TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Generative Adversarial Networks (GANs)

Continuous Cross Entropy is Problematic

(When the Entropy is Large)

Suppose we want to train a model of the probability distribution of natural images using cross-entropy loss.

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{y \sim \text{pop}} \; - \ln p_\Phi(y)$$

Images are continuous stuctured objects — a continuous value at every pixel.

It is difficult to build probability models for sounds or images (or other high-entropy continuous densities) that both accurately model the distribution and also allow us to calculate $p_\Phi(y)$.

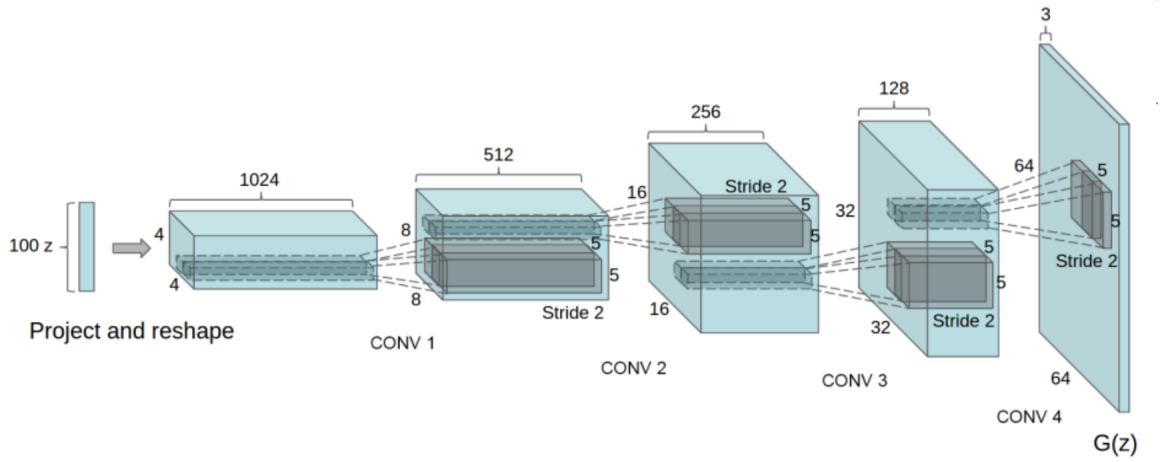
Generative Adversarial Networks (GANs)

GANs represent $p_\Phi(y)$ implicitly by constructing an image generator and abandon the ability to compute $p_\Phi(y)$.

The cross-entropy loss is replaced by an adversarial discriminator which tries to distinguish between generated images and real images.

Representing a Distribution with a Generator

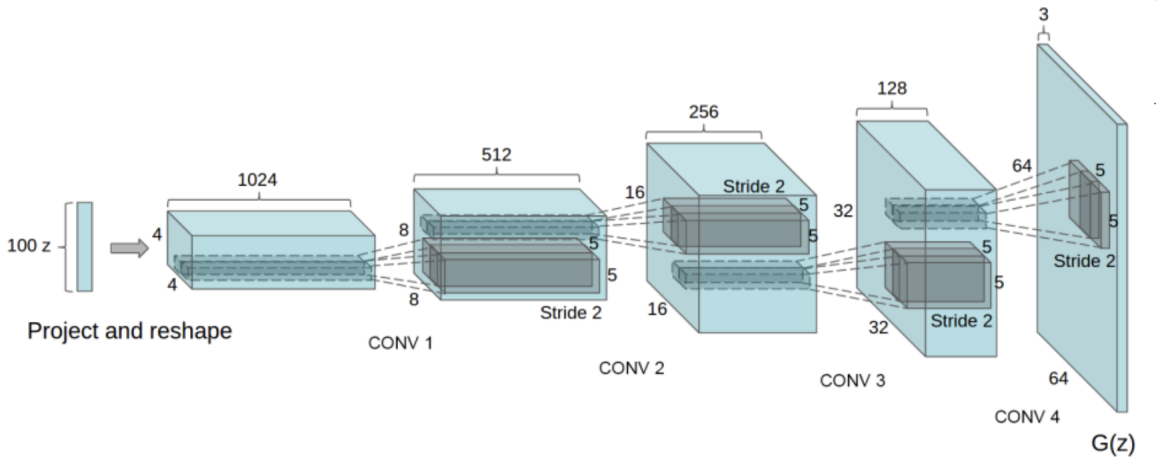
z $y_\Phi(z)$



The random input z defines a probability density on images $y_\Phi(z)$. We will write this as $p_\Phi(y)$ for the image y.

Representing a Distribution with a Generator

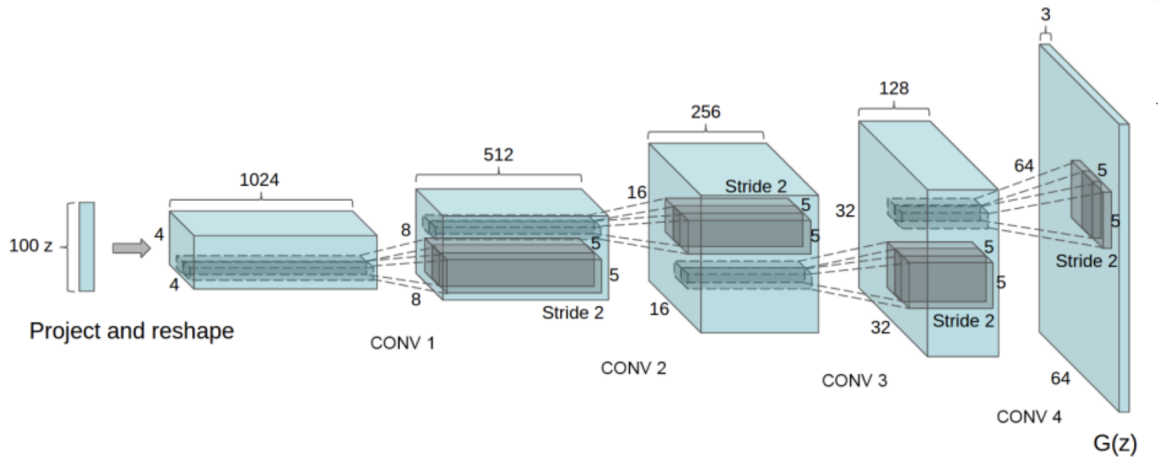
z $y_\Phi(z)$



We want $p_\Phi(y)$ to model a natural image distribution such as the distribution over human faces.

Representing a Distribution with a Generator

z $y_\Phi(z)$



We can sample from $p_\Phi(y)$ by sampling z. But we cannot compute $p_\Phi(y)$ for y sampled from the population.

Increasing Spatial Dimension

Reducing spatial dimention with strided convolution:

For $x, y, j, \Delta x, \Delta y, i$

$$L_{\ell+1}[x, y, j] \texttt{ += } W[\Delta x, \Delta y, i, j] L_\ell[s * x + \Delta x, s * y + \Delta y, i]$$

Increasing spatial dimension with PyTorch ConvTranspose2d:

For $x, y, j, \Delta x, \Delta y, i$

$$L_{\ell+1}[s * x + \Delta x, s * y + \Delta y, i] \texttt{ += } W[\Delta x, \Delta y, i, j] L_\ell[x, y, j]$$

Irrelevant Observation: ConvTranspose follows the "swap rule" for computing gradients for a spacially-reducing Conv layer.

Generative Adversarial Networks (GANs)

Let y range over images. We have a generator p_Φ. For $i \in \{-1, 1\}$ we define a probability distribution over pairs $\langle i, y \rangle$ by

$$\begin{aligned} \tilde{p}_\Phi(i = 1) &= 1/2 \\ \tilde{p}_\Phi(y|i = 1) &= \mathrm{pop}(y) \\ \tilde{p}_\Phi(y|i = -1) &= p_\Phi(y) \end{aligned}$$

We also have a discriminator $P_{\mathrm{disc}}(i|y)$ that tries to determine the source i given the image y.

The generator tries to fool the discriminator.

$$\mathrm{gen}^* = \operatorname*{argmax}_{\mathrm{gen}} \min_{\mathrm{disc}} E_{\langle i, y \rangle \sim \tilde{p}_{\mathrm{gen}}} - \ln P_{\mathrm{disc}}(i|y)$$

GANs

The generator tries to fool the discriminator.

$$\text{gen}^* = \underset{\text{gen}}{\text{argmax}} \ \min_{\text{disc}} \ E_{\langle i, y \rangle \sim \tilde{p}_{\text{gen}}} \ -\ln P_{\text{disc}}(i|y)$$

Assuming universality (next slide) of both the generator p_{gen} and the discriminator P_{disc} we have $p_{\text{gen}^*} = \text{pop}$.

Note that this involves only discrete cross-entropy.

The Universality Assumption

DNNs are universally expressive (can model any function) and trainable (the desired function can be found by SGD).

Universality assumption is clearly false but is useful.

The success of GANs (to the extent that they have been successful) is a tribute to the utility of the universality ssumption.

Jensen-Shannon Divergence

$$
\begin{aligned}
\text{gen}^* &= \underset{\text{gen}}{\text{argmax}} \ \min_{\text{disc}} \ E_{\langle i,\, y\rangle} \left[-\ln P_{\text{disc}}(i|y)\right] \\
&= \underset{\text{gen}}{\text{argmax}}\, E_{\langle i,\, y\rangle} \left[-\ln P(i|y)\right] \\
&= \underset{\text{gen}}{\text{argmax}}\, \frac{1}{2} E_{y\sim \text{Pop}} \left[-\ln P(1|y)\right] + \frac{1}{2} E_{y\sim p_{\text{gen}}} \left[-P(-1|y)\right] \\
&= \underset{\text{gen}}{\text{argmin}} \ KL\left(\text{pop}, \frac{\text{pop}+p_{\text{gen}}}{2}\right) + \ KL\left(p_{\text{gen}}, \frac{\text{pop}+p_{\text{gen}}}{2}\right) \\
\text{gen}^* &= \underset{\text{gen}}{\text{argmin}} \ \text{JSD}(\text{pop}, p_{\text{gen}})
\end{aligned}
$$

GAN Mode Collapse

A major concern is "mode collapse" where the learned distribution omits a significant fraction of the population distribution.

There is no quantitative performance measure that provides a meaningful guarantee against mode collapse.

In practice GANS are evaluated on FID score.

The Fréchet Inception Score (FID)

Consider two distributions P and Q on $\mathbb{R}^d$ (perhaps two distributions on images).

Generative image models are (still) evaluated using a certain measure of the "distance" between the population distribution and the generation distribution.

For GANs we cannot compute the probability of a generated image so we cannot use cross entropy or KL-divergence.

The Fr\'enchet Inception Score (FID)

The Fr\'enchet distance $F(P, Q)$ can be measured (approximately) by sampling.

Let μ range over distributions on pairs (x, y) such that

$$\sum_y \mu(x, y) = P(x)$$

$$\sum_x \mu(x, y) = Q(y)$$

$$\mathrm{F}(P, Q) = \inf_{\mu} \left(E_{(x,y)\sim\mu} ||x - y||^2 \right)^{\frac{1}{2}}$$

This is a form of earth movers distance. It is also known as the 2-Wasserstein distance.

The Fréchet Inception Score (FID)

But rather than measure the L_2 distance between images we measure the L_2 distance between the "inception feature vectors" $I(x)$ and $I(y)$.

For an image x the feature vector $I(x)$ is computed from a certain layer in the inception image classification network.

Generative Adversarial Nets

Goodfellow et al., June 2014

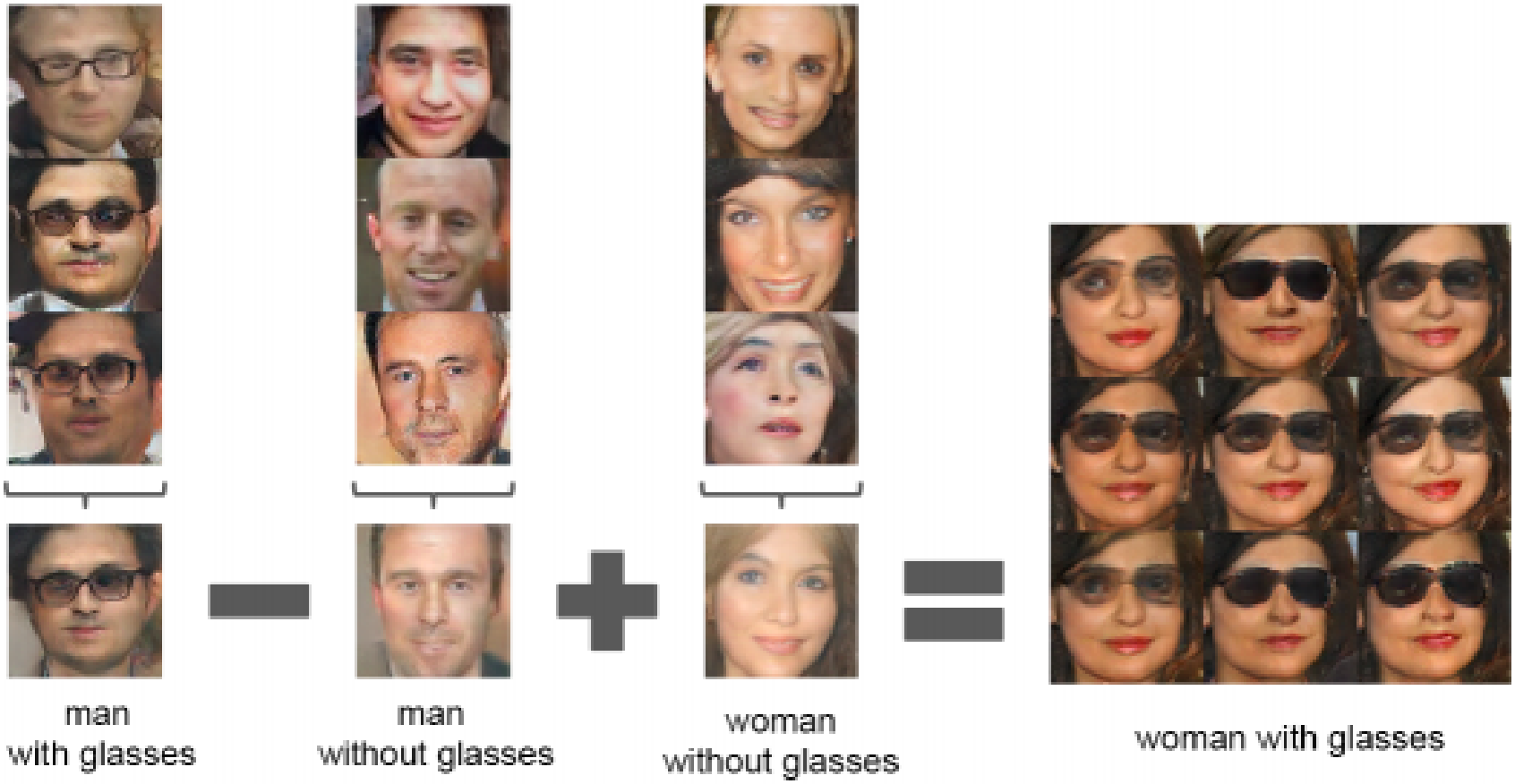
The rightmost column (yellow boarders) gives the nearest neighbor in the training data to the adjacent column.

Unsupervised Representation Learning ... (DC GANS)

Radford et al., Nov. 2015

Unsupervised Representation Learning ... (DC GANS)

Radford et al., Nov. 2015



Interpolated Faces

[Ayan Chakrabarti, January 2017]

Conditional Density Estimation with L_2 loss

Consider training a model P_Φ to predict the next frame in a video from the two previous frames.

$P_\Phi(y_{t+1}|y_t, y_{t+1})$

The pair y_t and y_{t+1} give both an image and the motion in the image so predicting y_{t+2} is possible.

Conditional Density Estimation with L_2 loss

Here we can train by L_2 loss

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \, ||\hat{y}_\Phi(y_t, y_{t-1}) - y_{t+2}||^2$$

L_2 loss is a special case of cross-entropy loss where $P_\phi(y_{t+2}|y_t, y_{t+1})$ is taken to be a Gaussian centered as $\hat{y}_\Phi(y_t, y_{t+1})$.

The General Conditional Case

We now consider a general conditional case we have a population distribution over pairs $\langle x, y \rangle$ where x provides enough information to train the model $\hat{y}_\Phi(y|x)$ by L_2 or L_1 loss.

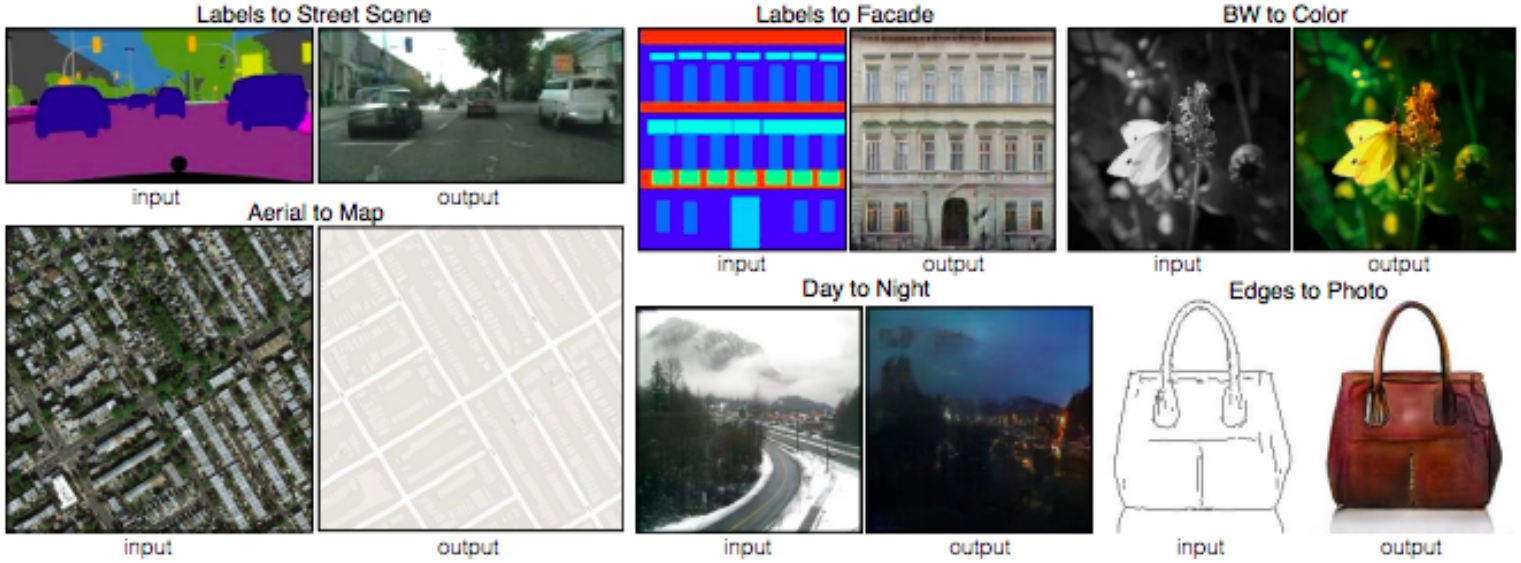
For the L_1 case we have the following.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \, ||\hat{y}_\Phi(y_t, y_{t-1}) - y_{t+2}||_1$$

Image-to-Image Translation (Pix2Pix)

Isola et al., Nov. 2016

We assume a corpus of "image translation pairs" such as images paired with semantic segmentations.



U-Nets

Pix2Pix uses a U-Net.

U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger, Fischer and Brox, May 2015.

U-Nets are fundamental to current diffusion models.

Adversarial Discrimination as an Additional Loss

$$\text{gen}^* = \underset{\text{gen}}{\text{argmin}} \; E_{(x,y)\sim\text{pop}} \; ||y - y_{\text{gen}}(x)||_1 + \lambda \, \mathcal{L}_{\text{Discr}}(\text{gen})$$

$$\mathcal{L}_{\text{Discr}}(\text{gen}) = \max_{\text{disc}} \; E_{x,y,i\sim\tilde{p}_{\text{gen}}} \; \ln P_{\text{disc}}(i|y,x)$$

Discrimination as an Additional Loss

$$\text{L1}: \quad \text{gen}^* = \text{argmin}_{\text{gen}} \; E_{(x,y)\sim\text{pop}} \; ||y - y_{\text{gen}}(x)||_1$$

$$\text{cGAN}: \quad \text{gen}^* = \text{argmin}_{\text{gen}} \; \mathcal{L}_{\text{Discr}}(\text{gen})$$

$$\text{L1} + \text{cGAN}: \quad \text{gen}^* = \text{argmin}_{\text{gen}} \; E_{(x,y)\sim\text{pop}} \; ||y - y_{\text{gen}}(x)||_1 \; + \; \lambda \, \mathcal{L}_{\text{Discr}}(\text{gen})$$

Image-to-Image Translation (Pix2Pix)

Isola et al., Nov. 2016

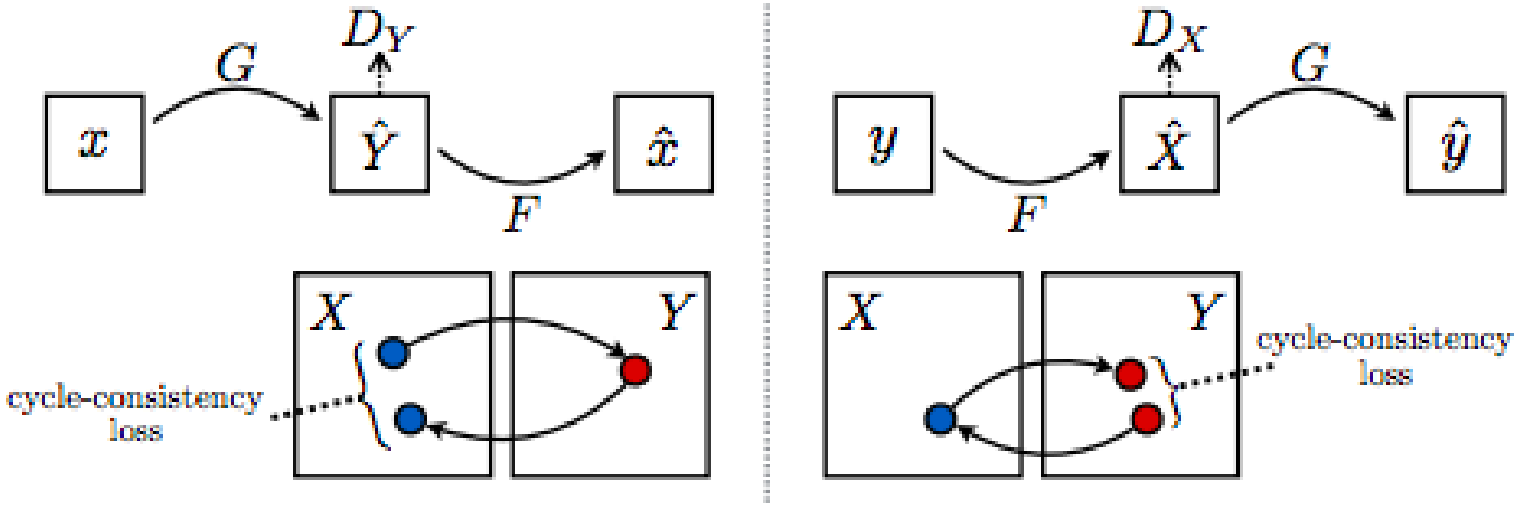
Arial Photo to Map and Back

Unpaired Image-to-Image Translation (Cycle GANs)

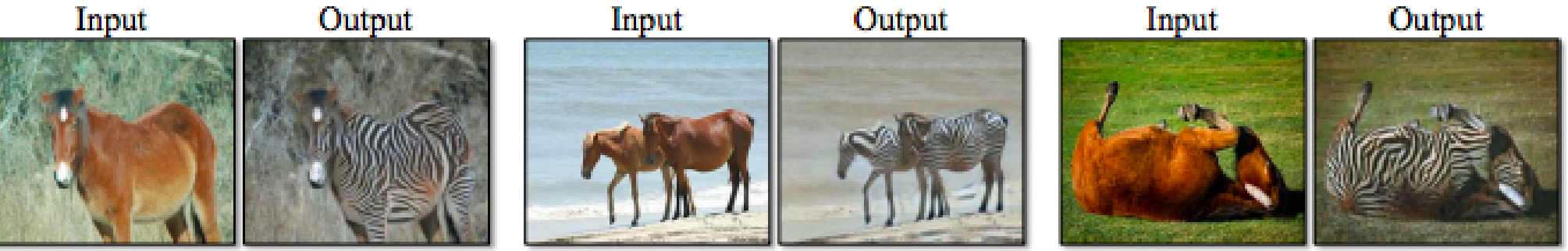
Zhu et al., March 2017

We have two corpora of images, say images of zebras and unrelated images of horses, or photographs and unrelated paintings by Monet.

We want to construct translations between the two classes.



Cycle Gans


Input
Output
Input
Output
Input
Output
horse → zebra
zebra → horse

Cycle Gans

Horse → Zebra

Unsupervised Machine Translation (UMT)

Lample et al, Oct. 2017, also Artetxe et al., Oct. 2017

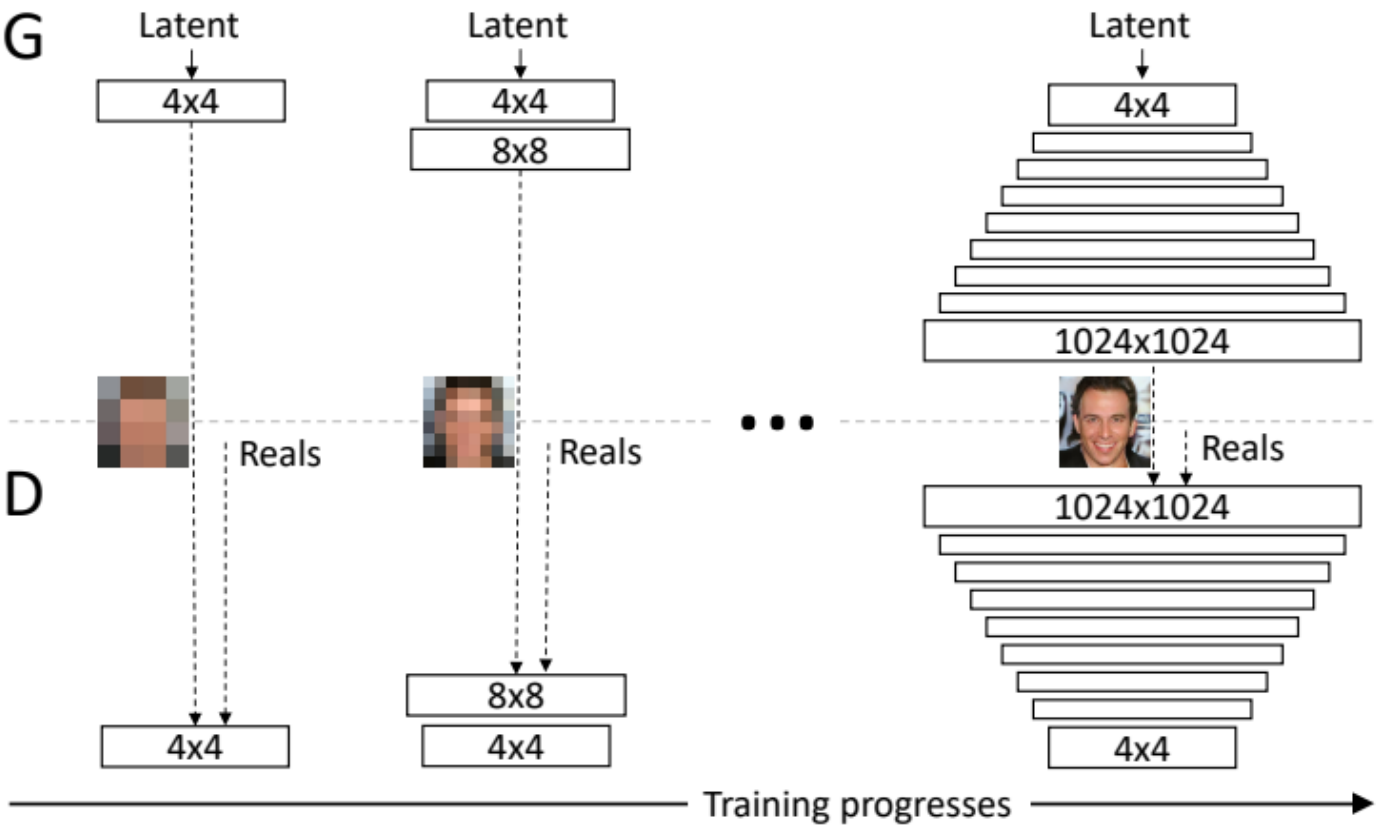
In unsupervised machine translation the cycle loss is called **back-translation**.

Progressive GANs

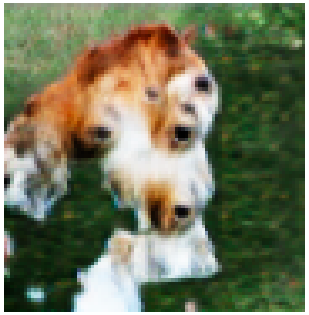
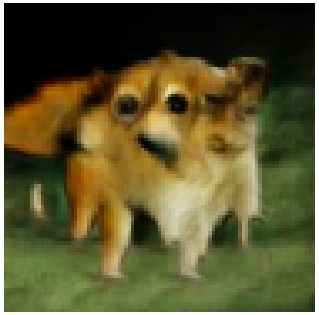
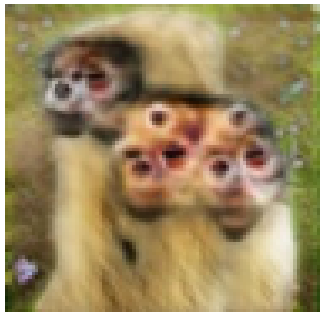
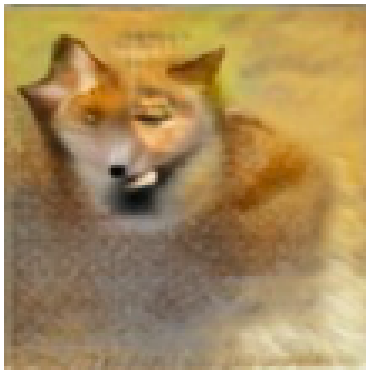
Progressive Growing of GANs, Karras et al., Oct. 2017

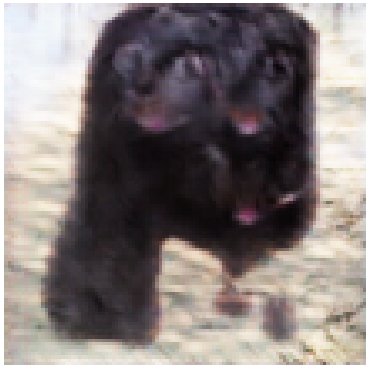
Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

Progressive GANs



Early GANs on ImageNet



BigGans

Large Scale GAN Training, Brock et al., Sept. 2018

Figure 1: Class-conditional samples generated by our model.

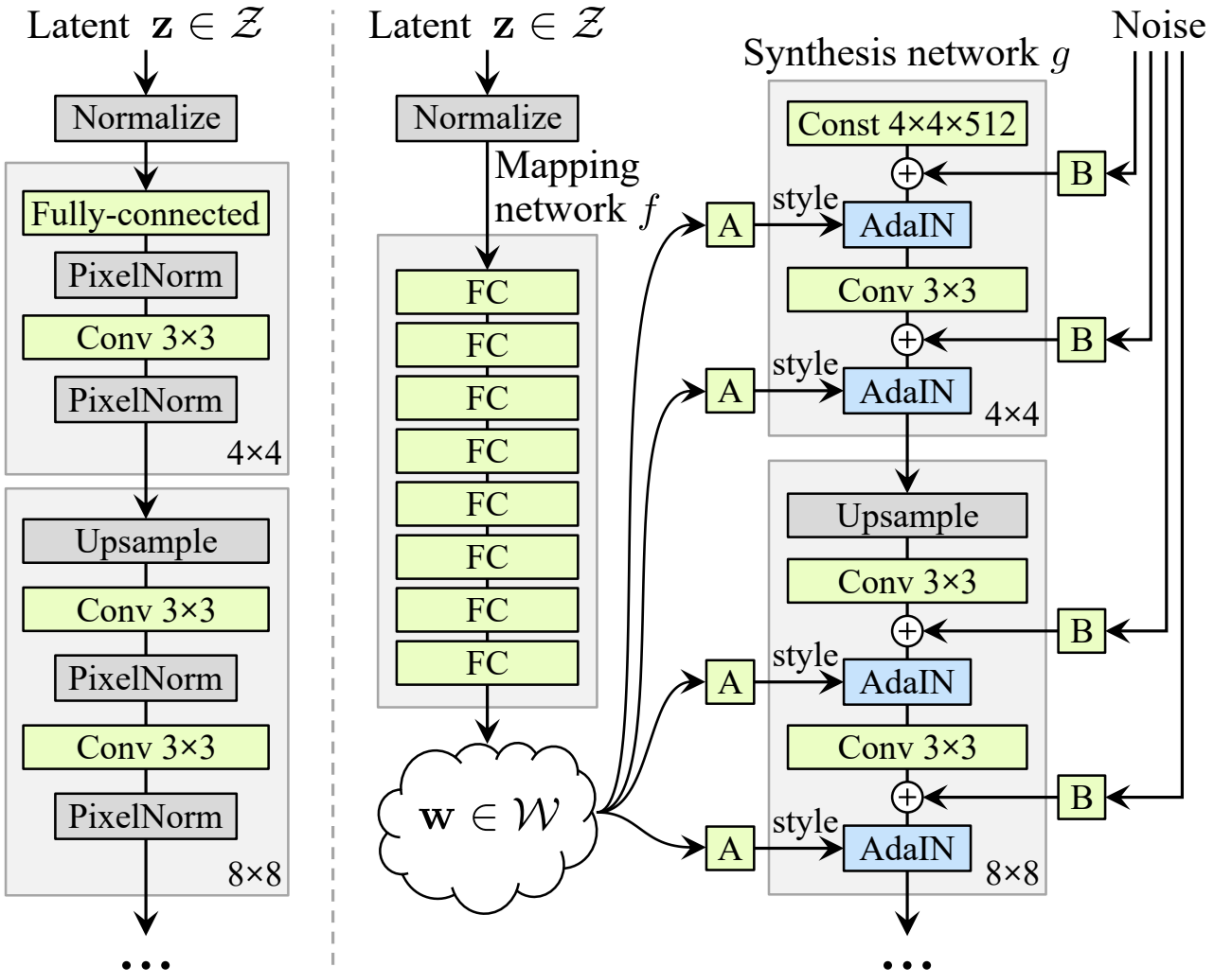
This is a class-conditional GAN — it is conditioned on the imagenet class label.

This generates 512 X 512 images without using progressive training.

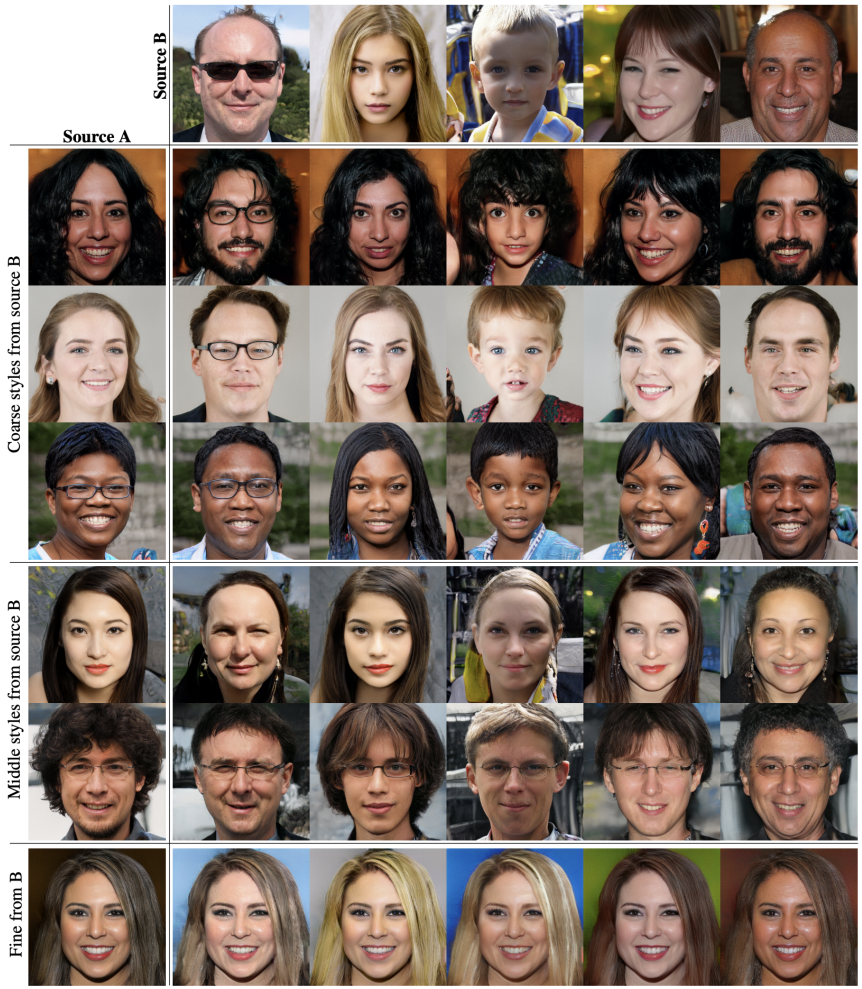
StyleGAN

A Style-Based Generator Architecture for Generative Adversarial Networks, Karras et al., Dec. 2018

StyleGAN: Architecture



StyleGAN: Style Transfer



StyleGAN: Noise Variation

StyleGAN2 and StyleGAN3

StyleGan2 appeared in December of 2019 with significant improvements.

It was demonstrated to work on many classes of images, not just faces.

StyleGAN3 appeared in June 2021.

Projecting Images into Latent Space

Given an image, can we find a noise vector (a latent vector) that generate a close approximation of the given image. Can we invert the generator?

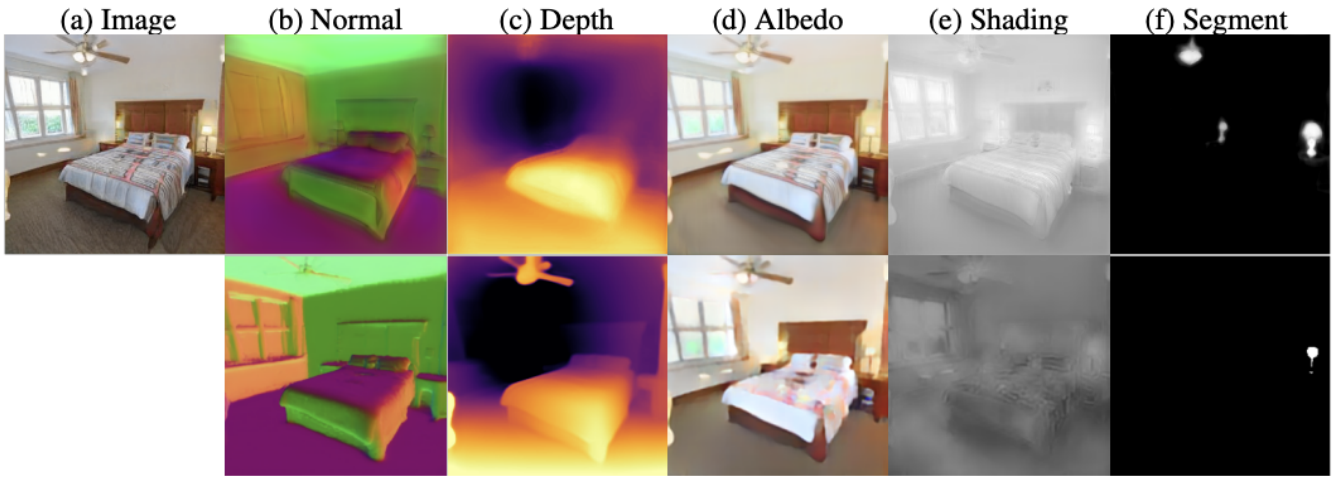
We can invert generated images. But we cannot invert (nearly as accurately) newly sampled natural images.

By measuring the match between an image y and $g(g^{-1}(y))$ we can determine whether y was generated by StyleGAN2.

June 2023: StyleGAN Seems to "Understand" Images

StyleGAN knows Normal, Depth, Albedo, and More

Anand Bhattad **Daniel McKee** **Derek Hoiem** **D.A. Forsyth**
University of Illinois Urbana Champaign



END