gradient in video synthesis training

- Our models are trained with GPU (DVDGAN[1]: **TPU** only)
  - Save computation cost by **15 ~ 40X**

[1] Adversarial Video Generation on Complex Datasets, DeepMind 2019

Experiments: In-Domain (UCF-101)

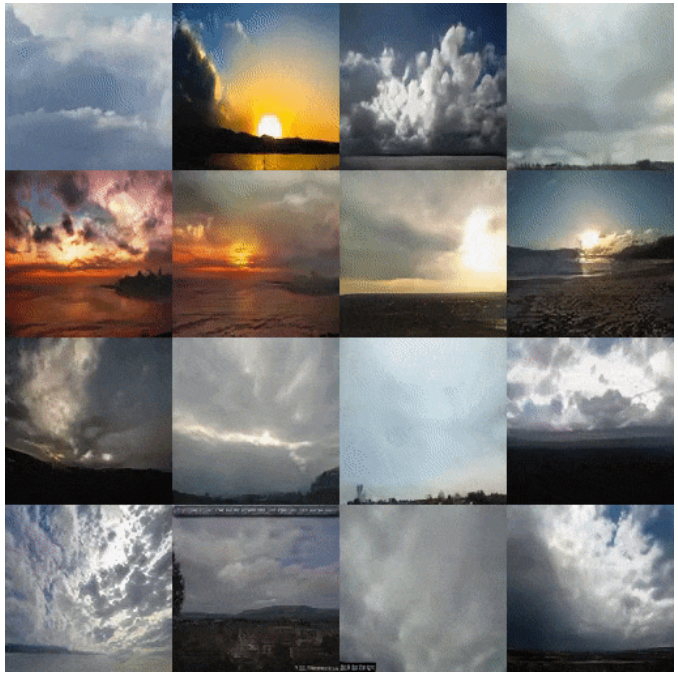
Method	IS (↑)	FVD (↓)
VGAN	8.31 ± .09	-
TGAN	11.85 ± .07	-
MoCoGAN	12.42 ± .07	-
ProgressiveVGAN	14.56 ± .05	-
TGANv2	26.60 ± .47	1209 ± 28
DVD-GAN	27.38 ± .53	-
Ours	**33.95 ± .25**	**700 ± 24**

Experiments: In-Domain (FaceForensics)

Method	FVD (↓)	ACD (↓)
GT	9.02	0.2935
TGANv2	58.03	0.4914
Ours	**53.26**	**0.3300**

[1] ACD from: MoCoGAN: Decomposing Motion and Content for Video Generation, CVPR 2018

Experiments: In-Domain (Sky Time-lapse)



Method	FVD (↓)	PSNR (↑)	SSIM (↑)
Up-B	-	25.367	0.781
MDGAN	840.95	13.840	0.581
DTVNet	451.14	21.953	0.531
Ours	**77.77**	**22.286**	**0.688**

Ablation: Mutual Info Loss

w/o mutual info loss

w/ mutual info loss

Ablation: Contrastive Loss

w/o contrastive discriminator

w/ contrastive discriminator

Summary

- Good image generator benefits video generation
  - LSTM motion generator + residual design
  - Pre-trained & fixed image generator
  - Contrastive 2D discriminator
  - Mutual information loss

- Code will be released at: https://github.com/snap-research/MoCoGAN-HD

Thanks!