11-695: AI Engineering
Conditional Generation

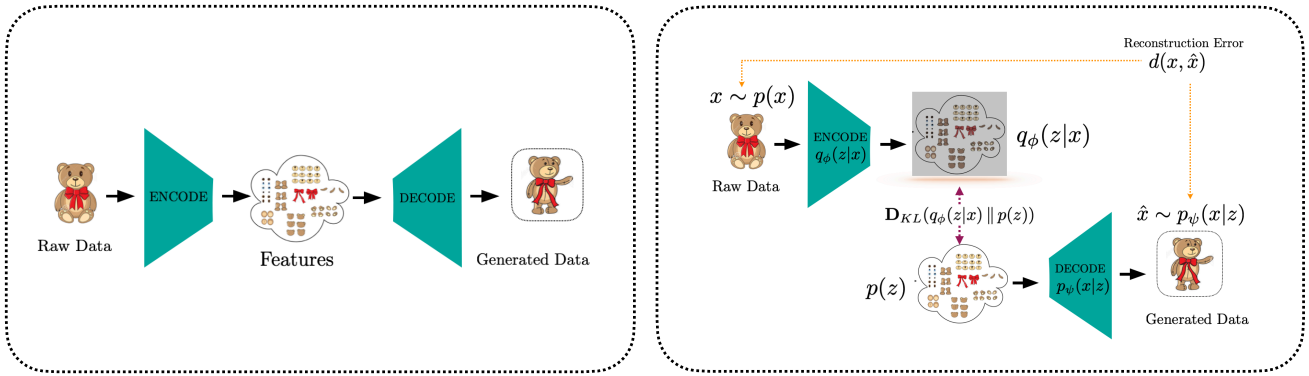
LTI/SCS

Spring 2020

Outline

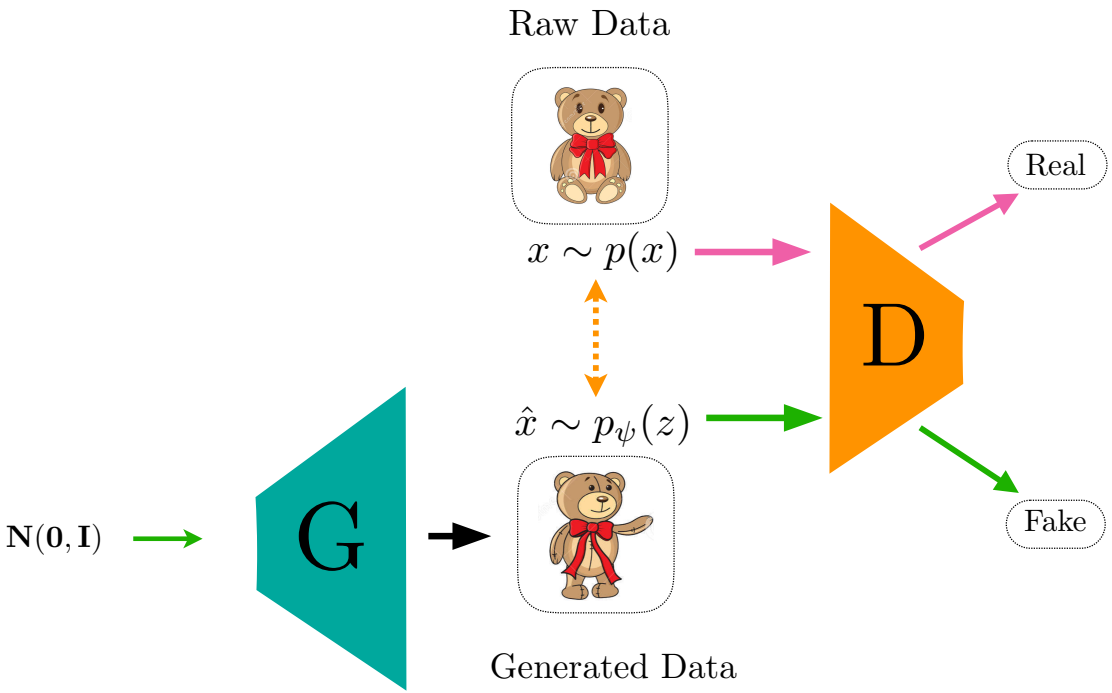
AutoEncoder Generation

Carnegie Mellon



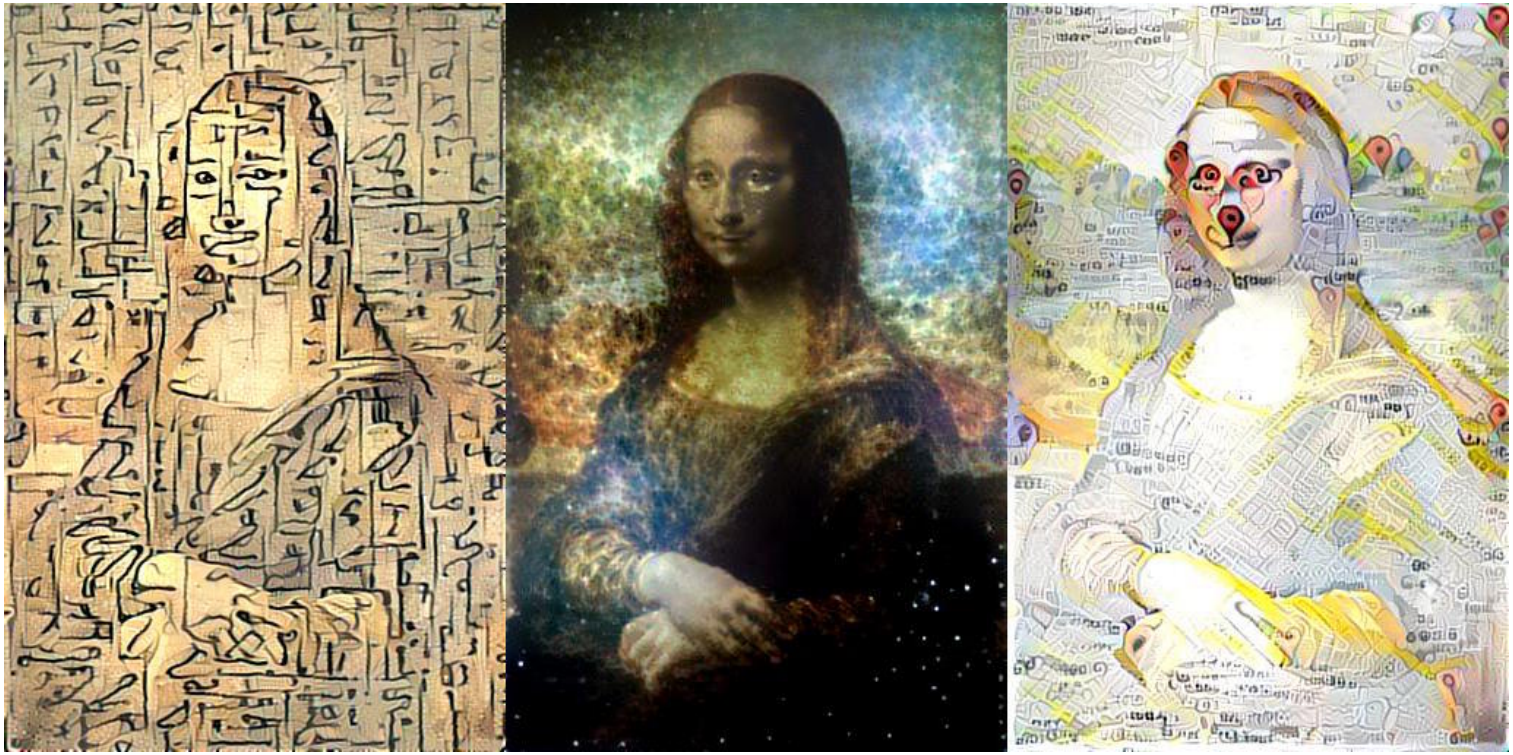
- Assume the model is well-trained. Take MNIST as an example
- If we want to generate digit 1
  - AE can do with feeding 1 into E, but it's not generative
  - VAE: we need to sample from random noise → no control!

GAN Generation



- Same as VAE, we generate from a noise → no control!
- Upside: they both can generate completely new, cool images

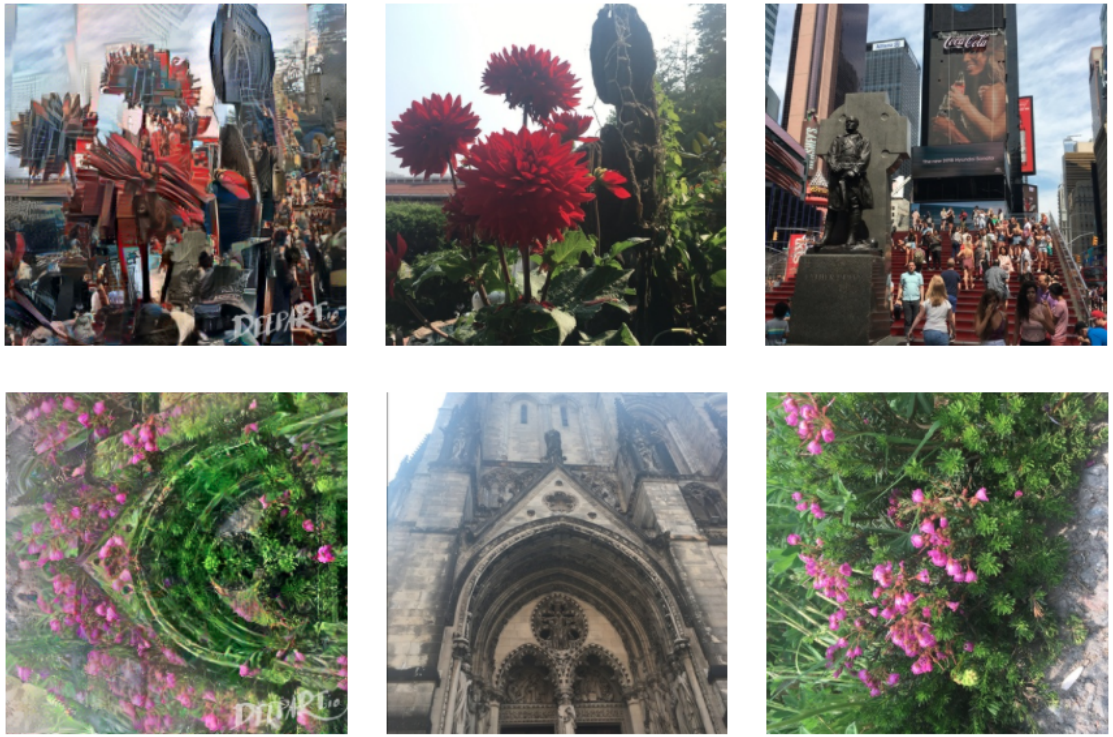
We Want More for Outputs



- Can we generate images of digits from a single signal, *e.g.* from a scalar input?
- Can we adopt, say Van Gogh's style into images w/o formal drawing training?

Image credit: Gene Kogan

How?



- Can we change the style of input based on an random template?

Different Input and Output Domains?

Carnegie Mellon

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

- Can we generate description based on images?
- Or vice versa?

Image credit: Kevin Xu *et al.* 2016

Outline

Idea: Condition on Attributes/Styles

Carnegie Mellon

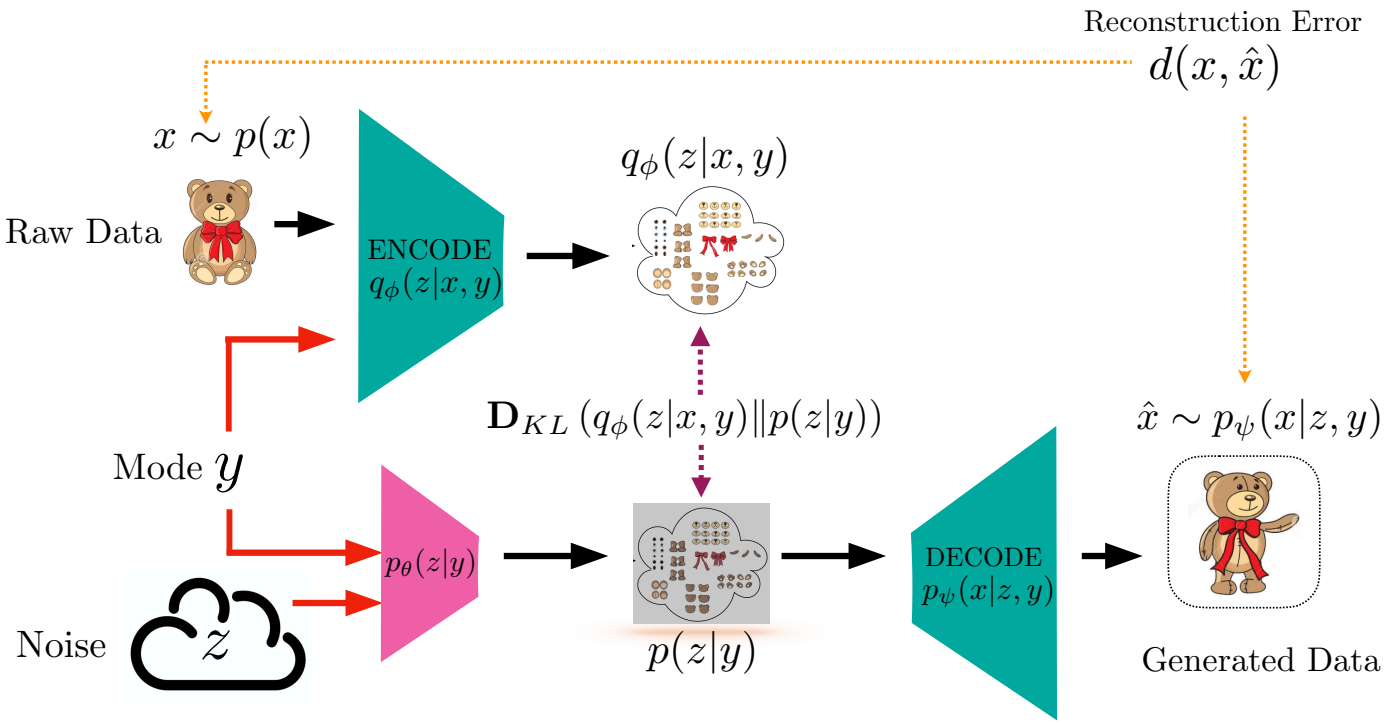
- So far: we generate from a latent representation

$$z \sim p(z), \qquad \hat{x} = p(x|z) \approx G_\psi(z) \tag{1}$$

- Now, we also have an attribute y as part of input:

$$z \sim p(z|y), \qquad \hat{x} = p(x|z, y) \approx G_\psi(z, y) \tag{2}$$

Conditional VAE (CVAE[1] [2])



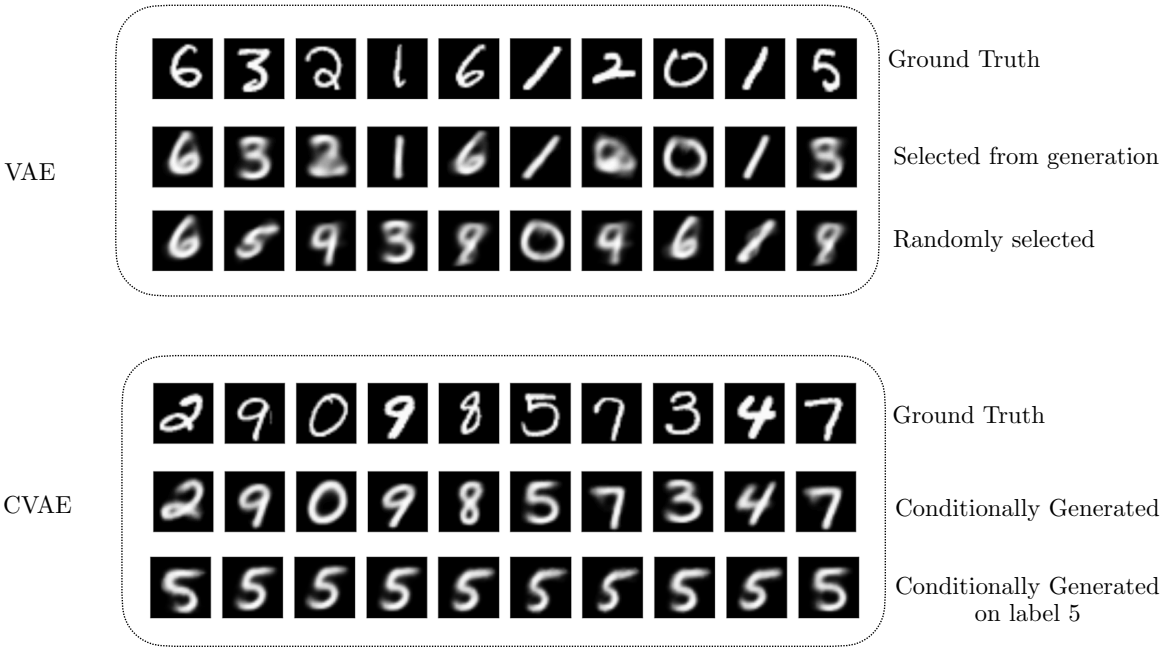
- VAE objective: $\text{ELBO} = \mathbf{E}_{q_\phi(z|x)}[\log p(x|z)] - \mathbf{D}_{KL}(q_\phi(z|x)\|p(z))$
- Now CVAE:

$$\text{ELBO} = \mathbf{E}_{q_\phi(z|x,y)}[\log p(x|z,y)] - \mathbf{D}_{KL}(q_\phi(z|x,y)\|p(z|y)) \quad (3)$$

[1] https://arxiv.org/abs/1406.5298

[2] https://pdfs.semanticscholar.org/3f25/e17eb717e5894e0404ea634451332f85d287.pdf

CVAE Results

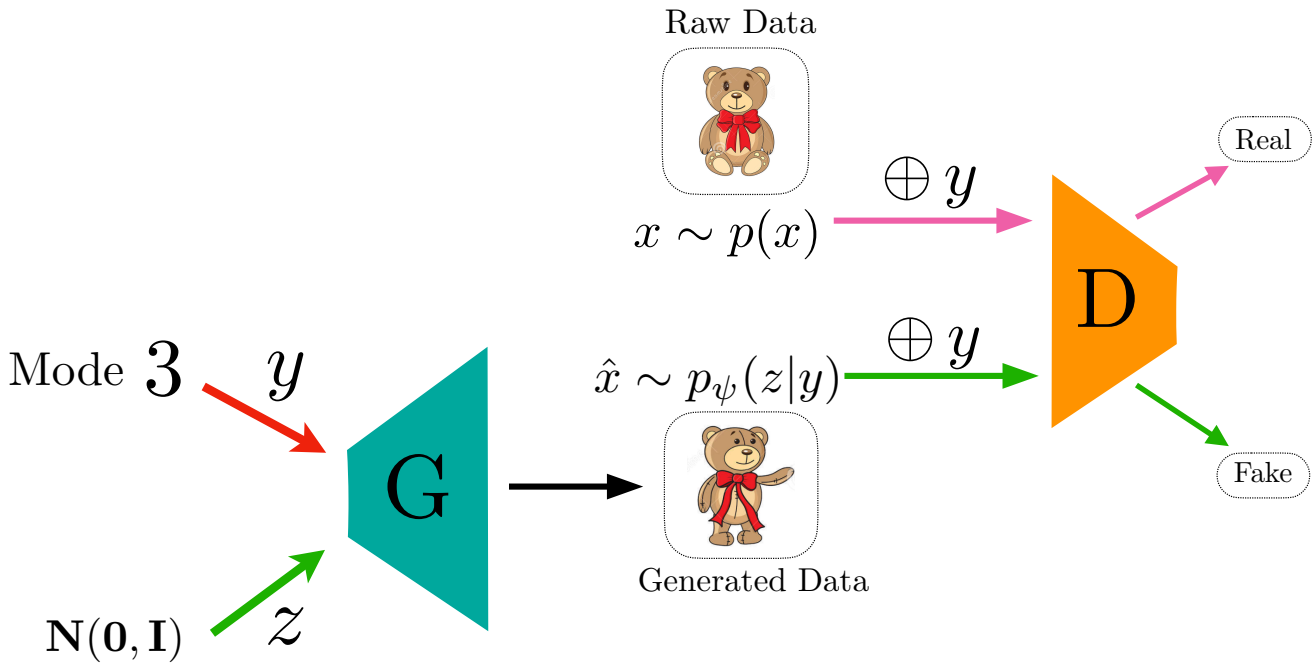


- How to edit VAE to make CVAE:
  - Fuse y and input to Encoder, *e.g.* as simple as concatenation
  - Fuse y and noise with learnable weights before Decoder
- Results seem sharper (we implicitly make use of labels)
- And we can now control the the modes of generation!

Image credit: Agustinus Kristiadi

Conditional GAN (CGAN[3]) - Similar!

Carnegie Mellon



- GAN: $\min_\psi \max_\theta \mathbb{E}_{x\sim p(x)}[\log D_\theta(x)] + \mathbb{E}_{z\sim p(z)}[1 - \log D_\theta(G_\psi(z)]$
- Now CGAN:

$$\min_\psi \max_\theta \mathbb{E}_{x\sim p(x)}[\log D_\theta(x|y)] + \mathbb{E}_{z\sim p(z|y)}[1 - \log D_\theta(G_\psi(z|y)] \quad (4)$$

[3] https://arxiv.org/pdf/1411.1784.pdf

CGAN Results

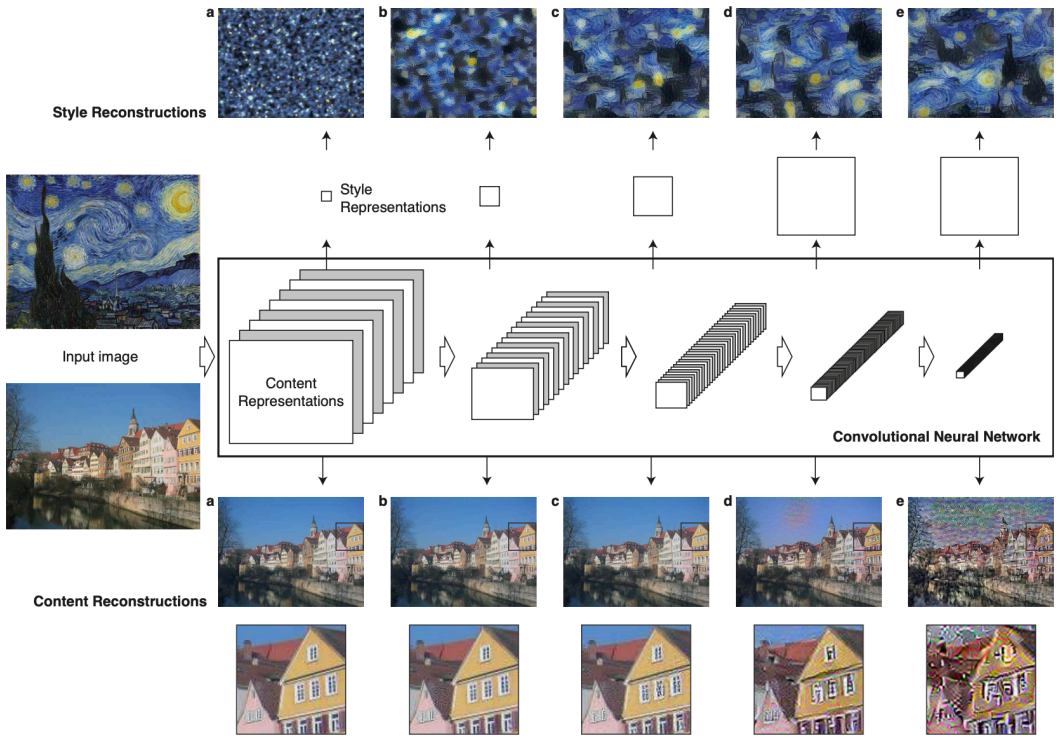
- How to edit GAN to make CGAN: similar to CVAE
- And also, we can now control the modes of generation

Image credit: Alec Radford *et al.* 2016

Outline

Neural Style Transfer[4]

Carnegie Mellon

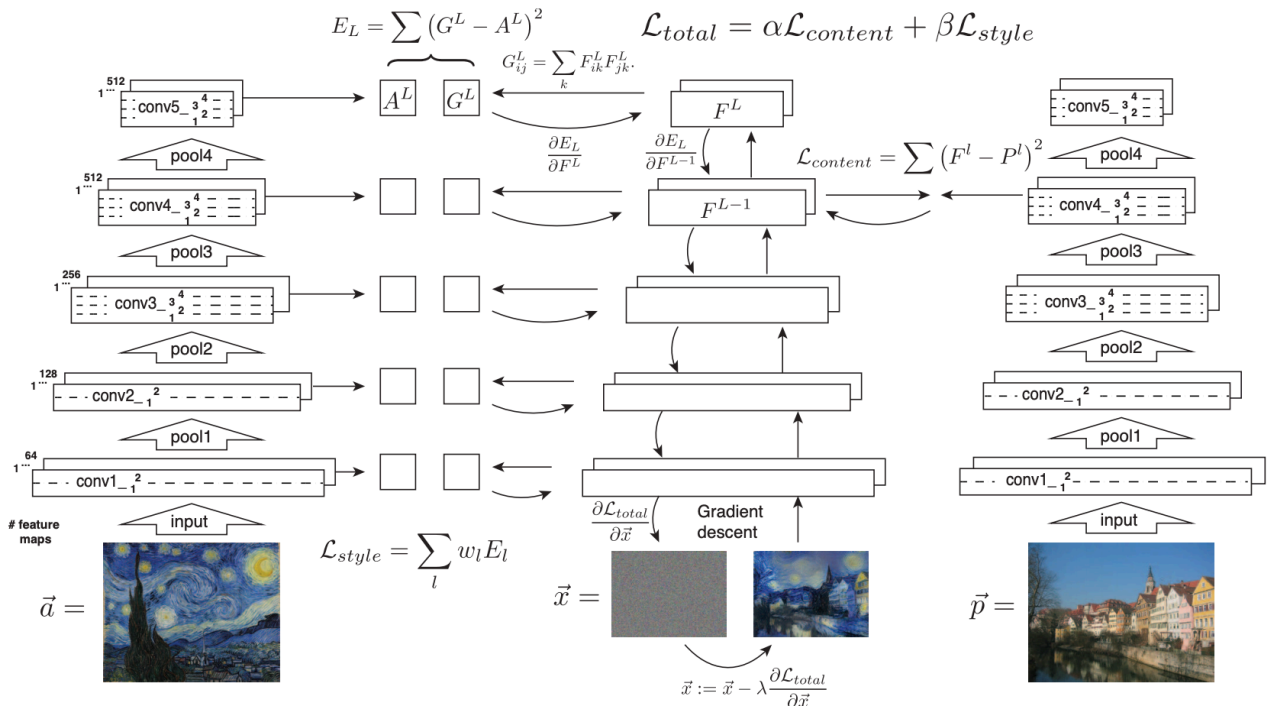


- Style Image an Content Image features are extracted by VGG

[4] https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf

Image credit: Leon A. Gatys et al. 2016

Neural Style Transfer using CNN

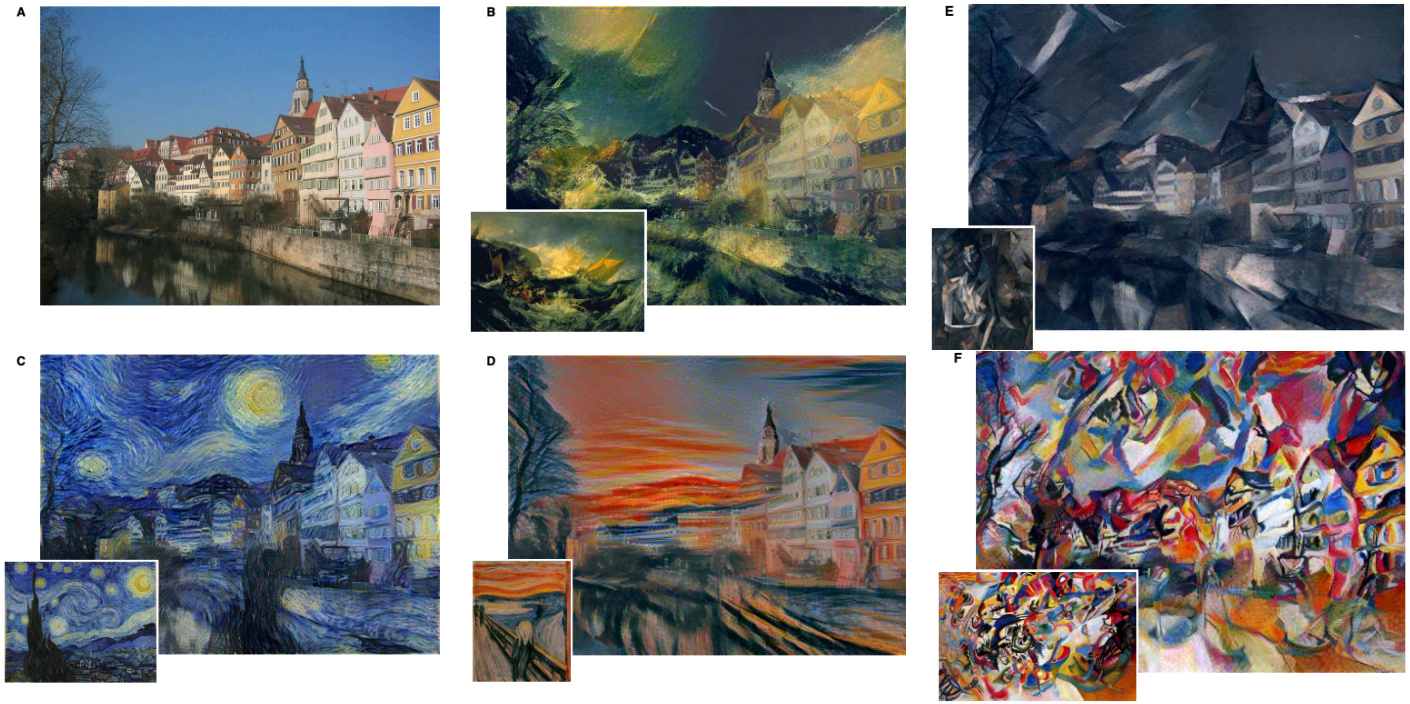


- White noise x is modeled into 2 components: content and style
- Each will have the respective loss
- Both losses are optimized together

Image credit: Leon A. Gatys *et al.* 2016

Neural Style Transfer using CNN

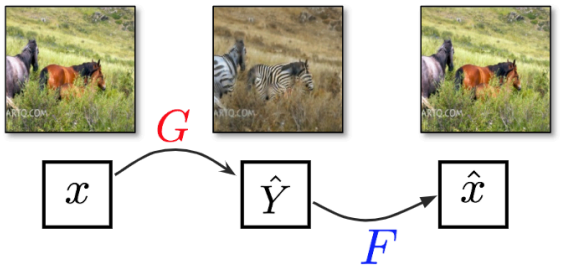
Carnegie Mellon

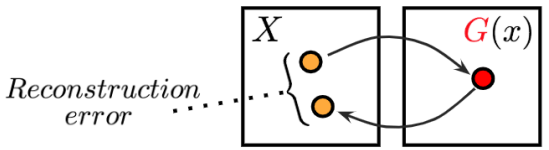


- Training converges when x matches content and style

Image credit: Leon A. Gatys *et al.* 2016

Cycle Consistency Loss



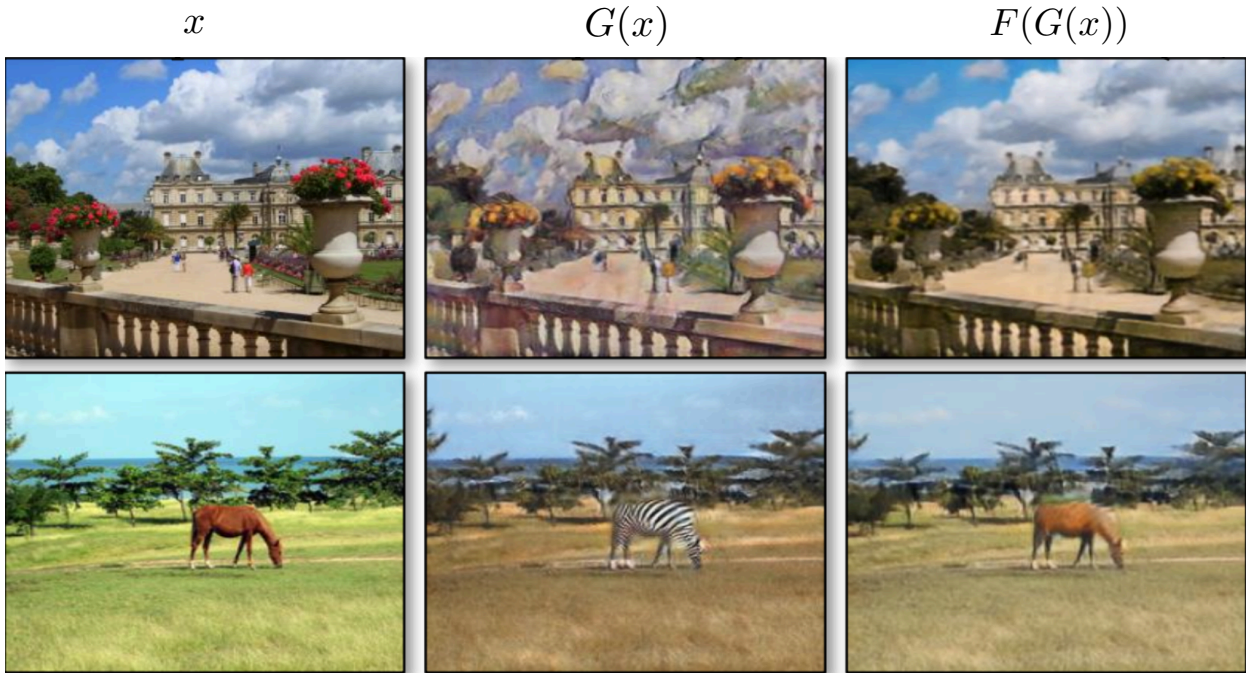


$$\|F(G(x)) - x\|_1$$

- Cycle loss is based on 2 phases: forward and back transfers
- Provides symmetric alignment intuition

Image credit: Jun-Yan Zhu *et al.* 2017 (Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks)

CycleGAN[5]



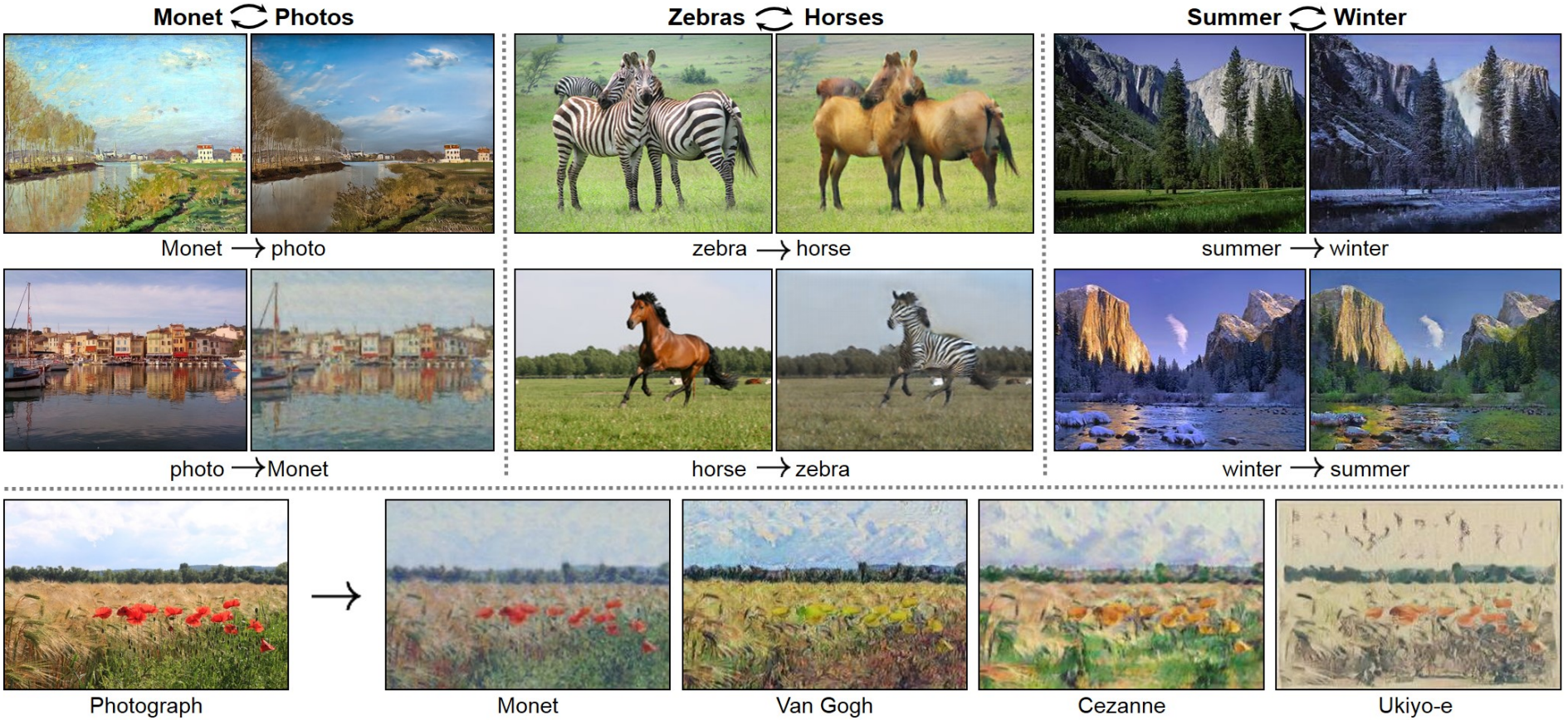
- 1 Generator,
- 2 Discriminators (1 per each style).
- Augmented by cycle loss

[5] https://arxiv.org/pdf/1703.10593.pdf

Image credit: Jun-Yan Zhu *et al.* 2017

CycleGAN

Carnegie Mellon

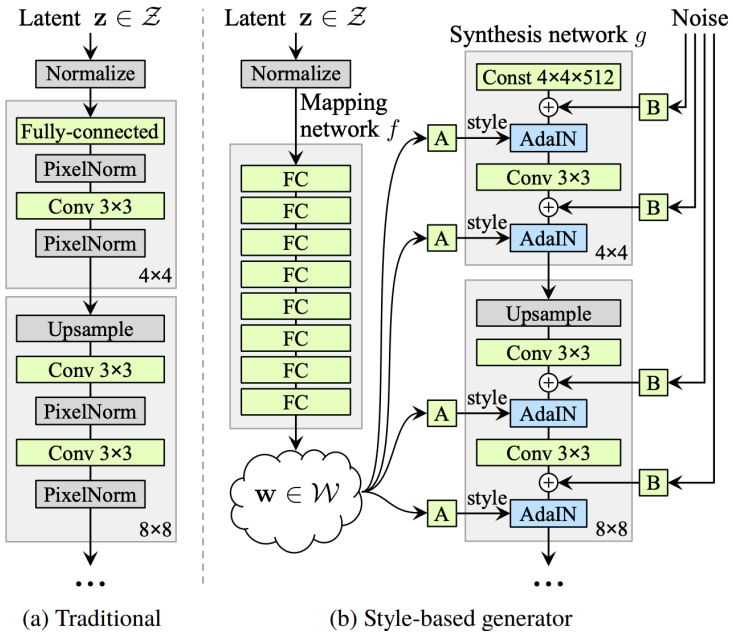


- Is also known as unpaired image-to-image translation

Image credit: Jun-Yan Zhu *et al.* 2017

StyleGAN[6]: Progressive GAN extended

Carnegie Mellon



(a) Traditional (b) Style-based generator

- 8-layer FNN mapping network f transforms z into styles w
- Each layer control each details: *e.g.* hair, pose, face, ...
- Adaptive instance norm (AdaIN) + noise are added to CNN layers
- G takes a constant instead of z directly

[6] https://arxiv.org/pdf/1812.04948.pdf

Image credit: Tero Karras *et al.* 2018

StyleGAN: Mixing Regulation

Carnegie Mellon

- Use two latent codes z_1 and z_2 during training
- Regulate contribution between them

Image credit: Tero Karras *et al.* 2018

StyleGAN: Noise and Results

Carnegie Mellon

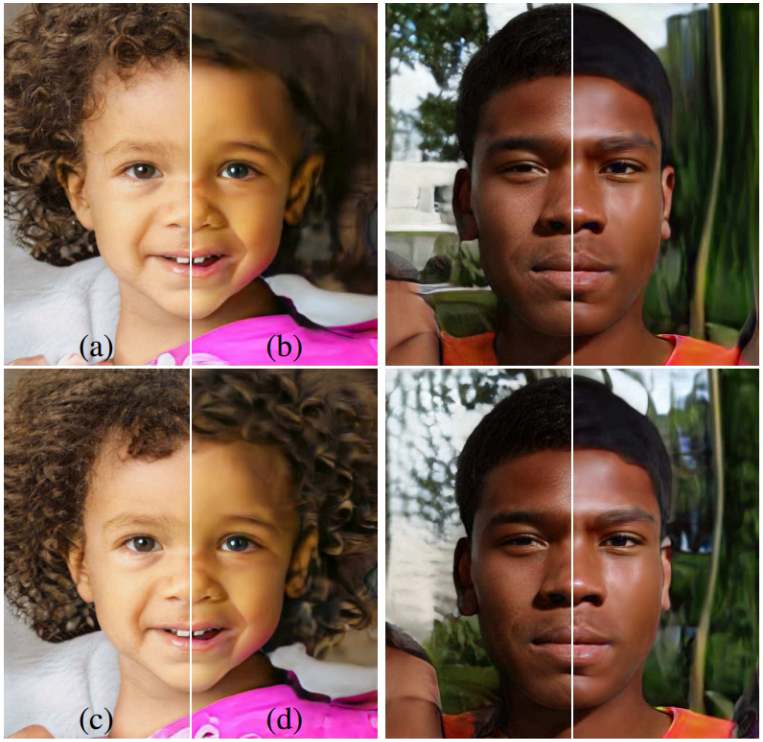
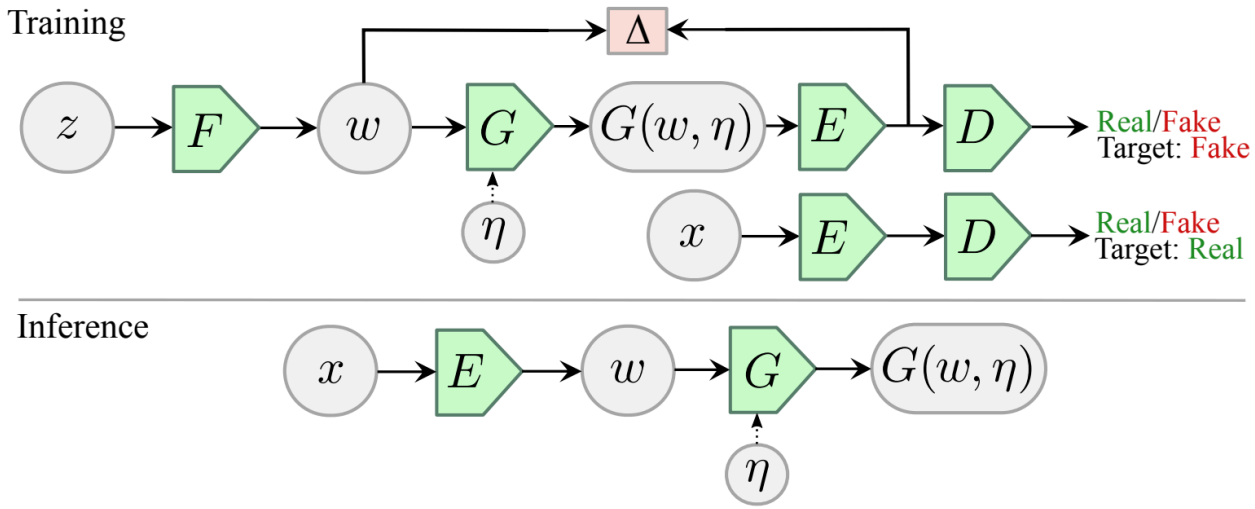


Figure 5. Effect of noise inputs at different layers of our generator. (a) Noise is applied to all layers. (b) No noise. (c) Noise in fine layers only (64^2 – 1024^2). (d) Noise in coarse layers only (4^2 – 32^2). We can see that the artificial omission of noise leads to featureless "painterly" look. Coarse noise causes large-scale curling of hair and appearance of larger background features, while the fine noise brings out the finer curls of hair, finer background detail, and skin pores.

Image credit: Tero Karras *et al.* 2018

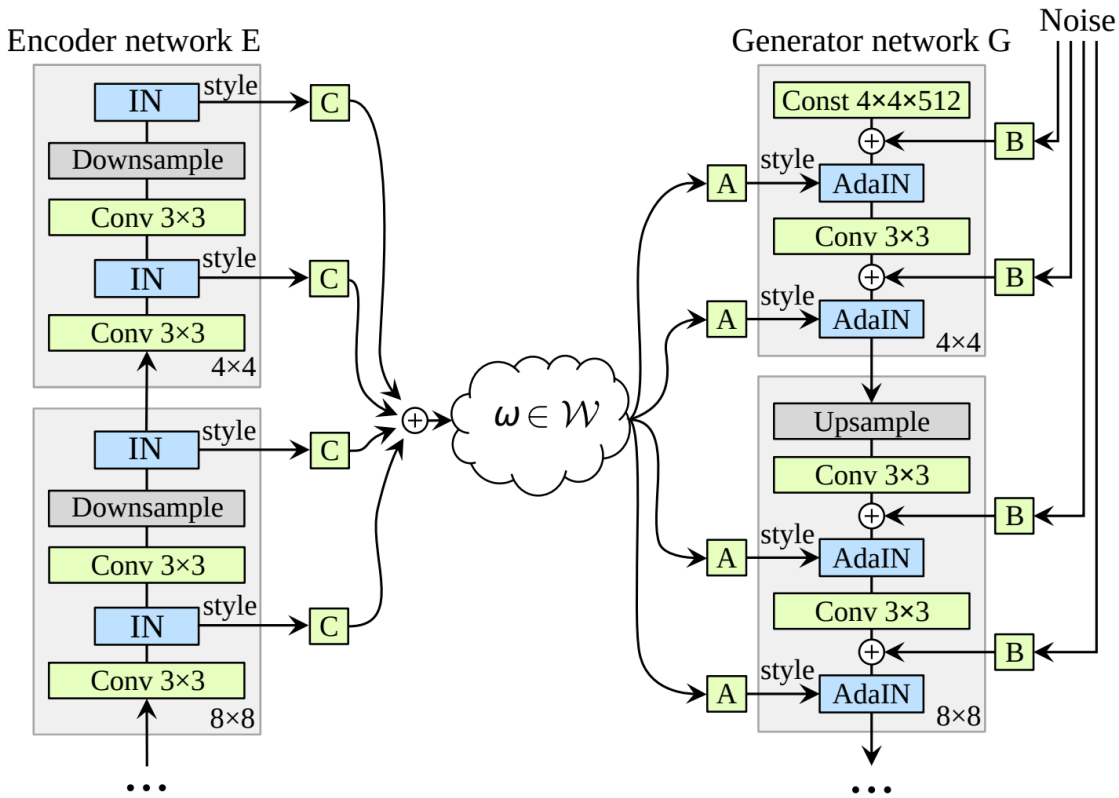
Adversarial Latent Autoencoders (ALAE)



- Idea: break GAN into smaller components
- Decompose: $\mathrm{G} = G \circ F$ and $\mathrm{D} = D \circ E$
- G and E act like Decoder and Encoder in auto-encoder

Image credit: Stanislav Pidhorskyi *et al.* 2020

From ALAE to StyleALAE[7]



- Now E and G act the same roles as f and g networks in StyleGAN

[7] https://arxiv.org/pdf/2004.04467.pdf

Image credit: Stanislav Pidhorskyi *et al.* 2020

StyleALAE: Generation

Carnegie Mellon

Figure 6: **FFHQ generations.** Generations with StyleALAE trained on FFHQ [24] at 1024 × 1024.

Image credit: Stanislav Pidhorskyi *et al.* 2020

StyleALAE: Mixing

Carnegie Mellon

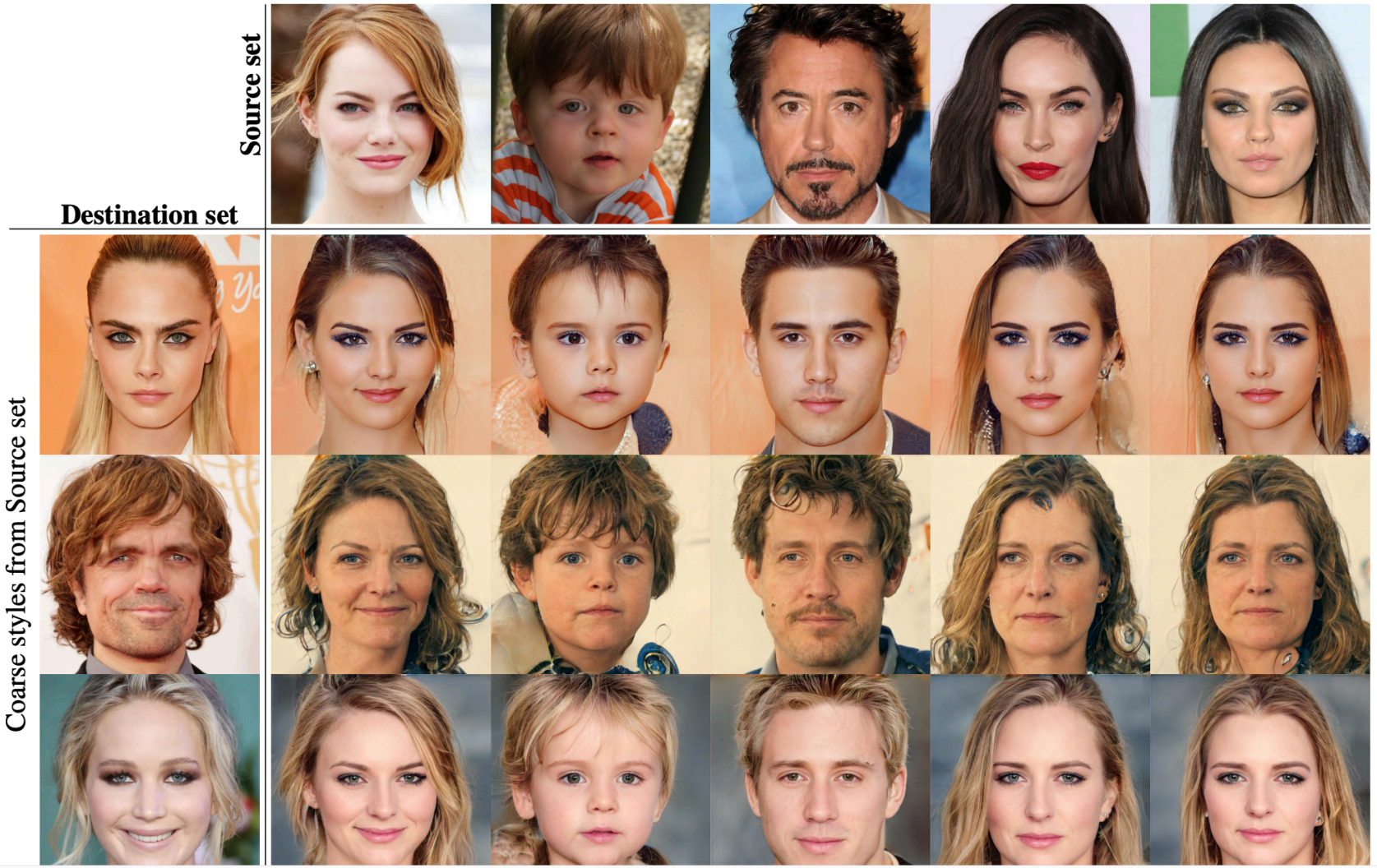


Image credit: Stanislav Pidhorskyi *et al.* 2020

Some other resources

Carnegie Mellon

- Multi-Content GAN for Few-Shot Font Style Transfer[8]
- Deep Photo Style Transfer[9]
- FSGAN: Subject Agnostic Face Swapping and Reenactment[10]
- Unsupervised Multimodal Video-to-Video Translation via Self-Supervised Learning[11]
- Unpaired Photo-to-manga Translation Based on The Methodology of Manga Drawing[12]
- ▸ Code: Neural Style Transfer
- ▸ Code: CycleGAN

[8] https://arxiv.org/pdf/1712.00516v1.pdf
[9] https://arxiv.org/pdf/1703.07511v3.pdf
[10] https://arxiv.org/pdf/1908.05932.pdf
[11] https://arxiv.org/pdf/2004.06502.pdf
[12] https://arxiv.org/pdf/2004.10634.pdf