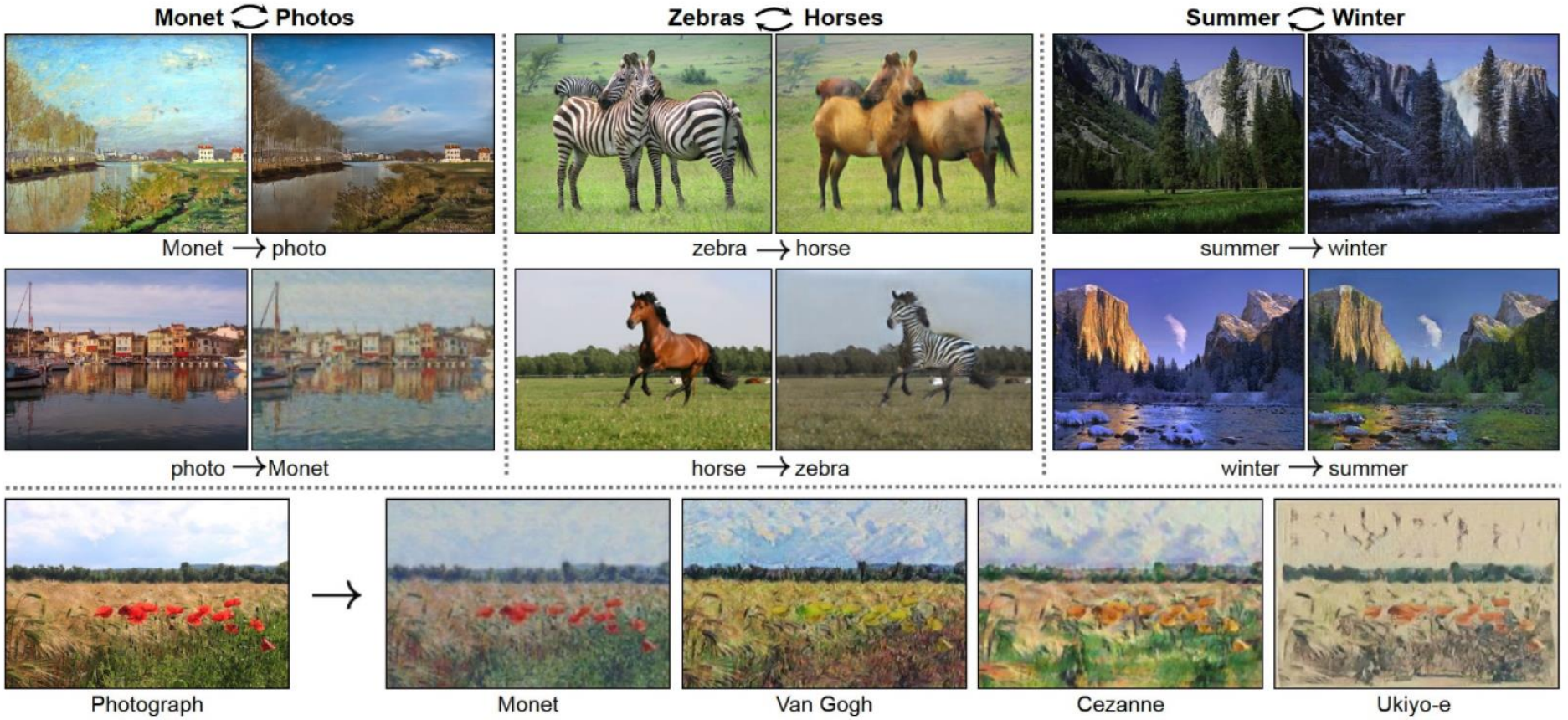
Lecture 22: Deep Generative Models VI: Variants of GANs

Lan Xu
SIST, ShanghaiTech
Fall, 2022

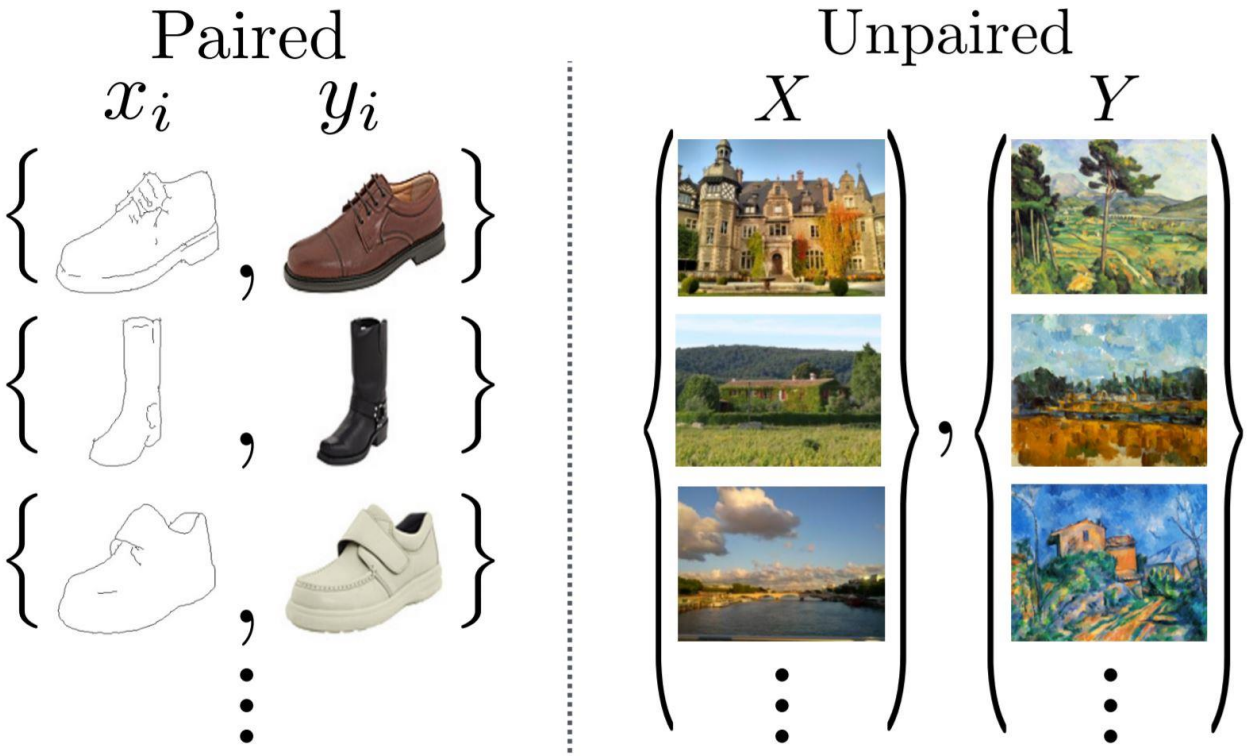
CycleGAN

- Image-to-image translation without paired data



CycleGAN

- If we had paired data (same content in both styles), this would be a supervised learning problem. But this is hard to find.



CycleGAN

- If we had paired data (same content in both styles), this would be a supervised learning problem. But this is hard to find.
- The CycleGAN architecture learns to do it from unpaired data.
  - Train two different generator nets to go from style 1 to style 2, and vice versa.
  - Make sure the generated samples of style 2 are indistinguishable from real images by a discriminator net.
  - Make sure the generators are **cycle-consistent**: mapping from style 1 to style 2 and back again should give you almost the original image.

CycleGAN

Real zebra image

The discriminator tries to distinguish generated zebra images from real ones

D

Discriminator loss: GAN generator objective, i.e. negative log probability D assigns to the sample being real

Reconstruction loss: squared error between the original image and the reconstruction

Input image (real horse image)

G_1

Generator 1 learns to map from horse images to zebra images while preserving the structure

Generated sample

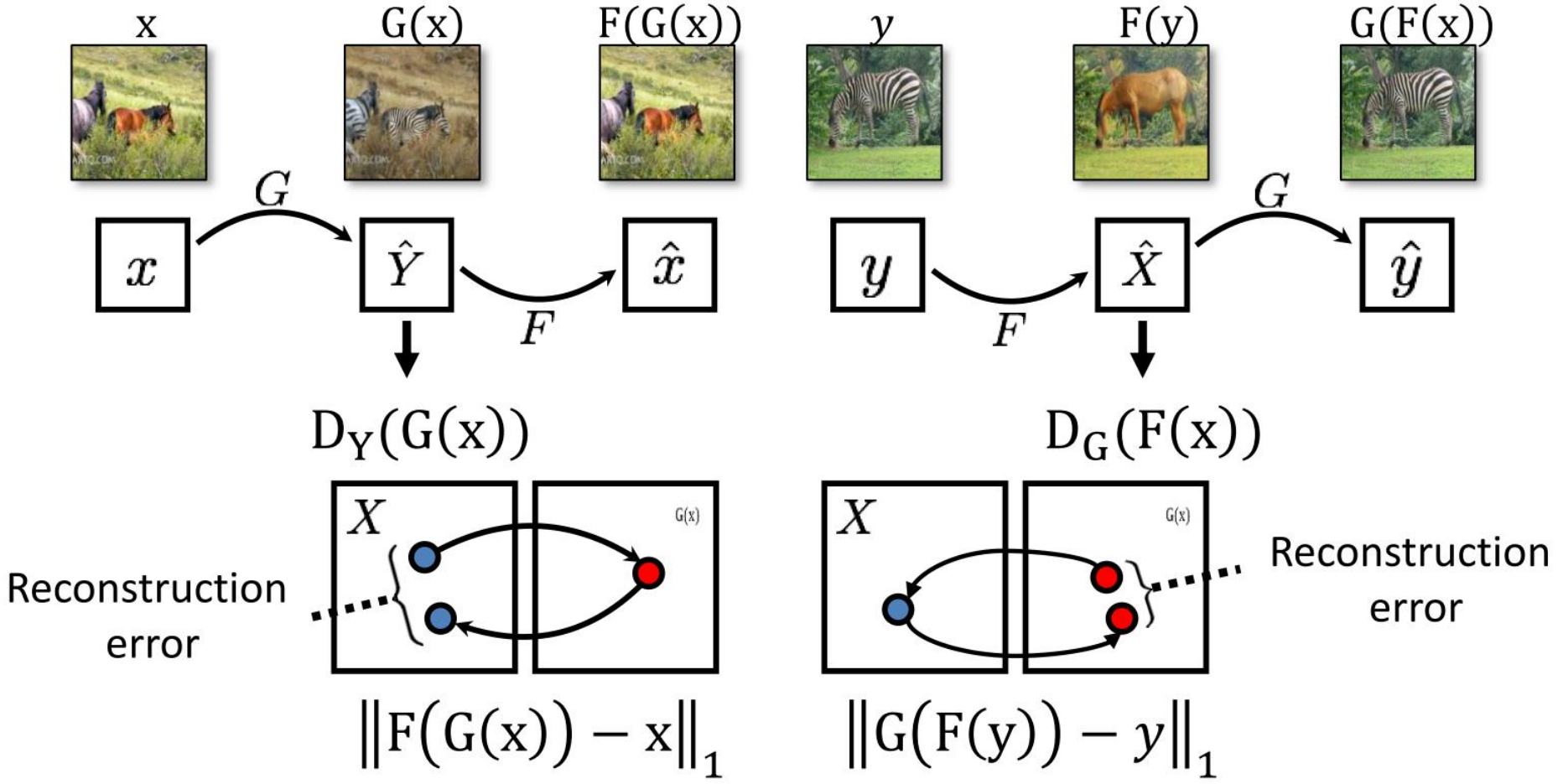
G_2

Generator 2 learns to map from zebra images to horse images while preserving the structure

Reconstruction

Total loss = discriminator loss + reconstruction loss

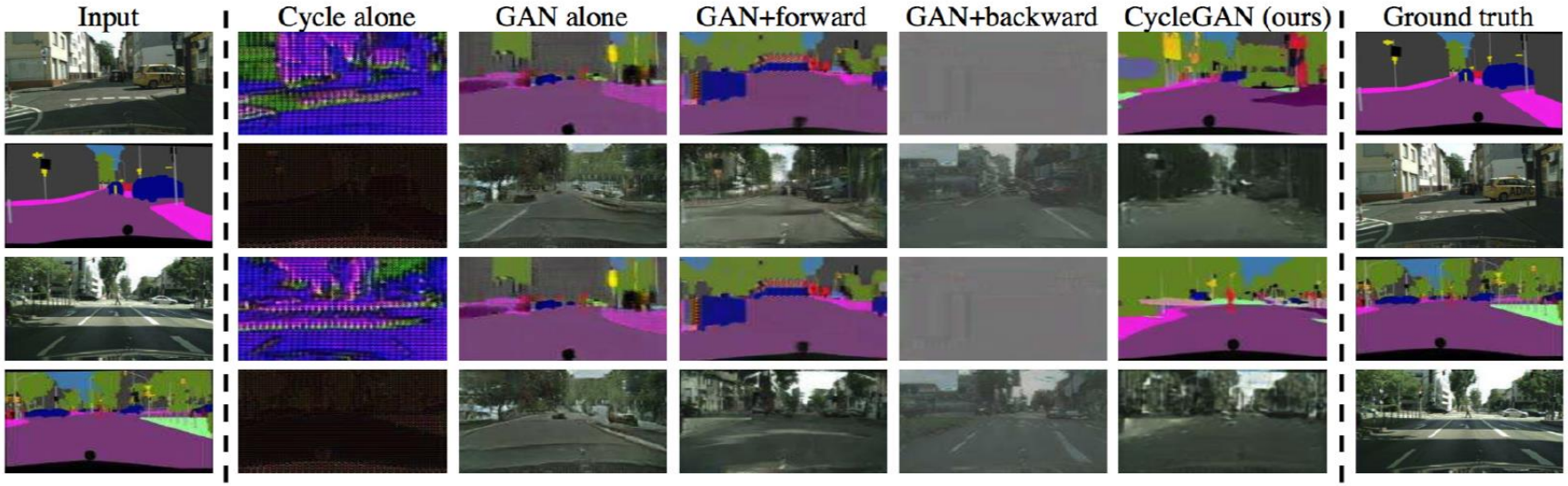
CycleGAN


x
G(x)
F(G(x))
y
F(y)
G(F(x))
G
F
G
F
x
$\hat{Y}$
$\hat{x}$
y
$\hat{X}$
$\hat{y}$
$D_Y(G(x))$
$D_G(F(x))$
X
G(x)
X
G(x)
Reconstruction error
Reconstruction error
$\|F(G(x)) - x\|_1$
$\|G(F(y)) - y\|_1$

CycleGAN

- Results

Style transfer between road scenes and semantic segmentations (labels of every pixel in an image by object category):



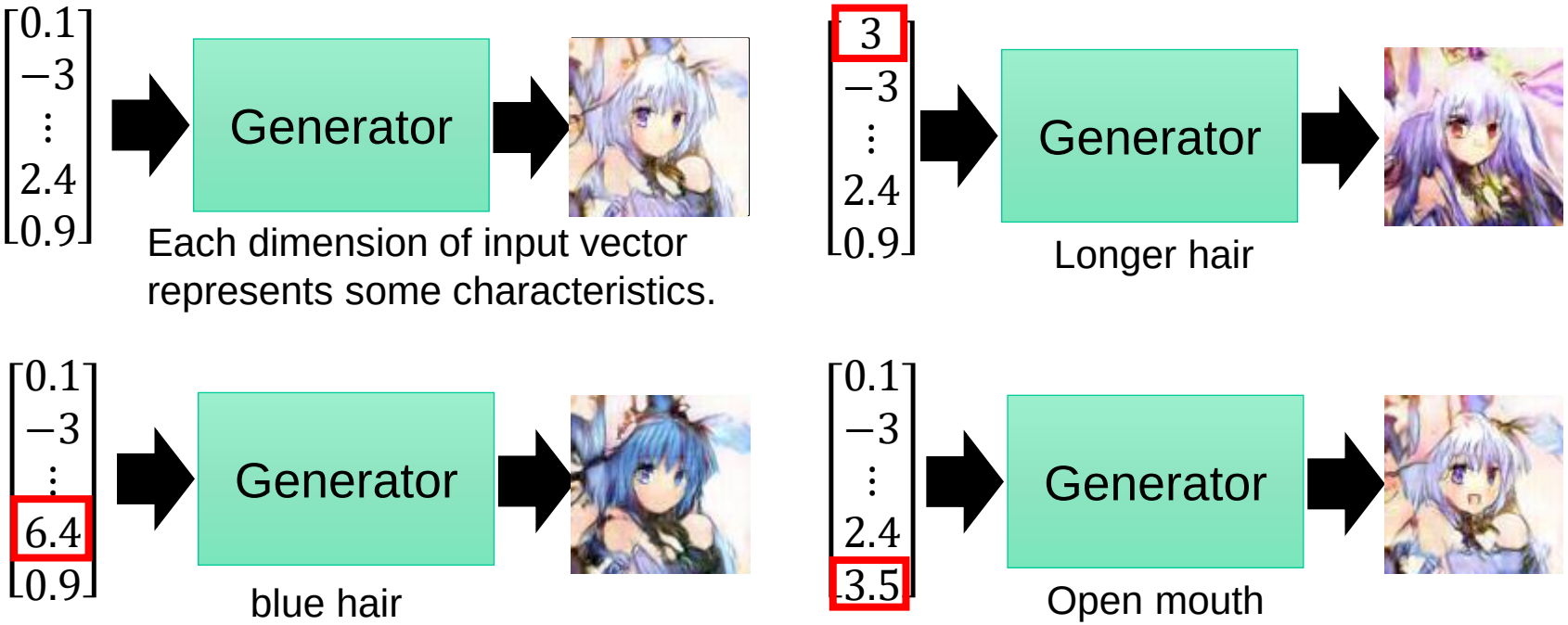
CycleGAN

- Results

- More details

https://hardikbansal.github.io/CycleGANBlog/

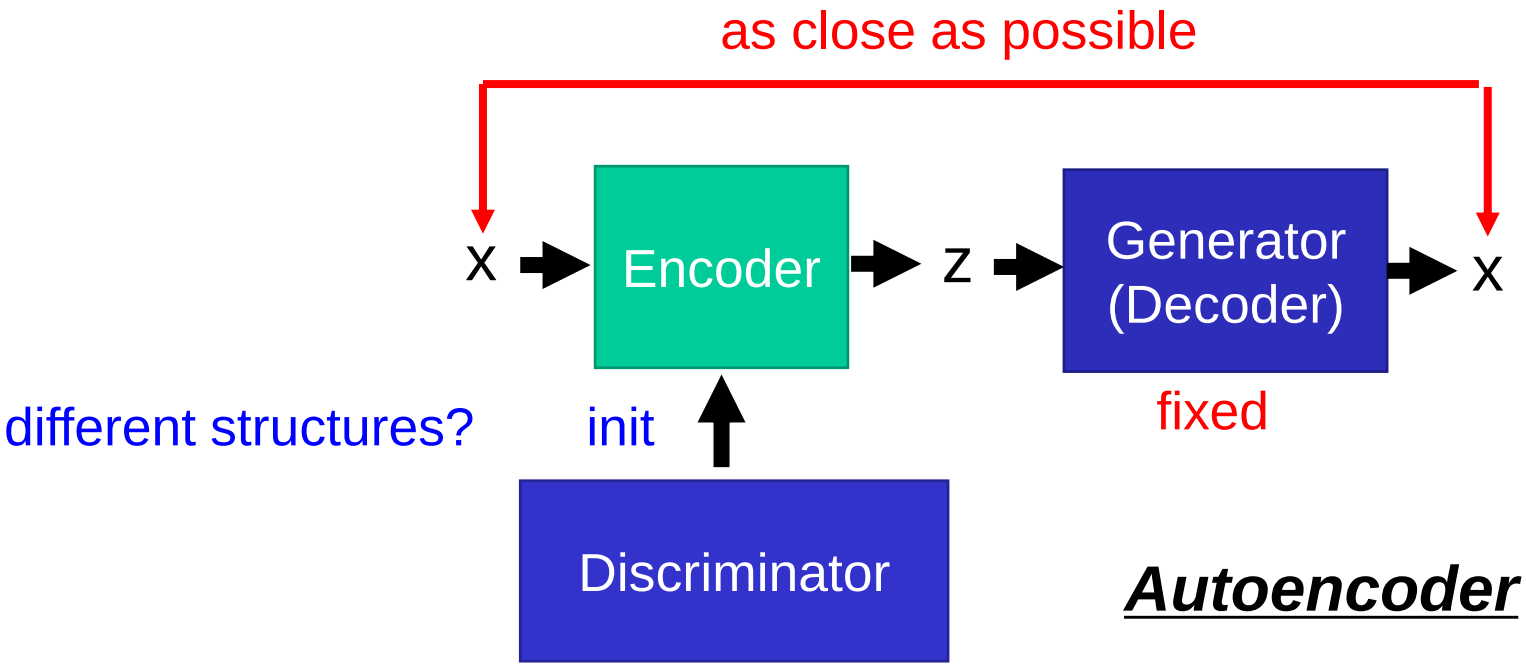
Modifying Input Code



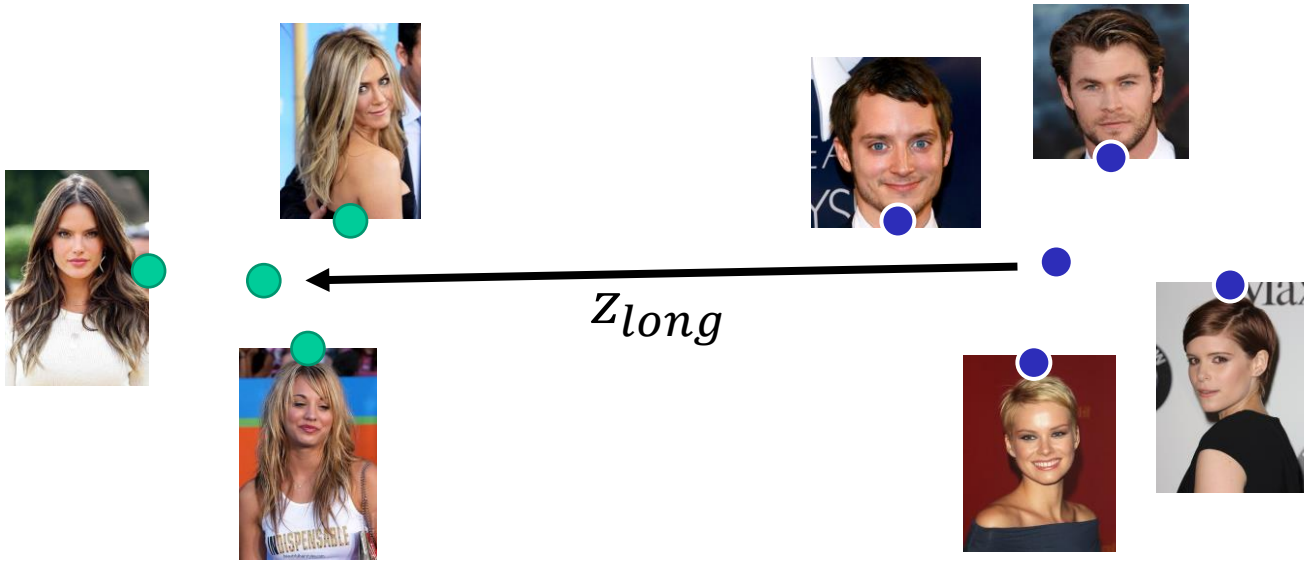
- The input code determines the generator output.
- Understand the meaning of each dimension to control the output.

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



$$z_{long} = \frac{1}{N_1}\sum_{x \in long} En(x) - \frac{1}{N_2}\sum_{x' \notin long} En(x')$$

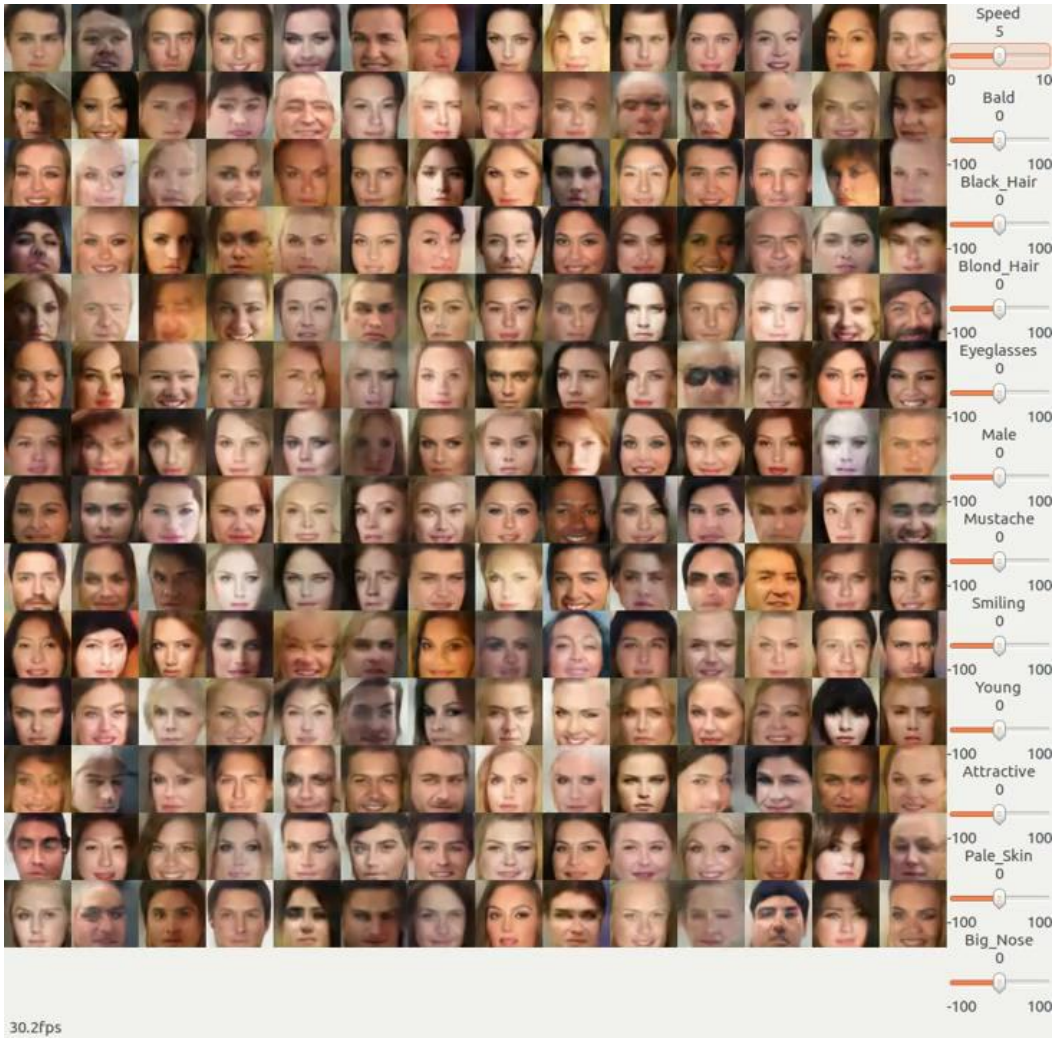
Short Hair

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

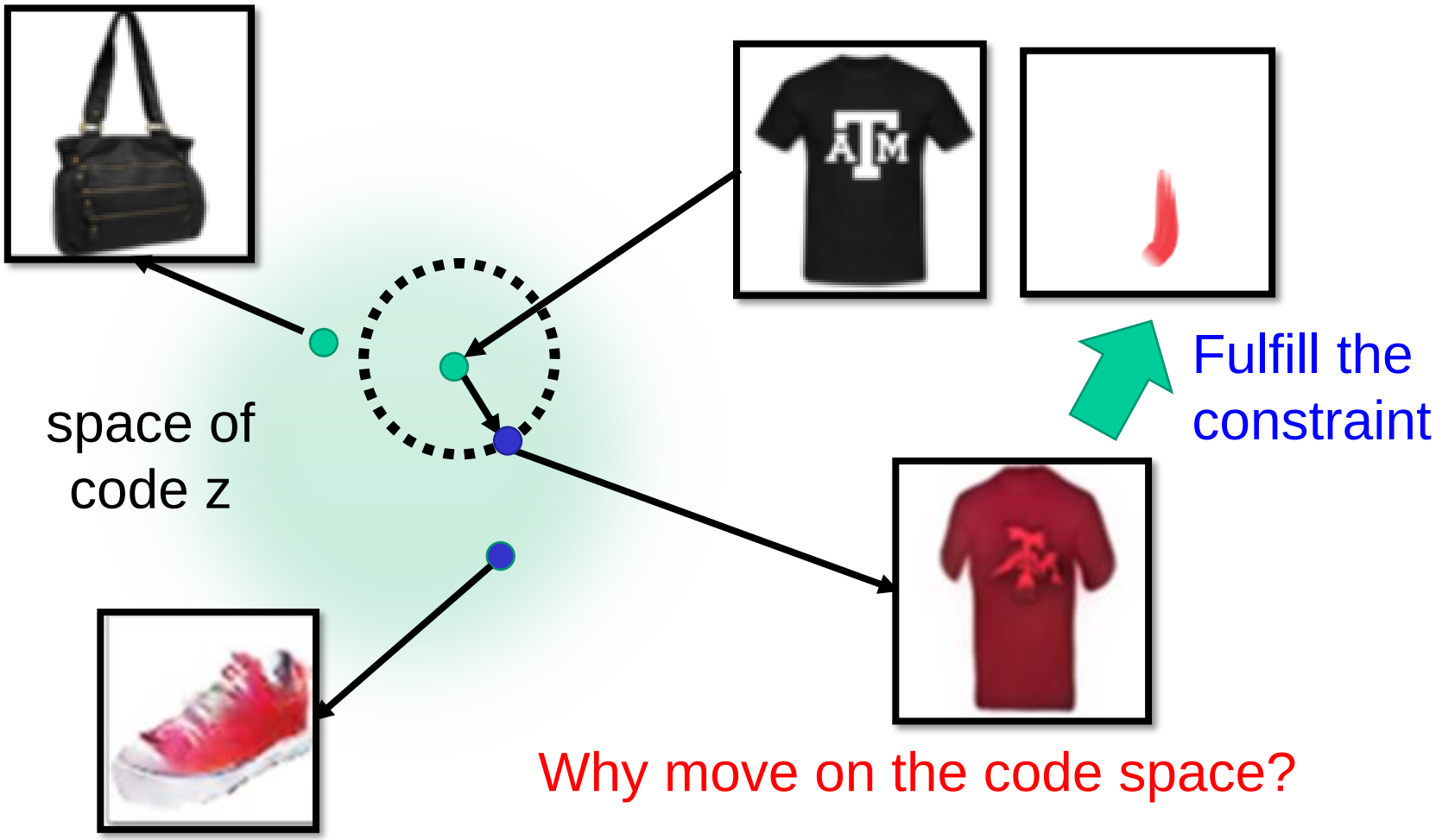
Long Hair

GAN + AE: attribute control

- Photo editing via attribute control

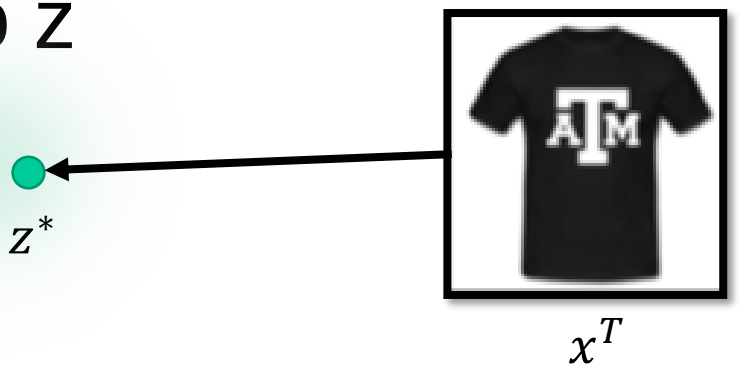


Attribute editing: basic idea



Back to z

- ***Method 1***



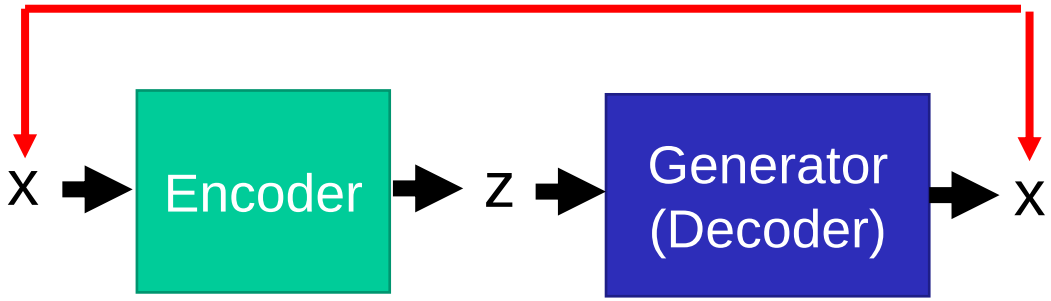
$$z^* = arg\min_z L\left(G(z), x^T\right)$$

Gradient Descent

Difference between G(z) and x^T

- Pixel-wise
- By another network

- ***Method 2***



- ***Method 3***

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

Editing Photos

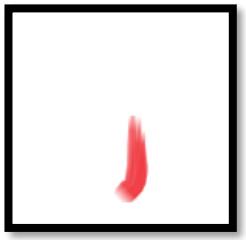
- z_0 is the code of the input image

Using discriminator to check the image is realistic or not

image

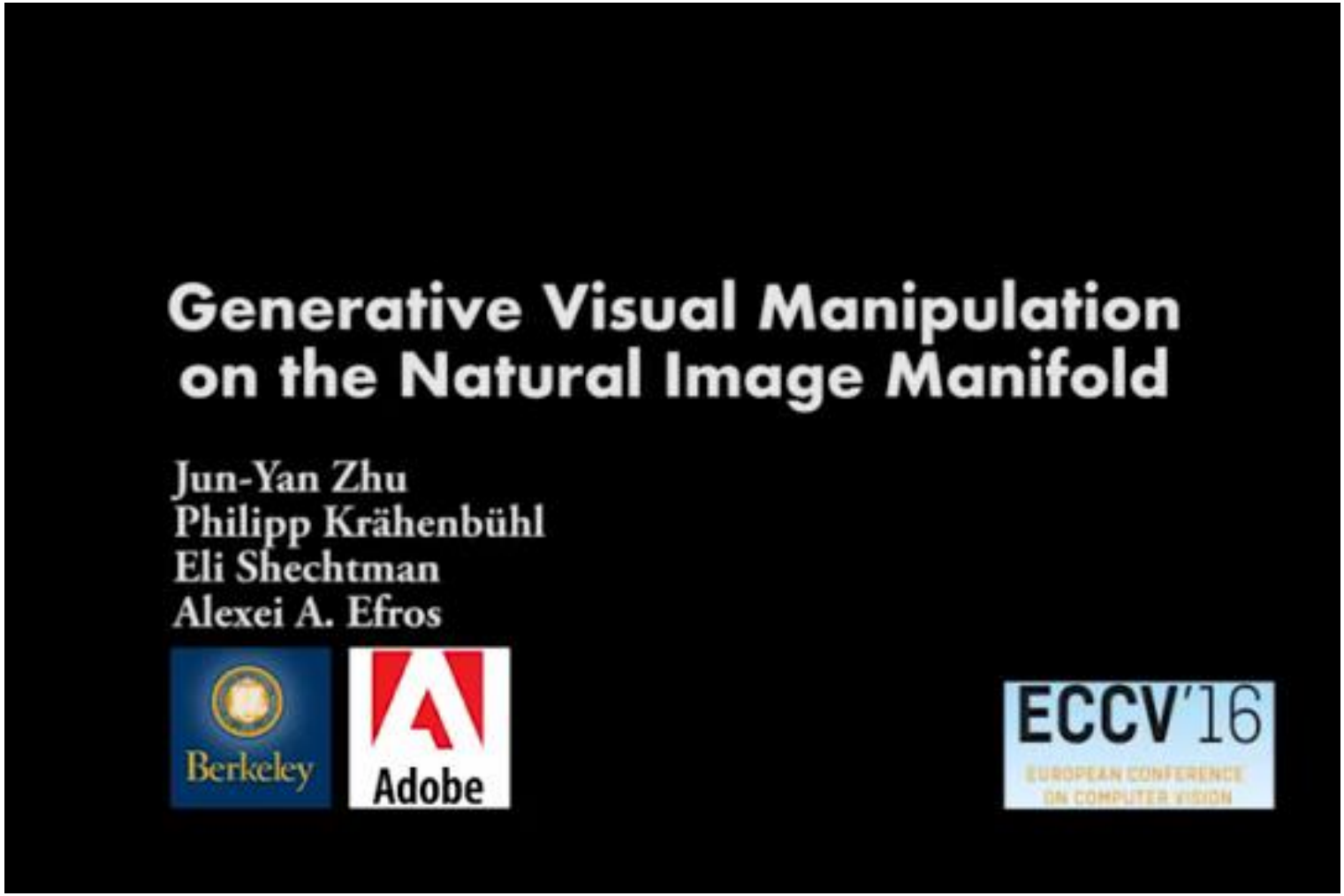
$$z^* = arg\min_z U\big(G(z)\big) + \lambda_1 \|z - z_0\|^2 - \lambda_2 D\big(G(z)\big)$$

Not too far away from the original image



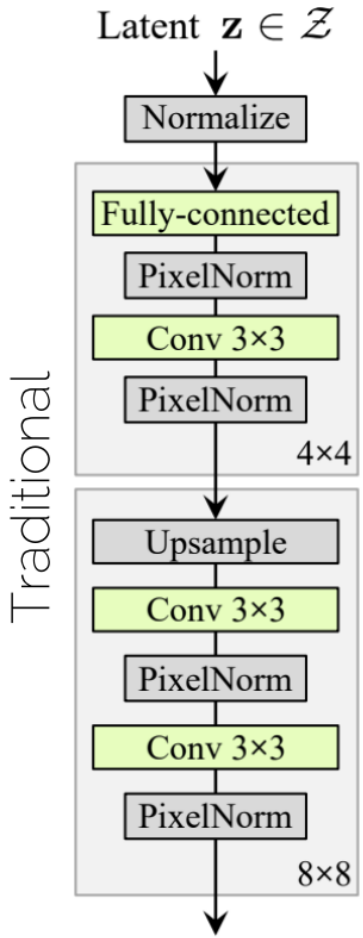
Does it fulfill the constraint of editing?

GAN + AE: attribute editing



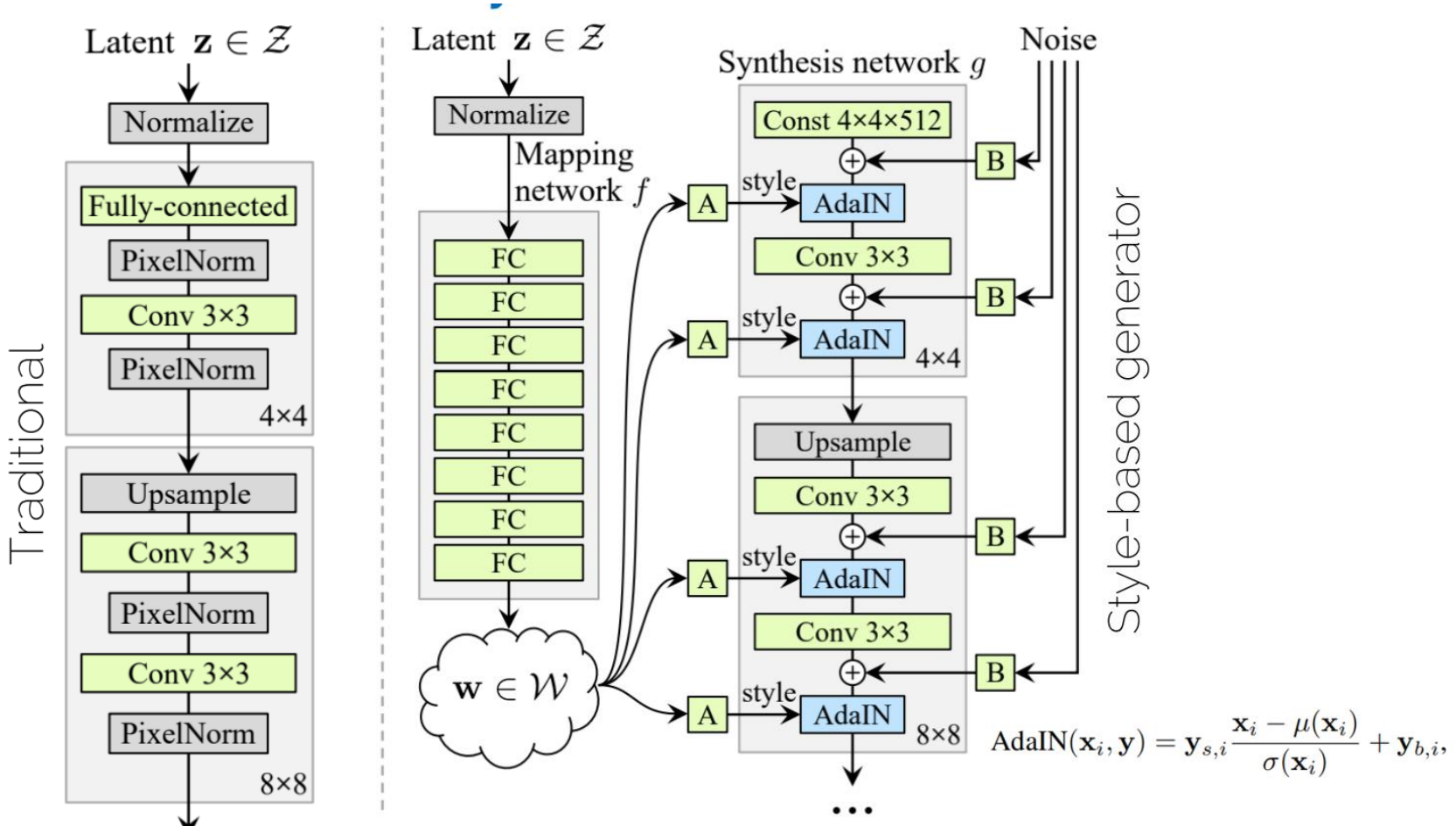
[Zhu et al. 2016] Generative Visual Manipulation on the Natural Image Manifold

StyleGAN



[Karras et al. 19]: StyleGAN

StyleGAN



[Karras et al. 19]: StyleGAN

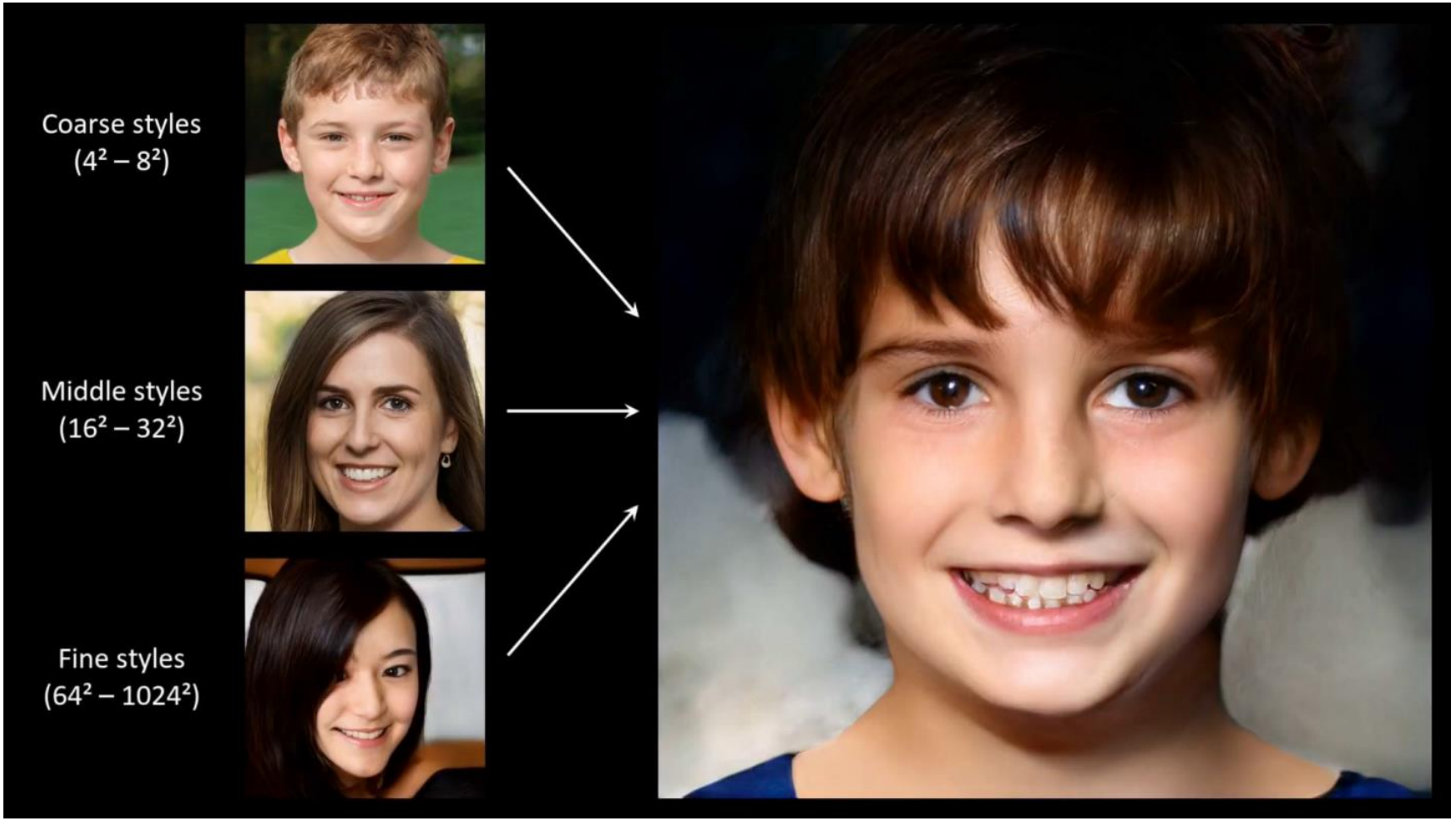
StyleGAN

- 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

StyleGAN

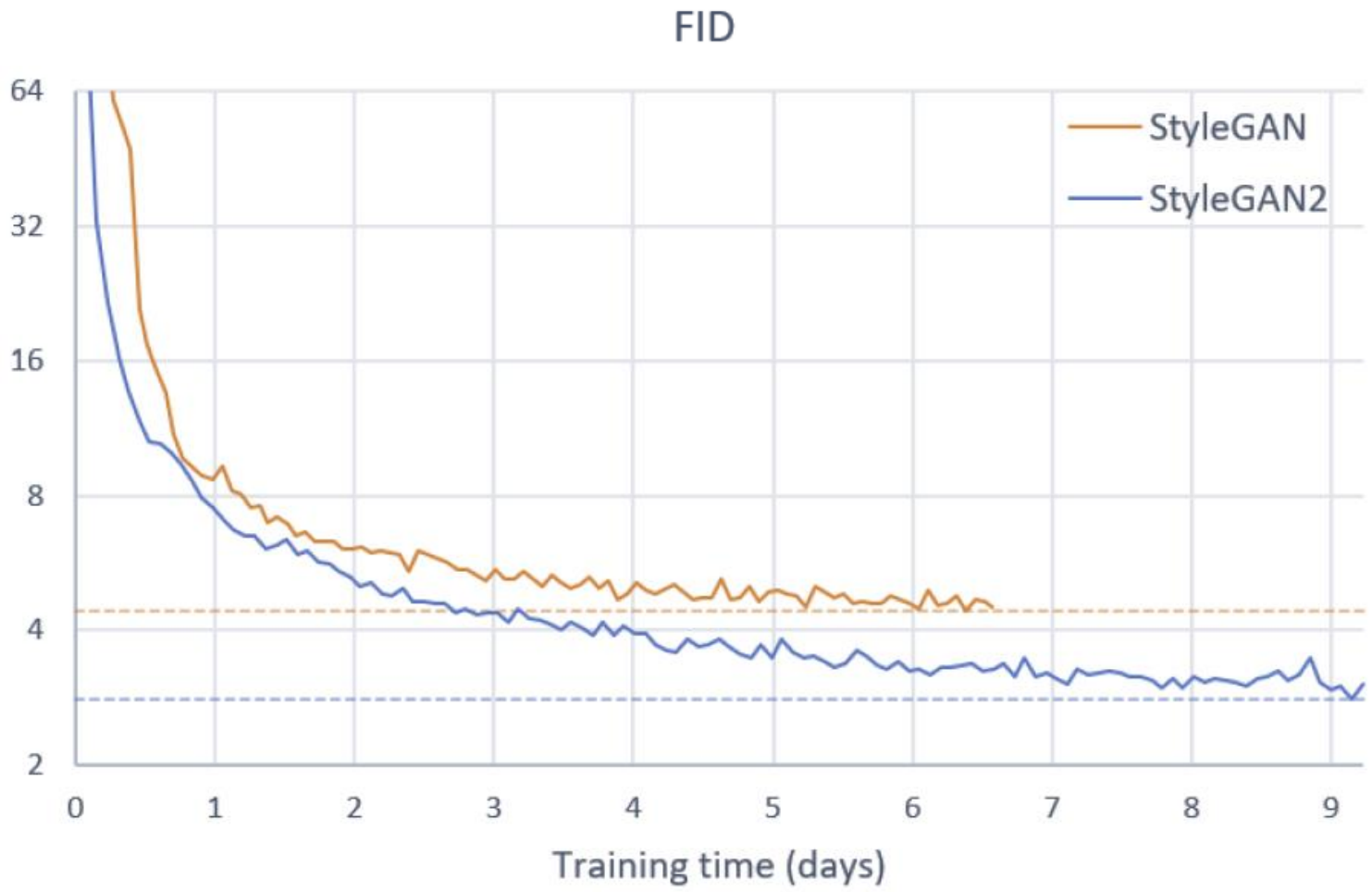


[Karras et al. 19]: StyleGAN

StyleGAN2

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

StyleGAN v.s StyleGAN2

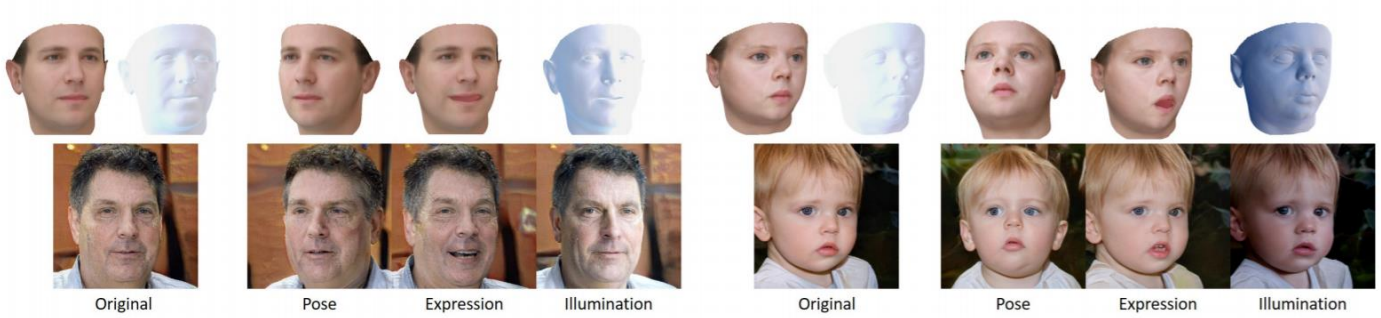


StyleGAN for style-transfer

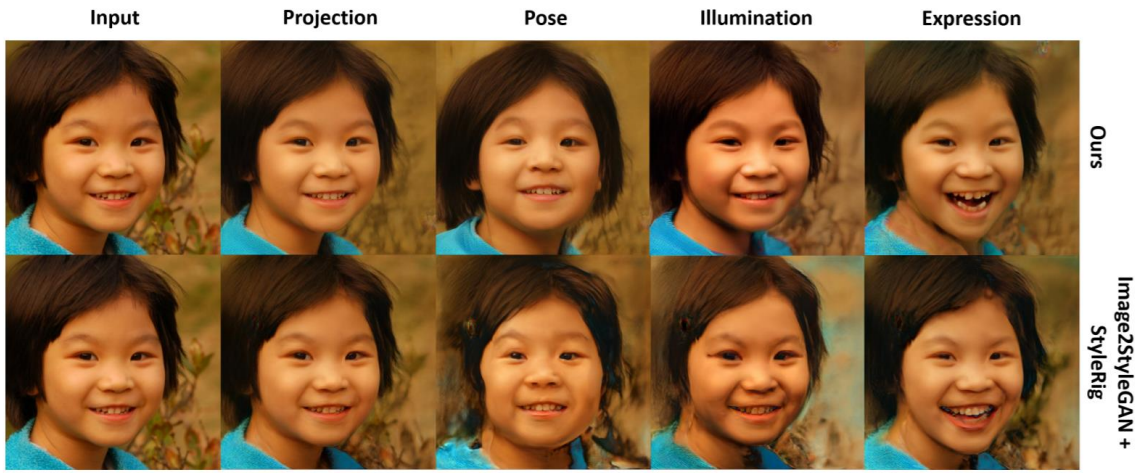
- Keep basic and medium levels
- Fine-tune it on the fine leves

Latent space of StyleGAN

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

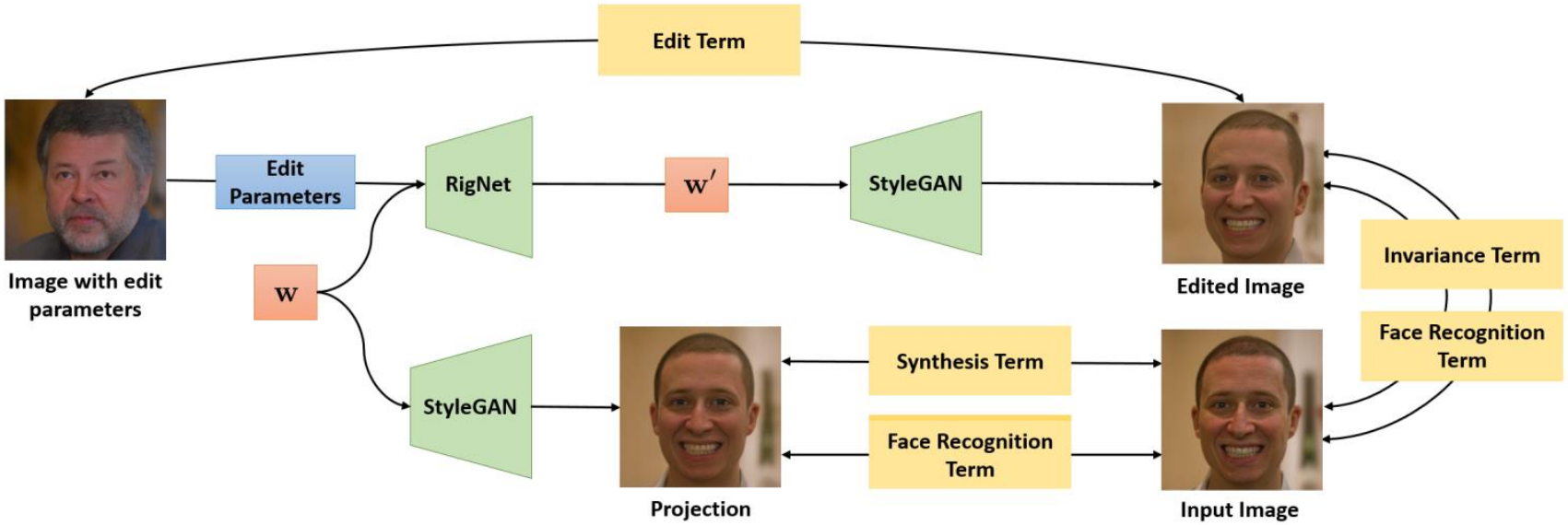


- PIE: Portrait Image Embedding for Semantic Control
  - Embedding in the latent space to edit real images



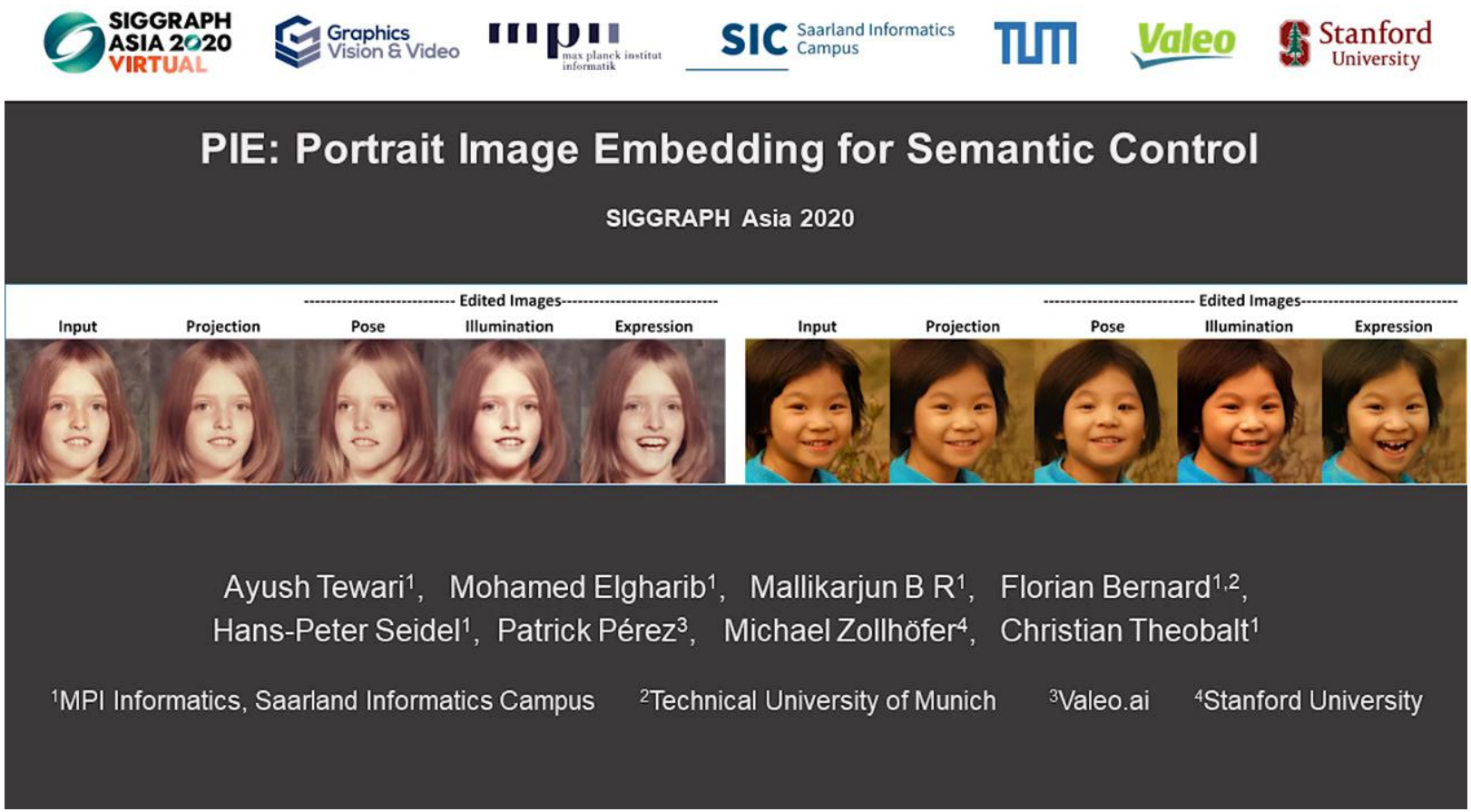
PIE: Portrait Image Embedding

- Utilize pretrained StyleGAN



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

PIE: Portrait Image Embedding

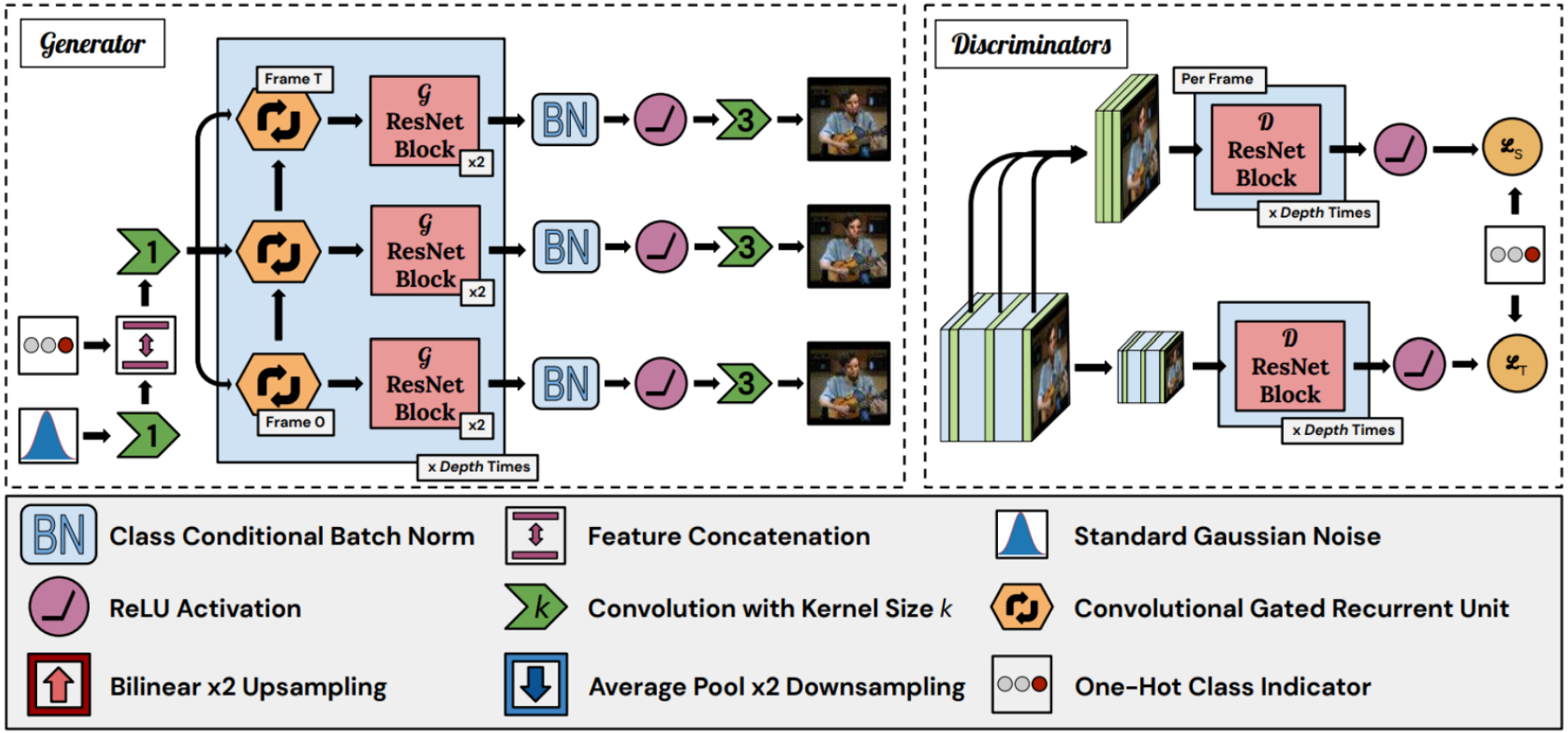


[TEWARI et al. 19] PIE: Portrait Image Embedding for Semantic Control

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)

GANs on Videos: DVD-GAN

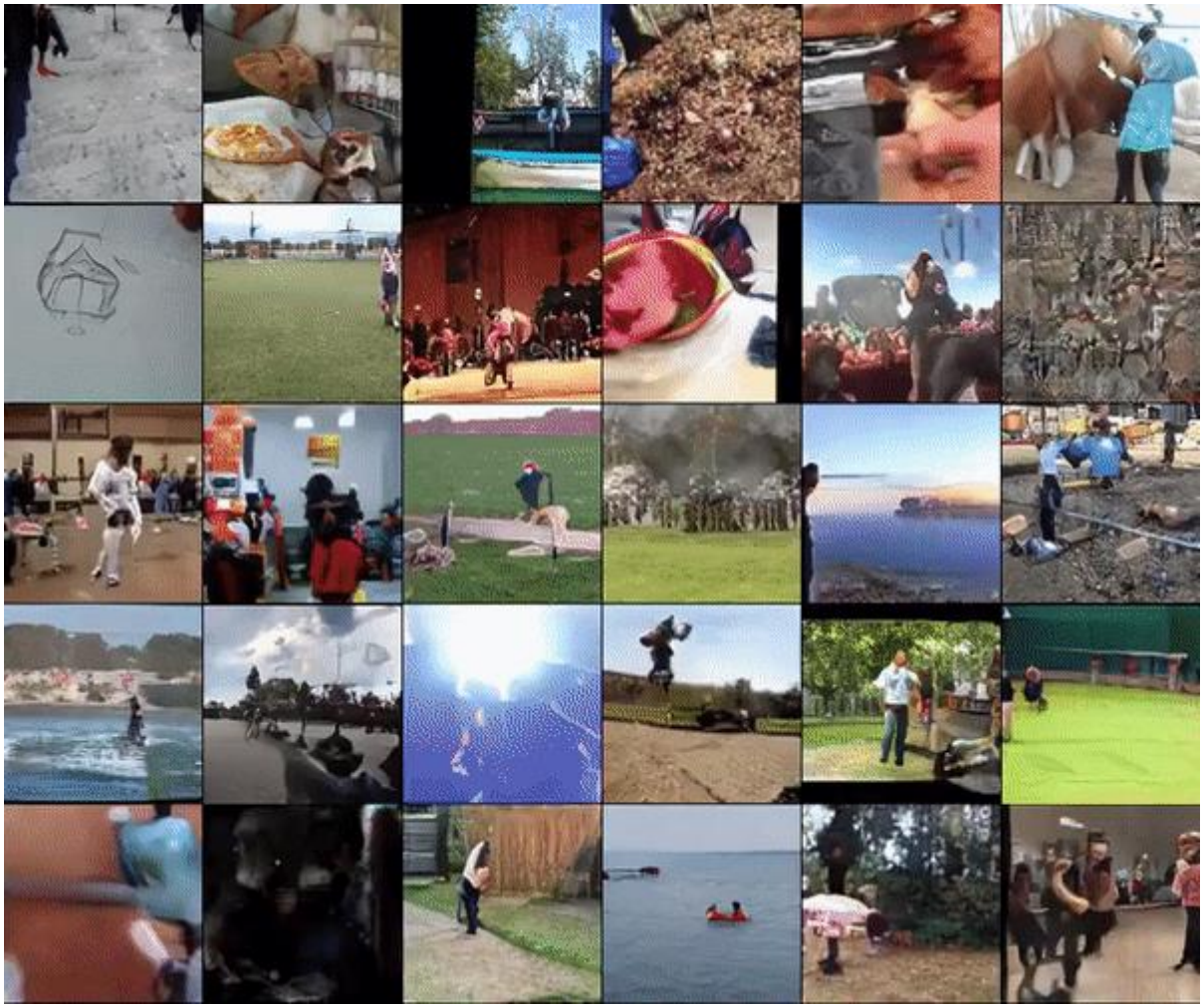


[Clark et al. 2019] Adversarial Video Generation on Complex Datasets

GANs on Videos: DVD-GAN

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

GANs on Videos: DVD-GAN



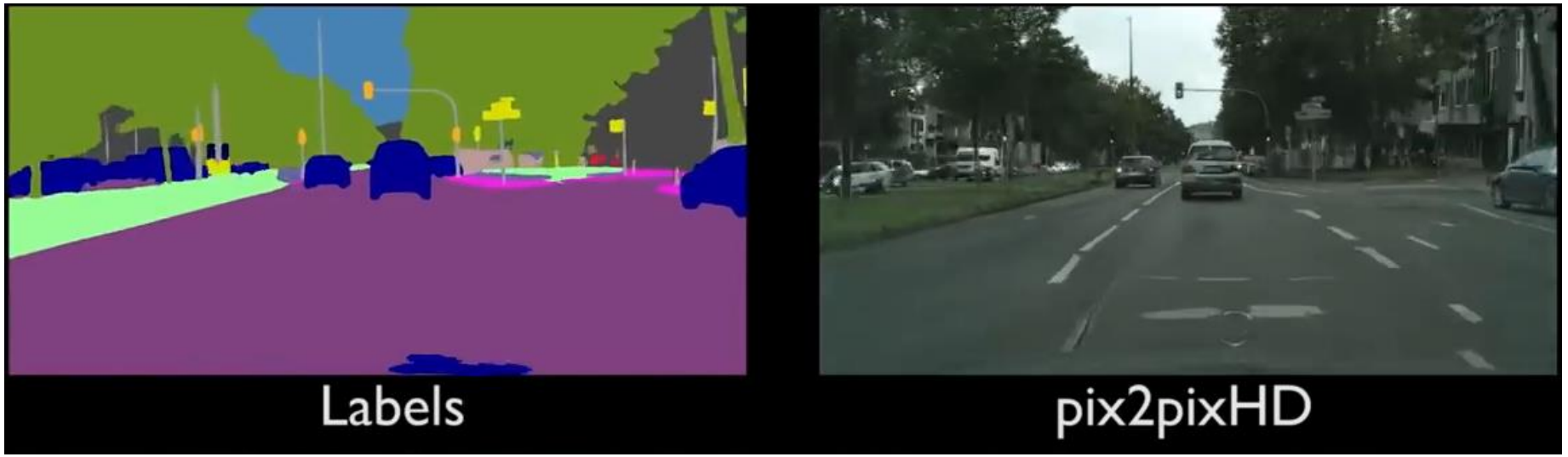
[Clark et al. 2019] Adversarial Video Generation on Complex Datasets

GANs on Videos: DVD-GAN

- 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

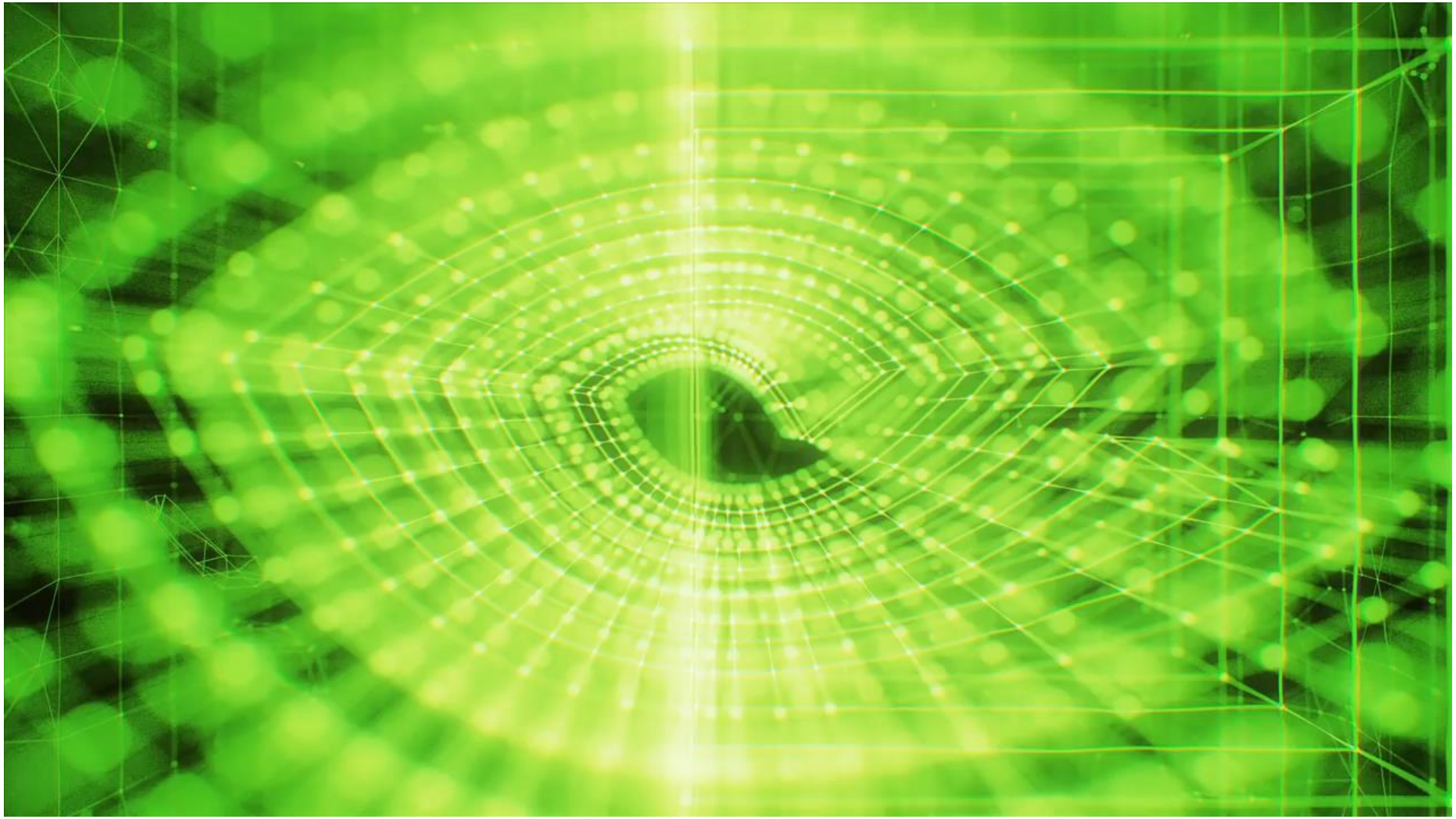
(Set L = 2)

- Conditional Image Discriminator D_i (is it real image)
- Conditional Video Discriminator D_v (temp. consistency via flow)
- Full Learning Objective:

$$\min_F \Big(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \Big) + \lambda_W \mathcal{L}_W(F)$$

[Wang et al. 2018] Vid2Vid

Video-to-Video Synthesis



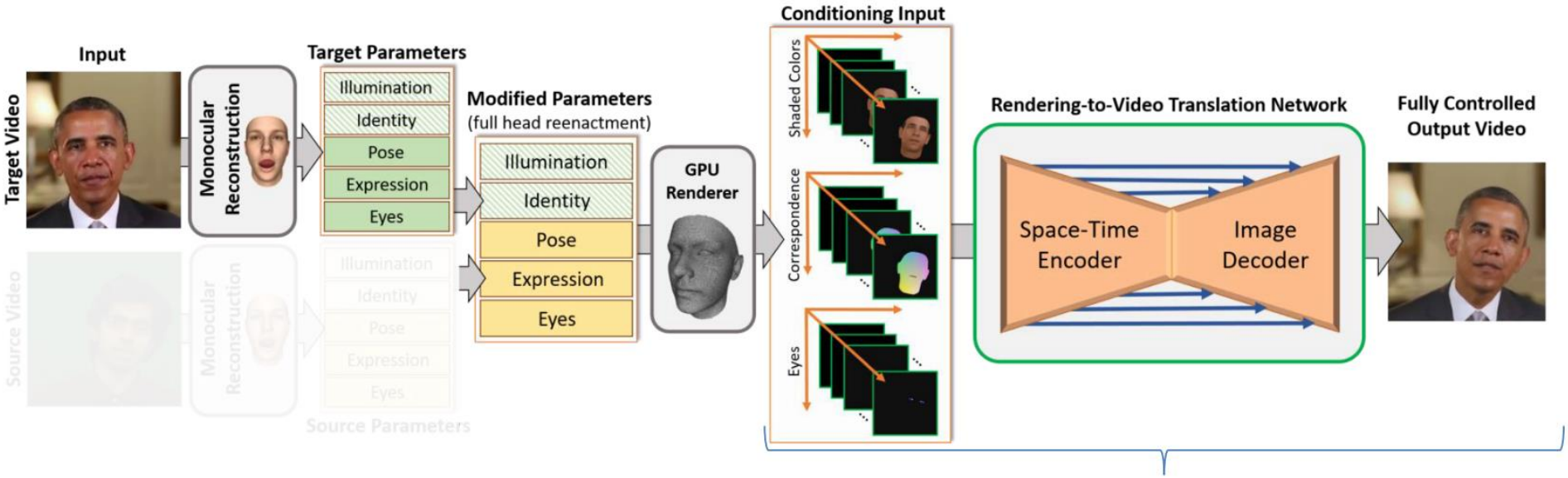
[Wang et al. 2018] Vid2Vid

Video-to-Video Synthesis

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

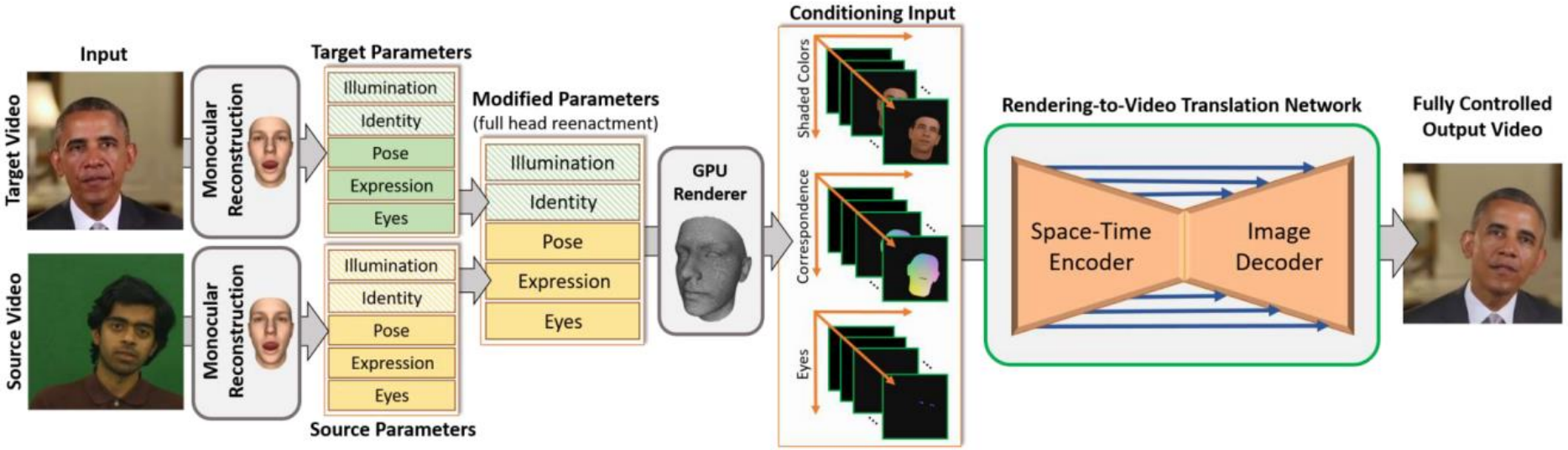
Deep Video Portraits



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

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

Deep Video Portraits



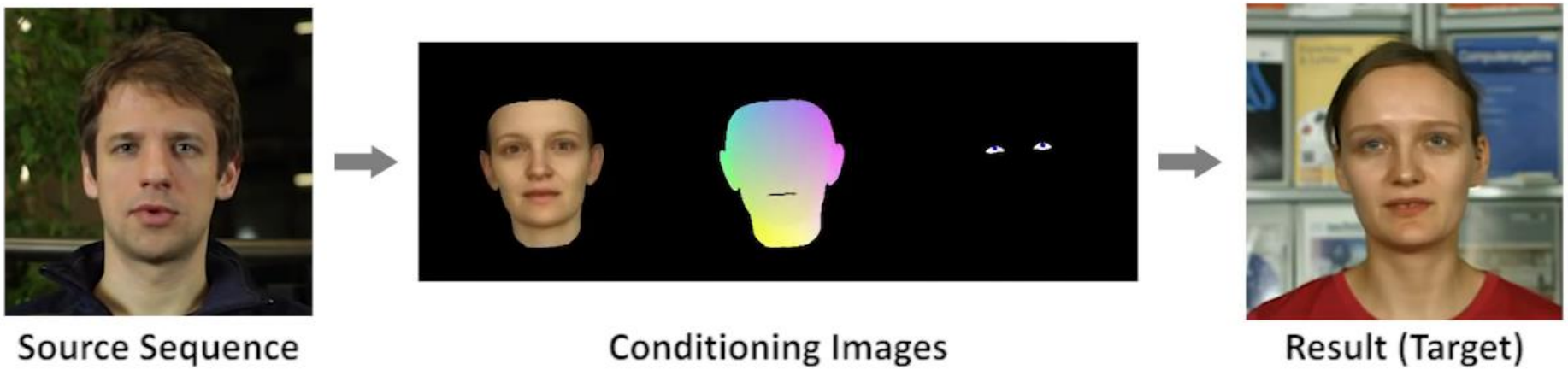
Siggraph'18 [Kim et al 2018]: Deep Portraits

Deep Video Portraits

Neural Network converts synthetic data to realistic video

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

Deep Video Portraits



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

Deep Video Portraits

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

Deep Video Portraits



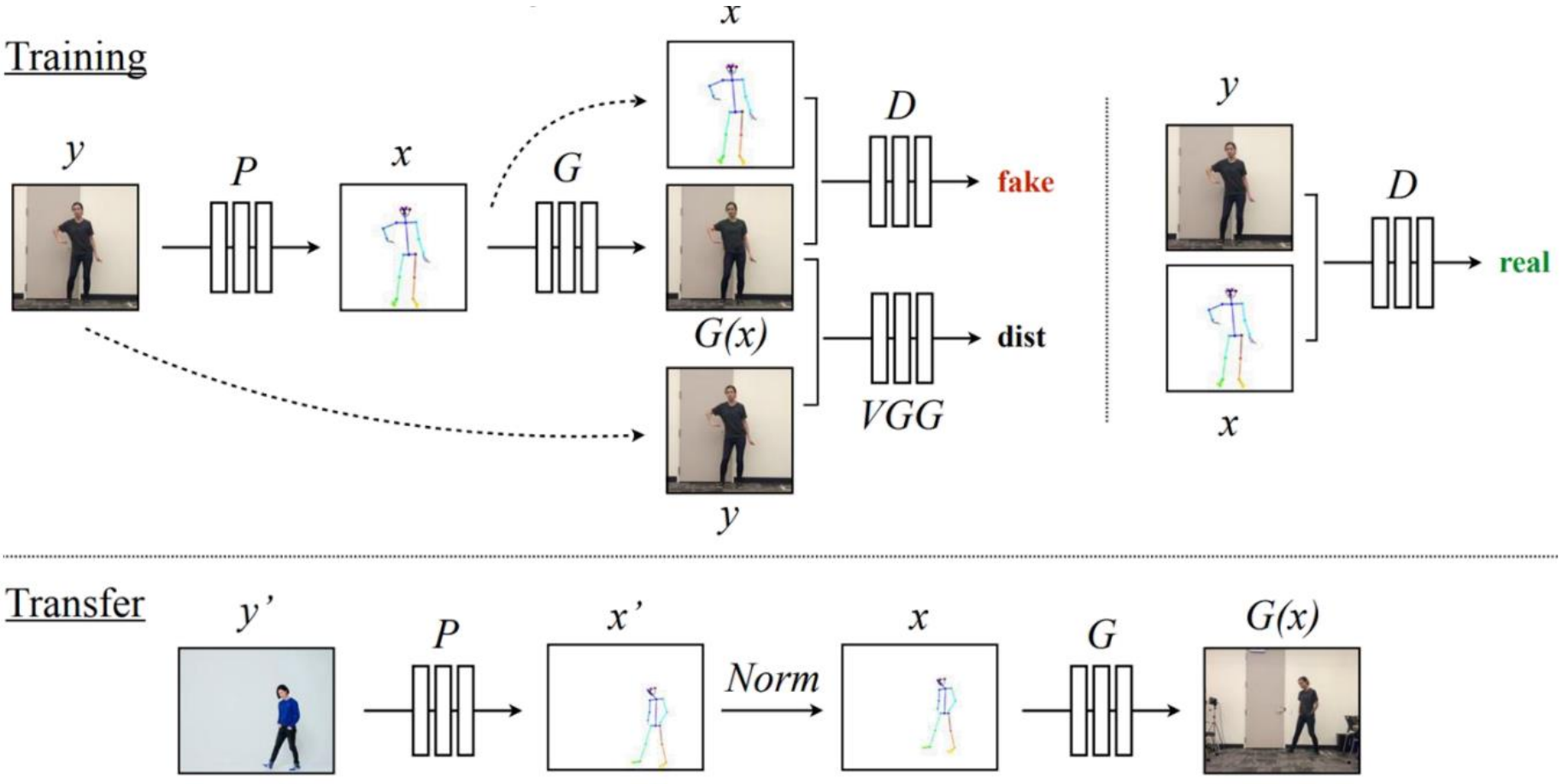
Siggraph'18 [Kim et al 2018]: Deep Portraits

Deep Video 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

Everybody Dance Now



[Chan et al. '18] Everybody Dance Now

Everybody Dance Now

[Chan et al. ’18] Everybody Dance Now

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

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
  - CycleGAN: Image-to-image translation with unpaired data
  - GANs on videos
- Next time
  - Diffusion

- Quiz8: send to wangchy8@shanghaitech.edu.cn
- Keep working on the projects!