

DeepMind

# Generative adversarial networks

Jeff Donahue & Mihaela Rosca

UCL x DeepMind Lectures



DeepMind

1

# Overview



# Generative models

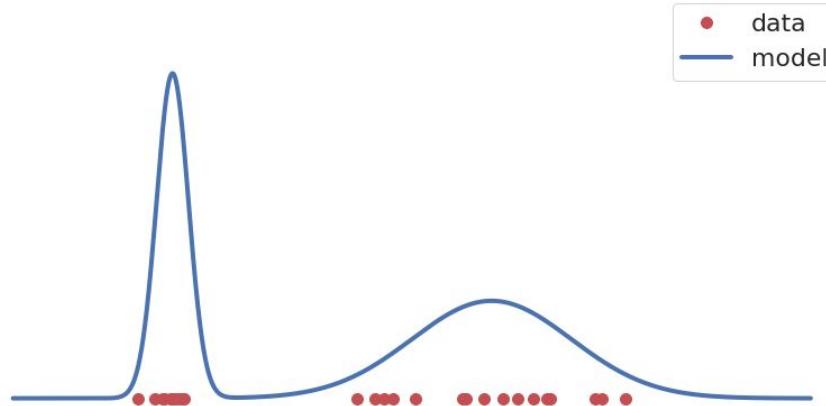
Learn a model of the true (unknown)  
underlying data distribution from samples



# Generative models



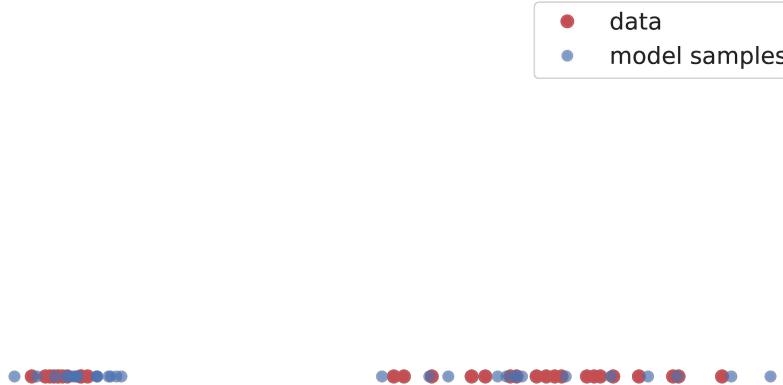
# Generative models



Learning an explicit distribution from data.



# Generative models



Learning an implicit distribution from data.



# Generative model zoo

## Explicit likelihood models:

- Maximum likelihood
  - PPCA, Factor Analysis, Mixture models
  - PixelCNN/PixelRNN
  - Wavenet
  - Autoregressive language models
- Approximate maximum likelihood
  - Boltzmann machines
  - Variational autoencoders

## Implicit models (no likelihoods):

- Generative adversarial networks
- Moment matching networks







# Generative adversarial networks

Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

Learning an implicit model through a two player game.



# Generative adversarial networks

## Discriminator

Learns to distinguish between real and generated data.

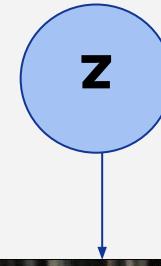


vs



## Generator

Learns to generate data to “fool” the discriminator.

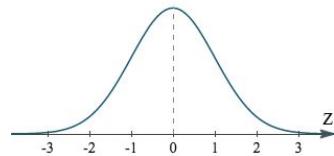
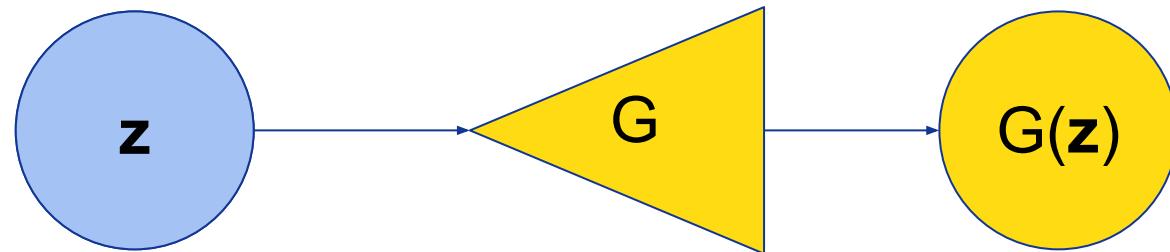


# Generator

latent (“noise”) vector  
 $\mathbf{z} \sim P(\mathbf{z})$

generator G:  
a deep neural network

generated data  
 $G(\mathbf{z})$

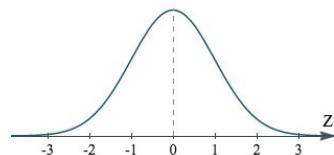
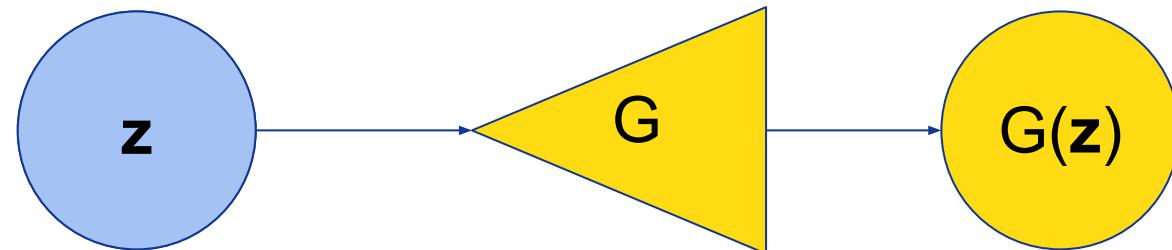


# Generator

latent (“noise”) vector  
 $z \sim P(z)$

generator G:  
a deep neural network

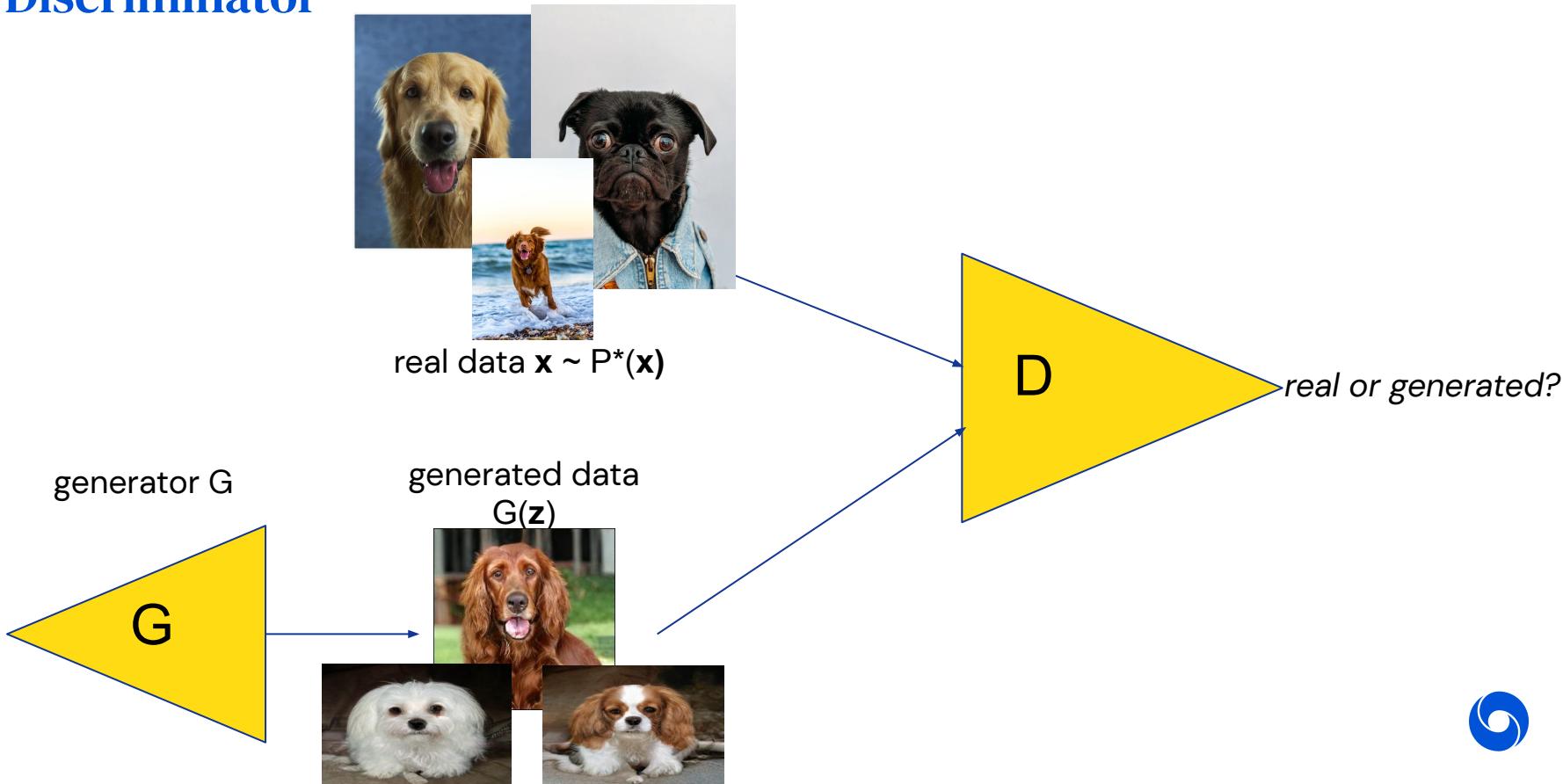
generated data  
 $G(z)$



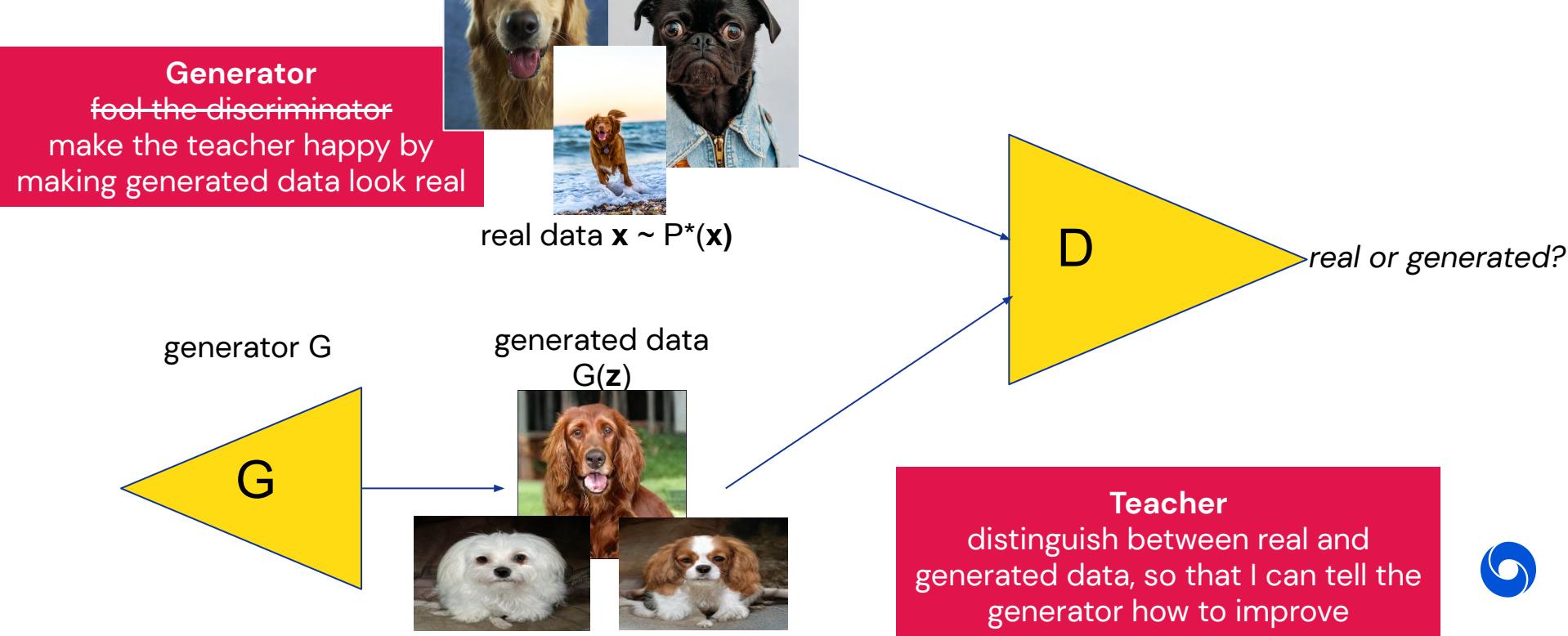
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness...



# Discriminator



# Discriminator Teacher (less adversarial view)



# Generative adversarial networks

Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

$$\min_G \max_D V(D, G) = \underbrace{\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})]}_{\text{log-probability that D correctly predicts real data } \mathbf{x} \text{ are real}} + \underbrace{\mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]}_{\text{log-probability that D correctly predicts generated data } G(\mathbf{z}) \text{ are generated}}$$



# Generative adversarial networks

Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$$

log-probability that D correctly predicts real data  $\mathbf{x}$  are real

log-probability that D correctly predicts generated data  $G(\mathbf{z})$  are generated

discriminator's (D) goal: **maximize** prediction accuracy

generator's (G) goal: **minimize** D's prediction accuracy, by **fooling** D into believing its outputs  $G(\mathbf{z})$  are real as often as possible



# Generative adversarial networks

Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

**for** number of training iterations **do**

**for**  $k$  steps **do**

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

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m \left[ \log D(\mathbf{x}^{(i)}) + \log (1 - D(G(\mathbf{z}^{(i)}))) \right].$$

**end for**

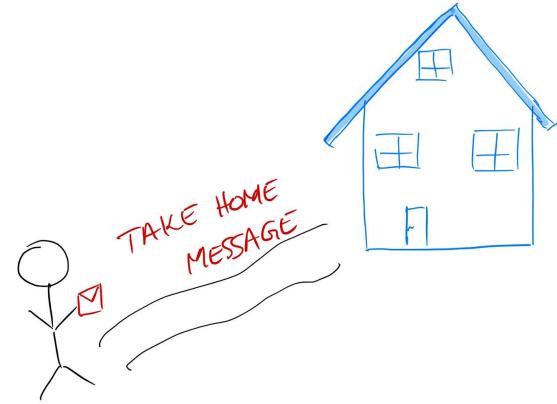
- Sample minibatch of  $m$  noise samples  $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$  from noise prior  $p_g(\mathbf{z})$ .
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log (1 - D(G(\mathbf{z}^{(i)}))).$$

**end for**



**GANs are an implicit generative model  
trained as a two player game.**



# Generative adversarial networks as zero sum game

$$\min_G \max_D V(D, G)$$

- Bi level optimization of the same loss function.
- Connection to game theory literature.
  - Nash equilibria
  - Strategies
  - Fictitious play



# Generative models as distance minimization

- The objective of generative models is often to minimize a divergence or distance.
- Most common: Maximum likelihood (KL divergence).

Why divergence/distance minimization?

$$D(p^* || p) = 0 \implies p = p^*$$



# Generative models as distance minimization

- ⇒ The objective of generative models is often to minimize a divergence or distance.
- ⇒ Most common: Maximum likelihood (KL divergence).

Maximum likelihood

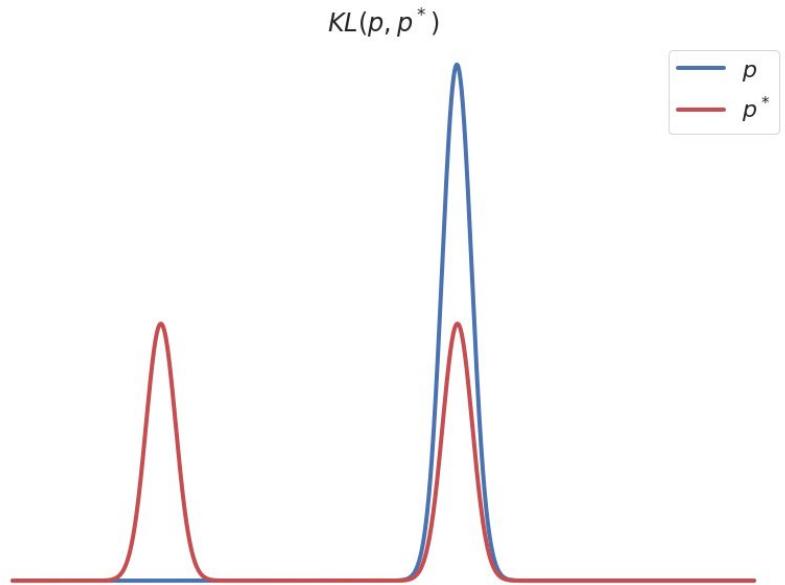
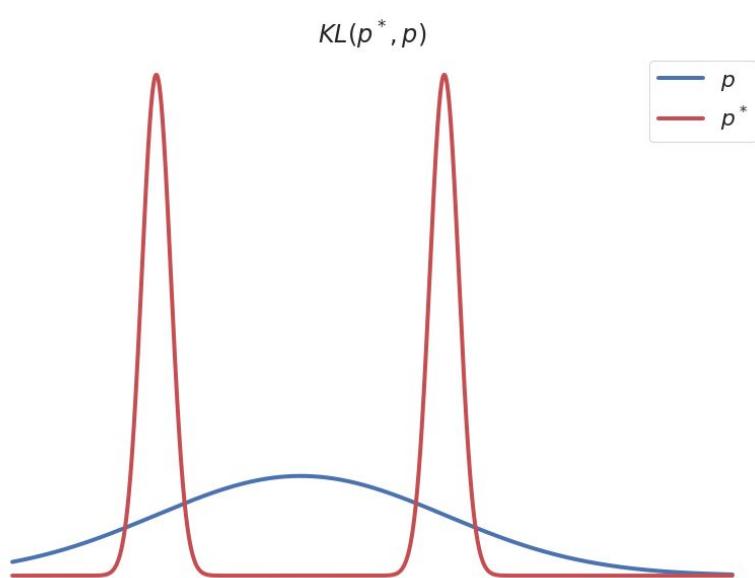
$$\text{KL}(p^*(\mathbf{x}) || p(\mathbf{x})) = \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

$$\text{KL}(p^*(\mathbf{x}) || p(\mathbf{x})) = 0 \implies p(\mathbf{x}) = p^*(\mathbf{x})$$





# Effects of the choice of divergence - learned models



Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

# Are GANs doing divergence minimization?

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$$

If the discriminator ( $D$ ) is optimal:  
the generator is minimizing the Jensen Shannon divergence  
between the true and generated distributions.



Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

# Are GANs doing divergence minimization?

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$$

If the discriminator ( $D$ ) is optimal:  
the generator is minimizing the Jensen Shannon divergence  
between the true and generated distributions.

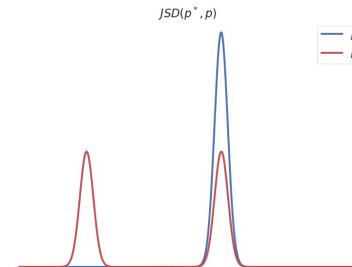
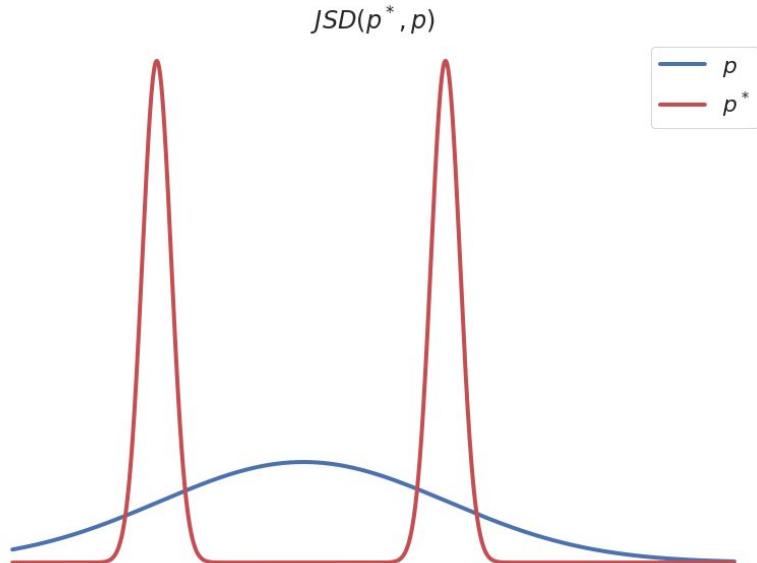
Connection to optimality:

$$JSD(p^* || p) = 0 \implies p = p^*$$



# Jensen Shannon divergence

$$\text{JSD}(p, p^*) = \frac{1}{2}\text{KL}(p, \frac{p + p^*}{2}) + \frac{1}{2}\text{KL}(p^*, \frac{p + p^*}{2})$$



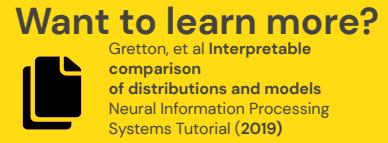
# GANs: More than Jensen Shannon divergence

In practice: D is not optimal:

- ⇒ limited computational resources
- ⇒ we do not have access to the true data distribution (just samples)



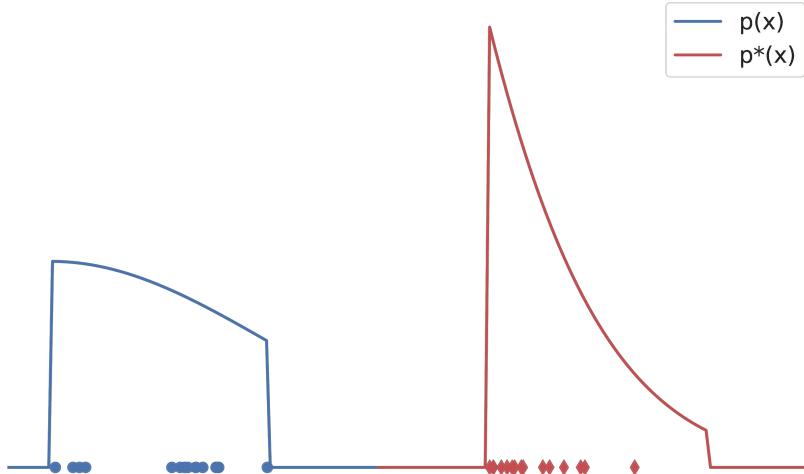
# Properties of KL & Jensen Shannon divergences



No learning signal from KL/JSD divergence if non overlapping support between the data and the model.

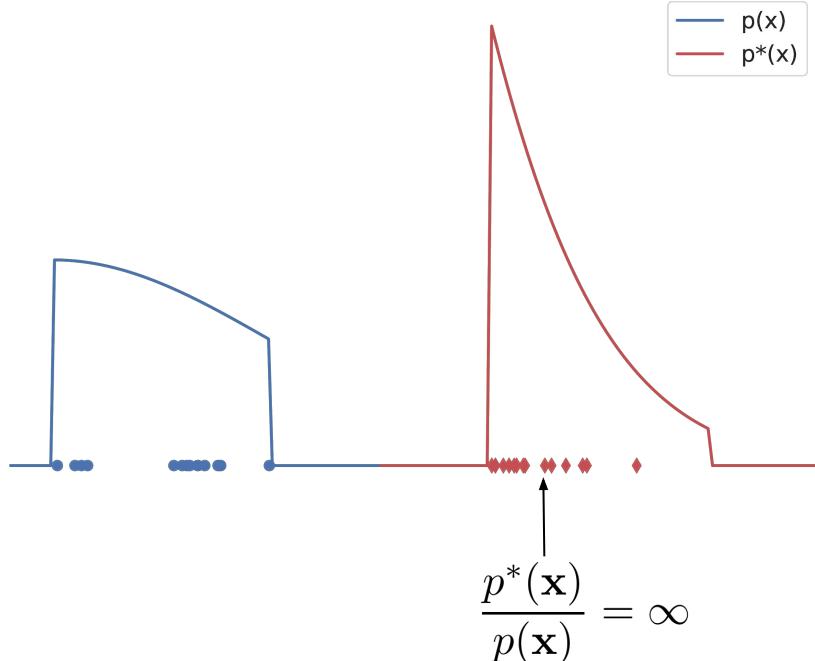
$$\text{KL}(p^*(\mathbf{x}) || p(\mathbf{x})) = \infty$$

$$\text{JSD}(p^*(\mathbf{x}) || p(\mathbf{x})) = \log 2$$



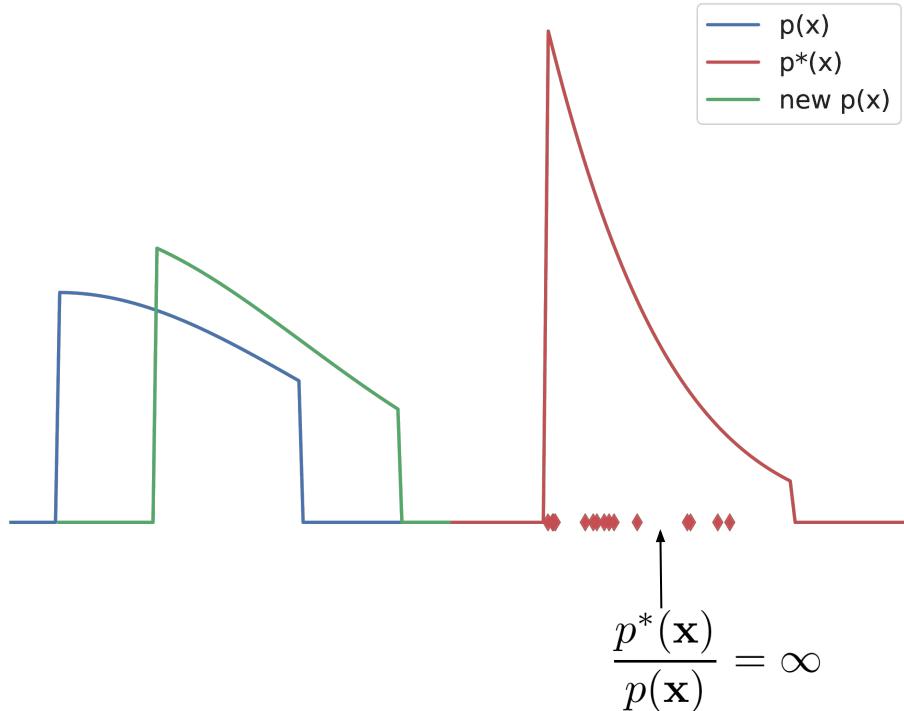
# Non overlapping support

$$\text{KL}(p^*(\mathbf{x}) \parallel p(\mathbf{x})) = \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$



# Non overlapping support

Moving the model closer to the true distribution (new p) results in no change in KL/JSD.



# Generative adversarial networks as zero sum game

$$\min_G \max_D V(D, G)$$

Can we choose another  $V$ ?



## Generative adversarial networks as zero sum game

$$\min_G \max_D V(D, G)$$

Will it correspond to a distributional divergence?



# Other divergences and distances

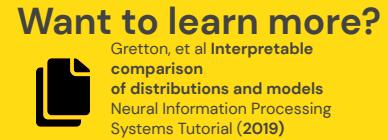
Wasserstein Distance

$$W(p^*, p) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(x)} f(x)$$

$$|f(x) - f(y)| \leq |x - y|$$



# Other divergences and distances

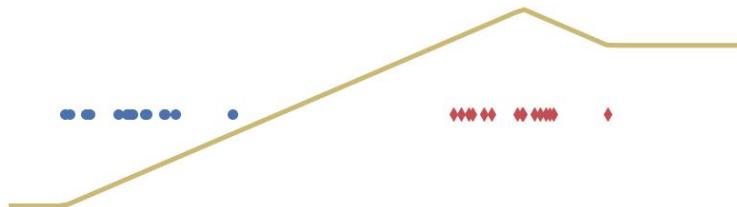


## Wasserstein Distance

$$W(p^*, p) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(x)} f(x)$$

$$W(p^*, p) = 1.78$$

—  $f^*$

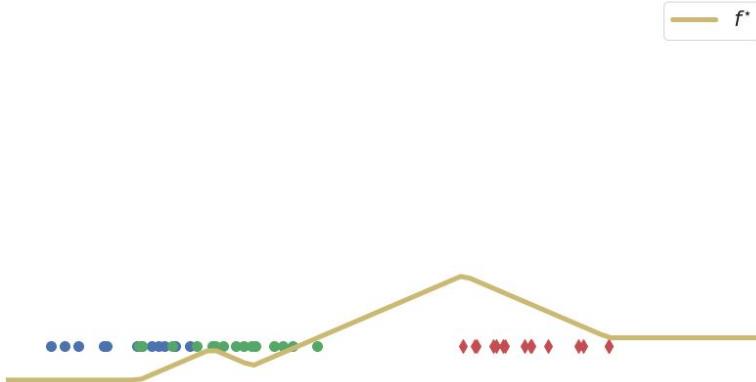


# Other divergences and distances

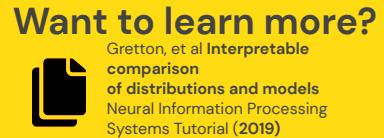
## Wasserstein Distance

$$W(p^*, p) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(x)} f(x)$$

$$W(p^*, p) = 1.6$$



# Other divergences and distances



Wasserstein Distance

Estimation

$$W(p^*, p) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(x)} f(x)$$

Learning

$$\min_G W(p, p^*) = \min_G \sup_{\|f\|_L \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(z)} f(G(z))$$





# Other divergences and distances

Wasserstein Distance

$$W(p, p^*) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p(x)} f(x) - \mathbb{E}_{p^*(x)} f(x)$$

Wasserstein GAN

$$\min_G \max_{\|D\|_L \leq 1} \mathbb{E}_{p^*(x)} D(x) - \mathbb{E}_{p(z)} D(G(z))$$



*Try to make D is 1-Lipschitz via gradient penalties, spectral normalization, weight clipping.*

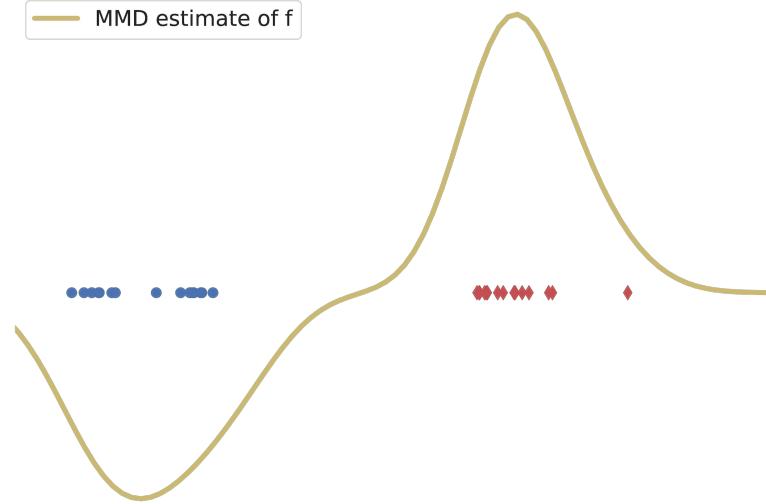
# Other divergences and distances

MMD

$$\text{MMD}(p^*, p) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(x)} f(x)$$

$\mathcal{H}$  is a RKHS.

— MMD estimate of  $f$



Want to learn more?



Li, et al MMD GAN: Towards Deeper Understanding of Moment Matching Network.  
Neural Information Processing Systems (2017)

## Other divergences and distances

MMD

$$\text{MMD}(p^*, p) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{p^*(x)} f(x) - \mathbb{E}_{p(x)} f(x)$$

$\mathcal{H}$  is a RKHS.

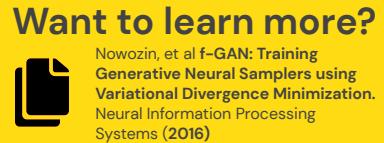
MMD-GAN

$$\min_G \max_{\|D\|_{\mathcal{H}} \leq 1} \mathbb{E}_{p^*(x)} D(x) - \mathbb{E}_{p(z)} D(G(z))$$

Choose kernel with learned features (via  $D$ )  $K_\phi(x, x') = K(\phi(X), \phi(X'))$



# Other divergences and distances



f-divergences

$$D_f(p^*||p) = \int p(x)f\left(\frac{p^*(x)}{p(x)}\right) dx$$

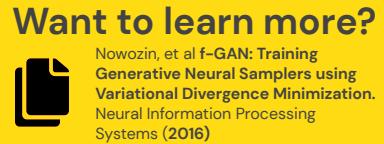
variational lower bound

$$\int p(x)f\left(\frac{p^*(x)}{p(x)}\right) dx \geq \sup_{T \in \mathcal{T}} (\mathbb{E}_{p(x)} T(x) - \mathbb{E}_{p^*(x)} f^*(T(x)))$$

optimal T for KL:  $f^*\left(\frac{p^*(x)}{p(x)}\right)$

  $f^*$  is the convex conjugate of  $f$

# Other divergences and distances



f-divergences

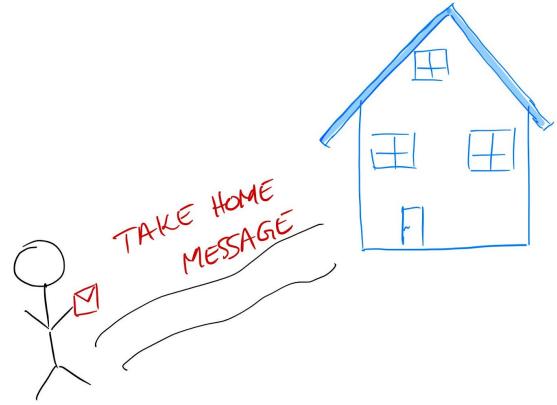
$$D_f(p^*||p) = \int p(x)f\left(\frac{p^*(x)}{p(x)}\right) dx$$

f-GAN

$$\min_G \max_D \mathbb{E}_{p(z)} D(G(z)) - \mathbb{E}_{p^*(x)} f^*(D(x))$$



**Can create GAN training criteria inspired by multiple divergences & distances.**



# Why train a GAN instead of doing divergence minimization?

- Model type
- Computational Intractability
- Smooth learning signal
- Learned “divergence”

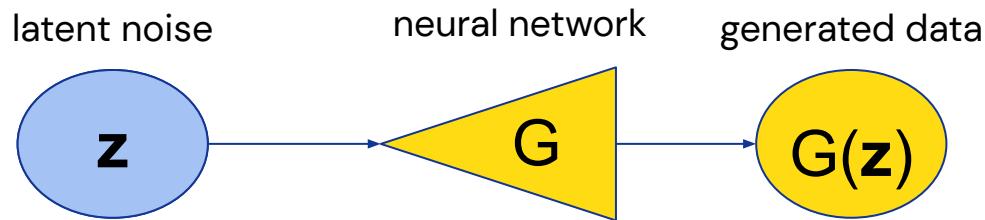


# Implicit models and KL divergence

Model type

$$\text{KL}(p^*(\mathbf{x}) || p(\mathbf{x})) = \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x})} dx$$

For implicit models, we do not have access to the explicit distribution  $p(x)$ .

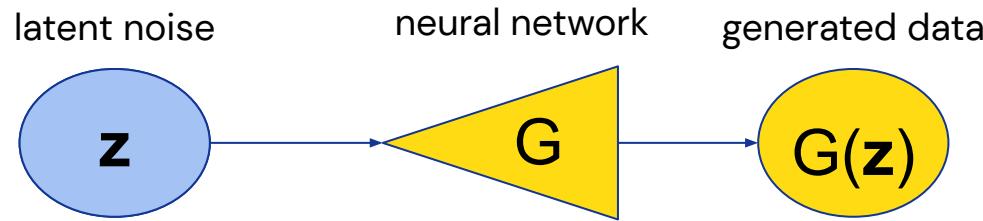


# Implicit models and KL divergence

Model type

$$\text{KL}(p^*(\mathbf{x}) || p(\mathbf{x})) = \int p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x})} dx$$

For implicit models, we do not have access to the explicit distribution  $p(x)$ .



f-GAN

$$\min_G \max_D \mathbb{E}_{p(z)} D(G(z)) - \mathbb{E}_{p^*(x)} f^*(D(x))$$



# Wasserstein distance & computational intractability

Computational intractability

$$W(p, p^*) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p(x)} f(x) - \mathbb{E}_{p^*(x)} f(x)$$

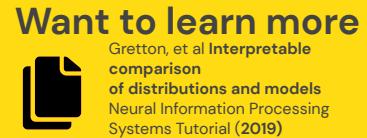
Computationally intractable for complex cases.

Wasserstein GAN

$$\min_G \max_{\|D\|_L \leq 1} \mathbb{E}_{p^*(x)} D(x) - \mathbb{E}_{p(z)} D(G(z))$$



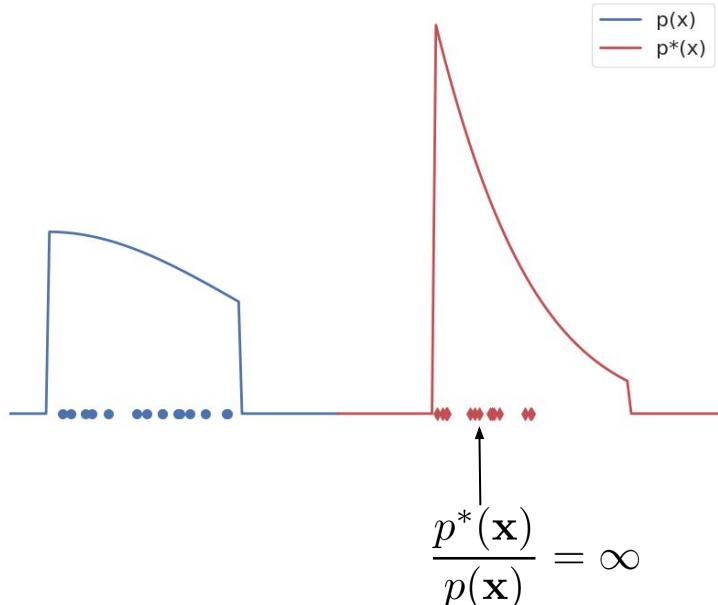
# Smooth learning signal



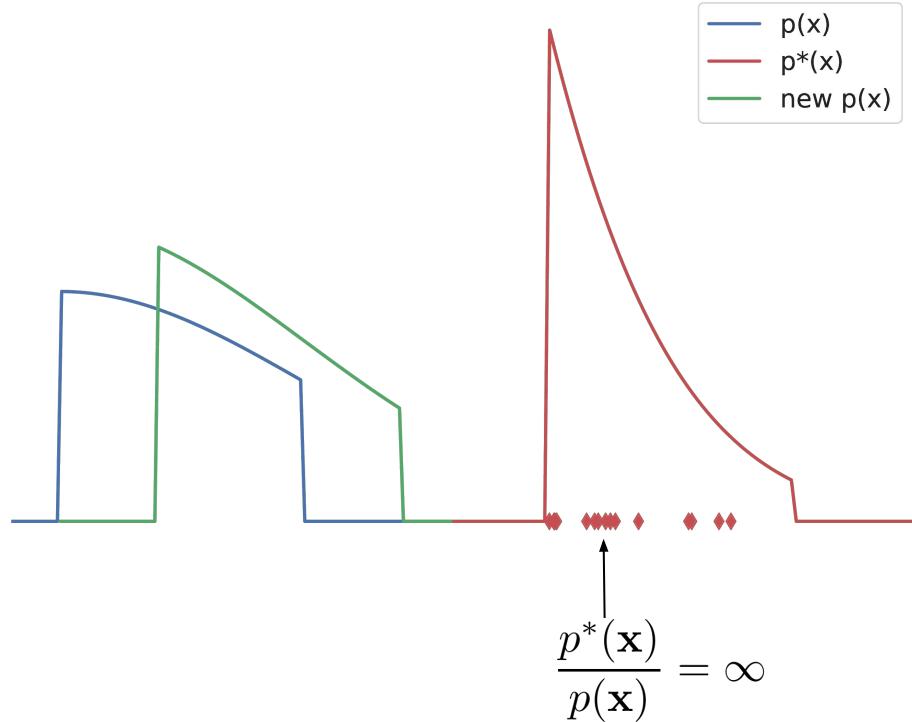
No learning signal from KL/JSD divergence if non-overlapping support between the data and the model.

$$\text{KL}(p^*(\mathbf{x}) \parallel p(\mathbf{x})) = \infty$$

$$\text{JSD}(p^*(\mathbf{x}) \parallel p(\mathbf{x})) = \log 2$$



# Smooth learning signal



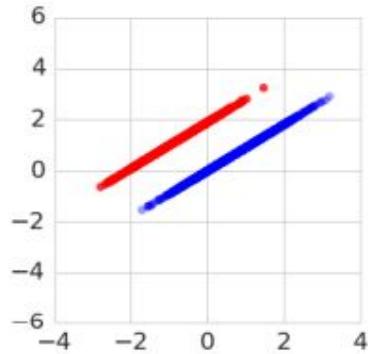
The density ratio jumps to infinity at the data distribution.



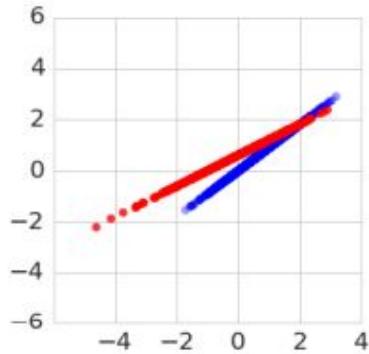


# Smooth learning signal

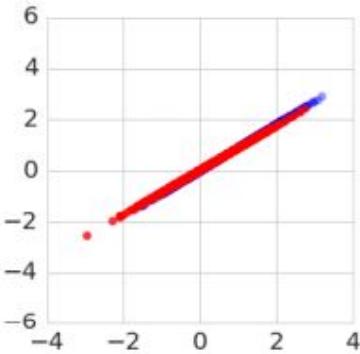
But GANs still learn!



(a) Step 0



(b) Step 5000



(c) Step 12500

Red = data

Blue = model (changes in training)



true ratio

$\downarrow$

$$KL[p^*(x) \parallel p(x)] = \int p^*(x) \log \left( \frac{p^*(x)}{p(x)} \right) dx \geq \sup_{D \in \mathcal{F}} \left( \mathbb{E}_{p^*(x)} D(x) - \mathbb{E}_{p(x)} e^{D(x)} \right)$$

$\uparrow$

ratio approximation used in GAN training



$$KL[p^*(x)||p(x)] = \int p^*(x) \log \left( \frac{p^*(x)}{p(x)} \right) dx \geq \sup_{D \in \mathcal{F}} \left( \mathbb{E}_{p^*(x)} D(x) - \mathbb{E}_{p(x)} e^{D(x)} \right)$$

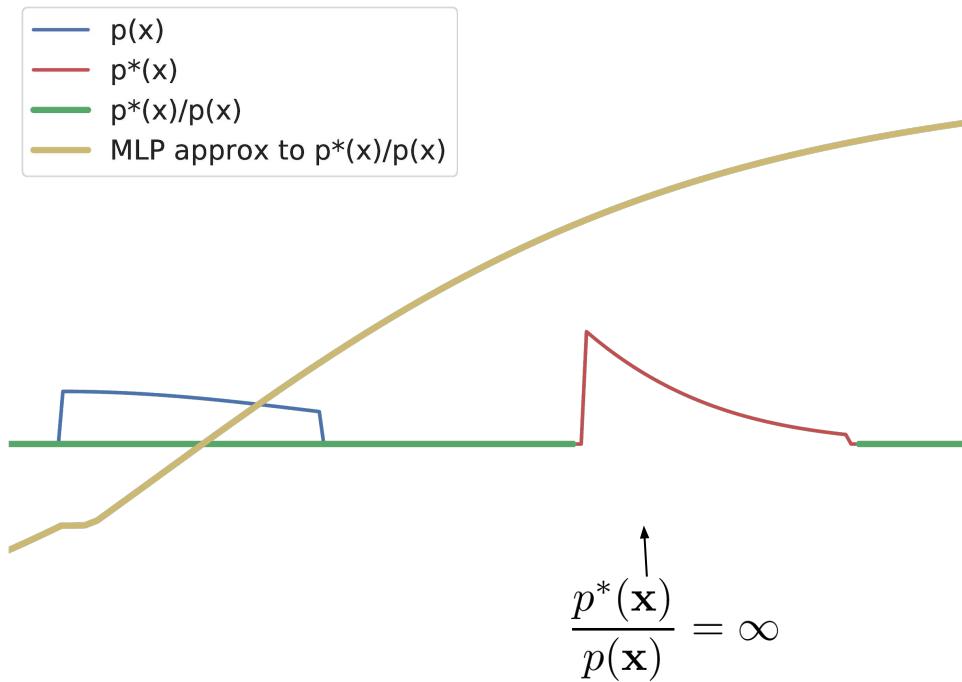
true ratio

ratio approximation used in GAN training

$\mathcal{F}$  is the family of functions used to approximate the ratio (deep neural networks, RKHS).



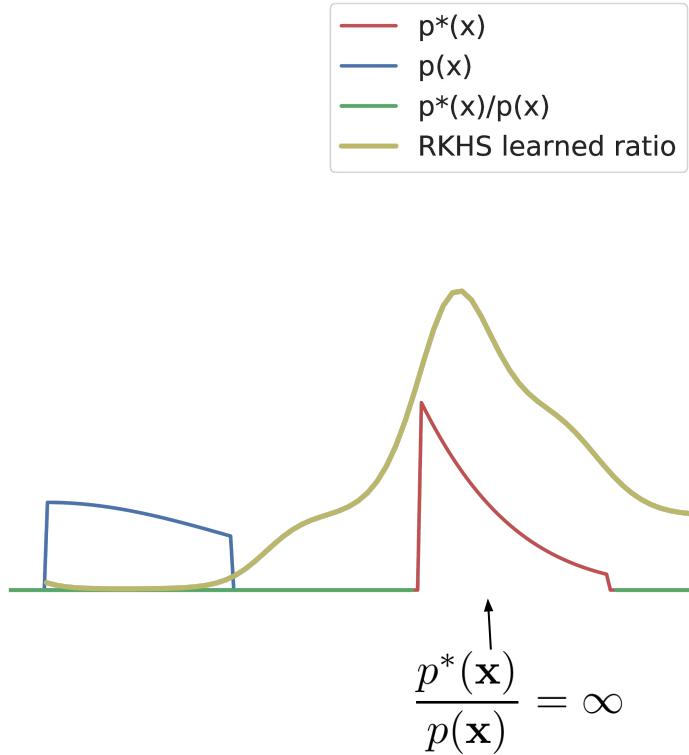
# Smooth learning signal



Smooth approximation of the density ratio does not go to infinity.



# Smooth learning signal



Smooth approximation of the density ratio does not go to infinity.



D is smooth approximation to the decision boundary of the underlying divergence:

- GANs do not do divergence minimization in practice
- GANs do not fail in cases where the underlying divergence would



# Discriminators as learned “distances”

Want to learn more?



Arora, et al Generalization and Equilibrium in  
Generative Adversarial Nets.  
International Conference for machine learning  
(2017)

We can think of D (the teacher) as learning a “distance” between the data and model distribution that can provide useful gradients to the model.



# Discriminators as “learned” distances

Original GAN

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$$

Wasserstein GAN

$$\min_G \max_{||D||_L \leq 1} \mathbb{E}_{p^*(x)} D(x) - \mathbb{E}_{p(z)} D(G(z))$$

$$\min_G \boxed{\max_D V(D, G)}$$



## Discriminators as “learned” distances

$$\min_G \max_D V(D, G)$$

D provides a learned distance between  
the data and sample distributions, using  
**learned neural network features.**



# GANs (learned distance) or divergence minimization?

## GANs

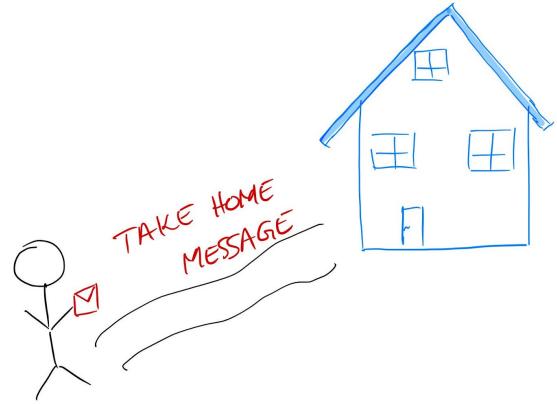
- good samples
- learned loss function
- hard to analyze dynamics (game theory)
- (in practice) no optimal convergence guarantees

## Divergence minimization

- optimal convergence guarantees
- easy to analyze loss properties
- hard to get good samples
- loss functions don't correlate with human evaluation



In practice, GANs do not do divergence minimization.  
The discriminator can be seen as a learned “distance”.



# Which GAN should I use?

Empirically, it has been observed that the underlying loss matters less than neural architectures, training regime, data.

Stay tuned!



# Unconditional and conditional generative models

## Unconditional

*provides a sample from the data distribution, but the user has no control over what kind of sample.*

## Conditional

*we can specify what sample we want (dog vs cat).*

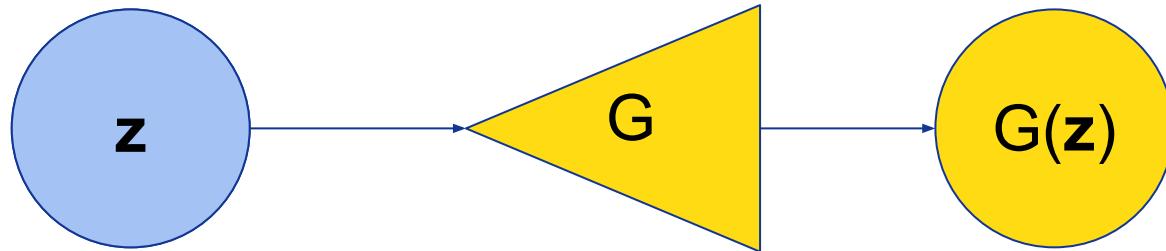


# So far... unsupervised GANs

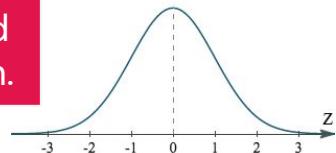
latent (“noise”) vector  
 $z \sim P(z)$

generator  $G$ :  
a deep neural network

generated data  
 $G(z)$



Generator input is random noise to account for spread of data distribution.

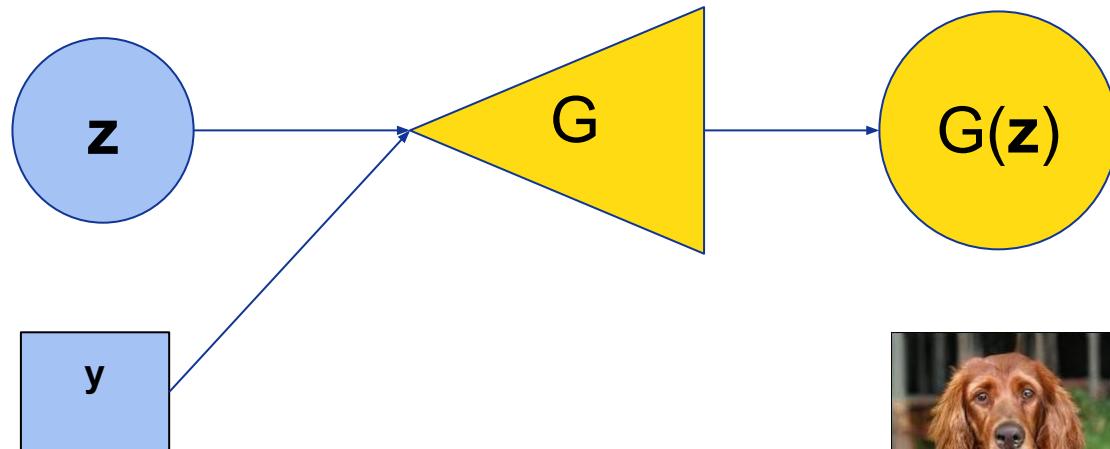
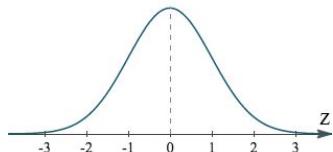


# Conditioning information for training GANs

latent (“noise”) vector  
 $\mathbf{z} \sim P(\mathbf{z})$

generator  $G$ :  
a deep neural network

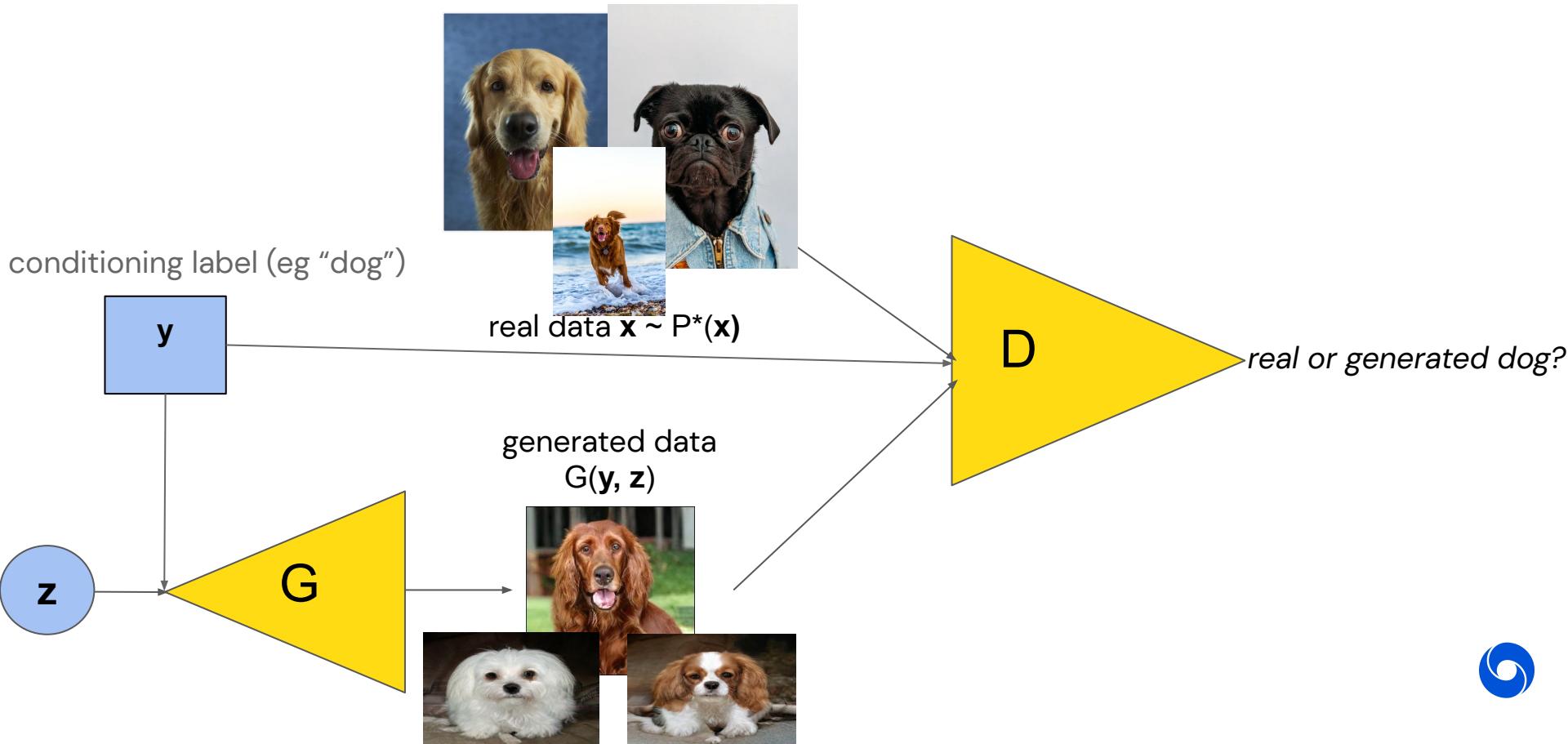
generated data  
 $G(\mathbf{z})$

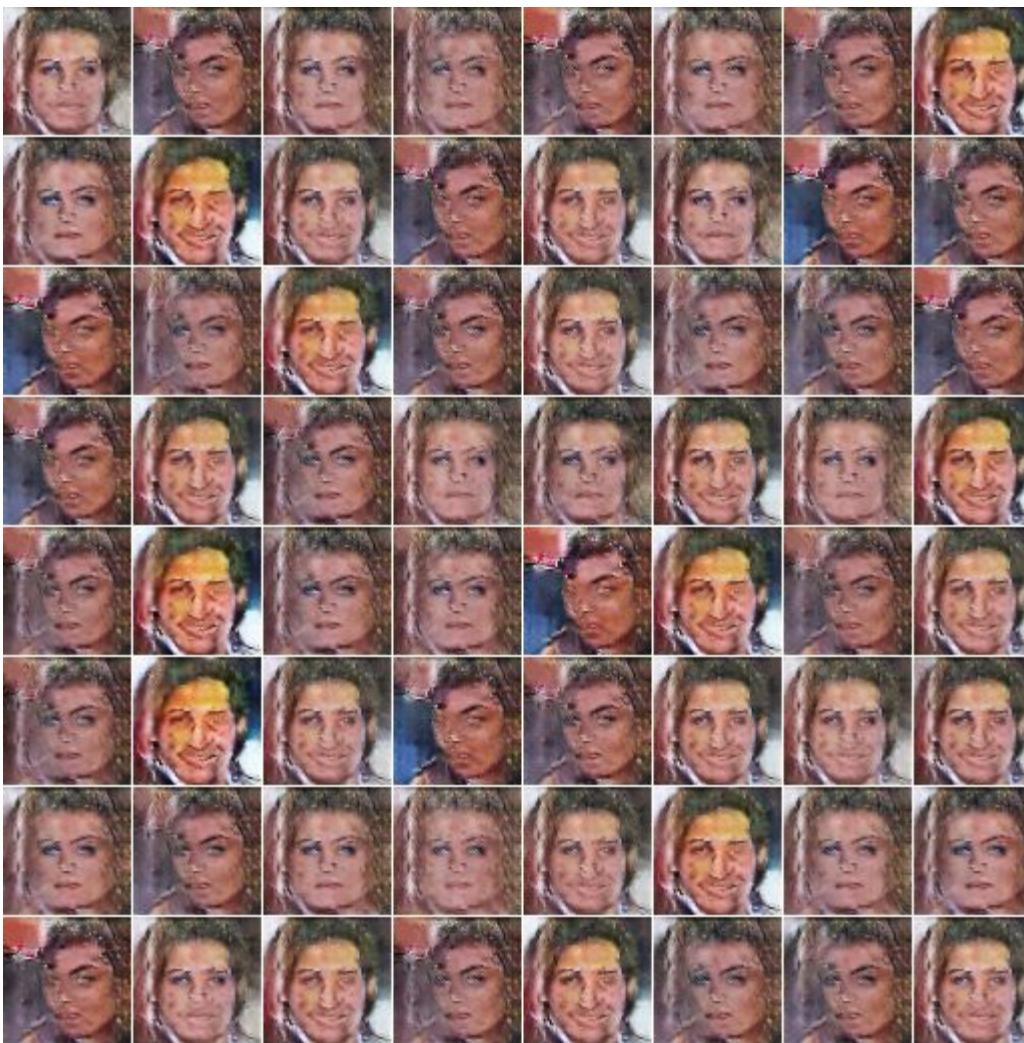


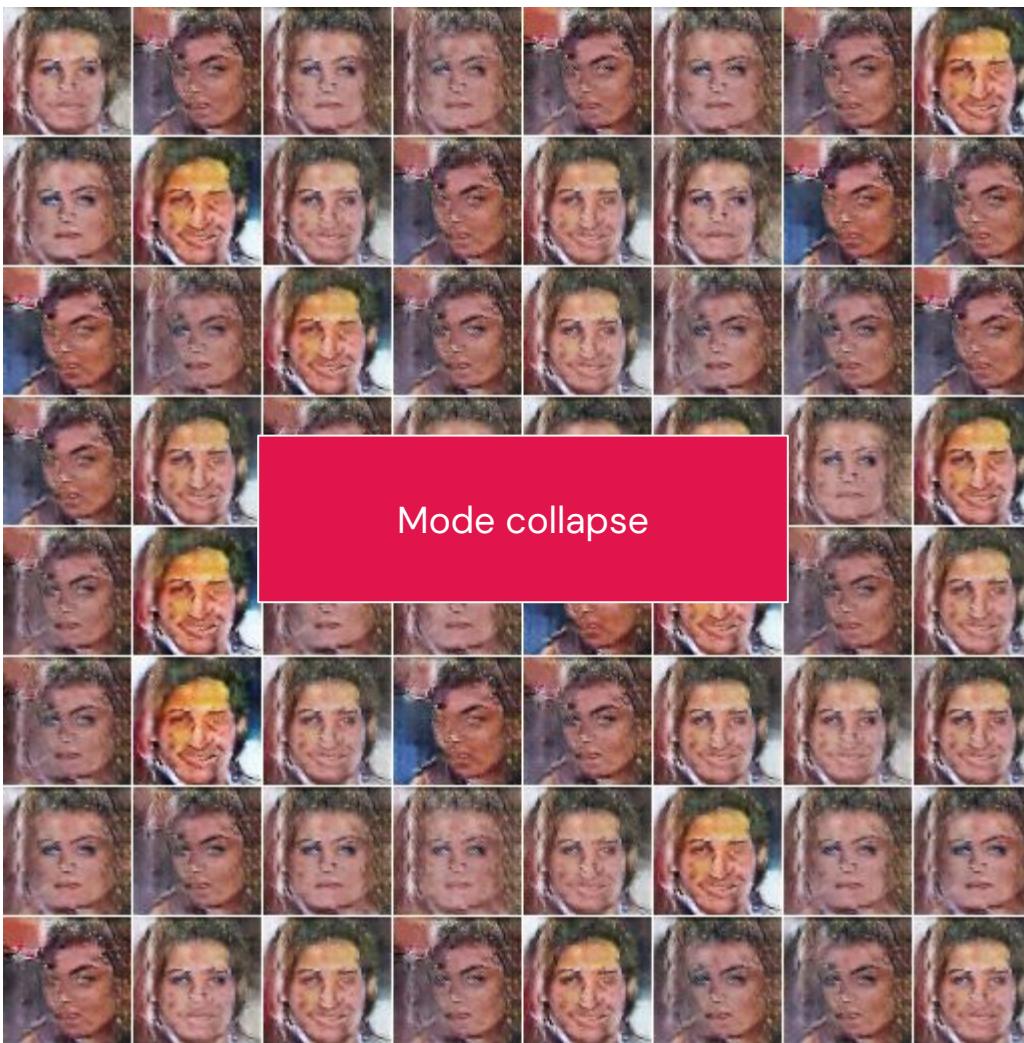
Add conditioning  
generation to  
specify information  
about generated  
sample.

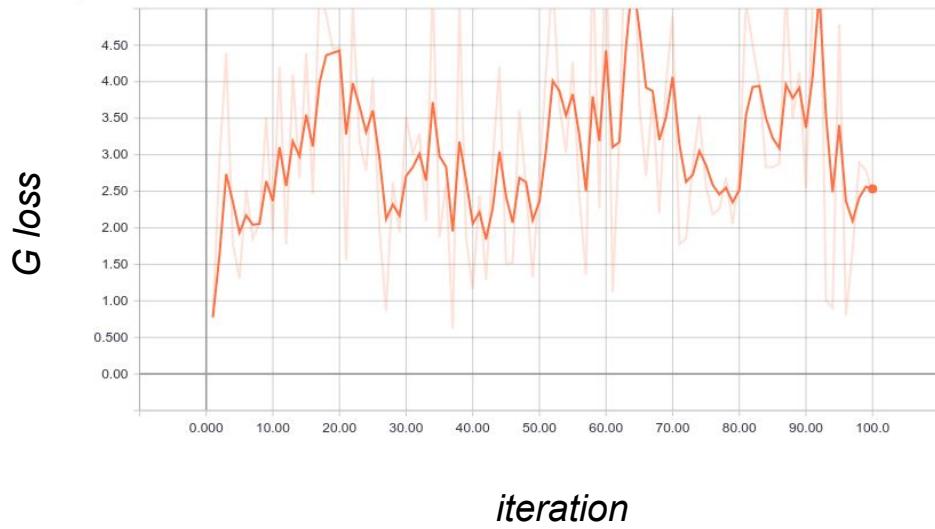


# Class conditional GANs









DeepMind

2

# Evaluating GANs



# Evaluating generative models



**No evaluation metric is able to capture all desired properties.**

- sample quality
- generalization
- representation learning

- Evaluate based on end goal
- semi supervised learning: classification accuracy
  - reinforcement learning: agent reward
  - data generation: human (user) evaluation



# GANs are implicit models

**Log likelihoods are not available (and are very expensive to approximate).**



# Inception score



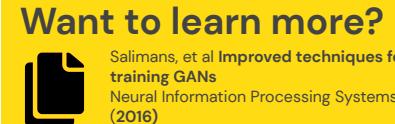
Data

---

Model



# Inception Score



Use a pretrained Imagenet classifier to compare (via KL divergence)  
the distribution of **labels** obtained from the data  
the distribution of **labels** coming from samples

Measures:

- sample quality
- dropping classes (no cats)
- correlates with human evaluation
- does not measure differences beyond class labels
- requires pretrained classifier

*Higher is better.*



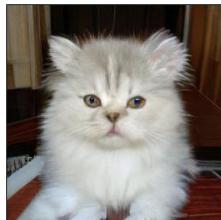
# Frechet Inception Distance



Data

---

Model



# Frechet Inception Distance

Want to learn more?



Heusel, et al GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

Neural Information Processing Systems (2017)

Use a pretrained Imagenet classifier to compare (via Frechet distance)  
the distribution of layer features obtained from the data  
the distribution of layer features coming from samples

Measures:

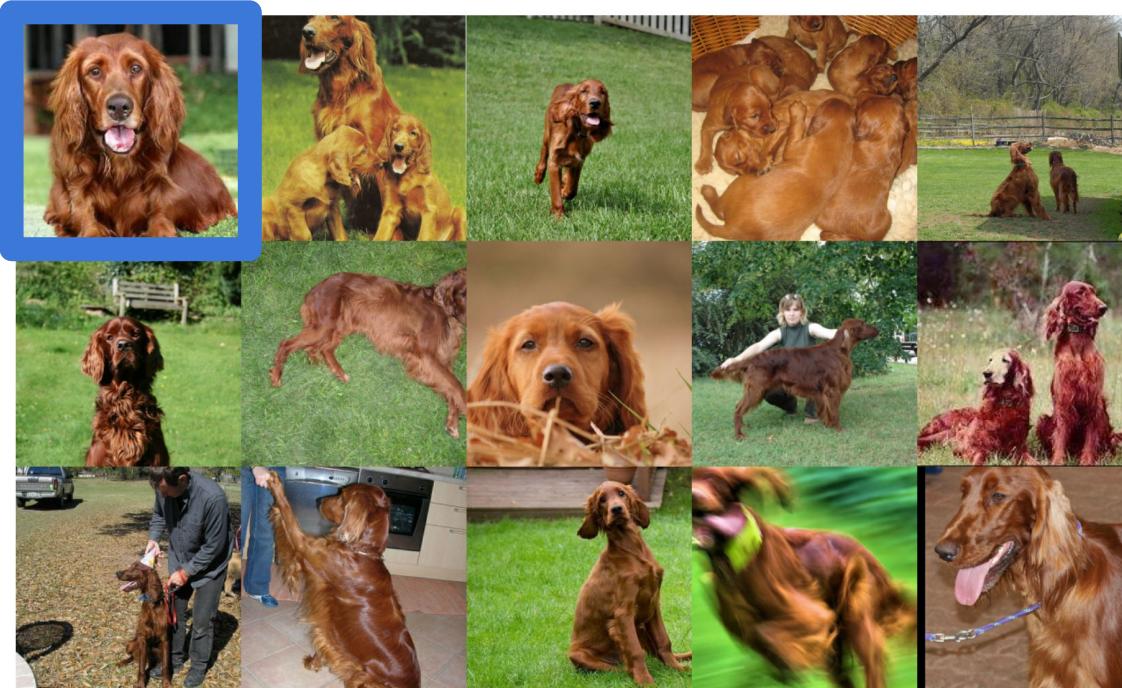
- sample quality
- dropping classes (no cats)
- captures feature level statistics (not just classes)
- correlates with human evaluation
- requires pretrained classifier
- biased for a small number of samples and KID for a fix (see Binkowski, et al., ICLR 2018)

*Lower is better.*

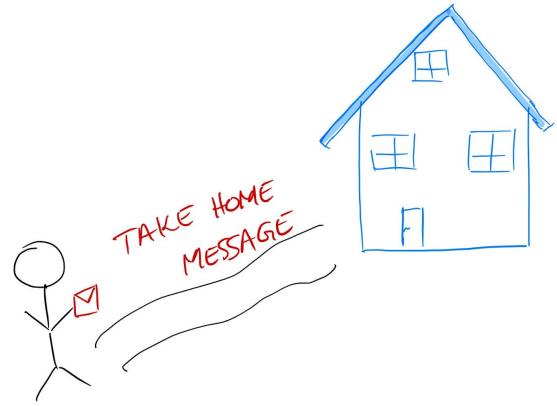


# Checking overfitting: Nearest neighbours

Nearest neighbors: most similar (in feature space of a pretrained ImageNet classifier) images in the dataset.



**Multiple metrics are needed to evaluate GAN samples.**



DeepMind

3

# The GAN Zoo



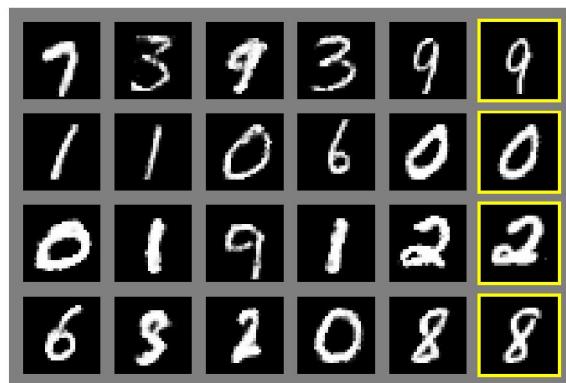
3.1

# Image Synthesis with GANs: MNIST to ImageNet



# The Original GANs (Goodfellow et al.)

- The original GAN paper (Goodfellow et al.)
- Simple data (~32x32 images)
- Simple models
- G and D in (a) and (b) were MLPs (not convolutional)
  - Images flattened to vectors for training, ignoring spatial structure



a)



c)



b)



d)

Want to learn more?



Goodfellow, et al. Generative adversarial networks.. Neural Information Processing Systems (2014)

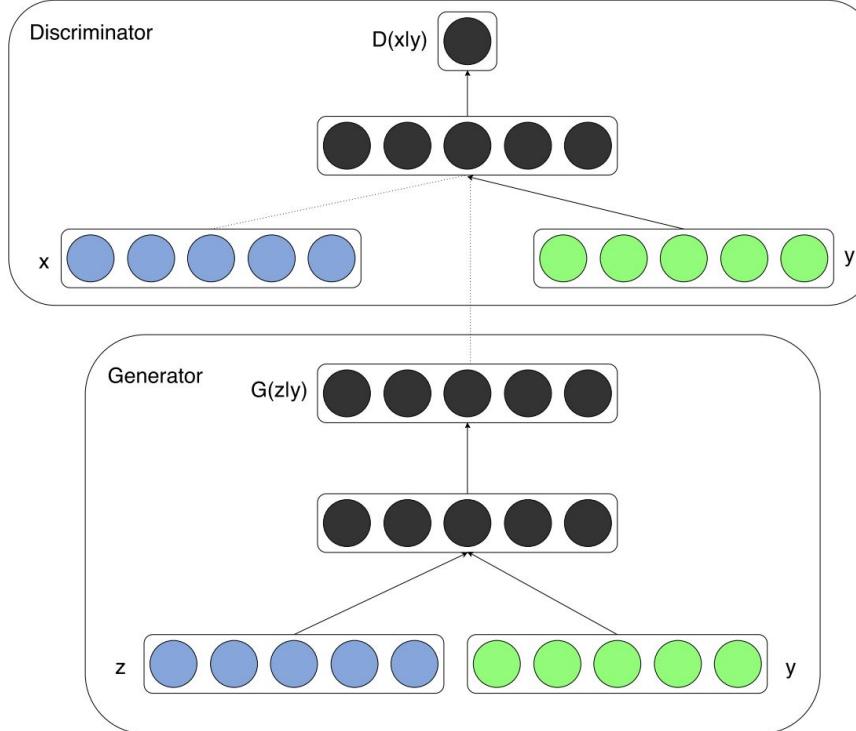
# Conditional GANs (Mirza and Osindero)

- Generalised GANs to the **conditional** setting where we have some extra information associated with each datum, e.g.,
  - a category ID ("cat", "dog", ...)
  - an input image from another domain

Want to learn more?



Mirza and Osindero. Conditional Generative Adversarial Nets.  
arXiv:1411.1784 (2014)



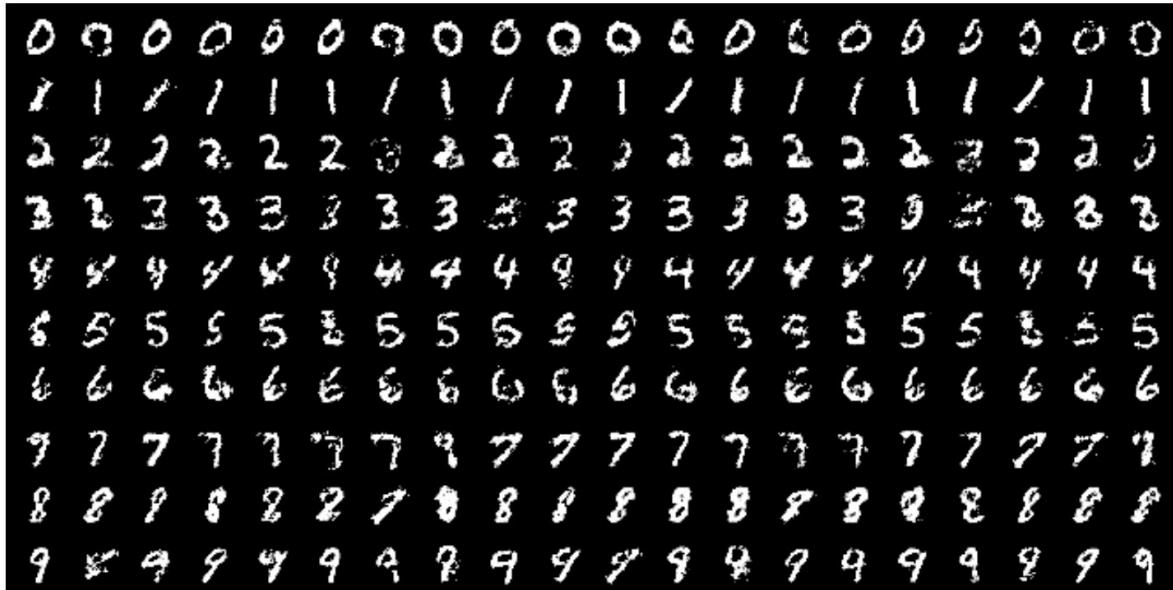
# Conditional GANs (Mirza and Osindero)

- Generalised GANs to the **conditional** setting where we have some extra information associated with each datum, e.g.,
  - a category ID ("cat", "dog", ...)
  - an input image from another domain

Want to learn more?

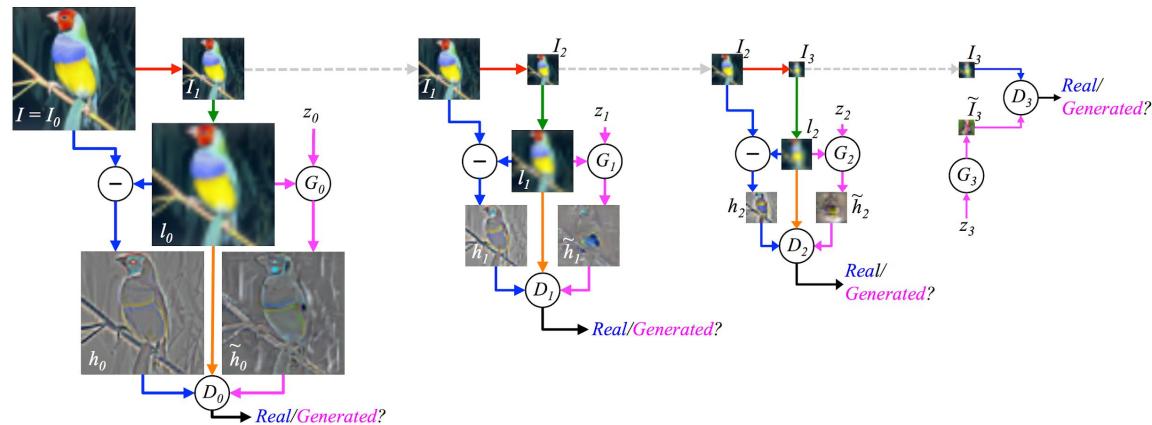


Mirza and Osindero. Conditional Generative Adversarial Nets.  
arXiv:1411.1784 (2014)



# Laplacian GANs (LAPGAN, Denton et al.)

- Start from a tiny image
- Upsample to a 2x larger image (blurry)
- Generate a Laplacian: the difference between the (blurry) upsampled image and the final image
- A *conditional* GAN after the initial resolution
  - G and D each take a lower resolution image as input, predicting e.g.:  
 $P(\text{is real } 64 \times 64 \text{ image} | 32 \times 32 \text{ image})$



Want to learn more?



Denton, et al. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. Neural Information Processing Systems (2015)



# Laplacian GANs (LAPGAN, Denton et al.)

Nice results at higher resolutions

Want to learn more?



Denton, et al. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. Neural Information Processing Systems (2015)



# Laplacian GANs (LAPGAN, Denton et al.)

Fully convolutional generator architecture

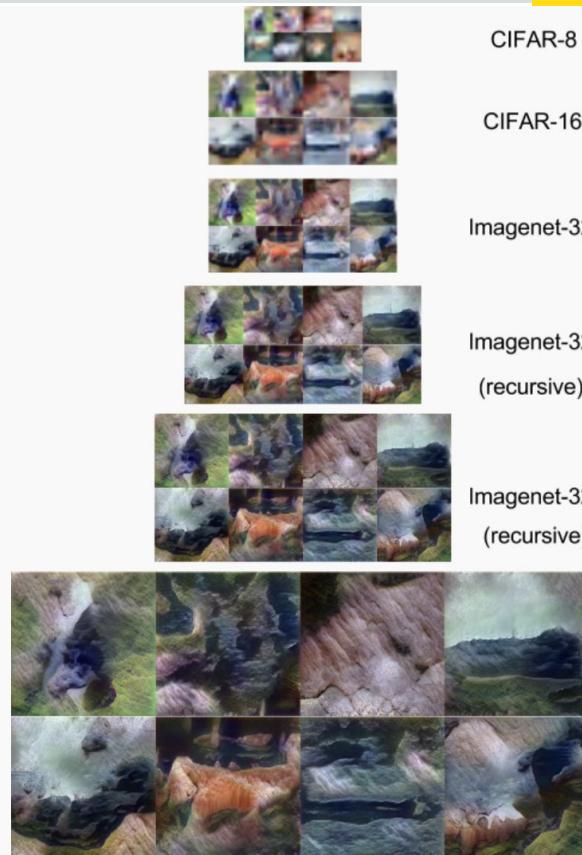
The model can be applied to produce arbitrarily high-resolution results

This model was trained on 32x32 images, but is applied recursively to upsample to 256x256.

Want to learn more?



Denton, et al. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. Neural Information Processing Systems (2015)



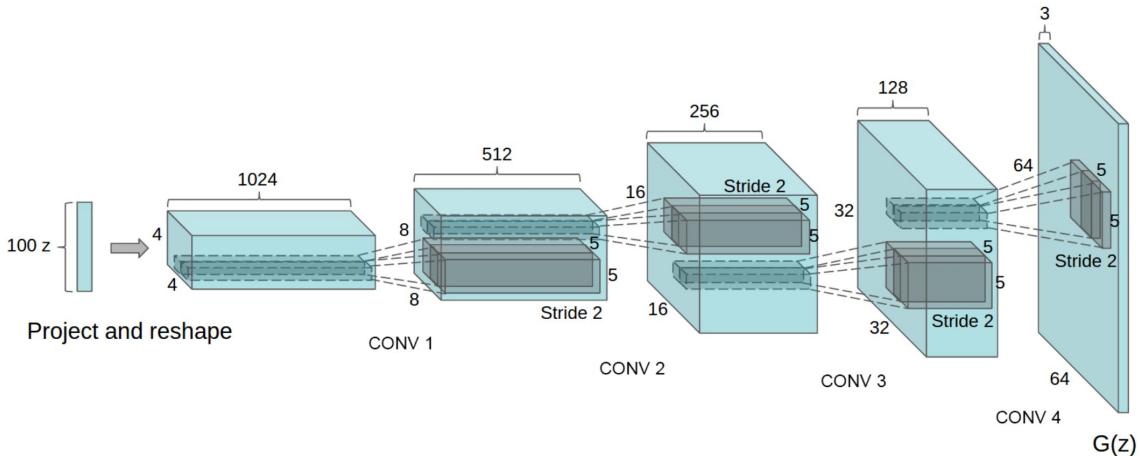
# Deep Convolutional GANs (DCGAN, Radford et al.)

- Simply use deep convnets for G and D
- Importantly, **batch normalization** (Ioffe and Szegedy, 2015) helped to stabilize the difficult learning process

Want to learn more?



Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. International Conference on Learning Representations (2016)



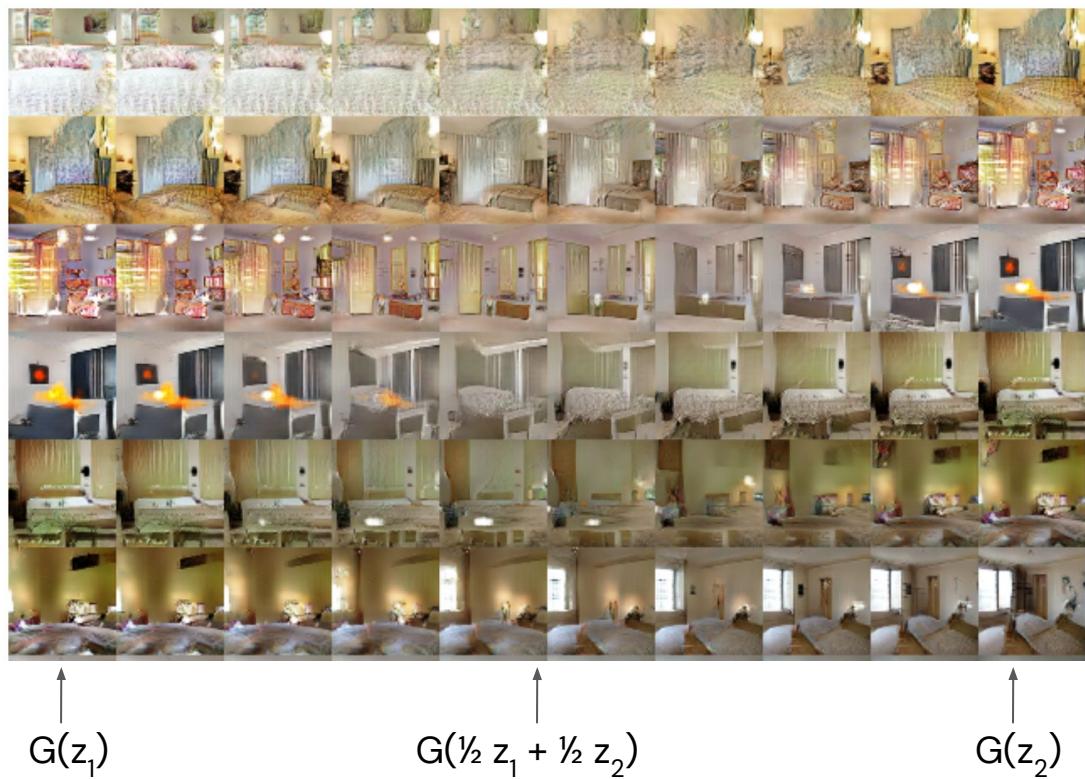
# Deep Convolutional GANs (DCGAN, Radford et al.)

- Simply use deep convnets for G and D
- Importantly, **batch normalization** (Ioffe and Szegedy, 2015) helped to stabilize the difficult learning process
- **Interpolation** between noise ( $z$ ) samples produces semantically reasonable images at every point

Want to learn more?



Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. International Conference on Learning Representations (2016)



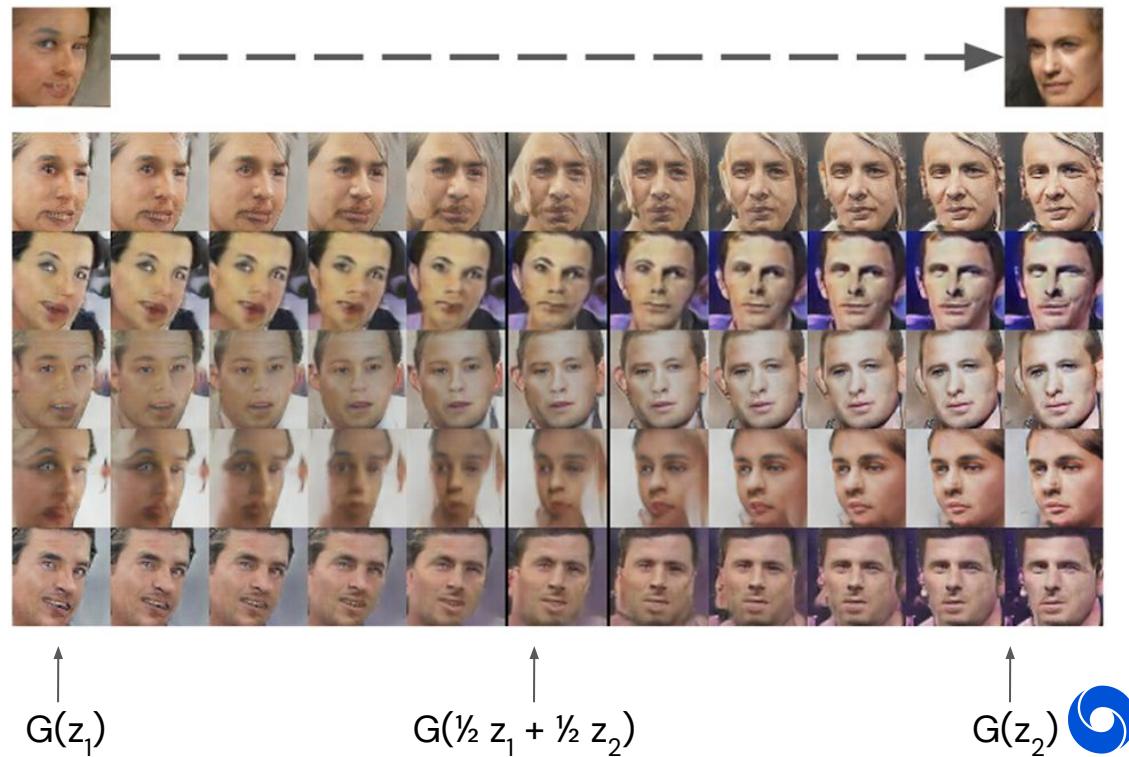
# Deep Convolutional GANs (DCGAN, Radford et al.)

- Simply use deep convnets for G and D
- Importantly, **batch normalization** (Ioffe and Szegedy, 2015) helped to stabilize the difficult learning process
- **Interpolation** between noise ( $z$ ) samples produces semantically reasonable images at every point

Want to learn more?

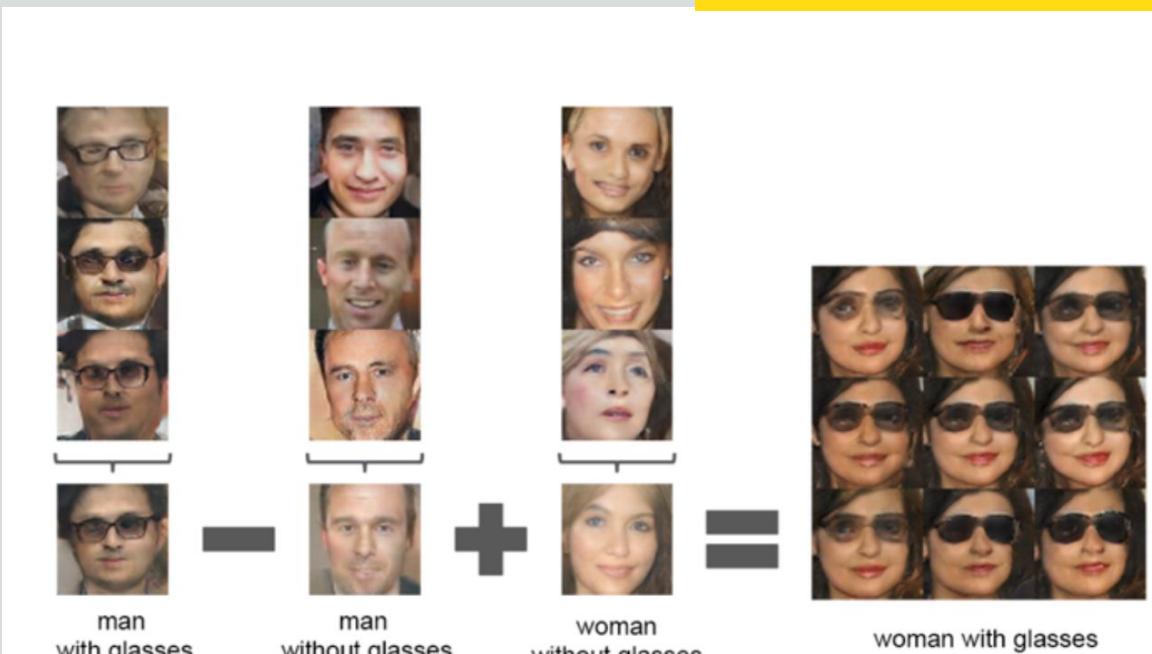


Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. International Conference on Learning Representations (2016)



# Deep Convolutional GANs (DCGAN, Radford et al.)

- The DCGAN generator's noise/latent space appears to have **meaningful semantics**



Want to learn more?



Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. International Conference on Learning Representations (2016)

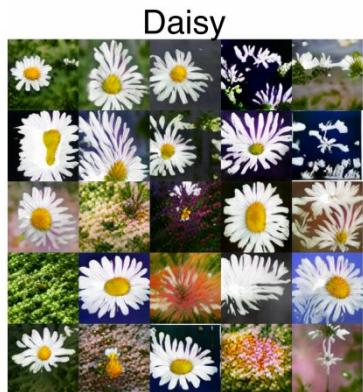


# Spectrally Normalised GANs (SNGAN, Miyato et al.)

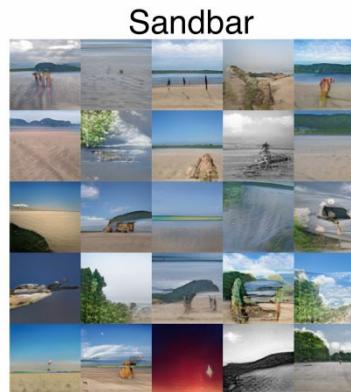
- Stabilise GAN training by clamping the singular values of D's weights to 1

$$\sigma(A) := \max_{\mathbf{h}: \mathbf{h} \neq \mathbf{0}} \frac{\|A\mathbf{h}\|_2}{\|\mathbf{h}\|_2} = \max_{\|\mathbf{h}\|_2 \leq 1} \|A\mathbf{h}\|_2$$

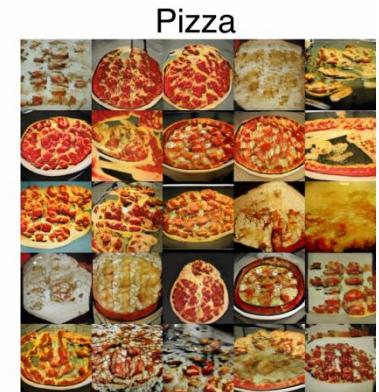
$$\bar{W}_{\text{SN}}(W) := W/\sigma(W)$$



Daisy



Sandbar



Pizza

Want to learn more?



Miyato, et al. Spectral Normalization for Generative Adversarial Networks. International Conference on Learning Representations (2018)



# Projection Discriminator (Miyato et al.)

- Novel formulation of the class-conditional discriminator
- Learnt class embedding is projected onto the final hidden representation
- Theoretically justified under the underlying probabilistic model
- Empirically, performs better than prior formulations

Want to learn more?



Miyato and Koyama. cGANs with Projection Discriminator. International Conference on Learning Representations (2018)

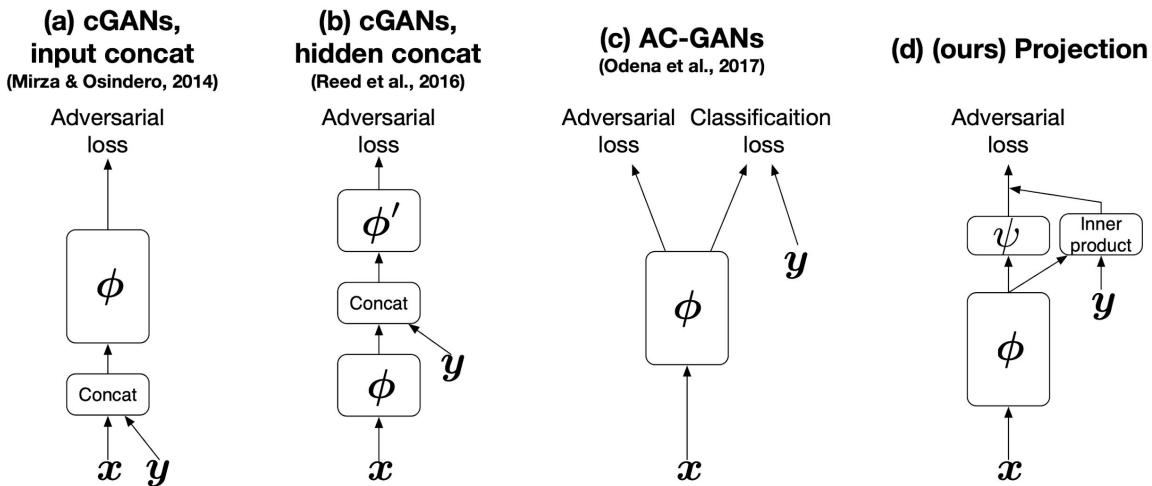


Figure 1: Discriminator models for conditional GANs



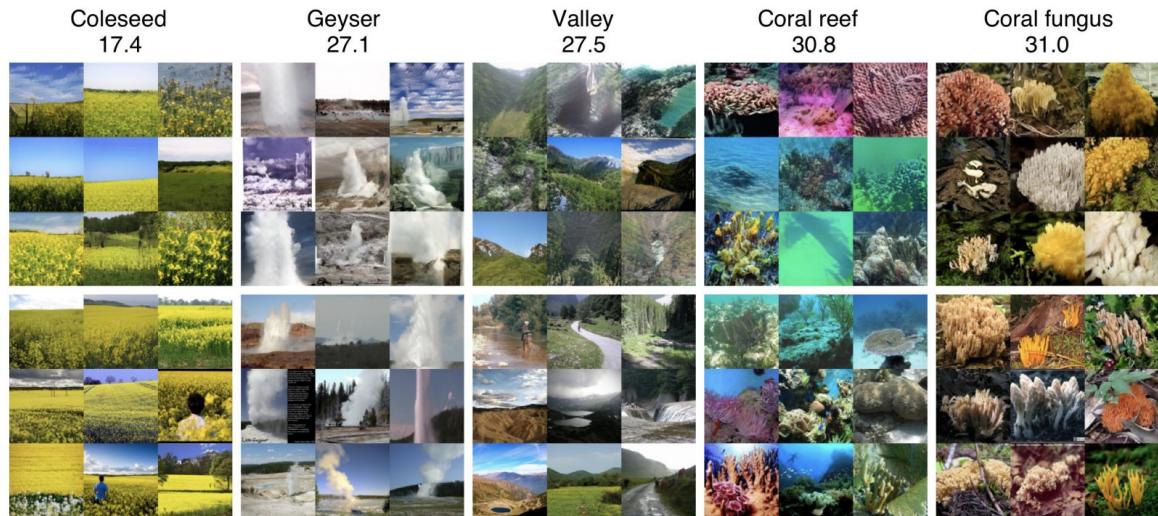
# Projection Discriminator (Miyato et al.)

- Novel formulation of the class-conditional discriminator
- Learnt class embedding is projected onto the final hidden representation
- Theoretically justified under the underlying probabilistic model
- Empirically, performs better than prior formulations

Want to learn more?



Miyato and Koyama. cGANs with Projection Discriminator. International Conference on Learning Representations (2018)



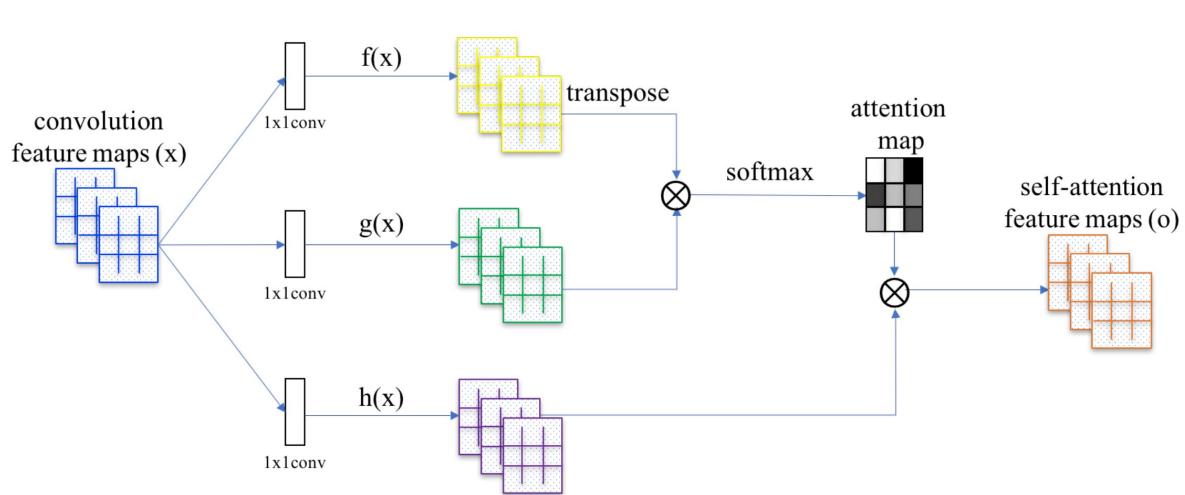
# Self-Attention GANs (SAGAN, Zhang et al.)

- Added **self-attention** to give images better **global structure** and coherence
- Self-attention has had a big impact in a number of domains (especially language modeling, translation)

Want to learn more?



Zhang, et al. Self-Attention Generative Adversarial Networks. International Conference on Machine Learning (2019)



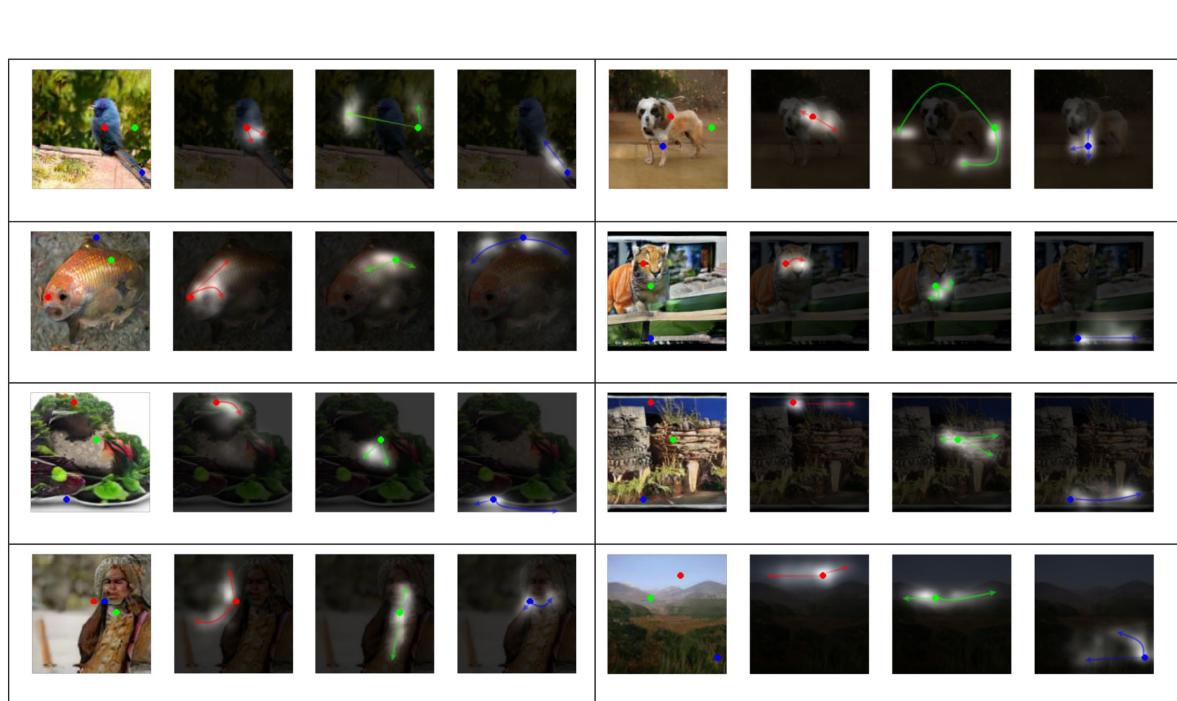
# Self-Attention GANs (SAGAN, Zhang et al.)

- Added **self-attention** to give images better **global structure** and coherence
- Self-attention has had a big impact in a number of domains (especially language modeling, translation)

Want to learn more?



Zhang, et al. Self-Attention Generative Adversarial Networks. International Conference on Machine Learning (2019)



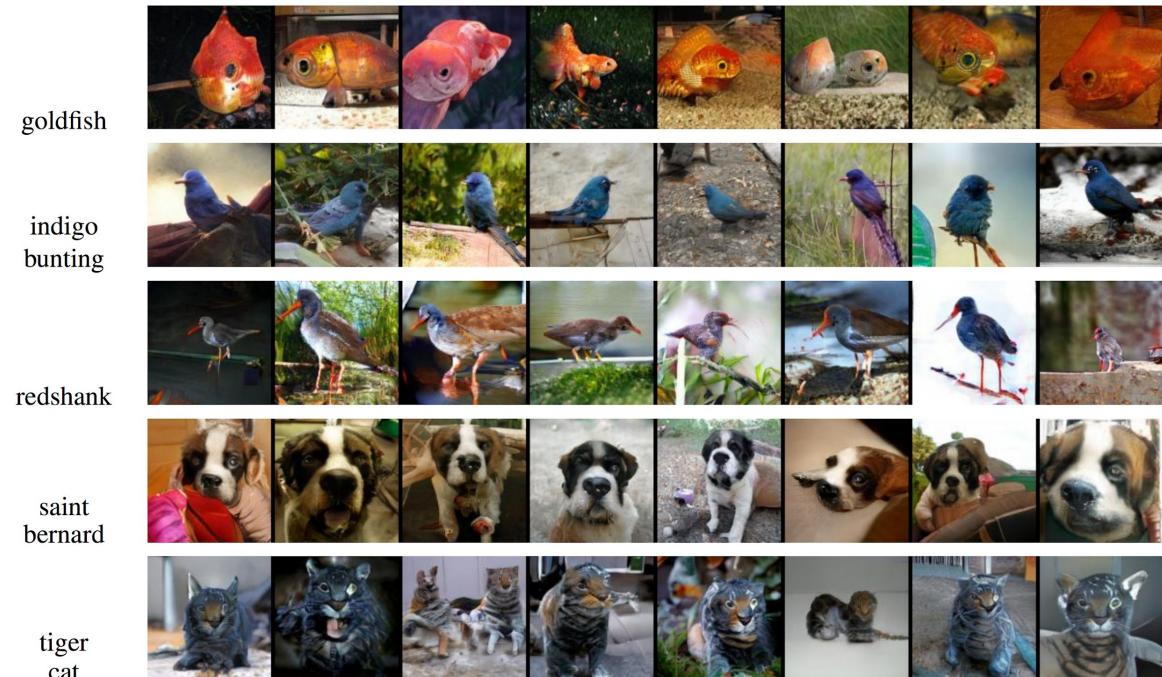
# Self-Attention GANs (SAGAN, Zhang et al.)

- Added **self-attention** to give images better **global structure** and coherence
- Self-attention has had a big impact in a number of domains (especially language modeling, translation)

Want to learn more?



Zhang, et al. Self-Attention Generative Adversarial Networks. International Conference on Machine Learning (2019)



# BigGANs (Brock et al.)

- Make GANs really big
  - Big batches
  - Big models
  - Big datasets
  - Big (high res) images
- Trained on ImageNet (1.2M images) and JFT (300M images)

Want to learn more?



Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations (2019)



# BigGANs (Brock et al.)

- Large empirical study to build a reliable recipe for large scale GAN training, including:
  - Hinge loss in D
  - Spectral norm
  - Self-attention
  - Projection disc
  - Orthogonal regularisation
  - "Skip connections" from noise
  - Class label embedding shared across layers

Want to learn more?



Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations (2019)



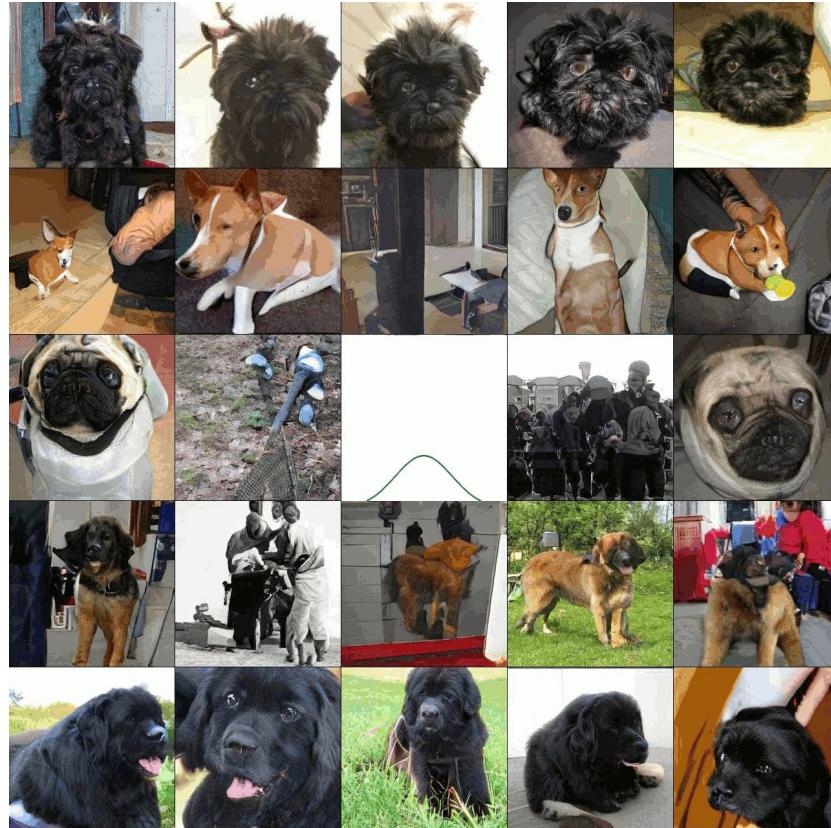
# BigGANs (Brock et al.)

- Introduced the **truncation trick**
- Change the scale of the noise  $z$  input to the generator
- Make the noise smaller (truncate) to increase image fidelity
  - Generates prototypical examples of each class
- Make the noise larger to increase variety
  - Generates the full class distribution

Want to learn more?



Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations (2019)



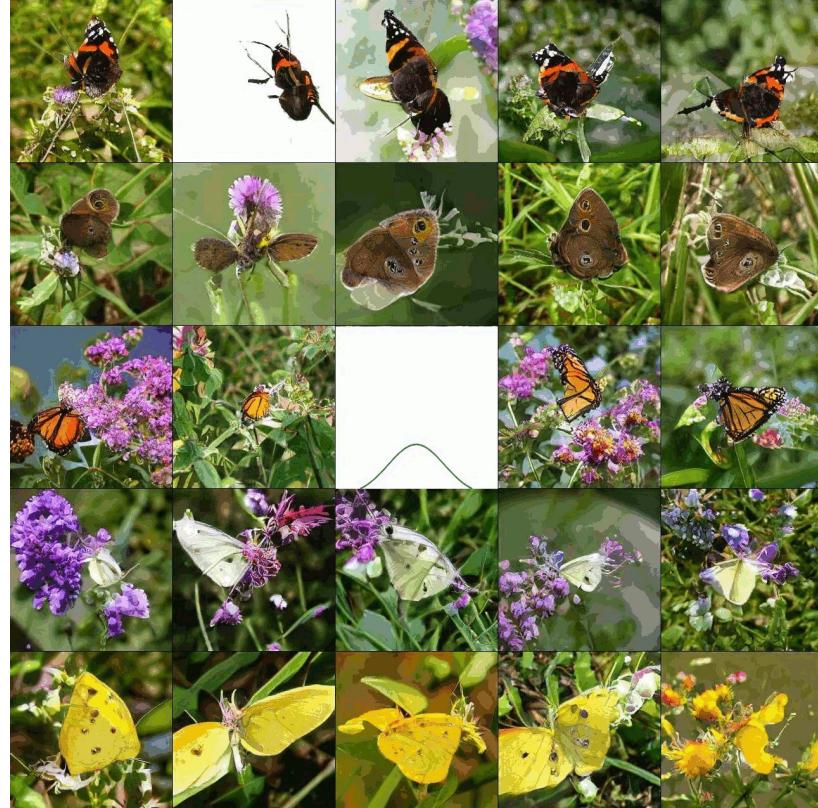
# BigGANs (Brock et al.)

- Introduced the **truncation trick**
- Change the scale of the noise  $z$  input to the generator
- Make the noise smaller (truncate) to increase image fidelity
  - Generates prototypical examples of each class
- Make the noise larger to increase variety
  - Generates the full class distribution

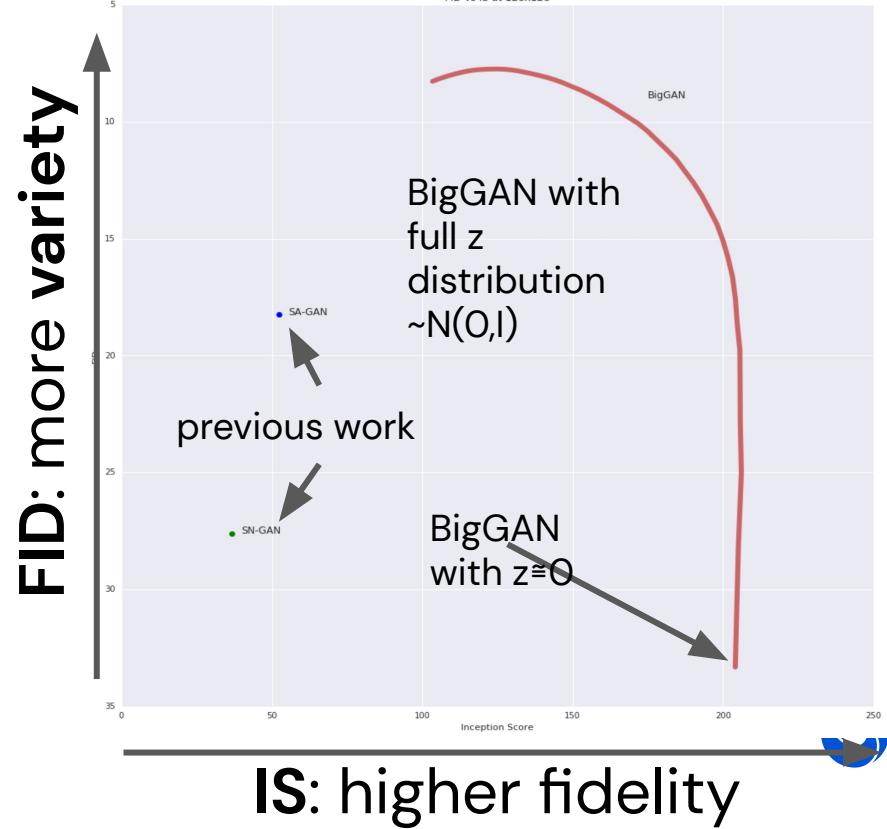
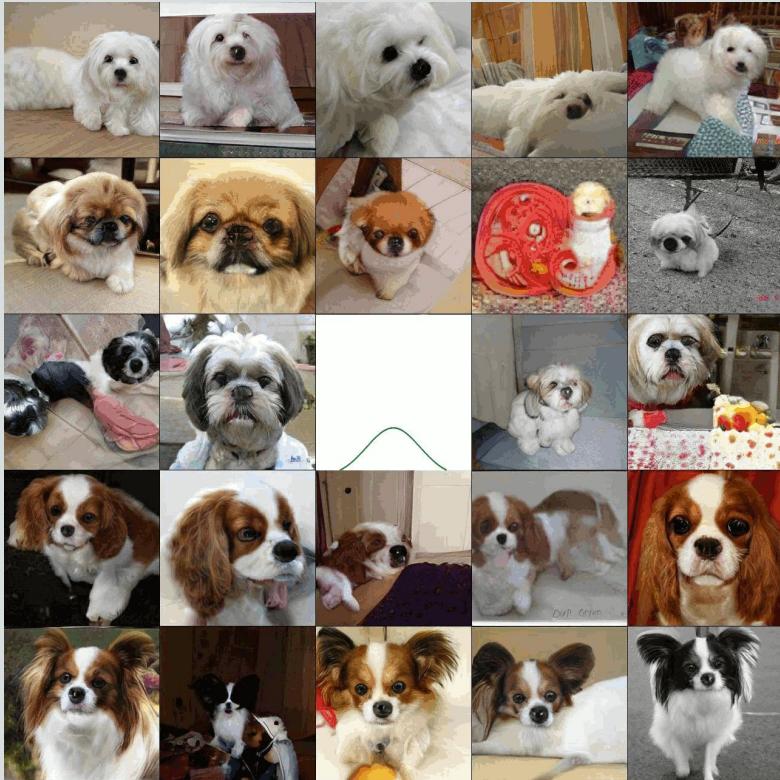
Want to learn more?



Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations (2019)



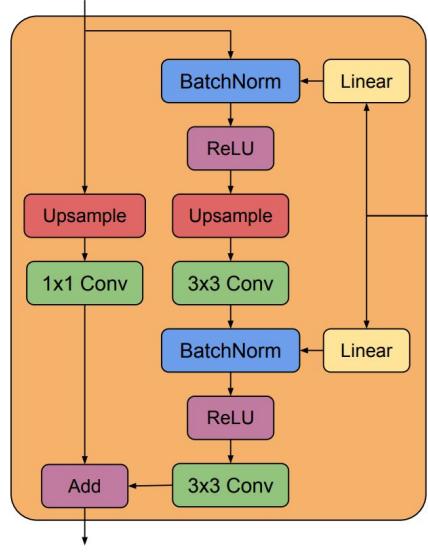
# BigGANs (Brock et al.)



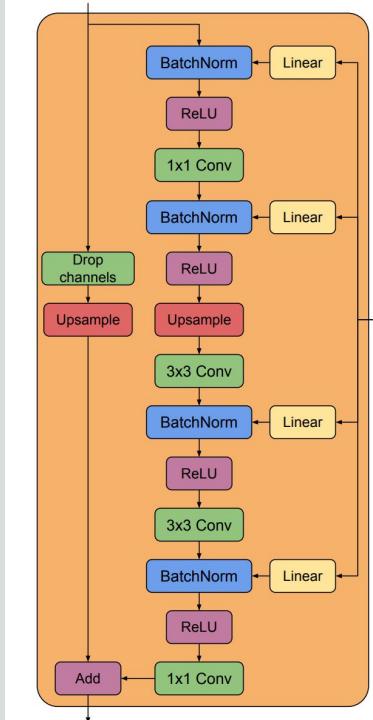
# BigGANs (Brock et al.)

4x deeper, but more efficient!

BigGAN (original) Block



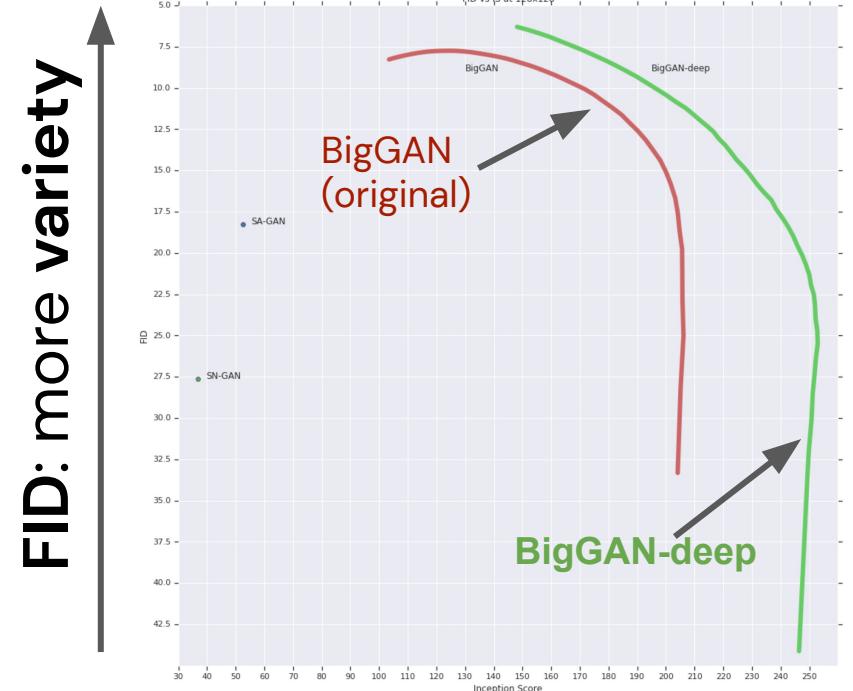
BigGAN-deep Block



Want to learn more?



Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations (2019)



IS: higher fidelity

# BigGANs (Brock et al.): failure modes

Want to learn more?



Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations (2019)



# LOGAN (Wu et al.)

- Uses **latent optimisation** to improve the adversarial dynamics between G & D
  - Natural gradient descent to optimise G's latent inputs
- Results in significant further improvements in BigGAN terms of fidelity and variety



BigGAN-deep  
IS = 259.4  
FID = 27.97

Want to learn more?



Wu, et al. LOGAN: Latent Optimisation for Generative Adversarial Networks.  
arXiv:1912.00953 (2019)

LOGAN  
IS = 259.9  
FID = 8.19



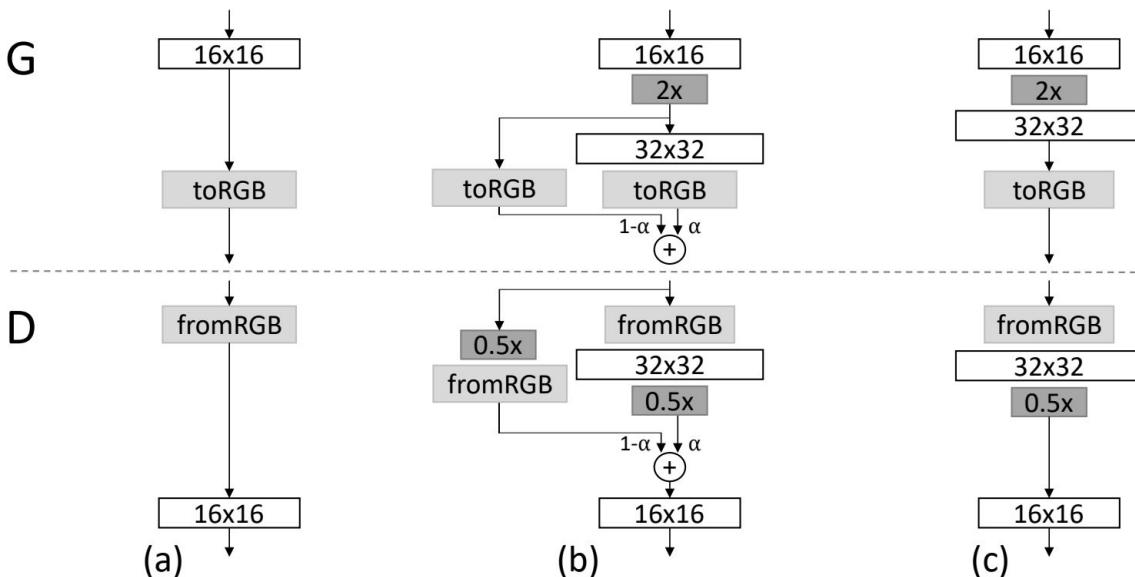
Want to learn more?



Karras, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation. International Conference on Learning Representations (2018)

# Progressive GANs (Karras et al.)

- First, train a GAN to generate tiny ( $4 \times 4$ ) images
- After convergence, add a new layer (in G & D) to generate  $8 \times 8$  resolution images
- Repeat for  $16 \times 16$ ,  $32 \times 32$ , ...
- Very compelling results in a restricted domain (faces)



# Progressive GANs (Karras et al.)

- First, train a GAN to generate tiny (4x4) images
- After convergence, add a new layer (in G & D) to generate 8x8 resolution images
- Repeat for 16x16, 32x32, ...
- Very compelling results in a restricted domain (faces)



Want to learn more?



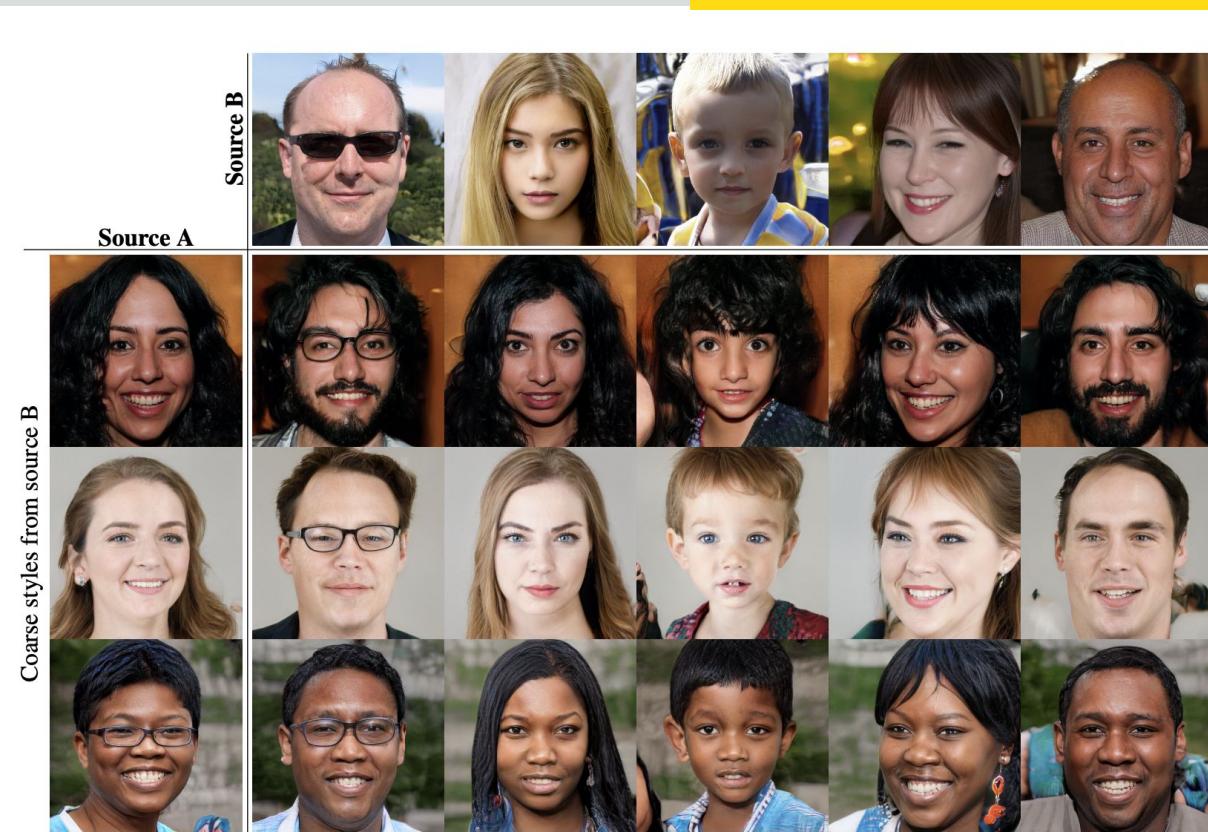
Karras, et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. International Conference on Learning Representations (2018)



# Style GANs (Karras et al.)

StyleGANs were shown to be capable of generating remarkably photorealistic face images

Structured latent inputs ( $\mathbf{z}$ ) to the generator can be used to control its outputs in various interesting ways.



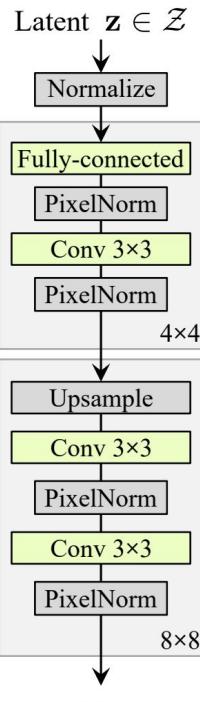
Want to learn more?



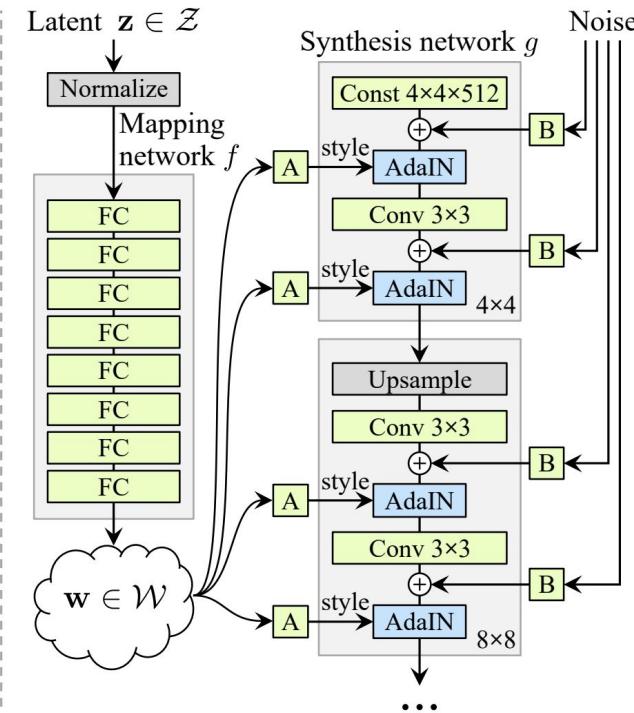
Karras, et al. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Conference on Computer Vision and Pattern Recognition (2019)

# Style GANs (Karras et al.)

- Global latents transformed via an 8 layer MLP
- Incorporates spatial **pixel noise** at each layer
  - Single-channel "image" of noise
  - Broadcast via learnt per-channel scaling factors
- Model learns to associate global latents with the overall **style** of the image
  - Pixel noise modulates the local appearance



(a) Traditional



(b) Style-based generator

Want to learn more?

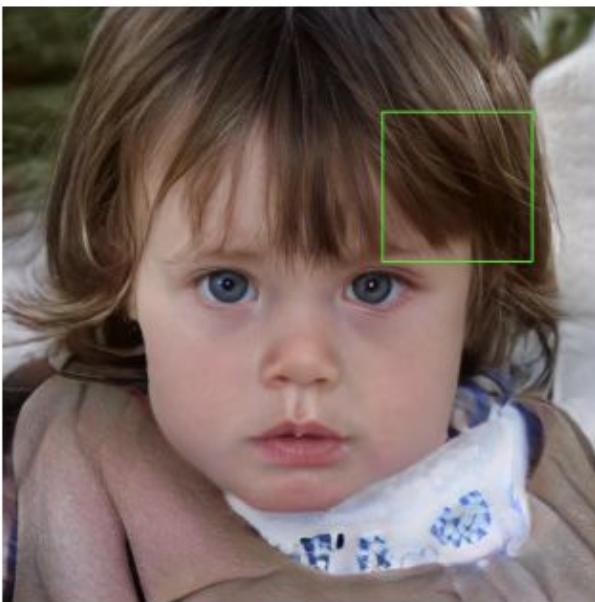


Karras, et al. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Conference on Computer Vision and Pattern Recognition (2019)



# Style GANs (Karras et al.)

- Global latents transformed via an 8 layer MLP
- Incorporates spatial **pixel noise** at each layer
  - Single-channel "image" of noise
  - Broadcast via learnt per-channel scaling factors
- Model learns to associate global latents with the overall **style** of the image
  - Pixel noise modulates the local appearance



(a) Generated image



(b) Stochastic variation

Want to learn more?



Karras, et al. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Conference on Computer Vision and Pattern Recognition (2019)



# Takeaways: Image Synthesis

- Rapid progress scaling up GANs from simple images (MNIST) to large-scale databases of high resolution images (ImageNet, Flickr Faces HQ)
- Improvements from a variety of sources
  - G & D architectures
  - Conditioning
  - Normalisation
  - D parametrization
  - Latent space structure
  - Loss functions
  - Algorithmic

1	3	9	3	9
1	1	0	6	0
0	1	9	1	2
6	3	2	0	8

Goodfellow et al. (2014)



Denton et al. (2015)



Radford et al. (2016)



Miyato et al. (2018)



Miyato et al. (2018)



Zhang et al. (2019)



Brock et al. (2019)



Karras et al. (2019)



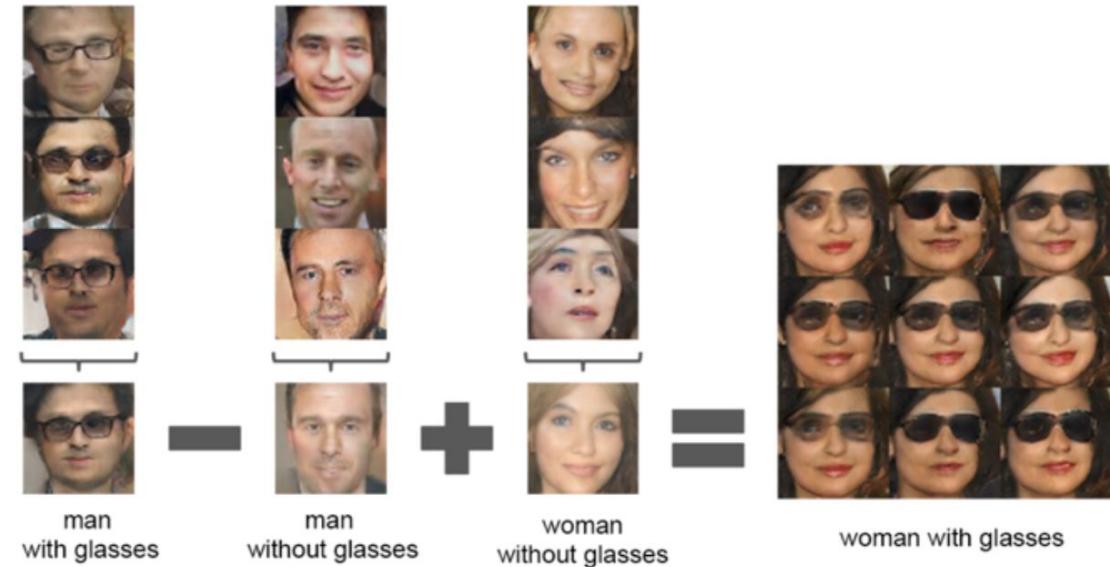
3.2

# GANs for Representation Learning



# Motivating Example #1: Semantics in DCGAN Latent Space (Radford et al.)

- The DCGAN generator's noise/latent space appears to have **meaningful semantics**



Want to learn more?

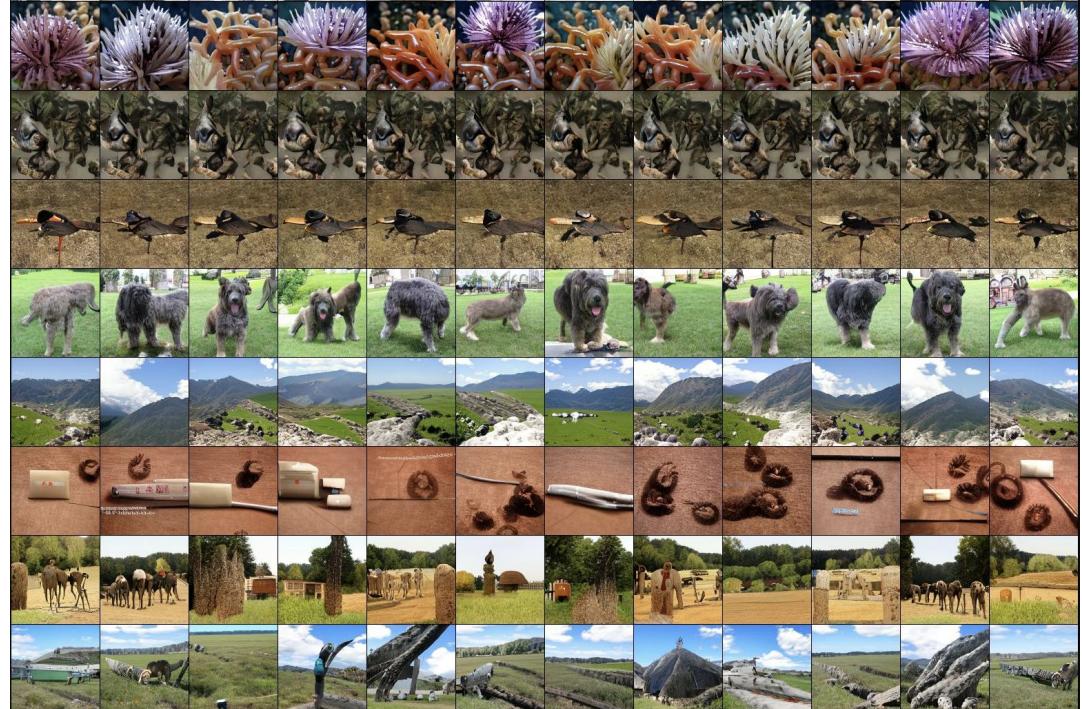


Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. International Conference on Learning Representations (2016)



# Motivating Example #2: Unsupervised Category Discovery with BigGANs

- **Unsupervised** BigGAN trained on unlabeled ImageNet learns to associate a discrete latent variable with interesting semantics
  - Qualitatively, the learnt clusters often resemble image categories
- This model was trained with a combination of discrete and continuous latents:
  - 120D Gaussian ( $N(0, 1)$ )
  - 1024-way uniform categorical
- Rows correspond to categorical values, columns to Gaussian values



[Unpublished Results]



# InfoGANs (Chen et al.)

- Information maximising GANs
- Adds an inference network to recover the latent codes  $\mathbf{z}$  given the generator output  $G(\mathbf{z})$

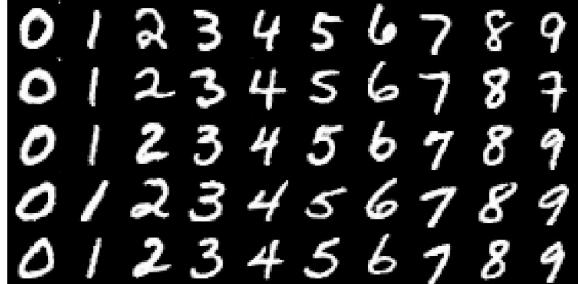
$$\min_G \max_D V_I(D, G) = V(D, G) - \lambda I(c; G(z, c))$$

- With information maximising objective, generator learns (unsupervised) to associate a discrete (10-way categorical) latent variable with digit category

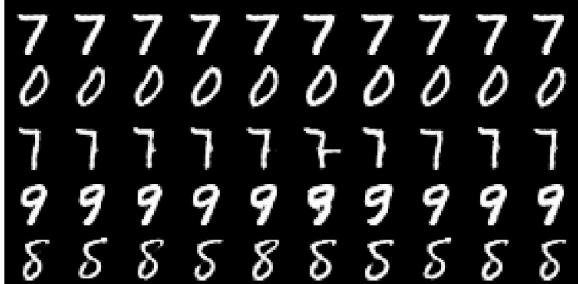
Want to learn more?



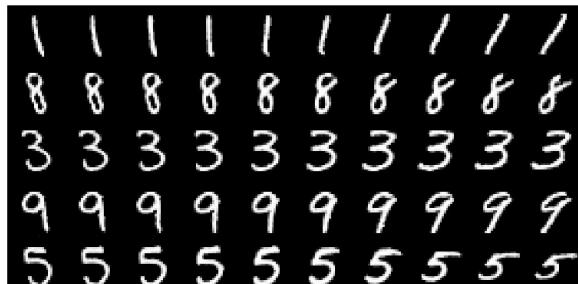
Chen, et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. arXiv:1606.03657 (2016)



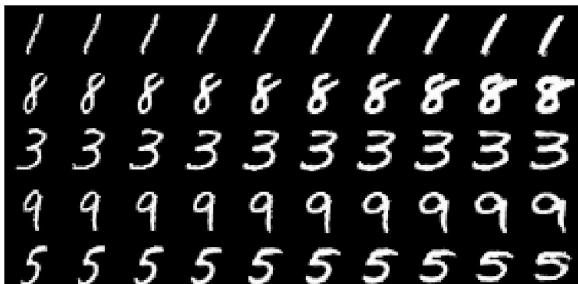
(a) Varying  $c_1$  on InfoGAN (Digit type)



(b) Varying  $c_1$  on regular GAN (No clear meaning)



(c) Varying  $c_2$  from  $-2$  to  $2$  on InfoGAN (Rotation)



(d) Varying  $c_3$  from  $-2$  to  $2$  on InfoGAN (Width)



# ALI / Bidirectional GANs

## (Dumoulin et al., Donahue et al.)

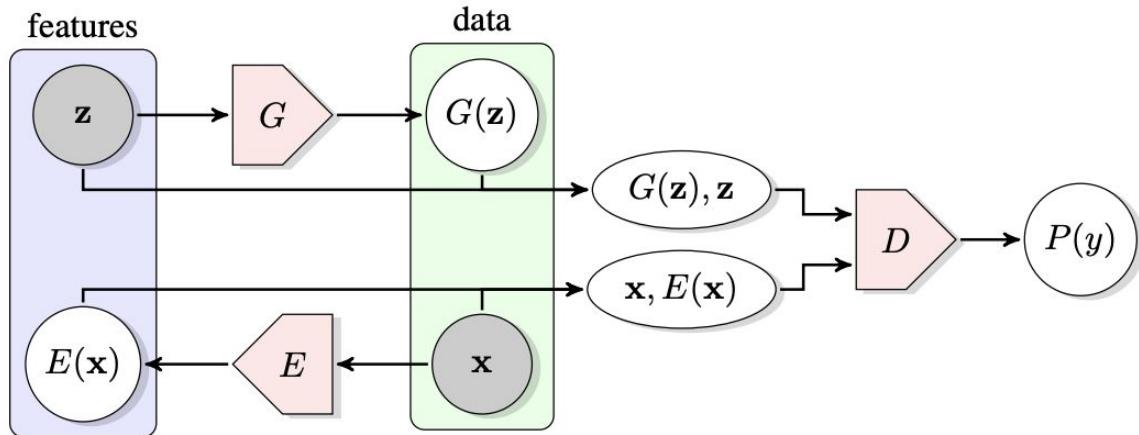
- Adversarial approach to feature representation learning and inference
- Adds an **encoder** network ( $E$ ) which learns the inverse mapping from  $G$ , mapping from data  $x$  to latents  $z$
- The **joint discriminator** sees tuples  $(x, z)$

Want to learn more?



Dumoulin, et al. *Adversarially Learned Inference*. International Conference on Learning Representations (2017)

Donahue, et al. *Adversarial Feature Learning*. International Conference on Learning Representations (2017)



# ALI / Bidirectional GANs

## (Dumoulin et al., Donahue et al.)

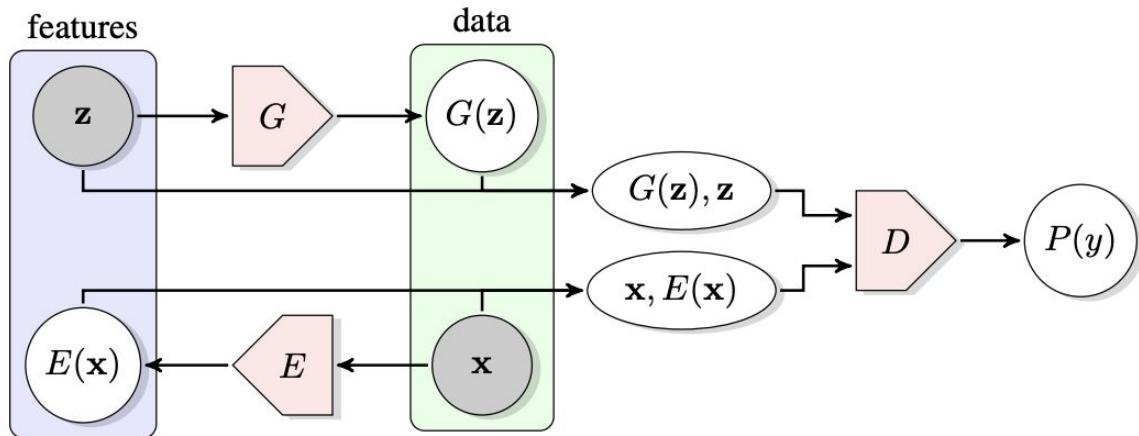
Want to learn more?



Dumoulin, et al. Adversarially Learned Inference. International Conference on Learning Representations (2017)

Donahue, et al. Adversarial Feature Learning. International Conference on Learning Representations (2017)

- The joint discriminator sees tuples  $(x, z)$ 
  - $z \sim P_{z'}, x = G(z)$
  - $x \sim P_{x'}, z = E(x)$
- In the global optimum,  $E$  and  $G$  are inverses; for all  $x$  and  $z$  we have
  - $x = G(E(x))$
  - $z = E(G(z))$



# ALI / Bidirectional GANs

## (Dumoulin et al., Donahue et al.)

- In the global optimum,  $E$  and  $G$  are inverses; for all  $x$  and  $z$  we have
  - $x = G(E(x))$
  - $z = E(G(z))$
- In practice, this inversion property does not hold perfectly
  - But reconstructions still often capture interesting semantics

Want to learn more?



Dumoulin, et al. Adversarially Learned Inference. International Conference on Learning Representations (2017)

Donahue, et al. Adversarial Feature Learning. International Conference on Learning Representations (2017)



(a) CelebA samples.



(b) CelebA reconstructions.



# BigBiGANs (Donahue et al.)

- BiGANs at scale:  
**BigBiGANs** are BiGANs trained using the BigGAN G and D architectures
- ResNet-style encoders E
- Reconstructions exhibit clear high-level semantics of the input images (despite being unsupervised), while clearly not being memorised copies

Want to learn more?



Donahue, et al. Large Scale Adversarial Representation Learning. Neural Information Processing Systems (2019)

real data  $\mathbf{x}$  (128x128)



BigBiGAN reconstructions  $G(E(\mathbf{x}))$



# BigBiGANs (Donahue et al.)

- BiGANs at scale:  
**BigBiGANs** are BiGANs trained using the BigGAN G and D architectures
- ResNet-style encoders E
- Reconstructions exhibit clear high-level semantics of the input images (despite being unsupervised), while clearly not being memorised copies

Want to learn more?



Donahue, et al. Large Scale Adversarial Representation Learning. Neural Information Processing Systems (2019)



BigBiGAN reconstructions  $G(E(\mathbf{x}))$



# BigBiGANs (Donahue et al.)

- BigBiGAN encoder learns ImageNet representations competitive with other unsupervised / self-supervised approaches
- Nearest neighbors (right) in BigBiGAN encoder feature space show the semantics present in the learnt representations



Want to learn more?



Donahue, et al. Large Scale Adversarial Representation Learning. Neural Information Processing Systems (2019)



3.3

# GANs for Other Modalities & Problems



# Pix2Pix (Isola et al.)

- Train a generator to **translate** between images of two different domains
- Standard GAN objective combined with reconstruction error

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_{x,z}[\log(1 - D(G(x, z)))].$$

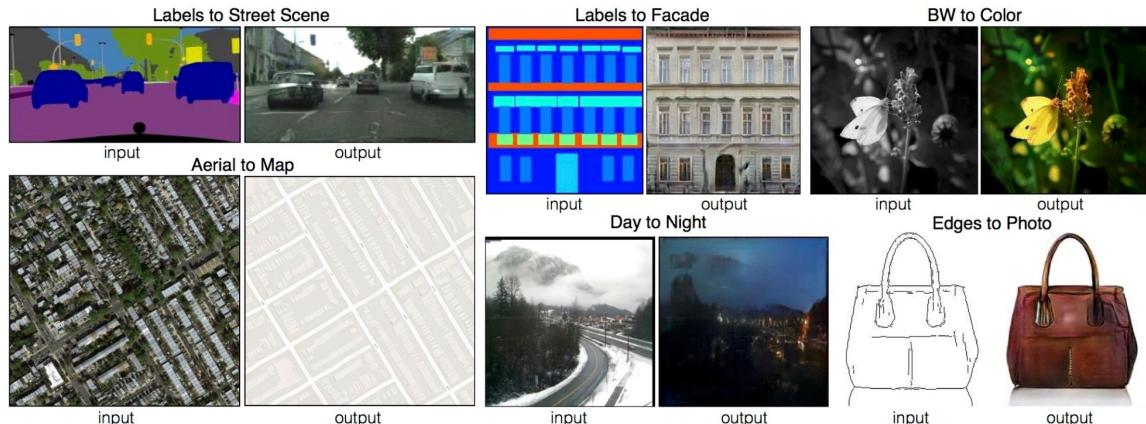
$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1].$$

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G).$$

Want to learn more?



Isola, et al. *Image-to-Image Translation with Conditional Adversarial Networks*. IEEE Conference on Computer Vision and Pattern Recognition (2017)



Example results on several image-to-image translation problems. In each case we use the same architecture and objective, simply training on different data.



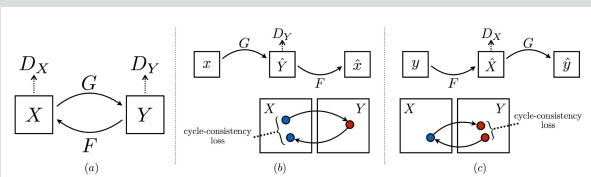
# CycleGAN (Zhu et al.)

Want to learn more?

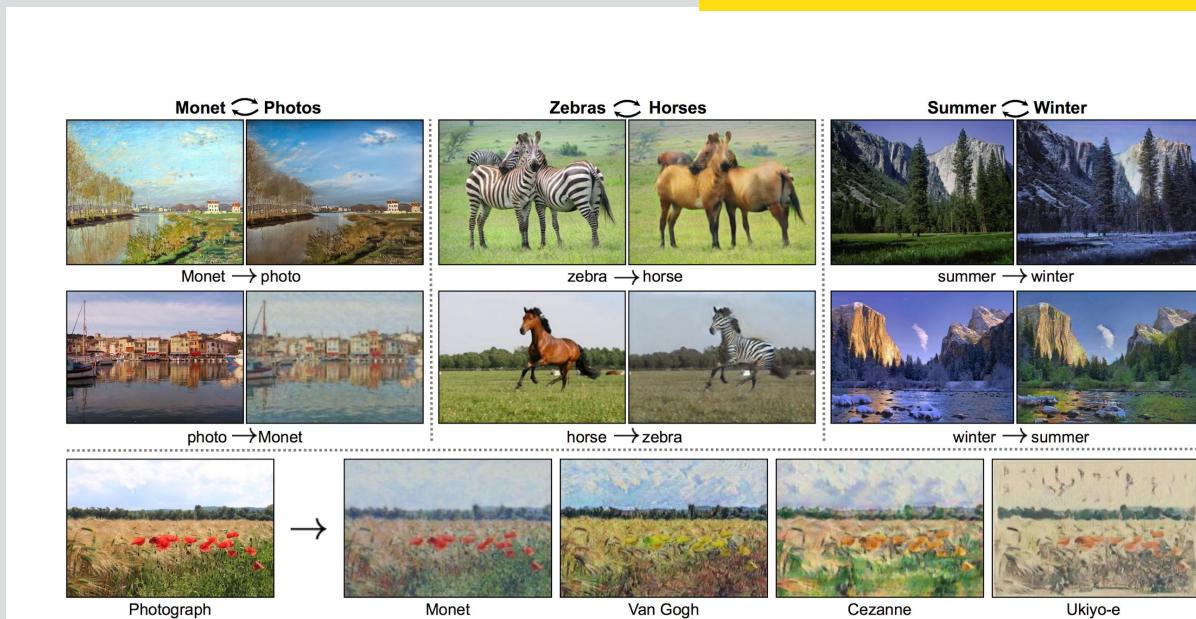


Zhu, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. International Conference on Computer Vision (2017)

- Train a generator to **translate** between images of two different domains
- But **without any paired samples!**

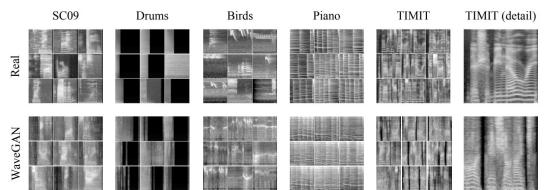


- Enforces **cycle consistency**:
  - Image  $x$  in domain A
  - Translate to domain B
  - Back to domain A  $\rightarrow x'$
  - Enforce  $x \approx x'$

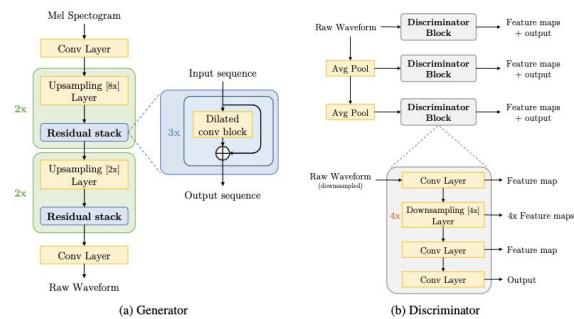


# GANs for Audio Synthesis

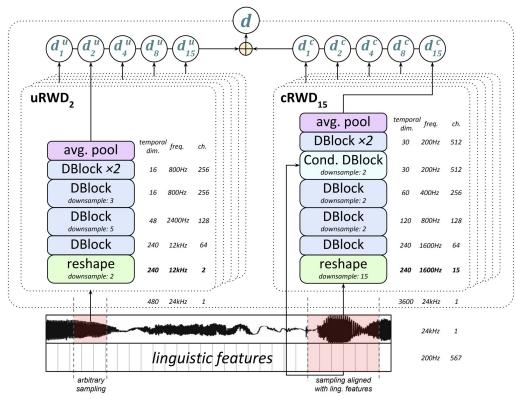
## WaveGAN (C. Donahue et al.)



## MelGAN (Kumar et al.)



## GAN-TTS (Bińkowski et al.)



Want to learn more?



C. Donahue et al. **Adversarial Audio Synthesis**. International Conference on Learning Representations (2019)

Want to learn more?



Kumar et al. **MeLGAN: Generative Adversarial Networks for Conditional Waveform Synthesis**. Neural Information Processing Systems (2019)

Want to learn more?



Bińkowski et al. **High Fidelity Speech Synthesis with Adversarial Networks**. International Conference on Learning Representations (2020)

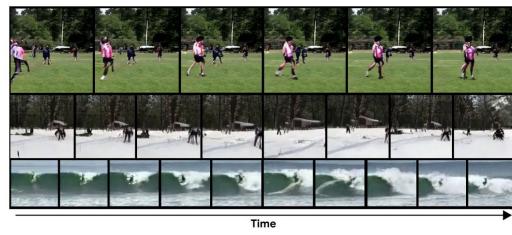


# GANs for Video Synthesis & Prediction

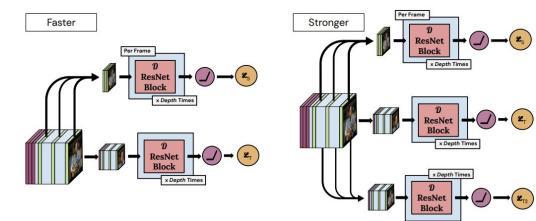
TGAN-v2 (Saito & Saito)



DVD-GAN (Clark et al.)



TriVD-GAN (Luc et al.)



Want to learn more?



Saito and Saito. **TGANv2: Efficient Training of Large Models for Video Generation with Multiple Subsampling Layers**. arXiv:1811.09245 (2018)

Want to learn more?



Clark et al. **Adversarial Video Generation on Complex Datasets**. arXiv:1907.06571 (2019)

Want to learn more?

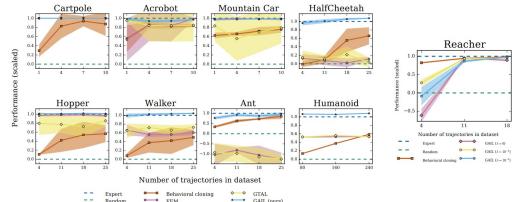


Luc et al. **Transformation-based Adversarial Video Prediction on Large-Scale Data**. arXiv:2003.04035 (2020)



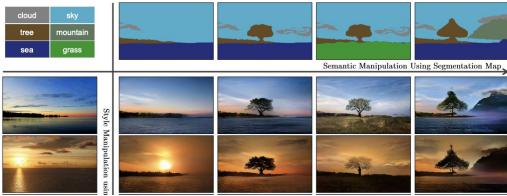
# GANs Everywhere!

## RL (Imitation Learning): GAIL

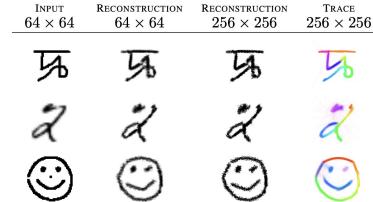


Ho and Erman. *Generative Adversarial Imitation Learning*. Neural Information Processing Systems (2016)

## Image Editing: GauGAN



## Program Synthesis: SPIRAL



Park et al. *Semantic Image Synthesis with Spatially-Adaptive Normalization*. IEEE Conference on Computer Vision and Pattern Recognition (2019)

## Motion Transfer: Everybody Dance Now



Chan et al. *Everybody Dance Now*. International Conference on Computer Vision (2019)

## Domain Adaptation: DANN



Ganin et al. *Domain-Adversarial Training of Neural Networks*. Journal of Machine Learning Research (2016)

## Art: Learning to See



Akten. *Learning To See*. <http://www.memo.tv/portfolio/learning-to-see/> (2017, accessed 2020)



3.4

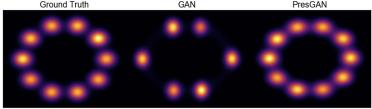
# More GANs at DeepMind

(Updated: 1 July 2020)



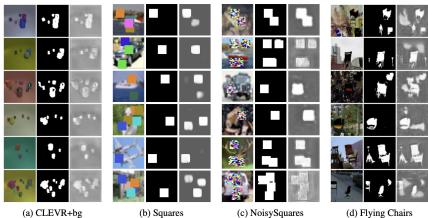
# More GANs at DeepMind (2019)

## PresGAN



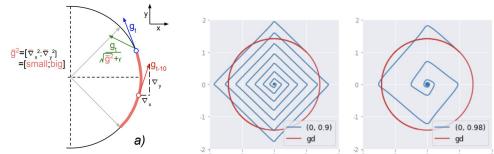
Dieng et al. **Prescribed Generative Adversarial Networks.** *arXiv:1910.04302* (2019)

## Copy-Pasting GAN



Arandjelović & Zisserman. **Object Discovery with a Copy-Pasting GAN.** *arXiv:1905.11369* (2019)

## Effectiveness of Adam on Cycles



Gemp & McWilliams. **The Unreasonable Effectiveness of Adam on Cycles.** *NeurIPS Smooth Games Optimization and Machine Learning Workshop* (2019)

## ScratchGAN



She's that result she believes that for Ms . Marco Rubio 's candidate and that is still become smaller than ever .

I hadn 't been able to move on the surface – if grow through , she said , given it at a time later that time .

If Iran wins business you have to win ( Iowa ) or Hillary Clinton 's survived nothing else since then , but also of all seeks to bring unemployment .

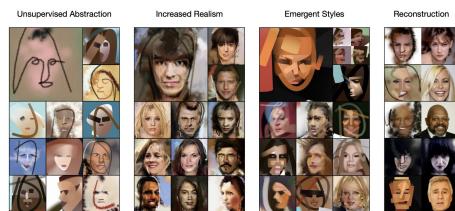
All the storm shows is incredible , most of the kids who are telling the girls the people we 're not turning a new study with a challenging group .

Six months before Britain went the EU we will benefit from the EU - it is meeting by auto , from London , so it 's of also fierce faith Freedom .



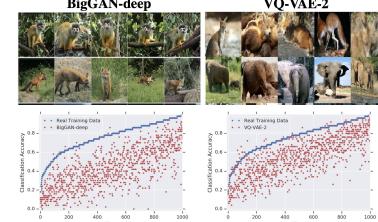
de Masson d'Autume et al. **Training language GANs from Scratch.** *Neural Information Processing Systems* (2019)

## Improved SPIRAL



Mellor et al. **Unsupervised Doodling and Painting with Improved SPIRAL.** *arXiv:1910.01007* (2019)

## Classification Accuracy Score

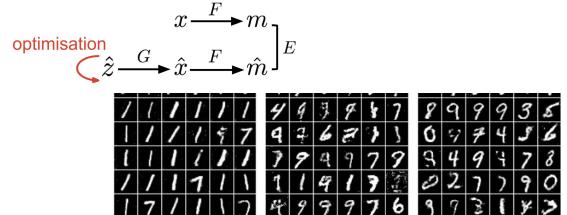


Ravuri & Vinyals. **Classification Accuracy Score for Conditional Generative Models.** *Neural Information Processing Systems* (2019)



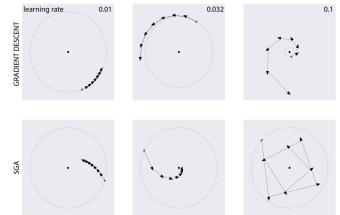
# More GANs at DeepMind (2017-19)

## Deep Compressed Sensing



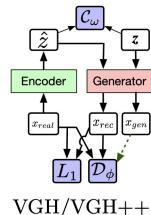
Wu et al. Deep Compressed Sensing. *International Conference on Machine Learning* (2019)

## n-Player Differentiable Games



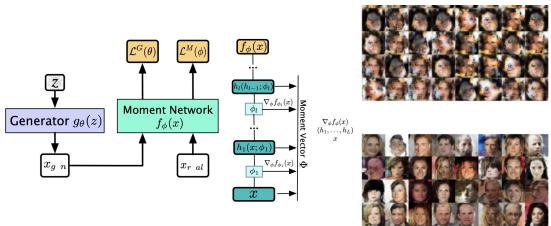
Baldazzi et al. The Mechanics of n-Player Differentiable Games. *International Conference on Machine Learning* (2018)

## Variational GAN Hybrids



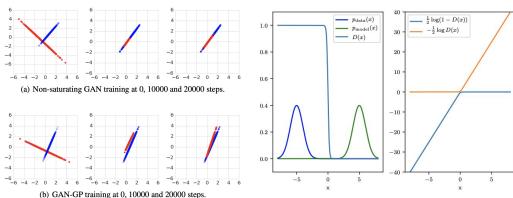
Rosca et al. Distribution Matching in Variational Inference. *arXiv:1802.06847* (2018)

## Method of Learned Moments



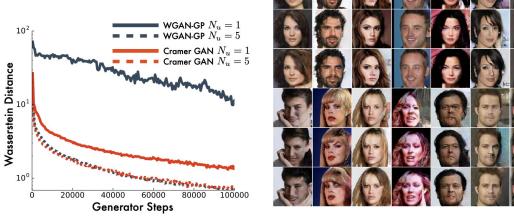
Ravuri et al. Learning Implicit Generative Models with the Method of Learned Moments. *International Conference on Machine Learning* (2018)

## Many Paths to Equilibrium



Fedus et al. Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step. *International Conference on Learning Representations* (2017)

## Cramér GAN

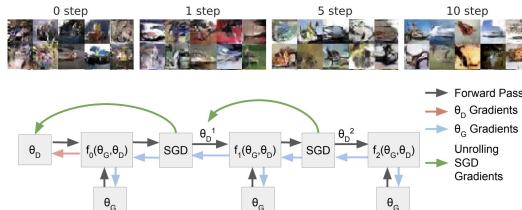


Bellemare et al. The Cramér Distance as a Solution to Biased Wasserstein Gradients. *arXiv:1705.10743* (2017)



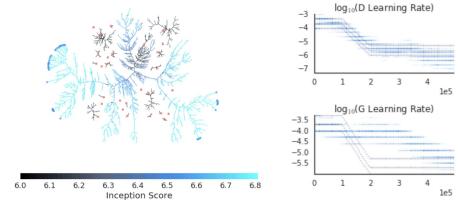
# More GANs at DeepMind (2016-17)

## Unrolled GAN



Metz et al. **Unrolled Generative Adversarial Networks**. International Conference on Learning Representations (2017)

## Population Based Training



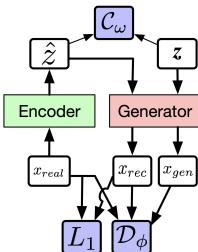
Jaderberg et al. **Population Based Training of Neural Networks**. arXiv:1711.09846 (2017)

## Likelihood vs. GAN NVP Training



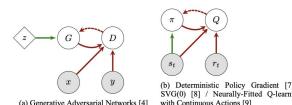
Danihelka et al. **Comparison of Maximum Likelihood and GAN-based training of Real NVPs**. arXiv:1705.05263 (2017)

## $\alpha$ -GAN



Rosca et al. **Variational Approaches for Auto-Encoding Generative Adversarial Networks**. arXiv:1706.04987 (2017)

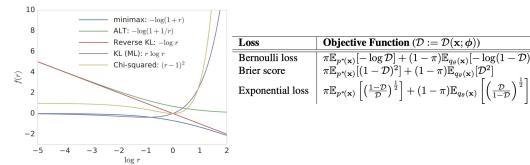
## Connecting GANs & Actor-Critic



Method	GANs	AC
Freezing learning	yes	yes
Label smoothing	yes	no
Historical averaging	yes	no
Minibatch discrimination	yes	no
Batch normalization	yes	yes
Target networks	n/a	yes
Replay buffers	no	yes
Entropy regularization	no	yes
Compatibility	no	yes

Pfau & Vinyals. **Connecting Generative Adversarial Networks and Actor-Critic Methods**. NeurIPS Workshop on Adversarial Training (2016)

## Learning in Implicit Gen. Models



Mohamed & Lakshminarayanan. **Learning in Implicit Generative Models**. arXiv:1610.03483 (2016)



Thank you

