Lecture 21: Deep Generative Models IV: Variants of GANs

Xuming He, Lan Xu SIST, ShanghaiTech Fall, 2020

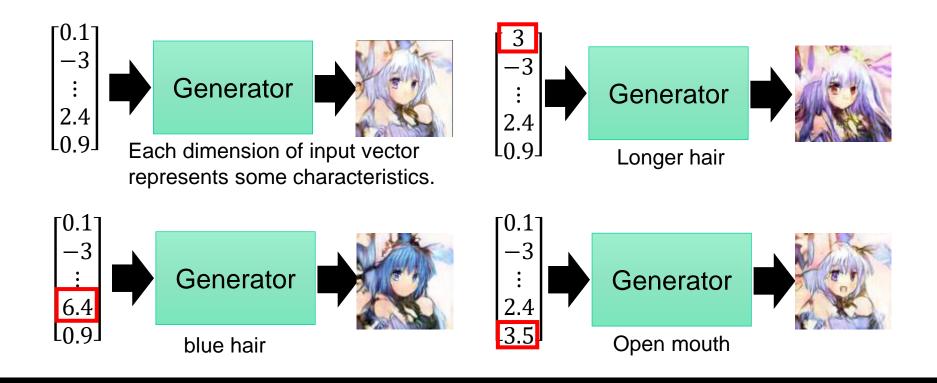


Outline

- More GAN variations
 - □ GAN + AE
 - StyleGAN
- GANs on Videos
 - DVD-GAN
 - □ Conditional GANs on Videos

Acknowledgement: TUM, NTU, Stanford notes

Modifying Input Code



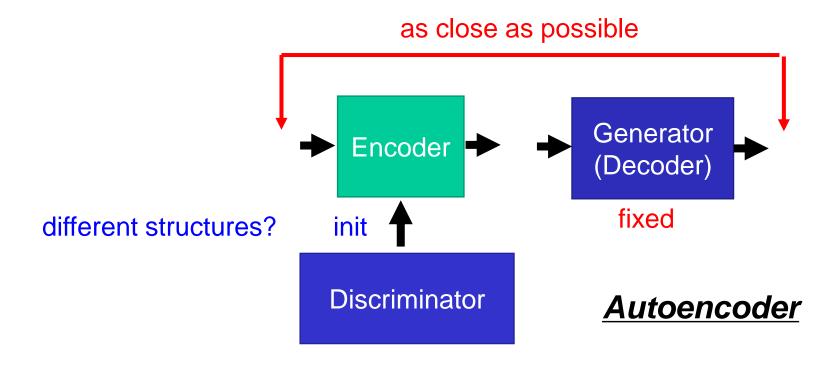
- > The input code determines the generator output.
- Understand the meaning of each dimension to control the output.

3

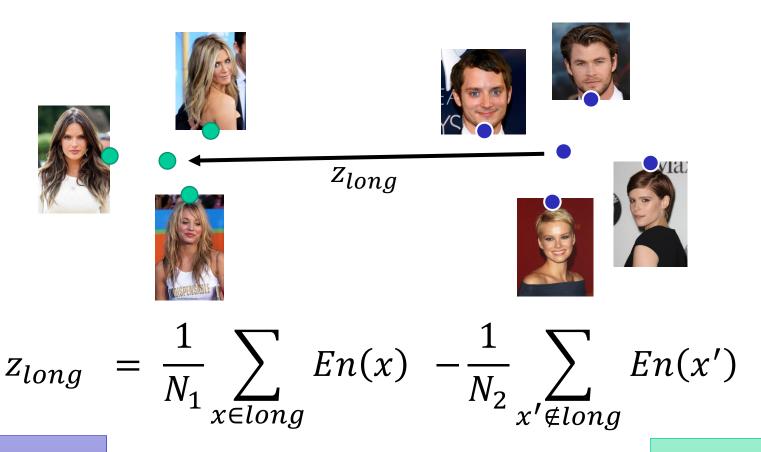


GAN + Autoencoder

- We have a generator (input z, output x)
- However, given x, how can we find z?
 - □ Learn an encoder (input x, output z)



GAN + AE: attribute representation



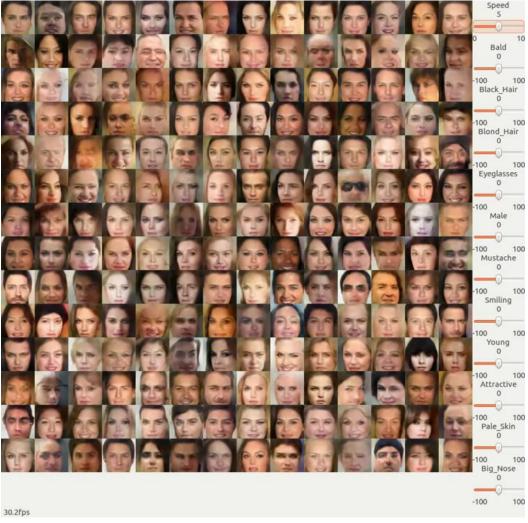
Short Hair

$$x \implies En(x) + z_{long} = z' \implies Gen(z')$$

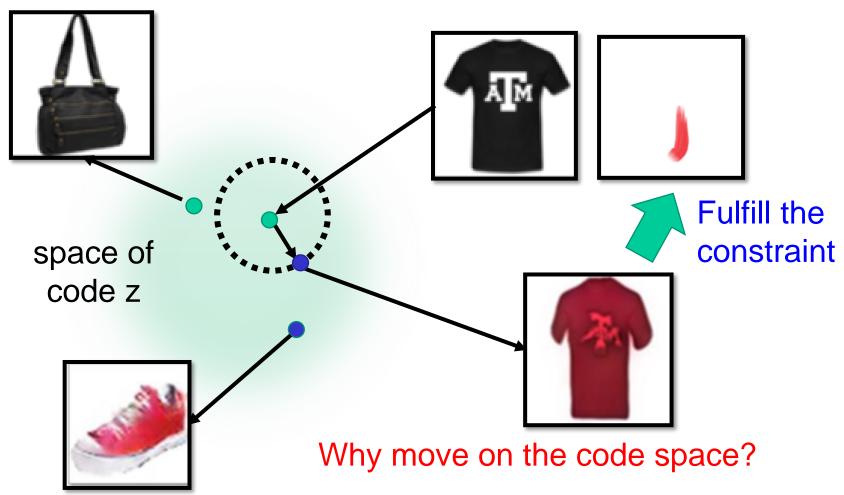
Long Hair

GAN + AE: attribute control

Photo editing via attribute control



Attribute editing: basic idea





Back to z

Method 1



$$z^* = arg \min_{z} \underline{L(G(z), x^T)}$$

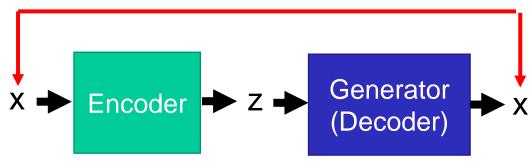
Gradient Descent

Difference between G(z) and x^T

- Pixel-wise
- By another network

as close as possible

Method 2

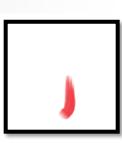


Method 3

Using the results from *method 2* as the initialization of *method 1*

Editing Photos





 \mathbf{z}_0 is the code of the input image

image

Using discriminator to check the image is realistic or not

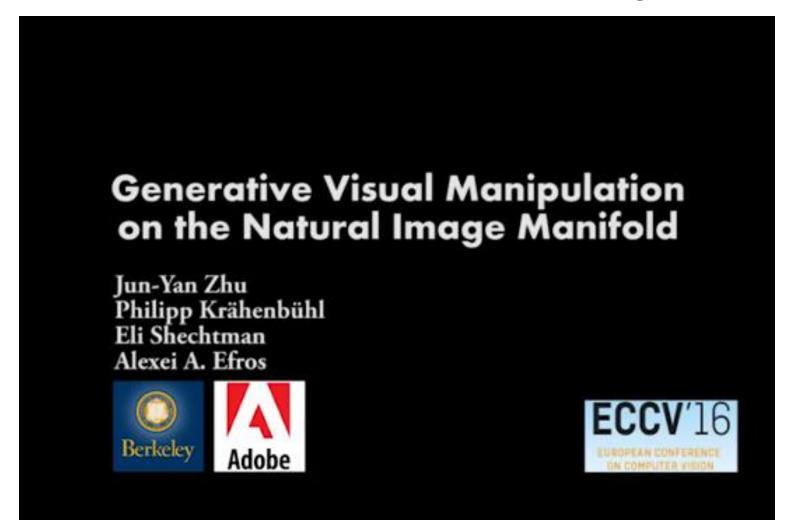
$$z^* = arg \min_{z} \underline{U(G(z))} + \lambda_1 \underline{||z - z_0||^2} - \lambda_2 \underline{D(G(z))}$$

Not too far away from the original image



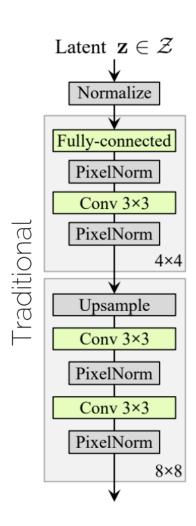
Does it fulfill the constraint of editing?

GAN + AE: attribute editing



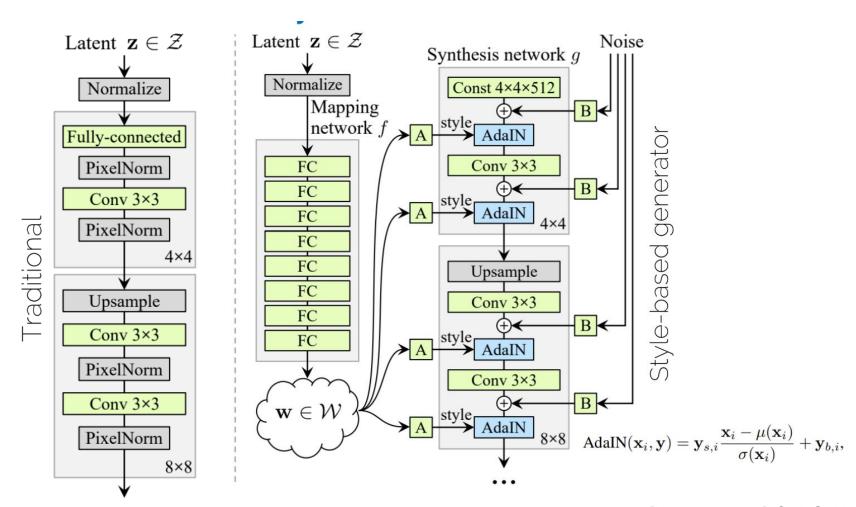
[Zhu et al. 2016] Generative Visual Manipulation on the Natural Image Manifold





[Karras et al. 19]: StyleGAN





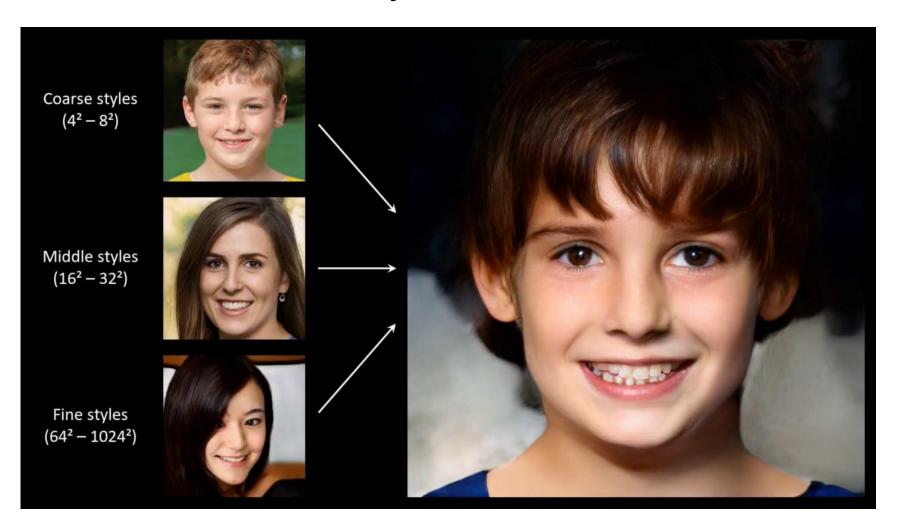
[Karras et al. 19]: StyleGAN



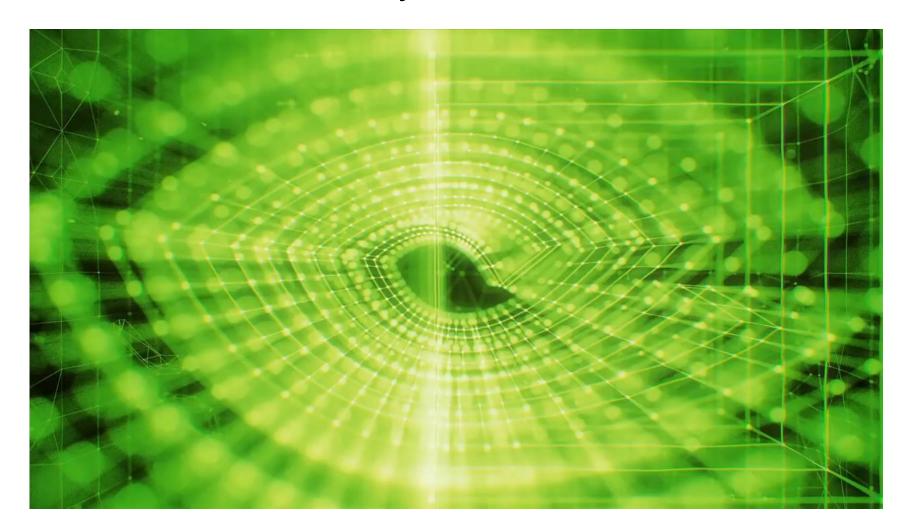
- FID on 50k gen. images
- Architecture is similar to Progressive Growing GAN

Method	CelebA-HQ	FFHQ
A Baseline Progressive GAN [30]	7.79	8.04
B + Tuning (incl. bilinear up/down)	6.11	5.25
C + Add mapping and styles	5.34	4.85
D + Remove traditional input	5.07	4.88
E + Add noise inputs	5.06	4.42
F + Mixing regularization	5.17	4.40

[Karras et al. 19]: StyleGAN



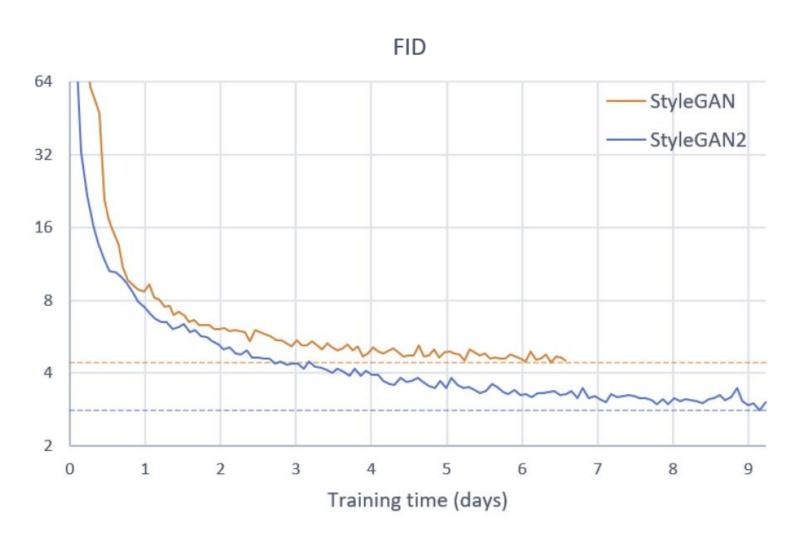
[Karras et al. 19]: StyleGAN



- Interesting analysis about design choices!
 - □ https://arxiv.org/pdf/1912.04958.pdf
 - https://github.com/NVlabs/stylegan2

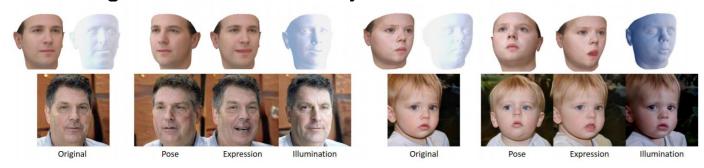


StyleGAN v.s StyleGAN2

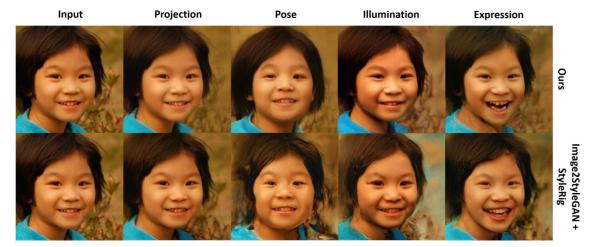


Latent space of StyleGAN

- StyleRig: Rigging StyleGAN for 3D Control over Portrait Images
 - Face rig-like control over StyleGAN

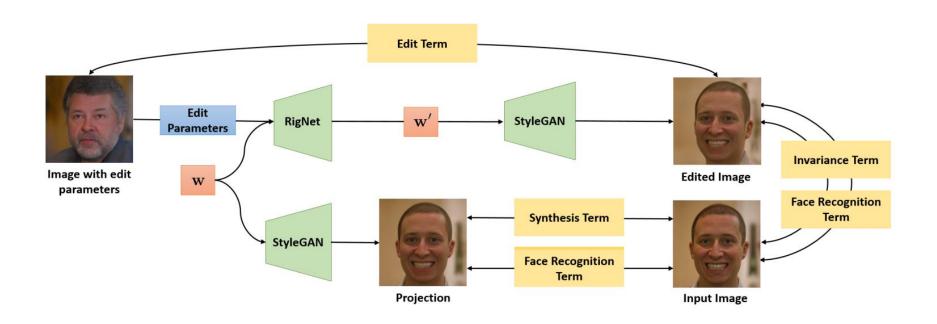


- PIE: Portrait Image Embedding for Semantic Control
 - Embedding in the latent space to edit real images



PIE: Portrait Image Embedding

Utilize pretrained StyleGAN



[TEWARI et al. 19] PIE: Portrait Image Embedding for Semantic Control

PIE: Portrait Image Embedding















PIE: Portrait Image Embedding for Semantic Control

SIGGRAPH Asia 2020



Ayush Tewari¹, Mohamed Elgharib¹, Mallikarjun B R¹, Florian Bernard^{1,2}, Hans-Peter Seidel¹, Patrick Pérez³, Michael Zollhöfer⁴, Christian Theobalt¹

¹MPI Informatics, Saarland Informatics Campus ²Technical University of Munich ³Valeo.ai ⁴Stanford University

[TEWARI et al. 19] PIE: Portrait Image Embedding for Semantic Control

Explicit Density Models

State of the art is pretty impressive

Vector Quantized Variational AutoEncoder



[Razavi et al. 19] Generating Diverse High-Fidelity Images with VQ-VAE-2

м

Outline

- More GAN variations
 - ☐ GAN + AE
 - □ StyleGAN
- GANs on Videos
 - □ DVD-GAN
 - □ Conditional GANs on Videos

Acknowledgement: TUM, NTU, Stanford notes



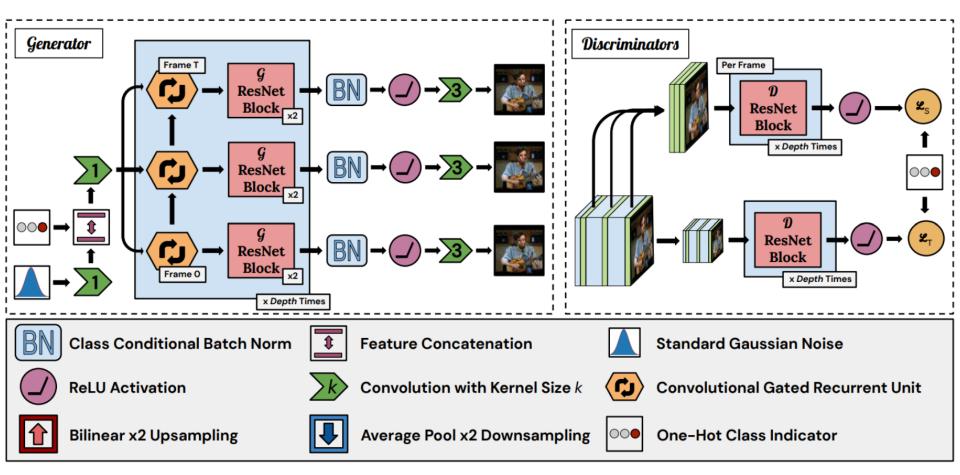
GANs on Videos

Two options

- □ Single random variable z seeds entire video (all frames)
 - Very high dimensional output
 - How to do for variable length?
 - Future frames deterministic given past
- Random variable z for each frame of the video
 - Need conditioning for future from the past
 - How to get combination of past frames + random vectors during training

General issues

- Temporal coherency
- Drift over time (many models collapse to mean image)



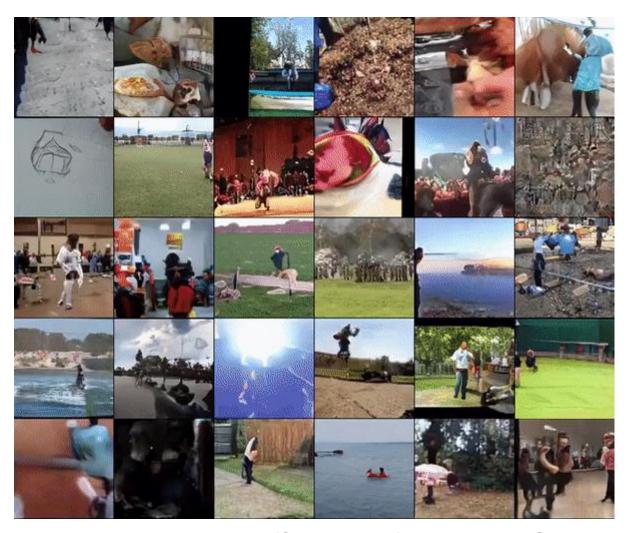
[Clark et al. 2019] Adversarial Video Generation on Complex Datasets



Time

[Clark et al. 2019] Adversarial Video Generation on Complex Datasets

25



[Clark et al. 2019] Adversarial Video Generation on Complex Datasets



- Trained on Kinetics-600 dataset
 - □ 256 x 256, 128 x 128, and 64 x 64
 - □ Lengths of up 48 frames
- This is state of the art!
- Videos from scratch still incredibly challenging

Conditional GANs on Videos

Challenge:

□ Each frame is high quality, but temporally inconsistent



Video-to-Video Synthesis

Sequential Generator:

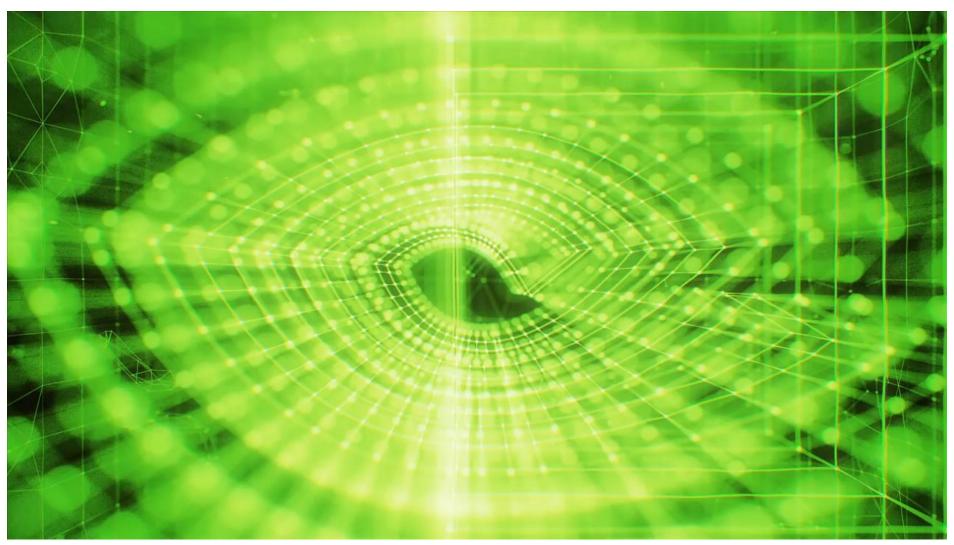
$$p(\tilde{\mathbf{x}}_1^T|\mathbf{s}_1^T) = \prod_{t=1}^T p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-L}^{t-1},\mathbf{s}_{t-L}^t).$$
 past L generated frames past L source frames (Set L = 2)

- lacktriangle Conditional Image Discriminator D_i (is it real image)
- lacktriangle Conditional Video Discriminator D_{v} (temp. consistency via flow)
- Full Learning Objective:

$$\min_{F} \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \mathcal{L}_W(F)$$

[Wang et al. 2018] Vid2Vid

Video-to-Video Synthesis



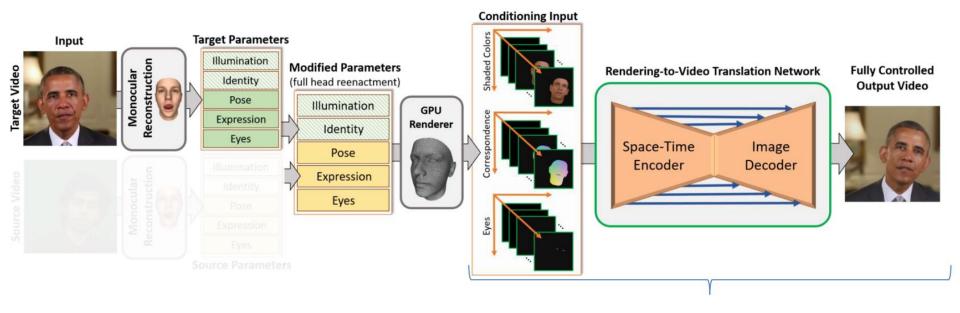
[Wang et al. 2018] Vid2Vid



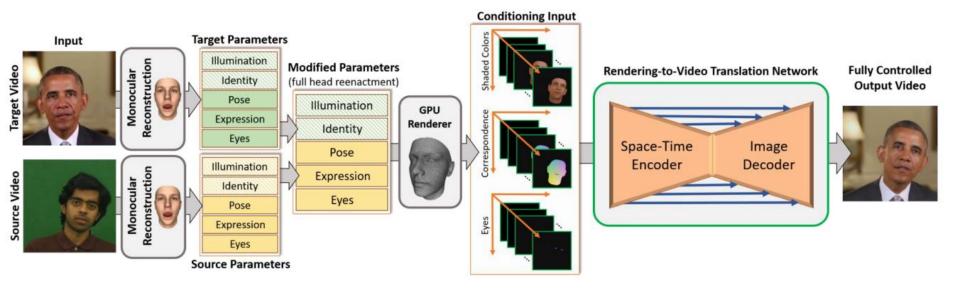
Key ideas:

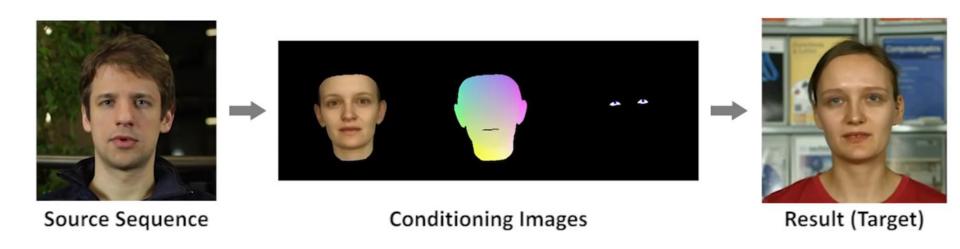
- Separate discriminator for temporal parts
 - □ In this case based on optical flow
- Consider recent history of previous frames

Train all of it jointly

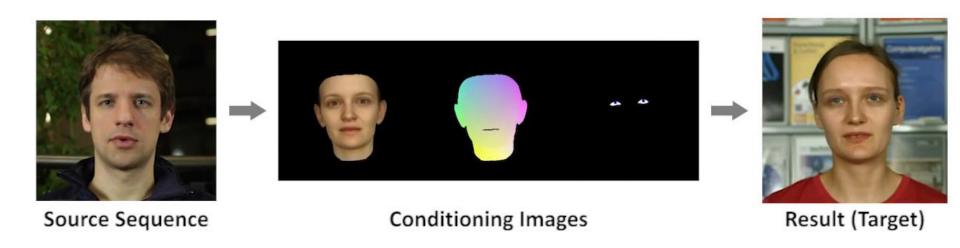


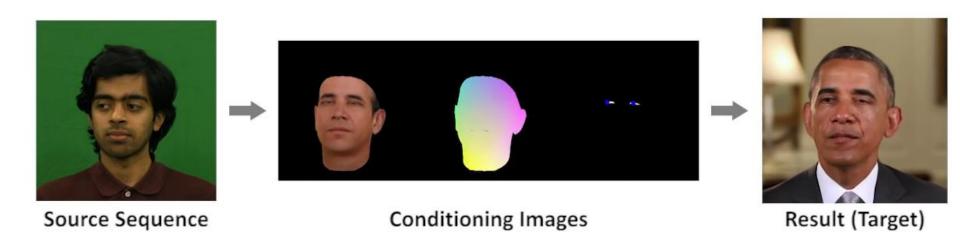
Similar to "Image-to-Image Translation" (Pix2Pix) [Isola et al.]





Neural Network converts synthetic data to realistic video



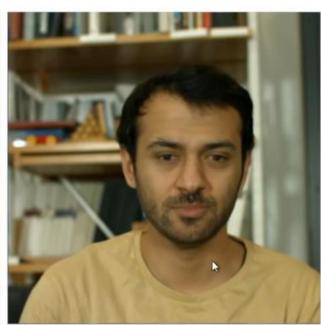




Siggraph'18 [Kim et al 2018]: Deep Portraits



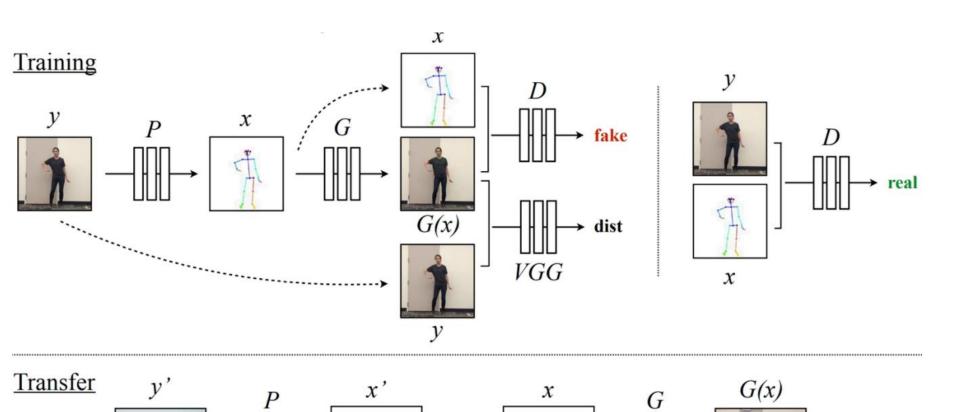






Key ideas:

- Synthetic data for tracking is great anchor / stabilizer
- Overfitting on small datasets works pretty well
- Need to stay within training set w.r.t. motions
- No real learning; essentially, optimizing the problem with SGD



Norm

[Chan et al. '18] Everybody Dance Now

Everybody Dance Now

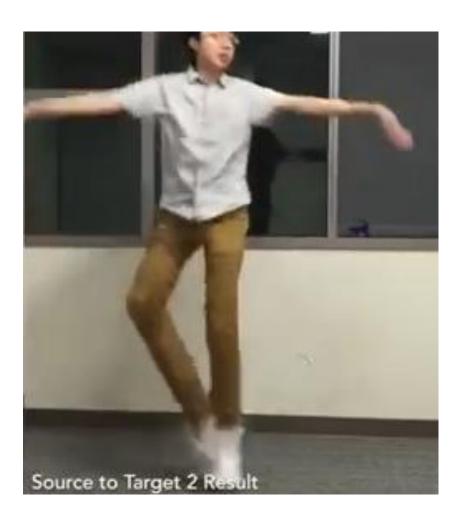
Motion Retargeting Video Subjects

Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros

UC Berkeley

[Chan et al. '18] Everybody Dance Now

- cGANs work with different input
- Requires consistent input i.e., accurate tracking
- Network has no explicit 3D notion



[Chan et al. '18] Everybody Dance Now



Key ideas:

- Conditioning via tracking seems promising!
- Tracking quality translates to resulting image quality
- Tracking human skeletons is less developed than faces
 - □ Temporally it's not stable... (e.g., OpenPose etc.)
- Fun fact, there were like 4 papers with a similar idea that appeared around the same time... (even more papers recently)



Summary

- Variants of GANs
- GANs on Videos
 - DVD-GAN: still incredibly challenging
 - Conditional GANs on Videos
- Next time
 - Advanced topics and guest lectures
- Keep working on the projects!

44