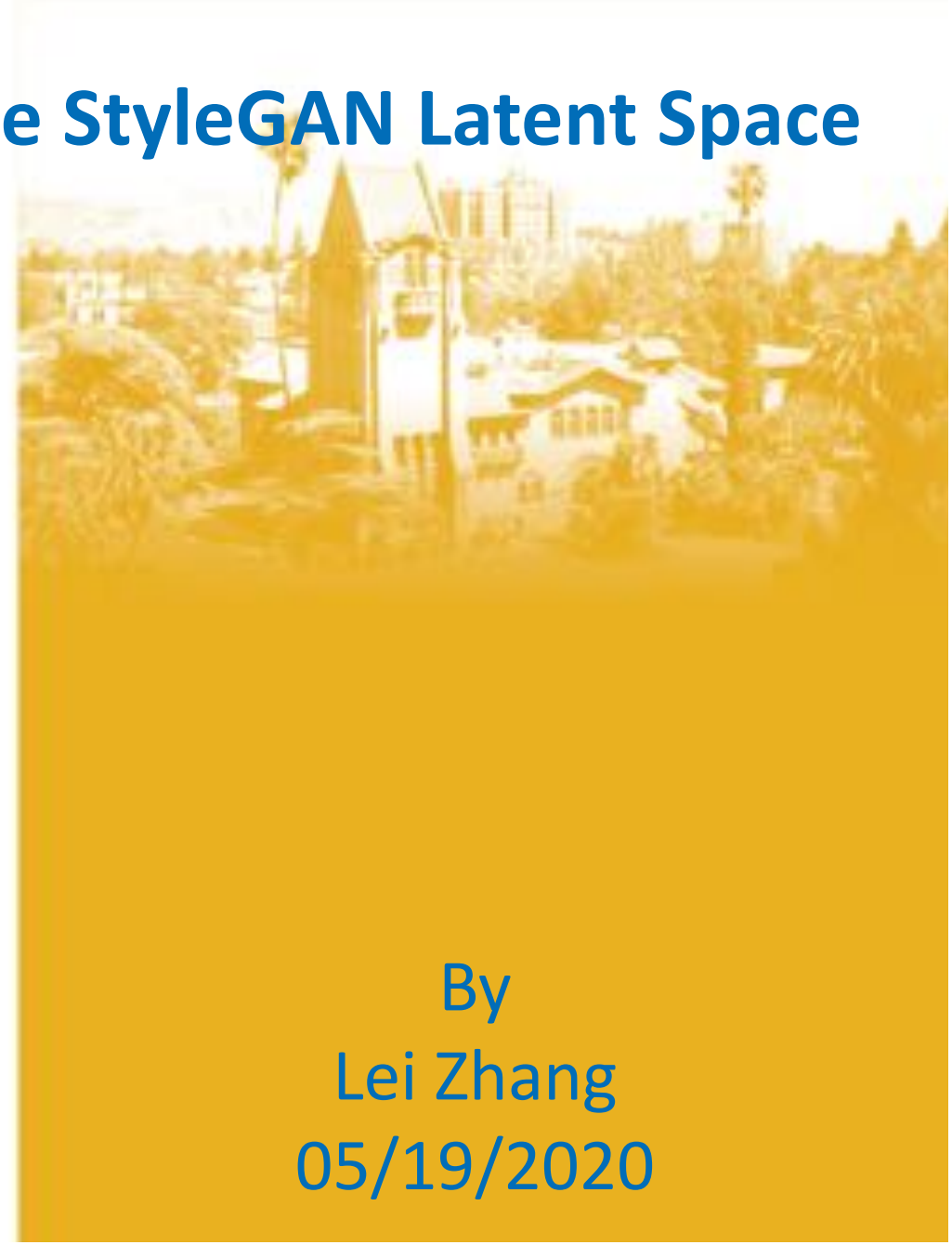
Video Synthesis from the StyleGAN Latent Space



SJSU SAN JOSÉ STATE UNIVERSITY

Advisor
Dr. Chris Pollett
Committee Members
Dr. Philip Heller
Dr. Leonard Wesley

By
Lei Zhang
05/19/2020

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

- Project Goals
- Video Generation Problems
- Related Works
- Implementation
- Experiments and Results
- Conclusion and Future Work
- References

SJSU SAN JOSÉ STATE UNIVERSITY

Project Goals

- Synthesize high-resolution and realistic video clips with artificial intelligence
- Inherit the success of image GANs researches and apply them to video generation
- Generate realistic videos of human facial expressions by using single start images

SJSU SAN JOSÉ STATE UNIVERSITY

Video Generation Problems

- Generative Adversarial Networks (GANs)
  - Invented by Ian Goodfellow in 2014
  - Used on image and video synthesis
  - Discriminator and generator
- GANs generate high-resolution images, but it cannot generate a good video
  - Video generation is more complex
  - Extra temporal layer to learn
  - Require more computation resources
  - Distorted images
  - Short video clips (usually a few seconds long)
  - Low-resolution (256x256) video from a single starting image

SJSU SAN JOSÉ STATE UNIVERSITY

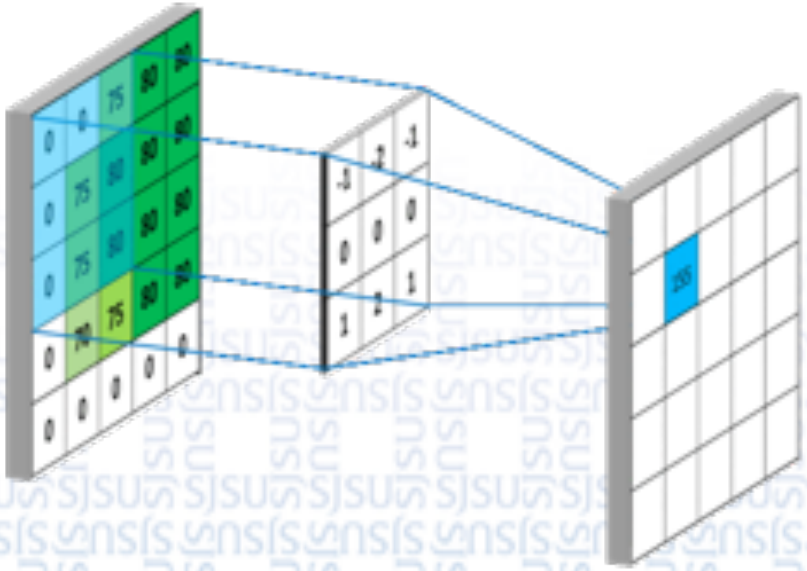
Related Works

- Convolutional Neural Networks (CNNs)
- Image GANs
- Video GANs
- VGG16 Network
- Embedding Images
- Sequence Prediction

SJSU SAN JOSÉ STATE UNIVERSITY

CNNs

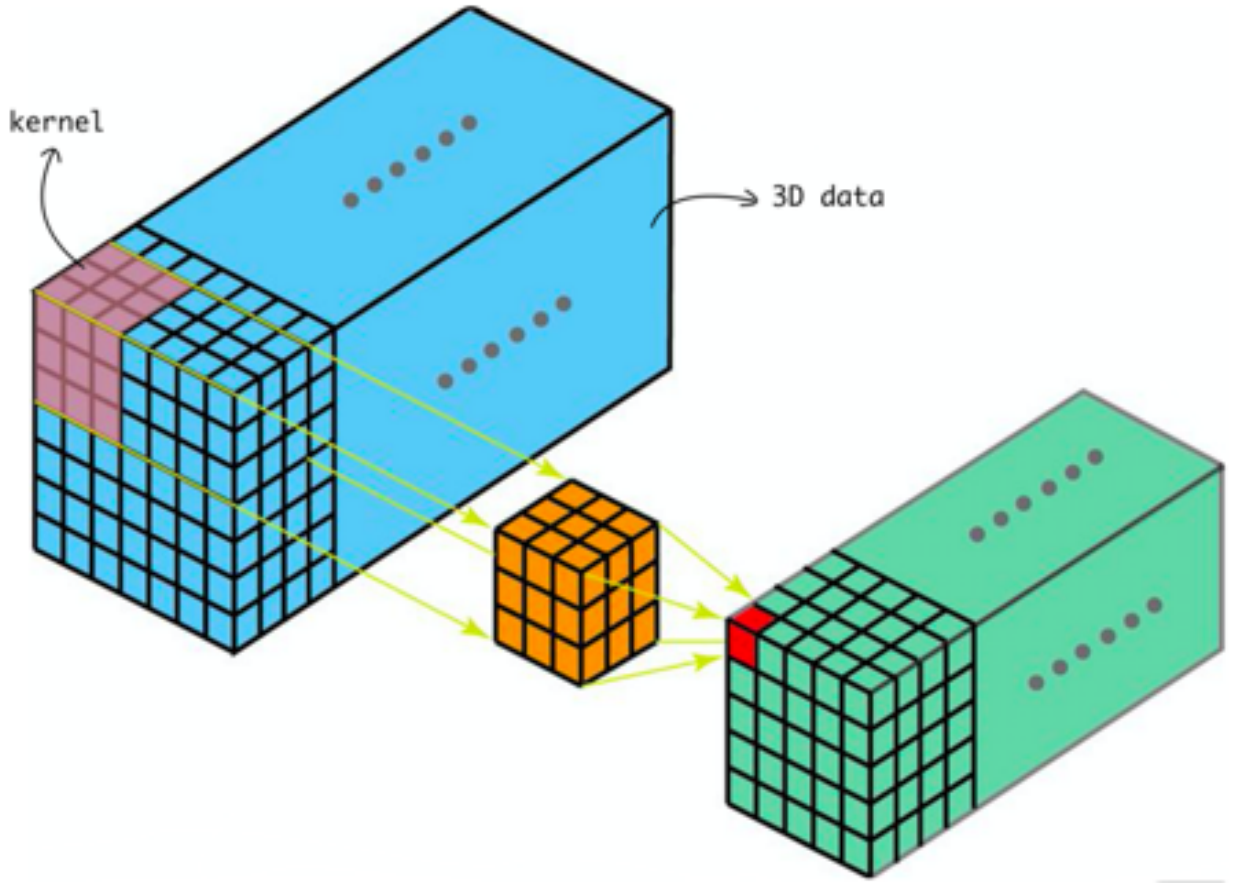
- One or more convolutional layers
- 2-dimentional (2D) CNNs
- Use a 3x3 kernel/filter
- Efficient in image recognition and classification



SJSU SAN JOSÉ STATE UNIVERSITY

CNNs

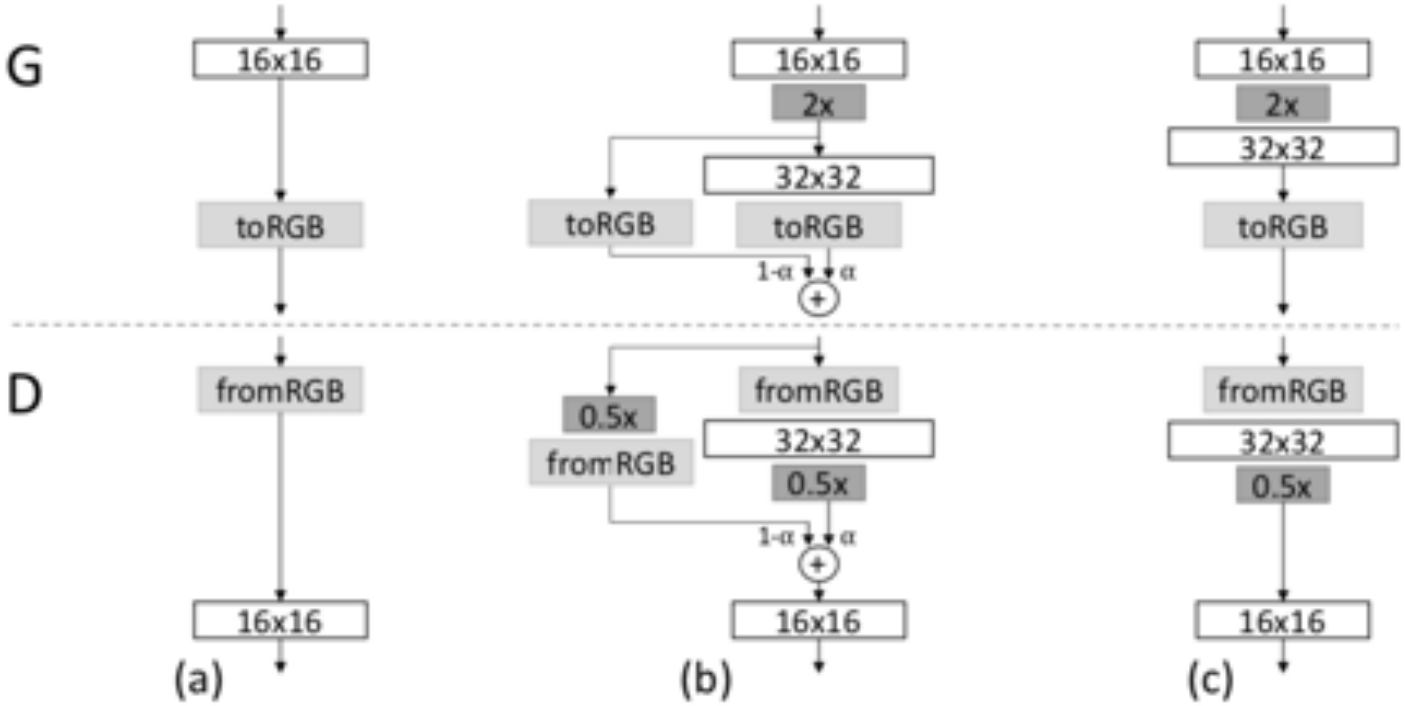
- 3-dimentional (3D) CNNs
- Use a 3x3x3 kernel/filter



SJSU SAN JOSÉ STATE UNIVERSITY

Image GANs

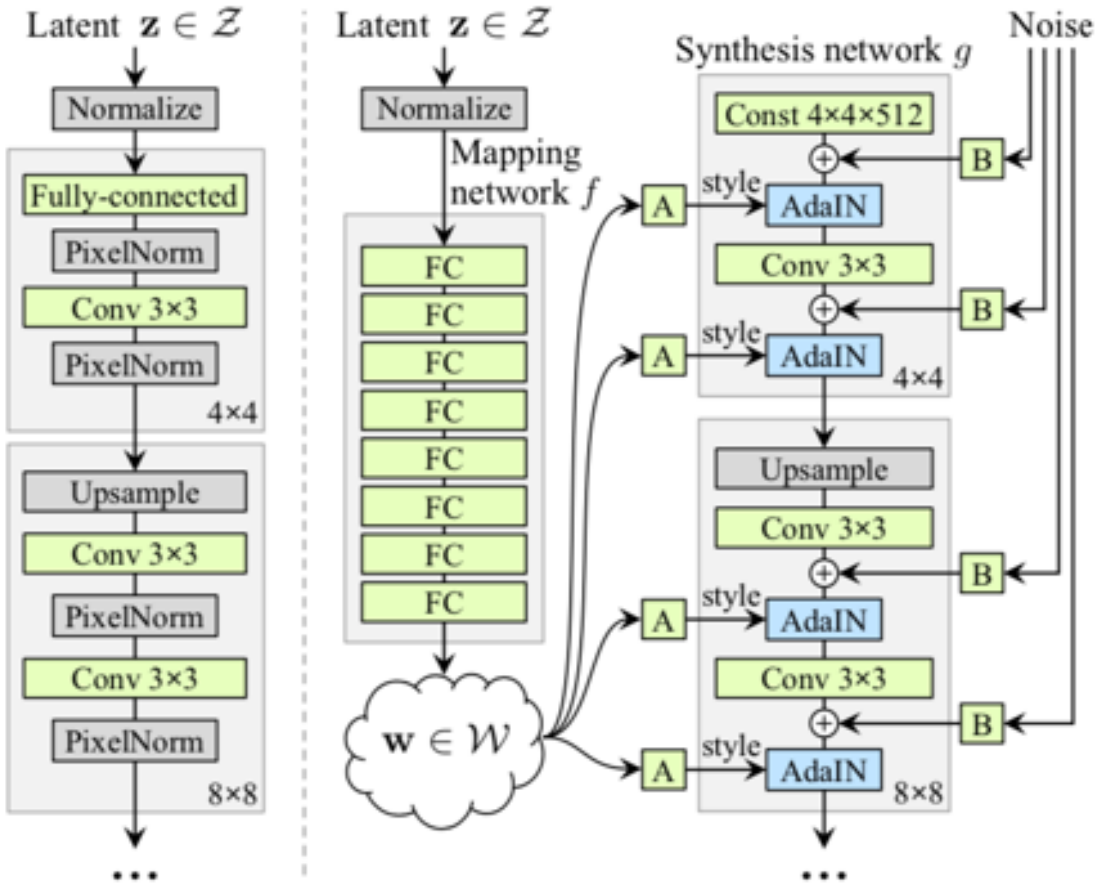
- Progressive Growing of GAN (ProGAN)
  - Proposed by Karras et al. in 2017
  - First time to generate images with 1024x1024 resolution
  - Propose a progressive growing method



SJSU SAN JOSÉ STATE UNIVERSITY

Image GANs

- StyleGAN
  - Based on the progressive growing concept of ProGAN
  - Further improved the ProGAN that suitable to do style mixing in the latent space
  - Adaptive Instance Normalization (AdaIN)



SJSU SAN JOSÉ STATE UNIVERSITY

Image GANs

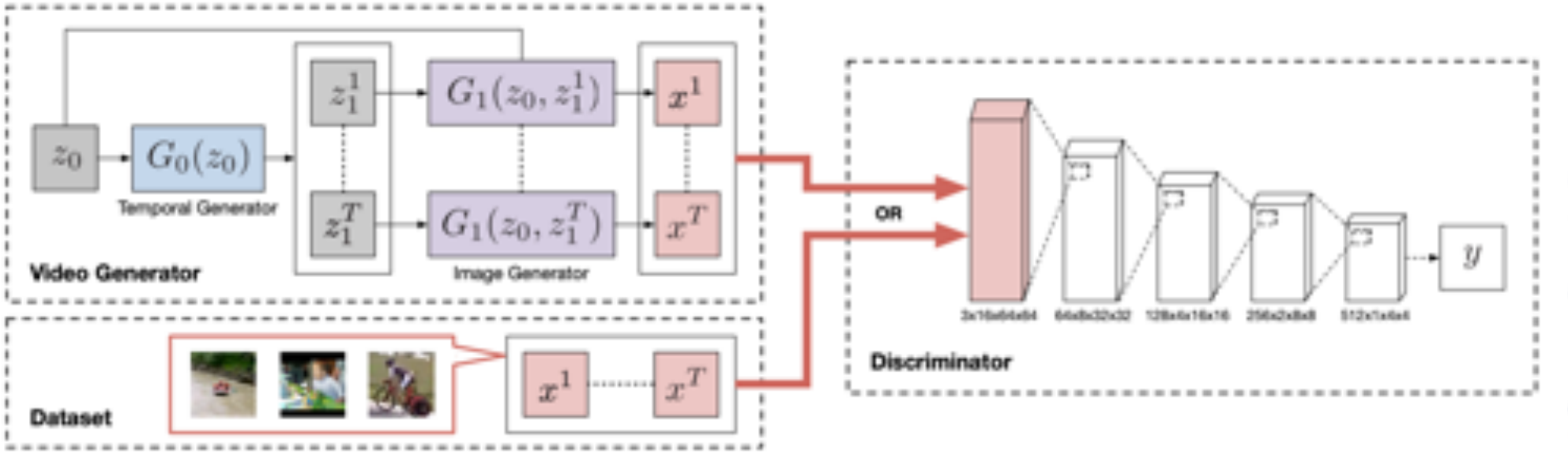
- StyleGAN – style mixing

Image GANs

- StyleGAN2
  - Fixed the water droplet -like artifacts issue in StyleGAN
  - Better training performance

Video GANs

- Temporal Generative Adversarial Nets (TGAN)
  - Abandon 3D CNN layers in its generator
  - Use a 1D deconvolutional layers to learn temporal features
  - Single stream 3D convolutional layers in the discriminator



SJSU SAN JOSÉ STATE UNIVERSITY

Video GANs

- TGAN
  - Generated videos

SJSU SAN JOSÉ STATE UNIVERSITY

Video GANs

- MocoGAN
  - Use a recurrent neural network (RNN)
  - Additional image based discriminator
  - Generated video clips with human facial expressions

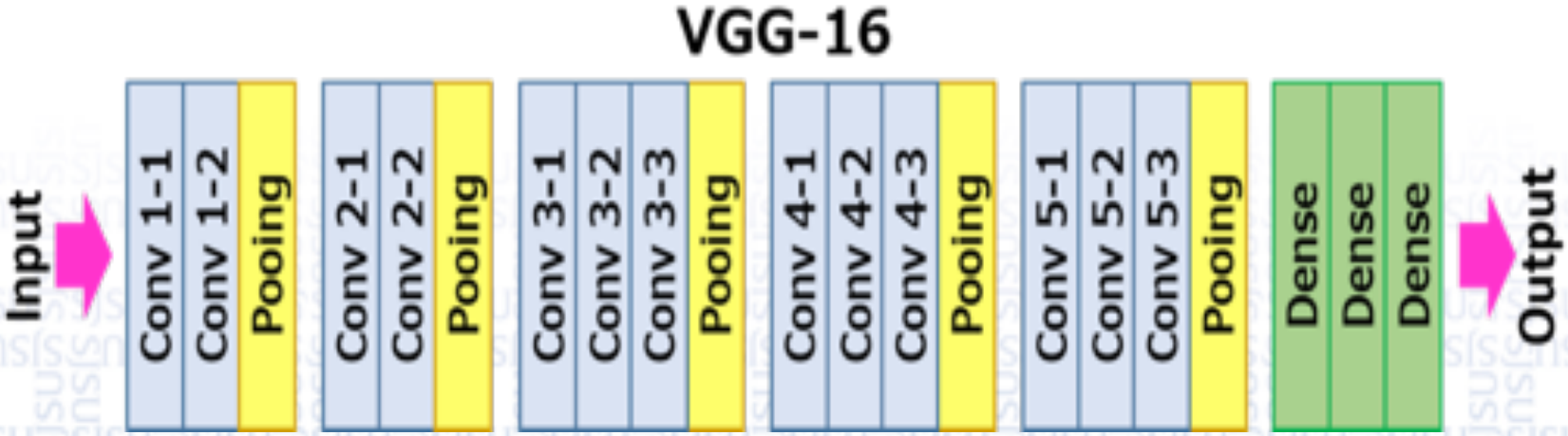
SJSU SAN JOSÉ STATE UNIVERSITY

Video GANs

- MocoGAN
  - Generated video clips with TaiChi dataset

VGG16

- Very Deep Convolutional Networks (VGG)
  - Use small (3x3) convolution filters instead of a larger one
  - VGG16 has 16 weight layers



SJSU SAN JOSÉ STATE UNIVERSITY

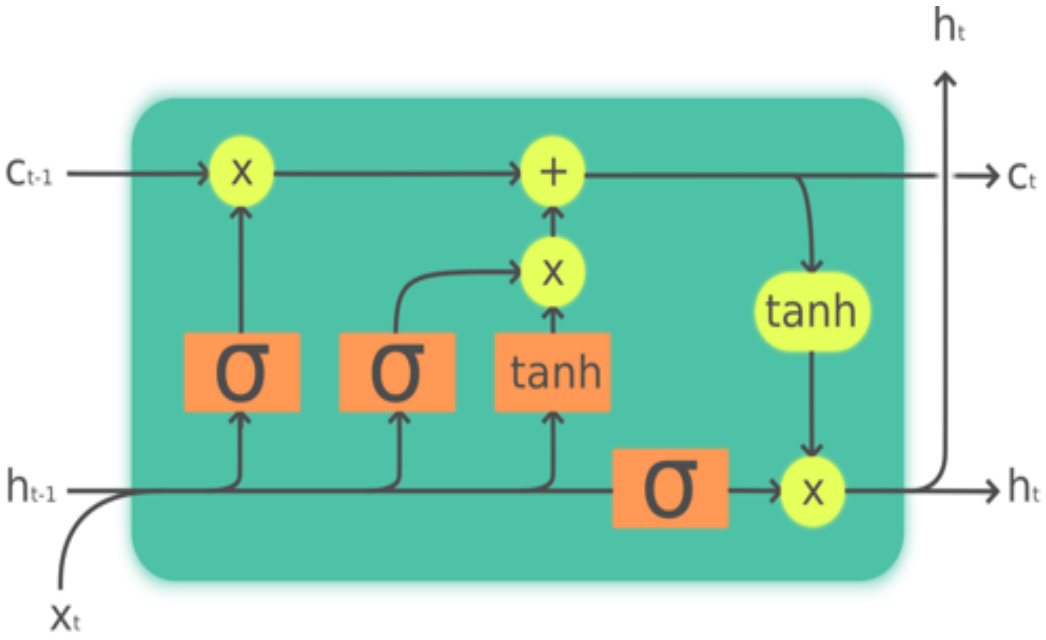
Embedding Images

- Image2StyleGAN
  - Use VGG image classification model as the feature extractor
  - Able to recover images from the StyleGAN latent space
  - Perfect embeddings are hard to reach

SJSU SAN JOSÉ STATE UNIVERSITY

Sequence Prediction

- Long short-term memory (LSTM)
  - For facial expressions prediction
  - Suitable to learn order dependence in sequence prediction problems
  - Output depends on three inputs: current input, previous output and previous hidden state
  - A LSTM unit is cell



SJSU SAN JOSÉ STATE UNIVERSITY

Implementation

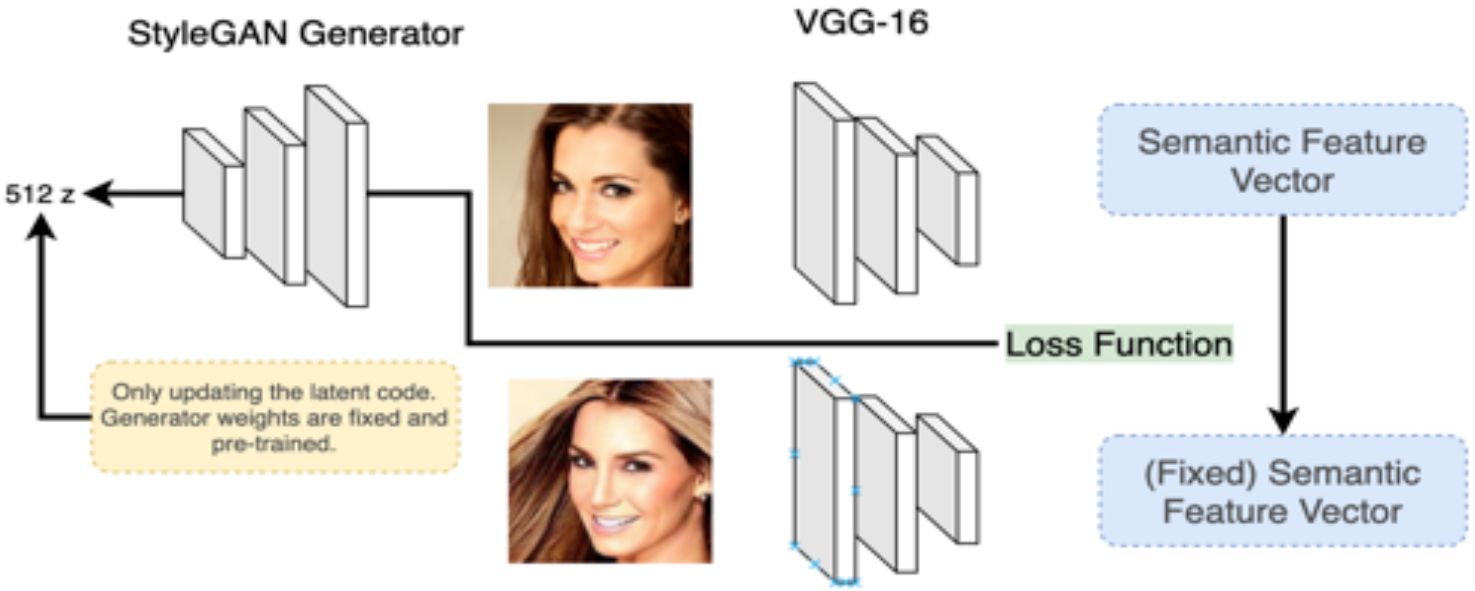
1. Create facial expression directions
   - Embed videos into StyleGAN latent space
   - Learn facial expression directions
2. Predict sequence of facial expressions
3. Generate videos of human facial expression
   - Generate keyframes
   - Latent vector based video Interpolation

SJSU SAN JOSÉ STATE UNIVERSITY

Implementation – Stage 1

Embed images into the StyleGAN latent space

- Use the pre-trained StyleGAN latent space
- Use StyleGAN generator
- Extract image features with VGG16 pretrained model on ImageNet dataset
- Back-Propagation only to update the latent vector



SJSU SAN JOSÉ STATE UNIVERSITY

Implementation – Stage 1

Learn facial expression directions

- Use logistic regression model to learn the directions
- Logistic regression is a classification machine learning model to predict binary results, which is an extension of linear regression
- Input: images - > latent vectors pairs
- Output: Facial expression directions
- Each direction represents a different emotion

SJSU SAN JOSÉ STATE UNIVERSITY

Implementation – Stage 2

Predict facial expressions in a movie

- Predict a sequence of emotions in natural order
- Use YouTube movie trailers
- Use EmoPy to predict facial expressions in the human faces
- EmoPy is a python tool which predicts emotions by providing images of people's faces
- Use 4 LSTM layers to train the emotion sequences
- Input: a sequence of emotions (0-6) from a YouTube movie trailer
- Output: predicted new emotion sequence in time order

SJSU SAN JOSÉ STATE UNIVERSITY

Implementation – Stage 3

Generate keyframes

- Generate face: Use a random noise vector z in the StyleGAN latent space
- Generate emotions: z + coefficent * directions
- Each keyframe represents a facial expression
- Reorder the keyframes with predicted emotion sequence

Implementation – Stage 3

Latent vector-based video interpolation

- Generate linear interpolation among all the keyframes
- Make the video looks smooth in transition
- Also called morphing
- Suppose w_1 and w_2 as two latent vectors

$$w = \lambda w_1 + (1 - \lambda) w_2, \lambda \in (0, 1)$$

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments

Datasets

- IMPA-FACE3D
  - The dataset collects 534 static images from 30 people with 6 samples of human facial expressions, 5 samples of mouth and eyes open and/or closed, and 2 samples of lateral profiles.

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments

Datasets

- MUG Facial Expression Database
  - Consists of 86 people of performing 6 basic expressions: anger, disgust, fear, happiness, sadness and surprise. Each video has a rate of 19 frames, and each image has 896x896 pixels.

Experiments

Datasets

- StyleGAN Flickr-Faces-HQ (FFHQ)
  - FFHQ is a human faces dataset which consists of 70,000 high-quality PNG images at 1024×1024 resolution. These aligned images were downloaded from Flickr and were used to train StyleGAN. I was created by the authors in StyleGAN paper.

Experiments

Embedding Images

- Acquire paired mappings of images to noise vectors

Experiments

Learn Facial Expressions Directions

- Effect of using different coefficients

Experiments

Morphing (Video Interpolation)

- Linear transition between two frames

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments

Comparison

- Average Content Distance (ACD)
- Calculate average L2 distance among all consecutive frames in a video
- A smaller ACD score is better which means a generated video is more likely to be the same person
- 17%, with an AVG of generated 210 videos

ACD	Facial Expressions
TGAN [5]	0.305
MoCoGAN [5]	0.201
My Model	0.167

SJSU SAN JOSÉ STATE UNIVERSITY

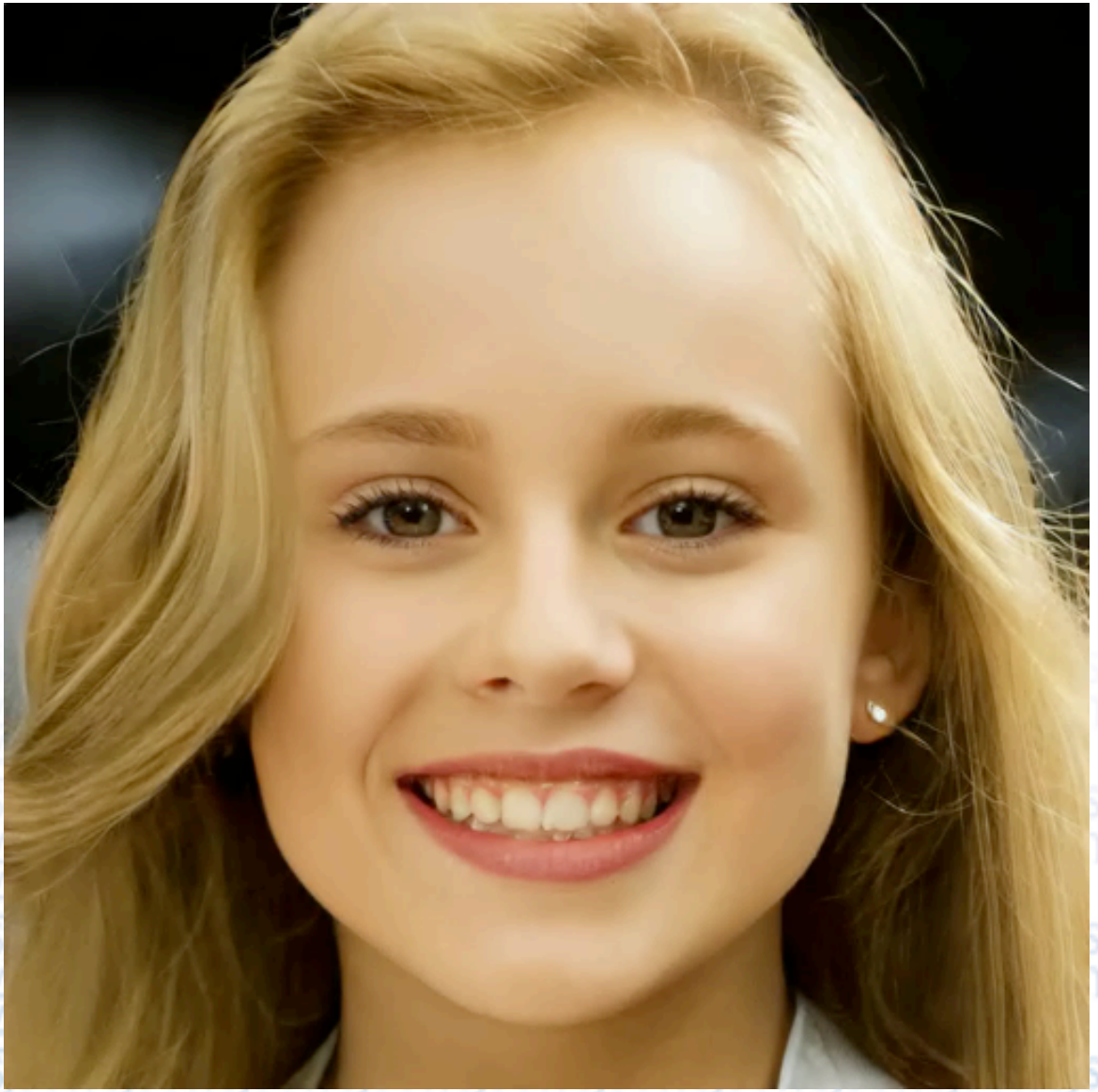
Experiments

Synthesis Video 1

SJSU SAN JOSÉ STATE UNIVERSITY

Experiments

Synthesis Video 2



SJSU SAN JOSÉ STATE UNIVERSITY

Conclusion and Future Work

- Transfer learning from image GANs saves a lot of training time to generate videos
- A well-trained image GANs latent space has enough frames to compose a video
- Use the model to generate other type of videos rather than facial expressions
- Explore different ways to find continuous frames in an image GAN latent space able to generate high-resolution videos

SJSU SAN JOSÉ STATE UNIVERSITY

THANK YOU!

SJSU SAN JOSÉ STATE UNIVERSITY

Reference

[1] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping, " In *ICCV*, 2017.

[2] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural computation 9.8*, 1997.

[3] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?," *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[4] T. Karras, et al., "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv*:1710.10196, 2017.

[5] S. Tulyakov, et al., "Mocogan: Decomposing motion and content for video generation," In *CVPR*, 2018.

SJSU SAN JOSÉ STATE UNIVERSITY

Reference

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG facial expression database," *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. IEEE*, 2010.

[8] T. Karras, et al., "Analyzing and improving the image quality of stylegan," *arXiv preprint arXiv*:1912.04958, 2019.

[9] S. Ji, et al., "3D convolutional neural networks for human action recognition," *TPAMI, 35(1):221– 231*, 2013.