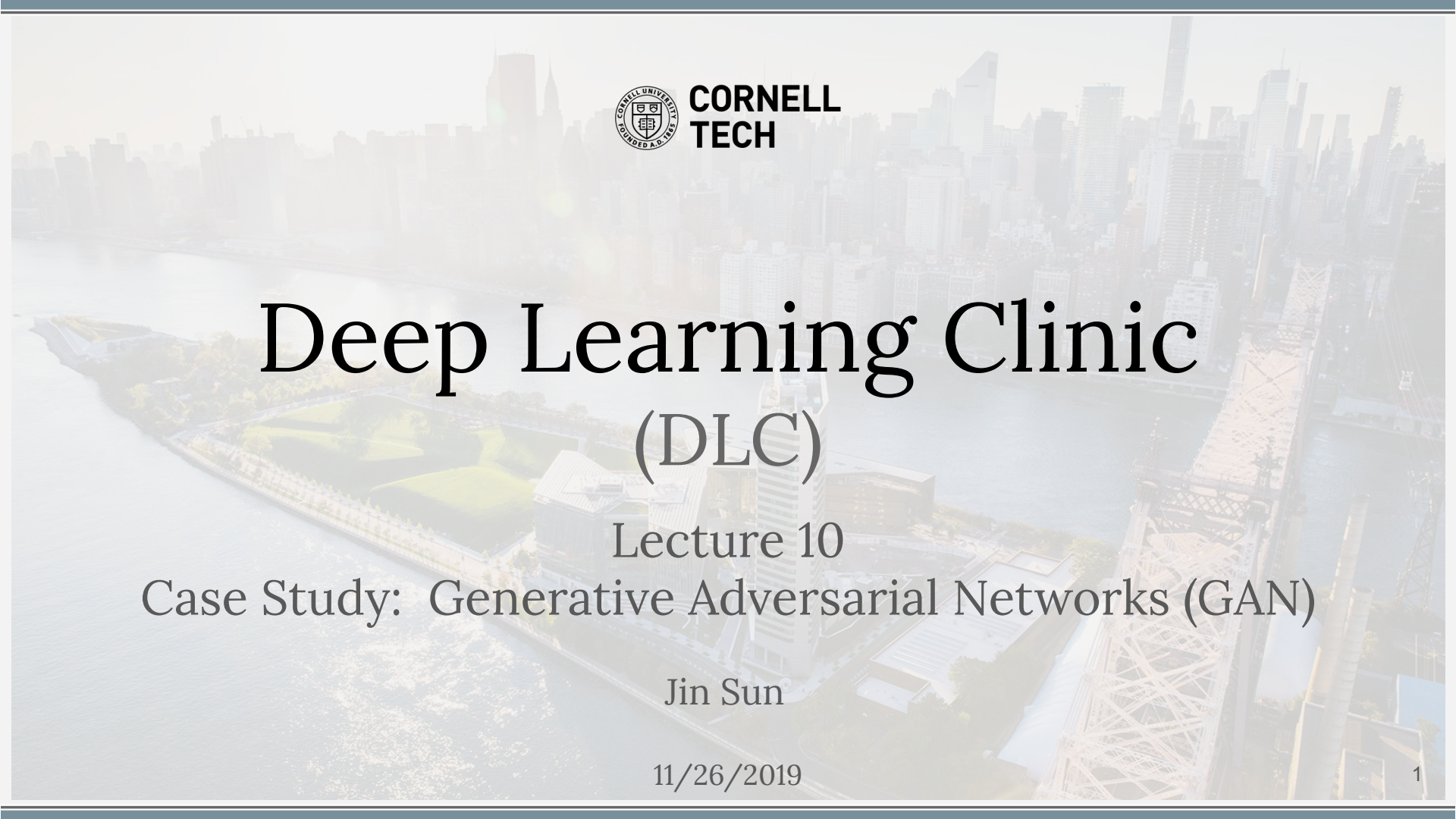


CORNELL TECH

Deep Learning Clinic (DLC)

Lecture 10
Case Study: Generative Adversarial Networks (GAN)

Jin Sun

11/26/2019

Today - GAN

- **Overview - Generative Models**
- Generative Adversarial Networks (GAN)
- Conditional GANs
- GAN Tricks
- Not Really GAN -- Adversarial Learning

Recall: Data and Neural Network Models

Static Data	Dynamic Data	Unsupervised Data
Convolutional Neural Networks	Recurrent Neural Networks	Generative Neural Networks

What to Learn When There Is No Label

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

DOG, **DOG**, **CAT**

Object Detection

Slides from Stanford CS231n

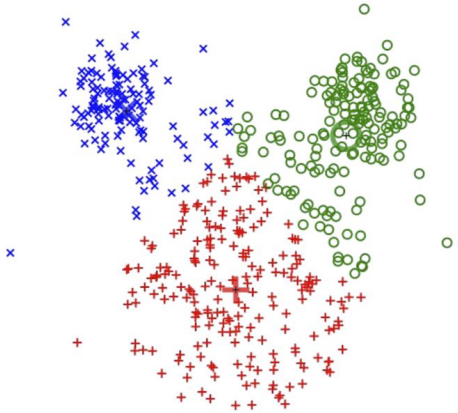
What to Learn When There Is No Label

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

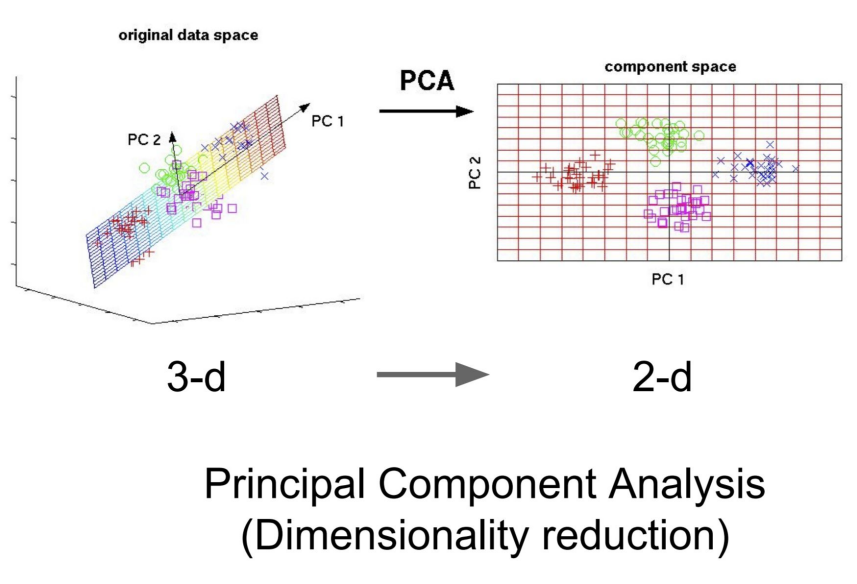
Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

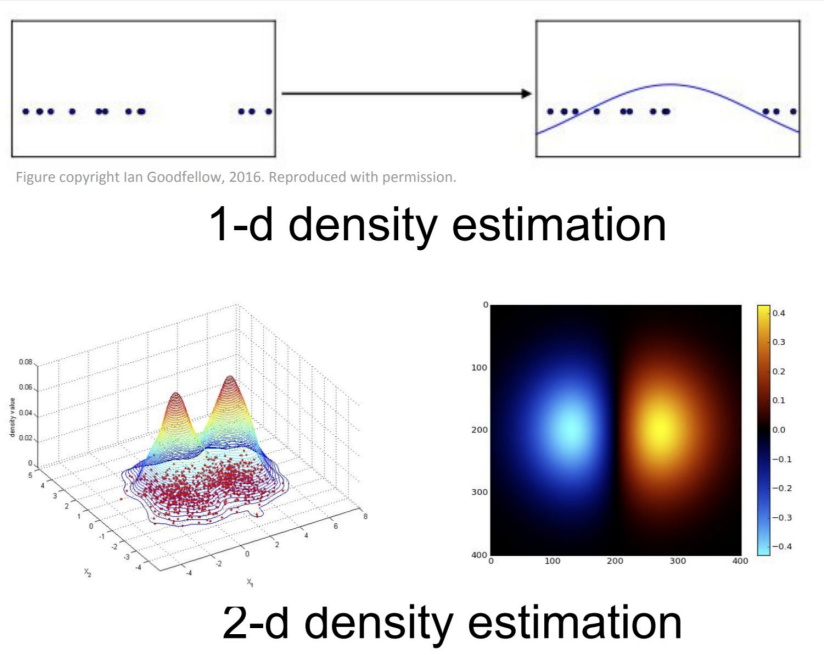


K-means clustering

Slides from Stanford CS231n

What to Learn When There Is No Label

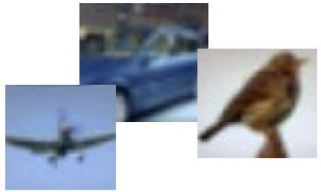




Slides from Stanford CS231n

Probabilistic View

Given training data, generate new samples from same distribution



Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

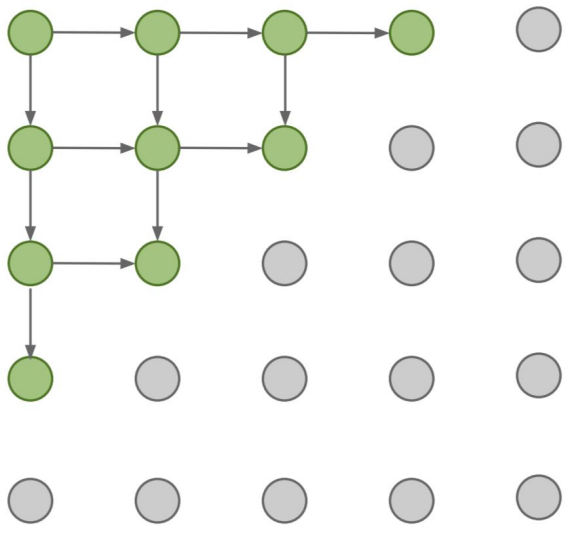
Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

Discriminative models and **Generative** models:

x: data, y: label
Joint probability: p(x, y) = p(y|x) **p(x)**

Slides from Stanford CS231n

Generative Modeling Is Hard



$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

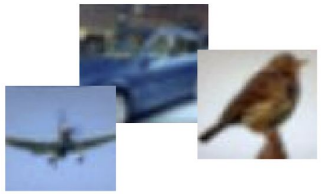
Likelihood of image x

Probability of i'th pixel value given all previous pixels

Slides from Stanford CS231n

Explicit Modeling

Given training data, generate new samples from same distribution



Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

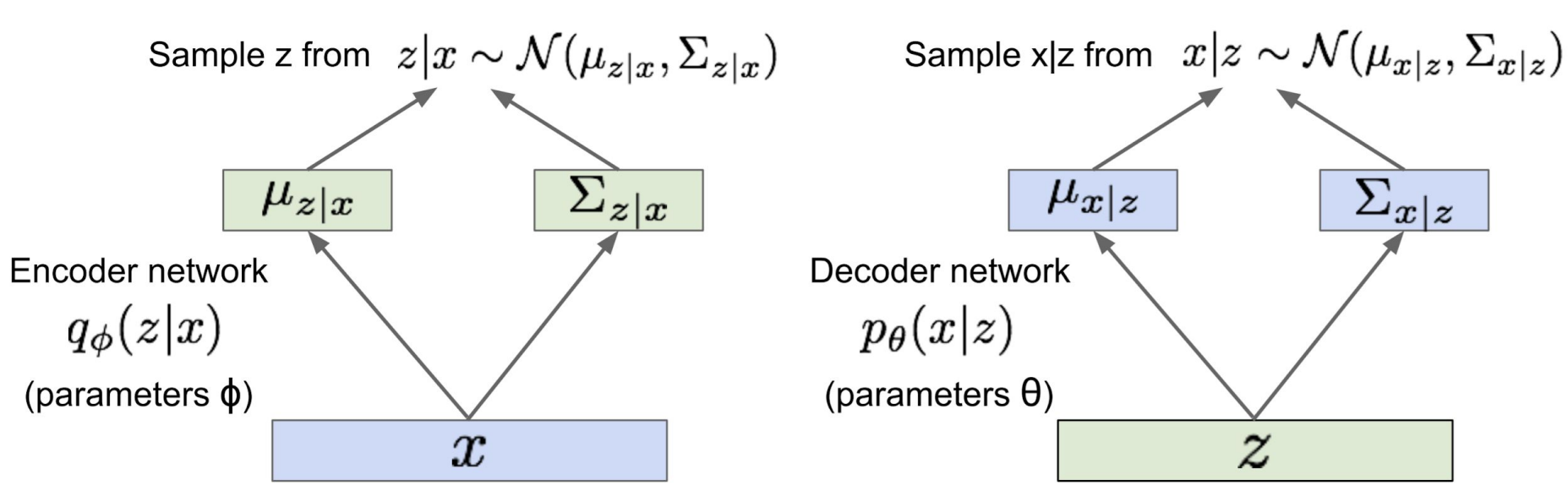
Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

Explicitly define the form of p(x):

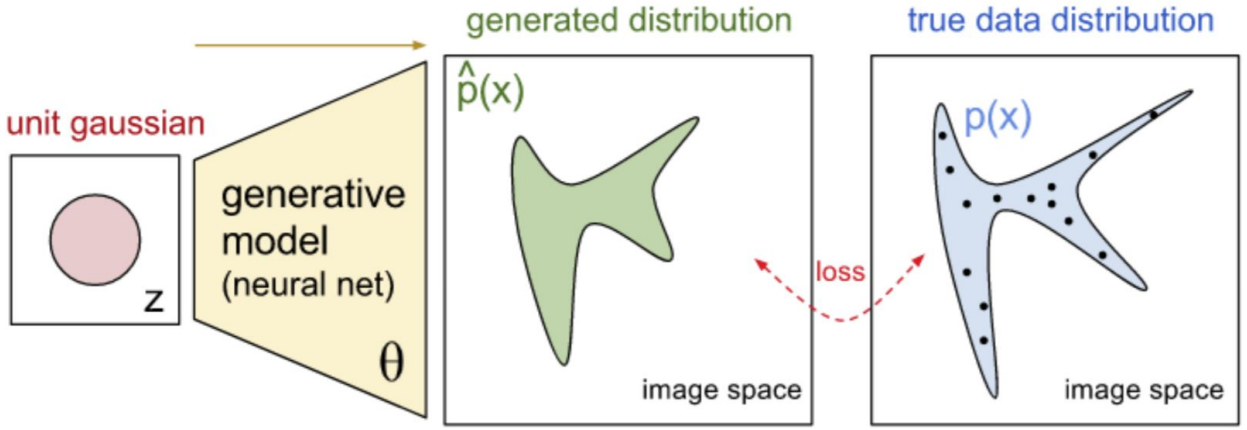
Belief Nets
Variational Autoencoder
Boltzmann Machine
...

Slides from Stanford CS231n

Variational Autoencoder

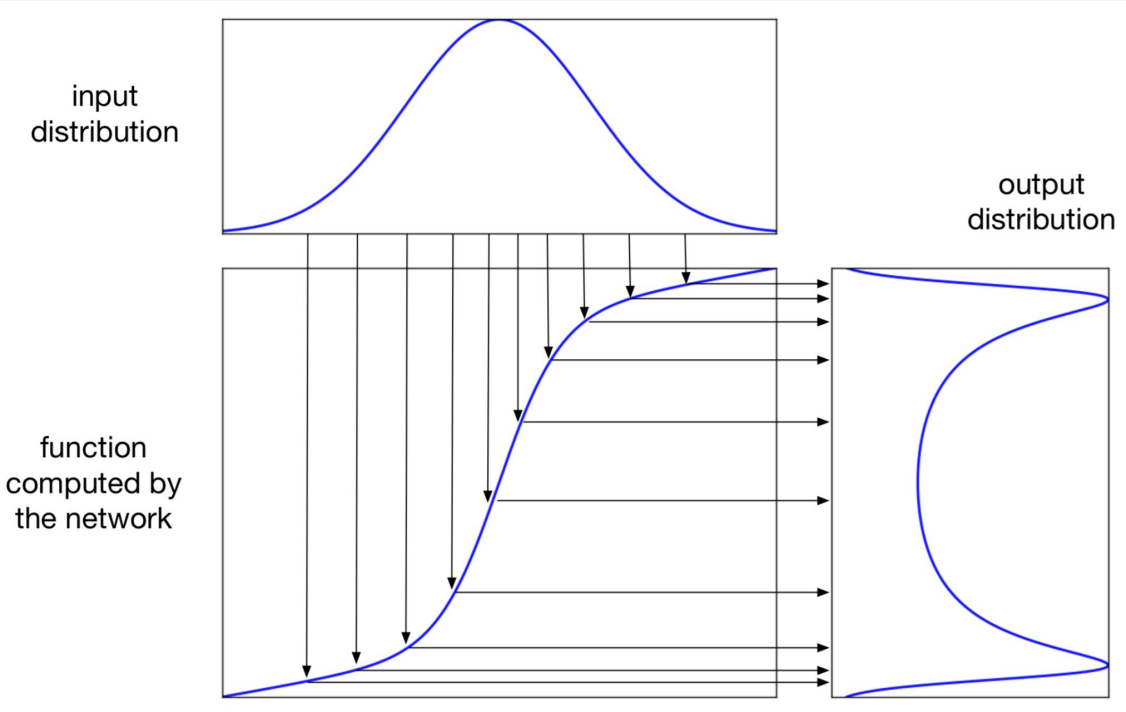


Implicit Modeling



https://blog.openai.com/generative-models/

Implicit Modeling



Slides from Stanford CSC321_2018

Taxonomy of Generative Models

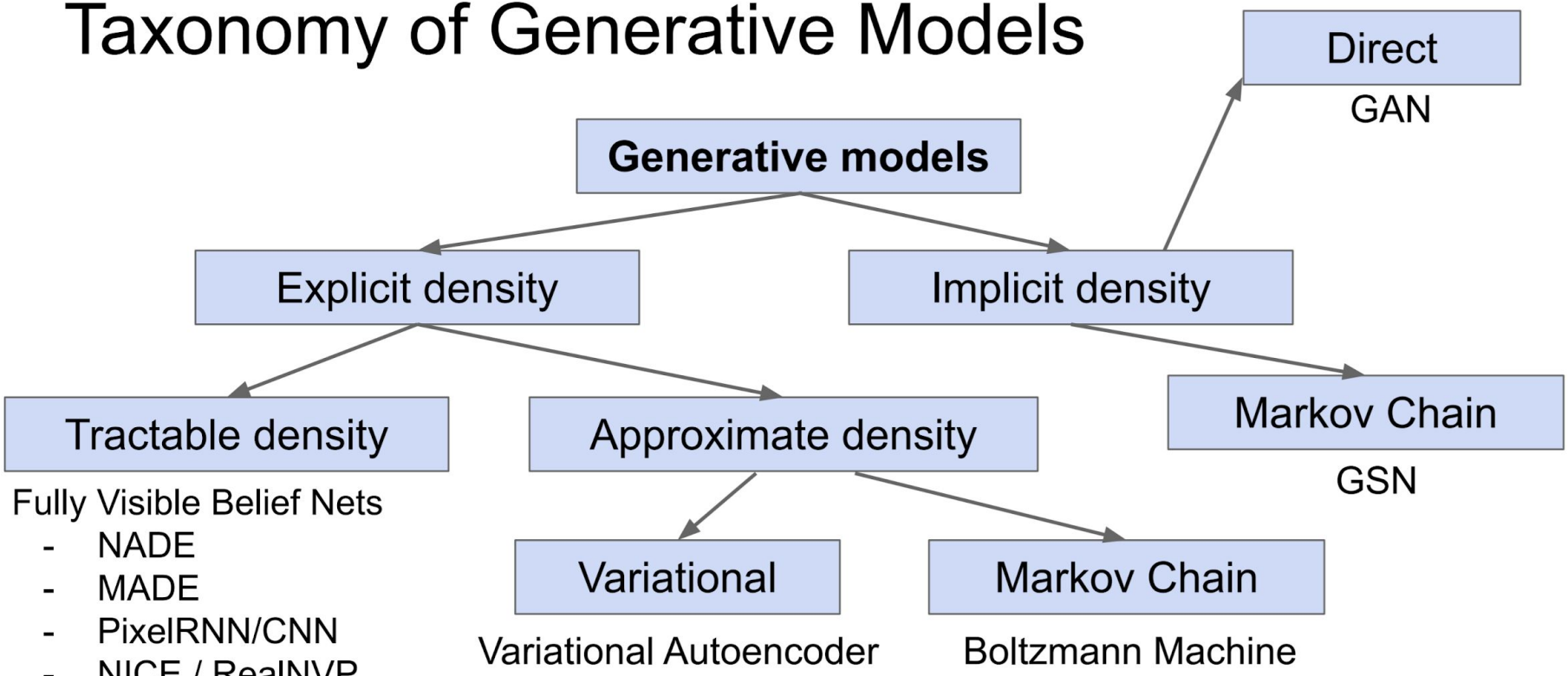


Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.

Why Generative Models?

Excellent test of our ability to use high-dimensional, complicated probability distributions

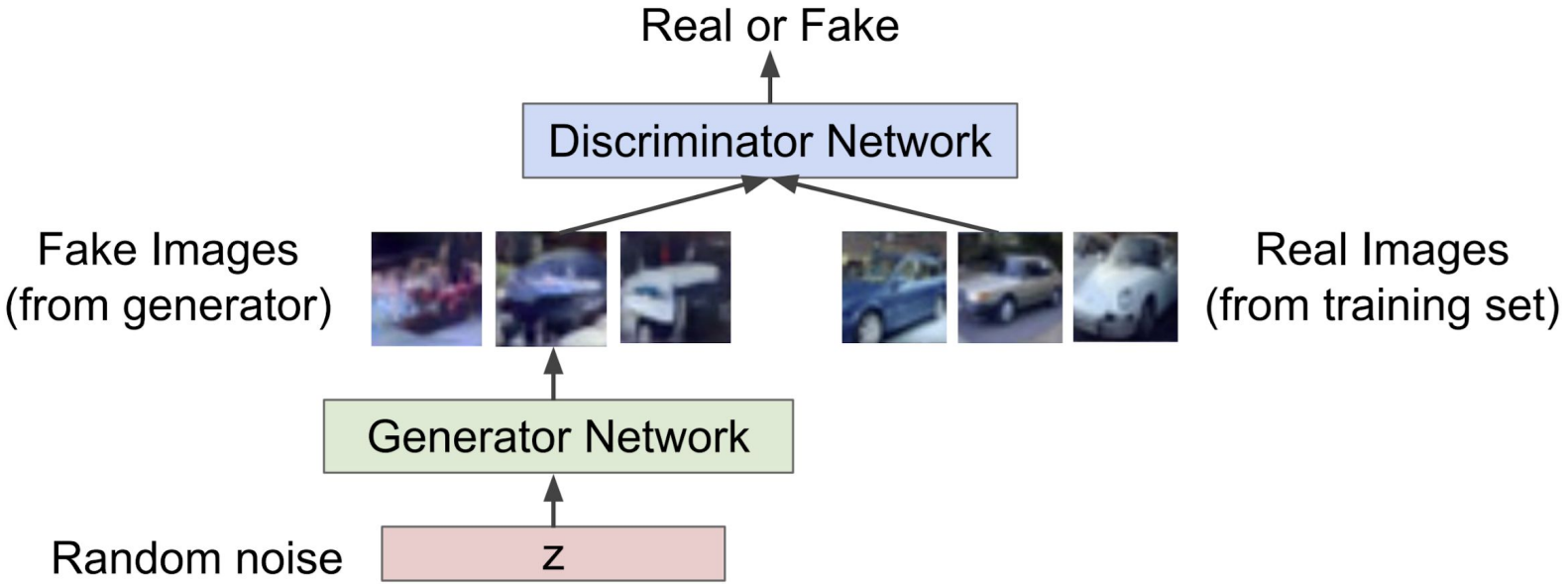
- Simulate possible futures for planning or simulated RL
- Missing data
    - Semi-supervised learning
- Multi-modal outputs
- Realistic generation tasks

Today - GAN

- Overview - Generative Models
- **Generative Adversarial Networks (GAN)**
- Conditional GANs
- GAN Tricks
- Not Really GAN -- Adversarial Learning

GAN Setting: A Two-Player Game

Generator network: try to fool the discriminator by generating real-looking images
Discriminator network: try to distinguish between real and fake images



Slides from Stanford CS231n

Generator network: try to fool the discriminator by generating real-looking images
Discriminator network: try to distinguish between real and fake images

Train jointly in **minimax game**

Discriminator outputs likelihood in (0,1) of real image

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log \underbrace{D_{\theta_d}(x)}_{\text{Discriminator output for real data x}} + \mathbb{E}_{z \sim p(z)} \log(1 - \underbrace{D_{\theta_d}(G_{\theta_g}(z))}_{\text{Discriminator output for generated fake data G(z)}}) \right]$$

Training Steps

Alternate between:

1. **Gradient ascent** on discriminator

$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

2. **Gradient descent** on generator

$$\min_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

Slides from Stanford CS231n

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)}))) \right]$$

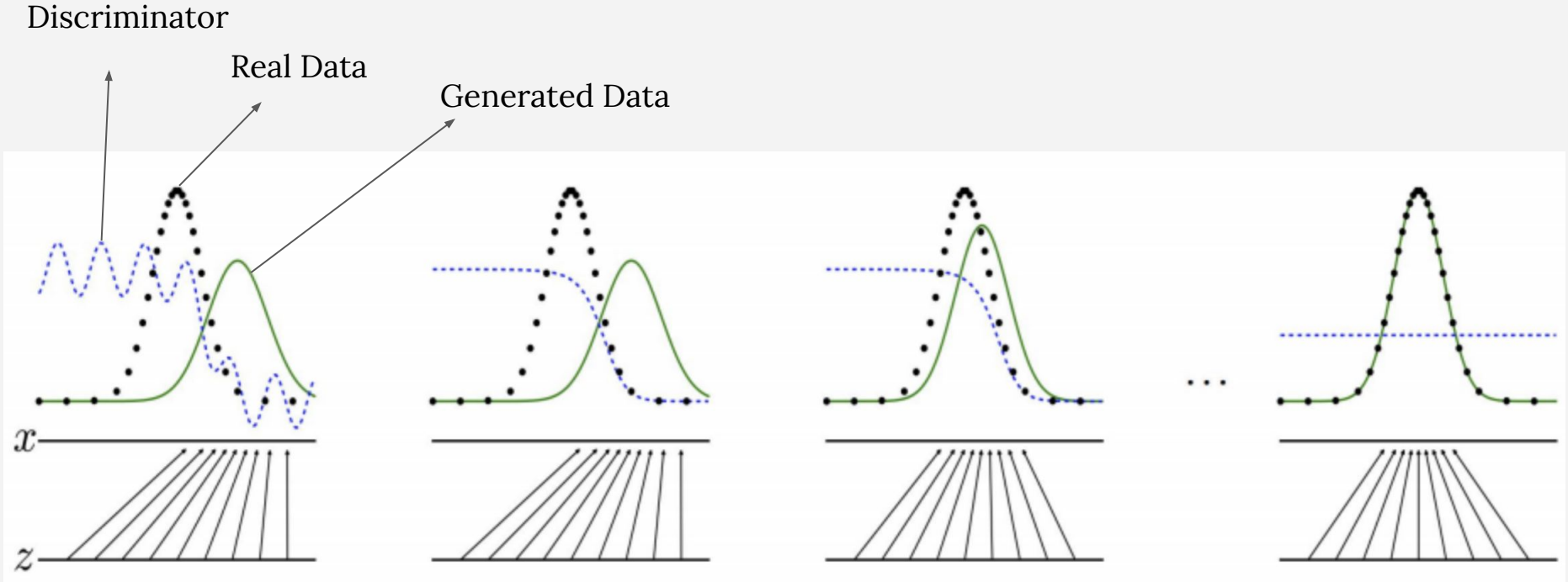
end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by ascending its stochastic gradient (improved objective):

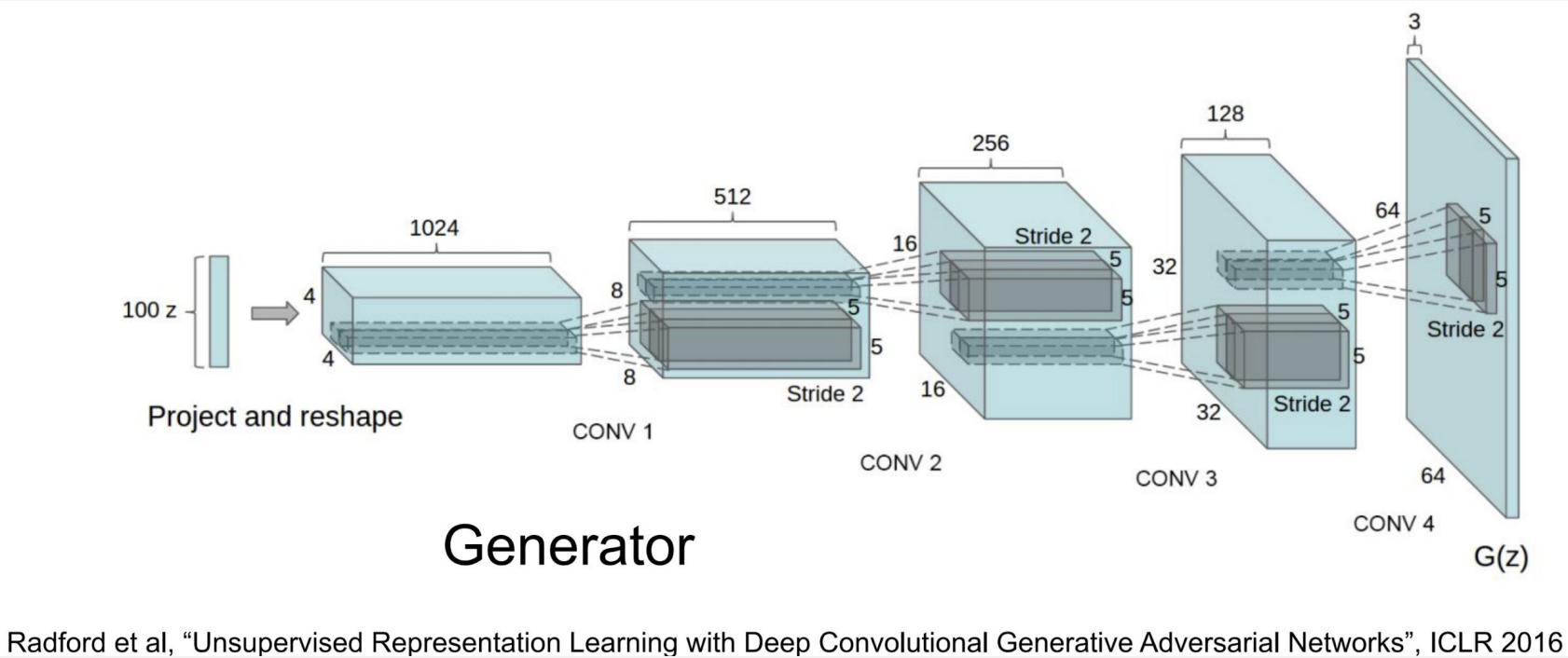
$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$$

end for

Training Dynamics



Network Structure for Generating Images



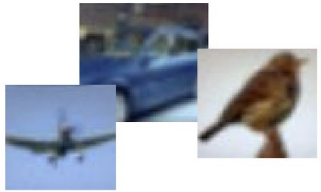
Radford et al, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, ICLR 2016

Today - GAN

- Overview - Generative Models
- Generative Adversarial Networks (GAN)
- **Conditional GANs**
- GAN Tricks
- Not Really GAN -- Adversarial Learning

Probabilistic View

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 $p_{data}(x)$

Discriminative models and **Generative** models:

Conditional Generative models

x: data, y: label

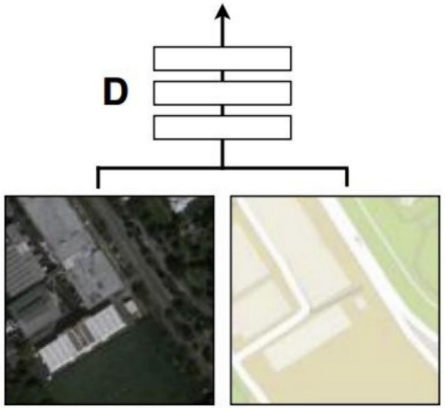
Joint probability: p(x, y) = p(y|x) **p(x)**

= **p(x|y)** p(y)

Slides from Stanford CS231n

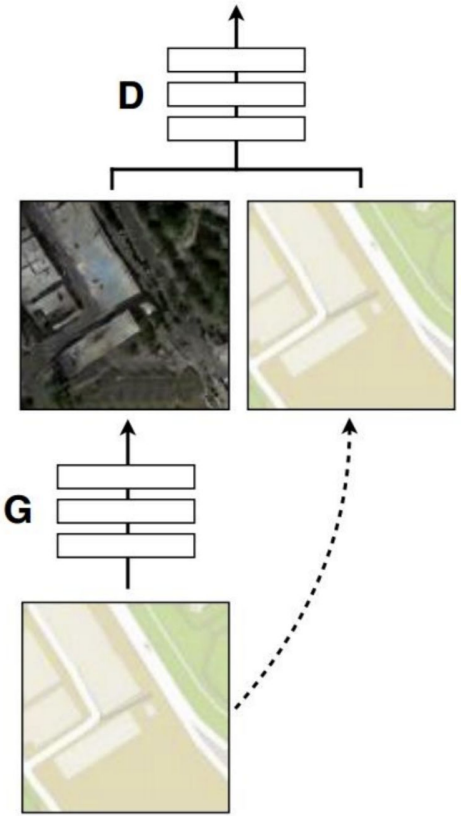
Positive examples

Real or fake pair?



Negative examples

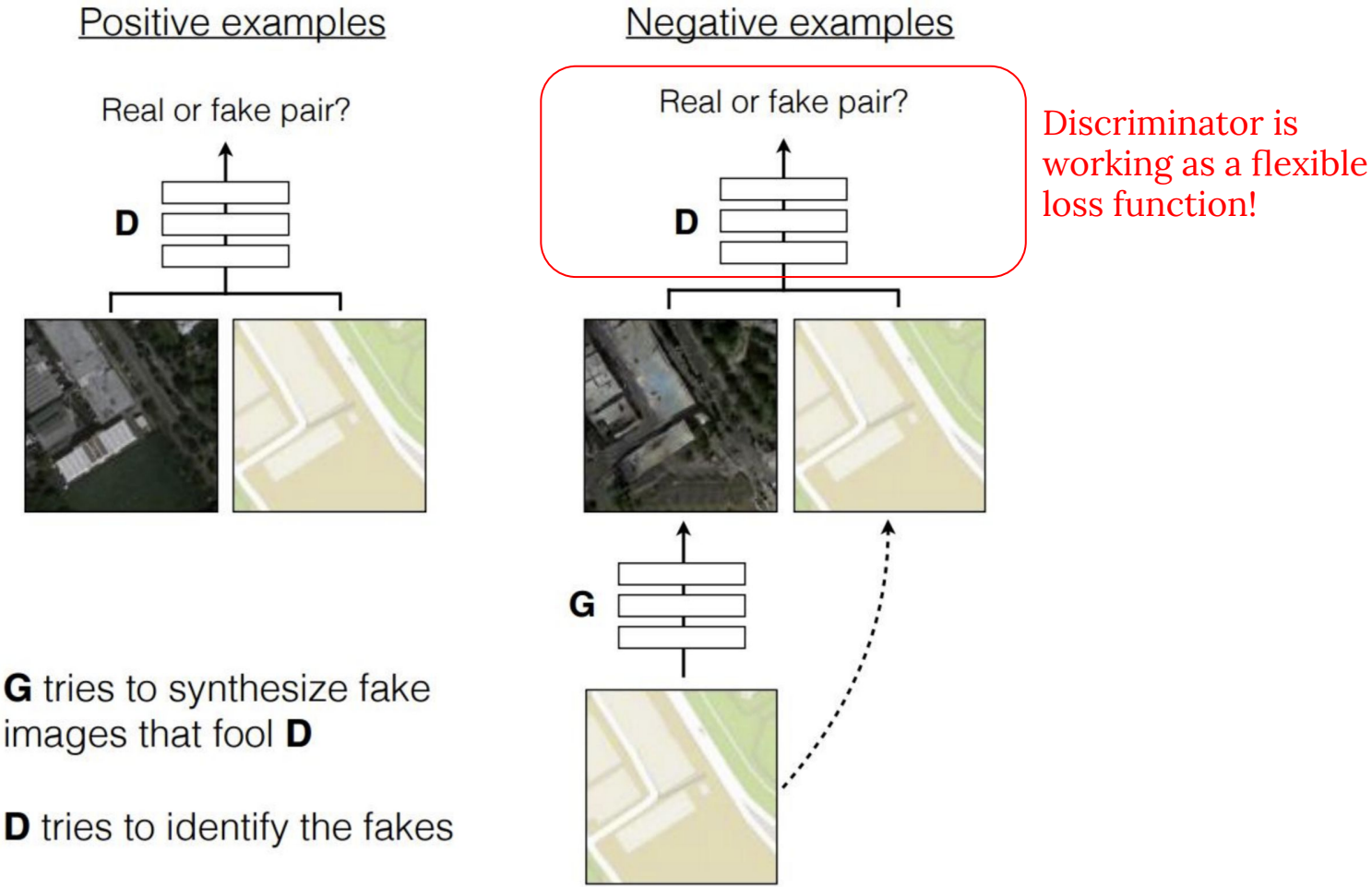
Real or fake pair?



G tries to synthesize fake images that fool **D**

D tries to identify the fakes

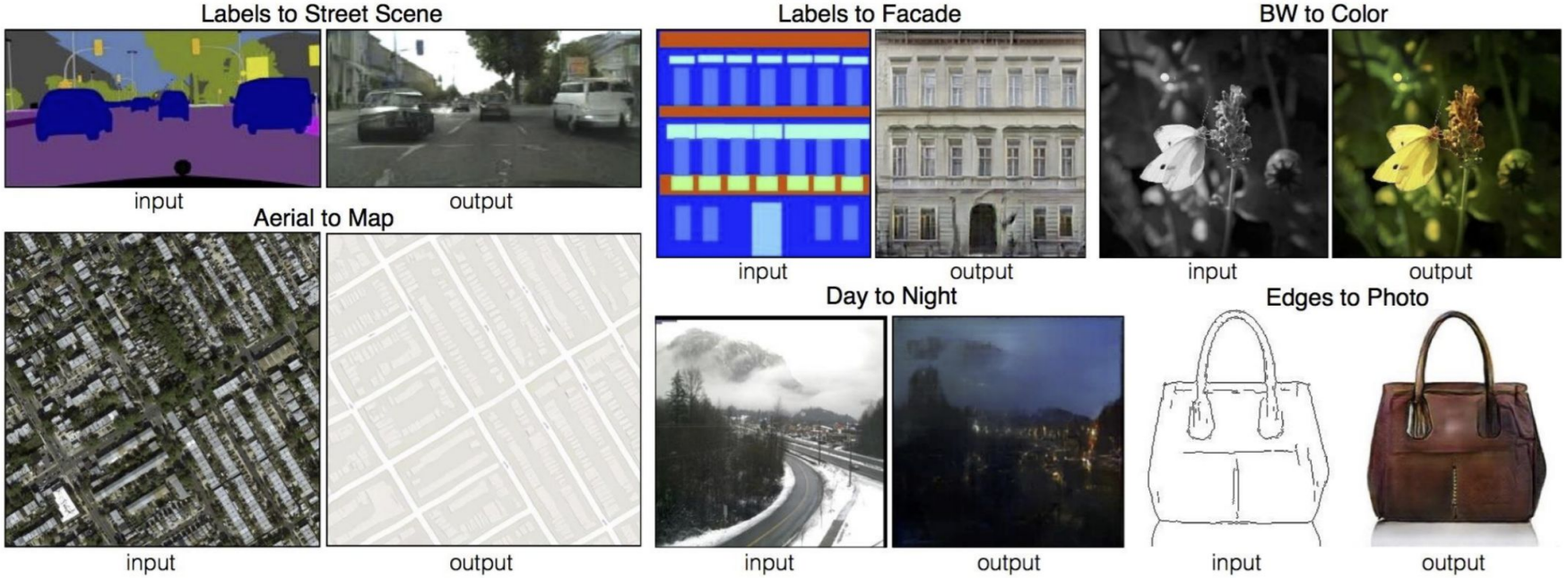
[Isola et al. CVPR 2017]



Discriminator is working as a flexible loss function!

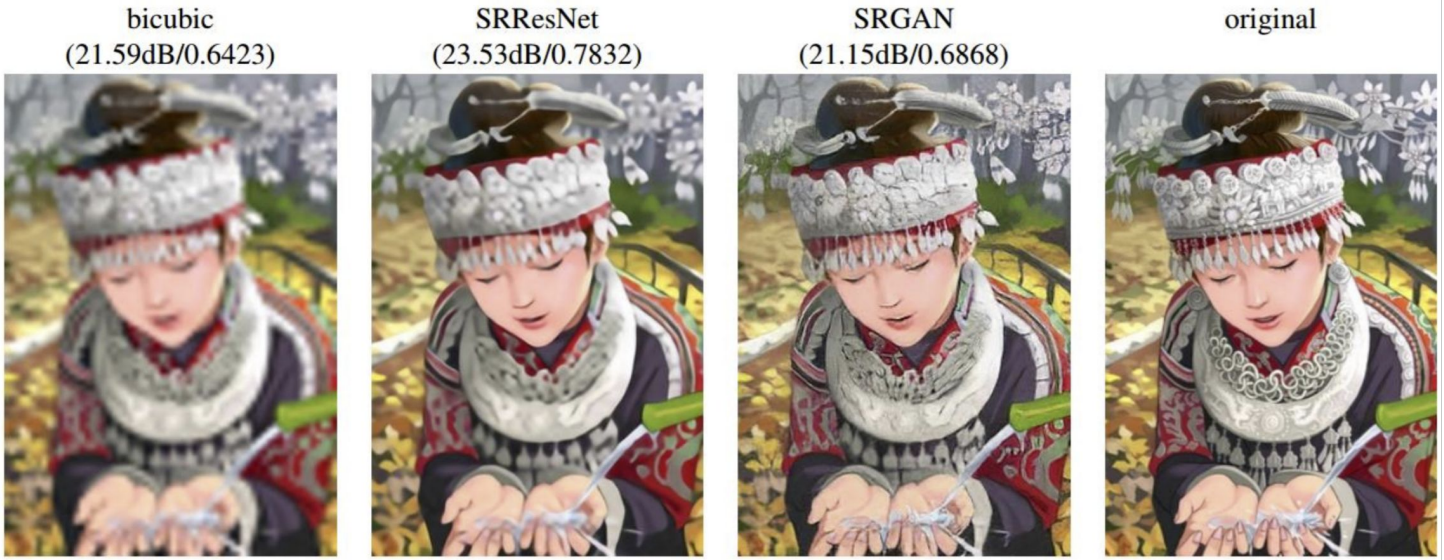
[Isola et al. CVPR 2017]

Image-to-Image Translation



Isola et al. CVPR 2017

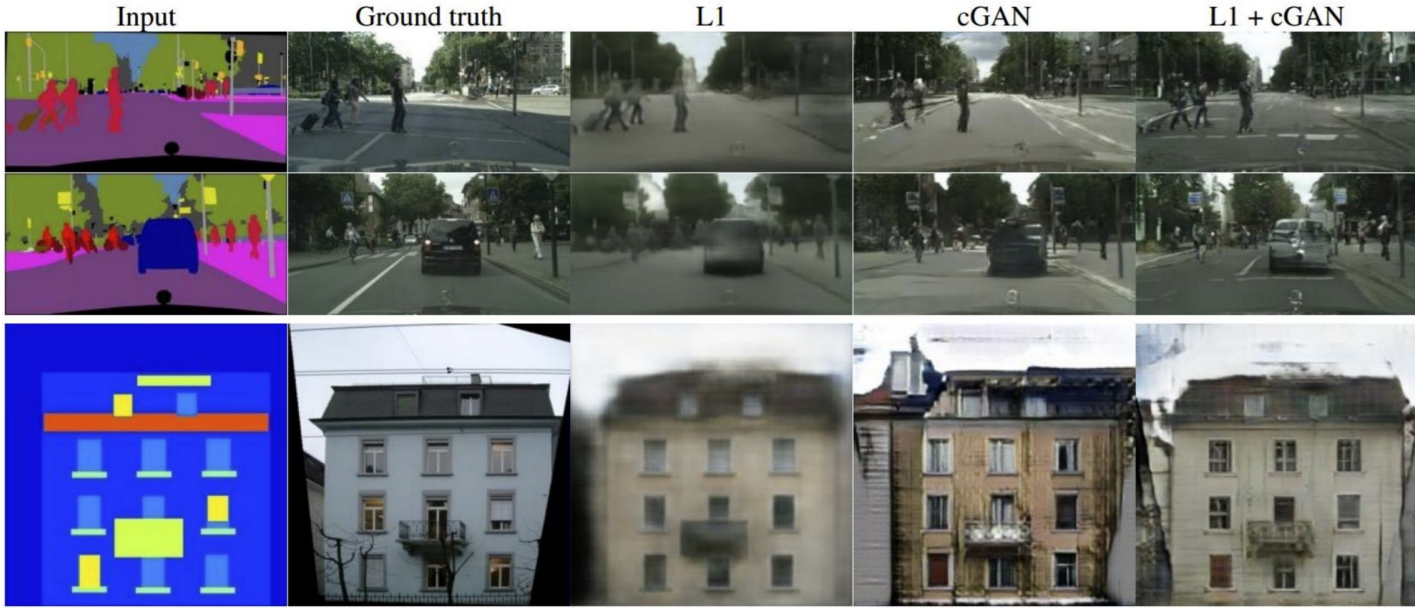
Image Super-Resolution



- Conditional on low-resolution input image

[Ledig et al. CVPR 2017]

Label2Image



[Isola et al. CVPR 2017]

Edges2Image

[Isola et al. CVPR 2017]

Text2Image

this small bird has a pink breast and crown, and black primaries and secondaries.

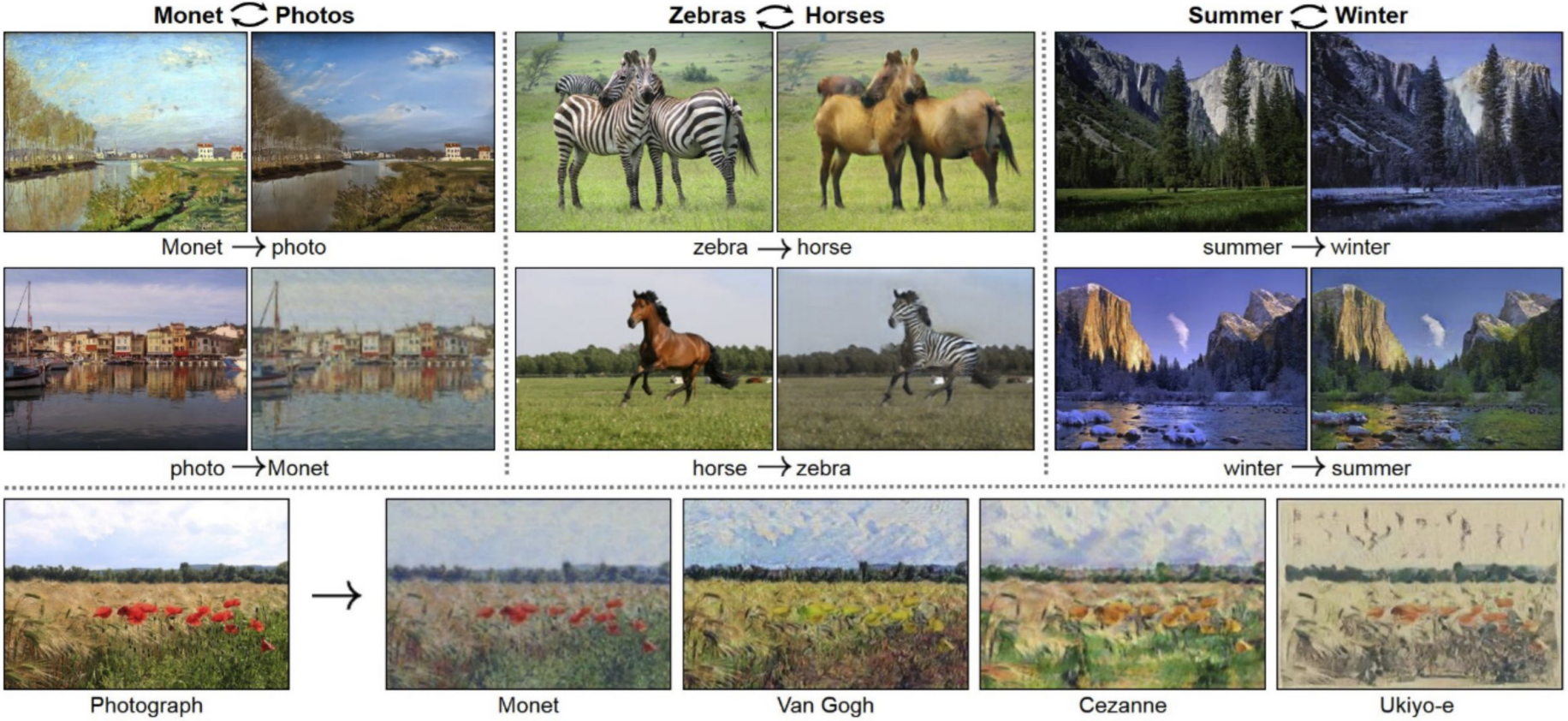
this magnificent fellow is almost all black with a red crest, and white cheek patch.

the flower has petals that are bright pinkish purple with white stigma

this white and yellow flower have thin white petals and a round yellow stamen

[Reed et al. ICML 2016]

Style Transfer



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

Today - GAN

- Overview - Generative Models
- Generative Adversarial Networks (GAN)
- Conditional GANs
- **GAN Tricks**
- Not Really GAN -- Adversarial Learning

GANs are Notorious Difficult to Train

Training one network is hard enough...

Now you have to train two at the same time!

2. Gradient descent on generator

$$\min_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

In fact the original formulation is not easy to learn.

GANs are Notorious Difficult to Train

Alternate between:

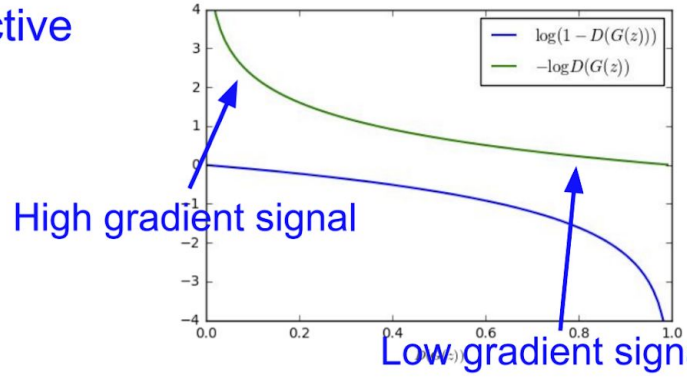
1. **Gradient ascent** on discriminator

$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

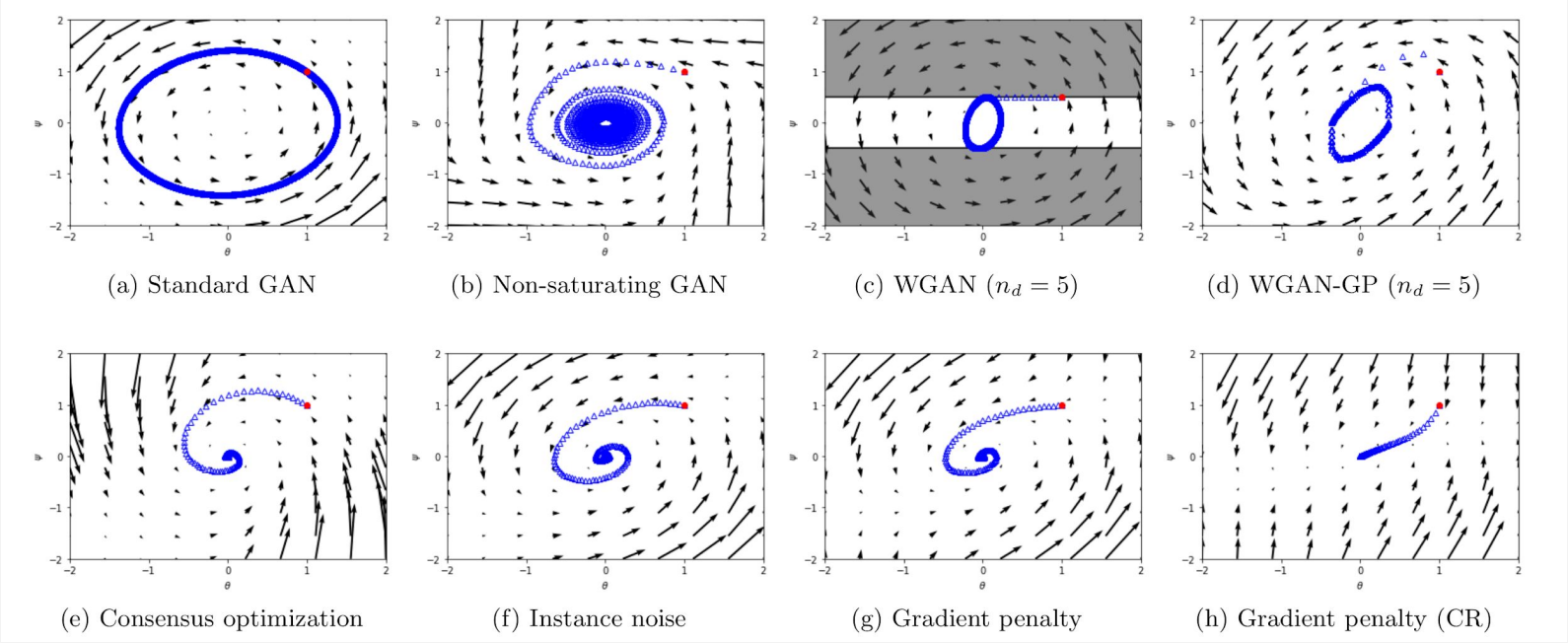
2. **Instead: Gradient ascent** on generator, different objective

$$\max_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z)))$$

Instead of minimizing likelihood of discriminator being correct, now maximize likelihood of discriminator being wrong.
Same objective of fooling discriminator, but now higher gradient signal for bad samples => works much better! Standard in practice.



(Non-)Convergence

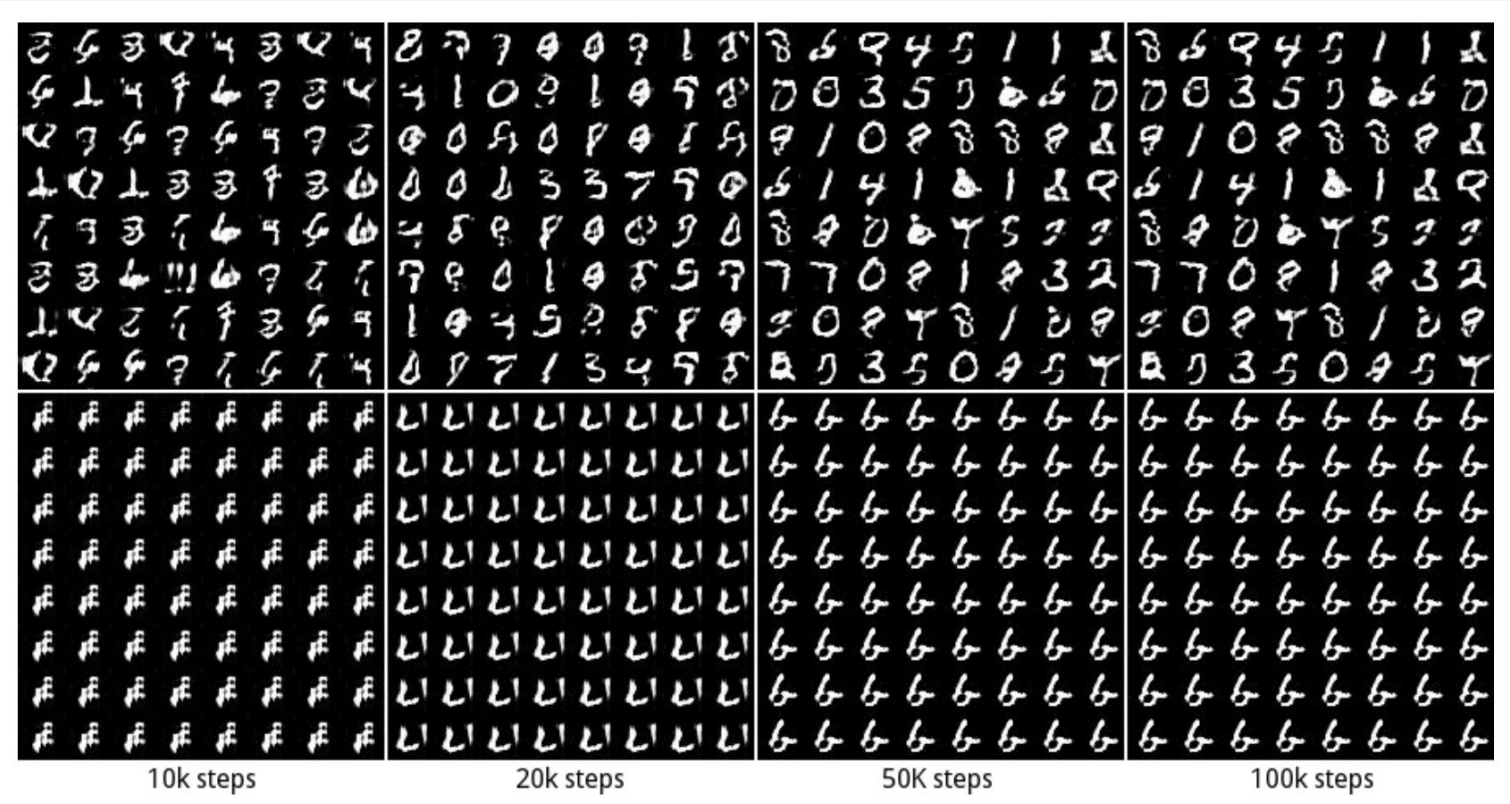


(a) Standard GAN (b) Non-saturating GAN (c) WGAN ($n_d = 5$) (d) WGAN-GP ($n_d = 5$)

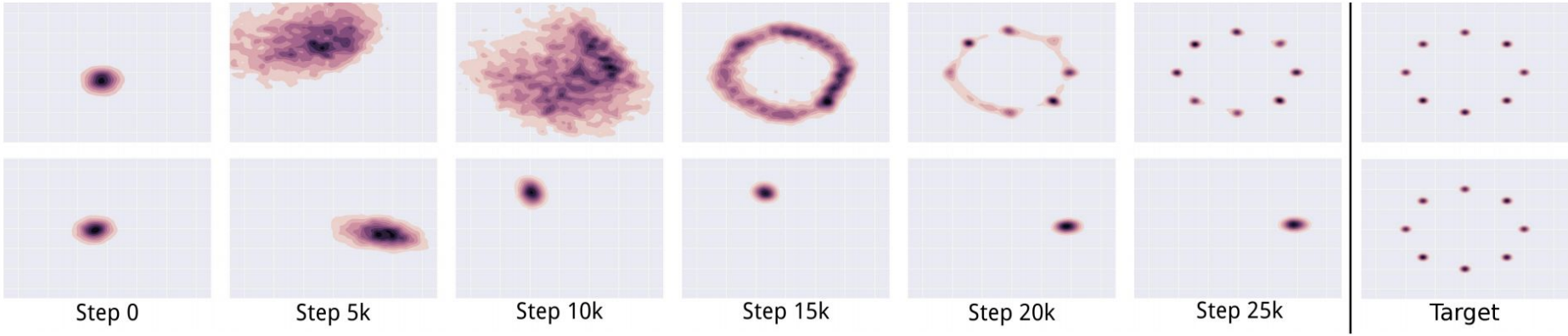
(e) Consensus optimization (f) Instance noise (g) Gradient penalty (h) Gradient penalty (CR)

https://avg.is.tuebingen.mpg.de/projects/convergence-and-stability-of-gan-training

Mode Collapse



Unroll GAN



GAN Loss

GAN	Discriminator Loss	Generator Loss
MM GAN	$\mathcal{L}_{\mathrm{D}}^{\mathrm{GAN}} = -\mathbb{E}_{x\sim p_d}[\log(D(x))] - \mathbb{E}_{\hat{x}\sim p_g}[\log(1-D(\hat{x}))]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{GAN}} = \mathbb{E}_{\hat{x}\sim p_g}[\log(1-D(\hat{x}))]$
NS GAN	$\mathcal{L}_{\mathrm{D}}^{\mathrm{NSGAN}} = -\mathbb{E}_{x\sim p_d}[\log(D(x))] - \mathbb{E}_{\hat{x}\sim p_g}[\log(1-D(\hat{x}))]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{NSGAN}} = -\mathbb{E}_{\hat{x}\sim p_g}[\log(D(\hat{x}))]$
WGAN	$\mathcal{L}_{\mathrm{D}}^{\mathrm{WGAN}} = -\mathbb{E}_{x\sim p_d}[D(x)] + \mathbb{E}_{\hat{x}\sim p_g}[D(\hat{x})]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{WGAN}} = -\mathbb{E}_{\hat{x}\sim p_g}[D(\hat{x})]$
WGAN GP	$\mathcal{L}_{\mathrm{D}}^{\mathrm{WGANGP}} = \mathcal{L}_{\mathrm{D}}^{\mathrm{WGAN}} + \lambda\mathbb{E}_{\hat{x}\sim p_g}[(\|\nabla D(\alpha x + (1-\alpha\hat{x})\|_2 - 1)^2]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{WGANGP}} = -\mathbb{E}_{\hat{x}\sim p_g}[D(\hat{x})]$
LS GAN	$\mathcal{L}_{\mathrm{D}}^{\mathrm{LSGAN}} = -\mathbb{E}_{x\sim p_d}[(D(x)-1)^2] + \mathbb{E}_{\hat{x}\sim p_g}[D(\hat{x})^2]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{LSGAN}} = -\mathbb{E}_{\hat{x}\sim p_g}[(D(\hat{x}-1)^2]$
DRAGAN	$\mathcal{L}_{\mathrm{D}}^{\mathrm{DRAGAN}} = \mathcal{L}_{\mathrm{D}}^{\mathrm{GAN}} + \lambda\mathbb{E}_{\hat{x}\sim p_d+\mathcal{N}(0,c)}[(\|\nabla D(\hat{x})\|_2 - 1)^2]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{DRAGAN}} = \mathbb{E}_{\hat{x}\sim p_g}[\log(1-D(\hat{x}))]$
BEGAN	$\mathcal{L}_{\mathrm{D}}^{\mathrm{BEGAN}} = \mathbb{E}_{x\sim p_d}[\|x - \mathrm{AE}(x)\|_1] - k_t\mathbb{E}_{\hat{x}\sim p_g}[\|\hat{x} - \mathrm{AE}(\hat{x})\|_1]$	$\mathcal{L}_{\mathrm{G}}^{\mathrm{BEGAN}} = \mathbb{E}_{\hat{x}\sim p_g}[\|\hat{x} - \mathrm{AE}(\hat{x})\|_1]$

[Lucic, Kurach et al. (2018): Are GANs Created Equal? A Large-Scale Study]

GAN Generated Image Evaluation

Inception Score

Using an Inception v3 Network pretrained on ImageNet.

$$\mathrm{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x} \sim p_g} D_{KL}(p(y|\mathbf{x}) \parallel p(y)) \right)$$

$$p(y) = \int_{\mathbf{x}} p(y|\mathbf{x}) p_g(\mathbf{x})$$

GAN Generated Image Evaluation

Inception Score

Using an Inception v3 Network pretrained on ImageNet.

$$\mathrm{IS}(G) \approx \exp\left(\frac{1}{N}\sum_{i=1}^{N} D_{KL}(p(y|\mathbf{x}^{(i)} \parallel \hat{p}(y))\right)$$

$$\hat{p}(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|\mathbf{x}^{(i)})$$

BigGAN

Scaling up GAN training! 2x, 4x parameters; 8x batch sizes

Many tricks: SAGAN, spectral norm, Hinge Loss, orthogonal init...

Batch size 8x

Width↑ 50%

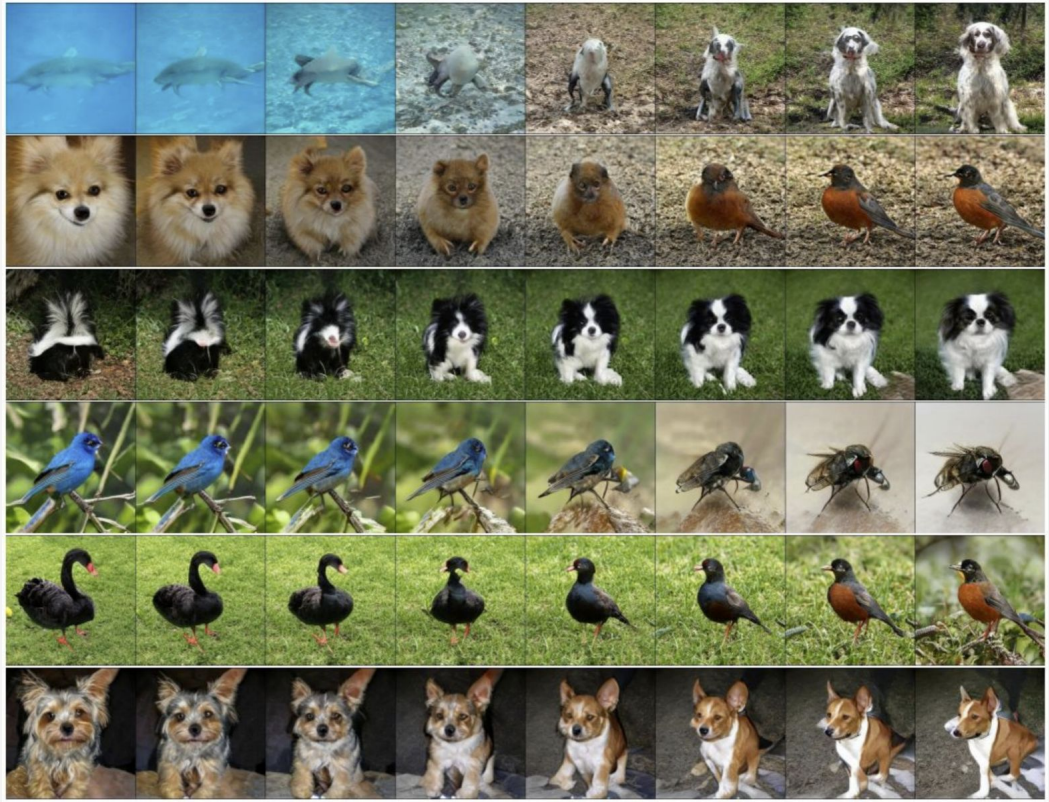
46%↑

21%↑

Batch	Ch.	Param (M)	Shared	Hier.	Ortho.	Itr $\times 10^3$	FID	IS
256	64	81.5	SA-GAN Baseline			1000	18.65	52.52
512	64	81.5	✗	✗	✗	1000	15.30	58.77(±1.18)
1024	64	81.5	✗	✗	✗	1000	14.88	63.03(±1.42)
2048	64	81.5	✗	✗	✗	732	12.39	76.85(±3.83)
2048	96	173.5	✗	✗	✗	295(±18)	9.54(±0.62)	92.98(±4.27)
2048	96	160.6	✓	✗	✗	185(±11)	9.18(±0.13)	94.94(±1.32)
2048	96	158.3	✓	✓	✗	152(±7)	8.73(±0.45)	98.76(±2.84)
2048	96	158.3	✓	✓	✓	165(±13)	8.51(±0.32)	99.31(±2.10)
2048	64	71.3	✓	✓	✓	371(±7)	10.48(±0.10)	86.90(±0.61)

BigGAN Examples

BigGAN Examples



“The GAN Zoo”

- GAN - Generative Adversarial Networks
- 3D-GAN - Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling
- acGAN - Face Aging With Conditional Generative Adversarial Networks
- AC-GAN - Conditional Image Synthesis With Auxiliary Classifier GANs
- AdaGAN - AdaGAN: Boosting Generative Models
- AEGAN - Learning Inverse Mapping by Autoencoder based Generative Adversarial Nets
- AffGAN - Amortised MAP Inference for Image Super-resolution
- AL-CGAN - Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts
- ALI - Adversarially Learned Inference
- AM-GAN - Generative Adversarial Nets with Labeled Data by Activation Maximization
- AnoGAN - Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery
- ArtGAN - ArtGAN: Artwork Synthesis with Conditional Categorial GANs
- b-GAN - b-GAN: Unified Framework of Generative Adversarial Networks
- Bayesian GAN - Deep and Hierarchical Implicit Models
- BEGAN - BEGAN: Boundary Equilibrium Generative Adversarial Networks
- BiGAN - Adversarial Feature Learning
- BS-GAN - Boundary-Seeking Generative Adversarial Networks
- CGAN - Conditional Generative Adversarial Nets
- CaloGAN - CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks
- CCGAN - Semi-Supervised Learning with Context-Conditional Generative Adversarial Networks
- CatGAN - Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks
- CoGAN - Coupled Generative Adversarial Networks
- Context-RNN-GAN - Contextual RNN-GANs for Abstract Reasoning Diagram Generation
- C-RNN-GAN - C-RNN-GAN: Continuous recurrent neural networks with adversarial training
- CS-GAN - Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets
- CVAE-GAN - CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training
- CycleGAN - Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
- DTN - Unsupervised Cross-Domain Image Generation
- DCGAN - Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
- DiscoGAN - Learning to Discover Cross-Domain Relations with Generative Adversarial Networks
- DR-GAN - Disentangled Representation Learning GAN for Pose-Invariant Face Recognition
- DualGAN - DualGAN: Unsupervised Dual Learning for Image-to-Image Translation
- EBGAN - Energy-based Generative Adversarial Network
- f-GAN - f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
- FF-GAN - Towards Large-Pose Face Frontalization in the Wild
- GAWWN - Learning What and Where to Draw
- GeneGAN - GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data
- Geometric GAN - Geometric GAN
- GoGAN - Gang of GANs: Generative Adversarial Networks with Maximum Margin Ranking
- GP-GAN - GP-GAN: Towards Realistic High-Resolution Image Blending
- IAN - Neural Photo Editing with Introspective Adversarial Networks
- iGAN - Generative Visual Manipulation on the Natural Image Manifold
- IcGAN - Invertible Conditional GANs for image editing
- ID-CGAN - Image De-raining Using a Conditional Generative Adversarial Network
- Improved GAN - Improved Techniques for Training GANs
- InfoGAN - InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets
- LAGAN - Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis
- LAPGAN - Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks

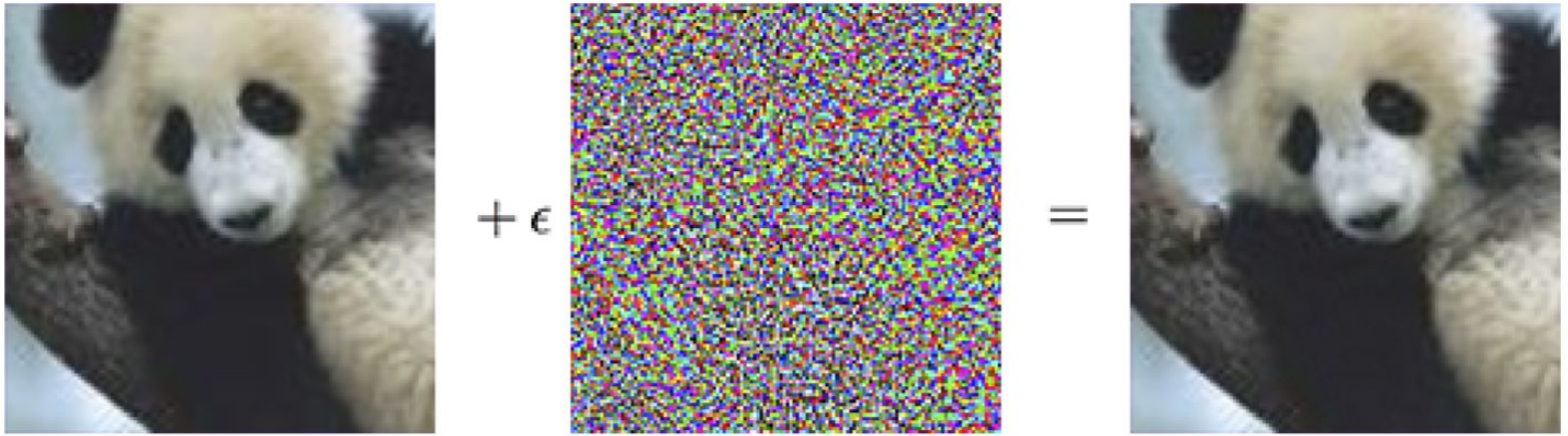
https://github.com/hindupuravinash/the-gan-zoo

Today - GAN

- Overview - Generative Models
- Generative Adversarial Networks (GAN)
- Conditional GANs
- GAN Tricks
- **Not Really GAN -- Adversarial Learning**

Adversarial Examples

Fooling GoogLeNet (Inception) on ImageNet.



"panda"

57.7% confidence

"gibbon"

99.3% confidence

Adversarial Examples in Real Life

[Athalye et al., 2018]

Adversarial example using 3D-printing . . .

Link

Adversarial Examples in Real Life

[Evtimov et al., 2017]

Figure : Before: Stop sign; After: 45 mph sign

Creating Adversarial Examples

The Fast Gradient Sign Method

$$J(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) \approx J(\boldsymbol{x}, \boldsymbol{\theta}) + (\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} J(\boldsymbol{x}).$$

Maximize

$$J(\boldsymbol{x}, \boldsymbol{\theta}) + (\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} J(\boldsymbol{x})$$

subject to

$$||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon$$

$$\Rightarrow \tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{x})\right).$$

(Goodfellow 2016)

Adversarial Defense

Prevent network overfitting

Adversarial Training

Generate adversarial examples and add them into training

...

Adversarial attack is easy, defense is hard!!

ARE ADVERSARIAL EXAMPLES INEVITABLE?

Ali Shafahi, Ronny Huang, Christoph Studer, Soheil Feizi & Tom Goldstein

ABSTRACT

A wide range of defenses have been proposed to harden neural networks against adversarial attacks. However, a pattern has emerged in which the majority of adversarial defenses are quickly broken by new attacks. Given the lack of success at generating robust defenses, we are led to ask a fundamental question: Are adversarial attacks inevitable? This paper analyzes adversarial examples from a theoretical perspective, and identifies fundamental bounds on the susceptibility of a classifier to adversarial attacks. We show that, for certain classes of problems, adversarial examples are inescapable. Using experiments, we explore the implications of theoretical guarantees for real-world problems and discuss how factors such as dimensionality and image complexity limit a classifier's robustness against adversarial examples.

Summary - GAN

- Overview - Generative Models
- Generative Adversarial Networks (GAN)
- Conditional GANs
- GAN Tricks
- Not Really GAN -- Adversarial Learning

Cool GAN Demos

Image-to-Image Translation

https://affinelayer.com/pixsrv/

NVIDIA Demos (GauGAN and more)

https://www.nvidia.com/en-us/research/ai-playground/