Generative Models

Analytics Club, IIT Madras

Quick Recap!

Supervised vs Unsupervised Learning

Supervised Learning

Given: (x, y) where x is data, y is label

Goal: Learn a function mapping from x to y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, **Conditional density estimation i.e P(X | Y)**

Unsupervised Learning

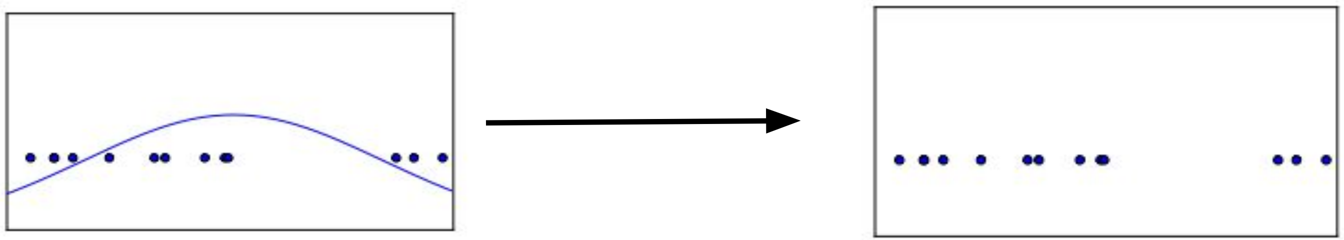
Given: x - unlabelled data

Goal: Learn some underlying hidden structure of the data

Examples: Clustering, dimensionality reduction, feature learning, **marginal density estimation i.e. P(x)**

Introduction

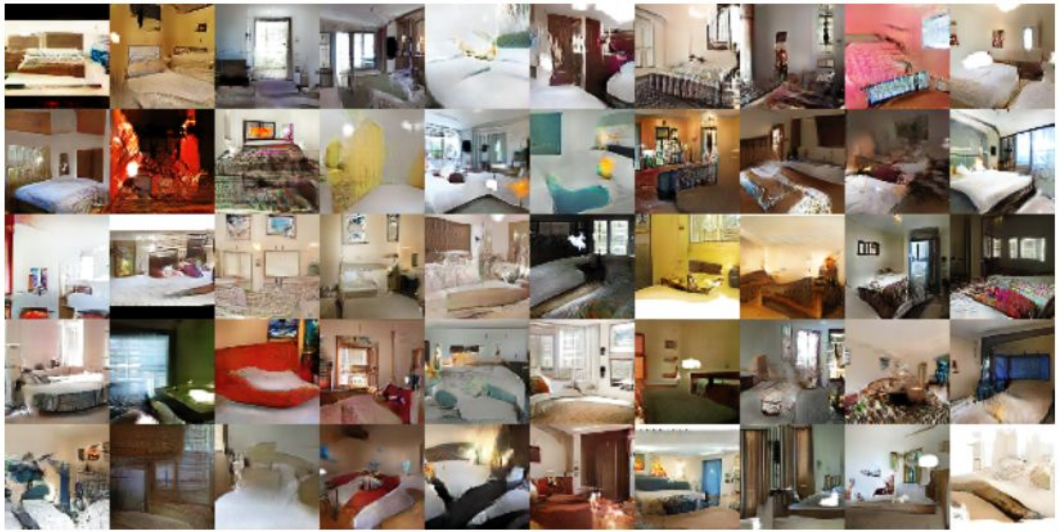
1-d density estimation



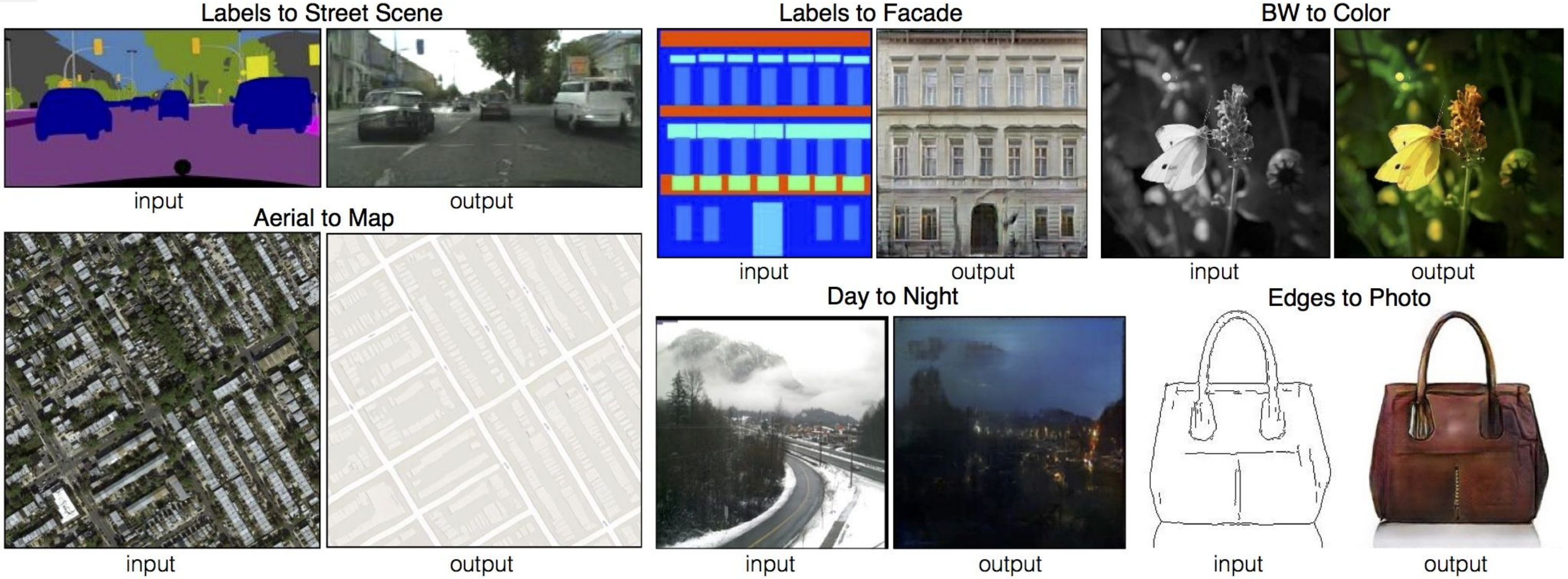
Generative Models: Given training data, generate new samples from same distribution

Training data ~ $p_{data}(x)$, Generated samples ~ $p_{model}(x)$, Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

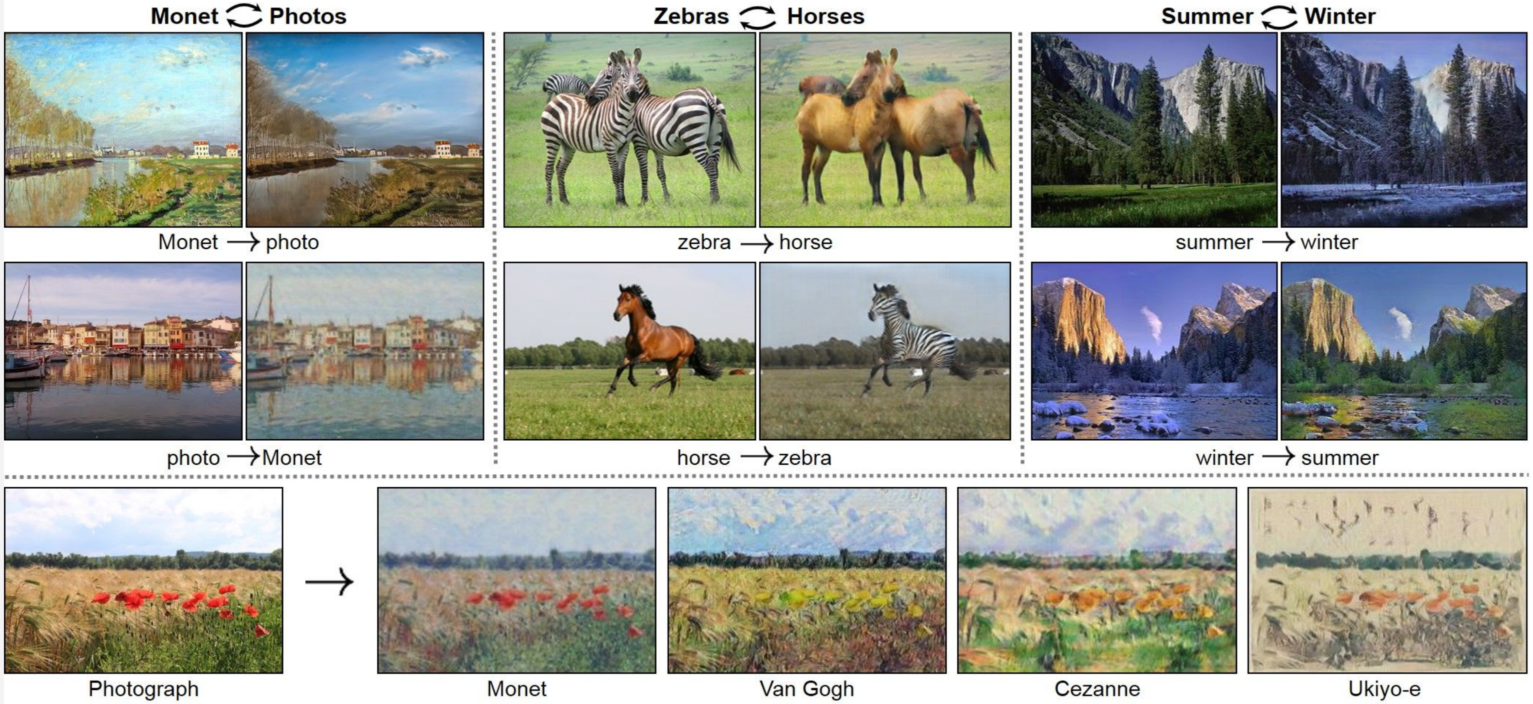
What can Generative Models do?



What can Generative Models do?



What can Generative Models do?



Some Math Basics

- **A Divergence** is a metric to measure how much probability distributions p, q differ from each other.

- KL(p||q) = ∫ p(x)(log(p(x)) - log(q(x))) dx

  Minimum value of KL divergence: 0
  Maximum value of KL divergence: ∞
  KL divergence is 0 iff p = q.

- Is KL divergence a symmetric metric?
  What happens to the KL divergence in the following cases?
    - $p(x) = 0$ but $q(x) \neq 0$. Nothing much.
    - $q(x) = 0$ but $p(x) \neq 0$. It explodes.

Some Math Basics

- $KL(p||q) = \int p(x) \log(p(x)) - \int p(x) \log(q(x))$
  $= -Entropy(p) + CrossEntropy(p, q)$.

- For many unsupervised learning problems we maximize the log-likelihood of the training data P

  $$\max LL(\text{real data}) = \max \sum_{x \in D} \log(p_{model}(x)) * 1/N$$
  $$\equiv \max -CrossEntropy(p_{data}, p_{model})$$
  $$\equiv \min CrossEntropy(p_{data}, p_{model})$$
  $$\equiv \min KL(p_{data} || p_{model})$$

- Therefore given enough samples of real data, if we maximize log-likelihood, we end up minimizing $KL(p_{data} || p_{model})$.

Some Math Basics

- The **Jensen-Shannon Divergence** (JSD) is a symmetric version of KL divergence defined as

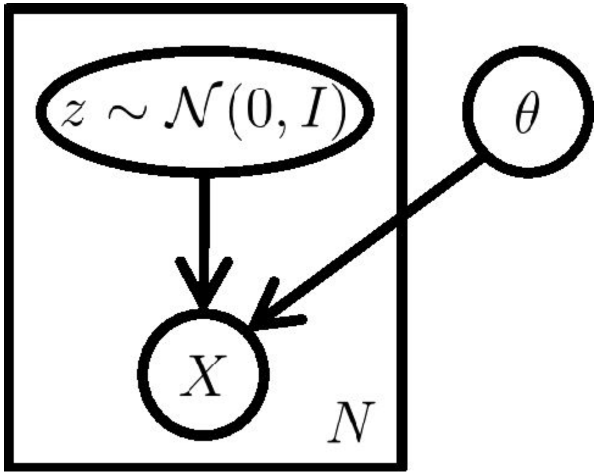
$$JSD(P \,||\, Q) = \tfrac{1}{2} KL(P \,||\, (P+Q)/2) + \tfrac{1}{2} KL(Q \,||\, (P+Q)/2\ )$$

Properties:

- $0 \le JSD \le \log 2$.
- JSD is 0 iff $p = q$.
- If p and q have disjoint supports, the $JSD(P \,||\, Q) = \log 2$

Latent Variable Models

- The Generative model makes a decision on what it is going to generate beforehand with the help of latent variables, eg MNIST digit generation

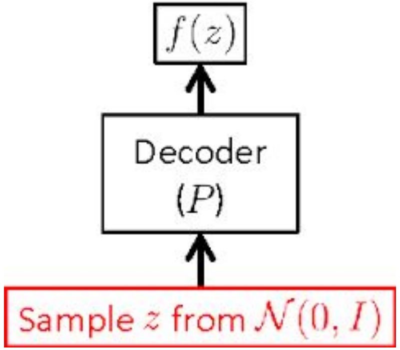


Variational Autoencoders (VAEs)

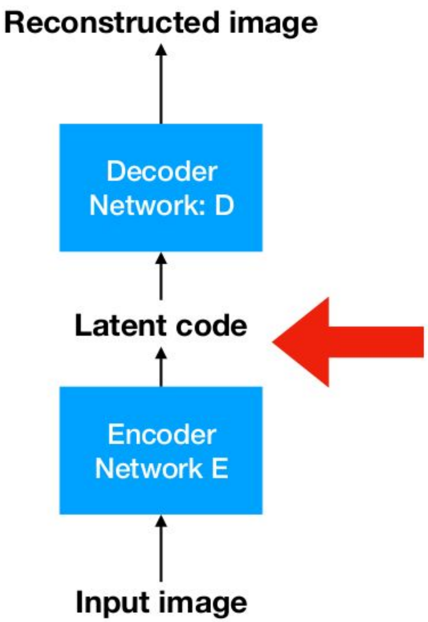
Maximum likelihood learning (KL Divergence)

$$P(X) = \int P(X|z;\theta)P(z)dz$$

- P(z) - Prior distribution - high dimensional standard gaussian
- In VAEs, output distribution is often Gaussian, i.e., $P(X|z;\theta) = N(X|\, f(z;\theta), \sigma^2 * I)$. That is, it has mean $f(z;\theta)$ and covariance equal to the identity matrix I times some scalar σ
- Integral is Intractable



VAEs



Ideally, we would like to maximize

$$\max_{\theta} L(\theta|\text{Dataset}) \equiv \sum_{x^{(j)}} \log p(\theta|x^{(j)})$$

But we maximize a lower bound for tractability

$$\max_{\theta} L_V(\theta|\text{Dataset}) \equiv$$

$$\sum_{x^{(j)}} -\mathrm{KL}(q_\theta(z|x^{(j)})||\mathcal{N}(z|0,I))+$$

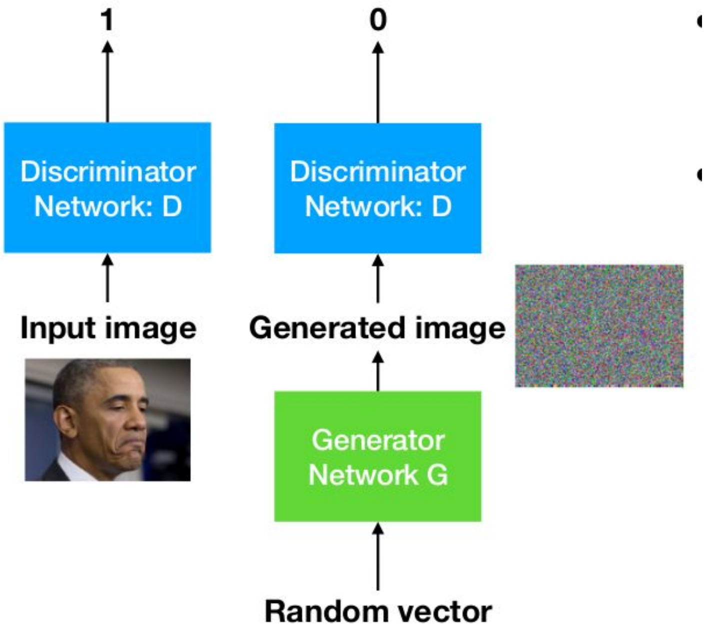
$$E_{z\sim(q_\theta(z|x^{(j)}))}[\log p_\theta(x^{(j)}|z)]$$

$$\leq L(\theta|\text{Dataset})$$

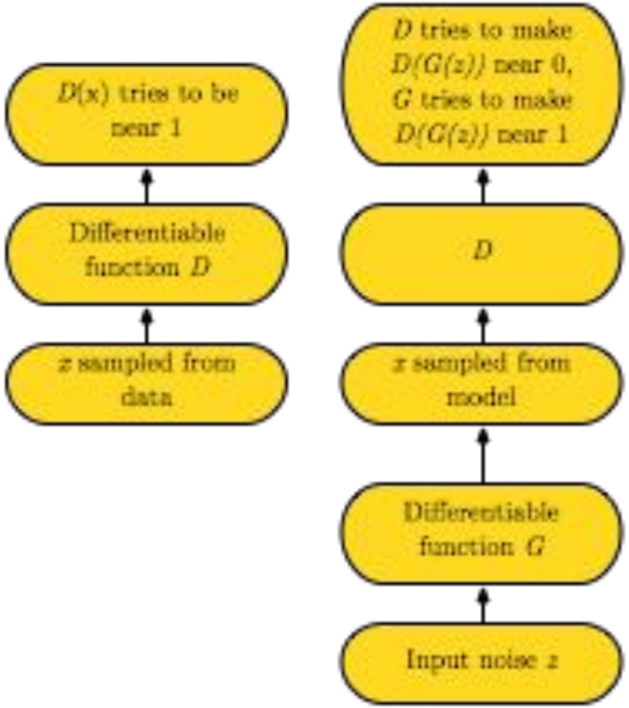
GANs

- Goodfellow et al. NIPS 2014

- Forget about designing a perceptual loss. Let's train a discriminator to differential real and fake image



GANs


D(x) tries to be near 1
Differentiable function D
x sampled from data
D tries to make D(G(z)) near 0, G tries to make D(G(z)) near 1
D
x sampled from model
Differentiable function G
Input noise z

GANs Objective

GANs solve a minimax objective

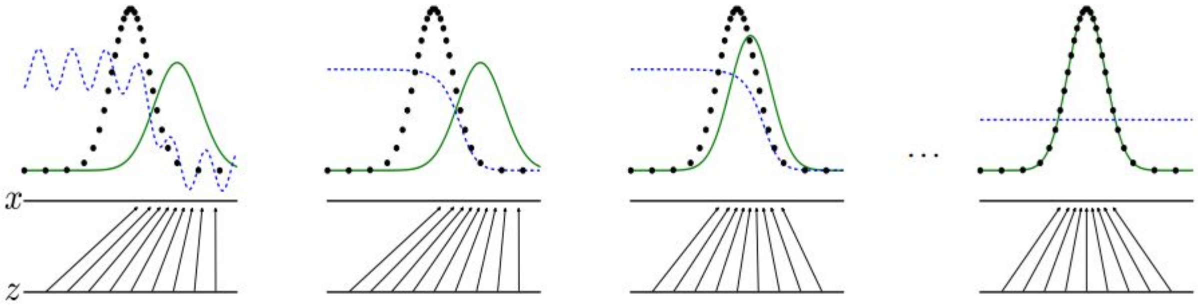
$$\min_G \max_D \quad E_{x \sim p_X}[\log D(x)] + E_{z \sim p_Z}[\log(1 - D(G(z))]$$

$$\begin{aligned} p_X &: \quad \text{Data distribution,} \\ &\quad\ \ \text{usually represented by samples.} \\ p_{G(Z)} &: \quad \text{Model distribution, where} \\ &\quad\ \ Z \text{ is usually modeled as uniform or Gaussian.} \end{aligned}$$

Discriminator strategy

Optimal discriminator (non-parametric)

$$D(x) = \frac{p_X(x)}{p_X(x) + p_{G(Z)}(x)}$$



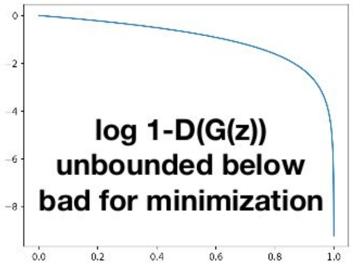
JS Divergence

Under an ideal discriminator, the generator minimizes the Jensen-Shannon divergence between p_X and $p_{G(Z)}$. This also requires that D and G have sufficient capacity and a sufficiently large dataset.

GANs in practice

- Step 1: Fix G and perform a gradient step to

$$\max_D E_{x\sim p_X}[\log D(x)] + E_{z\sim p_Z}[\log(1 - D(G(z))]$$

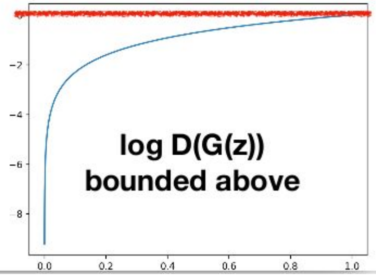


- Step 2: Fix D and perform a gradient step to
  (in theory)

$$\min_G E_{z\sim p_Z}[\log(1 - D(G(z))]$$

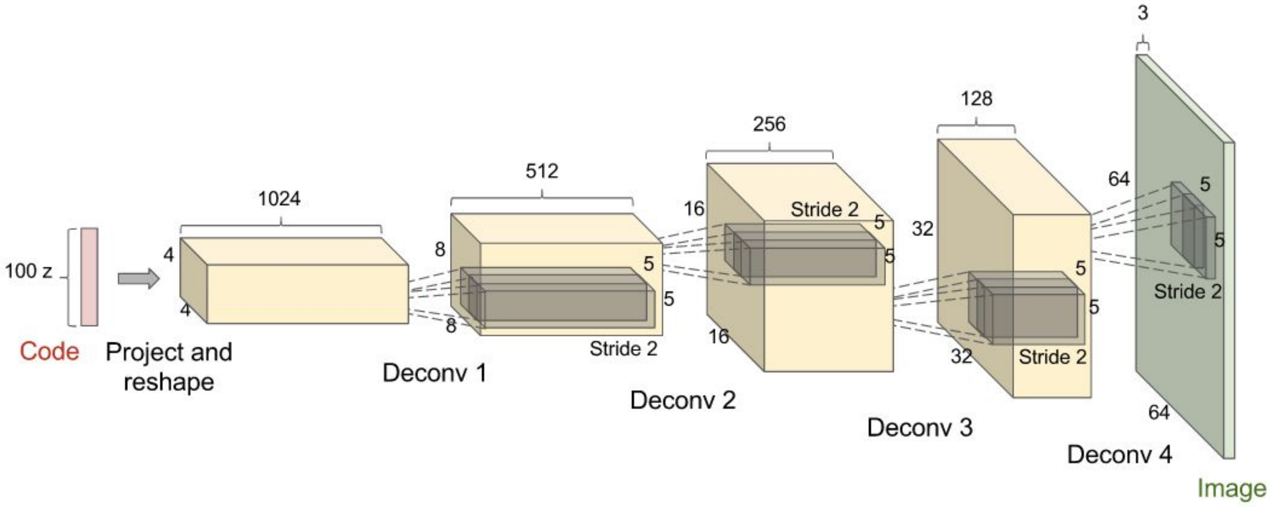
(in practice)

$$\max_G E_{z\sim p_Z}[\log D(G(z))]$$



36

DC-GAN Architecture



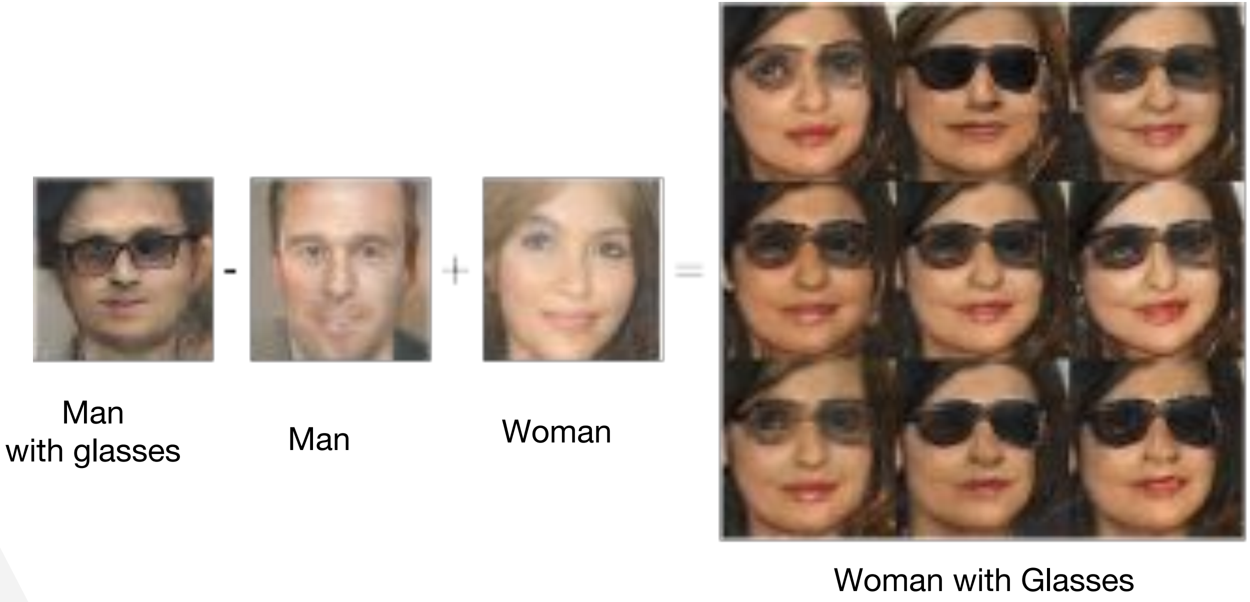
Results

DCGANs for LSUN Bedrooms

(Radford et al 2015)

Results

Vector Space Arithmetic



(Radford et al 2015)

Wasserstein GAN

M. Arjovsky, S. Chintala, L. Bottou “Wasserstein GAN” 2016
Replace classifier with a critic function

Discriminator

GAN $$\max_D E_{x\sim p_X}[\log D(x)] + E_{z\sim p_Z}[\log(1 - D(G(z))]$$

WGAN $$\max_D E_{x\sim p_X}[D(x)] - E_{z\sim p_Z}[D(G(z))]$$

Generator

GAN $$\max_G E_{z\sim p_Z}[\log D(G(z))]$$

WGAN $$\max_G E_{z\sim p_Z}[D(G(z))]$$

64

WGAN

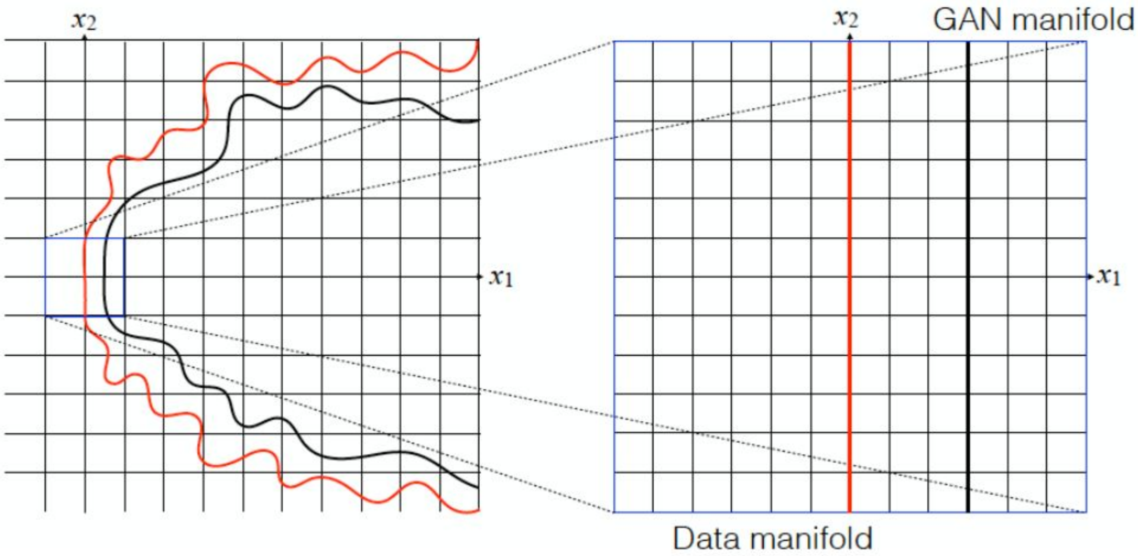
GAN: minimize Jensen-Shannon divergence between p_X and $p_{G(Z)}$

$$JS(p_X||p_{G(Z)}) = KL(p_X||\frac{p_X + p_{G(Z)}}{2}) + KL(p_{G(Z)}||\frac{p_X + p_{G(Z)}}{2})$$

WGAN: minimize earth mover distance between p_X and $p_{G(Z)}$

$$EM(p_X, p_{G(Z)}) = \inf_{\gamma \in \prod(p_X, p_{G(Z)})} E_{(x,y)\sim\gamma}[||x - y||]$$

WGAN

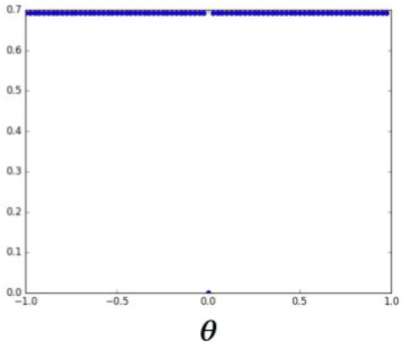


WGAN VS GAN

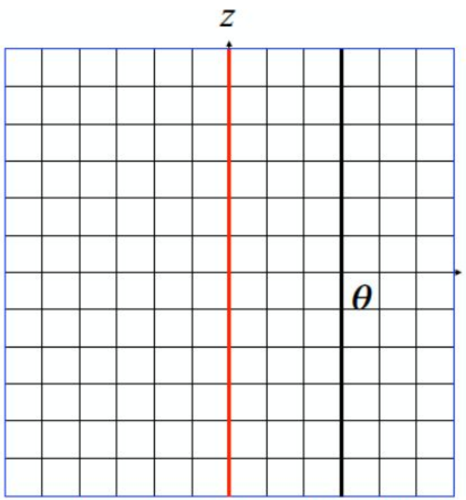
$$JS(p_X \| p_{G(Z)}) = KL(p_X \| \frac{p_X + p_{G(Z)}}{2}) + KL(p_{G(Z)} \| \frac{p_X + p_{G(Z)}}{2})$$

Jesen-Shannon divergence in this example

$$JS(p_X \| p_{G(Z)}) = \begin{cases} \log 2 & \text{if } \theta \neq 0, \\ 0 & \text{if } \theta = 0, \end{cases}$$



68



Example from Arjovsky et al. 2017

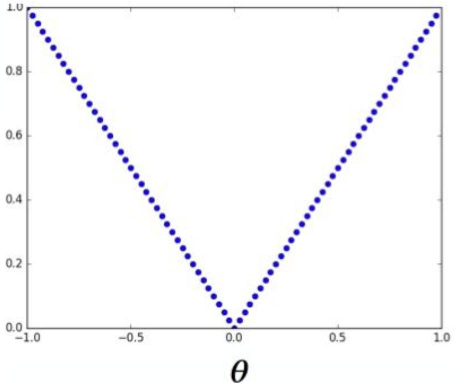
Slide credit, Courville 2017

WGAN Vs GAN

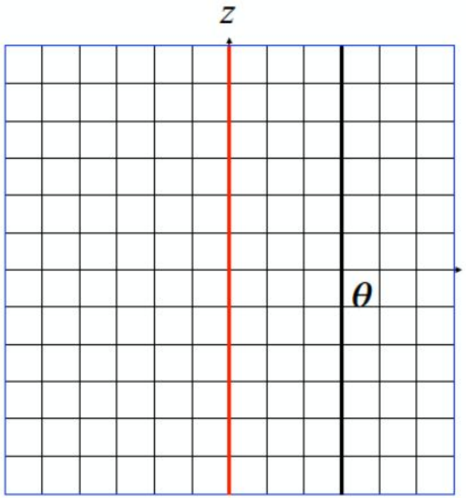
$$EM(p_X, p_{G(Z)}) = \inf_{\gamma \in \prod(p_X, p_{G(Z)})} E_{(x,y)\sim\gamma}[||x - y||]$$

Earth Mover distance in this example

$$EM(p_X, p_{G(Z)}) = |\theta|$$



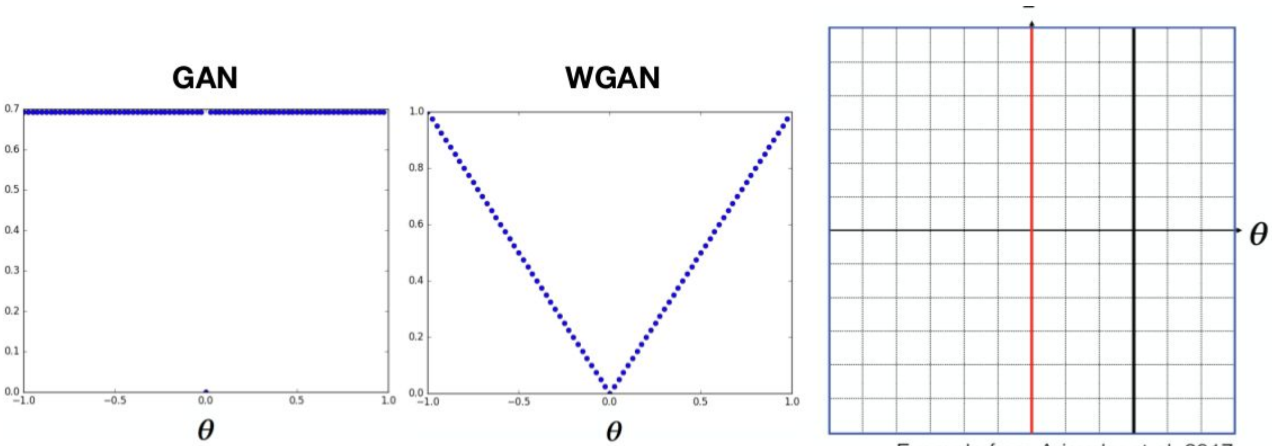
69



Example from Arjovsky et al. 2017

Slide credit, Courville 2017

WGAN Vs GAN



Example from Arjovsky et al. 2017

- If we can directly change the density shape parameter, the Earth Mover distance is smoother.
- But we do not directly change the density shape parameter, we change the generation function.

WGAN-GP

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Domoulin, A. Courville “Improved Training of Wasserstein GANs” 2017

$$\min_G \max_D E_{x\sim p_X}[D(x)] - E_{z\sim p_Z}[D(G(Z))] + \lambda E_{y\sim p_Y}[(||\nabla_y D(y)||_2 - 1)^2]$$

$$y = ux + (1-u)G(z)$$

- *y:* imaginary samples

Optimal critic has unit gradient norm almost everywhere

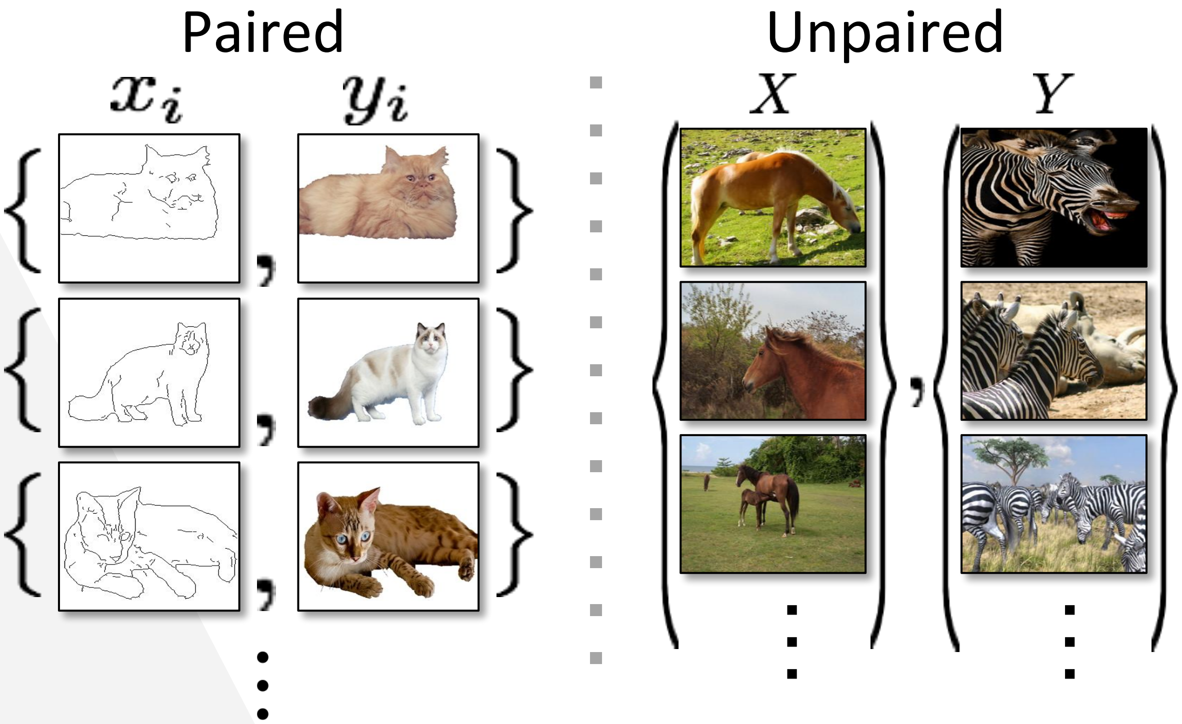
pix2pix

Paired Image-to-Image Translation

Cycle GANs

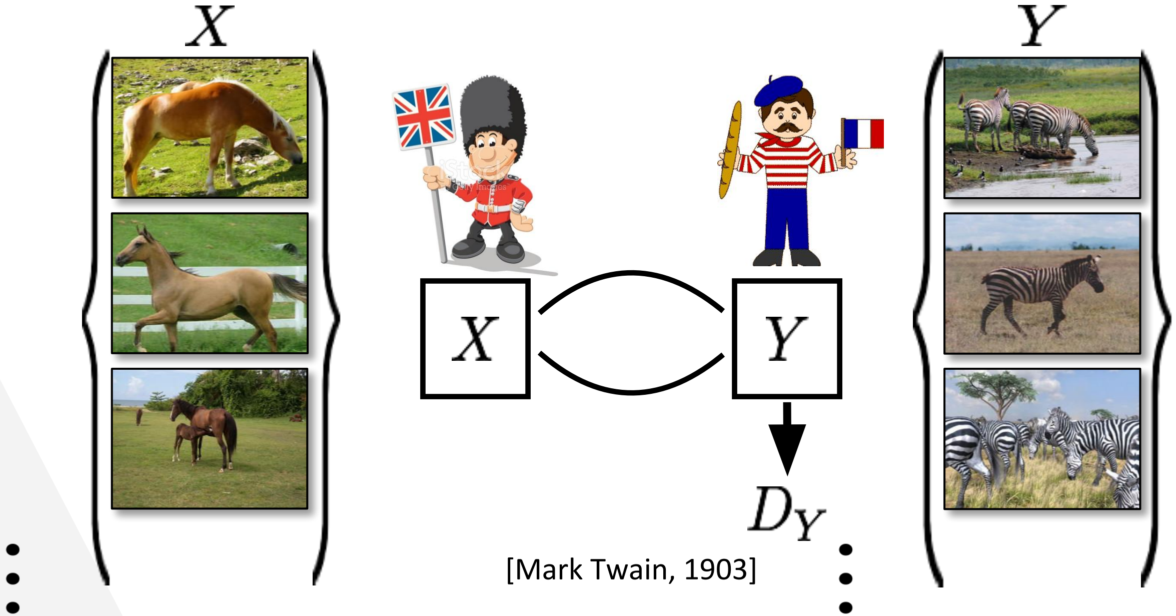
Unpaired Image-to-Image Translation with CycleGAN

Cycle GANs



Cycle GANs

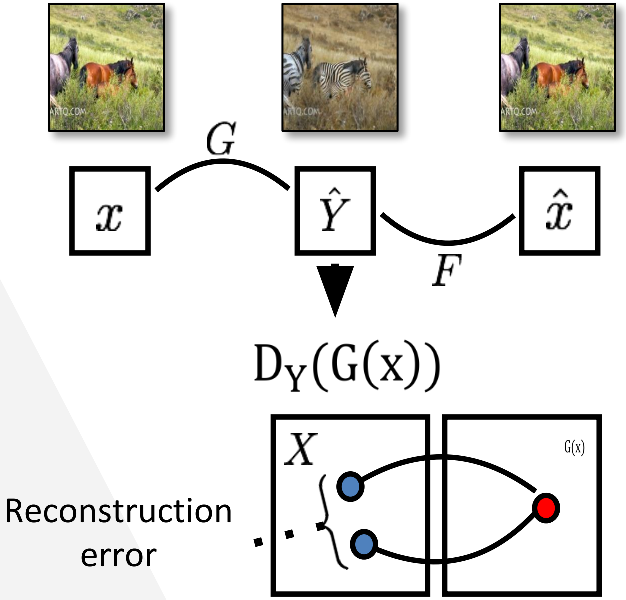
Cycle-Consistent Adversarial Networks



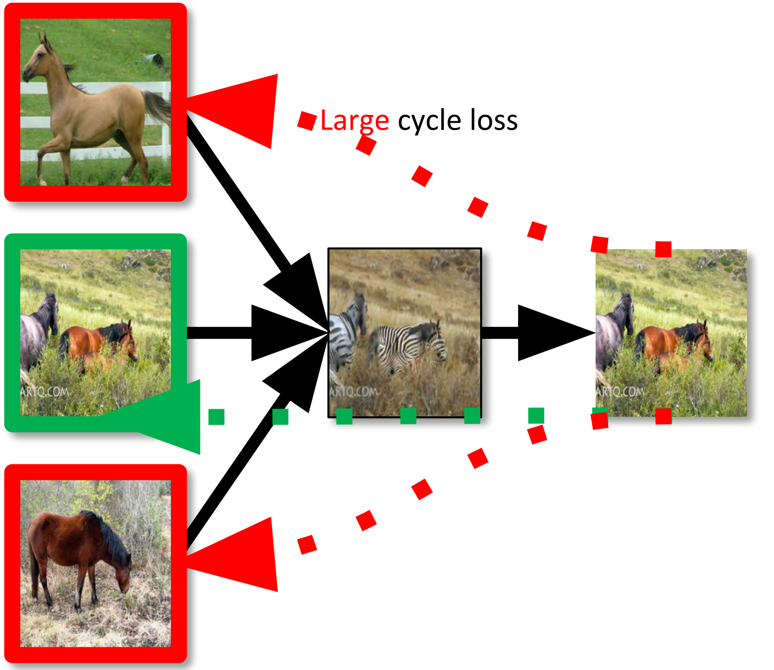
[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Cycle Consistency Loss

x

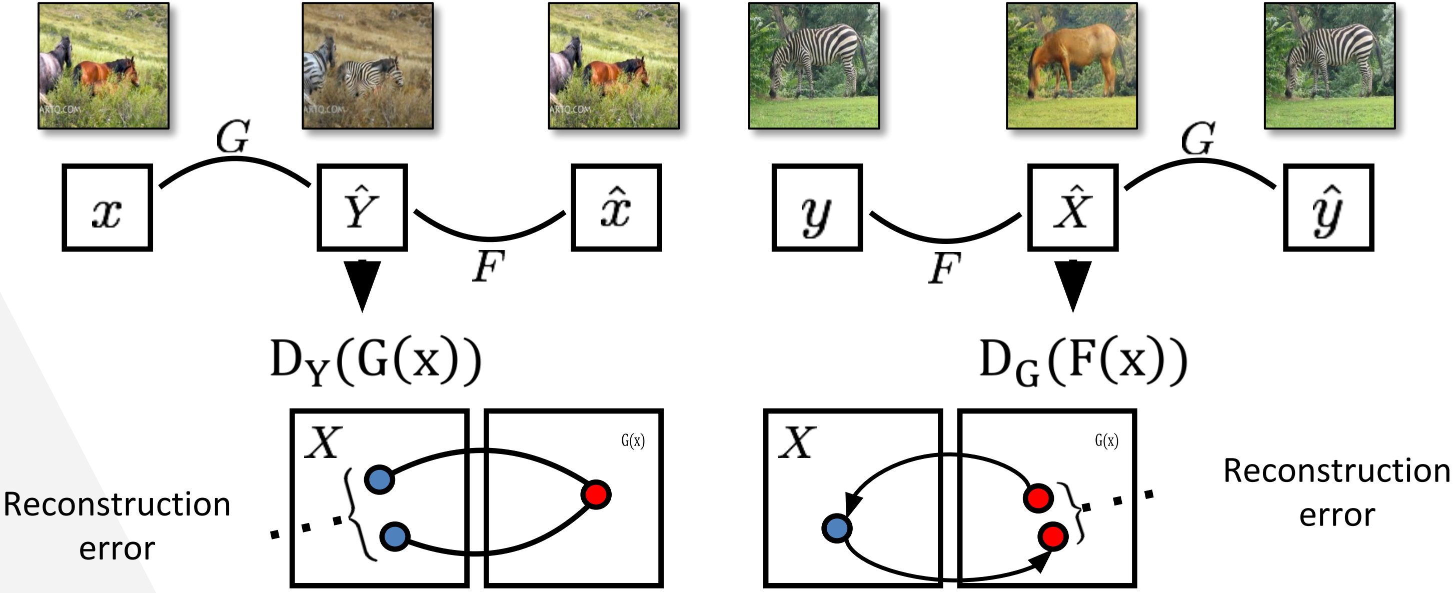


$$\|F(G(x)) - x\|_1$$



[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Cycle Consistency Loss

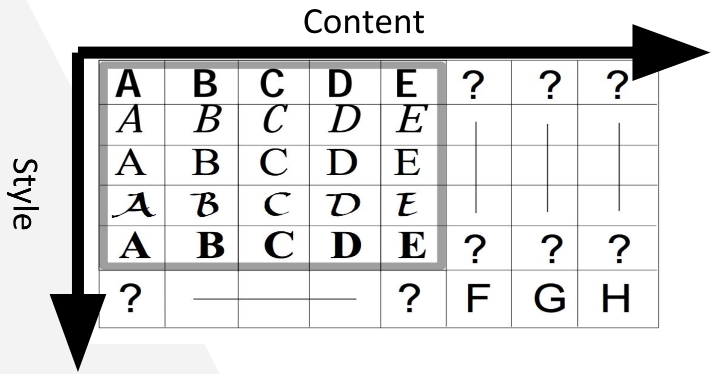


$$\|F(G(x)) - x\|_1 \qquad \|G(F(y)) - y\|_1$$

How do Cycle GANs Work?

Style and Content Separation

Paired Separation



Separating Style and Content with Bilinear Models
[Tenenbaum and Freeman 2000’]

Unpaired Separation

Adversarial Loss: change the Style

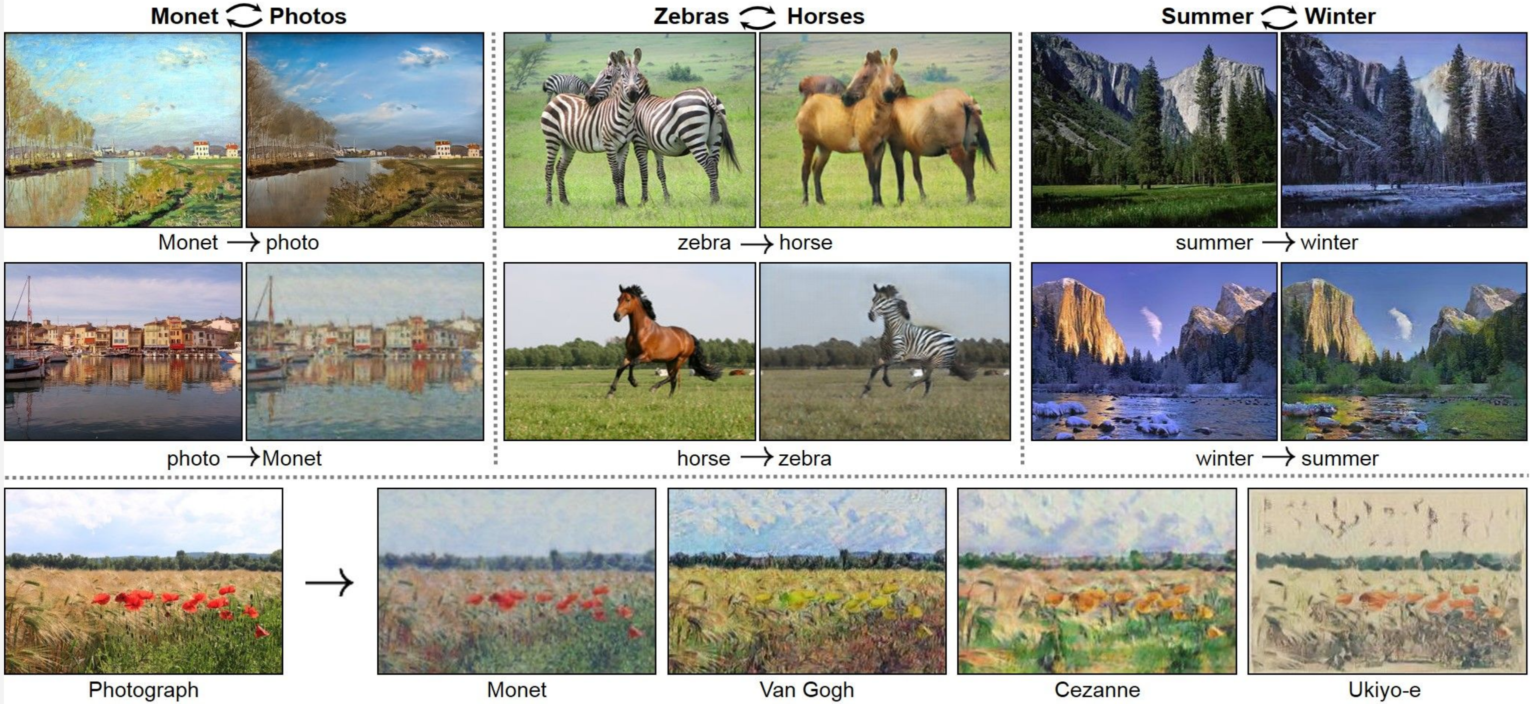
$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x)))]$$

Cycle Consistency Loss: preserve the con

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].$$

Two empirical assumptions:
- content is easy to keep.
- style is easy to change.

Results



Monet ⇄ Photos
Zebras ⇄ Horses
Summer ⇄ Winter
Monet → photo
zebra → horse
summer → winter
photo → Monet
horse → zebra
winter → summer
Photograph
Monet
Van Gogh
Cezanne
Ukiyo-e

Results

THANKS!

Any questions?

You can find us at analyticsclub.iitm@gmail.com